Model reduction, mechanistic modelling and transience in models of stochastic chemical kinetics

James Holehouse

THE UNIVERSITY
of EDINBURGH

Doctor of Philosophy

The University of Edinburgh

2022

Dedicated to my Dad,
who sparked my scientific curiosity
and who rarely failed to answer a 'why?'

Abstract

Now, it is long known that gene expression and chemical kinetics are subject to random fluctuations. These lead to deviations from deterministic models that do not account for the random nature of biochemical kinetics. Successfully incorporating these stochastic dynamics is of great interest so that one can better model, and more closely understand, the intricate phenomena inherent in biological mechanisms. Many previous studies have been conducted in modelling such processes stochastically, for instance processes such as genetic autoregulation, Michaelis-Menten enzyme action and ant recruitment models. However, the majority of these studies explore only the steady state solutions of such processes while assuming mass-action kinetics, without considering: (1) extrinsic noise, (2) transience from an initial condition, or even (3) the finite, non-continuous nature of molecule or agent numbers.

This thesis focuses on the aforementioned complex systems, with an emphasis on how to use toy models in responsible and informed ways. *Responsible* refers to a knowledge of how good our approximations of microscopic dynamics are and their limitations: Do we understand the assumptions that commonly employed approximations rely on? *Informed* refers to whether a model we design is sufficiently minimal or complex to represent the underlying biochemical (or economical) kinetics: Can we use alternative models of similar simplicity (possibly mechanistically informed) to more properly capture the dynamics of the system we are attempting to model? Further issues pursued in this thesis are whether common approximative methods can be extended to effectively include details of more complex underlying dynamics, or whether we can move beyond typical steady state solutions and explore transience from an initial condition.

There are several main findings from our studies. We find that for non mass-action Hill-type propensities, often used in biochemical kinetics, that typically only assume time scale separation as the basis of approximation, that finite molecule number effects can greatly perturb their accuracy. Then, we show that the addition of non-Gaussian colored noise to biochemical rate parameters can capture intricate characteristics of gene expression that are not explicitly modelled. For common two-state gene models, we explore why they seem to be so effective at approximating gene expression, where it is known that several key rate limiting steps are ignored. Finally, we develop transient solutions to master equations describing Michaelis-Menten enzyme kinetics and ant recruitment, and we show how to extend the solutions therein to more general forms.

Lay Summary

In order to understand the world around us, it is often convenient to construct *mathematical models* that allow one to make predictions about the future, or to understand why something has previously occurred. Broadly speaking, mathematical models come in two forms: those which include random behaviours, and those that do not. Newton's theory of gravity is an example of a model that does not include random behaviour. This is appropriate since planets and stars in space generally assume motion that is effectively unperturbed by random collisions with the low densities of molecules in the vacuum of space. Their motion is *deterministic*, meaning that if we know the position and velocity of a planet around a star, then we know everything about its future evolution. But what about the motion of an air molecule that bumps into many surrounding air molecules? Or the motion of a cell traversing a surface? Or the movement of people in a crowd? Knowing the position and velocity of an air molecule at one time will not give you anything other than a probabilistic measure of where it will be in the future as generally we do not know (and could not comprehend) the positions and velocities of all surrounding molecules. Therefore, the models we construct must take into account the complicated nature of interactions between molecules, cells and people.

To describe these processes we do not consider each individual agent interacting with all other agents (as this quickly becomes a computational nightmare!), but instead we consider simplified systems that capture only the relevant and essential biology or physics, whilst still giving us an intuition to the real-world system. The simplification of such systems allows us to gain insights into the behaviours that arise due to properties of interest, meaning that we do not become confused by the cacophony of contributing factors from all aspects of the real-world system. However, it is not always so clear which simplified model truly captures the essence of the real-world system of interest. Or indeed, what are the true limitations on the approximations we make? Further to this, many aspects of interacting systems are typically ignored in favour of more mathematically or computationally accessible ones.

Of interest in this thesis are three separate complex systems in each of which random behaviour is a key property of the phenomena observed: (1) gene expression, (2) enzyme kinetics, and (3) ant recruitment. In gene expression, the process by which genes produce mRNA and proteins inside every cell in every living organism, the interest is in how the random binding events between genes, mRNA and proteins affect the regulation of gene products in a cell. For example, a special property of auto-regulation, which is the

ability of a gene to make itself produce more or less of its own mRNA and proteins, is that often two modes of behaviour are observed—a low production state and a high production state. Genes that have this property are then capable of performing specific functions in the gene regulatory network, hence understanding how these behaviours arise is of vital interest. In enzyme kinetics, it is key to understand the efficiency of product formation such that engineers can make maximal amounts of product in short periods of time. Finally, ant recruitment, whereby ants are recruited by other ants to work different food sources, is an important toy model due to its implications for human herding behaviour. Particularly, it has been found that where imitation between ants is strong enough that multiple behavioural modes can arise solely due to random fluctuations in the system.

The work conducted in this thesis contributes to the above in three ways. (1) The limitations of common approximations to genetic auto-regulation are explored. (2) Mapping complex models to simpler ones using systematic methods (again in the context of gene expression). (3) Assessing a common toy model of gene expression and asking whether it truly captures the important dynamics of gene state change. (4) Time-dependence in systems of enzymes and ants are studied, where the initial state of a system impacts the behaviours one observes on long time scales.

Acknowledgements

First, I would like to thank my supervisor Ramon Grima for being a constant source of good ideas and for always being available for either scientific discussion, personal consultation, or else a good laugh. I would also like to thank my co-supervisor Meriem El Karoui, for not only being a figure of scientific rigour, but additionally for her strong emphasis on non-work matters such as mental health, which during the covid-years this has been of vital importance. I additionally thank Diego Oyarzún for being the chair of my PhD committee, and Richard Blythe for providing necessary advice at a crucial time.

Thanks also go to the members of the research groups of Ramon Grima and Meriem El Karoui, office mates in the C.H. Waddington building, and the EASTBIO 2018 cohort, for making the PhD such a joy, and for providing great discussion, both of scientific and non-scientific calibre. In particular, I would like to thank Svitlana Braichenko, Abhishek Gupta, Juraj Szavits-Nossan, Louis Headley, Rodrigo Garcia, Lyndsay Kerr, Xiaoming Fu, Sam Haynes, Alessia Lepore, Xavier Zaoui, Sebastian Jaramillo, Ira Iosub, James Broughton, Lorna McLaren, Wei Li, Connor Bruce, Randeep Samra, Liat Adler, Alice Scarpa and Christopher Jennings. Special thanks go to Kaan Öcal, Tatiana Filatova, Zhixing Cao, Irina Kalita and Augustinas Sukys for aiding me in various computational, mathematical and biological aspects of my PhD. I pay particular thanks to the EASTBIO cohort before covid struck, for several very good training and 'social' sessions.

I would like to thank Cambridge Econometrics for hosting me on my 4-month NPIF internship, in particular to Hector Pollitt and Pim Vercoulen for being great hosts. Additionally, thanks to José Moran for further engaging my foray into discrete choice models towards the end of my PhD.

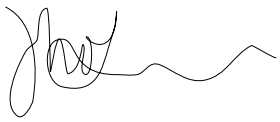
I am indebted to several of my teachers over the history of my education for inspiring me up to the point of writing this thesis, including (but by no means limited to) Hamish Harron, Dave Robinson, Colin Fallows, Sarah Thackray, Kristel Torokoff, Simon Tett and Alexander Morozov.

Finally, I would like to thank those who have most supported me on my PhD. First, to my Mum and Brother, Kathryn and Tom, for always believing in my ability to succeed and always being available for a chat. Second, to my partner Matilde for her extensive proofreading of this thesis, but more importantly: for always pushing me forward, for drowning my self-doubt in praise, and in never letting me settle for less than I am capable. Thank you to my friends from Edinburgh, Bridlington and beyond for many

necessary chill nights and weekends. Also, thanks go out to my extended family for supporting not just myself, but my Mum and Brother during challenging periods in the past few years. Finally, thanks go to my late Father, Michael Holehouse, for initially motivating me to study science, but more importantly for always having a wealth of knowledge from which to draw. It was through watching Horizon documentaries and stargazing with you that made me pursue this path.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.



The candidate confirms that the work submitted is their own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

This thesis contains previously peer-reviewed research, and a single submitted pre-print (* indicates equal contributions to the publication, † indicates the candidate's supervisor Ramon Grima, and ‡ indicates the corresponding author):

[1] **Holehouse, J.** and Grima, R.†,‡, 2019. Revisiting the reduction of stochastic models of genetic feedback loops with fast promoter switching. Biophysical journal, 117(7), pp.1311-1330. *Included in Chapter 3.*

[2] **Holehouse, J.***, Gupta, A.* and Grima, R.†,‡, 2020. Steady-state fluctuations of a genetic feedback loop with fluctuating rate parameters using the unified colored noise approximation. Journal of Physics A: Mathematical and Theoretical, 53(40), p.405601. *Included in Chapter 4.*

[3] Braichenko, S.*, **Holehouse, J.*** and Grima, R.†,‡, 2021. Distinguishing between models of mammalian gene expression: telegraph-like models versus mechanistic models. Journal of the Royal Society Interface, 18(183), p.20210510. *Included in Chapter 5.*

[4] **Holehouse, J.***, Sukys, A.* and Grima, R.†,‡, 2020. Stochastic time-dependent enzyme kinetics: Closed-form solution and transient bimodality. The Journal of Chemical Physics, 153(16), p.164113. *Included in Chapter 6.*

[5] **Holehouse, J.**‡ and Moran, J., 2022. Exact time-dependent dynamics of discrete binary choice models. Journal of Physics: Complexity, 3(3):035005. *Included in Chapter 7.*

The writing and research conducted in the above publications was a collaborative effort between the candidate and the indicated coauthors, on each of which a major part of the writing and research was conducted by the candidate.

Additionally, two other peer-reviewed articles have been published whose results are, for the most part, not included in this thesis:

[6] **Holehouse, J.**, Cao, Z. and Grima, R.[†,‡], 2020. Stochastic modeling of autoregulatory genetic feedback loops: A review and comparative study. Biophysical Journal, 118(7), pp.1517-1525.

[7] **Holehouse, J.**[‡] and Pollitt, H., 2022. Non-equilibrium time-dependent solution to discrete choice with social interactions. PloS one, 17(5), p.e0267083.

James Holehouse
14th October 2022

Contents

Nomenclature

ABS Activator binding site
CFPE Chemical Fokker-Planck equation
CME Chemical master equation
CV Coefficient of variation
dCME Delayed chemical master equation
dSSA Delayed stochastic simulation algorithm
esc Einstein summation convention
FF Fano factor
FPE Fokker-Planck equation
FSP Finite state projection
HD Hellinger distance
LHS Left-hand side
LNA Linear noise approximation
MM Michaelis-Menten
mRNA Messenger RNA
ODE Ordinary differential equation
PDE Partial differential equation
Pol II RNA polymerase II
QEA Quasi equilibrium approximation
QSSA Quasi steady state approximation
RHS Right-hand side
SDE Stochastic differential equation
smFISH Single molecule fluorescence *in situ* hybridisation
SSA Stochastic simulation algorithm (a.k.a. the Gillespie algorithm)
SSE System size expansion
UCNA Unified colored noise approximation
WKB Wentzel-Kramers-Brillouin

In history there are no control groups. There is no one to tell us what might have been. We weep over the might have been, but there is no might have been. There never was. It is supposed to be true that those who do not know history are condemned to repeat it. I don't believe knowing can save us.

Alfonsa, in *All the Pretty Horses* by Cormac McCarthy

Chapter 1

Introduction

1.1 General motivation and overview

You wake up and have the same breakfast as normal, and go to work at the normal time. The coffee in the office is a bit too cold, and because a meeting runs late into lunch all the good food in the cafeteria is gone. Your train home comes unexpectedly on time, and you arrive home in time for your favourite quiz show. Life is undoubtedly a mixture and deterministic and random events. Each day in your life follows a particular pattern, one that likely you have vague expectations for, but fluctuations in this plan make the day unpredictable. Sometimes these fluctuations lead to small changes, but more often than not, large macroscopic changes occur in your day due to the proverbial butterfly flapping its wings.

In certain regimes, deterministic modelling via rate equations provides a tractable way to quantitatively assess systems whose underlying dynamics are truly stochastic. For example, where chemical reactions are conducted in large enough quantities (the large molecule number regime where $N \gg 1$), since the magnitude of molecular fluctuations is typically $\propto \sqrt{N}$, the coefficient of variation scales as $1/\sqrt{N}$, meaning that the larger the system the lesser the effects of noise (where N denotes the molecule number) [8]. In these regimes, approximating molecule number as continuous quantity does not lead to large errors, since molecule numbers are large and $1/\sqrt{N}$ becomes negligible. However, for systems with small numbers of agents, or else where *stochastic multimodality* (multiple modes of behaviour whose origin is due to noise) occurs, the deterministic description is often not enough, and to rely on it results in unnecessary errors.

That fluctuations are important in a multitude of complex systems has paved the way for individual realisations that stochastic effects must be considered in many fields. In gene expression this was elucidated by Elowitz *et. al* [9, 10], who emphasised that noise (fluctuations in molecule numbers) puts limitations on the precision of gene expression due to low copy numbers of genes, mRNA and proteins. They broke down the extrinsic (noise due to cell-to-cell variability) and intrinsic (noise due to low copy number effects still observed in homogeneous populations of cells) noises observed in gene expression, and they provided a quantitative framework for modelling noise in gene

expression. By constructing an experiment that allows for the the measurement of two different fluorescent proteins, it was determined that gene expression noise is not always largely extrinsic, but that in many cases intrinsic noise is a non-negligible factor. This important finding made it clear that variation in gene expression is not entirely due to cell-to-cell variability, but that individual cells have fluctuations in their own gene expression profiles. The implications of this study have been far reaching. In its wake, many researchers now spend time investigating: (1) The properties of intrinsic noise in various network motifs [11, 12, 13], including autoregulation [14, 15, 16, 17, 18] and feed-forward loops [19, 20, 21]; (2) In incorporating cell-cycle variability into models of stochastic gene expression [22, 23, 24, 25, 26, 27, 28]; (3) Conducting experimental studies of gene expression in low copy-number regimes via single molecule fluorescence microscopy [29, 30, 31, 32, 33, 34]; (4) In the construction of minimal models of gene expression over a variety of scales that accurately capture the effects of intrinsic noise [35, 36, 37, 38]. It should be noted that although stochasticity limits the precision of gene expression, it also provides a mechanism through which phenotypic and cell-type diversification can occur [39].

The most commonly used analytically explored models of stochastic gene expression typically involve a single gene, and either mRNA or proteins, although generally not both together since this often makes analytics intractable. Furthermore, they often assume that the biological steady state (or cyclo-stationary state for cell-cycle models [22, 23]) has been reached so that transient effects become negligible. This is in correspondence with the idea that for most of the time gene expression happens in the so-called *biological steady-state* and that perturbations away from this are the exception rather than the rule [40]. The gene expression models of interest in this thesis are (1) the *telegraph model* of gene expression [35, 41], and (2) genetic *autoregulation* [42, 43]. The telegraph model is a simplification of gene expression that assumes the gene operates via an on/off switch, meaning that the production of mature mRNA can either occur in a transcriptionally active or inactive state. It is now common amongst experimental studies to directly infer the parameters of the telegraph model straight from the data [44, 45, 46, 47]. On the other hand, genetic autoregulation, despite being somewhat similar in structure to the telegraph model (two genes states and the same number of reactions), realises radically different behaviour since the bound state of the gene can only be activated by binding with a protein—in the case where no proteins (or few) are available then the system must remain in the unbound gene state. There are two forms of autoregulation, positive and negative. Both of them have their individual uses. Positive feedback promotes bimodality meaning that two differing modes of behaviour can be observed [15, 42, 39] being beneficial for cellular differentiation. Whereas, negative feedback can act as a control mechanism whereby when too many proteins are present the ability to produce more of them is restricted [39]. Notably, *bimodality seen in autoregulation is a purely*

stochastic phenomenon when cooperative binding is not present (i.e., when only a single protein needs to bind for the gene to move into the bound state) [6]. Together, positive and negative autoregulation contribute to the functioning of the circadian clock [48, 49] and autoregulation is a very common network motif. For example, in *E. coli* it is estimated that 40% of all transcription factors are self-regulated [11, 14].

Another relevant toy model in molecular biology is Michaelis-Menten enzyme kinetics [50, 51]. It provides possibly the simplest description of catalysis on a microscopic level. In the system, there are enzymes and substrates which can bind to each other to form a complex. This complex can then either unbind to give back the substrate, or else go on to produce the product, where in both cases the enzyme is conserved. Despite its simple structure, this model is mainly studied from deterministic perspectives [52, 53], often where only a single enzyme is present [54, 55]. Stochastic approaches to Michaelis-Menten kinetics are few [56, 57], and the exact time-dependent solution for a single enzyme although complete is practically difficult to use [58]. Since product formation is generally assumed to be non-reversible, there exists no steady state distribution at large times and only an absorbing state. Therefore, understanding the dynamics of the Michaelis-Menten mechanism comes from a comprehension of the transient dynamics, for which a solution with multiple enzymes is difficult not only from stochastic but also from deterministic perspectives.

A final model that will be considered in this thesis is the ant recruitment model introduced by Kirman (isomorphic to the Moran model used to model mutation in population genetics) [59, 60]. This model has been of intense interest over the past two decades, in particular due to the parallels it draws in how economic agents make decisions. It concerns two food sources in a population of ants, where each ant is associated with a single food source. The ants switch between the food sources due to two separate influences—one being a random switching event while the other being a recruitment interaction with other ants. The economic parallels arise in that under certain conditions the ants coalesce on a single food source, due to the recruitment, an effect that is entirely *endogenous* (meaning not due to external effects). This contradicts the standard economic idea of the *representative agent* [61], which assumes that a heterogeneous population of agents can be replaced, to a good approximation, by agents who take on the average of all economic behaviour. In the world of representative agents, there are no interactions and hence all changes in macrobehaviour (behavioural changes across the whole population) occur from forces external to the system known as *exogenous* forces. Kirman's model of ant recruitment illustrates that endogenous forces can be the primary force leading to large sways in economic decision making. Importantly, the coalescence seen in the ant recruitment model is another example of stochastic

bimodality that has no deterministic counterpart [62]. Many of the analytic results have been conducted either at steady-state [61, 63, 64]. If time-dependence is considered it is done so in the limit of infinite ants [65, 66, 67] or effectively includes a finite number of ants in a continuous agent number (non-discrete) setting [62].

Although stochastic methods allow us to model random processes more accurately and observe phenomena that do not occur from deterministic considerations alone, the downsides of their use often come from computational or analytic difficulties. Computationally, there are few methods through which simulations or numerical solutions to stochastic systems can be conducted. The two most popular methods being the stochastic simulation algorithm (SSA) [68] and finite state projection (FSP) [69]. However, these both suffer from the curse of dimensionality which restricts their use for problems of more than a few species (unless one has access to vast computing power). More recently, other methods reliant on neural networks have been explored [70, 71, 72, 73] which vastly reduce the computational burden, nevertheless this research is still in its nascent stages. As a result, the main problem is that even though computational approaches are available, we are limited to the computing power and methods currently available to us, which motivates more mathematical approaches. In general, analytics are favoured since they give both more intuition and at a reduced computational burden than simulation based approaches. Analytically, the starting point for stochastic modelling of chemical kinetics is the *chemical master equation* (CME), a set of coupled first-order ordinary differential equations describing the evolution of the probability distribution in the system of interest [8, 74]. In all but the simplest of reaction schemes the CME is difficult to solve, even at steady-state, and where solutions are possible they often come in the form of series expansions or the (relatively generalised) special functions that define them ([42, 43, 75, 76] among many others). To solve the CME exactly in time is an even more difficult task and it has only been done for a handful of systems [77, 78, 79, 80, 41]. Because of this, approximations to the CME are vital in order to extract analytics [81], and come in a variety of forms including (but not limited to) time scale separation [82, 83, 84, 85, 86], system size expansions [8, 87, 88, 89], Fokker-Planck equations [8, 74, 90] and the linear mapping approximation [91].

This thesis contributes to several of the aspects mentioned above. First, it addresses the validity of common approximations across different systems, from models of gene expression to enzyme kinetics, and further asserts the conditions in which we expect these approximations to hold. Secondly, we take an interest in solving and approximating stochastic kinetics not only at steady state, but in time as well. Such transient dynamics are often unstudied, but are of vital importance whenever a system is perturbed by an external force requiring the system to again relax back to the steady state, and additionally provide much more information regarding the underlying dynamical process.

1.2 Thesis structure

This thesis is structured as follows. Chapter 2 outlines the necessary preliminaries to understand the remaining chapters. In these preliminaries, there are introductions to the stochastic simulation algorithm, the chemical master equations and its various approximations used in this thesis, the finite state approximation, the delayed stochastic simulation algorithm and transient methods to solve the chemical master equation.

Chapter 3 focuses on Hill-type propensities often used to model genetic autoregulation, and provides an answer to the question: Under what conditions is the Hill function propensity a valid approximation? *A priori* this is not known, since its derivation relies on the deterministic rate equations, and its use in stochastic kinetics is a purely heuristic approximation.

Chapter 4 explores the use of the unified colored noise approximation to provide a reduced mapping of complex models of gene expression via the addition of Gaussian colored noise on the rate parameters. For a model of cooperative auto-regulatory feedback, we show how to appropriately choose the timescale and size of the colored noise to perform the mapping between the full and reduced models. Importantly, we show how to include fluctuations in the protein production rate due to multi-stage mRNA processing and how to effectively include more complex protein degradation mechanisms in models of autoregulation.

Chapter 5 studies why the telegraph model of gene expression has been so successful in predicting mRNA distributions given that it neglects RNA polymerase dynamics, which provide several key rate limiting steps. To conduct this study we use a first passage time analysis to compare analytically determined waiting time distributions of mRNA production between two models. The first model is a telegraph-like model, while the second is a mechanistic model that includes the key rate limiting steps in the RNA polymerase dynamics and deterministic elongation from nascent to mature mRNA. The results show that such a mapping based on the first passage time admits a region of uniquely mapped parameters where there is good correspondence between the two models.

In Chapter 6 the Michaelis-Menten mechanism of enzyme kinetics is studied. We provide an approximate closed-form solution to the CME describing its dynamics under a realistic time scale separation in the rate parameters. This closed-from solution is analysed in various regimes of the parameter space, where we observe the emergence of *transient bimodality*, which corresponds to the transient occurrence of two distinct populations over a finite time. We further extend the approximative method used on the Michaelis-Menten mechanism to solve the more complex mechanism of ternary complex formation, involving multiple substrates.

In Chapter 7 we continue the theme of transient kinetics showing how one can derive an exact time-dependent solution to Kirman's model of ant rationality for a finite discrete number of N ants. We show how to effectively use this solution in a practical way, and then extend it to solve the ant rationality model for increasing degrees of model asymmetry, and even show how it provides a semi-analytic solution to the vacillating voter model.

Finally, in Chapter 8 we illustrate some future directions that stem from the previous chapters, including the discovery of volume scaling laws for protein production rates in minimal gene expression models, and an exploration on the origins of transient bimodality in enzyme kinetics. We then conclude this thesis.

Chapter 2

Preliminaries

This chapter provides the analytical and computational framework that are necessary to understand the following chapters. The core references for this chapter are [8, 68, 69, 81].

We begin by acquainting the reader with the stochastic simulation algorithm (SSA) in Section 2.1, which provides the computational standard against which analytic solutions are compared, and which is utilised in every subsequent chapter of this thesis. In Section 2.2 we derive the chemical master equation (CME), which provides the analytic grounding for all of the subsequent chapters, followed by a discussion on the finite state projection (FSP) in Section 2.3, which provides an alternative to the SSA. In Section 2.4 we then introduce several relevant approximation methods often used to simplify the CME such that it can be solved. Section 2.5 defines the delayed SSA (dSSA) and the delayed CME (dCME) a modification of the CME that includes the presence of *deterministically delayed* reactions (which are non-Markovian) and shows how one can deal with these computationally. Such delayed reaction schemes are used in Chapter 5. Finally, Section 2.6 reviews two analytic methods for solving the CME in time. In this Chapter, boldface symbols indicate vectors and matrices. Note that a detailed account of the unified colored noise approximation (UCNA) is given in Chapter 4, hence we have not included it in these preliminaries.

2.1 Stochastic simulation algorithm

The SSA is a Monte Carlo algorithm and it is the standard tool with which we test our analytical solutions from the CME. To set up the SSA one must first grasp what a *chemical reaction network* is, and a reaction from that network is generally denoted in the following way [8, 81],

$$\sum_{i\in\mathcal{S}} a_{ir}X_i \xrightarrow{k_r} \sum_{i\in\mathcal{S}} b_{ir}X_i. \tag{2.1}$$

with $r \in \{1, 2, \ldots, R\}$, where R is the number of reactions. Here, $\mathcal{S}$ is the set of all N species, a_{ir} and b_{ir} are positive integers (and can be zero) denoting the stoichiometric coefficients for species i in reaction r, k_r is the rate at which reaction r occurs, and X_i is a chemical species. We define the state vector $\mathbf{n} = (n_1, n_2, \ldots, n_N)$ as the vector containing the number of each species in $\mathbf{X}$, i.e., n_i is the number of molecule of species X_i. The end result of the SSA is to simulate the probability distribution $P(\mathbf{n}, t)$, i.e., having $\mathbf{n}$ at a time t. Note that reaction rate k_r is not the propensity of the reaction (*propensity* defined as the probability for the reaction to occur per unit time), which will be defined below via mass-action kinetics, but a *rate constant* independent of $\mathbf{n}$ and t.

As an example of chemical kinetics, consider a system consisting of a catalytic reaction with two species, A and B, where A decays,

$$A + B \xrightarrow{k_1} 2A + B, \quad A \xrightarrow{k_2} \varnothing, \tag{2.2}$$

in which case,

$$\mathbf{a} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 2 & 0 \\ 1 & 0 \end{pmatrix},$$

$N = 2$, $R = 2$ and X_1 and X_2 are equivalently denoted as as A and B respectively. Note that the symbol $\varnothing$ is used to indicate a large reservoir into which A is removed from the system [8]. It is useful to define the stoichiometric matrix $\mathbf{S} \equiv \mathbf{b} - \mathbf{a}$, where S_{ij} defines the net gain or loss of species i in reaction j.

One now needs to assign the propensities with which reactions are fired to give the chemical reaction network some dynamics. The common way to do this involves making two key assumptions [92]: that diffusion is the fast timescale in the system and that molecules are point particles. These two assumptions result in *mass-action kinetics*, where the reaction propensities are dependent only on (1) the number of molecules in the system and (2) the size of the system Ω, *but not a spatial description of where particles are*. Further to this, since a spatial description is not needed in this regime, the state of the system is entirely specified by the number of molecules of each reacting species $\mathbf{n}$ (hence why we refer to it as the state vector). Mathematically, the propensity for reaction r, denoted $f_r(\mathbf{n})$, means that the probability of reaction r firing in the infinitesimal time interval $[t, t + dt)$ is $f_r(\mathbf{n})dt$.

The laws of mass-action kinetics are derived from combinatorial considerations. For example, if we have a first-order reaction $A \xrightarrow{k}$, then the propensity of this reaction is proportional to the number of molecules of A, hence the propensity for this reaction is simply $k\, n_A$. Similarly, if one has a second-order reaction $A + B \xrightarrow{k}$ then the propensity takes into account all combinations of A and B molecules that could interact, and is also inversely proportional to the system size, giving the propensity to be $k n_A n_B / \Omega$.

These same principles can be applied for all possible types of reactions, for example the propensity of the reaction $A + A \xrightarrow{k}$ is $kn_A(n_A - 1)/\Omega$ where the factor $n_A - 1$ accounts for the fact that in choosing a second molecule to interact with one has already been chosen (and that if one can $n_A = 1$ that no reaction occurs). For the example reaction above in Eq. (2.2), the propensities are $f_1(\mathbf{n}) = k_1 n_A n_B/\Omega$ and $f_2(\mathbf{n}) = k_2 n_A$. Note zeroth-order reactions of the form $\varnothing \xrightarrow{k}$ imply a propensity $\Omega\, k$, i.e., proportional to the system size.

The Markov property, which is the statement that the state of the system at time $t + dt$ depends only on the state of the system at time t, defines the waiting time for each reaction to occur. Markov processes have exponentially distributed waiting times since the differential probability, in our case $f_r(\mathbf{n})dt$, for an event to occur in $[t, t + dt)$ is time-independent. Gillespie refers to this assignment of the reaction propensity as the *fundamental premise*, since each $f_r(\mathbf{n})$ only depends on the current state of the system $\mathbf{n}$, therefore defining the Markov properties of the SSA [68]. Hence, the probability that at time τ in the future *any reaction* will occur is,

$$p(\tau|\mathbf{n}, t) = f_0(\mathbf{n})\, e^{-f_0(\mathbf{n})\tau}$$

where $f_0(\mathbf{n}) = \sum_{i=1}^{R} f_r(\mathbf{n})$ is the total propensity at which reactions occur. Reaction r will then be fired with a probability,

$$p_r(\mathbf{n}) = f_r(\mathbf{n})/f_0(\mathbf{n}).$$

Now all of the pieces of the SSA are in place, and one could use the following pseudocode to run the algorithm,

1. Instantiate the initial state, $\mathbf{n} = \mathbf{n}_0$, of the system at $t = t_0$.
2. Draw a reaction waiting time τ from $p(\tau|\mathbf{n}, t)$.
3. Draw the reaction r that is fired from $p_r(\mathbf{n})$.
4. Update the state vector for the stoichiometry: $\mathbf{n} \to \mathbf{n} + \mathbf{S}_r$ (where $\mathbf{S}_r$ is the r^{th} column of $\mathbf{S}$) and the reaction time $t \to t + \tau$ and store the it the matrix $\mathbf{n}(t)$.
5. Repeat process from Step 2 until some specified maximum time T is reached and output $\mathbf{n}(T)$.

Generally, one runs many parallel simulations of the SSA and then bins the data over the time intervals of interest giving a probability distribution that can then be compared to those from analytic calculation. Although slightly less intuitive, it is more computationally efficient to employ the 'direct method' [68] which bypasses the need to sample the distributions $p(\tau|\mathbf{n}, t)$ and $p_r(\mathbf{n})$ in favour of drawing two uniform random variables (a result which comes from inverse transform sampling [93]). The direct method replaces steps 2 and 3 above with,

2. Draw two uniform random numbers, u_1 and u_2, from the unit interval $[0, 1]$.
3. Assign $\tau = -\ln(u_1)/f_0(\mathbf{n})$ and $r = \min_r \left(\sum_{i=1}^{r} f_i(\mathbf{n}) > u_2 f_0(\mathbf{n})\right)$.

Throughout the thesis it is the direct method that we utilise for the SSA, and we implement it in Julia [94]. Julia packages `Catalyst.jl` and `DifferentialEquations.jl` [95] provide a simple and computationally fast framework with which one can simulate the SSA. Example trajectories and histograms from the SSA for an auto-regulatory reaction scheme are shown in Fig. 2.1(a)–(b), where the histogram shown is binned over 10^4 trajectories. Follow the link here for an example of coding up the SSA in Julia (made by the author).

2.2 Chemical master equation

The CME provides the analytic basis of this PhD thesis. Consider a chemical reaction network (as seen previously in Eq.(2.1)) evolving from a time $t \to t + dt$ (for infinitesimal time interval dt). If one has the same assumptions that underlie mass-action kinetics from the SSA, then following the fundamental premise, reaction events occur with exponential waiting times with mass-action propensities, and one can write the following equation for probability balance,

$$P(\mathbf{n}, t + dt) = P(\mathbf{n}, t) + \left(\sum_{r=1}^{R} P(\mathbf{n} - \mathbf{S}_r, t) f_r(\mathbf{n} - \mathbf{S}_r) - P(\mathbf{n}, t) \sum_{r=1}^{R} f_r(\mathbf{n}) \right) dt, \quad (2.3)$$

where $P(\mathbf{n}, t)$ is the probability to have $\mathbf{n}$ at time t. We remind the reader that $\mathbf{S}_r$ is the r^{th} column of the stoichiometric matrix $\mathbf{S}$, and is of length N, where N is the number of reacting species. In words, this equation states that the probability of having $\mathbf{n}$ at time $t + dt$ is the probability to have $\mathbf{n}$ at time t *plus the net gain or loss of probability to or from* $\mathbf{n}$ *over time interval* dt. Rearranging Eq. (2.3) and taking the limit of $dt \to 0$ we obtain the CME,

$$\partial_t P(\mathbf{n}, t) = \sum_{r=1}^{R} P(\mathbf{n} - \mathbf{S}_r, t) f_r(\mathbf{n} - \mathbf{S}_r) - P(\mathbf{n}, t) \sum_{r=1}^{R} f_r(\mathbf{n}). \quad (2.4)$$

Note that the CME has a natural boundary at all $n_i = 0$ since there is always zero flux towards the states with $n_i < 0$ given propensities of mass-action form (even the non mass-action Hill-type propensities considered later in the thesis satisfy this condition). There is no such upper boundary at some finite $n = N$ unless it is imposed by the dynamics of the reaction network itself. In general, the state space of $\mathbf{n}$ is infinite.

2.2.1 Generating functions

The method of generating functions is a very common and flexible way to solve the CME that we apply throughout this thesis [8, 74, 81]. We now show two examples of its application. The first example is very simple but allows one to see the elegance of the method, while the second is more relevant to our purpose.

Example 1

Here we consider the birth-death process,

$$\varnothing \underset{r_2}{\overset{r_1}{\rightleftharpoons}} X. \tag{2.5}$$

The CME describing this reaction scheme is,

$$\partial_t P(n,t) = r_1 P(n-1,t) + r_2(n+1)P(n+1,t) - (r_1 + r_2 n)P(n,t). \tag{2.6}$$

The generating function method introduces a generating function $G(z,t) = \sum_n P(n,t)z^n$. If one can then construct and solve an equation for $G(z,t)$, $P(n,t)$ and the factorial moments can be calculated as follows,

$$\begin{aligned} P(n,t) &= \frac{1}{n!}\partial_z^n G(z,t)|_{z=0}, \\ \mathbb{E}[(n)_r] &= \partial_z^r G(z,t)|_{z=1}, \end{aligned} \tag{2.7}$$

where $(n)_r = \prod_{x=1}^{r}(n-x+1)$ is the falling factorial function. We then find the raw moments from the factorial moments via,

$$\mathbb{E}[n^r] = \sum_{j=1}^{r} \left\{ {r \atop j} \right\} \mathbb{E}[(n)_r], \tag{2.8}$$

where $\left\{ {r \atop j} \right\}$ are Stirling's numbers of the second kind [96, p. 822]. Let's suppose the birth-death process has reached steady state, meaning that the probability distribution and moments are unchanging in time, i.e., $\partial_t P(n,t) = 0$. It can be shown that the corresponding generating function equation to Eq. (2.6) is,

$$r_1(z-1)G(z) + r_2(1-z)\partial_z G(z) = 0, \tag{2.9}$$

whose solution is $G(z) = Ce^{r_1 z/r_2}$. The constant of integration C is determined by the normalisation condition $G(1) = 1$, leading to $C = \exp(-r_1/r_2)$. Taking derivatives of $G(z)$ one then easily determines that the birth-death process is Poisson distributed with,

$$P(n,t) = \frac{1}{n!}\left(\frac{r_1}{r_2}\right)^n e^{-r_1/r_2},$$
$$\mathbb{E}[(n)_r] = \left(\frac{r_1}{r_2}\right)^n,$$

which fully specifies the steady state solution to the birth-death process with constant rates.

Example 2

In this example we consider a system containing a single gene that possesses two different gene states. mRNA is produced by both of the gene states and is removed from the system at some rate (either by degradation or dilution). This process in shown in the reaction scheme below,

$$G \xrightarrow{\rho_u} G + M,\ G^* \xrightarrow{\rho_b} G^* + M,$$
$$G \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*,\ M \xrightarrow{d} \varnothing. \tag{2.10}$$

Note that the CME for this reaction scheme has recently been solved in time (see SI of [91]). This reaction scheme is very similar to the telegraph model of gene expression [35, 41], aside from the fact that both gene states G and G^* admit the production of mRNA (whereas in the telegraph model only a single gene state is transcriptionally active). We show here how to solve this reaction scheme at steady state. First we write out the master equations for $P_0(n,t)$ and $P_1(n,t)$, which are the probabilities of having n mRNA at a time t in gene states G and G^* respectively (suppressing the time-dependence for brevity),

$$\partial_t P_0(n) = \rho_u P_0(n-1) + d(n+1)P_0(n+1) + \sigma_u P_1(n) - \sigma_b P_0(n) - (\rho_u + dn)P_0(n), \tag{2.11}$$

$$\partial_t P_1(n) = \rho_b P_1(n-1) + d(n+1)P_1(n+1) - \sigma_u P_1(n) + \sigma_b P_0(n) - (\rho_b + dn)P_1(n). \tag{2.12}$$

Defining a generating function for each probability distribution, the set of generating function equations we obtain is,

$$\partial_t G_0(z) = \rho_u(z-1)G_0(z) + d(1-z)\partial_z G_0(z) + \sigma_u G_1(z) - \sigma_b G_0(z), \tag{2.13}$$

$$\partial_t G_1(z) = \rho_b(z-1)G_1(z) + d(1-z)\partial_z G_1(z) - \sigma_u G_1(z) + \sigma_b G_0(z). \tag{2.14}$$

Summing these together, and defining the total generating function over both gene states as $G(z) = G_0(z) + G_1(z) = \sum_n z^n P(n)$, we get,

$$\partial_t G(z) = (z-1)(\rho_u G_0(z) + \rho_b G_1(z)) + d(1-z)\partial_z G(z). \tag{2.15}$$

Using the fact that $G_1 = G - G_0$ and rearranging the above for G_0, one can then substitute into Eq. (2.13) or (2.14) and at steady-state we obtain,

$$d^2(z-1)\partial_z^2 G(z) + (d\Sigma - d(z-1)(\rho_b + \rho_u))\partial_z G(z) + (\rho_u\rho_b(z-1) - \chi)G(z) = 0, \tag{2.16}$$

where we have defined,

$$\Sigma = \sigma_u + \sigma_b,$$
$$\chi = \rho_u\sigma_u + \rho_b\sigma_b.$$

To solve Eq. (2.16) we observe that it has a regular singularity at $z = 1$ and an irregular singularity at $z = \infty$. Hence, it will admit a solution of a confluent Hypergeometric function $w(z) = {}_1F_1(a, b; z)$, which is defined the differential equation by,

$$z\partial_z^2 w(z) + (b - z)\partial_z w(z) - aw(z) = 0, \tag{2.17}$$

which has an irregular singularity at $z = \infty$ and a regular singularity at $z = 0$. To map Eq. (2.16) into this form we let $x = z - 1$ and introduce the transformation $G(x) = e^{\alpha x}F(x)$, where we choose the value of α such to remove the z dependent term in the coefficient of $G(z)$ in Eq. (2.16). One finds that both $\alpha = \rho_u/d$ and $\alpha = \rho_b/d$ are admissible (since both gives the same generating function $G(z)$), from which we arbitrarily choose the former, which gives us an equation for $F(x)$,

$$x\partial_x^2 F(x) + (\Sigma - x(\rho_b - \rho_u))d^{-1}\partial_x F(x) + (\rho_u\Sigma - \chi)d^{-2}F(x) = 0. \tag{2.18}$$

A final change of variable $y = (\rho_b - \rho_u)x/d$ gives us an equation of confluent hypergeometric form, and we hence find the solution for $G(z)$ as,

$$G(z) = e^{(z-1)\rho_u/d}\,{}_1F_1\left(\frac{\sigma_b}{d}, \frac{\Sigma}{d}; \frac{(\rho_b - \rho_u)(z-1)}{d}\right), \tag{2.19}$$

from which one can calculate the moments and probabilities. Note that there is an additional solution to $G(z)$ which includes a *Tricomi function* [97], however this is not physically admissible since the generating function it describes does not represent a well-defined probability function. For example, the steady state mean of the process is

given by,

$$\langle n \rangle = \frac{\rho_u \sigma_u + \rho_b \sigma_b}{d(\sigma_u + \sigma_b)}, \tag{2.20}$$

which can be more intuitively expressed by defining the proportion of time in the u state $f = \sigma_u/(\sigma_u + \sigma_b)$, giving,

$$\langle n \rangle = f\frac{\rho_u}{d} + (1-f)\frac{\rho_b}{d}. \tag{2.21}$$

2.3 Finite State Projection

2.3.1 Time-dependent FSP

The finite state projection (FSP) provides an alternate, albeit approximative, method to the SSA for computing distributions of molecule numbers in time [69] or at steady state [98], which can be used to verify analytic calculation. The basic idea of the FSP is to solve the CME by matrix exponentiation. To do this, first let $\boldsymbol{\mathcal{P}}(\mathbf{n}, t) = (P(\mathbf{x}_1(t)), P(\mathbf{x}_2(t)), \dots)$ be a vector containing the probability of having each possible configuration of the state vector $\mathbf{n}$, where for $\mathbf{x}_i(t)$ are specific instances of $\mathbf{n}$. The order of the $P(\mathbf{x}_i(t))$ in $\boldsymbol{\mathcal{P}}(\mathbf{n}, t)$ is arbitrary, so long as one remains consistent with their placement. For example, in the one-dimensional case $\boldsymbol{\mathcal{P}}(\mathbf{n}, t) = (P(0,t), P(1,t), \dots)$, where $P(n,t)$ is the probability of having n molecules of the single species at time t. One can then write the CME in the following form,

$$\partial_t \boldsymbol{\mathcal{P}}(\mathbf{n}, t) = \mathbf{M} \cdot \boldsymbol{\mathcal{P}}(\mathbf{n}, t), \tag{2.22}$$

where the matrix $\mathbf{M}$, which we will call the *master operator*, is defined by,

$$\mathbf{M}_{ij} = \begin{cases} -\sum_m^R f_m(\mathbf{x}_i) & \text{for } i = j, \\ f_m(\mathbf{x}_i) & \text{for all } \mathbf{x}_j \text{ such that } \mathbf{x}_j = \mathbf{x}_i + \mathbf{S}_m, \\ 0 & \text{otherwise.} \end{cases} \tag{2.23}$$

Then, Eq. (2.22) can be formally solve by utilising the matrix exponential,

$$\boldsymbol{\mathcal{P}}(\mathbf{n}, t) = \exp(t\mathbf{M}) \cdot \boldsymbol{\mathcal{P}}(\mathbf{n}, 0). \tag{2.24}$$

However, in the general case of the CME there is no upper bound on the molecule number of each species, meaning that the state space is generally infinite, and that one cannot calculate the matrix exponential above. Note that in the case where the state space is naturally bounded, Eq. (2.24) provides an exact solution to the CME describing

that problem (although it may be computationally inefficient). It is now that we employ the 'finite state' projection to evade this issue. Where $n_i \in \mathbb{N}_0$, specify a molecule number 'cut-off' or truncation in the state space for species i denoted $M_i,\ i \in \{1, 2, \ldots, N\}$, then there are $K = \prod_i^N M_i$ unique configurations in the truncated state space of $\mathbf{n}$, and K is the size of the state space. We denote the finite state master operator (of dimensions $K \times K$) by $\tilde{\mathbf{M}}$, the set of all states in the truncated space as $\tilde{\mathcal{S}}$, and we denote the solution to the resulting truncated CME by $\tilde{\boldsymbol{\mathcal{P}}}(\mathbf{n}, t)$. Note that by construction the stochastic process defined by $\tilde{\mathbf{M}}$ leaks probability from all states $\mathbf{n}$ that are at the upper boundary of the state space (i.e., $\sum_{x_i} \tilde{\mathbf{P}}(\mathbf{x}_i, t) < 1$ for $t > 0$), and accounting for the error that this causes is a major aspect of the work conducted in [69]. The loss of probability to the absorbing state is shown in the schematic in Fig. 2.1(c).

As stated in the original paper [69], by specifying an acceptable error ϵ such that if the sum of elements in $\sum_{i=1}^K \tilde{\mathbf{P}}(\mathbf{x}_i, t) \geq 1 - \epsilon$ we accept the result of the FSP, then one could implement the following algorithm that the authors of [69] term the FSP,

1. Define the initial state of the system $\tilde{\boldsymbol{\mathcal{P}}}(\mathbf{n}, 0)$, acceptable error $0 < \epsilon < 1$, and the set of state space truncations $\{M_i\}$ for $i = 1, 2, \ldots, N$.
2. Calculate $\tilde{\boldsymbol{\mathcal{P}}}(\mathbf{n}, t) = \exp(t\tilde{\mathbf{M}}) \cdot \tilde{\boldsymbol{\mathcal{P}}}(\mathbf{n}, 0)$, and hence evaluate $s = \sum_{i=1}^K \tilde{\boldsymbol{\mathcal{P}}}(\mathbf{x}_i, t)$.
3. If $s \geq 1 - \epsilon$ then stop. Else move to step 4.
4. Increase the size of the state space in a systematic way such that $K_{\text{new}} \geq K_{\text{old}}$ and return to step 2.

This fully specifies the time-dependent FSP. In common practice however, it may be more efficient to gain a heuristic understanding of where in $\mathbf{n}$ that $\boldsymbol{\mathcal{P}}(\mathbf{n}, t)$ becomes very small such that it can be neglected. The downsides of the FSP lie almost solely in the curse of dimensionality—the algorithm becomes exponentially slower as K increases due to the computation of the matrix exponential.

2.3.2 Steady state FSP

The FSP algorithm as defined above cannot be used to approximate the steady state probability as $t \to \infty$ since all the probability will have 'leaked out', i.e., when $\partial_t \tilde{\boldsymbol{\mathcal{P}}}_s(\mathbf{n}, t) = \tilde{\mathbf{M}} \cdot \tilde{\boldsymbol{\mathcal{P}}}_s(\mathbf{n}, t) = 0$. To get around this, the steady state FSP was developed in [98], which essentially designates a state, or set of states, into which the probability normally lost in the time-dependent FSP re-enters the system. The chosen state is known as the *designated state*, shown in the schematic in Fig. 2.1(d). For a CME with a well-defined steady state, the sum of all the elements each column of $\mathbf{M}$ is 0, i.e., $\sum_{j=1}^{\infty} \mathbf{M}_{ij} = 0$ for all $i = 1, 2, \ldots, \infty$. Hence, the steady state is well-defined since this condition states that the system is ergodic and that all states remain accessible as $t \to \infty$.

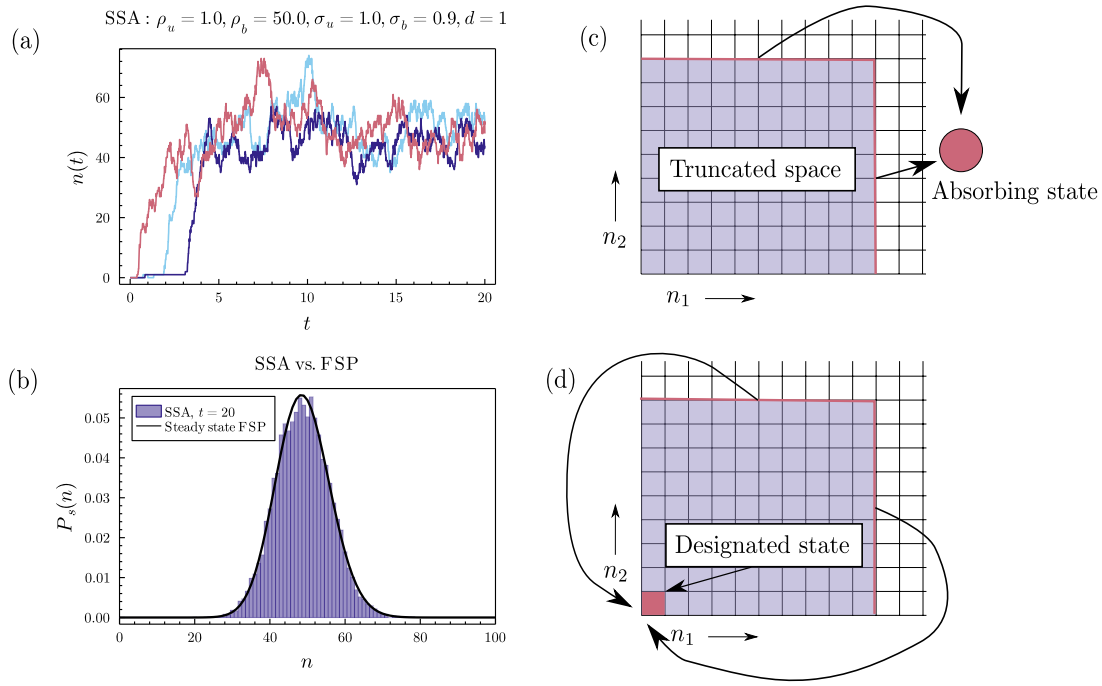


Figure 2.1: Figure showing examples of the SSA and FSP applied to the auto-regulatory reaction scheme in Eq. (2.26). (a) Three SSA trajectories for parameters shown on the figure. (b) Plot showing a histogram of the SSA at $t = 20$ (which is at steady state) versus the result from the steady state FSP. That the SSA is at steady state at $t = 20$ is verified by the very close agreement with the steady state FSP. The histogram from the SSA is binned over 10^4 trajectories, whereas the designated state of the steady state FSP is $n_G = 0, n_P = 0$ and the truncation chosen was $M = 200$ for the protein species (the gene is naturally bounded and so does not require a truncation). (c) Schematic showing the truncation of the state space in time-dependent FSP, and the flow of probability into the absorbing state (outside of the truncated region). (d) Schematic showing the truncation of the state space in steady state FSP, and the flow of probability into the designated state. (c) and (d) are based on Figs. 1 in [69] and [98] respectively.

In order to give our truncated system a well-defined steady state, we must enforce the condition $\sum_{j=1}^{\infty} \mathbf{M}_{ij} = 0$ for $j = 1, 2, \ldots, K$ on the matrix $\tilde{\mathbf{M}}$. Choose a designated state $\mathbf{x}_D(t) \in \tilde{\mathcal{S}}$ to which the 'leakage probability' will be assigned to. To enforce this as the designated state we need to modify $\tilde{\mathbf{M}}$ introduced above, specifically we assign the elements of row D such that $\tilde{\mathbf{M}}_{Dj} = -\sum_{i \neq D} \tilde{\mathbf{M}}_{ij}$ for $j = 1, 2, \ldots, K$, where the sum of all columns in $\tilde{\mathbf{M}}$ is now 0. We refer to this new matrix as $\overline{\mathbf{M}}$, and the steady state FSP is reduced to solving the set of simultaneous equations defined by,

$$\overline{\mathbf{M}} \cdot \tilde{\boldsymbol{\mathcal{P}}}_s(\mathbf{n}) = \mathbf{b}, \tag{2.25}$$

where $\mathbf{b}$ is a vector of zeros of length K. Unlike the time-dependent FSP we must apply the normalisation condition to fully determine the vector $\tilde{\boldsymbol{\mathcal{P}}}_s(\mathbf{n})$, since the row D of $\overline{\mathbf{M}}$ is not linearly-independent. The lack of linear-independence in this row means that one could indeed simplify the above calculation by replacing rows D of $\overline{\mathbf{M}}$ and $\mathbf{b}$ with the normalisation condition, i.e., let $\overline{\mathbf{M}}_{Dk} = 1$ and $\mathbf{b}_k = \delta_{kD}$ for $k = 1, 2, \ldots, K$, and where δ_{kD} is the Kronecker delta function. With these newly defined $\overline{\mathbf{M}}$ and $\mathbf{b}$ the enforcement of the normalisation condition *a posteriori* to solving Eq. (2.25) is no longer required.

Importantly, in conducting the steady state FSP one must have enforced any *conservation laws* present in the system upon construction of $\overline{\mathbf{M}}$. Simply, a conservation law is statement regarding that a certain quantity is conserved throughout the entire evolution of the system, and each conservation law reduces the effective number of species one has by 1. For example, consider the chemical reaction network describing the auto-regulating gene (3 species),

$$G \xrightarrow{\rho_u} G + P,\ P \xrightarrow{d} \varnothing,\ P + G \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*,\ G^* \xrightarrow{\rho_b} G^* + P. \tag{2.26}$$

In this reaction scheme we make the assumption that there is only one gene copy present that can fluctuate between two states G (unbound) and G^* (bound). Hence, the sum of the number of G and G^* present at any one time is 1, i.e., $n_G + n_{G^*} = 1$ (this is our conservation law). We can then enforce this on the reaction scheme above using this conservation law,

$$G \xrightarrow{\rho_u} G + P,\ P \xrightarrow{d} \varnothing,\ P + G \underset{\sigma_u(1-n_G)}{\overset{\sigma_b}{\rightleftharpoons}} \varnothing,\ \varnothing \xrightarrow{\rho_b(1-n_G)} P, \tag{2.27}$$

where we have assumed that the system size $\Omega = 1$, and clearly there are now only 2 species (with G^* being replaced with the conservation law). Note that although one can do similarly for the time-dependent FSP, it is actually not required that one does this since the initial condition explicitly enforces the conservation law. For this reaction scheme we show the performance of the steady state FSP against the SSA in Fig. 2.1(b) with a truncation of $M = 200$ for the protein number, and note that it is an order of $\mathcal{O}(10^3)$ times faster than the SSA used in the comparison.

Similar to the time-dependent FSP, work has been done on algorithmically determining the appropriate state space truncation [98]. In practice, it is likely more efficient to gain a heuristic understanding of where in $\mathbf{n}$ that $\boldsymbol{\mathcal{P}}(\mathbf{n}, t)$ becomes very small such that those states can be neglected.

A recent package has been published in Julia called `FiniteStateProjection.jl` that allows one to easily conduct the FSP in time and at steady state. Follow the link here for an example of coding the steady state FSP in Julia (made by the author). We utilise the steady state FSP in Chapters 3 and 4 of this thesis.

2.4 Approximations of the CME

Having defined the CME, we see that it consists of a set of coupled first-order ODEs of the number of the magnitude of the state space $\mathbf{n}$ (the number of unique state vectors possible). For unbounded systems this set of first-order ODEs is infinite in number, and where generating function approaches become intractable we are often forced to consider approximations of the CME to make analytic progress. Below we detail several approximations of the CME from the literature which we will use in this thesis.

2.4.1 Fokker-Planck and Langevin equation

It is often more convenient to deal with partial differential equations that have continuous state variables. The Fokker-Planck equation (FPE) performs this approximation for the CME. It consists of making two basic assumptions [8]: (1) only small jumps in the molecule numbers occur, and (2) that $P(\mathbf{n}, t)$ varies slowly with respect to $\mathbf{n}$. If these assumptions are valid then it is appropriate to assume that $\mathbf{n}$ is a continuous vector, and Taylor expand the RHS of Eq. (2.4) with respect to $\mathbf{n}$. Let $h(\mathbf{n}) = P(\mathbf{n}, t) f_r(\mathbf{n})$, then expanding to second order in $\mathbf{n}$ gives

$$h(\mathbf{n} - \mathbf{S}_r) \approx h(\mathbf{n}) - (\mathbf{S}_r \cdot \nabla) h(\mathbf{n}) + \frac{1}{2}(\mathbf{S}_r \cdot \nabla)^2 h(\mathbf{n}) + \mathcal{O}((\mathbf{S}_r \cdot \nabla)^3 h(\mathbf{n})).$$

where we have used the common denotation $\nabla_i \equiv \partial_{n_i}$ for $i = 1, 2, \ldots, N$. Use of this result gives us the so-called chemical Fokker-Planck (CFPE) equation,

$$\partial_t P(\mathbf{n}, t) \approx -\sum_{r=1}^{R} (\mathbf{S}_r \cdot \nabla) f_r(\mathbf{n}) P(\mathbf{n}, t) + \frac{1}{2} \sum_{r=1}^{R} (\mathbf{S}_r \cdot \nabla)^2 f_r(\mathbf{n}) P(\mathbf{n}, t). \tag{2.28}$$

If one instead continued the expansion of $\mathbf{h}$ *ad infinitum* then one recovers the Kramers-Moyal expansion which is formally identical to the CME itself [74]. We use FPEs in Chapter 4 to approximate the CME describing cooperative autoregulation.

In Section 4 we additionally use the Langevin equations corresponding to the FPE whose general form is given by [57, 99, 100],

$$\frac{d\mathbf{n}}{dt} = \mathbf{S} \cdot \mathbf{f}(\mathbf{n}) + \mathbf{S} \cdot \mathrm{Diag}\left(\sqrt{\mathbf{f}(\mathbf{n})}\right) \cdot \mathbf{\Gamma}(t), \tag{2.29}$$

where $\mathrm{Diag}\left(\sqrt{\mathbf{f}(\mathbf{n})}\right)$ is a $R \times R$ diagonal matrix with diagonal elements $(\sqrt{f_1(\mathbf{n})}, \ldots, \sqrt{f_R(\mathbf{n})})$, and $\mathbf{\Gamma}(t) = (\Gamma_1(t), \ldots, \Gamma_R(t))$ is a vector where each $\Gamma_i(t)$ is an independent Gaussian white noise, i.e., $\Gamma_i(t)\Gamma_j(t') = \delta_{ij}\delta(t - t')$. Note that this Langevin equation is an Itô stochastic differential equation (as opposed to a Stratonovich type).

2.4.2 Linear noise approximation

A further approximation one can make is the linear noise approximation (LNA), which is typically done by truncating the system size expansion (SSE) at order Ω^0 [8]. Generally, it is not as accurate as the FPE [90], but unlike the FPE it always admits a Gaussian solution, even in time. Although normally the SSE is conducted on the CME, up to order Ω^0, the SSE of the CME and the CFPE are identical [90]. Since we do not use higher orders of the SSE in this thesis we proceed with the derivation of the LNA from the CFPE below.

The general idea behind the SSE is to expand the CME or CFPE around the mean as given by the deterministic rate equations. Before introducing this expansion we first note that the deterministic rate equations are given by [81],

$$\frac{d\boldsymbol{\phi}(t)}{dt} = \mathbf{S} \cdot \mathbf{g}(\boldsymbol{\phi}(t)), \tag{2.30}$$

where $\boldsymbol{\phi}(t)$ is the vector of deterministic concentrations and $\mathbf{g}(\boldsymbol{\phi}(t))$ is the vector of macroscopic rates, which can be calculated from the propensity vector through $\mathbf{g}(\mathbf{x}) = \lim_{\Omega\to\infty}(\Omega^{-1}\mathbf{f}(\mathbf{n}, \Omega))$, where we have now explicitly included the dependence of $\mathbf{f}$ on the system size Ω and have defined $\mathbf{x} = \mathbf{n}/\Omega$ as the vector of exact concentrations (whereas $\boldsymbol{\phi}(t)$ is the solution of Eq. (2.30), and in general $\mathbf{x}(t) \neq \boldsymbol{\phi}(t)$). We note that generally $\mathbf{x}(t) \neq \boldsymbol{\phi}(t)$ since $\mathbf{x}(t)$ is still a discrete quantity, whereas $\boldsymbol{\phi}(t)$ is continuous, although their first two moments do coincide for linear reaction networks, and some second-order reaction networks of particular form [101]. Although the rate equations do not account for stochastic effects, their benefit lies in their simplicity (often analytically tractable and always computationally fast), meaning that $\boldsymbol{\phi}(t)$ provides the ideal starting point for the expansion of the CFPE.

We now expand the CFPE in three-steps. The first step is what van Kampen calls the *essential step* [8], and it is the ansatz from which the SSE follows,

$$\mathbf{n} = \Omega\boldsymbol{\phi}(t) + \Omega^{1/2}\boldsymbol{\xi}. \tag{2.31}$$

The first term in the ansatz is the molecule number corresponding to the deterministic concentrations, while the second terms accounts for fluctuations about these values, with the coefficient $\Omega^{1/2}$ simply stating that we expect fluctuations about the mean to be of order $\Omega^{1/2}$. We can then further define the probability distribution in terms of $\boldsymbol{\xi}$, i.e.,

$$\Pi(\boldsymbol{\xi}, t) = \left|\frac{d\mathbf{n}}{d\boldsymbol{\xi}}\right| P(\mathbf{n}, t) = \Omega^{N/2} P(\mathbf{n}, t), \tag{2.32}$$

which follows from the conservation of differential area under the change of variable $\mathbf{n} \to \boldsymbol{\xi}$ (where N is the number of species). The second step is to realise that the partial derivative on the LHS of the CFPE in Eq. (2.28) is taken at constant $\mathbf{n}$, i.e., in the plane defined by $\Omega^{-1/2}\partial_t \mathbf{n} = \Omega^{1/2}\partial_t \boldsymbol{\phi}(t) + \partial_t \boldsymbol{\xi} = 0$. The LHS of the CFPE then becomes,

$$\begin{aligned}\Omega^{-N/2}\partial_t P(\mathbf{n},t) \equiv \frac{d\Pi(\boldsymbol{\xi},t)}{dt}\Big|_{\mathbf{n}=\text{constant}} &= \partial_t \Pi(\boldsymbol{\xi},t) + (\partial_t \boldsymbol{\xi}\cdot\nabla_\xi)\Pi(\boldsymbol{\xi},t),\\ &= \partial_t \Pi(\boldsymbol{\xi},t) - \Omega^{1/2}(\partial_t \boldsymbol{\phi}(t)\cdot\nabla_\xi)\Pi(\boldsymbol{\xi},t),\end{aligned}$$

where we have simply used chain rule on the first line. The third step is to expand the propensities in terms of Ω. We know that in mass-action kinetics the largest power of Ω expected in $\mathbf{f}$ is of order one (for a zeroth-order reaction), and all integer powers of Ω below this are possible, hence we expand out the propensities as such [81, 8],

$$\mathbf{f}(\mathbf{n},\Omega) = \Omega \sum_{i=0}^{\infty} \Omega^{-i}\mathbf{a}_i(\mathbf{n}/\Omega). \tag{2.33}$$

Expansion of mass-action propensities in Ω then gives identification of the $\mathbf{a}_i$ (see [102] for mass-action identification up to second-order reactions), although this process can be easily extended for non mass-action propensities including the Michaelis-Menten reaction with a Hill-type propensity [103]. For the purpose of the LNA we only need the expansion of Eq. (2.33) up to order Ω^1, and one identifies $\mathbf{a}_0(\mathbf{n}/\Omega) = \mathbf{g}(\boldsymbol{\phi}(t))$—which is intuitive since we already know $\mathbf{g}(\mathbf{x}) = \lim_{\Omega\to\infty}(\Omega^{-1}\mathbf{f}(\mathbf{n},\Omega))$. Now one can Taylor expand $\mathbf{f}(\mathbf{n},\Omega)$ about $\mathbf{n} = \Omega\boldsymbol{\phi}(t)$,

$$\mathbf{f}(\mathbf{n},\Omega) \sim \Omega\mathbf{g}(\mathbf{x}) = \mathbf{g}(\boldsymbol{\phi}(t)) + \Omega^{-1/2}(\boldsymbol{\xi}\cdot\nabla_\phi)\mathbf{g}(\boldsymbol{\phi}(t)) + \mathcal{O}(\Omega^{-1}), \tag{2.34}$$

where we have defined $\nabla_{\phi_i} \equiv \partial_{\phi_i}$. Finally, we use the ansatz to identify $\nabla = \Omega^{-1/2}\nabla_\xi$, where $\nabla_{\xi_i} \equiv \partial_{\xi_i}$. We can now use all these elements of the SSE in the CFPE, upon which matching powers of Ω on either side of the transformed CFPE gives,

$$\begin{aligned}&\Omega^{1/2}:\ \partial_t\boldsymbol{\phi}(t) = \mathbf{S}\cdot\mathbf{g}(\boldsymbol{\phi}(t)),\\ &\Omega^{0}:\ \partial_t\Pi(\boldsymbol{\xi},t) \overset{\text{esc}}{=} \left(-\nabla_{\xi_i}\xi_j\nabla_{\phi_j}S_{ir}g_r(\boldsymbol{\phi}(t)) + \frac{1}{2}\sum_{r=1}^{R} S_{kr}S_{lr}g_r(\boldsymbol{\phi}(t))\nabla_{\xi_k}\nabla_{\xi_l}\right)\Pi(\boldsymbol{\xi},t),\end{aligned}$$

where esc implies the use of Einstein summation convention and we have prescribed the sum over r in line 2 since it is a contraction over three r indices. The terms of order $\Omega^{1/2}$ turn out to be simply the rate equations, whereas the terms of order Ω^0 describe a linear FPE equation in $\Pi(\boldsymbol{\xi},t)$. Often, one defines the expressions $J_{ij} = \nabla_{\phi_j}S_{ir}g_r(\boldsymbol{\phi}(t))$ (the Jacobian of the rate equations) and $D_{ij} = \sum_r S_{ir}S_{jr}g_r(\boldsymbol{\phi}(t))$ (the diffusion matrix)

which gives the more concise form,

$$\partial_t \Pi(\boldsymbol{\xi}, t) \overset{\text{esc}}{=} \left(-J_{ij} \nabla_{\xi_i} \xi_j + \frac{1}{2} D_{kl} \nabla_{\xi_k} \nabla_{\xi_l} \right) \Pi(\boldsymbol{\xi}, t). \tag{2.35}$$

This linear FPE has a Gaussian solution, with means of $\boldsymbol{\xi}$ given by the Jacobian of the rate equations, i.e., $\partial_t \langle \boldsymbol{\xi} \rangle = \mathbf{J} \cdot \langle \boldsymbol{\xi} \rangle$, and a covariance matrix $C_{ij} = \langle \xi_i \xi_j \rangle$ determined by the following Lyapunov equation [87, 102],

$$\partial_t \mathbf{C} = \mathbf{J} \cdot \mathbf{C} + \mathbf{C} \cdot \mathbf{J}^T + \mathbf{D}. \tag{2.36}$$

Note that the formal solution of the Jacobian is given by the matrix exponential $\langle \boldsymbol{\xi}(t) \rangle = \exp(t\mathbf{J}) \cdot \langle \boldsymbol{\xi}(0) \rangle$, although this can generally be simplified down to a sum of exponentials (where the argument of the exponentials is t times the eigenvalues of $\mathbf{J}$). Using van Kampen's definition of the steady state correlation matrix $K_{ij}(t) = \langle (n_i(0) - \langle n_i(\infty) \rangle)(n_j(t) - \langle n_j(\infty) \rangle) \rangle$, we find in the LNA that,

$$\begin{aligned} K_{ij}(t) = \Omega^{-1} \langle \xi_i(0) \xi_j(t) \rangle &= \Omega^{-1} \langle \xi_i(0) \langle \xi_j(t) \rangle \rangle, \\ &= \Omega^{-1} \langle \xi_i(0) [\exp(t\mathbf{J})]_{ij} \xi_j(0) \rangle, \end{aligned} \tag{2.37}$$

where $\xi_i(0)$ is some steady state fluctuation about $\langle n_i(\infty) \rangle$ at time 0, and $\xi_j(t)$ is some steady state fluctuation about $\langle n_j(\infty) \rangle$ at t (given the same initial state). Identification of $\langle \xi_i(0) \xi_j(t) \rangle = \langle \xi_i(0) \langle \xi_j(t) \rangle \rangle$ simply comes from the independence of the time evolution and the initial condition—as is clear from the Jacobian, since $\partial_t \langle \boldsymbol{\xi} \rangle = \mathbf{J} \cdot \langle \boldsymbol{\xi} \rangle$. Diagonal elements of $\mathbf{K}(t)$ are the autocorrelations, whereas off-diagonal components describe the cross-correlations between species. Since we have calculated this assuming steady state (even at $t = 0$), when $\exp(t\mathbf{J})$ is known more explicitly one can then use the steady state variances, $\mathbf{C}$, calculated from $\mathbf{J} \cdot \mathbf{C} + \mathbf{C} \cdot \mathbf{J}^T + \mathbf{D} = 0$ in place of $\langle \xi_i(0) \xi_j(0) \rangle$.

The LNA is used to calculate correlation functions between species in complex gene expression models in Section 4.6.

2.4.3 Deterministic based approximations

A very common method of approximating complicated stochastic dynamics is to use modified reaction schemes whose derivation arises from deterministic rate equations. The two most popular methods, the quasi equilibrium approximation (QEA) and the quasi steady state approximation (QSSA), have been used widely in approximating enzyme kinetics (see Chapter 6 and [86]) and genetic autoregulation (see Chapters 3, 4 and [104, 105, 106, 107, 108] for some examples). They have birthed several other approximation methods, notably the total QSSA [109] and the pre-factor QSSA [110, 111]. Notably, the QEA, QSSA and all their derivatives are all deterministic

approximations, which are then applied *ad hoc* to a stochastic model. Their usage bypasses more rigorous concerns regarding what constitutes a fast or slow species, being reliant only on the choice of the user in how they are integrated into a stochastic model. Later in Chapter 3, we look at exactly when the QEA is valid in the stochastic setting applied to autoregulation. For the moment, we introduce the QEA and QSSA, applied to Michaelis-Menten kinetics,

$$S + E \underset{k_1}{\overset{k_0}{\rightleftharpoons}} C \xrightarrow{k_2} E + P. \tag{2.38}$$

The deterministic rate equations for the Michaelis-Menten scheme is as follows,

$$\frac{d[S(t)]}{dt} = -k_0[S(t)]([E]_0 - [C(t)]) + k_1[C(t)], \tag{2.39}$$

$$\frac{d[C(t)]}{dt} = -(k_1 + k_2)[C(t)] + k_0[S(t)]([E]_0 - [C(t)]), \tag{2.40}$$

where terms in the $[\cdot]$ indicate the deterministic concentrations of the species in the bracket. We now detail the deterministic QEA and the QSSA approximations. The QEA proceeds by considering the reaction $S + E \rightleftharpoons C$ to be at equilibrium for all concentrations of $[S(t)]$, which intuitively means that the rates of the reversible reaction are much faster than the rate of product formation. This means the following is assumed,

$$\begin{aligned} k_0[S(t)]([E]_0 - [C(t)]) &\approx k_1[C(t)], \\ [C(t)] &\approx \frac{[S(t)][E]_0}{[S(t)] + k}, \end{aligned} \tag{2.41}$$

where $k = k_1/k_0$. Note that the QEA does imply that $d[S(t)]/dt \approx 0$, which effectively states that the concentration of substrate is assumed to always take the value that enforces the quasi-equilibrium of the reversible reaction. In other words, as the concentration of $[C(t)]$ changes due to product formation, the concentration $[S(t)]$ changes instantaneously to enforce the QEA. The QEA then states that,

$$\frac{d[P[(t)]}{dt} = k_2[C(t)] \overset{\text{QEA}}{\approx} \frac{V_{\max}[S(t)]}{[S(t)] + k}, \tag{2.42}$$

where $V_{\max} = k_2[E]_0$ is the maximum possible velocity of product formation. On the other hand, the QSSA does not assume equilibrium in the complex formation, but instead that $[C(t)]$ is at steady state for all concentrations of $[C(t)]$, i.e., $\partial_t[C(t)] \approx 0$. The result is the same as that of the QEA, but now $k = (k_1 + k_2)/k_0$. Intuitively, the QEA states that the reactions in the reversible reaction are fast, whereas the QSSA states that C is a fast species. Both give similar, albeit distinct, results.

The stochastic QEA and QSSA then proceed by asking the question: What stochastic kinetics approximately give rise to the deterministic kinetics of the QEA or QSSA? The most obvious choice of the user is to approximate the full reaction scheme in Eq. (6.1) by the reduced system,

$$S \xrightarrow{\frac{V_{\max}}{n+k}} \varnothing, \tag{2.43}$$

where n is the number of molecules of S, and where we have set the system size $\Omega = 1$ such that $[S(t)] = n$. From the CME of this scheme, one then determines the equation for the evolution of the mean of n as,

$$\partial_t \langle n(t) \rangle = -\left\langle \frac{V_{\max} n(t)}{n(t)+k} \right\rangle \approx -\frac{V_{\max}\langle n(t)\rangle}{\langle n(t)\rangle + k}, \tag{2.44}$$

where the final approximation is valid where fluctuations of $\langle n(t)\rangle$ are small compared its magnitude. One can explicitly conduct a small-noise expansion [112] to verify this by setting $n(t) = \langle n(t)\rangle + \delta n(t)$ with $\delta n(t) = n(t) - \langle n(t)\rangle$, and expanding in $\delta n(t)$ to give,

$$\left\langle \frac{n(t)}{n(t)+k} \right\rangle = \frac{\langle n(t)\rangle}{\langle n(t)\rangle + k} - \frac{k\sigma_n(t)^2}{(\langle n(t)\rangle + k)^3} + \mathcal{O}(\delta n^3), \tag{2.45}$$

where we have recognised $\langle \delta n(t)\rangle = 0$ and the variance of $n(t)$ as $\sigma_n(t)^2 = \langle \delta n(t)^2\rangle$. Therefore, when $\sigma_n(t)^2 \ll \langle n(t)\rangle$ Eq. (2.44) can be assumed to a good approximation. Since we have recovered the same equation for the evolution of the mean, we call this the stochastic QEA (or QSSA). However, there is no guarantee that the original reaction scheme will agree with the reduced system *beyond the mean level*. Conveniently, although the validity of the Hill function has not been understood mechanistically, it does generally, but not always, provide a good approximation of stochastic dynamics [110].

In Chapters 3 and 6 we explore the validity of the stochastic QEA applied to autoregulation and Michaelis-Menten enzyme kinetics respectively. In the case of autoregulation we realise that the validity of the stochastic QEA is not only related to the speed of the reversible switching rates, but also due to finite molecule number effects that are neglected on the deterministic level.

2.4.4 Averaging

Another approximative method based on time scale separation is 'averaging', for which a comprehensive review is given in [113]. The essential idea relies on the separation of Markov dynamics into blocks of *fast dynamics* and *slow dynamics*. What defines the fast dynamics is that they are assumed to reach steady state immediately given any change in the system due to the slow dynamics. This reduces the complexity of the full system since the calculation of transient solutions is then determined by two easier to solve problems: (1) a steady state solution for the fast dynamics and (2) a transient solution for the slow dynamics (as a function of the steady state of the fast dynamics).

Notably, averaging is much lesser known as a time scale separation method than other methods in the literature such as *singular perturbation theory*. Singular perturbation theory allows a user to rigorously invoke a physical/biological time scale separation on a set of ODEs or PDEs defining a the dynamics of a physical system in situations for which regular perturbation theory fails—e.g., the three-stage model of gene expression in conditions where mRNA is assumed to degrade on a much faster time scale than that of the proteins (leading to 'bursty' protein expression, see [38, 114]). The main differences between singular perturbation theory and averaging lie both in the regimes of applicability of each technique and in the quantification of errors. Singular perturbation theory requires the identification of a small expansion parameter, whereas averaging does not require such identification—in averaging, the time scale separation is applied to the Markov state diagram directly, where one can then identify the groups that are effectively in a quasi steady state. This often allows one to express the dynamics of multi-species systems in terms of an effectively reduced number of species, meaning that analytic techniques only applicable for low-dimensional systems can be used. However, the expansion parameter defining the singular perturbation expansion allows for error quantification, something that cannot currently be obtained through the method of averaging.

A substantial amount of time is given to discussing averaging in Chapter 6 in the application to enzyme kinetics, so we will limit our discussion here. Notably, averaging has been applied to an array of other stochastic problems, including autoregulation [115] and multi-state gene models [116].

2.5 Delayed SSA and CME

Biological systems are not always best modelled by Markov processes. Even though one can generally model a system as Markovian by introducing enough Markov states, this quickly becomes more challenging to solve. For this reason, other modelling methods may be more efficient. One recently popular non-Markovian modelling approach are *deterministically delayed reactions* [73, 117, 118, 119, 120], reactions that fire after a deterministic time τ which can be incorporated in conjunction with Markov reactions in a given reaction network. Although seemingly simple, their non-Markovian nature means that we cannot use the SSA and CME as stated above and must be more careful in our consideration of the history of the stochastic process. In the following Sections we discuss how to modify the SSA and the CME to account for this.

2.5.1 Delayed SSA

In addition to the general reaction scheme of Markov reactions in Eq. (2.1) we now also have a set of deterministically delayed reactions $\mathcal{D}$ (of length D) which we denote by,

$$\sum_{i\in\mathcal{D}} c_{ir}X_i \underset{\tau_r}{\Longrightarrow} \sum_{i\in\mathcal{D}} d_{ir}X_i. \tag{2.46}$$

for $r = 1, 2, \ldots, D$, where c_{ir} and d_{ir} are the stoichiometric coefficients of the delayed reactions, and the τ_r under the double-lined arrow is the time of the *deterministic* delay for reaction r. Further, we define the stoichiometry matrix for the delayed reactions as $\boldsymbol{\sigma} = \mathbf{d} - \mathbf{c}$. Now, at any time in the simulation of the full reaction scheme, including Markovian and delayed reactions, there will be a set of M delayed reactions $\mathcal{T} = \{T_1, T_2, \ldots, T_M\}$ ordered such that $T_i < T_{i+1}$. Given that the Markov reactions follow the same dynamics as before (from pseudocode in 2.1 where τ and r are drawn from there respective probability distributions), the question then becomes: How do we include the dynamics of the delayed reactions? For example, if we have some scheduled Markov reaction at $t + \tau$, but a delayed reaction will occur at $T_1 < t + \tau$, then after we've fired the delayed reaction, do we fire the Markov one too?

To answer this question, a rigorous derivation of the SSA for reaction schemes including delays was conducted in [121], where it is found that *algorithm 2* previously used in [117] gives the correct method of incorporating delays. The essential addition is that if $T_1 < t + \tau$, where τ is the drawn time at which the next Markov reaction would occur, then we reject the Markov reaction at $t + \tau$ and simply fire the delayed reaction at T_1. This is conveniently referred to as the *rejection method* by [121]. The pseudocode for this addition to the SSA (the dSSA) is then [117, 121],

1. Instantiate the initial state, $\mathbf{n} = \mathbf{n}_0$, of the system at $t = t_0$.
2. Draw a Markov reaction waiting time τ from $p(\tau|\mathbf{n}, t)$.

3. If $\mathcal{T}$ is an empty set move to Step 4. Otherwise, compare Markov reaction firing time $t+\tau$ to next upcoming delay reaction time T_1. If $t+\tau < T_1$ then the next reaction will be Markovian and move to Step 4. Else, the next reaction fired is the delay at $t = T_1$: set r to the label corresponding to the type of delayed reaction and move to Step 5 (and discard τ).
4. Draw the reaction r that is fired from $p_r(\mathbf{n})$.
5. Update the state vector for the stoichiometry. For a Markov reaction: update $\mathbf{n} \to \mathbf{n} + \mathbf{S}_r$ (where $\mathbf{S}_r$ is the r^{th} column of $\mathbf{S}$) and the reaction time $t \to t+\tau$ and store the it the matrix $\mathbf{n}(t)$. For a delayed reaction: update $\mathbf{n} \to \mathbf{n} + \boldsymbol{\sigma}_r$ (where $\boldsymbol{\sigma}_r$ is the r^{th} column of $\boldsymbol{\sigma}$) and the reaction time $t \to T_1$ and store it in the matrix $\mathbf{n}(t)$.
6. If the firing of $\mathbf{S}_r$ or $\boldsymbol{\sigma}_r$ initiates a future delayed reaction(s), then add the time(s) $T_r = t + \tau_r$ to $\mathcal{T}$ (and reorder $\mathcal{T}$).
7. Repeat process from Step 2 until some specified maximum time T is reached and output $\mathbf{n}(T)$.

As before, the Markov reactions can be simulated using the direct method, where one only needs to draw two random numbers on the unit interval (see Section 2.1). For the readers information, a package has recently been written in Julia that allows one to easy code up the dSSA for any reaction scheme [122]. Intuitively, the reason why this algorithm is exact comes from the memoryless property of Markov reaction, which states that,

$$P(t+\tau) = P_{>}(T_1)P(t+\tau-T_1), \tag{2.47}$$

where $T_1 < t+\tau$, $P(t)$ is the probability density of a reaction firing at t and $P_{>}(t)$ is the probability of the reaction firing after time t, i.e., $P_{>}(t) = \int_t^{\infty} P(t)dt$, and it is assumed the state of the system is the same over $[t, t+\tau)$. Hence, rejecting the drawn time τ in Step 3 does not affect the dynamics of the Markovian reactions (providing the delayed reaction at T_1 does not effect the propensity of the Markov reactions). Explicitly, one can verify using $P(t) = f_0(\mathbf{n}) \exp\left(f_0(\mathbf{n})t\right)$ that Eq. (2.47) follows. In the case where the delayed reaction at T_1 does effect the propensity of the Markov reactions the scheme is still exact, but it now accounts for the change of state caused by the delayed reactions.

We use the dSSA for modelling the elongation of nascent mRNA in Chapter 5.

2.5.2 Delayed CME

As we have seen, the dSSA is quite different from the standard SSA, and the same is true for the delayed CME (dCME) compared to the standard CME. In particular, one must be much more careful in the consideration of the history of a stochastic process with delays because the system is no longer Markovian. We explore how to account for this non-Markovian nature below in Chapter 5, where we solve the three-state gene model with transcriptional elongation modelled as a deterministic delay. The general form of the dCME is quite complicated (see Eq. (3.3) in [118]), so the dCME, as for the CME, is solved on a case-by-case basis.

2.6 Transient solutions of the CME

In this final Section of the preliminaries we detail two methods with which we solve the CME in time later in the thesis. Solving a CME in time means that, starting from a specified initial condition, one can see how the system relaxes towards the steady state. This type of analysis is less common that steady state analyses for two reasons. First, time-dependent problems are more difficult to solve and require more advanced methods. Second, it is often assumed that biological and economical processes inhabit the steady state. However, permanent steady states do not exist in nature, and often a greater understanding of complex systems can be obtained by seeing how they relax towards their steady state (information not contained in the steady state itself). Therefore, how systems respond to changes gives us a way to access more information regarding their underlying structure. This is the reason why perturbation experiments (experiments that perturb the cell away from its steady state) are popular in molecular biology [123, 124, 125].

2.6.1 Eigenfunction expansion and determination of eigenvalues

Beginning with the CME, it is clear that one can write it as a matrix equation of the form,

$$\partial_t \mathbf{P}(t) = \mathbf{M} \cdot \mathbf{P}(t), \tag{2.48}$$

where $\mathbf{P}(t)$ is a vector where each element corresponds to the probability of being in a corresponding state of $\mathbf{n}$, and $\mathbf{M}$ is the master operator which contains all the dynamics of the CME coming from $\mathbf{f}(\mathbf{n})$ (if unclear see Section 2.3). This form of the CME invites us to consider an eigendecomposition of $\mathbf{M}$, wherein we must calculate its eigenvalues $\{-\lambda\}$ and eigenfunctions $\mathbf{\Phi}_\lambda$. The eigenvalue equation is $\mathbf{M} \cdot \mathbf{\Phi}_\lambda = \partial_t \mathbf{\Phi}_\lambda = -\lambda \mathbf{\Phi}_\lambda$, from

which we find,

$$\mathbf{\Phi}_\lambda(t) = \boldsymbol{\phi}_\lambda \exp(-\lambda t). \tag{2.49}$$

We can now choose to express $\mathbf{P}(t)$ in terms of these eigenfunctions,

$$\mathbf{P}(t) = \sum_\lambda C_\lambda \boldsymbol{\phi}_\lambda \exp(-\lambda t), \tag{2.50}$$

where $\boldsymbol{\phi}_\lambda$ are λ dependent vectors of the same dimension of $\mathbf{P}$, and the C_λ are constants determined from the initial condition(s),

$$\mathbf{P}(0) = \sum_\lambda C_\lambda \boldsymbol{\phi}_\lambda. \tag{2.51}$$

Note that the expansion of $\mathbf{P}(t)$ in $\mathbf{\Phi}_\lambda$ assumes that $\mathbf{\Phi}_\lambda$ form a complete basis, i.e., that each $\boldsymbol{\phi}_\lambda$ is a linearly-independent basis function for each λ. If one can then define an inner product such that the $\boldsymbol{\phi}_\lambda$ are orthonormal then one can determine the C_λ by projecting $\mathbf{P}(0)$ onto each $\boldsymbol{\phi}_\lambda$. Additionally, from Perron-Frobenius theorem one can state that, providing a system is ergodic [126], the eigenvalues $\{-\lambda\}$ satisfy the following,

$$\lambda_0 = 0 < \mathrm{Re}(\lambda_1) \leq \mathrm{Re}(\lambda_2) \leq \ldots, \tag{2.52}$$

where the eigenvalue $\lambda_0 = 0$ corresponds to the steady state eigenfunction,

$$P_s(n) = P(n, t \to \infty) = C_0 \phi_0(n). \tag{2.53}$$

C_0 is simply a constant determined by normalisation of the steady state distribution.

In general, it is a difficult task to solve CMEs of two species via the eigenfunction expansion method. Now consider the one-dimensional CME describing the dynamics of a single species. The probability vector is simply given by $[\mathbf{P}(t)]_n = P(n, t)$, where n denotes the number of molecules of the only available species. As we have done previously, we can then introduce the generating function $G(z, t) = \sum_n z^n P(n, t)$, into which we can substitute the time-dependent form of $P(n, t)$ from Eq. (2.50) giving,

$$\begin{aligned} G(z, t) &= \sum_\lambda C_\lambda \exp(-\lambda t) f(z, \lambda), \\ f(z, \lambda) &= \sum_{n=0}^{\infty} z^n \phi_\lambda(n). \end{aligned} \tag{2.54}$$

One can use this form of $G(z,t)$ in the time-dependent generating function equations (for problems with a single species), transforming the generating function PDE for $G(z,t)$ in terms of z and t derivatives into an ODE in terms of z derivatives of $f(z,\lambda)$ alone, with the additional introduction of the spectral parameter λ.

There are then three aspects of this problem to solve, which depend on the specifics of the model:

1. Can one find the set of eigenvalues $\{-\lambda\}$ such that each $f(z,\lambda)$ is a physically admissible function.
2. Can one find the form of $f(z,\lambda)$ defined by the ODE?
3. Can one define an orthogonality condition on the on the functions $f(z,\lambda)$ such that the constants C_λ can be determined?

In Chapter 7 we show how to solve Kirman's model of ant rationality in time, as well as more complex versions of the model, by application of the above methods.

2.6.2 General transient solution to 1D 1-step master equation

In this section we show how one can analytically solve the master equation using a method from [127], which uses Cauchy's integral formula on the resolvent of the master operator. We note this same method has been derived with respect to the Laplace transform [128, 129]. The utility of this method is that by determining the eigenspectrum of the master operator one completely specifies the time-dependent solution for the probability distribution of the stochastic process.

We now detail the essential steps from the method of [127]. We first point out that the formal solution (using the matrix notation introduced above) for $\mathbf{P}(t)$ from Eq. (2.48) can be given as a matrix exponential in $t\mathbf{M}$,

$$\mathbf{P}(t) = e^{t\mathbf{M}} \cdot \mathbf{P}(0). \tag{2.55}$$

In order to calculate the eigenvectors of $\mathbf{M}$ necessary for the time-dependent solution, we can now employ Cauchy's integral formula for matrices [130], explicitly,

$$\mathbf{f}(\mathbf{M}) = \frac{1}{2\pi i} \oint_\gamma (z\mathbf{I} - \mathbf{M})^{-1} \cdot \mathbf{f}(z) dz, \tag{2.56}$$

where γ is a contour that contains all the eigenvalues of $\mathbf{M}$ and $\mathbf{I}$ is a $M \times M$ identity matrix, where M is the size of the state space (the number of unique state vectors). Choosing $\mathbf{f}(\mathbf{M}) = e^{t\mathbf{M}} \cdot \mathbf{P}(0)$, Cauchy's integral formula then gives us the solution for $\mathbf{P}(t)$,

$$\mathbf{P}(t) = \frac{1}{2\pi i} \oint_\gamma e^{zt} (z\mathbf{I} - \mathbf{M})^{-1} \cdot \mathbf{P}(0) dz, \tag{2.57}$$

where we assume the initial condition $[\mathbf{P}(0)]_n = \delta_{n,n_0}$, i.e, the initial state is given by the n_0-th element of the state vector. Using this initial condition in Eq. (2.57) gives us,

$$[\mathbf{P}(t)]_n = \frac{1}{2\pi i} \oint_\gamma e^{zt} [(z\mathbf{I} - \mathbf{M})^{-1}]_{n,n_0} dz \tag{2.58}$$

We now state Cramer's rule for matrix inverses which is given in its general form for some matrix $\mathbf{A}$ as $\mathbf{A}^{-1} = \text{adj}(\mathbf{A})/\det(\mathbf{A})$, where $\text{adj}(\mathbf{A})$ is the adjugate matrix (formally the transpose of the cofactor matrix) and $\det(\mathbf{A})$ is the determinant. In our case,

$$\det(z\mathbf{I} - \mathbf{M}) = \prod_{i=1}^{M} (z + \lambda_i),$$

and we denote the adjugate matrix of $z\mathbf{I} - \mathbf{M}$ as $\mathbf{B}(z)$, and the upper limit on the sum is M since there are M eigenvalues of $\mathbf{M}$. This gives us,

$$[\mathbf{P}(t)]_n = \frac{1}{2\pi i} \oint_\gamma \frac{e^{zt}}{\prod_{i=1}^{M}(z + \lambda_i)} \mathbf{B}(z)_{n,n_0} dz, \tag{2.59}$$

where $\mathbf{B}(z)_{n,n_0}$ is a polynomial in z and can be determined using standard methods [130], including a simple iterative formula for the case of tridiagonal $\mathbf{M}$, i.e., for a one-step birth death process in one variable [131]. The integrand in Eq. (2.59) has M simple poles, each centred at the eigenvalues of $\mathbf{M}$, and hence where Cauchy's integral formula is given by,

$$\oint_\gamma f(z) dz = 2\pi i \sum_m \text{Res}(f, a_m),$$

where the sum is over all poles a_m of $f(z)$, our final result for $[P(t)]_n$ is given by,

$$[\mathbf{P}(t)]_n = \sum_{m=1}^{M} \left\{ e^{-\lambda_m t} \frac{\mathbf{B}(-\lambda_m)_{n,n_0}}{\prod_{j \neq m}(\lambda_j - \lambda_m)} \right\}. \tag{2.60}$$

This completes the derivation from [127]. We use this method in Chapters 6 and 7 to provide a practical way to calculate time-dependent probability distributions for enzyme kinetics and ant rationality models respectively.

Chapter 3

Revisiting the reduction of stochastic models of genetic feedback loops with fast promoter switching

This chapter has been published as [1] entitled *Revisiting the reduction of stochastic models of genetic feedback loops with fast promoter switching* in the *Biophysical Journal*. Slight modifications have been made for its inclusion in this thesis.

3.1 Abstract

Propensity functions of the Hill-type are commonly used to model transcriptional regulation in stochastic models of gene expression. This leads to an effective reduced master equation for the mRNA and protein dynamics only. Based on deterministic considerations, it is often stated or tacitly assumed that such models are valid in the limit of rapid promoter switching. Here, starting from the chemical master equation describing promoter-protein interactions, mRNA transcription, protein translation and decay, we prove that in the limit of fast promoter switching, the distribution of protein numbers is different than that given by standard stochastic models with Hill-type propensities. We show the differences are pronounced whenever the protein-DNA binding rate is much larger than the unbinding rate, a special case of fast promoter switching. Furthermore we show using both theory and simulations that use of the standard stochastic models leads to drastically incorrect predictions for the switching properties of positive feedback loops and that these differences decrease with increasing mean protein burst size. Our results confirm that commonly used stochastic models of gene regulatory networks are only accurate in a subset of the parameter space consistent with rapid promoter switching.

3.2 Introduction

Many biochemical systems have one or more species with low molecule numbers which implies that the dynamics can be highly noisy and consequently a deterministic description may not be accurate [132, 133, 134, 135]. Rather a more appropriate mathematical description is stochastic and given by the chemical master equation [8]. When the system is made up of zero and first order reactions only, exact solutions at both steady state and in time are occasionally possible [136]. However, many systems have at least one bimolecular reaction and in such cases only a few exact steady state solutions of the CME are known (see for example [42, 75, 137, 138]). A common example of such systems are auto-regulatory feedback loops, whereby a protein produced by a gene binds to its own promoter region to activate or suppress its own production [11, 139, 140]. In the absence of exact solutions, we become either reliant: (i) on the stochastic simulation algorithm (SSA) [68] or (ii) on approximations of the original network so that analytic results become tractable [81, 89, 91]. Generally, it is a challenge to utilise approximations to simplify the CME such that the resulting reduced equation is representative of the true system dynamics.

A common set of approximation methods are based on time scale separation. At the microscopic level, there are several different scenarios which can lead to time scale separation conditions. Depending on the propensity at which reaction are fired, reactions can be classified as either slow or fast. Depending on the reaction system, it is possible that fast and slow reactions do not involve the same species but more commonly fast and slow reactions share some species and hence it is generally unclear what should be considered a fast or a slow species. Methods in the literature differ according to the definition of what is a slow or fast species. Zeron and Santillan [141] assume that fast species are only involved in fast reactions whereas slow species can participate in both slow and fast reactions. In contrast, Cao *et al.* [142] define slow species as those involved in slow reactions only and fast species as those participating in at least one fast reaction and any number of slow reactions. These two approaches lead to a reduced CME description in the slow species only. Other approaches due to Haseltine and Rawlings [143] and Goutsias [144] model the state of the system using extents of reaction (i.e., the count of the number of times each reaction has fired) as opposed to molecules of species. A singular perturbation theory based method has also recently been used to obtain a reduced stochastic description [145]. There are also several formal results that have been mathematically proven for reaction systems in various scenarios [146, 147, 148, 149].

Despite the wide breadth of rigorous approaches (e.g., perturbation theory [150]), by far the most popular approach in the literature of computational and systems biology to obtain a reduced master equation is heuristic. The key idea is to use the results of time scale separation for deterministic kinetics. Under the quasi-steady state or fast equilibrium approximations, the mean concentration of a subset of species (the fast species) reaches steady state on a much shorter time scale than the rest of the species (the slow species). Using the deterministic rate equations it is then possible to express the concentration of the fast species in terms of the concentration of the slow species. This leads to a reduced chemical system composed of effective reactions with non-mass action kinetics describing the dynamics of the slow species. The reduced chemical master equation is then obtained by writing effective propensities analogous to the non-mass action reaction rates obtained from the deterministic analysis. For example, Hill-type effective protein production rates in the deterministic rate equations result if the gene equilibrates on a much faster time scale than mRNA and protein, i.e., the fast promoter switching limit (see for example [33] for experimental evidence of this limit), and hence by analogy, Hill-type propensities for the protein production rates are commonly used in stochastic simulations of gene regulatory networks [105, 106, 107, 108, 15, 151, 152, 153]. All of these studies and many others assume that such effective propensities are justified in the limit of fast promoter switching.

The advantage of this heuristic approach is its simplicity and ease of use and this is the main reason for its widespread use. However, clearly the use of a reduced master equation obtained from deterministic considerations is doubtful. This has led to a number of studies evaluating the accuracy of these reduced master equations. Thomas *et al.* in a series of papers [84, 154, 155] showed using Langevin approximation theory that in the limit of large molecule numbers and for parameters consistent with the quasi-steady state approximation, the mean number of molecules of slow species predicted by the reduced master equation agrees with that predicted from the master equation of the full system but the variance of molecule number fluctuations does not. Similar results have been shown using stochastic simulations by Kim *et al.* (in particular see Fig. 1 of [156]). In contrast, Bundschuh *et al.* [157] have shown that the SSA corresponding to the heuristic reduced master equation of a negative feedback loop, whereby the DNA-protein binding reactions are assumed fast compared to the rest of the reactions and hence are eliminated, is in very good agreement with the SSA of the full system for parameter values specific to the phage λ system (this case is referred to as "Michaelis-Menten system" in their paper). At first sight the results of Thomas *et al.* and Bundschuh *et al.* may appear contradictory but in reality they are not: while the results of Thomas *et al.* prove that the heuristic approach of obtaining reduced master equations cannot be considered equivalent to the stochastic version of the quasi-steady state approximation (see also [158]), nevertheless it is possible that the error in the predictions of the heuristic

approach are small for specific parameter values which would be consistent with the results of Bundschuh *et al.* What is clear from these studies is that more work is needed to identify the precise regions of parameter space where the heuristic reduced master equation of gene regulatory networks can be safely used—the study reported in this chapter identifies such regions and hence fills a gap in the literature.

The structure of this chapter is as follows. In Section 3.3 we obtain the steady state solution of the heuristic reduced master equation with Hill-type protein production propensities for a non-bursty genetic feedback loop and prove that it is different than the solution of the master equation of the genetic feedback loop in the limit of fast promoter switching. It is then shown that the differences between the probability distributions of protein numbers predicted by the two master equations tend to zero when the rate of DNA-protein binding is much smaller than the unbinding rate. In contrast, the differences maximise when the rate of DNA-protein binding is much larger than the unbinding rate. The results are confirmed by stochastic simulations across large swathes of parameter space. In Section 3.4 we extend the analysis to 'bursty' feedback loops, which effectively account for the presence of fast degrading mRNA by modelling protein production as occurring in geometrically distributed bursts. We finish with Section 3.5, providing a discussion and then by concluding our results.

3.3 Model reduction for non-bursty feedback loops

3.3.1 Deterministic description and reduction

The reaction scheme for a genetic non-bursty feedback loop is given by:

$$G \xrightarrow{\rho_u} G + P,\ P \xrightarrow{1} \varnothing,\ P + G \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*,\ G^* \xrightarrow{\rho_b} G^* + P. \tag{3.1}$$

This models the production of proteins, their degradation, DNA-protein binding and unbinding. For simplicity we do not have an mRNA description (though this will be added in Section 3.4). The gene can be in one of two states: an unbound state G and a bound state G^*. The rate of protein production depends on the gene state. Note that here we have scaled all parameters by the protein degradation rate. Of course, given that there is only one copy of the gene, one expects fluctuations to be important and a stochastic model to be the most appropriate mathematical description. However, for the moment we shall ignore the inherent stochasticity and analyze the system using a

deterministic approach. The deterministic rate equations are:

$$\frac{d\langle g(t)\rangle_d}{dt} = -\sigma_b\langle g(t)\rangle_d\langle n(t)\rangle_d + \sigma_u(1-\langle g(t)\rangle_d), \tag{3.2}$$

$$\frac{d\langle n(t)\rangle_d}{dt} = -\sigma_b\langle g(t)\rangle_d\langle n(t)\rangle_d + \sigma_u(1-\langle g(t)\rangle_d) + \rho_u\langle g(t)\rangle_d + \rho_b(1-\langle g(t)\rangle_d) - \langle n(t)\rangle_d, \tag{3.3}$$

where $\langle n(t)\rangle_d$ denotes the mean number of molecules of protein P at time t, $\langle g(t)\rangle_d$ denotes the mean number of molecules of gene G at time t and $\langle g^*(t)\rangle_d$ denotes the mean number of molecules of gene G^* at time t. Since the gene can only be in either the bound or unbound state at any one time, one may also interpret $\langle g(t)\rangle_d$ and $\langle g^*(t)\rangle_d$ as the mean fraction of time spent in either gene state respectively. These mean molecule numbers are calculated within the deterministic approximation (hence the subscript d) and will generally be different than the mean molecule numbers of the system obtained from a stochastic description of the system [102]. Note that we have used the relation $\langle g(t)\rangle_d + \langle g^*(t)\rangle_d = 1$, i.e., there is one gene copy. Note also that t is non-dimensional time, i.e., actual time multiplied by the protein degradation rate. It can also be shown (see Appendix A.1) that the deterministic equations Eqs. (3.2)–(3.3) agree with the moment equations derived from the chemical master equation under the assumption of independence of fluctuations in the protein and gene numbers.

By the fast equilibrium approximation it follows that $\partial_t\langle g(t)\rangle_d \approx 0$ (and $\partial_t\langle g^*(t)\rangle_d \approx 0$) for all times which implies, from Eq. (3.2), that $\langle g(t)\rangle_d = L/(L+\langle n(t)\rangle_d)$ where $L = \sigma_u/\sigma_b$. The definition of L is used frequently throughout the text. Substituting the latter in the right hand side of Eq. (3.3) and suppressing the time dependence (for notational convenience) we obtain:

$$\frac{d\langle n\rangle_d}{dt} \approx \frac{L\rho_u + \rho_b\langle n\rangle_d}{L+\langle n\rangle_d} - \langle n\rangle_d. \tag{3.4}$$

This is an effective time evolution equation for the protein numbers, within the deterministic approximation. This corresponds to a system with two reactions: an effective zero-order reaction modeling the transcriptional regulation of protein production, and a first-order protein degradation reaction. The rate of protein production is a function of the mean number of proteins and three special cases can be distinguished: (i) If $\rho_u > \rho_b$ then the rate of protein production decreases with increasing $\langle n\rangle_d$; this is the case of negative feedback. (ii) If $\rho_u < \rho_b$ then the rate of protein production increases with increasing $\langle n\rangle_d$; this is the case of positive feedback. (iii) If $\rho_u = \rho_b$ then the rate of protein production is independent of $\langle n\rangle_d$ and effectively there is no feedback.

Intuitively one would expect the solution of Eq. (3.4) to be an excellent approximation to the time evolution of the protein in the full model given by Eqs. (3.2-3.3), in the limit of fast promoter switching, i.e., $\min(\sigma_u, \sigma_b) \gg \max(1, \rho_u, \rho_b)$. This can be explicitly verified by calculating the ratio of gene and protein time scales, as follows. Replacing σ_u by σ_u/ϵ and σ_b by σ_b/ϵ and taking the limit of $\epsilon \to 0$, it is straightforward to show that to leading order the two eigenvalues of the Jacobian matrix of the rate equations Eqs. (3.2-3.3) evaluated at steady-state, are given by:

$$\begin{aligned}\lambda_1 &= -\frac{(\langle g\rangle_d^2 + L)\sigma_b}{\epsilon\langle g\rangle_d} + O(\epsilon^0),\\ \lambda_2 &= -\frac{L + \langle g\rangle_d^2(\rho_u - \rho_b)}{\langle g\rangle_d^2 + L} + O(\epsilon),\end{aligned} \tag{3.5}$$

where $\langle g\rangle_d$ is the steady state mean gene number given by:

$$\langle g\rangle_d = \frac{L + \rho_b - \sqrt{(L-\rho_b)^2 + 4L\rho_u}}{2(\rho_b - \rho_u)}. \tag{3.6}$$

For completeness and since we will use it later, the steady state mean protein number is given by:

$$\langle n\rangle_d = \frac{1}{2}\left(\rho_b - L + \sqrt{(L-\rho_b)^2 + 4L\rho_u}\right). \tag{3.7}$$

Note that $\lambda_{1,2}$ are negative and hence the steady state of the system is stable to small perturbations. Furthermore Eq. (3.5) shows that as $\epsilon \to 0$, $\lambda_1 \to -\infty$ and λ_2 tends to a constant. Since the time scales of decay of transients in the mean protein and gene numbers are given by the absolute of the inverse of the eigenvalues, it follows that there is clear time scale separation in the limit of fast promoter switching. Note that in the calculation above we assumed that $\rho_u \neq \rho_b$; a similar calculation for the equality case also leads to time scale separation.

Hence to summarise, a deterministic rate equation analysis shows that in the limit of fast promoter switching, the reaction scheme (3.1) composed of five reactions (four first-order reactions and a bimolecular reaction) reduces to just two reactions: an effective zero-order reaction for the production of proteins with a rate which is a function of the mean number of proteins and a first-order reaction modeling protein degradation.

3.3.2 Heuristic stochastic model reduction

As mentioned in the Section A, one of the most popular stochastic model reduction approaches consists of directly writing the chemical master equation for the reduced reaction scheme deduced from the deterministic analysis in Section 3.3.1. In particular, given there are n proteins in the system then we define the effective propensities:

$$\begin{aligned} T^+(n) &= \frac{L\rho_u + \rho_b n}{L+n}, \\ T^-(n) &= n, \end{aligned} \tag{3.8}$$

where $T^+(n)dt$ is the probability, given n proteins, that a protein production reaction increasing the number of proteins by one, will occur in the time interval $[t, t+dt)$ and $T^-(n)dt$ is the probability, given n proteins, that a protein degradation event reducing the number of proteins by one will occur in the time interval $[t, t+dt)$. These probabilities are deduced directly from the form of the effective rate equation Eq. (3.4). Essentially the probability per unit time for a particular reaction is taken to be the same as the reaction rate in the effective deterministic rate equation with $\langle n \rangle$ replaced by n. The chemical master equation for this reduced reaction scheme is then given by:

$$\frac{dP_a(n,t)}{dt} = T^+(n-1)P_a(n-1,t) + T^-(n+1)P_a(n+1,t) - (T^+(n)+T^-(n))P_a(n,t). \tag{3.9}$$

Note that we have labelled the solution of this *approximate* heuristic master equation P_a to distinguish it from the solution of the full master equation P, which we discuss in Section 3.3.4. The equations for the mean number of protein $\langle n(t) \rangle_a = \sum_n n P_a(n,t)$ can be derived from the master equation:

$$\begin{aligned} \frac{d\langle n \rangle_a}{dt} &= \langle T^+(n) \rangle_a - \langle T^-(n) \rangle_a, \\ &= \left\langle \frac{L\rho_u + \rho_b n}{L+n} \right\rangle_a - \langle n \rangle_a, \\ &\approx \frac{L\rho_u + \rho_b \langle n \rangle_a}{L + \langle n \rangle_a} - \langle n \rangle_a. \end{aligned} \tag{3.10}$$

Note that in the last line we have made use of the fact that in the limit of small protein number fluctuations n can be replaced by its average. *Hence while the selection of the propensities stems from a heuristic rule with no fundamental microscopic basis, nevertheless it guarantees equivalence between the effective equation for the time evolution of the mean protein numbers of the heuristic master equation and the reduced deterministic*

rate equation in the limit of small protein number fluctuations (since Eq. (3.4) and Eq. (3.10) are the same upon interchanging $\langle n \rangle_d$ by $\langle n \rangle_a$). Note that however for the general case of non-vanishing protein fluctuations, the mean of the heuristic stochastic model is different than that predicted by the deterministic rate equations.

The exact solution of the one variable master equation Eq. (3.9) in steady state conditions can be obtained using standard methods [74] and is given by:

$$P_a(n) = P_a(0) \prod_{y=1}^{n} \frac{T^+(y-1)}{T^-(y)} = \frac{\rho_b^n [LN]_n}{n! [L]_n M(LN, L, \rho_b)}, \tag{3.11}$$

where $N = \rho_u/\rho_b$, $[x]_n = x(x+1)...(x+n-1)$ (the Pochhammer symbol) and M is the Kummer confluent hypergeometric function. The definition of N is used frequently throughout the text. To obtain insight into the discrepancies introduced by the heuristic approach, we now study two limiting cases.

The limit of large L

This is the limit in which the rate at which proteins bind DNA is much smaller than the unbinding rate. In this limit, the propensities given by Eq. (3.8) reduce to the simpler form:

$$\begin{aligned} T^+(n) &\approx \rho_u, \\ T^-(n) &= n. \end{aligned} \tag{3.12}$$

Hence in this limit, the propensity $T^+(n)$ is independent of n and the steady state solution of Eq. (3.9) is simply a Poisson with mean ρ_u:

$$P_a(n) \approx \frac{\exp(-\rho_u)\rho_u^n}{n!}. \tag{3.13}$$

Note that this derivation is intuitive but not formally precise because we have implicitly assumed the exchange of two limits: $\lim_{L\to\infty} \lim_{t\to\infty} P_a(n,t) = \lim_{t\to\infty} \lim_{L\to\infty} P_a(n,t)$. A formal proof of this result starting from the exact solution Eq. (3.11) can be found in Appendix A.3.

The limit of small L

This is the limit in which the rate at which proteins bind DNA is much larger than the unbinding rate. In this limit, the propensities given by Eq. (3.8) reduce to the simpler form:

$$\begin{aligned} T^+(n) &\approx (\rho_u - \rho_b)\delta(0,n) + \rho_b, \\ T^-(n) &= n, \end{aligned} \tag{3.14}$$

where $\delta(0,n)$ is the Kronecker delta. Substituting these in the heuristic reduced master equation Eq. (3.9), multiplying throughout by z^n and taking the sum over n on both sides of this equation we get the corresponding generating function equation:

$$\frac{\partial G(z,t)}{\partial t} \approx ((\rho_u - \rho_b)G(0,t) + \rho_b G(z,t))(z-1) + (1-z)\frac{\partial G(z,t)}{\partial z}, \tag{3.15}$$

where $G(z) = \sum_n z^n P_a(n,t)$. In steady-state, this equation has the solution:

$$G(z) = \frac{\rho_b + \rho_u(\exp(\rho_b z) - 1)}{\rho_b + \rho_u(\exp(\rho_b) - 1)}. \tag{3.16}$$

Hence the steady state probability distribution is given by:

$$P_a(n) \approx \begin{cases} \frac{1}{1+N(\exp(\rho_b)-1)}, & \text{if } n = 0, \\ \frac{\exp(-\rho_b)\rho_b^n}{n!}\left(1 + \frac{N-1}{1+N(\exp(\rho_b)-1)}\right), & \text{if } n \geq 1. \end{cases} \tag{3.17}$$

While intuitive, this proof suffers from the same looseness with exchange of limits as with the limit of small L. An alternative rigorous proof of this result starting from the exact solution Eq. (3.11) can be found in Appendix A.3). Eq. (3.17) is clearly not a Poisson when there is positive or negative feedback ($N \neq 1$). Note that when $N = 1$, the solution is a Poisson but this case is biologically unimportant because it implies that the rate of protein production is independent of the state of the promoter (bound or free) and consequently there is effectively no feedback mechanism at play.

3.3.3 Conditions for the validity of heuristic stochastic model reduction

To obtain further insight into the conditions under which the heuristic model reduction is correct, we consider the stochastic description of the non-bursty feedback loop (3.1) but ignoring fluctuations in the protein numbers stemming from the reversible binding of protein to gene. The neglection of protein binding fluctuations corresponds to the

following reaction scheme:

$$G \xrightarrow{\rho_u} G + P,\ P \xrightarrow{1} \varnothing,\ G \underset{\sigma_u}{\overset{\sigma_b n}{\rightleftharpoons}} G^*,\ G^* \xrightarrow{\rho_b} G^* + P, \tag{3.18}$$

where n in the reaction rate denotes the number of free proteins. This is a common approximation in the literature [153, 159, 160], the rationale being that since protein numbers are typically much larger than one hence the gain or loss of one molecule via the gene binding reactions can be safely ignored. Given this assumption, the chemical master equation of the non-bursty feedback loop (3.1) can be conveniently written as a set of two coupled equations:

$$\begin{aligned}\frac{dP_0(n,t)}{dt} =& \rho_u(P_0(n-1,t) - P_0(n,t)) + ((n+1)P_0(n+1,t) - nP_0(n,t)) \\ &+ \sigma_u P_1(n,t) - \sigma_b n P_0(n,t),\end{aligned} \tag{3.19}$$

$$\begin{aligned}\frac{dP_1(n,t)}{dt} =& \rho_b(P_1(n-1,t) - P_1(n,t)) + ((n+1)P_1(n+1,t) - nP_1(n,t)) \\ &- \sigma_u P_1(n,t) + \sigma_b n P_0(n,t),\end{aligned} \tag{3.20}$$

where $P_0(n,t)$ is the probability that at time t there are n proteins and the gene is in state G while $P_1(n,t)$ is the probability that at time t there are n proteins and the gene is in state G^*. Note that time t is non-dimensional and equal to the actual time multiplied by the protein degradation rate. The probability of n proteins is then given by $P(n,t) = P_0(n,t) + P_1(n,t)$. Defining the generating functions $G_0(z,t) = \sum_n z^n P_0(n,t)$ and $G_1(z,t) = \sum_n z^n P_1(n,t)$, the generating function differential equations corresponding to Eqs. (3.19) are given by:

$$\frac{\partial G_0(z,t)}{\partial t} = \rho_u(z-1)G_0(z,t) - (z-1)\frac{\partial G_0(z,t)}{\partial z} + \sigma_u G_1(z,t) - \sigma_b z \frac{\partial G_0(z,t)}{\partial z}, \tag{3.21}$$

$$\frac{\partial G_1(z,t)}{\partial t} = \rho_b(z-1)G_1(z,t) - (z-1)\frac{\partial G_1(z,t)}{\partial z} - \sigma_u G_1(z,t) + \sigma_b z \frac{\partial G_0(z,t)}{\partial z}. \tag{3.22}$$

We can solve for the total generating function $G(z) = G_0(z) + G_1(z)$ as follows. At steady state $\partial G_i(z,t)/\partial t = 0$, and utilising the relation $G_1(z) = G(z) - G_0(z)$ we can use the sum of Eq. (3.21) and (3.22) to find $G_0(z) = G_0(G(z), G'(z), z)$ and $G_0'(z) = G_0'(G'(z), G''(z), z)$ below

$$G_0 = \frac{1}{\rho_b - \rho_u}(\rho_b G - G'), \quad G_0' = \frac{1}{\rho_b - \rho_u}(\rho_b G' - G''), \tag{3.23}$$

where we suppress the z dependence for brevity. We then substitute Eq. (3.23) into Eq. (3.21), again using $G_1 = G - G_0$, to give a second order linear differential equation in terms of G,

$$((1+\sigma_b)z-1)G''+((\sigma_u+\rho_b+\rho_u)-(\rho_u+(1+\sigma_b)\rho_b)z)G'+\rho_u(\rho_b z-\sigma_u-\rho_b)G=0. \quad (3.24)$$

This differential equation has two singularities, a regular singularity at $z = 1/(1+\sigma_b)$ and an irregular singularity at $z = \infty$ and hence satisfies the differential equation defining the ${}_1F_1(\alpha;\beta;z)$ hypergeometric function up to a change in variable (otherwise known as the Kummer function $M(\alpha;\beta;z)$). Using a change of variable and an exponential transformation we confirm this, and the solution is given as

$$G(z) = \exp\left(\frac{\rho_u(z-1)}{1+\sigma_b}\right)\frac{M(\alpha,\beta,\frac{\gamma(z(1+\sigma_b)-1)}{\sigma_b})}{M(\alpha,\beta,\gamma)}, \quad (3.25)$$

where

$$\alpha = \frac{\rho_u\sigma_b(\rho_b(1+\sigma_b)-\rho_u+\sigma_u(1+\sigma_b))}{(1+\sigma_b)^2(\rho_b-\rho_u+\rho_b\sigma_b)},\ \beta = \frac{\sigma_u+\sigma_b(\rho_u+\sigma_u)}{(1+\sigma_b)^2},\ \gamma = \frac{\sigma_b(\rho_b-\rho_u+\rho_b\sigma_b)}{(1+\sigma_b)^2}. \quad (3.26)$$

In the limit of fast promoter switching, i.e., replacing σ_u by σ_u/ϵ and σ_b by σ_b/ϵ and taking the limit of $\epsilon \to 0$, one can show that the leading-order term in the series expansion of Eq. (3.25) in powers of ϵ is given by:

$$G(z) = \frac{M[LN;L;\rho_b z]}{M[LN;L;\rho_b]}, \quad (3.27)$$

where we remind the reader of definitions $L = \sigma_u/\sigma_b$ and $N = \rho_u/\rho_u$. It is easy to show that $P(n) = (1/n!)d^nG(z)/dz^n|_{n=0}$ precisely equals Eq. (3.11).

Hence, we have shown that if protein number fluctuations due to reversible binding can be ignored then the stochastic description agrees with that of the heuristic master equation in the limit of fast promoter switching. This indeed gives some credibility to the use of the heuristic master equation and explains the widespread belief, based on stochastic simulations, that the heuristic master equation is correct in the limit of fast promoter switching. This result is however surprising when one considers that the heuristic is often justified from deterministic arguments, and that the deterministic treatment ignores the sizeable fluctuations associated with gene switching.

3.3.4 Exact stochastic model reduction

The master equation we solved in the previous section is not the exact master equation since we have ignored protein binding fluctuations. In what follows we properly take these into account. For the non-bursty feedback loop (3.1), the stochastic description is given by the chemical master equation which can be conveniently formulated as a set of two coupled equations:

$$\begin{aligned}\frac{dP_0(n,t)}{dt} =&\rho_u(P_0(n-1,t)-P_0(n,t))+((n+1)P_0(n+1,t)-nP_0(n,t)) \\ &+\sigma_u P_1(n-1,t)-\sigma_b n P_0(n,t),\end{aligned} \tag{3.28}$$

$$\begin{aligned}\frac{dP_1(n,t)}{dt} =&\rho_b(P_1(n-1,t)-P_1(n,t))+((n+1)P_1(n+1,t)-nP_1(n,t)) \\ &-\sigma_u P_1(n,t)+\sigma_b(n+1)P_0(n+1,t).\end{aligned} \tag{3.29}$$

Note that these equations are the same as Eq. (3.19) except for the terms describing protein-gene binding, i.e., those proportional to σ_b and σ_u. In the limit of fast promoter switching, i.e., replacing σ_u by σ_u/ϵ and σ_b by σ_b/ϵ and taking the limit of $\epsilon \to 0$, one can show that the steady state solution of Eqs. (3.28) (to leading-order in ϵ) is given by:

$$P(n) = \frac{(1+L)N\rho_b^n(n\rho_b + L(L+n+N\rho_b))[1+LN]_n}{AM(1+LN,1+L,\rho_b)+BM(2+LN,2+L,\rho_b)}, \tag{3.30}$$

where

$$A = (LN+n)n!(1+L)(L+(N-1)\rho_b)[1+L]_n, \tag{3.31}$$

$$B = (LN+n)n!(1+LN)\rho_b[1+L]_n. \tag{3.32}$$

See Appendix A.2 for the details of the derivation. Comparing Eq. (3.30) with the solution of the heuristic reduced master equation Eq. (3.11), it is immediately obvious that the two are not equal. *Hence it follows that the solution to the heuristic reduced master equation is generally different than the solution of the full master equation under fast promoter switching conditions, contradicting a major assumption in the literature (as discussed in the Introduction).* It is straightforward to verify that the two agree only if $N = 1$ in which case the protein production rate is the same in state G or G^* and hence there is no effective feedback mechanism.

To understand the nature of the differences between Eq. (3.11) and Eq. (3.30), we consider two limiting cases of small and large L. Whilst results in these limits can be obtained directly from consideration of Eq. (3.30) (see Appendix A.3) it is both simpler and instructive to consider a different approach which does not need the exact solution of the master equation. The advantage of this approach is that as we shall see later on, it can be easily extended to the analysis of more complex feedback systems.

Since $\langle g \rangle$ is the fraction of time spent in state G for the full stochastic model, it follows that in the limit of small L, $\langle g \rangle$ is also very small, the gene spends most of its time in state G^* and consequently the principal reactions determining the protein dynamics are $G^* \xrightarrow{\rho_b} G^* + P, P \xrightarrow{1} \varnothing$. Similarly it can be argued that in the limit of large L, $\langle g \rangle \approx 1$ (the gene spends most of its time in state G) and hence the principal reactions determining the protein dynamics are $G \xrightarrow{\rho_u} G + P, P \xrightarrow{1} \varnothing$. The master equation for both sets of principal reactions is trivial to solve and implies that the steady state protein number distribution in both limits is a Poisson:

$$P(n) \approx \frac{\exp(-\rho_u)\rho_u^n}{n!}, \text{ if } L \to \infty, \tag{3.33}$$

$$P(n) \approx \frac{\exp(-\rho_b)\rho_b^n}{n!}, \text{ if } L \to 0. \tag{3.34}$$

A formal derivation of these results starting from the solution Eq. (3.30) can be found in Appendix A.3.

3.3.5 Comparison of heuristic and exact reduction for small & large L

Comparing Eqs. (3.34)–(3.33) with Eq. (3.13) and Eq. (3.17), it is immediately clear that the heuristic method of stochastic model reduction gives the correct answer in the limit of fast promoter switching for large L but the incorrect answer for small L. Note that time scale separation exists in both cases of small and large L (as can be verified using Eq. (3.5)) and hence, the lack of agreement of the heuristic and exact reduction is not expected. In Fig. 3.1 we verify that the heuristic and exact reductions agree with each other and with the Finite State Projection (FSP) of the full master equation for large L provided the fast promoter switching limit (large σ_u and σ_b compared to all other parameters) is also met. This is the case for both positive and negative feedback. Note that FSP is a computationally efficient non Monte Carlo method that solves the master equation to any desirable degree of accuracy [69].

To further understand the differences between the two protein distributions in the limit of small L we now look at the mean protein numbers, the Fano Factor (FF) and the Coefficient of Variation (CV) of protein number fluctuations:

$$\langle n \rangle = \rho_b, \qquad \langle n \rangle_a = \rho_b + \frac{(N-1)\rho_b}{1 + N(\exp(\rho_b) - 1)}, \tag{3.35}$$

$$\mathrm{FF} = 1, \qquad \mathrm{FF}_a = 1 + \frac{\rho_b(1-N)}{1 + N(\exp(\rho_b) - 1)}, \tag{3.36}$$

$$\mathrm{CV}^2 = \frac{1}{\rho_b}, \qquad \mathrm{CV}_a^2 = \frac{1 - (1+\rho_b)(1-N^{-1})\exp(-\rho_b)}{\rho_b}. \tag{3.37}$$

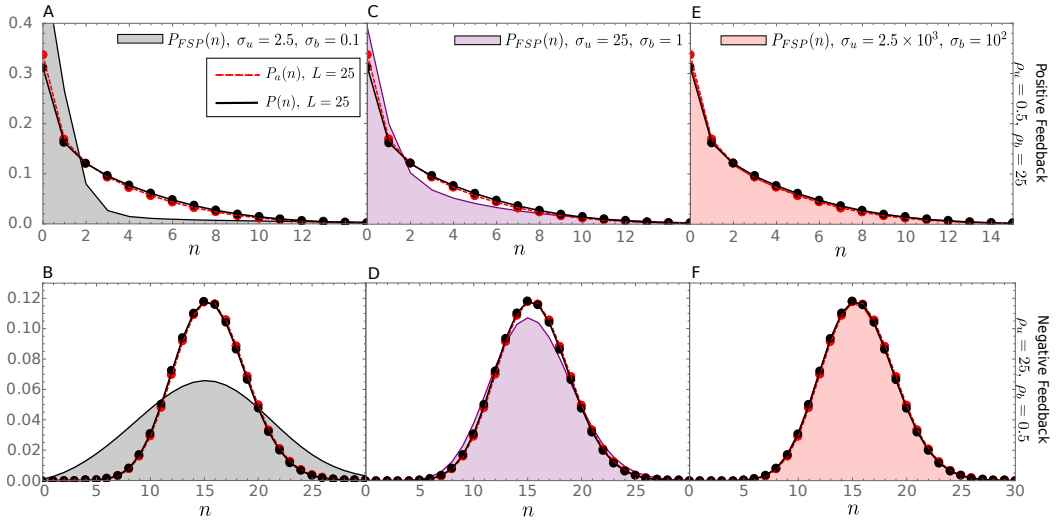


Figure 3.1: Plots comparing the probability distributions of proteins as predicted by the heuristic reduced master equation ($P_a(n)$), the exact reduced master equation ($P(n)$) and the full master equation in the limit of large L ($P_{FSP}(n)$). L is kept constant throughout the figure. Shaded regions indicate the solution of the full master equation Eq. (3.28) using FSP, dashed red lines indicate the heuristic probability distribution from Eq. (3.11), and black solid lines indicate the exact solution in the fast promoter switching limit from Eq. (3.30). Throughout this chapter FSP is used as the benchmark for our analytic results, with a state space truncation chosen such that the probability distributions are indistinguishable from SSA. Going from left-to-right one can observe how Eqs. (3.11) and (3.30) correctly describe the large L limit (here $L = 25$) when both σ_b and σ_u are themselves large, i.e., the fast promoter switching limit. The top row of plots show this for the case of positive feedback ($\rho_u = 0.5$ and $\rho_b = 25$) and the bottom row of plots show this for negative feedback ($\rho_u = 25$ and $\rho_b = 0.5$).

Note that the subscript a denotes calculation using Eq. (3.17) while no subscript implies calculation using Eq. (3.34).

From Eq. (3.35) we deduce that $\langle n \rangle_a < \langle n \rangle$ for $N < 1$ and $\langle n \rangle_a > \langle n \rangle$ for $N > 1$. This means that the solution of the approximate heuristic master equation *underestimates* the mean for positive feedback ($N < 1$) and *overestimates* the mean for negative feedback ($N > 1$). Since the deterministic rate equations also predict a steady state protein mean of ρ_b for the case $L \to 0$ (see Eq. (3.7)) it then follows that the approximate heuristic master equation also leads one to believe in noise-induced shifts of the mean which actually do not exist. From Eq. (3.36) we deduce that the approximate heuristic master equation artificially predicts *sub-Poissonian* ($\mathrm{FF_a} < 1$) fluctuations in molecule numbers for negative feedback ($N > 1$) and *super-Poissonian* ($\mathrm{FF_a} > 1$) fluctuations in molecule numbers for positive feedback ($N < 1$). These deviations from Poissonian behavior are most pronounced for intermediate ρ_b since for small and large ρ_b, $\mathrm{FF}_a \approx 1$. From Eq. (3.37) we deduce that $\mathrm{CV}_a^2 > \mathrm{CV}^2$ for $N < 1$ and $\mathrm{CV}_a^2 < \mathrm{CV}^2$ for $N > 1$, i.e., the approximate heuristic master equation overestimates the size of the protein number fluctuations for positive feedback and underestimates them for negative feedback. These observations are confirmed for positive feedback loops in Fig. 3.2. In particular note

that in Fig. 3.2(A) the heuristic reduced master equation predicts switch-like behavior as ρ_b is increased (from zero for ρ_b below approximately 5 to larger than zero for $\rho_b > 5$) while the exact reduced master equation predicts no such transition for this set of parameters—the lack of reliability in predicting the switching characteristics of positive feedback loops is notable because previous studies [15] have used the heuristic reduced master equation to study switching phenomena.

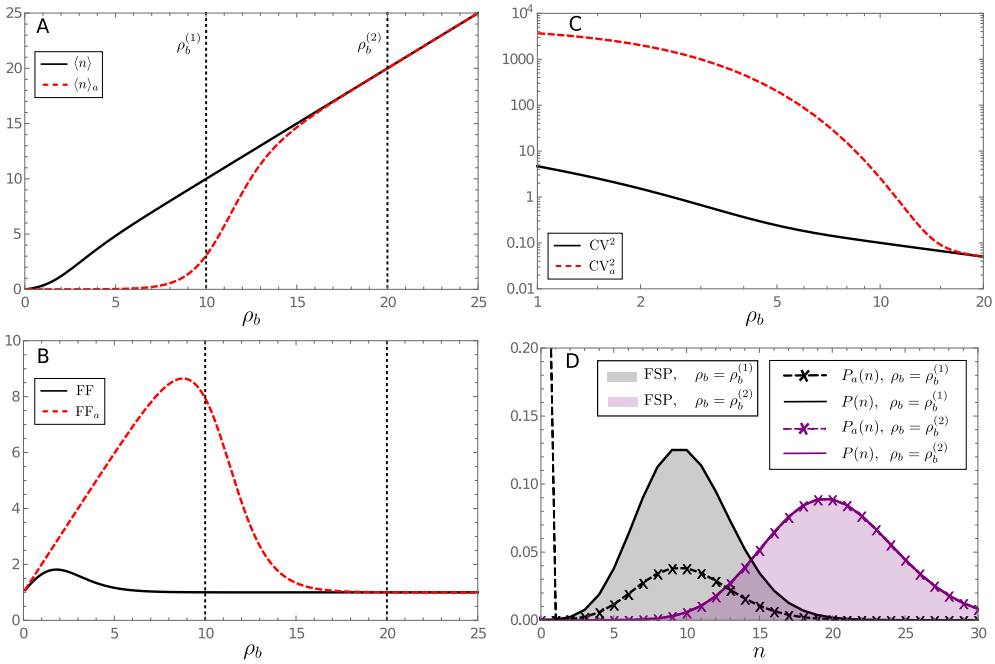


Figure 3.2: Plots showing the breakdown of the heuristic reduced master equation for fast promoter switching in the limit of small L and positive feedback in steady state conditions. The plots show the mean protein number (A), the Fano Factor of protein number fluctuations (B) and the Coefficient of Variation squared (C) as a function of ρ_b. In (D) we show the probability distribution of protein numbers corresponding to two different values of ρ_b. The rest of the parameters are fixed to $\sigma_u = 10^2$, $\sigma_b = 10^5$ and $\rho_u = 0.0002$; this implies $L = 10^{-3}$. Note that $\min(\sigma_u, \sigma_b) \gg \max(1, \rho_u, \rho_b)$ and hence fast promoter switching is ensured. Note that $\langle n \rangle_a$, $\mathrm{FF_a}$ and $\mathrm{CV_a^2}$ in (A)–(C) are calculated using the solution of the heuristic master equation Eq. (3.11) while their non-subscript versions are calculated using the solution of the exact reduced master equation Eq. (3.30). These are in good agreement with the moments calculated in the limit of small L and given by Eq. (3.35)–(3.37). The distributions $P_a(n)$ and $P(n)$ in (D) are calculated using Eq. (3.17) and Eq. (3.34), respectively. The plots verify the large differences between the heuristic and exact reduced master equation for positive feedback loops (see text for discussion), as well as showing agreement between the theoretical distribution for the exact reduced master equation and that obtained using FSP of the full master equation Eq. (A.1).

We can also show that generally positive feedback leads to larger deviations of the heuristic from the exact stochastic model reduction than is the case for negative feedback. Consider strong positive feedback $\rho_u \ll \rho_b$ ($N \ll 1$) with the additional constraint $\rho_u \ll \rho_b \exp(-\rho_b)$. From Eq. (3.35) it can then be shown that $\langle n \rangle = \rho_b, \langle n \rangle_a \approx 0$. If we now reverse the values of ρ_u and ρ_b such that we have strong negative feedback then ρ_b is very small and $N \gg 1$ then $\langle n \rangle = \rho_b, \langle n \rangle_a \approx 1$. Clearly, $|\langle n \rangle - \langle n \rangle_a|$ is much

larger for positive feedback than negative feedback (with ρ_u, ρ_b interchanged) and this difference is evident in the distributions as well. Finally we compute the conditions for the existence of a mode of the probability distribution at $n = 0$ using Eq. (3.34) and Eq. (3.17) respectively:

$$\begin{aligned} \frac{P(1)}{P(0)} < 1 &\Rightarrow \rho_b < 1, \\ \frac{P_a(1)}{P_a(0)} < 1 &\Rightarrow N\rho_b = \rho_u < 1. \end{aligned} \tag{3.38}$$

This implies that if $\rho_u < 1, \rho_b > 1$ (a special case of positive feedback), the approximate heuristic master equation predicts an artificial mode at $n = 0$ whereas if $\rho_u > 1, \rho_b < 1$ (a special case of negative feedback), the approximate heuristic master equation misses to predict an actual mode at $n = 0$. These predictions, contrasting the differences between the heuristic and exact model reduction for positive and negative feedback loops, are illustrated in Fig. 3.3. Note that multiple peaks in the protein distribution are often thought to describe switching between different phenotypes and hence of importance to understanding cellular decision-making [161, 162]—the lack of accuracy in the heuristic model predictions for the bimodality of the protein distribution shows that use of this model can lead to incorrect biological predictions.

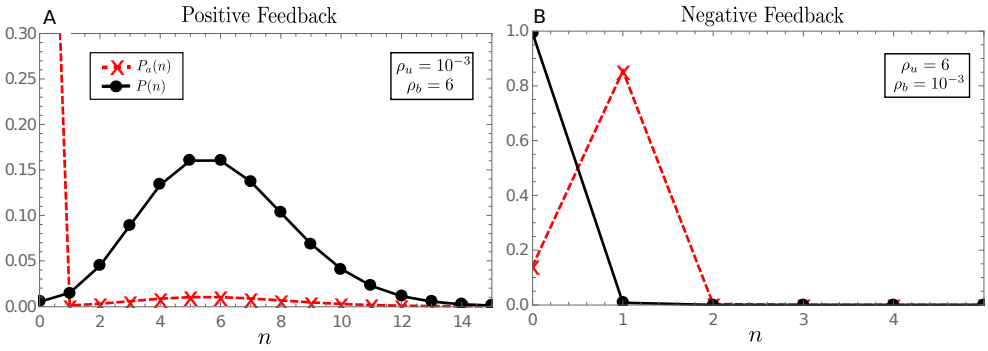


Figure 3.3: Plots comparing the steady state protein distributions predicted by the heuristic and exact reduced master equations for strong positive feedback (A) and strong negative feedback (B) for the case of small L. $P_a(n)$ is calculated using Eq. (3.11) while its subscript version is calculated using the solution of the exact reduced master equation Eq. (3.30). Note that the heuristic distribution ($P_a(n)$) predicts an artificial mode at zero for positive feedback and misses the prediction of a mode at zero for negative feedback, in line with the conditions Eq. (3.38). Note also that as predicted by theory, the differences between $P(n)$ and $P_a(n)$ are most significant for positive feedback loops since for negative feedback loops the differences amount to the order of a single molecule. The parameters $\sigma_u = 10^2$ and $\sigma_b = 10^5$ are fixed across both plots (implying $L = 10^{-3}$) while the values of ρ_u and ρ_b are stated on the figure. FSP distributions are indistinguishable from the theoretical distributions shown in the figure.

The differences between the heuristic and exact model reduction can be explained using the results of Section 3.3.3. There we showed that the heuristic master equation has the same solution, in the limit of fast promoter switching, as the master equation which ignores protein fluctuations due to the reversible protein-DNA binding reaction.

First consider the case of positive feedback (Fig. 3.3A). When proteins are present in the system there will be rapid switching between the G and G^* states. However, in the rare case of an extinction of proteins in the G^* state and where protein binding fluctuations are neglected, a transition from the bound state G^* to unbound state G does not release a protein (reaction scheme shown in Eq. (3.18)). The system then must wait for a protein to be produced via the low ρ_u firing rate if it is to leave state G. And hence, the waiting time for a protein to be produced at the low ρ_u firing rate dominates the steady state dynamics, leading to the mode at zero (Fig. 3.3A red curve). However, where protein binding fluctuations are included (reaction scheme shown in Eq. (3.1)), a transition from the G^* to G does release a protein which can immediately bind to G (due to the high σ_b firing rate, meaning $L \ll 1$) and hence the system does not so readily encounter an extinction of proteins (black curve Fig. 3.3A). We note that even for the black curve there exists a non-zero probability of having zero proteins, which accounts for the long waiting times in the extremely rare event (again note $\sigma_b \gg 1$) that the protein released from the G^* state decays before binding to G; clearly however this is not the dominating feature where protein binding fluctuations are taken into consideration.

Now consider the case of strong negative feedback (Fig. 3.3B). This implies that when a protein is produced in the active G state (now ρ_u is large, ρ_b is very small), the rapid promoter-protein binding reaction will occur forcing the system into the G^* state. Where protein binding fluctuations are neglected no protein is removed upon binding and hence the number of free proteins is still 1; the system will then flip back and forth between the G and G^* states, spending (on average) more time in the G^* state since $\sigma_b \gg \sigma_u$ and hence it is unlikely more than 1 protein will ever be present (due to very small ρ_b and $\sigma_b \gg 1$), hence the mode at $n = 1$ for the red curve in Fig. 3.3B. In the event of a protein extinction the unbound state G will quickly produce another protein in the G state. For strong negative feedback including the binding fluctuation, the rapid promoter-protein binding reaction will instead remove a protein from the system. Again, since it is unlikely the system will contain more than one protein (bound or otherwise), and since the system spends much more time in the G^* state ($\sigma_b \gg \sigma_u$) the probability distribution for the number of free proteins will have a mode at zero (black curve Fig. 3.3B).

The results stated thus far are for steady state conditions. It would also be interesting to understand the difference between the heuristic reduced master equation Eq. (3.9) and the exact master equation Eq. (A.1) for finite time. Since this is analytically intractable we use stochastic simulations to explore this question. Figure 3.4 summarises the results of such simulations for two different parameter sets: (i) $\rho_b = \rho_b^{(s)} = 10$ in which case the heuristic predicts a very different steady state mean number of proteins than the exact reduced master equation; (ii) $\rho_b = \rho_b^{(l)} = 15$ where the heuristic and exact reduced master equations are indistinguishable at steady state (see Fig. 3.4(A)). In Fig. 3.4B we show three independent trajectories of the SSA corresponding to the exact and heuristic reduced master equations for the two parameter choices; the vast difference between the trajectories of the heuristic and the exact for $\rho_b = \rho_b^{(s)}$ are particularly striking. In Fig. 3.4C we show the mean number of proteins as a function of time for the exact and heuristic reduced master equations (solid lines) and compare with the same predicted from the deterministic equations (dashed lines). Two observations can be made: (i) for both parameter sets, the deterministic reaches steady state at a much earlier time than the stochastic models; (ii) for $\rho_b = \rho_b^{(s)}$ the heuristic predicts that the difference in average protein numbers from the deterministic does not decrease with time, while the exact solution predicts that the differences from the deterministic decrease with time (compare top two sub-figures in Fig. 3.4C). In contrast, for $\rho_b = \rho_b^{(l)}$ both master equations predict that differences from the deterministic decrease with time. Taken together, the results indicate that the full time-dependent solution of the heuristic reduced master equation is an accurate reflection of the exact reduced master equation provided ρ_b, the protein production rate in state G^*, is large enough so that we are far away from the switching point of the positive feedback loop.

3.3.6 Numerical computation of the distance measure between steady state distributions

To further understand the regions of parameter space where the heuristic and exact reduced master equations differ, we numerically compute the Hellinger distance (HD) between the exact steady state solution of the heuristic master equation Eq. (3.11) and the exact steady state solution of the master equation Eq. (A.8) for a large region of parameter space for the positive feedback loop: N varying between 10^{-5} and 1, and L varying between 10^{-4} and 1. The results are shown as a heatmap in Fig. 3.5(A). Note that the Hellinger distance is a distance measure between two probability distributions; it is convenient for interpretation since the distance is a fraction, i.e., a HD value of 0 means that two distributions are identical and a HD value approaching 1 means that the distributions are very different from one another. Specifically, a maximum distance 1 is achieved when one of the distributions assigns probability zero to every set to which the other distribution assigns a positive probability. In Fig. 3.5(B) we

calculate the Hellinger distance between the steady state solution of the exact reduced master equation Eq. (3.30) and the exact steady state solution of the master equation Eq. (A.8). Note that there is clear time scale separation across the whole region of parameter space used for the heatmaps as demonstrated in Fig. 3.5(E) and hence based on conventional wisdom, one would expect the heuristic reduced master equation to be accurate at all points in this space. However, Fig. 3.5(A) shows this is not the case—the HD between the distributions predicted by the exact and heuristic reduced master equations varies widely between 0 and 1. In contrast Fig. 3.5(B) shows that the HD between the distributions predicted by the exact and exact reduced master equations is very close to zero across all of parameter space thus verifying that the latter is the correct form of the reduced master equation under fast promoter switching conditions. From Fig. 3.5(A) we see that there is a trend for the HD between the heuristic and exact master equations to decrease with increasing L which agrees with the theoretical prediction in previous sections that the differences are significant for very small L and disappear in the limit of large L. As well, there is a trend for the HD to decrease with increasing N and to be particularly small close to $N = 1$; this agrees with the theoretical prediction that for $N = 1$ the heuristic and exact precisely agree because in this case there is no effective feedback mechanism.

In Fig. 3.5(F) we plot the protein distributions predicted by the heuristic and exact master equations for the star points labeled (1) and (2) in Fig. 3.5(A) which are positioned in regions of high and low HD, respectively. Note that for point (1) the heuristic predicts that the probability that the protein numbers are zero is high whereas the exact predicts the probability that the protein numbers are zero is very small. Inspired by this observation, as well as the theoretical prediction of modes at zero for small L given by Eq. (3.38), in Fig. 3.5(C) we plot a heatmap of the absolute difference between the height of the zero modes of the exact master equation and the heuristic reduced master equation and find that this heatmap is in very good agreement with the heatmap for the HD shown in Fig. 3.5(A). *This verifies our intuition that the differences between the protein distributions of the two master equations is mostly due to differences in their prediction of the probability of zero proteins at steady-state.* Inspired by the result in Appendix A.1 that the deterministic time evolution equations can be obtained from the full master equation under the assumption $\langle n|G\rangle \approx \langle n\rangle$ (which is equivalent to independence of protein and gene fluctuations), we plot in Fig. 3.5(D) a heatmap of $|\langle n|G\rangle - \langle n\rangle|$ which also shows broad similarity to that in Fig. 3.5(A).

In Fig. 3.6 we show the results of the same analysis as in Fig. 3.5 but now for the case of negative feedback loops. As before, the largest HD between the heuristic reduced master equation and the exact master equation is found for N far away from the trivial case of $N = 1$ and for small L, in line with the theoretical predictions of the previous section.

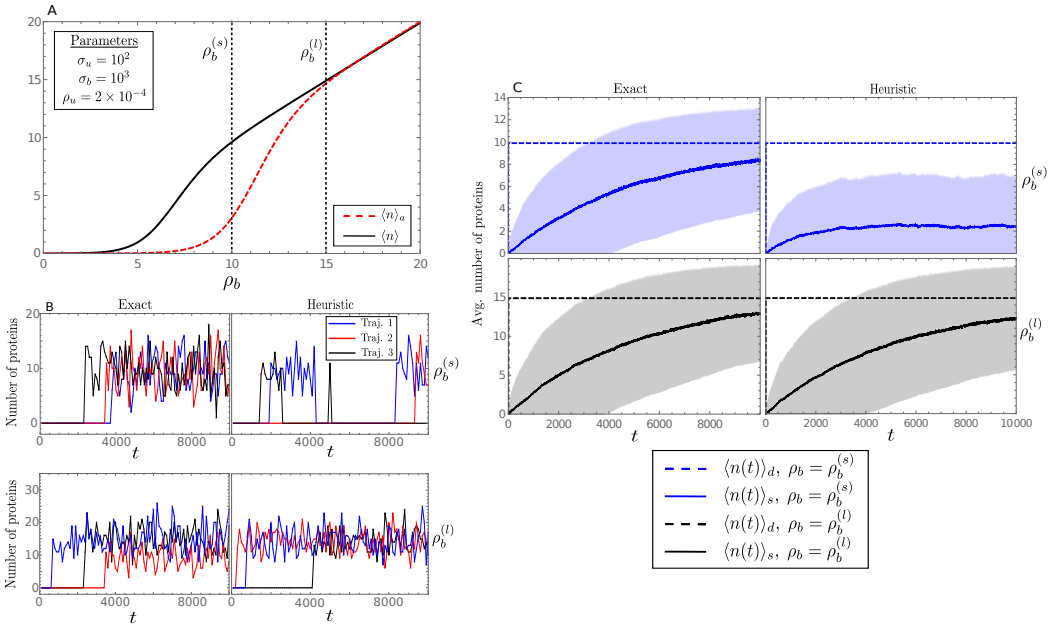


Figure 3.4: Plots comparing the time evolution of the heuristic reduced master equation Eq. (3.9) and the exact master equation Eq. (A.1) for fast promoter switching conditions, positive feedback and small L (0.1). In (A) we show the switching characteristics of the positive feedback loop as a function of ρ_b for the two master equations in steady state conditions. Here dotted lines define the values of $\rho_b^{(s)}$ (subscript s for small) and $\rho_b^{(l)}$ (l for large) used throughout the rest of the figure. In (B) we plot three independent trajectories from the SSA corresponding to the two master equations. Each trajectory shown is down-sampled 1:100 for visual clarity. The top row corresponds to $\rho_b = \rho_b^{(s)}$ and the bottom row corresponds to $\rho_b = \rho_b^{(l)}$. In (C) we plot the mean number of proteins as a function of time as predicted by the exact and heuristic reduced master equations (shown as solid lines and denoted by the subscript s in the legend) and by the deterministic rate equations (shown as dashed lines and denoted by the subscript d in the legend). The shaded regions show one standard deviation about the mean. The moments were calculated over 2×10^3 SSA trajectories. All sub-figures compare two different parameter sets, one for small ρ_b (where at steady state the heuristic and exact differ considerably) and one for large ρ_b (where at steady state the differences are negligible), as indicated in Fig. 3.4(A). Note that $\min(\sigma_u, \sigma_b) \gg \max(1, \rho_u, \rho_b)$ and hence fast promoter switching is ensured for both parameter sets. See text for discussion.

Both the heatmaps of the absolute difference between the heights of the zero modes (Fig. 3.6(C)) and of the absolute difference between the protein mean and the conditional protein mean (Fig. 3.6(D)) show high correlation with the HD heatmap in Fig. 3.6(A). The differences between the heuristic and exact protein distributions for the large HD and small HD in Fig. 3.6(A) (star point 1, 2 respectively) are shown in Fig. 3.6(F). Note that the differences between the two cases amount to the order of a single molecule and are far smaller than the differences found for positive feedback (compare Fig. 3.5(F)), verifying the theoretical predictions of the previous section, namely that the heuristic fails worst for positive feedback.

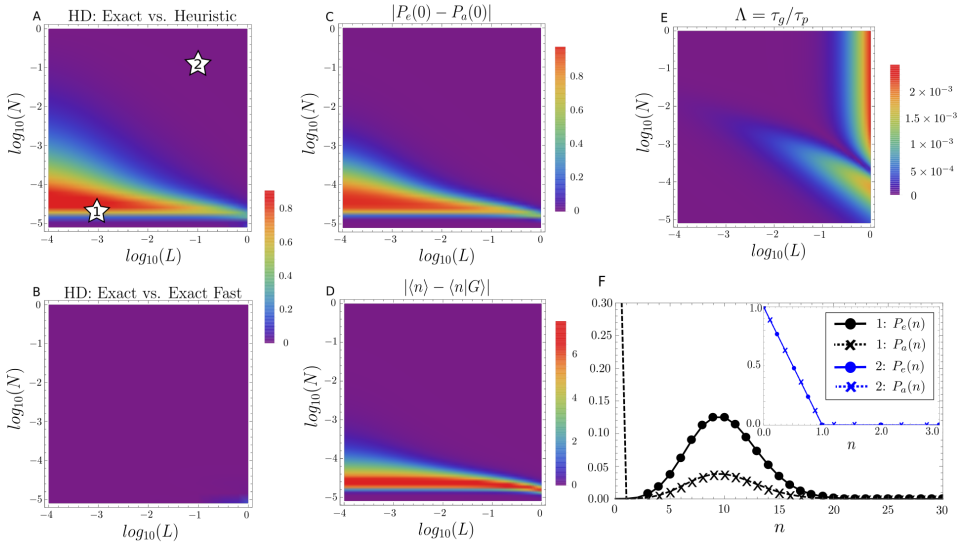


Figure 3.5: Quantifying the differences between the steady state protein distributions predicted by the exact master equation Eq. (A.8), the heuristic reduced master equation Eq. (3.11) and the exact reduced master equation Eq. (3.30) for positive feedback loops. $P_e(n)$ denotes the exact steady state solution of the exact master equation (i.e., the exact solution of the reaction scheme from Eq. (3.1), see derivation in Appendix A.2) and $P_a(n)$ denotes the distribution from the heuristic reduced master equation. We note that $P_e(n)$ takes the role of P_{FSP} used in previous figures. In (A) we show the Hellinger distance (HD) between the predictions of the exact master equation and heuristic reduced master equation. In (B) we show the HD between the predictions of the exact master equation and exact reduced master equation (denoted as exact fast). (C) Shows the absolute difference between the probability of zero protein molecules predicted by the exact master equation and the probability of zero protein molecules predicted by the heuristic reduced master equation. (D) Shows the absolute difference between $\langle n\rangle$ and $\langle n|G\rangle$ computed using the exact master equation. (E) Shows that the whole region of parameter space chosen has suitable deterministic time scale separation where $\tau_g = 1/\lambda_1, \tau_p = 1/\lambda_2$ and λ_i are the eigenvalues of the Jacobian of the deterministic rate equations Eqs. (3.2-3.3) evaluated at steady-state. Note that it is expected from Fig. 3.4 that the stochastic time scale separation will be much greater than the deterministic. Parameters $\sigma_u = 100$, $\rho_u = 2 \times 10^{-4}$ are fixed throughout the figure, with σ_b varying in the range $100 - 10^6$ (small L) and ρ_b varying in the range $2 \times 10^{-4} - 25$ (positive feedback). Numbered stars in (A) indicate the two points in parameter space whose corresponding probability distributions of protein numbers we show in (F). See text for discussion.

3.3.7 Extending results to the case of multiple protein binding

Here we briefly treat the more general case where multiple protein molecules can bind the promoter, a common case in nature often associated with cooperative behaviour. The reaction scheme is an extension of (3.1) and reads:

$$\begin{aligned} &G \xrightarrow{\rho_u} G + P, \;\; P \xrightarrow{1} \varnothing, \\ &P + G \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*, \;\; P + G^* \underset{\delta_u}{\overset{\delta_b}{\rightleftharpoons}} G^{**}, \\ &G^{**} \xrightarrow{\rho_b} G^{**} + P. \end{aligned} \tag{3.39}$$

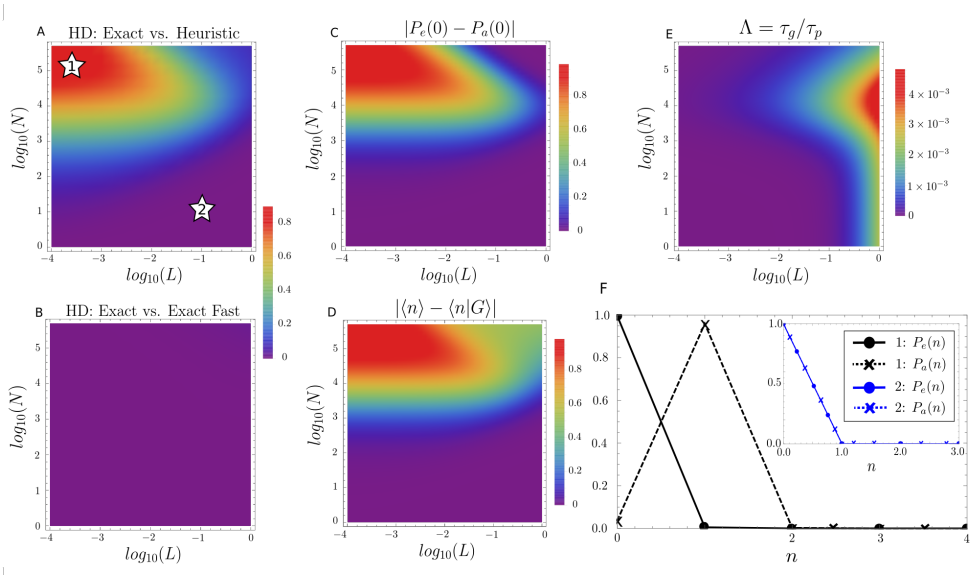


Figure 3.6: Quantifying the differences between the steady state protein distributions predicted by the exact master equation Eq. (A.8), the heuristic reduced master equation Eq. (3.11) and the exact reduced master equation Eq. (3.30) for negative feedback loops. In (A) we show the Hellinger distance (HD) between the predictions of the exact master equation and heuristic reduced master equation. In (B) we show the HD between the predictions of the exact master equation and exact reduced master equation. (C) Shows the absolute difference between the probability of zero protein molecules predicted by the exact master equation and the probability of zero protein molecules predicted by the heuristic reduced master equation. (D) Shows the absolute difference $\langle n\rangle$ and $\langle n|G\rangle$ computed using the exact master equation. (E) Shows that the whole region of parameter space chosen has suitable deterministic time scale separation where $\tau_g = 1/\lambda_1, \tau_p = 1/\lambda_2$ and λ_i are the eigenvalues of the Jacobian of the deterministic rate equations Eqs. (3.2)–(3.3) evaluated at steady-state. Note that it is expected from Fig. 3.4 that the stochastic time scale separation will be much greater than the deterministic. Parameters $\sigma_u = 100$ and $\rho_b = 2 \times 10^{-4}$ are fixed throughout the figure, with σ_b varying between 100 and 10^6 (small L) and ρ_u varying between 2×10^{-4} and 25 (negative feedback). Numbered stars in (A) indicate the two points in parameter space whose corresponding probability distributions of protein numbers we show in (F). See text for discussion.

Writing the deterministic rate equations for this system and making the assumption of fast promoter switching such that $\partial_t\langle g\rangle_d \approx 0$, $\partial_t\langle g^*\rangle_d \approx 0$ and $\partial_t\langle g^{**}\rangle_d \approx 0$ it is straightforward to show that the effective deterministic time evolution equation for the protein numbers has the form:

$$\frac{d\langle n\rangle_d}{dt} \approx \frac{LR\rho_u + \rho_b\langle n\rangle_d^2}{(L + \langle n\rangle_d)R + \langle n\rangle_d^2} - \langle n\rangle_d, \tag{3.40}$$

where $L = \sigma_u/\sigma_b$ and $R = \delta_u/\delta_b$. It follows by the same reasoning as in Section 3.3.2 that the corresponding heuristic reduced master equation for protein dynamics is given by Eq. (3.9) with the effective propensities:

$$\begin{aligned} T^+(n) &= \frac{LR\rho_u + \rho_b n^2}{(L+n)R + n^2}, \\ T^-(n) &= n. \end{aligned} \tag{3.41}$$

In the limit of small L and R, Eq. (3.41) reduces to the effective propensities given by Eq. (3.14) while in the limit of large L and R, Eq. (3.41) reduces to the effective propensities given by Eq. (3.12). Hence the solutions of the heurestic master equation in these two limits are given by Eq. (3.17) and Eq. (3.13).

By inspection of the reaction scheme (3.39) it is obvious that for small L and R, the gene will be mostly in state G^{**} and hence the principal reactions determining the protein dynamics are $G^{**} \xrightarrow{\rho_b} G^{**} + P, P \xrightarrow{1} \varnothing$. By the same reasoning, it follows that for large L and R, the gene will be mostly in state G and the principal reactions are $G \xrightarrow{\rho_u} G + P, P \xrightarrow{1} \varnothing$. Hence the solution of the exact master equation of (3.39) in the limit of small and large L, R is Poisson and given by Eq. (3.34) and Eq. (3.33) respectively. *Hence all the conclusions previously reached regarding the differences between the heuristic reduced master equation and the exact master equation for single protein binding for the case of small and large L also hold for multiple protein binding for the cases of small and large L, R.* Note that the derivations here assume the exchangeability of the limits of large/small L, R and large time; hence the proof here presented is not formal but the results are the expected ones and are confirmed by simulations (see later).

It is interesting to find the general conditions under which the heuristic reduced master equations generally agrees with the exact. For the case of single promoter binding, we showed in Appendix A.1 that the deterministic rate equations agreed with the mean of the exact master equation when $\langle n \rangle = \langle n|G \rangle$. Next we derive a similar condition for the case of multiple protein binding. We start by noting that under fast promoter switching conditions, the reactions $P + G \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*, P + G^* \underset{\delta_u}{\overset{\delta_b}{\rightleftharpoons}} G^{**}$ are in equilibrium and hence the deterministic rate equations yield:

$$\begin{aligned} \langle g \rangle_d &= \frac{LR}{\langle n \rangle_d^2 + (L + \langle n \rangle_d)R}, \\ \langle g^{**} \rangle_d &= \frac{\langle n \rangle_d^2}{\langle n \rangle_d^2 + (L + \langle n \rangle_d)R}. \end{aligned} \tag{3.42}$$

Next we derive equations for the same quantities but from the exact master equation (denoted $\langle g\rangle$, $\langle g^*\rangle$ and $\langle g^{**}\rangle$). Writing the master equation for the same two reversible reactions in equilibrium, one can deduce the moment equations:

$$\partial_t\langle g\rangle = 0 = -\sigma_b\langle g\rangle + \sigma_u(1-\langle g\rangle - \langle g^{**}\rangle), \tag{3.43}$$

$$\partial_t\langle g^{**}\rangle = 0 = \delta_b\langle n(1-g-g^{**})\rangle - \delta_u\langle g^{**}\rangle, \tag{3.44}$$

from which we can deduce:

$$\begin{aligned}\langle g\rangle &= \frac{L(R+\langle n|G^{**}\rangle - \langle n\rangle)}{\langle n|G\rangle(R+\langle n|G^{**}\rangle) + L(R+\langle n|G^{**}\rangle - \langle n|G\rangle)},\\ \langle g^{**}\rangle &= \frac{L(\langle n\rangle - \langle n|G\rangle) + \langle n\rangle\langle n|G\rangle}{\langle n|G\rangle(R+\langle n|G^{**}\rangle) + L(R+\langle n|G^{**}\rangle - \langle n|G\rangle)},\end{aligned} \tag{3.45}$$

where we used the definitions of conditional means: $\langle n|G\rangle = \langle ng\rangle/\langle g\rangle$ and $\langle n|G^{**}\rangle = \langle ng^{**}\rangle/\langle g^{**}\rangle$. Comparing Eq. (3.42) and Eq. (3.45), we see that they can only be equal if the following condition is true:

$$\langle n|G\rangle = \langle n|G^{**}\rangle = \langle n\rangle. \tag{3.46}$$

By means of the definition of the mean in terms of conditional means $\langle n\rangle = \langle n|G\rangle\langle g\rangle + \langle n|G^*\rangle\langle g^*\rangle + \langle n|G^{**}\rangle\langle g^{**}\rangle$, one can deduce the final condition required for the agreement of the time evolution equations for the mean protein number according to the deterministic rate equations and the exact master equation:

$$\langle n|G\rangle = \langle n|G^*\rangle = \langle n|G^{**}\rangle = \langle n\rangle. \tag{3.47}$$

Since the heuristic is based on the deterministic, we expect that this condition is also an indicator of when the heuristic and exact master equations agree. In Fig. 3.7 we verify this intuition using simulations: when the above condition is approximately met then the heuristic and exact master equations predict very similar distributions of protein numbers (see Fig. 3.7(A),(C)) whereas the largest differences between the two master equations (see Fig. 3.7(B),(D)) correlate with significant differences between the three conditional mean protein numbers $\langle n|G\rangle, \langle n|G^*\rangle, \langle n|G^{**}\rangle$.

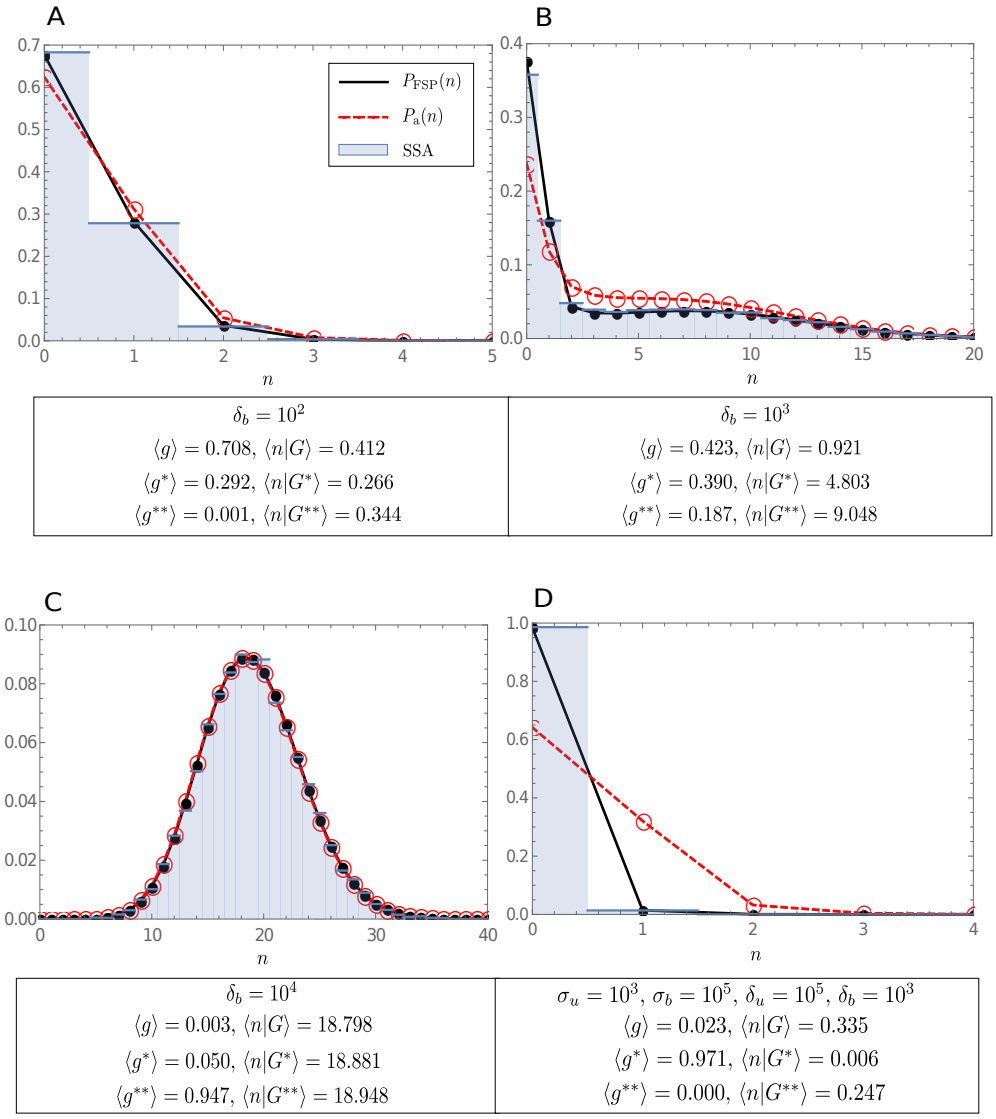


Figure 3.7: Plots comparing the heuristic master equation (Eq. (3.9) with Eq. (3.41)) and exact master equation predictions for the protein number distributions of a positive feedback loop ($\rho_u = 0.5, \rho_b = 20$) with multiple protein binding (3.39). All calculations done using FSP. Fast promoter switching is enforced by choosing $\min(\sigma_u, \sigma_b, \delta_u, \delta_b) \gg \max(1, \rho_u, \rho_b)$. In (A)–(C) parameter values are fixed to $\sigma_u = 10^4, \sigma_b = 10^4, \delta_u = 10^4$ and δ_b is varied in the range $10^2 - 10^4$. Plot (A) shows a case where the system spends most of its time in state G, in (B) we show a case for which all states G, G^* and G^{**} are frequently accessed by the system and in (C) we show a case where the state G^{**} dominates. Sub-figure (D) shows a case where the systems spends most of its time in state G^*. Note that the differences between the heurestic and exact master equation are reflected in the differences between the values of the mean number of proteins conditional on each state, with the smallest differences occurring for cases (A) and (C), in agreement with the condition given by Eq. (3.47)

3.4 Model reduction for bursty feedback loops

In this section we consider model reduction for feedback loops in which there is an implicit mRNA description. It has been rigorously shown that when mRNA degrades much quicker than proteins (a common situation for bacteria and yeast cells) then the mRNA does not need to be explicitly described but rather implicitly manifests through protein bursts [38]. Studies have elucidated the implications of taking into account protein bursting on downstream pathways and shown its importance [163]. Hence we now consider a feedback loop with an implicit mRNA description which has the effective reaction scheme:

$$G \xrightarrow{\rho_u} G + wP,\ P \xrightarrow{1} \varnothing,\ P + G \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*,\ G^* \xrightarrow{\rho_b} G^* + wP, \tag{3.48}$$

where w is the protein burst size which is a random positive integer drawn from the geometric distribution with mean $b = k/d_M$, k is the rate at which mRNA is translated into proteins and d_M is the mRNA degradation rate. Note that the geometric form of the protein burst distribution has been shown theoretically [164] and verified experimentally [30].

3.4.1 Heuristic stochastic model reduction

We are interested in a reduced description of protein fluctuations in the limit of fast promoter switching. Clearly, the effective master equation has to have the general form:

$$\begin{aligned}\frac{dP_a(n,t)}{dt} =& \sum_{i=0}^{\infty}(T_i^+(n-i)P_a(n-i,t) - T_i^+(n)P_a(n,t)) \\ &+ T^-(n+1)P_a(n+1,t) - T^-(n)P_a(n,t),\end{aligned} \tag{3.49}$$

where $T_i^+(n)dt$ is the probability, given n proteins, that a protein burst of size i will occur in the time interval $[t, t+dt)$ and $T^-(n)dt$ is the probability, given n proteins, that a protein degradation event reducing the number of proteins by one will occur in the time interval $[t, t+dt)$. Next we use the deterministic rate equations to guess the equations for $T_i^+(n)$ and $T^-(n)$. The deterministic rate equations corresponding to (3.48) are given by:

$$\frac{d\langle g(t)\rangle_d}{dt} = -\sigma_b\langle g(t)\rangle_d\langle n(t)\rangle_d + \sigma_u(1 - \langle g(t)\rangle_d), \tag{3.50}$$

$$\begin{aligned}\frac{d\langle n(t)\rangle_d}{dt} =& -\sigma_b\langle g(t)\rangle_d\langle n(t)\rangle_d + \sigma_u(1 - \langle g(t)\rangle_d) + \rho_u b\langle g(t)\rangle_d \\ &+ \rho_b b(1 - \langle g(t)\rangle_d) - \langle n(t)\rangle_d.\end{aligned} \tag{3.51}$$

Note that these equations are the same as the deterministic rate equations for the non-bursty case given by Eqs. (3.2-3.3) except that ρ_u is replaced by $\rho_u b$ and ρ_b is replaced by $\rho_b b$; this directly follows from the definition of b as the mean protein burst size. Assuming fast promoter switching, $\partial_t \langle g(t) \rangle_d \approx 0$, it follows that an effective reduced rate equation for the mean protein numbers is:

$$\frac{d\langle n \rangle_d}{dt} \approx \frac{L\rho_u + \rho_b \langle n \rangle_d}{L + \langle n \rangle_d} b - \langle n \rangle_d. \tag{3.52}$$

The form of this effective equation combined with the fact that we know that burst size is distributed according to a geometric distribution with mean b suggests a one-variable master equation of the form Eq. (3.49) with effective propensities:

$$\begin{aligned} T_i^+(n) &= \frac{L\rho_u + \rho_b n}{L + n} \psi_i, \\ T^-(n) &= n, \end{aligned} \tag{3.53}$$

where ψ_i is the probability that a burst has size i which is given by $b^i/(1+b)^{i+1}$. If we denote the angled brackets with subscript a as the statistical averages calculated using the heuristic master equation Eq. (3.49) with propensities given by Eq. (3.53) then it follows that:

$$\begin{aligned} \frac{d\langle n \rangle_a}{dt} &= \sum_{i=0}^{\infty} i \langle T_i^+(n) \rangle_a - \langle T^-(n) \rangle_a, \\ &= b \left\langle \frac{L\rho_u + \rho_b n}{L + n} \right\rangle_a - \langle n \rangle_a, \\ &\approx \frac{L\rho_u + \rho_b \langle n \rangle_a}{L + \langle n \rangle_a} b - \langle n \rangle_a. \end{aligned} \tag{3.54}$$

Note that this equation is the same as the reduced rate equation Eq. (3.52) (upon replacing $\langle n \rangle_a$ by $\langle n \rangle$) and hence verifies that the form of the effective propensities given by Eq. (3.53) guarantees equivalence between the effective equation for the time evolution of the mean protein numbers of the heuristic master equation and the reduced deterministic rate equation in the limit of small protein number fluctuations when $\langle n \rangle_a \approx n$.

The heuristic stochastic model given by Eqs. (3.49,3.53) is difficult to solve exactly in steady state because there are no known general solutions for one species reaction systems with multi-step reactions, i.e., reactions leading to the production of more than one molecule at a time [74]. However, as we now show, provided we can assume exchangeability of the limits of large/small L and large time then closed-form solutions can be obtained for the steady state distributions of protein numbers.

The limit of large L

In this limit, Eq. (3.53) reduces to the simpler form:

$$\begin{aligned} T_i^+(n) &\approx \rho_u \psi_i, \\ T^-(n) &= n. \end{aligned} \tag{3.55}$$

Substituting these in the heuristic reduced master equation Eq. (3.49), multiplying both sides by z^n and taking the sum over n on both sides of this equation we get the generating function equation:

$$\frac{\partial G(z,t)}{\partial t} \approx \rho_u G(z,t)\left(\frac{1}{1+b(1-z)} - 1\right) + (1-z)\frac{\partial G(z,t)}{\partial z}, \tag{3.56}$$

where $G(z,t) = \sum_n z^n P_a(n,t)$. This equation can be solved in steady state yielding $G(z) = (1-b(z-1))^{-\rho_u}$ which implies that:

$$\begin{aligned} P_a(n) = \frac{1}{n!}\frac{d^n G(z)}{dz^n}\bigg|_{z=0} &\approx \left(\frac{b}{1+b}\right)^n \left(1 - \frac{b}{1+b}\right)^{\rho_u} \frac{\Gamma(\rho_u + n)}{\Gamma(n+1)\Gamma(\rho_u)} \\ &= \mathrm{NB}\left(\rho_u, \frac{b}{(1+b)}\right), \end{aligned} \tag{3.57}$$

where $\mathrm{NB}(x,y)$ stands for a negative binomial distribution with parameters x, y and mean $xy/(1-y)$.

The limit of small L

In this limit, Eq. (3.53) reduces to the simpler form:

$$\begin{aligned} T_i^+(n) &\approx ((\rho_u - \rho_b)\delta(0,n) + \rho_b)\psi_i, \\ T^-(n) &= n, \end{aligned} \tag{3.58}$$

where $\delta(0,n)$ is the Kronecker delta. Substituting these in the heuristic reduced master equation Eq. (3.49), multiplying both sides by z^n and taking the sum over n on both sides of this equation we get the corresponding generating function equation:

$$\frac{\partial G(z,t)}{\partial t} \approx ((\rho_u - \rho_b)G(0,t) + \rho_b G(z,t))\left(\frac{1}{1+b(1-z)} - 1\right) + (1-z)\frac{\partial G(z,t)}{\partial z}. \tag{3.59}$$

In steady-state, this equation has the solution:

$$G(z) = \frac{1 + N(-1 + (1+b)^{\rho_b}(1-b(z-1))^{-\rho_b})}{1 + N(-1 + (1+b)^{\rho_b})}. \tag{3.60}$$

Hence the steady state probability distribution is given by:

$$P_a(n) \approx \frac{1}{n!}\frac{d^n G(z)}{dz^n}\Big|_{z=0} = \begin{cases} \frac{1}{1+N((1+b)^{\rho_b}-1)}, & \text{if } n=0, \\ \mathrm{NB}\left(\rho_b, \frac{b}{1+b}\right)\frac{N}{N-(N-1)(1+b)^{-\rho_b}}, & \text{if } n \geq 1. \end{cases} \tag{3.61}$$

3.4.2 Exact stochastic model reduction

To determine how accurate is the heuristic model reduction we need to compare it with the reduction done on the exact model in the limit of fast promoter switching. Unlike the case of a non-bursty feedback loop, the exact solution of the chemical master equation for reaction scheme (3.48) is unknown. However, by taking the same approach as we did in Section 3.3, it is easy to find the solution of the chemical master equation for the case of fast promoter switching and L being either very small or very large.

The limit of fast promoter switching implies that the reaction $P + G \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*$ in the reaction scheme (3.48) is approximately in equilibrium for all times. From the chemical master equation for this reversible reaction one finds that the fraction of time that the gene is ON is given by Eq. (A.6). Hence it follows that in the limit of small L, $\langle g \rangle$ is also very small, the gene spends most of its time in state G^* and consequently the only effective reactions determining the protein dynamics are:

$$G^* \xrightarrow{\rho_b} G^* + wP, P \xrightarrow{1} \varnothing, \tag{3.62}$$

where w is the protein burst size which is a random positive integer drawn from the geometric distribution with mean b. The chemical master equation for these two reactions can be easily solved in steady state leading to a negative binomial distribution, $P(n) \approx \mathrm{NB}(\rho_b, b/(1+b))$. In the opposite limit of large L, $\langle g \rangle$ is approximately 1, the gene spends most of its time in state G and consequently the only effective reactions determining the protein dynamics are:

$$G \xrightarrow{\rho_u} G + wP, P \xrightarrow{1} \varnothing. \tag{3.63}$$

Solving the chemical master equation in steady state (using the generating function method) for these two reactions leads to another negative binomial solution, $P(n) \approx \mathrm{NB}(\rho_u, b/(1+b))$.

Hence summarising the results of Sections 3.4.1 and 3.4.2, we can state that the heuristic and exact stochastic model reduction in the limit of fast promoter switching agree for large L (both predict a negative binomial distribution, $P(n) = P_a(n) = \mathrm{NB}(\rho_u, b/(1+b))$) but disagree for small L: the exact reduction predicts a negative binomial distribution, $P(n) = \mathrm{NB}(\rho_b, b/(1+b))$ while the heuristic reduction predicts the different distribution given by Eq. (3.61). These results qualitatively parallel those previously obtained for a non-bursty feedback loop.

To further understand the differences between these two distributions for small L we now look at the mean protein numbers, the Fano Factor and the Coefficient of Variation of protein number fluctuations:

$$\langle n \rangle = b\rho_b, \quad \langle n \rangle_a = b\rho_b + \frac{b(N-1)\rho_b}{1 + ((1+b)^{\rho_b} - 1)N}, \tag{3.64}$$

$$\mathrm{FF} = 1 + b, \quad \mathrm{FF}_a = 1 + b + \frac{b(1-N)\rho_b}{1 + ((1+b)^{\rho_b} - 1)N}, \tag{3.65}$$

$$\mathrm{CV}^2 = \frac{1+b}{b\rho_b}, \quad \mathrm{CV}_a^2 = \frac{1+b}{b\rho_b} - \frac{(1+b)^{-\rho_b}(1-N^{-1})(1+b+b\rho_b)}{b\rho_b}. \tag{3.66}$$

Hence for $N < 1$ (positive feedback), the heuristic underestimates the mean protein number and over-estimates the Fano Factor and the Coefficient of Variation of protein number fluctuations and the opposite occurs when $N > 1$ (negative feedback). The deterministic rate equations predict a mean of $b\rho_b$ (can be deduced from Eq. (3.52) in limit of small L) which agrees with $\langle n \rangle$ but not with $\langle n \rangle_a$ and hence the heuristic artificially predicts noise-induced deviations from the deterministic mean. These are the same conclusions that we reached in Section 3.3.4 for the case of a non-bursty feedback loop. The exact reduction predicts super-Poissonian fluctuations (FF > 1) while the heuristic predicts the same for $N < 1$ and either super- or sub-Poissonian fluctuations for $N > 1$ ($\mathrm{FF}_a > 1$ and $\mathrm{FF}_a < 1$ respectively). We note that the prediction of sub-Poissonian fluctuations is a surprising illogical output of the heuristic model since naturally the production of proteins in bursts has to lead to number distributions which are wider than Poisson. The theoretical predictions for the mean, FF, CV are corroborated using FSP in Fig. 3.8(A)–(C).

The relative errors (made by the heuristic reduction method) for the mean, FF and CV^2 can be computed using $e_m = |\langle n \rangle - \langle n \rangle_a|/\langle n \rangle$, $e_{\mathrm{FF}} = |\mathrm{FF} - \mathrm{FF}_a|/\mathrm{FF}$ and $e_{\mathrm{CVs}} = |\mathrm{CV}^2 - \mathrm{CV}_a^2|/\mathrm{CV}^2$, respectively. The errors for the bursty feedback loop (computed using Eqs. (3.64-3.66)) are smaller than the errors for the non-bursty feedback loop (computed using Eqs. (3.35-3.37)) provided the mean burst size $b \gg 1$. *Hence a major prediction of our theory is that bursts in protein expression generally reduce the size of the discrepancies between heuristic and exact stochastic model reduction in the limit of small L.* This theoretical prediction is verified using FSP in Fig. 3.8(D) .

Finally, we compute the conditions for the existence of a mode of the probability distribution at $n = 0$ using the distribution obtained from the exact method $P(n) = \mathrm{NB}(\rho_b, b/(1+b))$ and the distribution from the heuristic reduction Eq. (3.61) respectively:

$$\frac{P(1)}{P(0)} < 1 \Rightarrow \rho_b < \frac{1+b}{b}, \tag{3.67}$$

$$\frac{P_a(1)}{P_a(0)} < 1 \Rightarrow N\rho_b = \rho_u < \frac{1+b}{b}. \tag{3.68}$$

This implies that if $\rho_u < (1+b)/b,\ \rho_b > (1+b)/b$ (a special case of positive feedback), the approximate heuristic master equation predicts an artificial mode at $n = 0$ whereas if $\rho_u > (1+b)/b,\ \rho_b < (1+b)/b$ (a special case of negative feedback), the approximate heuristic master equation misses to predict an actual mode at $n = 0$.

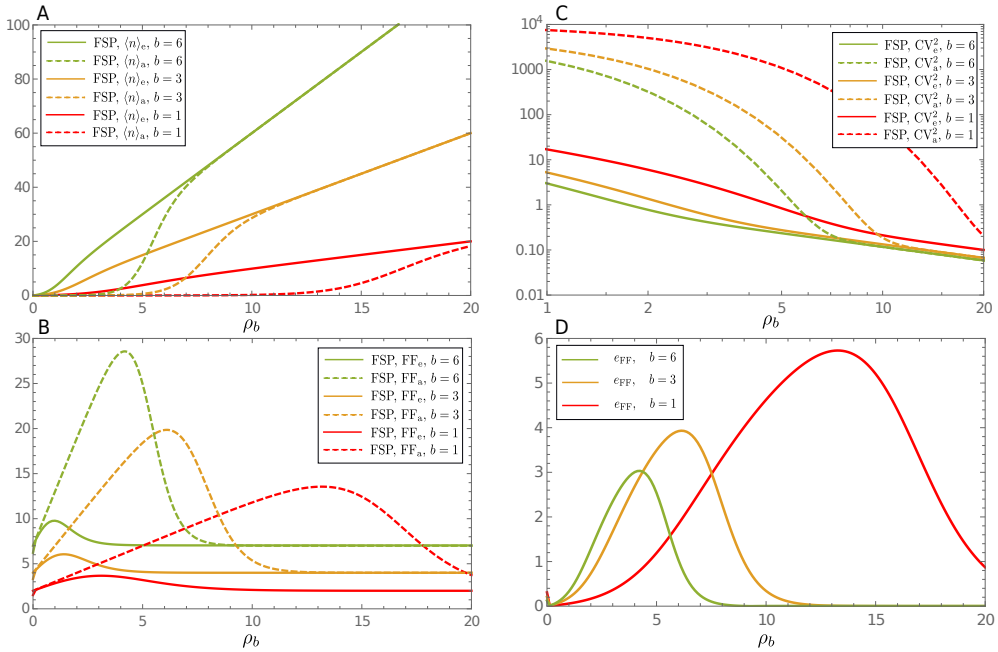


Figure 3.8: Plots showing the breakdown of the heuristic reduced master equation for fast promoter switching in the limit of small L for a bursty positive feedback loop. The plots show the mean protein number number (A), the Fano Factor of protein number fluctuations (B) and the Coefficient of Variation squared (C) as a function of ρ_b and the mean burst size b. In these plots we compare the FSP solution of the full master equation corresponding to reaction scheme (3.48) with the FSP solution of the heuristic reduced master equation Eq. (3.49) with Eq. (3.53); the latter is distinguished from the former by the subscript e. These are in good agreement with the moments calculated in the limit of small L and given by Eq. (3.64)–(3.66)—for example the means of the exact solution in (A) are very well approximated by $\langle n\rangle = b\rho_b$. In (D) we show the relative error in the heuristic reduced master equation's FF computed using the data in (B) where $e_{\mathrm{FF}} = |\mathrm{FF}_e - \mathrm{FF}_a|/\mathrm{FF}_e$. Note that the relative error decreases with increasing mean burst size b. In all cases $\rho_u = 0.0002$, $\sigma_u = 100$ and $\sigma_b = 10^5$ were chosen such that L is small, there is positive feedback and fast promoter switching is guaranteed.

3.5 Conclusion

In this chapter we have conclusively shown that heuristic stochastic models with Hill-type propensities for transcriptional regulation are not generally valid under fast promoter switching conditions, as commonly assumed. Rather we show that they are valid only over a subset of parameter space consistent with the fast promoter switching assumption, namely when the rate of protein-DNA binding reaction is much less than the unbinding reaction. Our work shows that when this condition is not met, the protein distributions predicted by the heuristic models can be considerably different than the true protein distributions. These differences exist for both negative and positive feedback loops but are particularly pronounced for the latter—in this case we have shown that the heuristic model can predict an artificial mode at zero proteins, an incorrect switching point from low to high protein expression as a parameter is varied, artificial deviations of the mean number of proteins from that predicted by the rate equations and a huge overestimation of the size of protein number fluctuations and of the Fano Factor. Surprisingly, we found that the heuristic solution exactly corresponds to the fast gene switching limit of the autoregulatory system that ignores protein number fluctuations due to the protein-promoter binding reaction. Our work further builds on previous work by other authors [157, 165, 156] but has the advantage of using theory to precisely deduce the region of validity of the heuristic approach.

A number of open questions remain: (i) Is there a simple way of constructing a different type of reduced stochastic model which avoids the pitfalls of the common heuristic models and which also avoids the use of sophisticated mathematical analysis to derive it? The requirement of simplicity is essential because typically only such methods are widely adopted and indeed this is a main reason why the problematic heuristic reduced stochastic models treated in this chapter are so widespread. (ii) What would be the differences between heuristic and correctly reduced stochastic spatial models of genetic feedback loops in the limit of fast gene switching? Would the differences between the two models increase or decrease with the diffusion coefficient of protein molecules? Spatial modeling of such systems is relatively rare but recent work in this direction [166, 167, 168, 169, 48] shows that these models are richer in complex behavior than their non-spatial counterparts and of course they are closer to reality. (iii) Say one used a heuristic reduced stochastic model to construct a likelihood function and then use the latter within a Bayesian approach to infer parameters of auto-transcriptional feedback loops from experimental data: how would these differ from parameters inferred using a likelihood built from a non-reduced model? A recent study [170] shows that inference from moment-based approaches is very sensitive to the type of approximation used

to construct the likelihood function and hence suggests large differences between the parameters inferred using heuristic reduced or exact master equations. In conclusion, our study shows that care must be exerted in the interpretation of the results of heuristic stochastic models.

Chapter 4

Steady-state fluctuations of a genetic feedback loop with fluctuating rate parameters using the unified colored noise approximation

This chapter has been published as [2] entitled *Steady-state fluctuations of a genetic feedback loop with fluctuating rate parameters using the unified colored noise approximation* in the *Journal of Physics A: Mathematical and Theoretical*. Slight modifications have been made for its inclusion in this thesis.

4.1 Abstract

A common model of stochastic auto-regulatory gene expression describes promoter switching via cooperative protein binding, effective protein production in the active state and dilution of proteins. Here we consider an extension of this model whereby colored noise with a short correlation time is added to the reaction rate parameters—we show that when the size and time scale of the noise is appropriately chosen it accounts for fast reactions that are not explicitly modelled, e.g., in models with no mRNA description, fluctuations in the protein production rate can account for rapid multiple stages of nuclear mRNA processing which precede translation in eukaryotes. We show how the unified colored noise approximation can be used to derive expressions for the protein number distribution that is in good agreement with stochastic simulations. We find that even when the noise in the rate parameters is small, the protein distributions predicted by our model can be significantly different than models assuming constant reaction rates.

4.2 Introduction

Proteins perform a large range of cellular functions and hence it is of great interest to understand how the genes that produce them operate. Autoregulation is a mechanism to regulate gene expression whereby proteins expressed by a certain gene can subsequently bind to the same gene and cause an increase or a decrease in its expression (positive and negative feedback, respectively) [171]. Autoregulation is common; for example in *E. coli* it is estimated that 40% of all transcription factors are self-regulated [11, 14].

For at least two decades, it has been known that gene expression is inherently stochastic [172, 9], and as such the modelling of gene regulatory networks must account for this stochasticity. Following van Kampen [8], given a system of interest, noise can be seen as originating from two different sources: (i) noise that is inherent to the system itself and cannot be turned off, also called internal or intrinsic noise; (ii) noise coming from a source outside the system of interest, known as external or extrinsic noise. If we specify a system of interest that is described by a set of reactions with constant rate parameters, then it follows that any fluctuations in the molecule numbers must be due to the inherent randomness in the time at which the reactions fire, and hence the noise is intrinsic. In contrast, if we add fluctuations to the rate parameters to account for external processes, then it follows that this noise is extrinsic. For example, if one models mRNA transcription from a gene by a first-order reaction with a constant rate parameter then one is only modelling intrinsic noise. However, if one adds noise to the transcription rate to account for fluctuating numbers of polymerases and transcription factors that are not explicitly described in the system, then one is modelling both intrinsic and extrinsic noise sources. Note that these definitions of intrinsic and extrinsic noise are generally different from, and not to be confused with, the definitions proposed using dual-reporter methods in [10].

The division of noise into these two categories is of course artificial but it is useful from a conceptual and modelling point of view. The simulation of stochastic biochemical processes is most commonly done using the stochastic simulation algorithm (SSA) [173] which assumes that the rate parameter of a reaction will not change in the interval between two successive reaction events, i.e., it models intrinsic noise only. While this may be the case in many situations, it is not generally true. This is because whenever we have an effective reaction that lumps together a large number of intermediate reactions (a multi-stage reaction process), we are making the inherent assumption that these intermediate reactions occur very fast and hence naturally the effective rate parameter is fluctuating on a fast time scale.

Taking into account these fluctuations is however not a simple feat. The chemical master equation (CME, [74, 8]) describing the Markov process simulated by the SSA has been solved exactly or approximately to obtain the protein number distribution in steady-state for a wide variety of models of autoregulation [6, 42, 75, 17, 174, 76, 91, 175, 176, 84, 151, 177, 178], provided the rate parameters are assumed to be constant. There are however a number of studies that have analyzed stochastic models with fluctuating rate parameters. The importance of studying these models stems from the fact that they potentially offer a compromise between model precision (how well can the model capture the complexity of the underlying biochemical dynamics by means of fluctuating parameters) and analytical tractability (how easy it is to solve the stochastic model). Modifications of the linear noise approximation (a type of Fokker-Planck approximation of the CME) incorporating noise in the rate parameters have proved popular to approximate moments for systems subject to small magnitudes of noise with certain properties: (i) for time-independent Gaussian colored noise [179, 180] and (ii) more realistic lognormally distributed noise [181]. Wentzel-Kramers-Brillouin (WKB) methods have also been utilised for cases where the correlation time of the colored noise is tending either to zero or to infinity [182]. These methods provide probability distributions for systems where the noise on the rate parameters is drawn from a negative binomial distribution, however their analysis does not easily translate to finding good approximations for steady states probability distributions where the correlation time of colored noise is neither small or large.

The focus of the present article is threefold: (i) to provide a general method by which one can obtain analytical expressions for the steady-state protein distributions of auto-regulatory gene circuits with fluctuating rate parameters, through the use of the *unified colored noise approximation* (UCNA) [183], (ii) to use this method to investigate the effects that extrinsic noise of different magnitude and time scales has on auto-regulatory gene expression and (iii) to show how the colored noise formalism can be used to describe complex models of autoregulation that involve multi-stage protein production and multi-stage protein degradation. We note that the UCNA was previously utilised in a gene expression context [184] for linear reaction networks that are deterministically monostable and in which there is no feedback mechanism. Our analysis goes further, exploring the addition of colored noise to a non-linear reaction network which expresses deterministic bistability, whilst also incorporating intrinsic fluctuations from the core gene expression processes.

The structure of this chapter is as follows. In Section 4.3 we introduce the cooperative auto-regulatory reaction scheme that we will study in this article. We also show that for non-fluctuating rate parameters, the analytical protein distribution given by the chemical Fokker-Planck equation provides an excellent approximation of the protein distribution solution of the CME, in the limit of fast gene switching. In Section 4.4 we

add colored noise to each reaction in the auto-regulatory reaction scheme (assuming fast gene switching) and use the UCNA to derive the protein number distribution solution of the chemical Fokker-Planck equation. The solution is shown to be in good agreement with a stochastic simulation algorithm modified to account for extrinsic noise on the rate parameters. We also use the solution to investigate the effect that extrinsic noise has on the number of modes of the protein distribution and clarify the limits of the UCNA derivation, including the three main conditions which cause it to breakdown. In Section 4.5 we extend the analysis to the limit of slow gene switching by introducing a conditional version of the UCNA. In Section 4.6 we show two examples of how one can successfully model complex auto-regulatory systems by means of simpler ones with colored noise on the reaction rate parameters, here done for multi-stage protein production and multi-stage degradation. We conclude in Section 4.7 with a discussion of our results and further problems to be addressed on this topic.

4.3 Approximate solution for autoregulation with non-fluctuating rates

We consider the reaction scheme for a genetic non-bursty cooperative feedback loop, where for simplicity we neglect the presence of mRNA:

$$G \xrightarrow{\rho_u} G + P,\ G^* \xrightarrow{\rho_b} G^* + P,\ G + 2P \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*,\ P \xrightarrow{d} \varnothing. \tag{4.1}$$

The reactions $G \xrightarrow{\rho_u} G + P$ and $G^* \xrightarrow{\rho_b} G^* + P$ model the production of protein P in each gene state, $G + 2P \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*$ models the binding and unbinding of the gene to the proteins (with cooperativity 2), and $P \xrightarrow{d} \varnothing$ models the dilution/degradation of proteins inside the cell. For simplicity we assume that there is only one gene copy present in the system and it can be in one of two states, G or G^*, at any one time (some models in the literature consider more states than two [185, 186, 187]). Note also that the reaction modelling protein binding to the gene is to be understood as an effective reaction in cases where the protein binds to enhancer regions rather than directly to the promoter [188]. Before considering the addition of colored noise to the reaction rate parameters above, we first consider the solution with constant rate parameters to provide a reference point for approximations made in Section 4.4, and to clarify the approximation of a CME by a one variable chemical Fokker-Planck equation (FPE).

The CME for the reaction scheme in Eq. (4.1) does not have a known exact solution, even at steady-state for constant reaction rate parameters, and so approximations are necessary. Note that in what follows, we will use the terminology "reaction rate parameters" and "rates" interchangeably. We first consider the limit of fast gene

switching—i.e., the frequency of gene activation and inactivation events is much larger than the frequency of any other reaction in the system. Later in Section 4.5 we will discuss approximations for the slow switching limit. Where $[g^*]$ and $[g]$ are the deterministic mean number of bound and unbound gene respectively and $[n]$ is the mean protein number, the rate equations for the reaction scheme in Eq. (4.1) are:

$$\frac{d[g]}{dt} = \sigma_u[g^*] - \frac{\sigma_b}{\Omega^2}[g][n]^2, \tag{4.2}$$

$$\frac{d[n]}{dt} = 2\sigma_u[g^*] - \frac{2\sigma_b}{\Omega^2}[g][n]^2 + \rho_u[g] + \rho_b[g^*] - d[n], \tag{4.3}$$

where $[g] + [g^*] = 1$. For clarity we state the units of each rate parameter: ρ_u, ρ_b, d and σ_u have units of s^{-1}, and σ_b has units of $\mathrm{Volume}^2 \cdot \mathrm{s}^{-1}$. This ensures a matching of the units with the left hand side of Eqs. (4.2–4.3), which has units of s^{-1}. In the fast switching limit, the gene rapidly equilibrates to quasi-steady state conditions, i.e., $d[g]/dt \approx d[g^*]/dt \approx 0$ and hence the deterministic rate equation for mean protein number reduces to a much simpler form:

$$\frac{d[n]}{dt} = \frac{L\rho_u + \rho_b([n]/\Omega)^2}{L + ([n]/\Omega)^2} - d[n], \tag{4.4}$$

where $L = \sigma_u/\sigma_b$. Note that the reaction scheme here described exhibits deterministic bistability over some regions of the parameter space. This equation is consistent with a birth-death process where proteins are produced via a reaction with a rate that is dependent on the number of proteins and are destroyed by a first-order reaction [1]. The CME for this reduced process is given by:

$$\frac{dP_a(n,t)}{dt} = T^+(n-1)P_a(n-1,t) + T^-(n+1)P_a(n+1,t) - (T^+(n) + T^-(n))P_a(n,t), \tag{4.5}$$

where $P_a(n,t)$ is the probability that at a time t there are n proteins in the system; $T^+(n)$ and $T^-(n)$ are the propensities of protein production and degradation respectively. The subscript a denotes that this is the probability for the reduced system, an *approximate* solution to the master equation of the full system. $T^+(n)dt$ is the probability, given n proteins are in the system, that a protein production reaction occurs, increasing the protein number of the system by 1, in the time interval $[t, t+dt)$. Similarly, $T^-(n)dt$ is the probability, given n proteins are in the system, that a protein degradation event occurs, decreasing the protein number by 1, in the time interval $[t, t+dt)$. These propensities

are given by:

$$T^+(n) = \frac{\rho_u L + \rho_b (n/\Omega)^2}{L + (n/\Omega)^2}, \tag{4.6}$$

$$T^-(n) = d\, n. \tag{4.7}$$

These propensities are deduced directly from the form of the effective rate equation in Eq. (4.4). Essentially, the probability for a particular reaction per unit time is taken to be the same as the reaction rate in the effective deterministic rate equation with $[n]$ replaced by n. *We emphasise that while this appears to be a heuristic rule with no apparent fundamental microscopic basis, it has been shown that the reduced master equation based on it provides an accurate approximation to the SSA of the full reaction system in fast gene switching conditions provided the low protein number states are rarely visited [76, 1].*

The exact steady state solution of the one variable master equation given by Eq. (4.5) can be found using standard methods [74]:

$$P_a(n) = P_a(0) \prod_{z=1}^{n} \frac{T^+(z-1)}{T^-(z)}, \tag{4.8}$$

where $P_a(0)$ is the steady state probability evaluated at $n = 0$ (acting effectively here as a normalisation constant). We can further approximate the reduced master equation in Eq. (4.5) by a Fokker-Planck equation [189, 8, 74]:

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n}(a_1(n)P(n,t)) + \frac{1}{2}\frac{\partial^2}{\partial n^2}(a_2(n)P(n,t)), \tag{4.9}$$

where $a_1(n)$ and $a_2(n)$ are the first two jump moments, given by $a_1(n) = T^+(n) - T^-(n)$ and $a_2(n) = T^+(n) + T^-(n)$ respectively, and $P(n,t)$ denotes the FPE solution (a notation used throughout this chapter). The purpose of this further approximation by means of a FPE will be made clear in Section 4.4.1. Eq. (4.9) has a steady state solution of the form [8]:

$$P(n) = \frac{N}{T^+(n) + T^-(n)} \exp\Big(2 \int^n \frac{T^+(z) - T^-(z)}{T^+(z) + T^-(z)}\, dz\Big), \tag{4.10}$$

where N is a normalisation constant. Although the integral in the exponent of Eq. (4.10) can be solved exactly with propensities of the form of Eq. (4.6) and Eq. (4.7) since it is the integral of the ratio of two cubic polynomials, the solution is too complicated to be detailed here. The approximations made by the FPE approximation are that (i) fluctuations in the protein number are small and (ii) we are in the fast switching regime between the gene states. Fig. 4.1 compares the FPE solution Eq. (4.10) with

the solution of the heuristic CME in Eq. (4.8) and the solution of the full CME of the reaction scheme in Eq. (4.1) using the finite space projection method (FSP) [69]. Note that provided the state space is truncated large enough, the FPE solution matches the solution of the heuristic CME almost exactly. Clearly, when gene switching is fast (bottom plot of Fig. 4.1) all three solutions agree with each other. However, when gene switching is not fast (top and middle plots on Fig. 4.1) both the reduced CME and FPE solutions are a poor approximation of the true distribution from FSP.

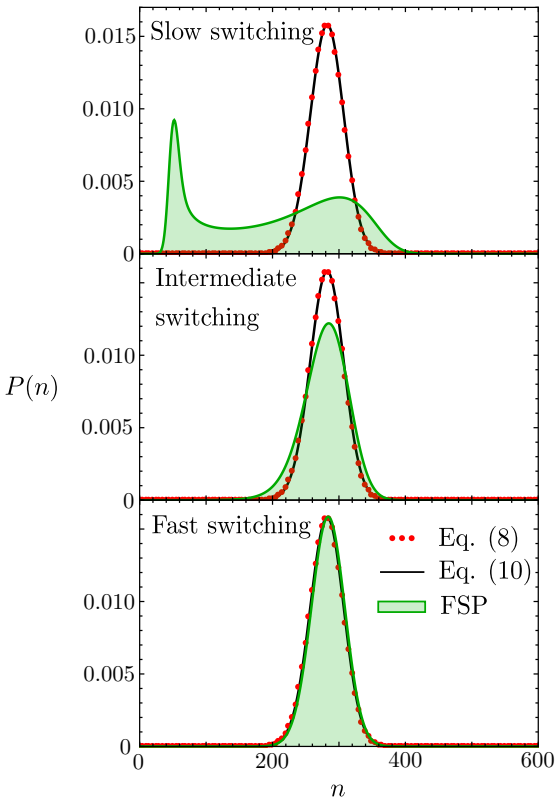


Figure 4.1: Comparison of the heuristic reduced master equation solution from Eq. (4.8) (red dots), the FPE solution from Eq. (4.10) (black line) and the solution of the full cooperative network using FSP (green shaded region). Shared parameters in each plot are $\rho_u = 50$, $\rho_b = 400$, $\Omega = 200$ and $d = 1$. The FSP gives the exact solution for a truncated state space chosen such that the neglected probability mass is negligible. The top plot shows distributions for the case $\sigma_u = \sigma_b = 5$, where clearly the heuristic master equation and FPE solutions are a poor approximation of the FSP. The middle plot shows distributions for the case $\sigma_u = \sigma_b = 50$ where we can observe a convergence of the heuristic master equation and FPE solutions towards the FSP solution. The bottom plot shows excellent agreement of the FSP with the heuristic master equation and FPE solutions for fast switching where $\sigma_u = \sigma_b = 5 \times 10^3$.

4.4 Accounting for fluctuating rates using the UCNA

Fluctuating rate parameters can be used to include a description of processes not explicitly taken into account in the formulation of a model. In Fig. 4.2 we illustrate this idea. In this section, we add fluctuations to the rate parameters of the FPE description derived earlier and use the UCNA to obtain a new effective FPE that is valid when the time scale of the noise on the rates is either very small or very large. We remind the reader that in this section we consider gene switching to be fast, and we consider the case of slow switching in Section 4.5.

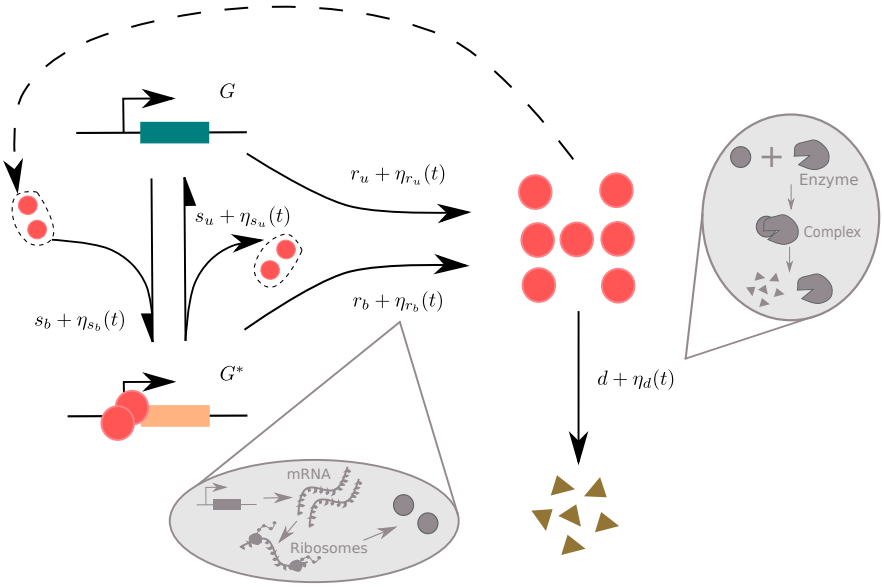


Figure 4.2: Illustration of the cooperative auto-regulatory reaction scheme, with colored noise included on each individual reaction. For the case of non-fluctuating rates explored in Section 4.3 the noise terms, η_i, on the rate parameter can be set to zero. Where colored noise is included in Section 4.4 these noise terms are not set to zero. The addition of noise onto rate parameters can be thought of as accounting for processes that are not explicitly included in the gene expression model. Here we show two examples, where colored noise on the rate parameters of the reduced model can be used to account for mRNA number fluctuations during protein translation, or the degradation of proteins via an enzyme catalytic mechanism.

4.4.1 Fluctuating degradation rate

We begin by considering the case of a fluctuating degradation rate. These fluctuations could for example stem from details of the degradation machinery that are not explicitly described in the model, e.g multi-step degradation mediated by enzymatic reactions.

The equivalent Langevin equation to the Fokker-Planck equation from Eq. (4.9) using the propensities from Eqs. (4.6) and (4.7) is given by [100, 8]:

$$\frac{dn}{dt} = \frac{\rho_u L + \rho_b (n/\Omega)^2}{L + (n/\Omega)^2} - d\, n + \sqrt{\frac{\rho_u L + \rho_b (n/\Omega)^2}{L + (n/\Omega)^2} + d\, n}\ \cdot \Gamma(t), \tag{4.11}$$

where $\Gamma(t)$ is Gaussian white noise with zero mean and correlator $\langle\Gamma(t)\Gamma(t')\rangle = \delta(t-t')$. Now we introduce a fluctuating degradation rate by setting $d = d_0(1+\eta(t))$, where $\eta(t)$ is Gaussian colored noise with a mean of zero and correlator $\langle\eta(t)\eta(t')\rangle = (D/\tau)\exp(-|t-t'|/\tau)$ [183, 190]. Here, τ is the *correlation time* of the colored noise, D/τ is the *noise strength* (the variance of fluctuations) and d_0 is the mean degradation rate. Since D/τ is the noise strength, i.e., D scales the noise strength at constant τ, we occasionally refer to D itself as the *noise strength* (where τ is a fixed parameter). In the limit of $\tau \to 0$ colored noise becomes white noise since $\lim_{\tau\to 0}\langle\eta(t)\eta(t')\rangle = D\delta(t-t')$. Note that $\eta(t)$ must satisfy $|\eta(t)| \ll 1$ such that d is a positive quantity (and hence admits physical interpretation as a rate parameter). The inclusion of colored noise can be approximated by the following two component system [183]:

$$\frac{dn}{dt} = \frac{\rho_u L + \rho_b (n/\Omega)^2}{L + (n/\Omega)^2} - d_0(1+\eta)\, n + \sqrt{\frac{\rho_u L + \rho_b (n/\Omega)^2}{L + (n/\Omega)^2} + d_0\, n}\ \cdot \Gamma(t), \tag{4.12}$$

$$\frac{d\eta}{dt} = -\frac{1}{\tau}\eta + \frac{1}{\tau}\theta(t), \tag{4.13}$$

where $\theta(t)$ is Gaussian white noise with zero mean and correlator $\langle\theta(t)\theta(t')\rangle = 2D\delta(t-t')$, and the time dependence on the protein number $n(t)$ and noise $\eta(t)$ is suppressed for notational convenience. Note that in the argument of the square root above we have replaced $\eta(t)$ by its mean of zero; this constitutes a mean-field type of approximation, and is useful such that one can solve Eqs. (4.12)–(4.13) analytically—however, where the noise is small, i.e., $\eta(t) \ll 1$, this is generally a good approximation since $d \sim d_0$ (however, this is not always true as explored in *Condition 3* in Section 4.4.4). Note that we also use this mean-field assumption in Sections 4.4.2 and 4.4.3. For transparency, we rewrite Eqs. (4.12)–(4.13) as:

$$\frac{dn}{dt} = h(n) + g_1(n)\eta + g_2(n)\Gamma(t), \tag{4.14}$$

$$\frac{d\eta}{dt} = -\frac{1}{\tau}\eta + \frac{1}{\tau}\theta(t), \tag{4.15}$$

with

$$h(n) = \frac{\rho_u L + \rho_b (n/\Omega)^2}{L + (n/\Omega)^2} - d_0 n, \tag{4.16}$$

$$g_1(n) = -d_0 n, \tag{4.17}$$

$$g_2(n) = \sqrt{\frac{\rho_u L + \rho_b (n/\Omega)^2}{L + (n/\Omega)^2} + d_0\, n}. \tag{4.18}$$

In order to approximately solve Eqs. (4.14)–(4.15) we next utilise the UCNA to obtain reduced Langevin equations when the noise η is either very fast or very slow. For completeness, we present a non-rigorous but intuitive proof of the UCNA along the lines found in [183] which essentially consists of a direct adiabatic elimination on the stochastic differential equations (SDEs) in Eqs. (4.14)–(4.15). For a more rigorous derivation of a UCNA-like FPE we advise reader to read the seminal work of Fox, who introduced a functional calculus approach to the study of colored noise SDEs [191, 192, 193, 194]. A review of the differing UCNA-like derivations can be found in [195].

It has been discussed in [183, 195] that the adiabatic elimination we employ below is exact for $\tau \to 0$ (white noise) or $\tau \to \infty$ (highly correlated noise) but that it should give a useful approximation for intermediate values of τ. We note that the theory provided by Roberts *et al.* [182] does not provide such a result as they consider separately the cases of $\tau \to 0$ and $\tau \to \infty$. For the biological applications we consider in Section 4.6 the limit of $\tau \to \infty$ is not of interest, and we will later focus on the limit of τ small, although the derivation shown here holds for large τ too. First, where we use overdots to represent derivatives with respect to time t, one should proceed in rearranging Eq. (4.14) for η:

$$\eta(n, \dot{n}) = \frac{1}{g_1(n)}(\dot{n} - h(n) - g_2(n)\Gamma(t)). \tag{4.19}$$

In what follows we will utilise a mean-field approximation (denoted by the subscript *mf*) to approximately calculate the time derivative of $\eta(n, \dot{n})$. We start by defining the mean-field approximation of $\eta(n, \dot{n})$ as:

$$\eta_{mf}(n_{mf}, \dot{n}_{mf}) = \frac{1}{g_1(n_{mf})}(\dot{n}_{mf} - h(n_{mf})). \tag{4.20}$$

Taking the time derivative with respect to non-dimensional time $\hat{t} = t/\tau$ (denoted by the overdot) we obtain:

$$\dot{\eta}_{mf} = \frac{1}{g_1(n_{mf})}\left(\frac{h(n_{mf})g_1'(n_{mf})}{g_1(n_{mf})} - h'(n_{mf})\right)\dot{n}_{mf} + \frac{\tau^{-1}}{g_1(n_{mf})}\left(\ddot{n}_{mf} - \frac{g_1'(n_{mf})}{g_1(n_{mf})}\dot{n}_{mf}^2\right), \tag{4.21}$$

where the prime on each function of n_{mf} denotes the derivative with respect to n_{mf}. In the limit of $\tau \to 0$, the second term on the right hand side of Eq. (4.21) goes to infinity and hence the only way to keep the time derivative finite is to impose the condition:

$$\ddot{n}_{mf} - \frac{g_1'(n_{mf})}{g_1(n_{mf})}\dot{n}_{mf}^2 = 0. \tag{4.22}$$

This then implies that in this limit we have:

$$\dot{\eta}_{mf} \approx \frac{1}{g_1(n_{mf})}\left(\frac{h(n_{mf})g_1'(n_{mf})}{g_1(n_{mf})} - h'(n_{mf})\right)\dot{n}_{mf}. \tag{4.23}$$

Note that taking the limit of $\tau \to \infty$ gives the same result and hence the approximation Eq. (4.23) is valid in both the limit of small and large τ. This can be shown to be self-consistently true; taking the time-derivative of Eq. (4.14) alongside a mean-field approximation we get,

$$\ddot{n}_{mf} = \left(h'(n_{mf}) + g_1'(n_{mf})\eta_{mf}\right)\dot{n}_{mf} + g_1(n_{mf})\dot{\eta}_{mf}, \tag{4.24}$$

Assuming Eqs. (4.20) and (4.23) to be true one then recovers

$$\ddot{n}_{mf} - \frac{g_1'(n_{mf})}{g_1(n_{mf})}\dot{n}_{mf}^2 = 0, \tag{4.25}$$

which means that if Eq. (4.23) holds true then so does Eq. (4.22) (and *vice versa*).

In Eq. (4.15) we can now substitute η from (4.19) and $\dot{\eta}_{mf}$ for $\dot{\eta}$ from Eq. (4.23) giving us the UCNA for the system with colored noise on the degradation rate, which is *exact* in the limits $\tau \to 0$ or $\tau \to \infty$:

$$\dot{n} \approx \frac{h(n)}{C(n,\tau)} + \frac{1}{C(n,\tau)}(g_1(n)\theta(t) + g_2(n)\Gamma(t)), \tag{4.26}$$

where

$$C(n,\tau) = 1 + \tau\left(\frac{g_1'(n)h(n)}{g_1(n)} - h'(n)\right). \tag{4.27}$$

Note that we have dropped off the *mf* subscript for clarity. Finally, in order to get a simplified Langevin equation, we modify Eq. (4.26) such that we only have one effective Gaussian white noise term. We begin by proposing:

$$g(n)\tilde{\Gamma}(t) = g_1(n)\theta(t) + g_2(n)\Gamma(t), \tag{4.28}$$

where $\tilde{\Gamma}(t)$ is Gaussian white noise with mean zero and correlator $\langle\tilde{\Gamma}(t)\tilde{\Gamma}(t')\rangle = 2\delta(t-t')$, and then use relations between the correlators to find our unknown $g(n)$. Note that we assume zero correlation between $\Gamma(t)$ and $\theta(t)$, i.e., $\langle\Gamma(t)\theta(t')\rangle = \langle\Gamma(t')\theta(t)\rangle = 0$. Explicitly, utilising the correlators, we find:

$$g(n)^2\langle\tilde{\Gamma}(t)\tilde{\Gamma}(t')\rangle = g_1(n)^2\langle\theta(t)\theta(t')\rangle + g_2(n)^2\langle\Gamma(t)\Gamma(t')\rangle, \tag{4.29}$$

which gives us

$$g(n) = \sqrt{Dg_1(n)^2 + \frac{1}{2}g_2(n)^2}. \tag{4.30}$$

Hence, our final reduced Langevin equation is given by:

$$\dot{n} = \frac{h(n)}{C(n,\tau)} + \frac{g(n)}{C(n,\tau)}\tilde{\Gamma}(t), \tag{4.31}$$

which corresponds to the result in [190]. Note that Eqs. (4.26) and (4.31) are *identical*. Here we pause to make a couple of comments on $C(n,\tau)$, which can be interpreted as a renormalisation of the Langevin equation in Eq. (4.14) to account for the addition of colored noise to the rate parameters. In fact, when $\tau = 0$, Eq. (4.31) recovers the correct Langevin equation for a process with white noise on the rate parameters. One should also note the independence of $C(n,\tau)$ from the strength of the noise D; the renormalisation with respect to the addition of colored noise on the degradation rate is not specific to the size of the noise, it simply accounts for the finite correlation time.

The FPE corresponding to this SDE should be chosen in the Stratonovich form, following from [193, 196, 197], as this is the physical implementation of an SDE with colored noise having a non-zero correlation time τ. This FPE is:

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n}\left[\left(\tilde{h}(n) + \tilde{g}(n)\tilde{g}'(n)\right)P(n,t)\right] + \frac{\partial^2}{\partial n^2}\left[\tilde{g}(n)^2 P(n,t)\right], \tag{4.32}$$

where $\tilde{h}(n) = h(n)/C(n,\tau)$ and $\tilde{g}(n) = g(n)/C(n,\tau)$. Following Eqs. (4.9)–(4.10) in Section 4.3 and [8], the steady state solution to this equation is then given by:

$$P(n) = \frac{N}{\tilde{g}(n)^2}\exp\left(\int^n \frac{\tilde{h}(z) + \tilde{g}(z)\tilde{g}'(z)}{\tilde{g}(z)^2}dz\right) = \frac{N}{\tilde{g}(n)}\exp\left(\int^n \frac{\tilde{h}(z)}{\tilde{g}(z)^2}dz\right), \tag{4.33}$$

where N is the normalisation constant, chosen over the domain $n \in [0,\infty)$.

Having made various approximations to arrive at Eq. (4.33) we now pause and summarise the approximations made thus far, clarifying the conditions under which we expect Eq. (4.33) to produce meaningful distributions. We started by considering Eq. (4.11) which is the chemical Langevin equation describing protein dynamics and which was derived from the CME describing the reaction scheme in Eq. (4.1) under the approximations of large protein numbers and fast promoter switching. Subsequently we added colored noise to the degradation rate in Eq. (4.11) and made a mean-field approximation (valid for small fluctuations about the mean degradation rate) to obtain the coupled Langevin equations Eqs. (4.12)–(4.13). These equations were then reduced to a single effective Langevin equation Eq. (4.31) by the UCNA under the assumption that the correlation time of colored noise is very small or very large. Finally the solution of this Langevin

equation is given by Eq. (4.33). Hence summarizing, we expect the latter solution to be an accurate stochastic description of the protein fluctuations in reaction scheme (4.1) with a fluctuating degradation rate provided the protein numbers are large, promoter switching is fast and the correlation time of fluctuations in the degradation parameter is either very small or very large.

To test the accuracy of the distributions for colored noise provided by the UCNA in Eq. (4.33), we compare the UCNA solution to a distribution produced from a modified SSA that explicitly accounts for the colored noise on the degradation rate. This modification is given in full detail in Appendix B.1. Essentially, the dilution/degradation reaction $P \to \varnothing$ is replaced by three new reactions alongside the introduction of a ghost species Y, these being (i) $\varnothing \rightleftharpoons Y$ and (ii) $P + Y \to Y$. The rates of these new reactions are then chosen to ensure the magnitude of *effective* external noise on the degradation reaction, due to fluctuations in molecule numbers of the ghost species, match the colored noise SDE given in Eq. (4.13).

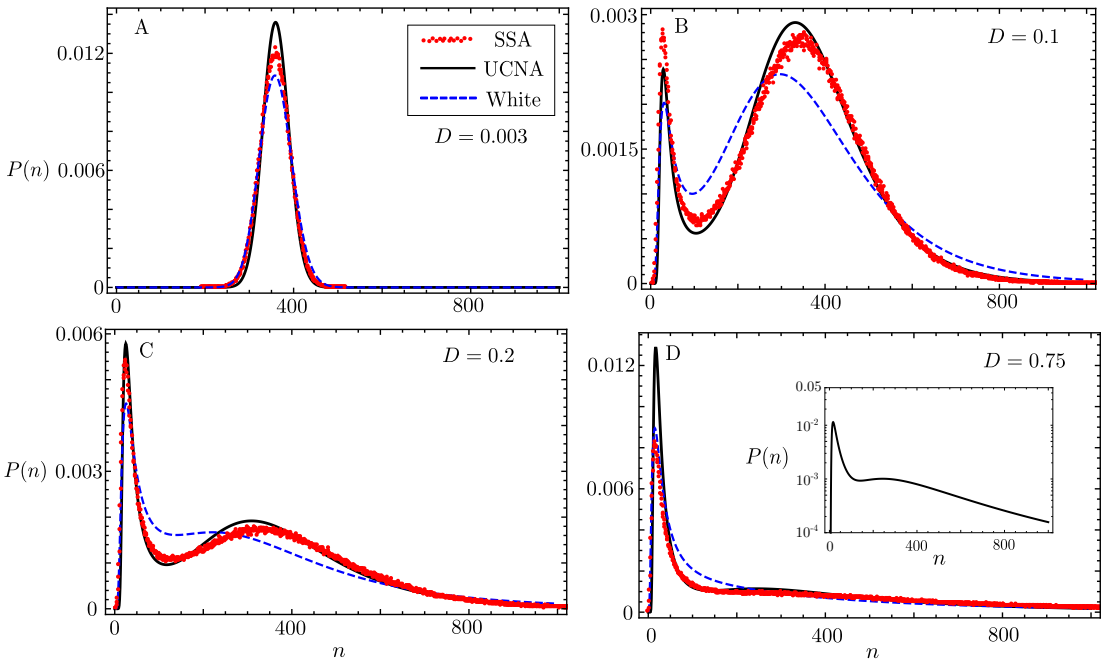


Figure 4.3: Comparison of the UCNA (black line) from Eq. (4.33) and white extrinsic noise (UCNA with $\tau = 0$, dashed blue line) with stochastic simulations using the modified SSA (red points) of the cooperative reaction scheme in Eq. (4.1), where the colored noise is added to the degradation rate. Aside from variation in the strength of noise D (shown on each plot), the shared parameters are $\rho_u = 24$, $\rho_b = 464$, $\sigma_b = \sigma_u = 1000$, $d_0 = 1, \Omega = 200$ and $\tau = 1$. Parameters σ_b and σ_u are chosen to be large compared to other system parameters such that the frequency of gene activation and inactivation events is much larger than the frequency of other reaction events, i.e., the fast gene switching assumption. Note that for this choice of rate parameters, the rate equations are bistable with equilibrium points at $n = 47.4, 360.4$. The criterion $\sqrt{D/\tau} < 1$ is required to ensure positivity of the degradation rate. As the extrinsic noise is increased, the mass of the distribution shifts from the mode at 360.4 to the mode at 47.4. The inset of D shows the same distribution but with the y-axis on a log scale, emphasising the exponential tail of the distribution for large n. SSA data in each case comes from a single steady state trajectory of 9×10^6s.

Fig. 4.3 shows steady state probability distributions produced by the UCNA for various values of D for a deterministically bistable set of parameters. The UCNA correctly captures the shift of the probability mass from the equilibrium point of higher molecule number (referred to as the *upper mode*) to the lower equilibrium point (referred to as the *lower mode*) as D is increased. Importantly, this shows that when gene switching is assumed to be fast, colored noise can induce bimodality—one should keep this in mind for when we look at slow gene switching in Section 4.5. Readers should also note that the parameter choices have been selected such that the Fokker-Planck approximation is good, notably that the system size is large, i.e., $\Omega \gg 1$, and the mean number of proteins in the system is also large. In all cases $\sqrt{D/\tau} < 1$ so that the degradation rate remains positive. The behaviour seen as D increases in Fig. 4.3 can be explained as follows. When D is small (Fig. 4.3A) the colored noise η in Eq. (4.15) is also small compared to the mean number of molecules in the system, and the noise cannot force the system out of the upper mode. As D gets larger (Figs. 4.3B and 4.3C) the fluctuations η at the upper mode also become larger, allowing the system to explore the lower mode. When the system is found in the lower mode the pre-factor of the coloured noise in Eq. (4.14), $g_1(n) \propto n$, is lesser in magnitude, and the fluctuations in η are much smaller than when the system inhabits the upper mode hence the increased probability mass at the lower mode. That the system is less noisy at the lower mode means that it is much less likely that a large fluctuation will propel the system into the upper mode. These properties of the system as D increases can be further seen through (i) the increase in probability mass found at the lower mode as D increases thoroughout all of Fig 4.3(A–D), and (ii) the increased probability mass found in the tail of the distribution for large n (Fig. 4.3D); while the tail is very slowly decaying it is still exponential and hence the distribution is not heavy-tailed (see the inset of Fig. 4.3D). This ability to induce bimodality through a more detailed description of the details of the degradation process is important in the context of *cellular decision-making*. It is hence possible for regions of the reaction rate parameter space previously thought unable to induce multiple phenotypic states to do so with an increasing influence of more complex degradation mechanisms. Note that for the majority of cases in Fig. 4.3, the UCNA provides a much better approximation than the white noise approximation, hence one cannot simply assume that since the correlation time τ is relatively small that it can be approximated as zero.

Fig. 4.4 shows how the UCNA responds to increasing correlation time τ while the noise strength, D/τ, remains fixed. For all cases where τ is small, the UCNA performs very well. As τ increases however the UCNA starts to predict ever increasing negative probabilities for some values of n. Notably though, Fig. 4.4B shows that even where significant negative probability is predicted at large τ, the UCNA still manages to capture the rest of the distribution. This negativity of $C(n, \tau)$ is commented on in both [183] and [193]. The former deals with this negativity by taking the absolute magnitude

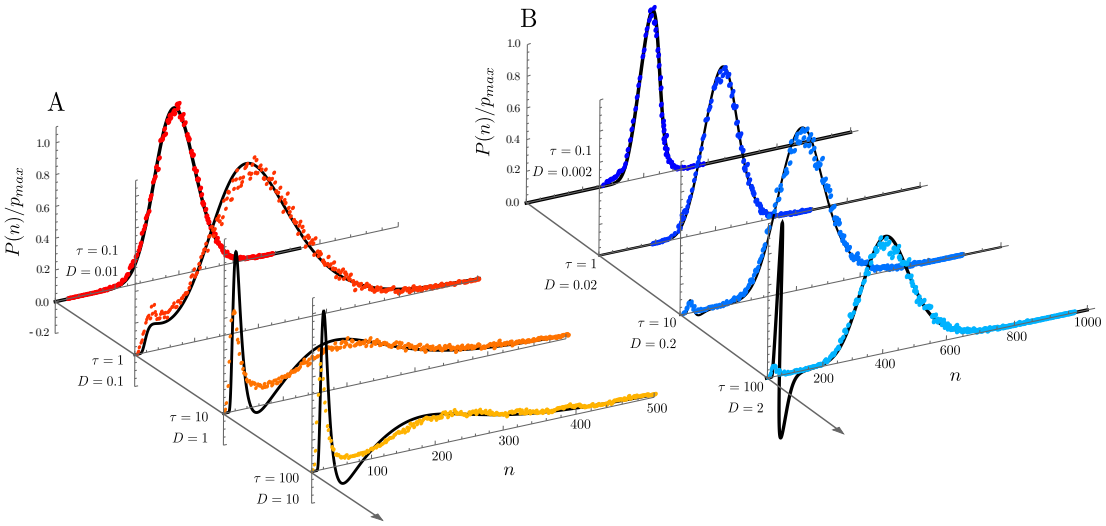


Figure 4.4: Comparison of the UCNA (black line) against the modified SSA (colored dots) as the correlation time τ is increased at constant noise size D/τ. Note that the y-axis shows $P(n)/p_{max}$, where $P(n)$ is defined in Eq. (4.33) and p_{max} is equal to the maximum value of $P(n)$. (A) Shows the performance of increasing τ for a system with parameters $\rho_u = 20$, $\rho_b = 250$, $d_0 = 1$, $\sigma_u = 3 \times 10^2$, $\sigma_b = 10^3$ and $\Omega = 100$. Deterministically this system is monostable with an equilibrium point at $n = 194.7$, however as τ is increased a shift towards a lower mode is observed. When τ is sufficiently large, the UCNA predicts a negative probability. (B) Shows similar to (A) but with parameters $\rho_u = 25$, $\rho_b = 480$, $d_0 = 1$, $\sigma_u = 8 \times 10^2$, $\sigma_b = 10^3$ and $\Omega = 200$. This too is a deterministically monostable system with equilibrium point $n = 406.0$. As τ increases, the breakdown of the UCNA is more apparent than for (A) with the prediction of negative probability for small n more drastic. Both (A) and (B) show that unless τ is large, while D/τ is small, the UCNA provides a very good approximation, even where the colored noise induces bimodality in deterministically monomodal systems. SSA data in each case comes from a single steady state trajectory of 9×10^5s.

of the pre-factor of the exponential in Eq. (4.33), while the latter comments that the proof of their UCNA-like FPE is only formally valid where $C(n, \tau) > 0$, $\forall\, n$. Here we choose not to take the magnitude of the pre-factor in Eq. (4.33), since although this leads to a positive probability for all n it is nonetheless a poor approximation; but we take careful note of the comment made by Fox in [193], as this indicates where the UCNA will perform well. The intuition behind the argument of Fox can be stated as: if for some n, $C(n, \tau) < 0$ there must be a transitory value of n for which $C(n, \tau) = 0$, at this point the Eq. (4.31) becomes physically ill-defined and our solution is invalid.

Finally, we observe that the parameter values chosen for both plots in Fig. 4.4 correspond to deterministically monostable systems. The bimodality that is observed in Fig. 4.4 is hence *noise induced bimodality*. The mode that appears for small τ corresponds to the deterministic equilibrium point, whereas the noise induced mode does not correspond to an equilibrium point of the deterministic system. We notice that the ability to exhibit a noise induced mode as τ becomes large is especially true for monostable parameter sets which are in close proximity to bistable parameter sets in the parameter space. This can be explained by occasional jumps between the monostable and bistable regimes due to sufficiently large fluctuations in the degradation rate. Hence a measure of the

distance here is the difference in the magnitude of d_0 needed such that the system is deterministically bistable divided by the noise strength, defined as $\Delta d_0 = |d_0 - d_c|/(D/\tau)$, where d_c is the closest value of the mean degradation rate to d_0 expressing bistability. For example, the parameter set chosen in Fig. 4.4A, although monostable, is very close to a parameter set that exhibits deterministic bistability ($\Delta d_0 = 2.12$). On the other hand, the parameter set of Fig. 4.4B is far from the bistable parameter regime ($\Delta d_0 = 57.5$)—and hence the bimodality shown is very limited as τ becomes large. The reason for this noise induced bimodality then can be seen by the ability of a system, through fluctuations in the rate parameters, to access parameter regimes which in fact do exhibit deterministic bistability. Importantly, even when it seems bimodality is not induced (e.g., Figs. 4.3A or 4.4B), using the extremal equation of $P(n)$ from [190], i.e., $\tilde{h}(n) = \tilde{g}(n)\tilde{g}'(n)$, one can show that the UCNA still predicts the presence of two modes. This explanation of the induced bimodality in cooperative autoregulation is further supported by the lack of noise induced bimodality when colored noise is included on the degradation rate of the FPE describing non-cooperative autoregulation; here the UCNA's extremal equation only ever predicts the existence of one mode for the probability distribution.

4.4.2 Fluctuating effective protein production rates

We now extend the analysis from Section 4.4.1 to the effective protein production rates. Colored noise on the effective production rates can be used to implicitly model multi-step protein production, including multiple stages of mRNA processing before translation (see Fig. 4.2). We add colored noise onto the effective protein production rates via, $\rho_u = \rho_u^{(0)}(1 + \eta_1(t))$ and $\rho_b = \rho_b^{(0)}(1 + \eta_2(t))$, which upon substituting in the Langevin equation describing the feedback loop Eq. (4.11) we obtain the following set of SDEs:

$$\frac{dn}{dt} = \frac{\rho_u^{(0)} L + \rho_b^{(0)}(n/\Omega)^2}{L + (n/\Omega)^2} - d\,n + \frac{\rho_u^{(0)} L\eta_1 + \rho_b^{(0)}(n/\Omega)^2\eta_2}{L + (n/\Omega)^2} + \sqrt{\frac{\rho_u^{(0)} L + \rho_b^{(0)}(n/\Omega)^2}{L + (n/\Omega)^2} + d\,n}\ \cdot \Gamma(t), \tag{4.34}$$

$$\frac{d\eta_1}{dt} = -\frac{1}{\tau}\eta_1 + \frac{1}{\tau}\theta_1(t), \tag{4.35}$$

$$\frac{d\eta_2}{dt} = -\frac{1}{\tau}\eta_2 + \frac{1}{\tau}\theta_2(t), \tag{4.36}$$

where $\theta_1(t)$ and $\theta_2(t)$ are Gaussian white noise terms with zero mean and correlators $\langle\theta_1(t)\theta_1(t')\rangle = 2D_1\delta(t-t')$ and $\langle\theta_2(t)\theta_2(t')\rangle = 2D_2\delta(t-t')$ respectively. Note that here we have used a mean-field approximation for the terms under the square root, as was done in Section 4.4.1. In a similar style to Eq. (4.28) we now propose a new noise term

$\tilde{\eta}(t)$, which couples $\eta_1(t)$ and $\eta_2(t)$, satisfying:

$$F(n)\tilde{\eta}(t) = f_1(n)\eta_1(t) + f_2(n)\eta_2(t), \tag{4.37}$$

where $f_1(n) = \rho_u^{(0)} L/(L+(n/\Omega)^2)$, $f_2(n) = \rho_b^{(0)}(n/\Omega)^2/(L+(n/\Omega)^2)$ and $\tilde{\eta}(t)$ is colored noise with zero mean and correlator $\langle\tilde{\eta}(t)\tilde{\eta}(t')\rangle = e^{-|t-t'|/\tau}/\tau$, satisfying the following equation:

$$\frac{d\tilde{\eta}}{dt} = -\frac{1}{\tau}\tilde{\eta} + \frac{1}{\tau}\theta(t), \tag{4.38}$$

where $\theta(t)$ is Gaussian white noise with correlator $\langle\theta(t)\theta(t')\rangle = 2\delta(t-t')$. The correlators for $\eta_1(t)$ and $\eta_2(t)$ are $\langle\eta_1(t)\eta_1(t')\rangle = D_1 e^{-|t-t'|/\tau}/\tau$ and $\langle\eta_2(t)\eta_2(t')\rangle = D_2 e^{-|t-t'|/\tau}/\tau$, where we have assumed that the colored noise on both production rates has the same correlation time but a differing magnitude of noise strength. Using the properties of the correlators of η_1, η_2 and $\tilde{\eta}$ we then find:

$$F(n) = \sqrt{f_1(n)^2 D_1 + f_2(n)^2 D_2}. \tag{4.39}$$

Sharing the notation adopted in Section 4.4.1, we define the following:

$$h(n) = \frac{\rho_u^{(0)} L + \rho_b^{(0)}(n/\Omega)^2}{L+(n/\Omega)^2} - dn, \tag{4.40}$$

$$g_2(n) = \sqrt{\frac{\rho_u^{(0)} L + \rho_b^{(0)}(n/\Omega)^2}{L+(n/\Omega)^2} + dn}. \tag{4.41}$$

This gives us the following SDE which is coupled to Eq. (4.38):

$$\frac{dn}{dt} = h(n) + F(n)\tilde{\eta} + g_2(n)\Gamma(t). \tag{4.42}$$

Then, following the same UCNA procedure as in Eqs. (4.19)–(4.26), we obtain the following approximate Langevin equation:

$$\dot{n} \approx \frac{h(n)}{C(n,\tau)} + \frac{1}{C(n,\tau)}(F(n)\theta(t) + g_2(n)\Gamma(t)), \tag{4.43}$$

where

$$C(n,\tau) = 1 + \tau\left(\frac{F'(n)h(n)}{F(n)} - h'(n)\right). \tag{4.44}$$

In this case it is interesting to note that unlike the case of a fluctuating degradation rate, here $C(n,\tau)$ does depend on both the correlation time τ and the strength of the colored noise D_1, D_2 (unless $D_1 = D_2$ in which case there is only dependence on τ). This occurs since the strengths of the noise on each production rate are independent, and hence do not cancel out in $F'(n)/F(n)$. To simplify Eq. (4.43) further, we again propose:

$$g(n)\tilde{\Gamma}(t) = F(n)\theta(t) + g_2(n)\Gamma(t), \tag{4.45}$$

where $\tilde{\Gamma}(t)\tilde{\Gamma}(t') = 2\delta(t-t')$, and find using the correlators that $g(n) = \sqrt{F(n)^2 + g_2(n)^2/2}$. This leads to the final approximate SDE:

$$\dot{n} = \frac{h(n)}{C(n,\tau)} + \frac{g(n)}{C(n,\tau)}\tilde{\Gamma}(t), \tag{4.46}$$

which is identical in notation to Eq. (4.31) but where $h(n)$, $C(n,\tau)$ and $g(n)$ are all defined in this section. The equivalent FPE for this SDE is then:

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n}\left[\left(\tilde{h}(n) + \tilde{g}(n)\tilde{g}'(n)\right)P(n,t)\right] + \frac{\partial^2}{\partial n^2}\left[\tilde{g}(n)^2 P(n,t)\right]. \tag{4.47}$$

Again our solution for the probability distribution will then be:

$$P(n) = \frac{N}{\tilde{g}(n)}\exp\left(\int^n \frac{\tilde{h}(z)}{\tilde{g}(z)^2}dz\right), \tag{4.48}$$

with $\tilde{h}(n) = h(n)/C(n,\tau)$ and $\tilde{g}(n) = g(n)/C(n,\tau)$.

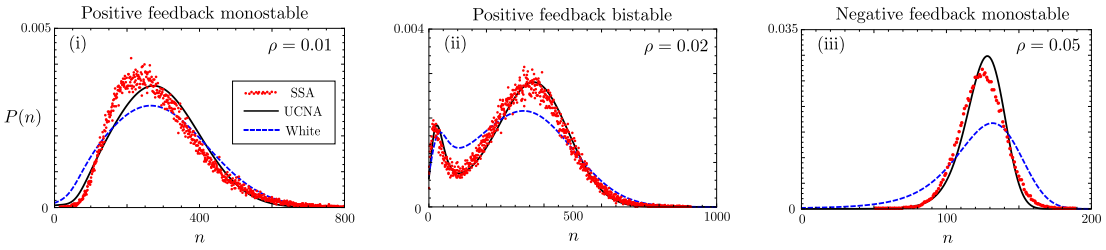


Figure 4.5: This figure shows the agreement of the UCNA on the protein production rates to the modified SSA (detailed in Section 4.4.2), also compared to the case of white extrinsic noise ($\tau = 0$). The plots show agreement of the UCNA over the three main qualitative regimes of cooperative autoregulation at large molecule number, showing respectively: (i) monostable positive feedback with parameters $\rho_u^{(0)} = 150$, $\rho_b^{(0)} = 300$, $\sigma_u = \sigma_b = 10^3$, $d = 1$, $D_1 = 0.25$, $D_2 = 0.25$, $\tau = 0.5$, $\Omega = 100$; (ii) bistable positive feedback with parameters $\rho_u^{(0)} = 24$, $\rho_b^{(0)} = 468$, $\sigma_u = \sigma_b = 10^3$, $d = 1$, $D_1 = 0.75$, $D_2 = 0.1$, $\tau = 1$, $\Omega = 200$; (iii) monostable negative feedback with parameters $\rho_u^{(0)} = 470$, $\rho_b^{(0)} = 20$, $\sigma_u = \sigma_b = 10^3$, $d = 1$, $D_1 = 0.1$, $D_2 = 0.1$, $\tau = 1$, $\Omega = 70$. In the top right hand corner of each plot is the value of ρ for the distribution, defined and discussed later in Section 4.4.4, here showing that for good UCNA performance ρ should be small to satisfy condition 3. SSA data in each case comes from a single steady state trajectory of 9×10^5s.

We now describe the modified SSA that takes into account extrinsic noise on the effective protein production rates. This modification replaces the protein production reaction in each gene state, i.e., $G_k \to G_k + P$ where G_k represents either G or G^*, by three new reactions alongside the introduction of a ghost species Y_k for each gene state. These new reactions are $\varnothing \underset{r_2}{\overset{r_1}{\rightleftharpoons}} Y_k$ and $G_k + Y_k \xrightarrow{r_3} G_k + Y_k + P$. Utilising the LNA (assuming Y_k to be abundant), as was done for colored noise on the degradation rate in Appendix B.1, one finds these rates to be $r_1 = 1/(D_k\Omega)$, $r_2 = 1/\tau$ and $r_3 = r_k^{(0)} D_k \Omega/\tau$, which ensure matching to the colored noise SDE given in Eq. (4.34), where $r_k^{(0)}$ represents $\rho_u^{(0)}$ or $\rho_b^{(0)}$ in G and G^* respectively.

Figure 4.5 shows a good performance of the UCNA when compared to the modified SSA described above. This performance is shown for each differing qualitative behaviour expressed by cooperative bimodality, i.e., (i) monostable positive feedback, (ii) bistable positive feedback, and (iii) monostable negative feedback. In all three plots shown the UCNA matches the modified SSA well, and clearly performs better than if one were to approximate the colored noise with white noise (i.e., $\tau = 0$).

We find the same qualitative behaviour of the creation and eventual destruction of bimodality (see Fig. 4.6A(i–iii)) as the noise strengths, D_1 and D_2, become large for the colored noise on the protein production rates as was found in Fig. 4.3 for colored noise on the degradation rate. Note that for the chosen parameter set in Fig. 4.6A that the white noise approximation performs generally very well compared to the UCNA. For $\tau \leq 1$, the white extrinsic noise approximation can typically perform quite well compared to the modified SSA, but note that this is not always the case especially in situations for which deterministic bistability leads to bimodality of the UCNA solution (see Fig. 4.3).

4.4.3 Fluctuating binding/unbinding rates

Finally, we apply the UCNA to the case of colored noise added to the binding and unbinding rates of the protein to the gene. This could be utilised to implicitly model the effect of multiple gene states in the transition of G to G^*, as has been experimentally and theoretically investigated [198, 170, 199], accounting for DNA looping via distal enhancers or chromatin conformational states. For convenience we define $\sigma_b = \sigma_b^{(0)}(1 + \eta_1(t))$, $\sigma_u = \sigma_u^{(0)}(1 + \eta_2(t))$ and

$$L_\eta = L_0 \left(\frac{1 + \eta_1(t)}{1 + \eta_2(t)} \right), \text{ with } L_0 = \frac{\sigma_u^{(0)}}{\sigma_b^{(0)}}. \tag{4.49}$$

Substituting Eq. (4.49) in the Langevin equation describing the feedback loop Eq. (4.11) (and making a mean-field approximation for the terms under the square root) we obtain the following set of SDEs:

$$\frac{dn}{dt} = \frac{L_\eta \rho_u + \rho_b (n/\Omega)^2}{L_\eta + (n/\Omega)^2} + \sqrt{\frac{\rho_u L_0 + \rho_b (n/\Omega)^2}{L_0 + (n/\Omega)^2} + d\,n}\ \cdot \Gamma(t), \tag{4.50}$$

$$\frac{d\eta_1}{dt} = -\frac{1}{\tau}\eta_1 + \frac{1}{\tau}\theta_1(t), \tag{4.51}$$

$$\frac{d\eta_2}{dt} = -\frac{1}{\tau}\eta_2 + \frac{1}{\tau}\theta_2(t), \tag{4.52}$$

where $\theta_1(t)$ and $\theta_2(t)$ are Gaussian white noise terms with zero mean and correlators $\langle \theta_1(t)\theta_1(t')\rangle = 2D_1\delta(t-t')$ and $\langle \theta_2(t)\theta_2(t')\rangle = 2D_2\delta(t-t')$ respectively. In order to proceed using the UCNA we must linearise the drift term in Eq. (4.50) with respect to η_1 and η_2 through the small noise approximation η_1,$\eta_2 \ll 1$:

$$\begin{aligned}\frac{dn}{dt} \approx \frac{\rho_u L_0 + \rho_b (n/\Omega)^2}{L_0 + (n/\Omega)^2} - d\,n + \left(\frac{L_0 n^2 \Omega^2 (\rho_u - \rho_b)}{(L_0\Omega^2 + n^2)^2}\right)(\eta_1 - \eta_2) \\ + \sqrt{\frac{\rho_u L_0 + \rho_b (n/\Omega)^2}{L_0 + (n/\Omega)^2} + d\,n}\ \cdot \Gamma(t).\end{aligned} \tag{4.53}$$

For convenience we now define:

$$h(n) = \frac{\rho_u L_0 + \rho_b (n/\Omega)^2}{L_0 + (n/\Omega)^2} - d\,n, \tag{4.54}$$

$$g_1(n) = \frac{L_0 n^2 \Omega^2 (\rho_u - \rho_b)}{(L_0\Omega^2 + n^2)^2}, \tag{4.55}$$

$$g_2(n) = \sqrt{\frac{\rho_u L_0 + \rho_b (n/\Omega)^2}{L_0 + (n/\Omega)^2} + d\,n}, \tag{4.56}$$

$$F(n) = g_1(n)\sqrt{D_1 + D_2}, \tag{4.57}$$

$$g(n) = \sqrt{F(n)^2 + g_2(n)^2/2}. \tag{4.58}$$

In terms of these new functions Eq. (4.53) becomes,

$$\frac{dn}{dt} = h(n) + g_1(n)(\eta_1 - \eta_2) + g_2(n)\Gamma(t). \tag{4.59}$$

Following Section 4.4.2 we then arrive at the UCNA for colored noise on the binding rates where η_1,$\eta_2 \ll 1$:

$$\frac{dn}{dt} = \frac{h(n)}{C(n,\tau)} + \frac{1}{C(n,\tau)}\left(F(n)\theta(t) + g_2(n)\Gamma(t)\right). \tag{4.60}$$

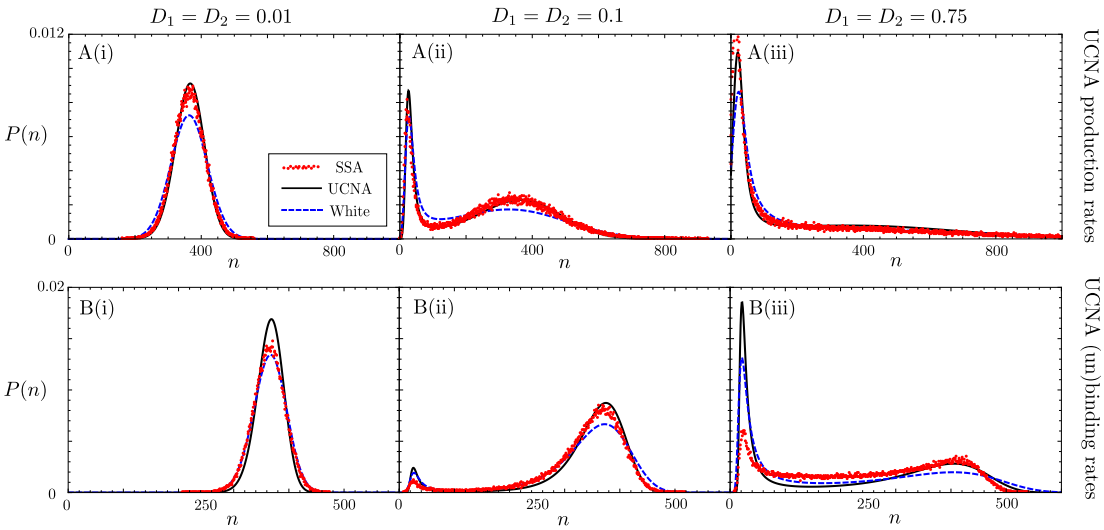


Figure 4.6: Plots showing the creation and eventual destruction of bimodality in the probability distributions for colored noise on the (A) protein production rates, (B) binding/unbinding rates (denoted on the figure by (un)binding rates), analogously to what was observed in Fig. 4.3 for colored noise on the degradation rate. For A it is clear that the UCNA performs well where the noise strength is both small in A(i) and large in A(iii). For B we see that the low (B(i)) and intermediate (B(ii)) noise cases are well predicted by the UCNA and white noise approximation, however where the noise becomes large (B(iii)) the UCNA breaks down, whereas the white noise approximation still performs well compared to the modified SSA prediction. Other than the noise strengths given on the figure, the parameters for both A and B are $\rho_u^{(0)} = 24$, $\rho_b^{(0)} = 468$, $\sigma_u = \sigma_b = 10^3$, $d = 1$, $\tau = 1$, $\Omega = 200$ (i.e., the same parameters used in Fig. 4.3 which express deterministic bistability). SSA data in each case comes from a single steady state trajectory of 9×10^5s.

Then, using the properties of the correlators of $\theta(t)$ and $\Gamma(t)$ we arrive at:

$$\frac{dn}{dt} = \frac{h(n)}{C(n,\tau)} + \frac{g(n)}{C(n,\tau)}\tilde{\Gamma}(t), \tag{4.61}$$

where,

$$C(n,\tau) = 1 + \tau\left(\frac{g_1'(n)h(n)}{g_1(n)} - h'(n)\right), \tag{4.62}$$

and $\tilde{\Gamma}(t)$ is Gaussian white noise with mean zero and correlator $\langle\tilde{\Gamma}(t)\tilde{\Gamma}(t')\rangle = 2\delta(t-t')$. Here, as for the UCNA applied to the degradation rate, $C(n,\tau)$ is again independent of the strengths of the colored noise terms. This UCNA, as we shall see, should be a good approximation where both D_1 and D_2 are small—by 'small' we explicitly mean that D_1 and D_2 should be smaller than noise strengths used on the UCNA for protein production rates or the degradation rate. The solution to Eq. (4.61) is given by:

$$P(n) = \frac{N}{\tilde{g}(n)}\exp\left(\int^n \frac{\tilde{h}(z)}{\tilde{g}(z)^2}dz\right), \tag{4.63}$$

with $\tilde{h}(n) = h(n)/C(n,\tau)$ and $\tilde{g}(n) = g(n)/C(n,\tau)$.

Now we evaluate the performance of the UCNA on the binding and unbinding rates, and compare it with the modified SSA. In this case the modified SSA replaces the binding and unbinding reactions, $G+2P \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*$, by the following: $\varnothing \underset{r_2}{\overset{r_1}{\rightleftharpoons}} Y_1$, $G+Y_1+2P \xrightarrow{r_3} Y_1+G^*$, $\varnothing \underset{r_5}{\overset{r_4}{\rightleftharpoons}} Y_2$, and $G^*+Y_2 \xrightarrow{r_6} G+Y_2+2P$, where Y_1 and Y_2 are ghost species. The rates of these reactions are determined via the LNA (assuming the ghost species to be numerous) and are $r_1 = 1/(D_1\Omega)$, $r_2 = 1/\tau$, $r_3 = \sigma_b^{(0)}D_1\Omega/\tau$, $r_4 = 1/(D_2\Omega)$, $r_5 = 1/\tau$ and $r_6 = \sigma_u^{(0)}D_2\Omega/\tau$.

In Fig. 4.6B we test the UCNA on the binding and unbinding rates compared to the modified SSA described above. Clearly, the same qualitative behaviour of the creation and destruction of bimodality, as noise strength is increased, is observed, as was also observed for colored noise on the degradation rate (Fig. 4.3) and protein production rates (Fig. 4.6A). The resultant expression of bimodality however, is clearly different than for these cases. Notably, this UCNA does ascribe to an additional limitation compared to the UCNA of degradation or production rates; a limitation due to the further small noise approximation made in Eq. (4.53). This limitation is seen in Fig. 4.6B(iii), showing that the UCNA applied to the binding and unbinding rates is much more sensitive to increased noise strength than the other UCNA applications. One also observes that the white noise approximation in Fig. 4.6B performs almost as well as the UCNA (Figs.

4.6B(i–ii)) or in some cases better than the UCNA (Fig. 4.6B(iii)); hence, the white noise approximation may be a safer approximation than the UCNA for colored noise applied to the binding and unbinding rates since it approximates the SSA very well and is less susceptible to the numerical instabilities that can result from the UCNA.

4.4.4 Breakdown conditions of the UCNA

Having now applied the UCNA to approximate distributions for colored noise on the (i) degradation rate, (ii) protein production rates and (iii) the binding/unbinding rates, we now assess the conditions which cause the UCNA to breakdown. The application of the UCNA to colored noise on the protein production rates presents a somewhat more complex problem than the application of the UCNA to colored noise on the degradation rate or the binding/unbinding rates; hence, we more easily see that there are *three main conditions for the breakdown of the UCNA*—conditions beside the large system size or large molecule number requirement needed to approximate the discrete master equation by a one variable FPE], or even the need for τ to be chosen small or large enough such that Eqs. (4.22) and (4.23) are approximately satisfied. Below we detail these three conditions, in each case explaining why the disagreement occurs. Note that although the analysis of breakdown conditions below is done for the UCNA on the protein production rates, the same arguments hold for the other applications of the UCNA previously presented.

Condition 1

The first of these conditions concerns the positivity condition required on $C(n,\tau)$, that is $C(n,\tau) > 0 \,\forall\, n$. We refer to this as *condition 1*. Since we have already discussed this condition in a previous section we will not repeat the discussion here, and refer the reader to Section 4.4.1. In Fig. 4.7A(i) we see a disagreement between the UCNA and the modified SSA for a parameter set that exhibits bimodality, and in Fig. 4.7A(ii) it is verified that this is because $C(n,\tau) < 0$ where $n \approx 100$. Note however, that if $C(n,\tau)$ becomes negative outside of the region containing most of the probability mass that the UCNA can still provide a good approximation to the true modified SSA solution.

Condition 2

The second condition observed for the breakdown of the UCNA concerns the violation of the *characteristic 'length' scale* (the length here being a distance measure in the n space), which we now discuss. In Appendix B.2 we show in more detail why the arguments we present below hold. Based on the noise intensity of the noise term arising from the colored noise in Eq. (4.43), we can introduce the characteristic length scale L

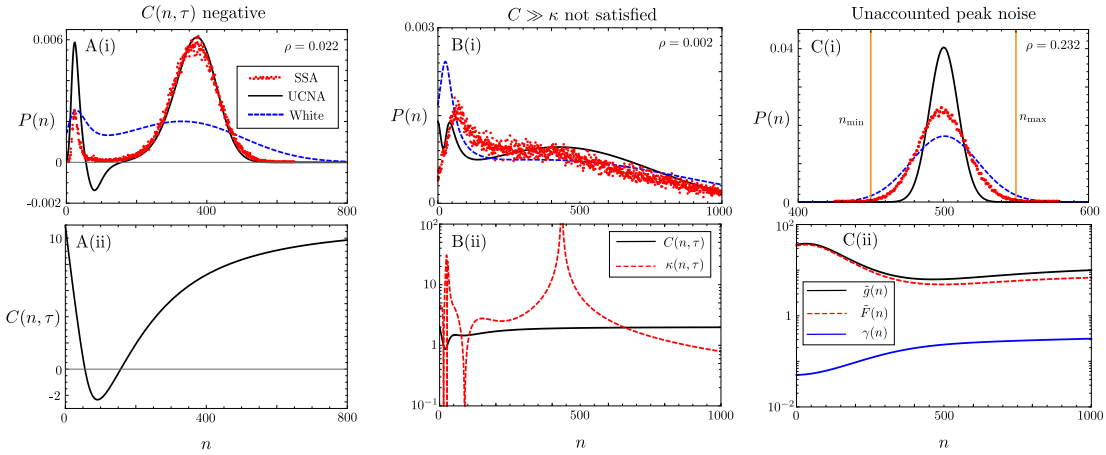


Figure 4.7: This figure shows the disagreement of the UCNA on the protein production rates to the ground truth modified SSA predictions (detailed in Section 4.4.2), also compared to the case of white extrinsic noise ($\tau = 0$). Each disagreement corresponds to a single breakdown condition of the UCNA being violated. The legend in A(i) applies to A(i), B(i) and C(i). Plots in A show the breakdown of the UCNA due to condition 1. A(i) shows the prediction of negative probability due to the negativity of $C(n,\tau)$ in A(ii) around the same value of n. Parameters for A are $\rho_u^{(0)} = 20$, $\rho_b^{(0)} = 470$, $\sigma_u = \sigma_b = 10^3$, $d = 1$, $D_1 = 1$, $D_2 = 0.1$, $\tau = 10$, $\Omega = 200$. Plots in B show the breakdown of the UCNA due to condition 2. B(ii) shows that $\kappa(n,\tau) > C(n,\tau)$ over a large range of n, corresponding to the poor UCNA prediction seen in B(i) over this entire region. Parameters for B are $\rho_u^{(0)} = 50$, $\rho_b^{(0)} = 450$, $\sigma_u = \sigma_b = 10^3$, $d = 1$, $D_1 = 1$, $D_2 = 1$, $\tau = 1$, $\Omega = 100$. Plots in C show the breakdown of the UCNA due to condition 3. C(ii) shows a relatively large value of $\gamma(n)$ over most of the defined region $\mathcal{D}$, and also shows the the pre-factors of the total UCNA noise $\tilde{g}(n)$ and that arising only from the colored noise $\tilde{F}(n) = F(n)/C(n,\tau)$. Vertical orange lines in C(i) indicate the limits of the region $\mathcal{D}$. The value ρ in the top right-hand corner of C(i) can be compared to the smaller values of ρ seen for other parameter sets in A(i) and B(i), indicating that the breakdown observed is truly associated to condition 3. The parameters for C are $\rho_u^{(0)} = 2300$, $\rho_b^{(0)} = 120$, $\sigma_u = \sigma_b = 10^4$, $d = 1$, $D_1 = 0.002$, $D_2 = 0.04$, $\tau = 2$, $\Omega = 230$. SSA data in each case comes from a single steady state trajectory of 9×10^5s.

over which fluctuations in the colored noise term are damped:

$$L(n,\tau) = \frac{F(n)}{C(n,\tau)}, \tag{4.64}$$

noting that the requirement of condition 1 means that this length is always positive. Our approximate one variable FPE in Eq. (4.47) will then be valid under the condition that the drift term varies slowly with respect to L (following Appendix B.2), meaning that one needs to satisfy

$$L\left|\partial_n\left(\tilde{h}(n) + \tilde{g}(n)\tilde{g}'(n)\right)\right| \ll \left|\tilde{h}(n) + \tilde{g}(n)\tilde{g}'(n)\right| \tag{4.65}$$

in order for the UCNA to hold. More succinctly, this condition is:

$$C(n,\tau) \gg \kappa(n,\tau), \tag{4.66}$$

where we henceforth define the function

$$\kappa(n,\tau) = F(n)\left|\frac{\partial_n\left(\tilde{h}(n)+\tilde{g}(n)\tilde{g}'(n)\right)}{\tilde{h}(n)+\tilde{g}(n)\tilde{g}'(n)}\right|. \tag{4.67}$$

We refer to Eq. (4.66) as *condition 2*. In Fig. 4.7B we explore this breakdown for a parameter set that breaks condition 2 over a large region of the parameter space, between $0 < n < 650$. Clearly the UCNA provides a poor approximation in this regime; note however that, similar to condition 1, if condition 2 is violated (i) outside of the domain where most of the probability mass is contained, or (ii) over a small region of the domain containing most of the probability mass, then the UCNA can still provide a good approximation.

Condition 3

The final condition resulting in the breakdown of the UCNA concerns the underestimation of noise. We refer to this as *unaccounted peak noise*, and this forms our final breakdown condition, *condition 3*. The explanation behind condition 3 is that the UCNA in general will always underestimate the Poisson noise for a particular value of n, arising from the necessary neglection of Poisson noise terms in the derivation of the UCNA: (i) neglection of the noise terms under the square root of the Poisson noise pre-factor in Eqs. (4.34) (a form of mean-field approximation), and (ii) neglection of Poisson noise term $g_2(n)\Gamma(t)$ and its time derivative from the $\dot{\eta}$ term in Eqs. (4.19–4.21) via the use of another mean-field approximation. However, the error on the UCNA caused by condition 3 will be small when colored noise dominates the Poisson noise. To investigate the degree to which colored noise is dominant, identifying $F(n)/C(n,\tau)$ from Eq. (4.42) and $g(n)/C|(n,\tau)$ from Eq. (4.46), we define

$$\gamma(n) = \left|\frac{g(n)/C(n,\tau) - F(n)/C(n,\tau)}{g(n)/C(n,\tau)}\right| = \left|\frac{g(n)-F(n)}{g(n)}\right| \tag{4.68}$$

where, for some n, if $\gamma(n) \approx 1$ then Poisson noise dominates, else if $\gamma(n) \approx 0$ then colored noise dominates. Intermediate values of $\gamma(n)$ mean that both Poisson and colored noise is apparent in the system. To investigate whether noise is underestimated generally over the region containing most of the probability, defined as $\mathcal{D} = [n_{\min}, n_{\max}]$, we further define

$$\rho = \frac{1}{|\mathcal{D}|}\int_{n_{\min}}^{n_{\max}} \gamma(n)dn. \tag{4.69}$$

Here, if $\rho \approx 1$ then Poisson noise dominates over the entire region $\mathcal{D}$, else if $\rho \approx 0$ then colored noise dominates over the entire region $\mathcal{D}$. Fig. 4.7C explores this disagreement, where Fig. 4.7C(i) shows the clear underestimation of noise in the UCNA distribution when compared to the modified SSA distribution. Sample values of $n_{\min}$ and $n_{\max}$ are also shown on Fig. 4.7C(i). Fig. 4.7C(ii) shows how the total UCNA noise $\tilde{g}(n)$ varies with respect to the contribution of colored noise $\tilde{F}(n) = F(n)/C(n,\tau)$. Also shown on Fig. 4.7C(ii) is the variation of $\gamma(n)$. Values of ρ are shown in the top right-hand corner for all probability distributions in Fig. 4.5; unlike the other distributions shown in Figs. 4.5 and 4.7, in Fig. 4.7C(i) $\rho \not\approx 0$ does not hold, clarifying that the reason for the UCNA's disagreement for this parameter set is due to condition 3.

Large τ UCNA distributions

Having successfully identified the three main conditions causing the breakdown of the UCNA, we are now able to determine where the UCNA will perform well, *even in the large τ limit*. In Figure 4.8 we explore an example of the UCNA performing exceptionally well for $\tau = 10^2$ (see Fig. 4.8(i)). Clearly the UCNA does not violate any of the three conditions here: (1) $C(n,\tau)$ is not negative in $\mathcal{D}$ (see Fig. 4.8(ii)); (2) $C(n,\tau) \gg \kappa(n,\tau)$ in $\mathcal{D}$ (again, see Fig. 4.8(ii)); (3) $\gamma(n)$ is small for all n in $\mathcal{D}$ (see Fig. 4.8(iii)) as evidenced by the small value of $\rho = 0.001$. As expected, the prediction of white noise on the protein production rates is very poor in the regime where $\tau \gg 1$.

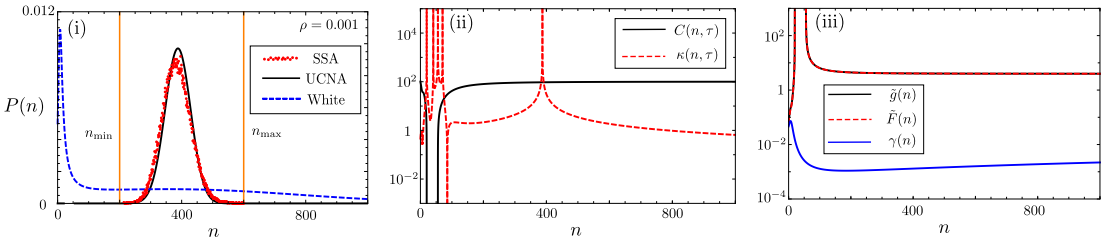


Figure 4.8: Plots showing a good performance of the UCNA for $\tau = 100$. (i) Shows the probability distributions from the modified SSA, UCNA and white noise approximation. Ther vertical orange lines show the limits of the region $\mathcal{D}$ in this case. (ii) Shows that in the region $\mathcal{D}$ that condition 1 is satisfied since $C(n \in \mathcal{D}) > 1$, and condition 2 is satisfied since $C(n \in \mathcal{D}, \tau) \gg \kappa(n \in \mathcal{D}, \tau)$. (iii) Shows that condition 3 is satisifed in $\mathcal{D}$, i.e., $\gamma(n) \approx 0$, since $\tilde{g}(n) \approx \tilde{F}(n)$, which is corroborated by the small value of ρ shown in (i). Other parameters here are $\rho_u^{(0)} = 10$, $\rho_b^{(0)} = 400$, $\sigma_u = \sigma_b = 10^3$, $d = 1$, $D_1 = 0.5$, $D_2 = 1$, $\Omega = 70$. SSA data in each case comes from a single steady state trajectory of 9×10^5s.

4.5 Slow gene switching: the conditional UCNA

In the previous sections we have focused on fast gene switching, whereby a Hill function can then be used to approximate the production of proteins from two different gene states, shown in the reaction scheme of Eq. (4.1). We now consider the case where the switching rates σ_u and σ_b are very small; small enough that the system has two dominant modes of behaviour, one pertaining to each gene state. The approach followed here is very similar to the *conditional linear noise approximation* (cLNA) studied in [161], but instead of approximating the distribution conditional on each gene state as a Gaussian we instead utilise the UCNA in each gene state. We shall refer to this method as the *conditional UCNA* (cUCNA). We begin by stating the law of total probabiity for the marginal distribution of proteins that we are interested in approximating:

$$P(n,t) = \sum_{\underline{G}} P(G_i,t)P(n|G_i,t). \tag{4.70}$$

Here $\underline{G}$ is the set of possible gene state (in our case $\underline{G} = \{G, G^*\}$), $P(G_i,t)$ is the marginal distribution of being in gene state G_i at a time t and $P(n|G_i,t)$ is the conditional probability of having n proteins at a time t given that the system is in state G_i. Our task now is to find suitable approximations for $P(G_i,t)$ and $P(n|G_i,t)$ that allow us then to construct an approximation of the full steady state distribution in Eq. (4.70). In our case we have two different gene states, G and G^*, and hence we can construct the reaction schemes conditional on each gene state. The reaction scheme conditional on gene state G is (i) $G \xrightarrow{\rho_u} G + P$, $P \xrightarrow{d} \varnothing$, and the reaction scheme conditional on gene state G^* is (ii) $G^* \xrightarrow{\rho_b} G^* + P$, $P \xrightarrow{d} \varnothing$. This then allows us to approximately find the steady state mean number of proteins conditional on each gene state when σ_u and σ_b are very small (where the subscript a denotes approximate): $\langle n|G\rangle_a = \rho_u/d$ and $\langle n|G^*\rangle_a = \rho_b/d$. We can use these conditional means to find the marginal probabilities of being in a specific gene state at steady state. Note that in this calculation we will ignore the influence of noise on the rate parameters; the inherent assumption is that extrinsic noise does not much influence the probability of being in each gene state. First we write an approximate master equation for the transitions between differing gene states:

$$\frac{d}{dt}P(G,t) \approx \sigma_u P(G^*,t) - \frac{\sigma_b \langle n|G\rangle_a^2}{\Omega^2} P(G,t). \tag{4.71}$$

We can then solve the above equation at steady state (denoted by the subscript s) by utilising conservation of probability, $P_s(G) = 1 - P_s(G^*)$, giving:

$$P_s(G^*) = \left(1 + \frac{\sigma_u}{\sigma_b}\left(\frac{d\cdot\Omega}{\rho_u}\right)^2\right)^{-1}, \tag{4.72}$$

$$P_s(G) = \left(1 + \frac{\sigma_b}{\sigma_u}\left(\frac{\rho_u}{d\cdot\Omega}\right)^2\right)^{-1}. \tag{4.73}$$

Since now we have the $P_s(G_i)$ needed for Eq. (4.70) we need to find the $P_s(n|G_i)$ terms. Here we show how to calculate these terms for noise on the degradation rate, although this can be easily extended to the case where we have noise on the protein production rates. In each gene state, the system we are concerned to study is $G_i \xrightarrow{r_i} G_i + P$, $P \xrightarrow{d_i} \varnothing$, where G_i, d_i and r_i represent either gene state G or G^*, the corresponding gene state dependent decay rate, and production rate ρ_u or ρ_b respectively. Adding colored noise to the degradation rate $d_i = d_0(1+\eta_i)$, where d_i is the degradation rate in gene state G_i given colored noise η_i, we then have the following set of SDEs in each gene state (here we have applied the mean-field approximation to the terms in the square root):

$$\frac{dn}{dt} = r_i - d_0\, n - (d_0\, n)\eta_i + \sqrt{r_i + d_0 n}\cdot\Gamma(t), \tag{4.74}$$

$$\frac{d\eta_i}{dt} = -\frac{1}{\tau_i}\eta_i + \frac{1}{\tau_i}\theta_i(t), \tag{4.75}$$

where $\Gamma(t)$ and $\theta_i(t)$ are Gaussian white noise terms, each with zero mean and correlators $\langle\Gamma(t)\Gamma(t')\rangle = \delta(t-t')$ and $\langle\theta_i(t)\theta_i(t')\rangle = 2D_i\delta(t-t')$ respectively. Processing the usual steps of the UCNA method, detailed explicitly in Section 4.4, we find the approximate steady state probability for each gene state:

$$P_s(n|G_i) \approx N\exp\left(u(n,r_i)\right) n^{2r_i\tau_i-1}(r_i + d_0\, n(2d_0D_in+1))^{-\frac{1}{2}-\frac{1}{2d_0D_i}-r_i\tau_i}(n + r_i\tau_i), \tag{4.76}$$

where N is a normalisation constant and we have defined,

$$u(n,r_i) = \frac{\left(D_i\left(4 - 6d_0\tau_i\right)r_i + 1\right)\tan^{-1}\left(\frac{4d_0D_in+1}{\sqrt{8D_ir_i-1}}\right)}{d_0D_i\sqrt{8D_ir_i-1}}. \tag{4.77}$$

Hence, using Eqs. (4.72)–(4.73) and (4.76) we can now approximate Eq. (4.70) as:

$$P(n) \approx P_s(G)P_s(n|G) + P_s(G^*)P_s(n|G^*). \tag{4.78}$$

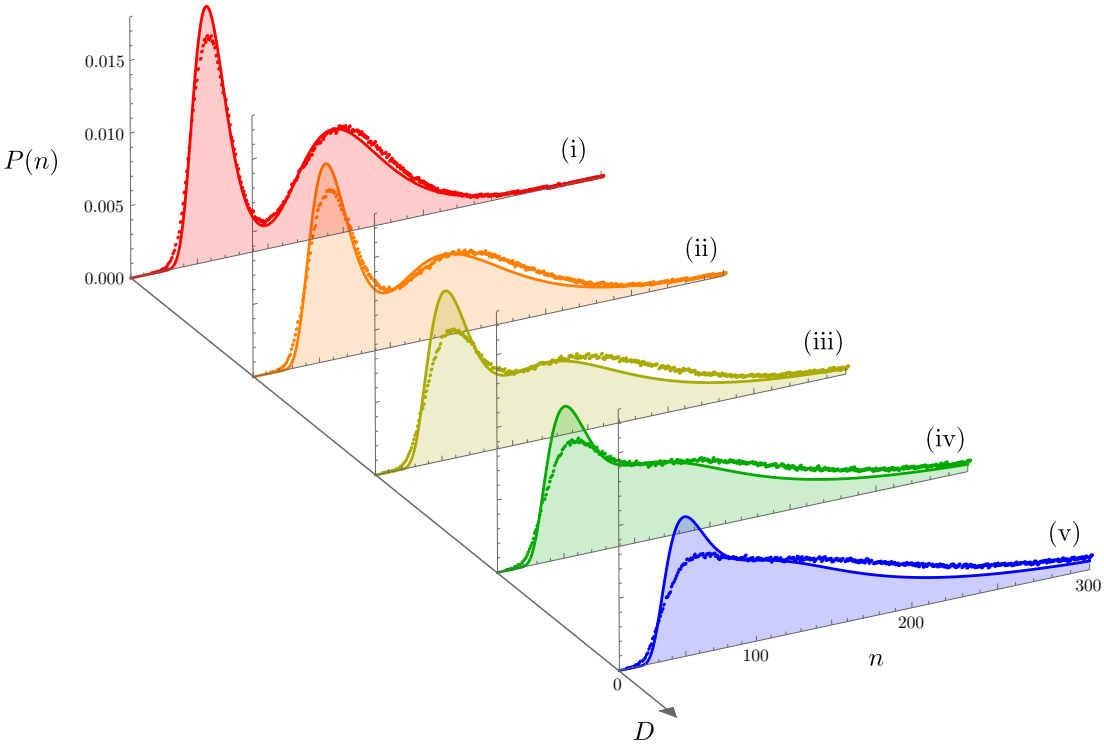


Figure 4.9: Comparison of the cUCNA with the modified SSA for increasing values of the colored noise strength for both gene states, D. It is seen that the cUCNA (solid lines) is a good approximation to the true distribution (dots, simulated using the modified SSA described in Appendix B.1), especially for small values of D. The noise strengths for each plot are (i) $D = 0.1$, (ii) $D = 0.2$, (iii) $D = 0.3$, (iv) $D = 0.4$, (v) $D = 0.5$, and the shared parameters are $\rho_u = 30$, $\rho_b = 75$, $\sigma_b = 0.01$, $\sigma_u = 0.001$, $d_0 = 0.5$ and $\tau = 1$. Clearly as D gets larger the bimodality exhibited by the slow switching between the gene states is destroyed by the extrinsic noise added to the degradation rate.

Figure 4.9 compares the cUCNA with the modified SSA—which is the same as the modified SSA found in Section 4.4.1. Fig. 4.9(i) shows that for small switching rates, the cUCNA can correctly capture the bimodality exhibited where the colored noise on the degradation rate is small. As the noise on the degradation rate gets larger the cUCNA still provides a decent approximation to the true distributions; it is also clear that the bimodality of the protein distribution is destroyed as the size of this noise increases. One can contrast this to the cases observed in Figs. 4.3 and 4.6 which showed that where the gene switching rates are fast, increased colored noise strength can in fact induce bimodality. *In summary, we find that extrinsic noise on the degradation rate of a slow switching auto-regulatory system generally destroys bimodality, but for fast switching it is common to observe the opposite phenomenon.*

4.6 Applications

In this section we explicitly show, by means of two examples, how one can use the colored noise formulation that was introduced earlier to describe intricate molecular details of cooperative autoregulation. We first show this for multi-stage protein production with fast gene switching, and then for multi-stage protein degradation with slow gene switching.

4.6.1 Multi-stage protein production

The first example of using colored noise as a form of model reduction is that of mapping multistage protein production onto a simpler system, where colored noise accounts for processes not explicitly considered in the simpler model. Consider multi-stage protein production on the cooperative auto-regulatory feedback loop:

$$\begin{aligned} &G \xrightarrow{\rho_u} G + M_1,\ G^* \xrightarrow{\rho_b} G^* + M_1, \\ &M_i \xrightarrow{\Lambda_i} M_{i+1},\ i \in [1, N-1],\ M_N \xrightarrow{\Lambda_N} \varnothing,\ M_N \xrightarrow{r_1} M_N + P, \\ &G + 2P \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*,\ P \xrightarrow{d} \varnothing, \end{aligned} \tag{4.79}$$

where it is assumed the system contains only one gene copy, either in state G or in state G^*. The simpler model that we will then map this system onto the cooperative auto-regulatory feedback loop:

$$\begin{aligned} &G \xrightarrow{\rho_u} G + P,\ G^* \xrightarrow{\rho_b} G^* + P, \\ &G + 2P \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*,\ P \xrightarrow{d} \varnothing, \end{aligned} \tag{4.80}$$

where $\rho_u = \rho_u^{(0)}(1+\eta_1(t))$ and $\rho_b = \rho_b^{(0)}(1+\eta_2(t))$, and assigning the properties of colored noises $\eta_1(t)$ and $\eta_2(t)$ such that Eq. (4.79) can be mapped onto Eq. (4.80) is the task we have assigned ourselves. One can think of the different M_i for $i < N$ as the various stages of nascent mRNA, before it is eventually fully transcribed in stage M_N (mature mRNA) where it can then begin translation [200, 201, 202]. Utilising the slow scale linear noise approximation [84] one can show that if $\Lambda_i \gg \max\{\Lambda_N, \rho_u, \rho_b\}$ for $i \in [1, N-1]$ then the nascent mRNA $M_1, ..., M_{N-1}$ are fast species, and the reaction system in Eq. (4.79) is consistent with the following reaction scheme describing fluctuations in the slow species G, G^*, M_N and P:

$$\begin{aligned} &G \xrightarrow{\rho_u} G + M_N,\ G^* \xrightarrow{\rho_b} G^* + M_N, \\ &M_N \xrightarrow{\Lambda_N} \varnothing,\ M_N \xrightarrow{r_1} M_N + P, \\ &G + 2P \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*,\ P \xrightarrow{d} \varnothing. \end{aligned} \tag{4.81}$$

We now apply the van Kampen ansatz to the number of mature mRNA, M_N. In gene state G this gives us $n_1(t) = \Omega\phi_1 + \Omega^{1/2}\epsilon_1(t)$, and in gene state G^* this gives us $n_2(t) = \Omega\phi_2 + \Omega^{1/2}\epsilon_2(t)$, where $\phi_1 = \rho_u/(\Lambda_N\Omega)$ and $\phi_2 = \rho_b/(\Lambda_N\Omega)$ are the steady state solutions to the rate equation describing the mature mRNA in the gene states G and G^* respectively, and $\epsilon_1(t)$ and $\epsilon_2(t)$ describe small fluctuations about these means. Note the occurrence of $1/\Omega$ in ϕ_1 and ϕ_2 follows since the concentration of a single gene in a volume Ω is $1/\Omega$. Using these ansatzes allows us to construct the effective protein production rates in gene states G and G^* respectively:

$$\rho_u = r_1 n_1(t) = \frac{r_1\rho_u}{\Lambda_N}\left(1 + \Omega^{1/2}\frac{\Lambda_N}{\rho_u}\epsilon_1(t)\right), \tag{4.82}$$

$$\rho_b = r_1 n_2(t) = \frac{r_1\rho_b}{\Lambda_N}\left(1 + \Omega^{1/2}\frac{\Lambda_N}{\rho_b}\epsilon_2(t)\right). \tag{4.83}$$

One can then see that $\rho_u^{(0)} = r_1\rho_u/\Lambda_N$, $\rho_b^{(0)} = r_1\rho_b/\Lambda_N$ and that the noise terms have the form:

$$\eta_1(t) = \Omega^{1/2}\frac{\Lambda_N}{\rho_u}\epsilon_1(t), \tag{4.84}$$

$$\eta_2(t) = \Omega^{1/2}\frac{\Lambda_N}{\rho_b}\epsilon_2(t). \tag{4.85}$$

In order to fully specify $\eta_1(t)$ and $\eta_2(t)$ we need to find the correlators $\langle\eta_1(t)\eta_1(t')\rangle$ and $\langle\eta_2(t)\eta_2(t')\rangle$, which can be done by application of the linear noise approximation (LNA) [8]. Note that since we are already restricted to the large system size, large molecule number regime following the FPE approximation to the CME (discussed in Section 4.4.1), we can apply the LNA without further restricting the validity of the final solution. The same can also be said for the use of the LNA in Section 4.6.2. Applying the LNA to $n_1(t)$ and $n_2(t)$, whose fluctuations are fully specified by the reactions $G \xrightarrow{\rho_u} G + M_N$, $G^* \xrightarrow{\rho_b} G^* + M_N$ and $M_N \xrightarrow{\Lambda_N} \varnothing$, gives us the two following one variable FPEs:

$$\frac{\partial\Pi(\epsilon_1,t)}{\partial t} = \Lambda_N\frac{\partial}{\partial\epsilon_1}\left(\epsilon_1\Pi(\epsilon_1,t)\right) + \frac{1}{2}\left(\frac{2\rho_u}{\Omega}\right)\frac{\partial^2\Pi(\epsilon_1,t)}{\partial\epsilon_1^2}, \tag{4.86}$$

$$\frac{\partial\Pi(\epsilon_2,t)}{\partial t} = \Lambda_N\frac{\partial}{\partial\epsilon_2}\left(\epsilon_2\Pi(\epsilon_2,t)\right) + \frac{1}{2}\left(\frac{2\rho_b}{\Omega}\right)\frac{\partial^2\Pi(\epsilon_2,t)}{\partial\epsilon_2^2}, \tag{4.87}$$

where $\Pi(\epsilon_i, t)$ is the probability of having a fluctuation of size ϵ_i at a time t. These FPEs, combined with Eq. (4.84) and (4.85), admit equivalent Langevin equations for $\eta_1(t)$ and $\eta_2(t)$, given by:

$$\frac{d\eta_1(t)}{dt} = \Lambda_N \left(-\eta_1(t) + \sqrt{\frac{2}{\rho_u}}\beta_1(t) \right), \tag{4.88}$$

$$\frac{d\eta_2(t)}{dt} = \Lambda_N \left(-\eta_2(t) + \sqrt{\frac{2}{\rho_b}}\beta_2(t) \right), \tag{4.89}$$

where $\beta_1(t)$ and $\beta_2(t)$ are independent Gaussian white noises with zero mean and correlator $\langle \beta_1(t)\beta_1(t')\rangle = \langle \beta_2(t)\beta_2(t')\rangle = \delta(t-t')$. From here one can find the correlators of $\eta_1(t)$ and $\eta_2(t)$:

$$\langle \eta_1(t)\eta_1(t')\rangle = \frac{\Lambda_N}{\rho_u} \exp\left(-\Lambda_N |t-t'|\right), \tag{4.90}$$

$$\langle \eta_2(t)\eta_2(t')\rangle = \frac{\Lambda_N}{\rho_b} \exp\left(-\Lambda_N |t-t'|\right). \tag{4.91}$$

Comparing to the results of Section 4.4.2 it is clear that $\eta_1(t)$ and $\eta_2(t)$ satisfy the definition of colored noise, with noise strengths $D_1 = 1/\rho_u$, $D_2 = 1/\rho_b$ and shared correlation time $\tau = 1/\Lambda_N$. This completes the mapping between the full complex system in Eq. (4.79) and our reduced process in Eq. (4.80). We can hence utilise our solution for the probability distribution with colored noise on the effective protein production rates in Eq. (4.48). Note that the colored noise in this case can model transcriptional bursting, namely the production of proteins in bursts due to rapid translation from short lived mRNA [203, 45]; bursty expression has been previously modelled in the literature by an effective first-order reaction with constant rate parameter but with the special property that when the reaction fires, the number of proteins produced is sampled from a geometric distribution [204, 1].

In Fig. 4.10A we show how effective the UCNA can be in approximating the protein distribution from the full system described in Eq. (4.79), where we have for simplicity assumed that there are three mRNA states: M_1, M_2 and M_3 (i.e., $N = 3$). Fig. 4.10A(i) shows the approximation for a parameter set exhibiting bimodality: the red points represent the *standard SSA* of the full system in Eq. (4.79); the black line represents the distribution predicted from the UCNA (i.e., using Eq. (4.48) with $D_1 = 1/\rho_u$, $D_2 = 1/\rho_b$ and $\tau = 1/\Lambda_N$); the blue dotted line represents the distribution if one put white noise of the same magnitude on the protein production rates (i.e., the UCNA at $\tau = 0$); and the orange line with circles shows the distribution if one was to neglect noise on the reaction rates entirely (i.e., $\rho_u = \rho_u^{(0)}$ and $\rho_b = \rho_b^{(0)}$). Clearly, in Fig. 4.10A(i) the UCNA is the only distribution that fits the SSA prediction, showing both the effectiveness of our model reduction as well as the need to properly account for the correlation time of

colored noise in model reduction. This makes sense, since one would expect processes occurring in the full system to be correlated over short times, *i.e., that noise events in close temporal proximity are not independent*, and one cannot simply neglect these effects. Fig. 4.10A(ii) instead shows the various approximations for a monomodal parameter set. In this case, white noise is a poor approximation, and it is clear that one cannot neglect the finite correlation time. However, it is interesting to note that properly accounting for the correlation time using the UCNA returns the same distribution as if one had not added noise to the production rates at all—this is due to the small magnitudes of D_1/τ and D_2/τ respectively. These two examples shows that where correlation time is finite, it is imperative that one models it correctly.

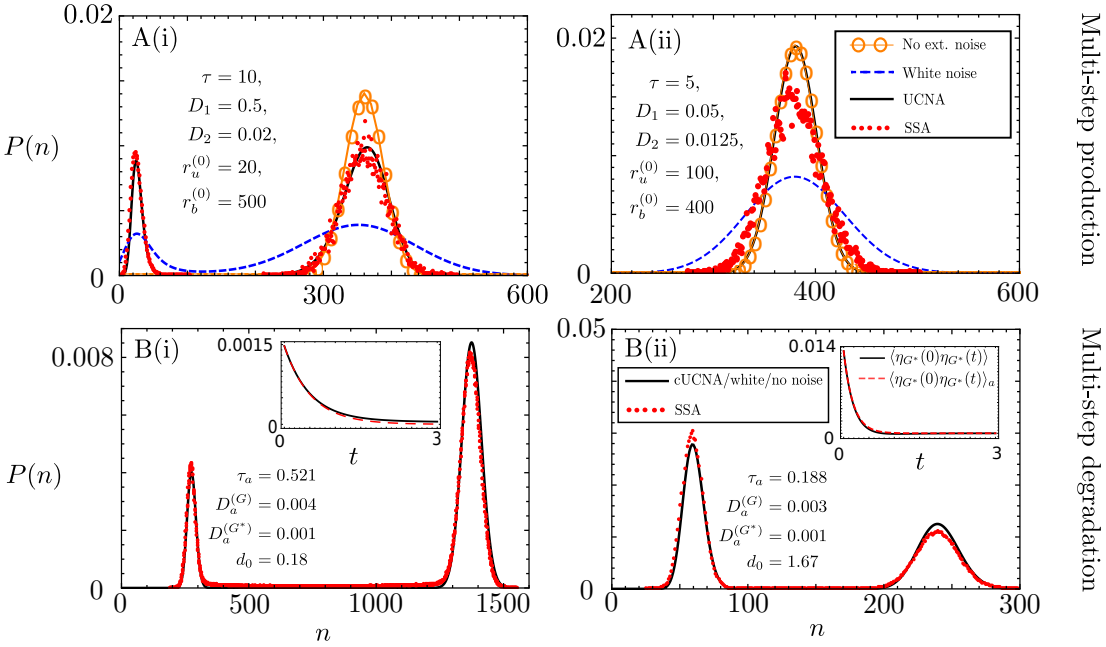


Figure 4.10: (A) Shows distributions of the *standard SSA* of the reaction scheme in Eq. (4.79) for multistage protein production in three intermediate species M_1, M_2 and M_3, against (1) the UCNA, (2) white noise (i.e., $\tau = 0$) and (3) no extrinsic colored noise (i.e., the solution given by Eq. (4.10)). Each plot shows the colored noise parameters used for the UCNA solution from Eq. (4.48), which are determined from the full multi-stage protein production process in Eq. (4.79). Note the legend in A(ii) applies only to distributions in A(i–ii). Parameter values for the standard SSA in A(i) are $\rho_u = 2$, $\rho_b = 50$, $\Lambda_1 = 1000$, $\Lambda_2 = 1000$, $\Lambda_3 = 0.1$, $r_1 = 1$, $\sigma_u = \sigma_b = 1000$, $d = 1$ and $\Omega = 230$. Parameter values for the SSA in A(ii) are $\rho_u = 20$, $\rho_b = 80$, $\Lambda_1 = 1000$, $\Lambda_2 = 1000$, $\Lambda_3 = 0.2$, $r_1 = 1$, $\sigma_u = \sigma_b = 1000$, $d = 1$ and $\Omega = 100$. (B) Shows distributions of the *standard SSA*, with protein decay following the multi-step process of Eq. (4.92), against the cUCNA. The colored noise parameters used for the cUCNA solution of Eq. (4.118) are shown on each plot, with these values being determined from the full model using Eqs. (4.99,4.116,4.117). The insets show a comparison of the approximate correlator (see Eq. (4.115)) and the full double exponential correlator (see Eq. (4.114)) for gene state G^*. Note the legend in B(ii) applies only to distributions in B(i–ii), and the legend on the inset of B(ii) applies also to the inset in B(i). Parameter values for the SSA in B(i) are $\rho_u = 50$, $\rho_b = 250$, $\sigma_b = 2.5 \times 10^{-3}$, $\sigma_u = 10^{-3}$, $\Omega = 200$, $d_1 = 1$, $k = 1$ and $d_2 = 0.1$. Parameter values for the SSA in B(ii) are $\rho_u = 100$, $\rho_b = 400$, $\sigma_b = 10^{-3}$, $\sigma_u = 10^{-4}$, $\Omega = 200$, $d_1 = 1$, $k = 1$ and $d_2 = 5$. SSA data for A(i) and A(ii) come from a single steady state trajectories of length 10^8s and 9×10^5s respectively. Note that A(i) presents a very long relaxation to the steady state due to the systems inertia in staying in one of the two modes of the distribution. SSA data for B(i) and B(ii) come from a single steady state trajectory of 9×10^6s.

4.6.2 Multi-stage protein degradation

Proteins in cells are often degraded via multi-step processes. For example, a major degradation pathway in eukaryotic cells is the ubiquitin-proteasome degradation pathway [205], and more recent experiments have shown that a subset of proteins in the mammalian proteome have age-dependent degradation rates [206, 207]. From Fig. 2 in [206] we consider a system with two different stages of protein with differing degradation rates combined with the cooperative auto-regulatory feedback loop:

$$\begin{aligned} &G \xrightarrow{\rho_u} G + P_1,\ G^*_{1,2} \xrightarrow{\rho_b} G^*_{1,2} + P_1, \\ &P_1 \xrightarrow{\kappa} P_2,\ P_1 \xrightarrow{d_1} \varnothing,\ P_2 \xrightarrow{d_2} \varnothing, \\ &G + 2P_1 \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*_1,\ G + 2P_2 \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*_2, \end{aligned} \tag{4.92}$$

where $G^*_{1,2}$ indicates *either* the state G^*_1 or G^*_2. This reaction system models age dependent protein states, since the protein P_1 is always produced from the gene, and eventually undergoes a transition to protein state P_2, where P_1 and P_2 have differing degradation rates. We will show how to map this system to the reduced system:

$$G + 2P \underset{\sigma_u}{\overset{\sigma_b}{\rightleftharpoons}} G^*,\ G \xrightarrow{\rho_u} G + P,\ G^* \xrightarrow{\rho_u} G^* + P,\ P \xrightarrow{d} \varnothing, \tag{4.93}$$

where the total number of P is given as the sum of the number of P_1 and P_2, i.e., $n = n_1 + n_2$, G^* is simply the sum of G^*_1 and G^*_2, and $d = d_0(1 + \eta(t))$, where $\eta(t)$ is colored noise. Our task is to find the properties of the noise $\eta(t)$ such that one can map the full system in Eq. (4.92) onto the reduced system in Eq. (4.93). Note that although we here look at two different stages of protein, the analysis presented below can be easily extended for several different stages of protein degradation.

One finds that the effective degradation rate of the sum of protein number P_1 and P_2 is:

$$d = \frac{n_1 d_1 + n_2 d_2}{n_1 + n_2}. \tag{4.94}$$

In the following analysis we consider gene switching to be slow, which allows us to apply the cUCNA from Section 4.5. We first consider the probability of being in each gene state at steady state $P_s(G_k)$, where G_k represents either gene state G or G^*. Note that we assume both protein stages P_1 and P_2 can bind and unbind to the gene at the same respective rates, and note that $\langle n|G\rangle = \langle n_1|G\rangle + \langle n_2|G\rangle$. Following the analysis from

Section 4.5 in Eqs. (4.72–4.73) we find:

$$P_s(G^*) = \left(1 + \frac{\sigma_u \Omega^2}{\sigma_b}\left(\frac{d_2(\kappa + d_1)}{\rho_u(\kappa + d_2)}\right)^2\right)^{-1}, \tag{4.95}$$

$$P_s(G) = \left(1 + \frac{\sigma_b}{\sigma_u \Omega^2}\left(\frac{\rho_u(\kappa + d_2)}{d_2(\kappa + d_1)}\right)^2\right)^{-1}. \tag{4.96}$$

We can now proceed to find the probability distribution conditional on each gene state $P_s(n_1, n_2)$. In gene state G_k the conditional reaction system is:

$$G_k \xrightarrow{r_k} G_k + P_1,\ P_1 \xrightarrow{d_1} \varnothing,\ P_1 \xrightarrow{\kappa} P_2,\ P_2 \xrightarrow{d_2} \varnothing, \tag{4.97}$$

where the protein is always produced in stage P_1, and $r_k \equiv \rho_u$ in gene state G, and $r_k \equiv \rho_b$ in gene state G^*. Now we employ the van Kampen ansatz [8] on n_1 and n_2 in gene state G_k, i.e., $n_1^{(k)}(t) = \Omega\phi_1^{*(k)} + \Omega^{1/2}\epsilon_1^{(k)}(t)$ and $n_2^{(k)}(t) = \Omega\phi_2^{*(k)} + \Omega^{1/2}\epsilon_2^{(k)}(t)$, where $\phi_1^{*(k)}$ and $\phi_2^{*(k)}$ are the deterministic steady state mean concentrations of P_1 and P_2 in gene state G_k respectively, and $\epsilon_1^{(k)}(t)$ and $\epsilon_2^{(k)}(t)$ are fluctuations about these mean values. In the following we drop the superscript (k) notation for aesthetic reasons, although one should keep in mind that the process below must be individually conducted on each gene state. The purpose of using the van Kampen ansatz can be seen upon its substitution into Eq. (4.94) which for a large system size, Ω, gives:

$$d = \frac{d_1\phi_1^* + d_2\phi_2^*}{\phi_1^* + \phi_2^*}\Bigg(1 + \Omega^{-1/2}\Big(\epsilon_1(t)\Big(\frac{d_1}{d_1\phi_1^* + d_2\phi_2^*} - \frac{1}{\phi_1^* + \phi_2^*}\Big) + \epsilon_2(t)\Big(\frac{d_2}{d_1\phi_1^* + d_2\phi_2^*} - \frac{1}{\phi_1^* + \phi_2^*}\Big)\Big)\Bigg) + \mathcal{O}(\Omega^{-1}). \tag{4.98}$$

By comparing to the effective degradation from the reduced model in gene state G_k, $d = d_0(1 + \eta_k(t))$, one can see that to match the two models we must have

$$d_0 = (d_1\phi_1^* + d_2\phi_2^*)/(\phi_1^* + \phi_2^*), \tag{4.99}$$

and

$$\eta_k(t) = \Omega^{-1/2}\left(\epsilon_1(t)y_1(d_1, d_2, \phi_1^*, \phi_2^*) + \epsilon_2(t)y_2(d_1, d_2, \phi_1^*, \phi_2^*)\right), \tag{4.100}$$

where we have defined the functions,

$$y_1(d_1, d_2, \phi_1^*, \phi_2^*) = \frac{d_1}{d_1\phi_1^* + d_2\phi_2^*} - \frac{1}{\phi_1^* + \phi_2^*}, \tag{4.101}$$

$$y_2(d_1, d_2, \phi_1^*, \phi_2^*) = \frac{d_2}{d_1\phi_1^* + d_2\phi_2^*} - \frac{1}{\phi_1^* + \phi_2^*}. \tag{4.102}$$

If the correlators $\langle\epsilon_1(0)\epsilon_1(t)\rangle$, $\langle\epsilon_2(0)\epsilon_2(t)\rangle$, $\langle\epsilon_1(0)\epsilon_2(t)\rangle$ and $\langle\epsilon_2(0)\epsilon_1(t)\rangle$ are known, then one can also find the correlator of $\eta_k(t)$, i.e., $\langle\eta_k(0)\eta_k(t)\rangle$, given by:

$$\langle\eta_k(0)\eta_k(t)\rangle = \frac{1}{\Omega}\left(y_1^2\langle\epsilon_1(0)\epsilon_1(t)\rangle + y_2^2\langle\epsilon_2(0)\epsilon_2(t)\rangle + y_1y_2(\langle\epsilon_1(0)\epsilon_2(t)\rangle + \langle\epsilon_2(0)\epsilon_1(t)\rangle)\right). \tag{4.103}$$

Note in Eq. (4.100) that if $d_1 = d_2 = d$, then the magnitude of $\eta_k(t)$ is zero for all t since the system $P_1 \xrightarrow{\kappa} P_2$, $P_1 \xrightarrow{d} \varnothing$, $P_2 \xrightarrow{d} \varnothing$ is equivalent to $P \xrightarrow{d} \varnothing$ where one is only interested in the total number of proteins.

To proceed in finding $\eta_k(t)$ in Eq. (4.100), we first need to find the steady state concentrations ϕ_1^* and ϕ_2^* from the deterministic rate equations. These are,

$$\frac{d\phi_1}{dt} = \frac{r_k}{\Omega} - (\kappa + d_1)\phi_1, \tag{4.104}$$

$$\frac{d\phi_2}{dt} = \kappa\phi_1 - d_2\phi_2, \tag{4.105}$$

where again the $1/\Omega$ in Eq. (4.104) follows since the concentration of a single gene in a volume Ω is $1/\Omega$. Enforcing the steady state condition allows us to find ϕ_1^* and ϕ_2^*,

$$\phi_1^* = \frac{r_k}{\Omega(\kappa + d_1)},\ \phi_2^* = \frac{\kappa r_k}{d_2(\kappa + d_1)}.$$

Note that the linear dependence of ϕ_1^* and ϕ_2^* on r_k means that the effective degradation rate d_0 from Eq. (4.99) is independent of the gene state. Assuming that both P_1 and P_2 are numerous, we now proceed to the LNA [8, 87] of the system in Eq. (4.97), which will allow us to find the correlators $\langle\epsilon_1(0)\epsilon_1(t)\rangle$, $\langle\epsilon_2(0)\epsilon_2(t)\rangle$, $\langle\epsilon_1(0)\epsilon_2(t)\rangle$ and $\langle\epsilon_2(0)\epsilon_1(t)\rangle$. Where $\mathbf{S}$ is the stoichiometric matrix, $\underline{\phi} = (\phi_1, \phi_2)$ and $\underline{f}(\underline{\phi})$ is the macroscopic rate vector one can computationally find the required matrices needed to perform the LNA: (i) the steady state Jacobian matrix $A_{ij} = d(\mathbf{S}\cdot\underline{f}(\underline{\phi}))_j/d\phi_i|_{\underline{\phi}=\underline{\phi}^*}$, and (ii) the steady state diffusion matrix $(\mathbf{B}\cdot\mathbf{B}^T)_{ij} = \mathbf{S}\cdot\mathrm{Diag}(\underline{f}(\underline{\phi}))\cdot\mathbf{S}^T|_{\underline{\phi}=\underline{\phi}^*}$. The Jacobian matrix then allows us to find the time evolution of both $\langle\epsilon_1(t)\rangle$ and $\langle\epsilon_2(t)\rangle$ since $\partial_t\langle\underline{\epsilon}\rangle = \mathbf{A}\cdot\langle\underline{\epsilon}\rangle$, where $\langle\underline{\epsilon}\rangle = (\langle\epsilon_1(t)\rangle, \langle\epsilon_2(t)\rangle)$. Solving these coupled first order ODEs gives us:

$$\langle\epsilon_1(t)\rangle = \langle\epsilon_1(0)\rangle e^{-(d_1+\kappa)t}, \tag{4.106}$$

$$\langle\epsilon_2(t)\rangle = \frac{-\kappa\langle\epsilon_1(0)\rangle e^{-(d_1+\kappa)t} + (\kappa\langle\epsilon_1(0)\rangle + (d_1 - d_2 + \kappa)\langle\epsilon_2(0)\rangle)e^{-d_2t}}{d_1 - d_2 + \kappa}, \tag{4.107}$$

where $-d_2$ and $-(d_1+\kappa)$ are eigenvalues of $\mathbf{A}$. Clearly, in the limit $t \to \infty$ the fluctuations about the steady state concentrations ϕ_1^* and ϕ_2^*, $\langle\epsilon_1(t)\rangle$ and $\langle\epsilon_2(t)\rangle$, tend to zero as required. The final step of the LNA then requires us to find the covariance matrix $\mathbf{C}$ at steady state, which has the steady state variances $\langle\epsilon_1^2\rangle$ and $\langle\epsilon_2^2\rangle$ as diagonal components and covariance $\langle\epsilon_1\epsilon_2\rangle = \langle\epsilon_2\epsilon_1\rangle$ in the off-diagonal components. $\mathbf{C}$ is then given by the

Lyapunov equation [87]:

$$\mathbf{A}\cdot\mathbf{C}+\mathbf{C}\cdot\mathbf{A}^T+\mathbf{B}\cdot\mathbf{B}^T=0, \tag{4.108}$$

whose solution is given by:

$$\mathbf{C}=\begin{pmatrix}\frac{r_k}{(d_1+\kappa)\Omega} & 0\\ 0 & \frac{\kappa r_k}{d_2(d_1+\kappa)\Omega}\end{pmatrix}. \tag{4.109}$$

From van Kampen [8] p. 259 we assert that for some fluctuation ϵ_i, $\langle\epsilon_i(0)\epsilon_j(t)\rangle = \langle\epsilon_i(0)\langle\epsilon_j(t)\rangle\rangle$, and that at $t=0$ we have $\underline{\phi}=\underline{\phi}^*$ so that $\langle\epsilon_i(0)\epsilon_j(0)\rangle=\langle\epsilon_i\epsilon_j\rangle$. For example, for $\langle\epsilon_1(0)\epsilon_1(t)\rangle$ we have, using $\langle\epsilon_1(t)\rangle$ from Eq. (4.106) and $\langle\epsilon_1^2\rangle$ from Eq. (4.109), $\langle\epsilon_1(0)\epsilon_1(t)\rangle=\langle\epsilon_1(0)\langle\epsilon_1(t)\rangle\rangle=\langle\epsilon_1^2\rangle e^{-(d_1+\kappa)t}$. Explicitly, one can then calculate all the correlators, which are given by:

$$\langle\epsilon_1(0)\epsilon_1(t)\rangle=\frac{r_k}{(d_1+\kappa)\Omega}e^{-(d_1+\kappa)t}, \tag{4.110}$$

$$\langle\epsilon_2(0)\epsilon_2(t)\rangle=\frac{\kappa r_k}{d_2(d_1+\kappa)\Omega}e^{-d_2t}, \tag{4.111}$$

$$\langle\epsilon_2(0)\epsilon_1(t)\rangle=0, \tag{4.112}$$

$$\langle\epsilon_1(0)\epsilon_2(t)\rangle=\frac{\kappa r_k(e^{-d_2t}-e^{-(d_1+\kappa)t})}{(d_1+\kappa)(d_1-d_2+\kappa)\Omega}. \tag{4.113}$$

Now that these correlators have been determined, we can substitute them into Eq. (4.103) giving us the following for the correlator of $\eta_k(t)$:

$$\langle\eta_k(0)\eta_k(t)\rangle=\frac{(d_1-d_2)^2\kappa\left(\kappa(d_1+\kappa)e^{-(d_1+\kappa)t}+(d_1-d_2)d_2e^{-d_2t}\right)}{(d_1+\kappa)(d_1-d_2+\kappa)(d_2+\kappa)^2r_k}, \tag{4.114}$$

noting the only dependence on the gene state G_k comes from the pre-factor $1/r_k$. Comparing this equation to the colored noise seen in Eq. (4.12) in Section 4.4.1 we see however that we have two exponentials in the correlator. This sum of exponentials in Eq. (4.114) can be approximated by a single exponential through a small t expansion. This gives us:

$$\langle\eta_k(0)\eta_k(t)\rangle\approx\langle\eta_k(0)\eta_k(t)\rangle_a=\frac{D_a^{(k)}}{\tau_a}e^{-t/\tau_a}, \tag{4.115}$$

where $D_a^{(k)}$ and τ_a are the approximate noise strength and correlation time given by:

$$D_a^{(k)} = \frac{(d_1 - d_2)^2 \kappa}{(d_1 + \kappa)(d_1\kappa + d_2(d_2 + \kappa) + \kappa^2) r_k}, \tag{4.116}$$

$$\tau_a = \frac{d_2 + \kappa}{d_1\kappa + d_2(d_2 + \kappa) + \kappa^2}. \tag{4.117}$$

Clearly, the small t expansion allows us to roughly interpret the noise $\eta_k(t)$, present in gene state G_k, as colored noise with strength D_a/τ_a and correlation time τ_a. Note that even when both exponentials equally contribute to the correlator in Eq. (4.114), this is generally a very good approximation for few protein stages so long as Eq. (4.115) is also a good approximation. Knowing $D_a^{(k)}$ and τ_a for $\eta_k(t)$ we can now substitute them into Eqs. (4.76–4.77) in Section 4.5, then using Eqs. (4.95–4.96) we find

$$P(n) \approx P_s(G)P_s(n|G) + P_s(G^*)P_s(n|G^*). \tag{4.118}$$

Fig. 4.10B shows two different cases of the cUCNA predicting distributions for multi-stage degradation and slow gene switching: in B(i) for the case of $d_1 > d_2$ (true for around 80% of proteins in [206]); in B(ii) for the case of $d_2 > d_1$ (true for around 20% of proteins in [206]). On the main plots red dots show the standard SSA prediction of the full reaction scheme in Eq. (4.92), and the black lines show the cUCNA from Eq. (4.118), which in both cases is almost indistinguishable from the white noise (cUCNA with $\tau = 0$) and no external colored noise predictions (discussed further in the following paragraph). The insets show the correlators in gene state G^*, where the red dashed line represents $\langle \eta_{G^*}(0)\eta_{G^*}(t)\rangle_a$ and the black line shows $\langle \eta_{G^*}(0)\eta_{G^*}(t)\rangle$. Note that the correlators for gene state G are not shown because they show very similar to what is seen for state G^*. Even given the complex model reduction from two protein species to one effective protein species the cUCNA performs very well in predicting distributions from the standard SSA of the full system in Eq. (4.92). Note that as one considers more protein stages with differing degradation rates it becomes more different to fit the correlator to a single exponential, which presents a limitation of this method for more protein stages.

However, we find that since our analysis is restricted to the large protein number regime, and the noise strength D_a is inversely proportional to the production rate r_k which is typically large, that D_a is typically very small in both gene states. This means that the cUCNA probability distribution is almost identical to probability distributions that assume white noise (cUCNA with $\tau = 0$) or even no colored noise. However, what the analysis in this section provides is the *quantitative reasons why one could necessarily neglect the contribution of colored noise in model reduction from the full system in Eq. (4.92) to the simpler system in Eq. (4.93).*

4.7 Conclusion

In this chapter we have explored the addition of colored noise onto the reaction rates for a cooperative auto-regulatory circuit. Starting from a reduced chemical Fokker-Planck description, we used the UCNA to derive approximate expressions for the probability distribution of protein numbers in the limits of fast and slow promoter switching. The approximation is valid provided the colored noise on the reaction rates is small and the correlation time is short. By means of stochastic simulations, we verified the accuracy of the approximate distributions; we also verified the predictions of the UCNA, namely that under fast promoter switching conditions the addition of colored noise can induce bimodality whereas under slow promoter switching conditions, noise can destroy bimodality.

We also have shown how complex models of gene expression can be mapped onto simpler models with noisy rates. In particular we have shown that: (i) An auto-regulatory feedback loop with multi-stage protein production, including different stages of mRNA processing, can be mapped onto an auto-regulatory feedback loop with a single protein production reaction step having colored noised added to its reaction rate. (ii) A feedback loop with multi-stage protein degradation can be mapped onto a feedback loop with a single protein degradation reaction with a fluctuating rate. We have also verified that in many instances, one cannot simply approximate colored noise with white noise, or else neglect it entirely, since this does not match behaviour seen from the full underlying models of multi-stage protein production or degradation.

While here we focused on a self-regulatory example, the UCNA and its conditional variant (cUCNA) provide an easily extendable analysis to model more complex gene regulatory networks with fluctuating parameters such as those with cross-regulation [208, 209, 37]. Our analysis is the first to our knowledge, to analytically find steady state probability distributions where colored noise is added to a non-linear reaction (the protein-gene binding reaction) in a gene regulatory context; a previous study applied the UCNA to study the effects of extrinsic noise in genetic circuits composed of purely linear reactions [184]. Given that our calculations show that the protein distributions for auto-regulatory circuits with extrinsic noise on reaction parameters can be dramatically different than models assuming constant reaction rates, an interesting future research direction would be to develop UCNA based methods that can directly infer the properties of colored noise on reaction rates from protein expression data.

Chapter 5

Distinguishing between models of mammalian gene expression: telegraph-like models versus mechanistic models

This chapter has been published as [3] entitled *Distinguishing between models of mammalian gene expression: telegraph-like models versus mechanistic models* in the *Journal of The Royal Society Interface*. Slight modifications have been made for its inclusion in this thesis.

5.1 Abstract

Two-state models (telegraph-like models) have a successful history of predicting distributions of cellular and nascent mRNA numbers that can well fit experimental data. These models exclude key rate limiting steps, and hence it is unclear why they are able to accurately predict the number distributions. To answer this question, here we compare these models to a novel stochastic mechanistic model of transcription in mammalian cells that presents a unified description of transcriptional factor, polymerase and mature mRNA dynamics. We show that there is a large region of parameter space where the first, second and third moments of the distributions of the waiting times between two consecutively produced transcripts (nascent or mature) of two-state and mechanistic models exactly match. In this region, (i) one can uniquely express the two-state model parameters in terms of those of the mechanistic model, (ii) the models are practically indistinguishable by comparison of their transcript numbers distributions, and (iii) they are distinguishable from the shape of their waiting time distributions. Our results clarify the relationship between different gene expression models and identify a means to select between them from experimental data.

5.2 Introduction

One of the most popular models of gene expression is the *telegraph model*, a two-state model where genes are assumed to be either *on* or *off*, being able to produce mature messenger RNA (mRNA) in the on state and having no mature mRNA production in the off state [35, 41, 210]. Since gene expression is inherently stochastic [9], mathematical models of the telegraph model often employ probabilistic modelling techniques such as the chemical master equation [8, 81] or the stochastic simulation algorithm (SSA) [68]. By fitting the steady-state analytical solution of the telegraph model to experimentally measured distributions of the number of cellular mRNA in single cells, several studies have estimated the rates of gene switching and of initiation for several mammalian genes [44, 45, 46, 47, 37]. However, mapping cellular mRNA number to the underlying transcription kinetics is difficult because fluctuations in this number reflect noise due to many processes downstream of transcription [211, 120].

In contrast, the number of actively transcribing RNA polymerase II (Pol II) on a gene is not subject to these complex processes, and hence reveals more information on the details of transcription [212, 213, 214]. Therefore, unlike mature mRNA statistics, fluctuations of actively transcribing Pol II provide a direct readout of transcription. Since the speed of actively transcribing Pol II is approximately constant along a gene and since its premature detachment is not frequent, it follows that the loss of actively transcribing Pol II (leading to the production of a mature mRNA transcript) cannot be described by a first-order reaction (as assumed in the telegraph model for cellular mRNA). Rather it is much better captured by a delayed degradation reaction where the removal of an actively transcribing Pol II occurs after a fixed elapsed time since its binding to the promoter. A recent paper [120] has modified the telegraph model to account for the aforementioned speciality, a model that we shall refer to as the *delayed telegraph model*. This alternative two-state model, unlike the telegraph model, is non-Markovian; while its mathematical analysis is complex, it can be solved exactly in steady-state to obtain distributions of the number of bound Pol II. Transcriptional parameters can then be obtained by fitting these distributions to those obtained experimentally using electron microscopy [215] or nascent RNA sequencing [216]. Alternatively, since each actively transcribing Pol II has attached to it an incomplete nascent mRNA, one can also use the delay telegraph model to numerically calculate the steady-state distribution of nascent mRNA numbers which can then be fit to distributions obtained using single-molecule fluorescence *in situ* hybridisation (smFISH) [217].

Despite their success in predicting distributions of transcript numbers that match those calculated from experimental data, it is important to remember that both the telegraph model and the delayed telegraph model do not include a description of all the key rate limiting steps. In the past decade, several experimental papers have shown the necessity

of including Pol II pausing and release in models of transcription. Bartman *et al.* [218] show experimentally that it is the release of Pol II from the pausing state, and not the Pol II recruitment rate, that is a key control point for gene expression. In fact, it is found universally amongst all metazoan genes that the rate of release of Pol II from pausing is the rate limiting step in transcription [219]. In mammalian cells the release of Pol II from the paused state is dependent on the activity of several molecules, including the transcription elongation factor P-TEFb [219, 220, 221]. Specifically in embryonic stem cells, ChIP-Seq data has revealed that Pol II peaks near genes are at the promoter-proximal region, and that inhibiting the P-TEFb causes Pol II to remain in the promoter-proximal region genome-wide [221]. Figures 1 and 2 in [219] provide a good overview of the key step of transcription, including Pol II pausing and release. The mechanism of Pol II pausing, in addition to the binding of Pol II and other transcription factors to the promoter, provides two layers of control over the production of nascent and mature mRNA. It is also found that expressed genes without a peak of paused Pol II in one cell type can acquire pausing in a different cell type, therefore genes have the potential of being regulated by proximal pausing even when the Pol II pausing peak is absent [220]. Clearly, if Pol II pausing and release is such a key feature of transcriptional models, the current ambiguity of the mechanisms' roles in the standard and delayed telegraph models is a problem in need of a solution.

Thus far, the modelling literature contains few studies where transcription is modelled incorporating Pol II pausing and release. One model, found in [170], includes pausing and release in a three-state gene model based on the findings of [218], where the three states represent (i) an inactive gene state D_0, (ii) a "burst initiated" state D_{10} where the gene is bound to transcription factors and enhancers, and (iii) a gene state D_{11} in which the Pol II is bound and paused. Mature mRNA is produced in the transition from $D_{11} \to D_{10}$; this reaction should actually produce nascent mRNA but in this model, it is assumed that the nascent lifetime is so short that a nascent mRNA description can be ignored. By ignoring nascent mRNA fluctuations and assuming that the pausing and unpausing of the Pol II is very fast, it was shown in [170] that the mature mRNA distribution from this model is well approximated by that from the telegraph model. Two other recent studies [222, 223] also explore similar models albeit in the context of transcription reinitiation [224].

In summary, it is currently not so clear why the telegraph model is so successful in fitting experimental mature mRNA distributions, even though it misses important reaction steps which are key control points for gene expression. It is unclear if the assumptions made in [170] are necessary to guarantee that the true mature mRNA distribution is well approximated by the telegraph model; it could well be that these are sufficient but not necessary conditions. Since this study did not derive nascent mRNA statistics, nothing

can be inferred about the reasons underlying the success of the delayed telegraph model in fitting experimental nascent mRNA distributions. A related and important question still remains: if the two-state and more detailed mechanistic models of transcription cannot be distinguished from distributions of the number of transcripts, is there another statistic that is useful to distinguish between them? In this study we take a first step at answering these questions.

The chapter is divided as follows. In Section 5.3 we introduce the standard and delayed telegraph models (two-state models), as well as a mechanistic multi-state gene model that provides a stochastic description of transcription factor, Pol II and mature mRNA dynamics. Then, in Section 5.5 we explore the relationship between the two-state and mechanistic models by comparing the distributions of their waiting times between two consecutive transcripts. We show that two-state models can always be told apart from the mechanistic model from the shape of the waiting time distribution, even when they are indistinguishable from a comparison of their number distributions. We also derive conditions under which the moments of the waiting time distribution (up to third order) of the mechanistic model agree with those of two-state models, leading to expressions relating the parameters of the latter with those of the former. In Section 5.6 we perform a sensitivity analysis using the aforementioned expressions to understand which parameters of the mechanistic model are the parameters of two-state model most sensitive to. This uncovers non-trivial correlations between the parameters of two-state models. In Section 5.7 we show that the conclusions previously based on waiting time distributions agree with those obtained using model reduction methods based on number statistics. Finally, in Section 5.8 we conclude the study and discuss our results in the context of the literature.

5.3 Models of transcription

In this section, we start by introducing an effective reaction scheme for a mechanistic model of transcription describing activator, Pol II, and mature mRNA dynamics. Then, we introduce the standard and delayed telegraph models as the two-state models whose dynamics we will attempt to match to that of mature mRNA and actively transcribing Pol II in the mechanistic model, respectively.

5.3.1 A non-Markovian mechanistic model of transcription

The mechanistic model of transcription in metazoan cells that we henceforth consider is defined in terms of the following effective reactions:

$$U \underset{a'}{\overset{a}{\rightleftharpoons}} U^{\star} \underset{b'}{\overset{b}{\rightleftharpoons}} U^{\star\star}, \quad U^{\star\star} \xrightarrow{c} U^{\star} + A, \quad A \underset{\tau}{\Rightarrow} M \xrightarrow{d} \varnothing. \tag{5.1}$$

State U describes a gene state in which Pol II cannot access the promoter region at the beginning of a gene since activator binding is impaired by chromatin [225, 226]. In contrast, state $U^{\star}$ describes a state where activator binding has reorganised the local nucleosome structure [226], allowing Pol II to access the promoter region along with all transcription factors, co-activators, unphosphorylated Pol II and initiation factors needed for transcription initiation to start. This state is coincident with the dynamic promoter condensate (or transcription factories) proposed in various papers [227, 228, 229]. Transcription factors recruit cofactors and Pol II, and hence drive the (reversible) change of state from U to $U^{\star}$.

Initiation starts with the binding of Pol II to the promoter; it then pauses promoter-proximally [230]. These processes are modelled by the change of state from $U^{\star}$ to $U^{\star\star}$, where the latter is the paused state. Once the pause is released, Pol II begins moving away from the promoter region, thus starting productive elongation that leads to a Pol II molecule with a nascent mRNA tail (even paused Pol II has a tail but it is very short and we will hence ignore it). Note that the nascent transcript is not a fully formed mRNA transcript since the length of the tail attached to Pol II increases as elongation progresses. The number of Pol II bound to the gene is equal to the number of nascent mRNA irrespective of their lengths. We call any Pol II undergoing productive elongation as an active Pol II (A), which implies that the change of state from paused $U^{\star\star}$ to unpaused $U^{\star}$ must simultaneously lead to the production of an A particle. Note that the binding of another Pol II to the promoter is not possible when there is already a Pol II paused promoter-proximally due to volume exclusion imposed by the latter [231].

Elongation (and termination) finishes after a fixed elapsed time τ leading to the detachment of Pol II from the gene and the dissociation of the mRNA tail from Pol II. We hence call the fully formed mRNA a mature transcript M and elongation is described by the effective reaction $A \Rightarrow M$ (the double horizontal arrow is here used to denote delayed degradation which occurs after a fixed time τ). Note that the change from A to M cannot be modelled by a first-order reaction because elongation involves the movement of Pol II along the gene with an approximately constant velocity and hence the lifetime of an active Pol II molecule is not exponentially distributed [120, 201]. Note

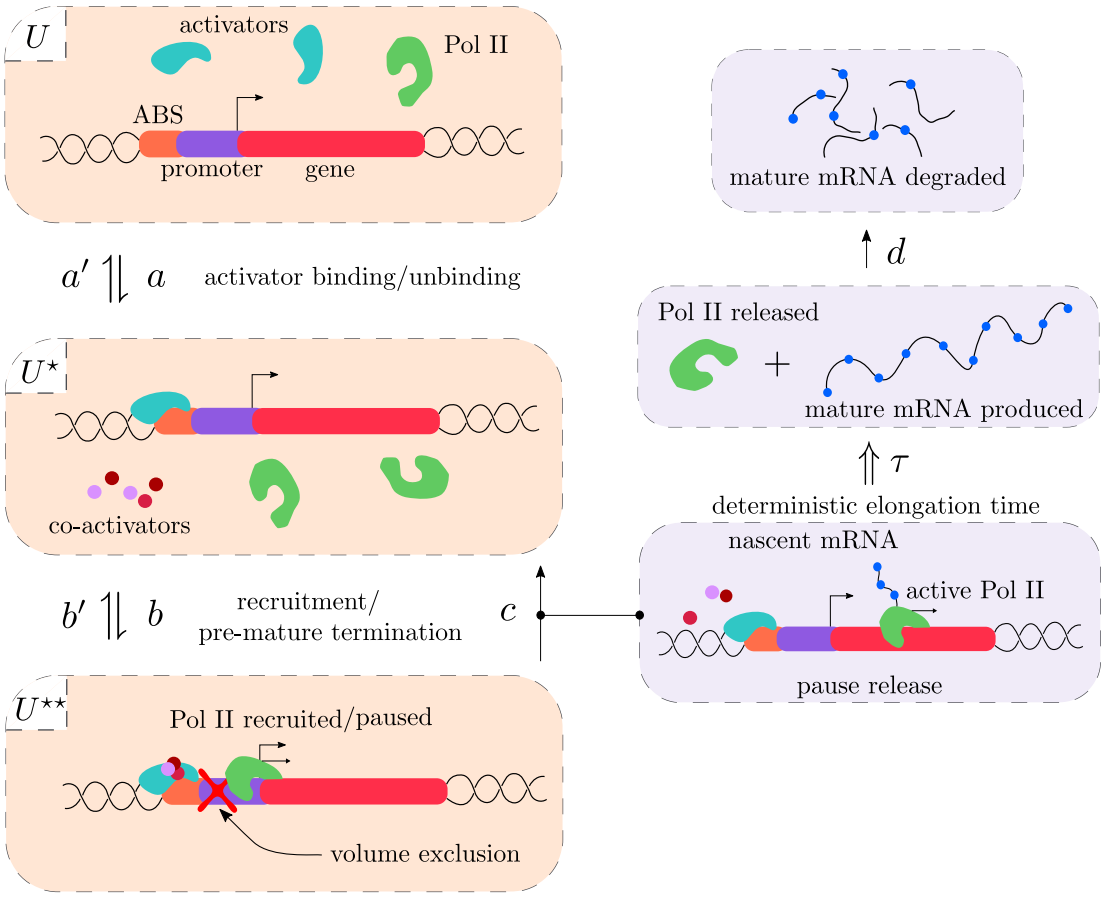


Figure 5.1: Illustration of system (1). The U state describes the state where both the activator binding site (ABS) and the promoter are unbound. In the U^* state, the activator is bound to the ABS meaning the Pol II can bind to the promoter. Pol II has been recruited to the promoter and pauses in state U^{**}. Transitions from U^{**} to U^* either result from premature termination or else pause release and the subsequent production of an actively transcribing Pol II. Elongation (and termination) takes a deterministic time τ after which the mature mRNA is produced. The latter is subsequently degraded in the cytoplasm. For more details see the main text.

that likely there are fluctuations in the elongation time (the lifetime of an active Pol II molecule) but we will ignore them since (i) we could not find single-cell measurements of the distribution of the elongation time for a given gene; (ii) theory suggests these fluctuations are very small for long genes with low transcription rates [201].

Note that paused Pol II instead of leading to productive elongation can also undergo premature termination [219] i.e., the paused Pol II releases the short nascent mRNA tail attached to it (which is rapidly degraded) and the polymerase is recycled into the free Pol II pool. These reactions may happen quite often [232, 233]; it is thus quite unlikely that they simultaneously lead to a dissociation of the dynamic promoter condensate since otherwise the efficiency of gene expression would become extremely low. Hence we assume that premature termination leads to a change from the paused state $U^{\star\star}$ to the unpaused state $U^{\star}$ but do not consider transitions of the type $U^{\star\star}$ to the non-permissive/inactive state U.

Finally, the mature transcripts are removed via various mRNA decay pathways in the cytoplasm [234, 235]. Since many mammalian genes follow single-exponential decay kinetics [236], we model mature mRNA turnover via a first-order reaction of the form $M \to \varnothing$.

We emphasise that a speciality of our model is the reaction $U^{\star\star} \to U^{\star} + A$. This involves a change of gene state each time a transcription event occurs, whereas common models of gene expression in the literature do not have such a coupling [35, 210, 120, 22, 237, 238]. As explained above, the change of state is necessary to model the fine-scale details of the molecular biology, namely the fact that unpausing a Pol II frees the promoter and enables the binding of a new Pol II to it. Unpausing of Pol II is a key rate limiting step since the mapping of Pol II using chromatin immunoprecipitation (ChIP) revealed peaks of Pol II near many promoters [239, 240, 221]. In fact, this accumulation of Pol II near the promoters indicates that the relative rates of premature termination (b') and pause release (c) are much slower than rates of recruitment and entry into the paused state (b). Since regulatory processes often target rate-limiting steps, the release of paused Pol II has emerged as a central point of gene expression control [219, 218].

We note that while the mechanistic model described incorporates more biological detail than the standard two-state models, nevertheless it is to be kept in mind that it is still based on some assumptions because the actual mechanisms of pausing and how variable it is between species is still an ongoing discussion in the experimental community. For instance Pol II volume exclusion may not be enough to avoid the immediate recruitment of other polymerases.

A master equation can be written down which describes the stochastic dynamics of the mechanistic model; its form is quite different than the standard chemical master equation that is popular in the literature of gene expression [81]. The right hand side of the latter equation is only a function of the present time t. In contrast, the master equation describing our model has a right-hand side that is a function of not only the present time t but of the history of the process in the interval $[t-\tau, t]$. This is because of the fixed time τ between the release from the paused state and the production of a mature transcript. The dependence of the dynamics of the system on its history means that our model is non-Markovian [73].

5.3.2 Two-state models of transcription: telegraph and delay telegraph models

In the literature, two models are commonly used to (separately) describe active Pol II and mature mRNA dynamics:

$$G \underset{\sigma_b}{\overset{\sigma_u}{\rightleftharpoons}} G^\star, \quad G \xrightarrow{\rho} G + A, \quad A \underset{\tau}{\Rightarrow} \varnothing, \tag{5.2}$$

and

$$G \underset{\sigma_b}{\overset{\sigma_u}{\rightleftharpoons}} G^\star, \quad G \xrightarrow{\rho} G + M, \quad M \xrightarrow{d} \varnothing. \tag{5.3}$$

The chemical master equation describing the stochastic dynamics of the systems defined by schemes (5.2) and (5.3) were exactly solved in steady-state by Xu *et al.* [120] and Peccoud and Ycart [35], respectively. Model (5.3) also has a transient solution which is reported in [41]. Model (5.3) is often called the telegraph model of gene expression; by analogy, we shall refer to Model (5.2) as the delayed telegraph model. Note that the former is a Markov model while the latter is non-Markov in character for the same reason as described above for the mechanistic model.

Clearly, the difference between the two models is how the transcripts are removed from the system: active Pol II is removed after a fixed elapsed time τ (due to the termination of elongation which results in a mature transcript), whereas mature mRNA degradation follows first-order kinetics. Both models postulate that at any point in time, the gene is in one of two states: an active state G from which transcription can occur and an inactive state G^*. As argued by Bartman *et al.* [218] it is unclear what is the precise biological meaning that should be associated with these two states because the reaction in these models cannot be clearly associated with polymerase processes that are central to transcription.

As we argued in the Introduction, both telegraph and delayed telegraph models have been shown to accurately replicate experimental distributions of mature and nascent mRNA numbers. This leads us to the following question: could it be possible that the stochastic dynamics of our mechanistic model defined by (5.1) are accurately approximated by these simpler models? To be more precise, is there a set of effective transcriptional parameters ρ, σ_u, σ_b of the two-state models that predict the same (or very similar) distributions of active Pol II and mature mRNA numbers in the mechanistic model. If there is such a set of effective parameters, ideally we would also want expressions showing their relationship to the parameters a, a', b, b', c of the mechanistic model.

5.4 Exact solutions of the mechanistic model

In this Section we show how to solve the dCME describing the mechanistic model in Eq. (5.1). We note that the solution to the delay telegraph model (in Eq. (5.2)) has been provided in [241], and that the solution to standard telegraph model (in Eq. (5.3)) is a special case of *Example 2* explored in Section 2.2 of this thesis.

5.4.1 Marginal steady state solution for M

We begin by solving for the marginal steady state generating function of M. We do this by identifying M with a simpler reaction network that removes the presence of A,

$$U \underset{a'}{\overset{a}{\rightleftharpoons}} U^{\star} \underset{b'}{\overset{b}{\rightleftharpoons}} U^{\star\star}, \quad U^{\star\star} \xrightarrow{c} U^{\star} + M, \quad M \xrightarrow{d} \varnothing. \tag{5.4}$$

Note that identification of the dynamics of M with this reaction scheme applies *only at steady state*, and results from the key property that A is elongated deterministically. The reason that this holds at steady state is that since A is a deterministic intermediate, the dynamics of M in Eq. (5.1) are simply a time delayed version of the steady state dynamics observed in Eq. (5.4). The CME describing this reaction scheme can be solved using standard generating function methods introduced in the preliminaries. In this case the generating function $G(z) = \sum_n P(n) z^n$, where $P(n)$ is the steady state distribution of n, is given by,

$$G(z) = {}_0F_2\left(;\{\lambda_1 - 1, \lambda_2 - 1\}; \frac{\beta_3}{d^3}(z-1)\right), \tag{5.5}$$

where ${}_0F_2$ is a type of hypergeometric function [242], and where we have defined,

$$\begin{aligned}
\beta_1 &= 3d + a + a' + b + b' + c, \\
\beta_2 &= d(\beta_1 - 2d) + a'(b + c) + a(b + b' + c) + bc, \\
\beta_3 &= abc,
\end{aligned}$$

and λ_1 and λ_2 are the solutions to the following simultaneous equations,

$$\lambda_1 + \lambda_2 + 1 = \beta_1/d,$$
$$\lambda_1\lambda_2 = \beta_2/d^2.$$

The probability distribution is then given by,

$$P(n) = \frac{1}{n!}\frac{d^n G(z)}{dz^n}\Big|_{z\to 0} \tag{5.6}$$
$$= \frac{\left(\beta_3/d^3\right)^n}{n!(\lambda_1-1)_n(\lambda_2-1)_n}\cdot {}_0F_2\left(;\{\lambda_1+(n-1),\lambda_2+(n-1)\};-\frac{\beta_3}{d^3}\right), \tag{5.7}$$

and $(a)_n$ is a falling factorial.

5.4.2 Marginal steady state solution for A

In this Section we derive and solve the dCME describing active Pol II dynamics at steady state. The dynamics of A is described by the following reaction scheme,

$$U \underset{a'}{\overset{a}{\rightleftharpoons}} U^*, \quad U^* \underset{b'}{\overset{b}{\rightleftharpoons}} U^{**}, \quad U^{**} \xrightarrow{c} U^* + A, \quad A \underset{\tau}{\Longrightarrow} M. \tag{5.8}$$

In the following we utilise the notation that $P_0(n,t)$, $P_1(n,t)$ and $P_2(n,t)$ are the respective probabilities of being in gene states U, U^* and U^{**} with partial state vector (n,t). Note that the only non-Markovian reaction here is the delay τ accounting for transcription elongation in which nascent mRNA is processed and becomes mature mRNA.

Derivation of the dCME

We now derive the dCME from first principles. One can split the contributions to $P_i(n,t+\Delta t)$ into three parts: (a) probability of an instantaneous reaction occurring in the interval $[t,t+\Delta t)$; (b) the probability of a delayed reaction occurring in $[t,t+\Delta t)$; and (c) the probability that no reaction occurs in $[t,t+\Delta t)$. Contributions (a) and (c) are easy to calculate, and one can do so as they would for the standard CME. Contribution (b) however is more difficult and requires one to consider the history of the process.

A careful consideration of the effect of an A production event at $t-\tau+\Delta t$, and it's effect on $P_i(n,t+\Delta t)$ involves following the state vector transitions from $t-\tau$ to $t+\Delta t$:

1. First we have: $(2,n',t-\tau)\to(1,n'+1,t-\tau+\Delta t)$.
2. Then: $(1,n'+1,t-\tau+\Delta t)\to(i,n+1,t)$.
3. Finally: $(i,n+1,t)\to(i,n,t+\Delta t)$.

The first of these contributions occurs with probability $d \cdot \Delta t \cdot P_2(n', t-\tau)$, the second occurs with the conditional probability $P_i(n+1, t|1, n'+1, t-\tau+\Delta t)$, and the third occurs with probability 1 (since the reaction is *deterministically* delayed). One can simplify the conditional probability $P_i(n+1, t|1, n'+1, t-\tau+\Delta t)$ by noting that *all* n' active Pol IIs present prior to $t-\tau+\Delta t$ will have already been processed to M by time t; hence, $P_i(n+1, t|1, n'+1, t-\tau+\Delta t)$ is equivalent to the probability of producing n active Pol IIs in a time $\tau - \Delta t$. We denote this probability as:

$$\tilde{P}_i(n, \tau - \Delta t) = P_i(n+1, t|1, n'+1, t-\tau+\Delta t),$$

with the initial conditions $\tilde{P}_1(n,0) = \delta_{n,0}, \tilde{P}_0(n,0) = \tilde{P}_2(n,0) = 0$. Taking the product of these three contributions to (b) together, we find that the probability of having $(n+1, t)$ *and* $(n, t+\Delta t)$ is:

$$\begin{aligned} P_i(n, t+\Delta t; n+1, t) &= d \cdot \Delta t \cdot \tilde{P}_i(n, \tau - \Delta t) \sum_{n'} P_2(n', t-\tau), \\ &= d \cdot \Delta t \cdot \tilde{P}_i(n, \tau - \Delta t) P_2(t-\tau), \end{aligned} \tag{5.9}$$

where $P_2(t-\tau)$ is the total probability of being in gene state U^{**} at $t-\tau$. Finally, taking the contributions from (a), (b) and (c) we state the probability conservation equation for each $P_i(n, t+\Delta t)$,

$$\begin{aligned} P_0(n, t+\Delta t) =& (1 - a\Delta t)P_0(n,t) + a'\Delta t P_1(n,t) \\ &+ d \cdot \Delta t \cdot P_2(t-\tau)(\tilde{P}_0(n, \tau - \Delta t) - \tilde{P}_0(n-1, \tau - \Delta t)), \qquad (5.10) \\ P_1(n, t+\Delta t) =& (1 - (a'+b)\Delta t)P_1(n,t) + a\Delta t P_0(n,t) + b'\Delta t P_2(n,t) \\ &+ d \cdot \Delta t \cdot P_2(t-\tau)(\tilde{P}_1(n, \tau - \Delta t) - \tilde{P}_1(n-1, \tau - \Delta t)), \qquad (5.11) \\ P_2(n, t+\Delta t) =& (1 - (b'+d)\Delta t)P_2(n,t) + b\Delta t P_1(n,t) \\ &+ d \cdot \Delta t \cdot P_2(t-\tau)(\tilde{P}_2(n, \tau - \Delta t) - \tilde{P}_2(n-1, \tau - \Delta t)), \qquad (5.12) \end{aligned}$$

and use them to derive the dCMEs for the activator in the limit $\Delta t \to 0$:

$$\begin{aligned} \partial_t P_0(n,t) =& a'P_1(n,t) - aP_0(n,t) + d \cdot P_2(t-\tau)(\mathbb{E}^1 - 1)\tilde{P}_0(n-1, \tau), \qquad (5.13) \\ \partial_t P_1(n,t) =& b'P_2(n,t) - (a'+b)P_1(n,t) + aP_0(n,t) \qquad (5.14) \\ &+ d \cdot \mathbb{E}^{-1} P_2(n,t) + d \cdot P_2(t-\tau)(\mathbb{E}^1 - 1)\tilde{P}_1(n-1, \tau), \\ \partial_t P_2(n,t) =& bP_1(n,t) - (b'+d)P_2(n,t) + d \cdot P_2(t-\tau)(\mathbb{E}^1 - 1)\tilde{P}_2(n-1, \tau), \qquad (5.15) \end{aligned}$$

where $\mathbb{E}^x$ is the step operator defined by $\mathbb{E}^x f(n) = f(n+x)$. Coupled to this set of dCMEs we will additionally have a set of conditional master equations describing the probabilities of having produced n active Pol II in a time t for each respective gene state:

$$\partial_t \tilde{P}_0(n,t) = a'\tilde{P}_1(n,t) - a\tilde{P}_0(n,t), \tag{5.16}$$

$$\partial_t \tilde{P}_1(n,t) = b'\tilde{P}_2(n,t) - (a'+b)\tilde{P}_1(n,t) + a\tilde{P}_0(n,t) + d\cdot\mathbb{E}^{-1}\tilde{P}_2(n,t), \tag{5.17}$$

$$\partial_t \tilde{P}_2(n,t) = b\tilde{P}_1(n,t) - (b'+d)\tilde{P}_2(n,t), \tag{5.18}$$

with the initial conditions $\tilde{P}_1(n,0) = \delta_{n,0}, \tilde{P}_{0,2}(n,0) = 0$ as argued above.

Generating function solution to the dCME

Our first task is to find $P_2(t-\tau)$. Taking the sum of Eqs. (5.13)–(5.15) over all n we get the set of coupled equations defining the marginal probabilities of being in each gene state, $P_i(t) = \sum_n P_i(n,t)$, explicitly:

$$\partial_t P_0(t) = a'P_1(t) - aP_0(t), \tag{5.19}$$

$$\partial_t P_1(t) = (b'+d)P_2(t) - (a'+b)P_1(t) + aP_0(t), \tag{5.20}$$

$$\partial_t P_2(t) = bP_1(t) - (b'+d)P_2(t). \tag{5.21}$$

Solving these equations at steady state and invoking the normalisation condition $\sum_{i=1}^{3} P_i(t) = 1$, we find the probability of being in the U^{**} state is:

$$h \equiv P_2(t\to\infty) = \frac{ab}{(a'+a)(b'+d)+ab}. \tag{5.22}$$

Our main interest is in solving Eqs (5.13)–(5.15), but since $P_0(n,t)$, $P_1(n,t)$ and $P_2(n,t)$ are coupled to $\tilde{P}_0(n,\tau)$, $\tilde{P}_1(n,\tau)$ and $\tilde{P}_2(n,\tau)$ (and not vice-versa) we must first solve Eqs. (5.16)–(5.18). Defining the generating functions $\tilde{G}_i(z,t) = \sum_n z^n \tilde{P}_i(n,t)$ for $i \in \{0,1,2\}$, Eqs (5.16)–(5.18) transform into:

$$\partial_t \tilde{G}_0 = a'\tilde{G}_1 - a\tilde{G}_0, \tag{5.23}$$

$$\partial_t \tilde{G}_1 = a\tilde{G}_0 + b'\tilde{G}_2 + dz\tilde{G}_2 - (a'+b)\tilde{G}_1, \tag{5.24}$$

$$\partial_t \tilde{G}_2 = b\tilde{G}_1 - (b'+d)\tilde{G}_2, \tag{5.25}$$

where we have dropped the functional dependence of G_i on (z,t) for brevity. Defining the total generating function as $\tilde{G} = \sum_i \tilde{G}_i$, one can sum together Eqs. (5.23)–(5.25)and rearrange to give:

$$\tilde{G}_2 = \frac{\partial_t \tilde{G}}{d \cdot (z-1)}. \tag{5.26}$$

Further manipulating Eqs. (5.23)–(5.25) we find an equation exclusively in terms of $\tilde{G}$:

$$\partial_t^3 \tilde{G} + q_1 \partial_t^2 \tilde{G} + (q_2 + q_3 z)\partial_t \tilde{G} + q_4(1-z)\tilde{G} = 0, \tag{5.27}$$

where we have defined

$$\begin{aligned} q_1 =& a + a' + b + b' + d, \\ q_2 =& a'(b'+d) + a(b+d+b'), \\ q_3 =& -bd, \quad q_4 = abd. \end{aligned}$$

Using the exponential ansatz $\tilde{G}(z,t) \sim e^{\lambda(z)t}$ (since the ODE is homogeneous and coefficients of the derivatives are independent of t) one finds the solution to (5.27):

$$\tilde{G}(z,t) = \sum_{i=1}^{3} Q_i(z) e^{\lambda_i(z)t}, \tag{5.28}$$

where the $\lambda_i(z)$ are the solutions to the cubic equation:

$$\lambda(z)^3 + q_1\lambda(z)^2 + (q_2 + q_3(z-1))\lambda(z) + q_4(1-z) = 0,$$

and the $Q_i(z)$ are given by:

$$Q_i(z) = \frac{bd(z-1) + \lambda_j(z)\lambda_k(z)}{(\lambda_j(z) - \lambda_i(z))(\lambda_k(z) - \lambda_i(z))}, \tag{5.29}$$

where i,j,k are distinct and $i,j,k \in \{1,2,3\}$, and $Q_i(z)$ were chosen such that the initial conditions are satisfied. Note that one can verify computationally that the normalisation condition $\tilde{G}(z=1,t) = 1$ is satisfied.

We are now in place to solve the dCMEs (5.13)–(5.15). First, we transform them into their corresponding generating function equations at steady-state:

$$a'G_1 - aG_0 - d \cdot h \cdot (z-1)\tilde{G}_0(\tau) = 0, \tag{5.30}$$

$$aG_0 + b'G_2 + dzG_2 - (a'+b)G_1 - d \cdot h \cdot (z-1)\tilde{G}_1(\tau) = 0, \tag{5.31}$$

$$bG_1 - (b'+d)G_2 - d \cdot h \cdot (z-1)\tilde{G}_2(\tau) = 0. \tag{5.32}$$

Summing together these generating function equations one arrives at:

$$G_2 = h \cdot \tilde{G}(\tau). \tag{5.33}$$

Through further manipulation of Eqs. (5.30)–(5.32), and defining $G = G_0 + G_1 + G_2$, we arrive at the solution to the activator delay chemical master equation (for the total generating function):

$$G(z,\tau) = \frac{h}{ab} \cdot \sum_{i=1}^{3} \left(\lambda_i(z)(q_1 + \lambda_i(z)) + q_2 + q_3(z-1)\right) Q_i(z) e^{\lambda_i(z)\tau}. \tag{5.34}$$

Through z derivatives of $G(z,\tau)$ one can obtain the probability distribution and the moments as seen in Section 2.2.

5.5 Relationship between two-state and mechanistic models

5.5.1 When can the two-state and mechanistic models be matched? A waiting time distribution perspective

In Appendix **??** and C.2, we calculate the distribution of the time between the production of two consecutive M (A) molecules for the mechanistic and (delayed) telegraph models. Using these distributions we can compute the square of the coefficient of variation of the time between two consecutive M (or A) production events. Throughout this chapter, we will refer to this time between production events as the waiting time. In line with previous usage in the single enzyme molecule literature [54], we shall refer to the coefficient of variation of the waiting time distribution as the *randomness parameter*, which is given by:

$$R^{tele} = \frac{\langle t^2 \rangle - \langle t \rangle^2}{\langle t \rangle^2} = 1 + \frac{2\rho\sigma_u}{(\sigma_b + \sigma_u)^2}. \tag{5.35}$$

for the *telegraph or delayed telegraph models* and by:

$$R^{mec} = \frac{\langle t^2 \rangle - \langle t \rangle^2}{\langle t \rangle^2} = 1 + \frac{2bc\left(a'\left(b' + c - a\right) - a^2\right)}{\left(a'\left(b' + c\right) + a(b + b' + c)\right)^2}, \tag{5.36}$$

for the mechanistic model. Note that the waiting time statistics for A and M in the two-state models are the same because the waiting time distribution calculation is not sensitive to the mode of degradation (first-order or delayed) since the absorbing state corresponds to the production of a new mature mRNA transcript or a new active Pol II

which necessarily always precedes its degradation or removal. In addition to this reason, the statistics are the same for active Pol II and mature mRNA in the mechanistic model also because of the fixed time τ between the unpausing of a Pol II and the production of a mature mRNA.

Note that while the randomness parameter for A or M is greater than 1 for all parameter values (see Eq. (5.35)) in the two-state models, the same statistical measure can be less than or greater than 1 in the mechanistic model (see Eq. (5.36)). In fact, evaluating the latter equation for over a million random values of parameters suggests that $R^{mec} \geq 1/2$. The 2 appears because in our model, it is the smallest number of reaction steps between A production events ($U^\star \to U^{\star\star} \to U^\star + A$). Similar results have been derived in the context of single molecule enzyme kinetics [54].

It follows that the two-state models can only capture the waiting time statistics of the mechanistic model (up to second order) when $R^{mec} \geq 1$ which is the case when the following condition is satisfied

$$b' + c \geq \frac{a}{a'}(a + a'). \tag{5.37}$$

This implies that the conditions which favour a description of the mechanistic model by the two-state models are: (i) premature termination and unpausing from the paused promoter-proximal state must be fast i.e., large $b' + c$; (ii) transcription factor binding to DNA elements and the reverse unbinding reaction must be slow i.e., small $a + a'$; (iii) transcription factor unbinding is fast compared to transcription factor binding i.e., a/a' is small. Note that the condition given by Eq. (5.37) is not a function of b, the rate at which polymerase binds the promoter and moves to the proximal paused state (see later for an explanation of the role of b).

5.5.2 Analytical expressions for the effective parameters of the two-state models

Matching the first three moments of the waiting time distribution of the times between consecutive M or A production events of the telegraph/delayed telegraph model (given by Eqs. (C.5)) with those calculated using the mechanistic model (given by Eq. (C.18) evaluated for $i = 1, 2, 3$), we obtain a set of 3 simultaneous equations for the effective parameters of the two-state models ρ, σ_u, σ_b. The solution of these equations gives:

$$\begin{aligned}
\rho &= \frac{bca'\left(\Delta a' + a^2\right)}{a'\left(a'\left(\Delta a' + a^2 + 3a\Delta + \Delta(b+\Delta)\right) + a^2(2a + b + 2\Delta)\right) + a^4}, \\
\sigma_u &= \frac{\Delta^3\left(a'\right)^4}{\left(\Delta a' + a^2\right)\left(a'\left(a'\left(\Delta a' + a^2 + 3a\Delta + \Delta(b+\Delta)\right) + a^2(2a + b + 2\Delta)\right) + a^4\right)}, \\
\sigma_b &= \frac{a\Delta a'}{\Delta a' + a^2},
\end{aligned} \tag{5.38}$$

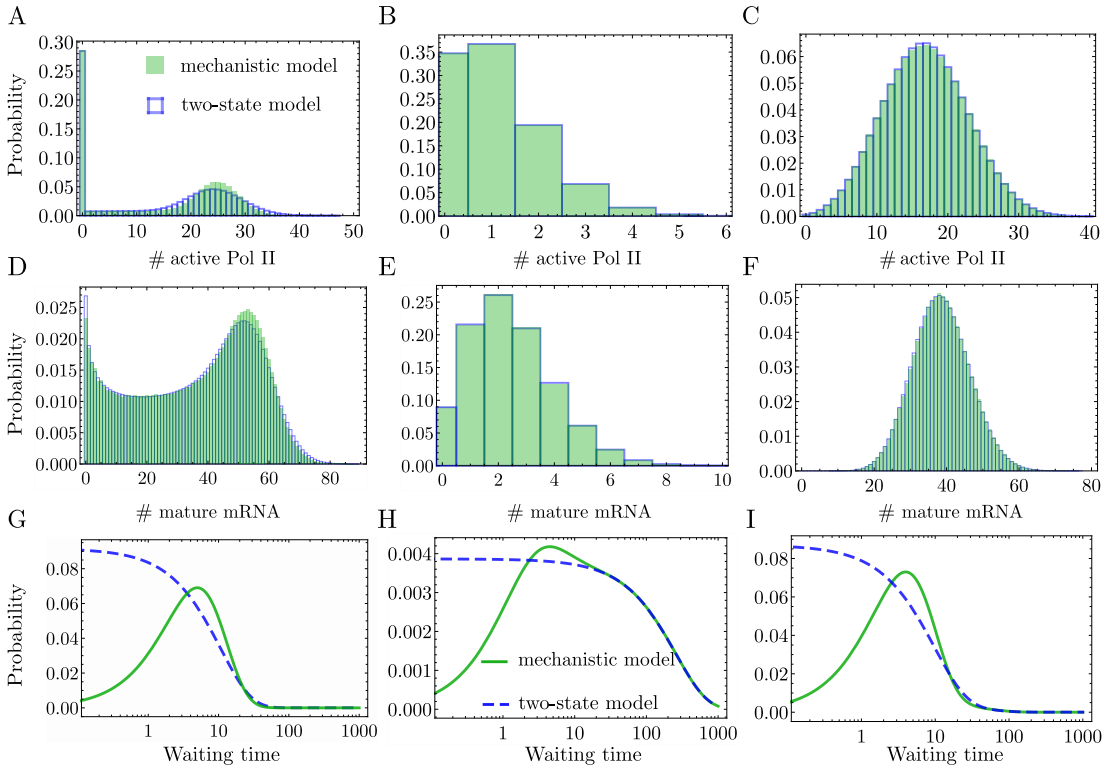


Figure 5.2: Comparison between the molecule number distributions of active Pol II and mature mRNA distributions of the two-state (reduced) models ((5.2) and (5.3)) and the mechanistic model (5.1). For a particular choice of parameters of the mechanistic model, Eq. (5.38) gives the two-state model parameters. In (A-C), for three different parameter sets (one per column), we show that the mechanistic model describing the number of active Pol II molecules is well approximated by the exact steady-statef solution of the delay telegraph model [120] evaluated with the effective parameters. (D-F) show a similar level of agreement between the mechanistic and two-state models but instead for the mature mRNA distributions, where the two-state model is now the telegraph model whose exact steady-state solution can be found in [35]. Steady-state distributions of the mechanistic model were obtained using the delay SSA (Algorithm 2 of [117]) with 10^4 samples. In (G-I) we show the corresponding distributions of the waiting time between two consecutive active Pol II (or mature mRNA) production events for the two-state and mechanistic models. The waiting time distributions for the mechanistic and two-state models are calculated by taking the inverse Laplace transform of Eq. (C.4) and Eq. (C.17) respectively. Clearly, the models can be distinguished through their waiting time distributions *even when their number distributions are almost indistinguishable*. Parameters of the mechanistic model and the corresponding effective parameters for two-state models are: (A) $a = 0.001$ s^{-1}, $a' = 0.001$ s^{-1}, $b = 0.16$ s^{-1}, $b' = 0.016$ s^{-1}, $c = 0.24$ s^{-1} mapped to $\sigma_u = 0.0007$ s^{-1}, $\sigma_b = 0.001$ s^{-1}, $\rho = 0.092$ s^{-1}; (B) $a = 0.144$ s^{-1}, $a' = 0.032$ s^{-2}, $b = 0.016$ s^{-1}, $b' = 0.56$ s^{-1}, $c = 0.24$ s^{-1} mapped to $\sigma_u = 3.8 \times 10^{-8}$ s^{-1}, $\sigma_b = 0.002$ s^{-1}, $\rho = 0.004$ s^{-1}; (C) $a = 0.032$ s^{-1}, $a' = 0.032$ s^{-1}, $b = 0.16$ s^{-1}, $b' = 0.016$ s^{-1}, $c = 0.32$ s^{-1} mapped to $\sigma_u = 0.012$ s^{-1}, $\sigma_b = 0.029$ s^{-1}, $\rho = 0.086$ s^{-1}. The mature mRNA decay rate is $d = 0.0016$ s^{-1}, and the delay time due to elongation is $\tau = 273.62$ s.

where $\Delta = b' + c - a - (a^2/a')$. Note that if the condition given by Eq. (5.37) is satisfied, then $\Delta \geq 0$ and hence the effective parameters defined by Eqs. (5.38) are positive and physically meaningful. If the condition is not satisfied, then one of these effective parameters is negative which means that there are no two-state models that can approximate the mechanistic model's waiting time moments up to third-order. We emphasise that the effective parameters are the same for the telegraph and delay telegraph models because the waiting time calculation is insensitive to the mode of decay (first-order or delayed). In Fig. 2(A-F) we compare the steady-state number distribution of the two-state models (which is analytically derived in [35] and [120]) evaluated with these effective parameters (for $\Delta > 0$) and the steady-state number distribution of the mechanistic model (which is obtained from stochastic simulations modified to take into account fixed time delays [117]). In the cases shown, the two-state models provide an excellent match to the mechanistic model for both unimodal and bimodal distributions of active Pol II and mature mRNA numbers. Note that since most of the parameters in the mechanistic model have not been measured directly, we chose parameters such that the number distributions looked similar to those measured experimentally and such that the average number of mRNA is larger than the average number of active Pol II (the former can range from few tens to few hundreds whereas the latter is at most few tens) [32, 45].

The case of fast switching between $U^\star$ and $U^{\star\star}$

Where $\min(b, b') \gg \max(a, a')$, the two states $U^\star$ and $U^{\star\star}$ can be effectively subsumed into a single super state W and the system dynamics amounts to switching between an inactive state U and an active state W. Physically, one sees that this arises since in this limit transitions between $U^\star$ and $U^{\star\star}$ occur almost instantaneously compared to transitions between U and $U^\star$. The transition rate from U to W is the same as from U to $U^\star$ and hence in the two-state model this implies

$$\sigma_b = a. \tag{5.39}$$

The transition rate from W to U must be equal to the transition rate from $U^\star$ to U multiplied by the probability of being in state $U^\star$ given that currently the effective system is in state W. This implies

$$\sigma_u = a' \frac{b'}{b + b'}. \tag{5.40}$$

Similarly, the effective production rate is the rate of producing active Pol II from state $U^{\star\star}$ multiplied by the probability of being in this state given that currently the effective system is in state W. This implies

$$\rho = c\frac{b}{b+b'}. \tag{5.41}$$

These results can be formally obtained from Eqs. (5.38) by choosing $b' = \gamma b$ (where γ is a constant) and taking the limit $b \to \infty$. The case of fast switching in a similar three-state model of gene expression (without an explicit description of active Pol II dynamics) has been previously studied in [170]. While it is obvious that fast switching between $U^{\star}$ and $U^{\star\star}$ simplifies to an effective two-state model, our condition (5.37) shows that *fast switching is sufficient but not a necessary condition for a two-state model to describe the dynamics of the mechanistic model*. We note that fast switching between $U^{\star}$ and $U^{\star\star}$ is unlikely to be the general case since the average time scale of Pol II pausing is $\sim$ 7min [220] and almost 1 hour in a small subset of genes[243]. This indicates Pol II pausing is very stable and "not the consequence of fast, repeated rounds of initiation and termination" [220].

Distinguishing between two-state and mechanistic models using waiting time distributions

It is interesting to note that while for $\Delta \geq 0$ the two-state and mechanistic models are practically indistinguishable by comparison of their number distributions, *they can be always distinguished by the distribution of the time between consecutive active Pol II or mRNA production events*. In particular, in Fig. 5.2(G-I) we show that while $f(t)$, the waiting time distribution between consecutive production events, is a monotonically decreasing function for the two-state models, it has a peak at a non-zero value of time for the mechanistic model. Another distinguishing feature is that for two-state models, $f(0)$ is non-zero while for the mechanistic model it is exactly zero. The latter feature can be explained as follows. For two-state models, since there is no change in the gene state when production occurs, hence there is no lower bound on how short the time between two consecutive production events can be. However, in the mechanistic model, a production event is accompanied by a change of state (from $U^{\star\star}$ to $U^{\star}$), therefore there is a finite non-zero time to switch back to state $U^{\star\star}$ from which the next production event occurs. Consequently, for the mechanistic model $f(0)$ must be zero.

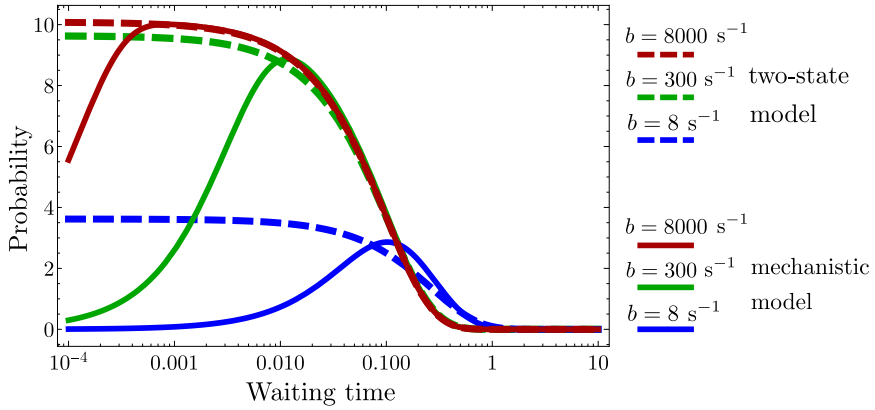


Figure 5.3: The waiting time distribution of the mechanistic model as a function of the rate parameter b (which controls the binding of Pol II to the promoter and the entry into the promoter-proximal state). The waiting time distribution of the mechanistic model is calculated by taking the inverse Laplace transform of Eq. (C.17). Note that as b increases, the peak moves closer to zero and the waiting time distribution of the mechanistic model approaches the waiting time distribution of the two-state model (calculated by taking the inverse Laplace transform of Eq. (C.4)). The parameter b is changed as described in the legend and the rest of the parameters are $a = 0.1$ s^{-1}, $a' = 0.1$ s^{-1}, $b' = 4$ s^{-1}, $c = 10$ s^{-1}. The parameters of the two-state model are calculated from Eqs. (5.38).

By this reasoning, it follows that the mode should be close to zero whenever the state $U^{\star\star}$ is recovered rapidly after an active Pol II production event, which occurs when b is large. In Fig. 5.3 we confirm this intuition and show that for the mechanistic model as we increase b, the waiting time distribution of the two-state model better approximates the waiting time distribution of the mechanistic model. Note that a log-scale is used on the x-axis to emphasise that there are always differences between the mechanistic and two-state models for small values of t.

5.6 Sensitivity analysis

Equations (5.38) allow us to understand how the parameters of the mechanistic model influence the effective parameters of the two-state models. We define the ordered set of mechanistic model parameters as $\theta^{mec} = \{a, a', b, b', c\}$ and the ordered set of the two-state model parameters as $\theta^{tele} = \{\rho, \sigma_u, \sigma_b\}$. In Table 5.1, we show the sign of the derivative of a parameter in a two-state model with respect to changes in the parameter of the mechanistic model (when $\Delta \geq 0$). For example, the first row shows the sign of the derivative of ρ with respect to a, a', b, b' and c. A positive (negative) sign for the pair (ρ, a) indicates that an increase in a leads to an increase (decrease) in ρ. We also show the same but for the burst size $\beta = \rho/\sigma_u$, a commonly cited measure equal to the amount of mRNA produced while the gene is in the on state (in the two-state models). Note that while the sign is fixed for most cases, in three instances the sign can flip. There is also a case where one of the two-state model parameters is independent of one of the

	a	a'	b	b'	c
ρ	+/-	-	+	-	+
σ_u	-	+	-	+	+
σ_b	+/-	+	0	+	+
$\beta{=}\rho/\sigma_u$	+	-	+	-	+/-

Table 5.1: Signs of the derivatives of the two-state effective parameters ρ, σ_u and σ_b with respect to the mechanistic model parameters a, a', b, b' and c. Expressions for the effective parameters are given by Eq. (5.38).

parameters of the mechanistic model (marked with a 0). Due to the complicated nature of Eqs. (5.38) it is difficult to deduce the signs in Table 5.1 using simple arguments, however in some cases it can be done. For example, the relationship of ρ with respect to parameters b, b' and c is intuitive since: (i) increasing b increases the time in the $U^{\star\star}$ state meaning production of A (or M) happens more often; (ii) decreasing b' has the opposite effect, meaning production of A (or M) occurs less often; and (iii) increasing c obviously increases the production rate of A (or M) and hence increases the predicted value of ρ.

Next, we investigate the sensitivities of the parameters θ^{tele} of the two-state model to the parameters of the mechanistic model θ^{mec}. For this purpose, we randomly selected 10^3 parameter sets from a log-scaled space in the θ^{mec} parameters, accepting only those parameter set combinations that came within 2 experimental errors of the measurements of Oct4 gene: $\rho = 3.2 \times 10^{-2} \pm 1.0 \times 10^{-2}\,\mathrm{s}^{-1}$, $\sigma_u = 3 \times 10^{-3} \pm 2 \times 10^{-3}\,\mathrm{s}^{-1}$ and $\sigma_b = 1.5 \times 10^{-4} \pm 0.5 \times 10^{-4}\,\mathrm{s}^{-1}$ [32]. We also did this for the Nanog gene whose measurements were: $\rho = 1.3 \times 10^{-2} \pm 0.3 \times 10^{-2}\,\mathrm{s}^{-1}$, $\sigma_u = 1.2 \times 10^{-4} \pm 0.2 \times 10^{-4}\,\mathrm{s}^{-1}$ and $\sigma_b = 3.2 \times 10^{-5} \pm 0.3 \times 10^{-5}\,\mathrm{s}^{-1}$. A log-scaled parameter space was used such that various combinations of mechanistic model parameter timescales could be easily explored. The ranges of the mechanistic model parameters that we explored were $\theta_i^{mec} \in [10^{-4}, 10]\,\mathrm{s}^{-1}$ for the Oct4 gene and $\theta_i^{mec} \in [10^{-5}, 10^{-1}]\,\mathrm{s}^{-1}$ for the Nanog gene. The sensitivities calculated are the absolute values of the relative sensitivities given by,

$$\mathrm{sen}(\theta_i^{tele}, \theta_j^{mec}) = \left| \frac{\theta_j^{mec}}{\theta_i^{tele}} \frac{d\theta_i^{tele}}{d\theta_j^{mec}} \right| = \left| \frac{d(\log(\theta_i^{tele}))}{d(\log(\theta_j^{mec}))} \right|, \tag{5.42}$$

where $\mathrm{sen}(\theta_i^{tele}, \theta_j^{mec})$ is the magnitude of the relative sensitivity of θ_i^{tele} with respect to θ_j^{mec}.

As we show in Fig. 5.4, we find that for both genes, (i) the initiation rate ρ of the two-state models is most sensitive to parameters b and c in the mechanistic model i.e., parameters that control the rate of Pol II binding, of entering and leaving the promoter-proximal paused state; (ii) the rate of switching *off* of the two-state models σ_u is most sensitive to parameter a' (controlling transcription factor unbinding) and also

to parameters b, c which control the initiation rate; (iii) the rate of switching *on* of the two-state models σ_b is most sensitive to parameter a (controlling transcription factor binding) but also to parameters a', c which control the the rate of switching *off* and the initiation rate. In Fig. 4, we also show which parameters of the mechanistic model are the three parameters of the two-state model least sensitive to. This analysis identifies how "microscopic" parameters of the mechanistic model affect the "macroscopic" parameters of the two-state models. *More importantly, it shows that the latter are typically correlated due to their dependence on common microscopic parameters.*

5.7 Model reduction using number statistics or three-state models

Thus far, we have explored model reduction solely using waiting time statistics. Alternatively, one can match two-state and mechanistic models using moments of the number of molecules. As well, one can match three-state models and mechanistic models using waiting time or number statistics. In this section, we explore these alternative perspectives.

5.7.1 Obtaining reduced models with two states using number statistics

We begin by finding the Fano factor (defined as the variance divided by the mean) of the active Pol II and mature mRNA numbers in both the two-state and mechanistic models. In Appendices C.3 and C.4, we derive expressions for the mean and variance of active Pol II and mature mRNA numbers in steady-state conditions for both the mechanistic and two-state models (for a test of their accuracy versus stochastic simulations using the delay SSA see Table C.1). The Fano factor of the two-state models is easily proved to be always greater than 1. Specifically, for the delayed and standard telegraph models, we have respectively:

$$\mathrm{FF}_A^{dtele} = 1 + \frac{2\rho\sigma_u\left(e^{-(\sigma_b+\sigma_u)\tau}-1\right)}{\tau\left(\sigma_b+\sigma_u\right)^3} + \frac{2\rho\sigma_u}{\left(\sigma_b+\sigma_u\right)^2}, \tag{5.43}$$

$$\mathrm{FF}_M^{tele} = 1 + \frac{\rho\sigma_u}{\left(\sigma_b+\sigma_u\right)\left(\sigma_b+d+\sigma_u\right)}. \tag{5.44}$$

The Fano factor of the number of active Pol II in the mechanistic model is given by:

$$\begin{aligned}\mathrm{FF}_A^{mec} = R^{mec} - \frac{1}{\gamma\tau}A_0 + \frac{A_1}{\gamma\tau}\exp\left(-\frac{1}{2}\tau\left(S-\sqrt{(S-2a)^2+4a'\left(a-b'-c\right)}\right)\right) \\ + \frac{A_2}{\gamma\tau}\exp\left(-\frac{1}{2}\tau\left(S+\sqrt{(S-2a)^2+4a'\left(a-b'-c\right)}\right)\right),\end{aligned} \tag{5.45}$$

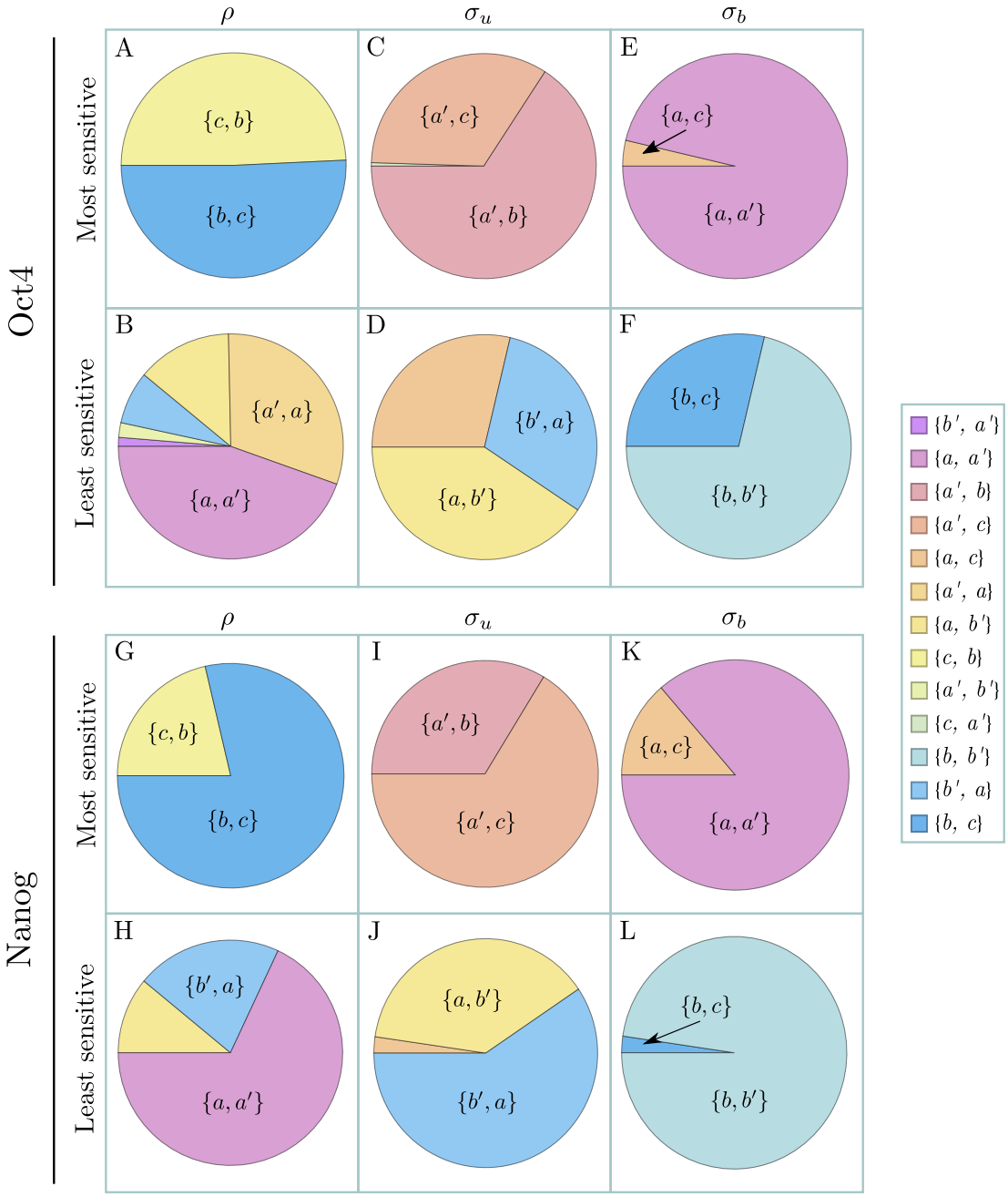


Figure 5.4: Pie charts showing the most and least sensitive of the telegraph model parameters $\theta^{tele} = \{\rho, \sigma_u, \sigma_b\}$ with respect to mechanistic model parameters $\theta^{mec} = \{a, a', b, b', c\}$, for the Oct4 gene in (A-F) and for the Nanog gene in (G-L) [32]. We chose 10^3 parameter sets θ^{mec} at random, accepting only parameter sets for which the predicted telegraph model parameters ρ, σ_u and σ_b from Eq. (5.38) were within 2 experimental errors of values reported in [32]. From these randomly chosen parameter sets, we then calculated the relative sensitivity $\text{sen}(\theta_i^{tele}, \theta_j^{mec})$ which is given by Eq. (5.42). The proportions on the pie charts show the proportion of parameter sets for which $\{i, j\}$ were the most/least sensitive parameters, where $\{i, j\}$ states that i is the most/least sensitive parameter followed by j. (A) for Oct4, the most sensitive parameters for ρ are b and c, with the majority of parameter sets being most sensitive to b and second-most to c. (B) for Oct4, the least sensitive parameters for ρ are a and a', with the majority of parameter sets being least sensitive to a and second-least sensitive to a'. (C-F) follow similar interpretations as made for (A) and (B) for the Oct4 gene, and (G-L) follow similar interpretations for the Nanog gene.

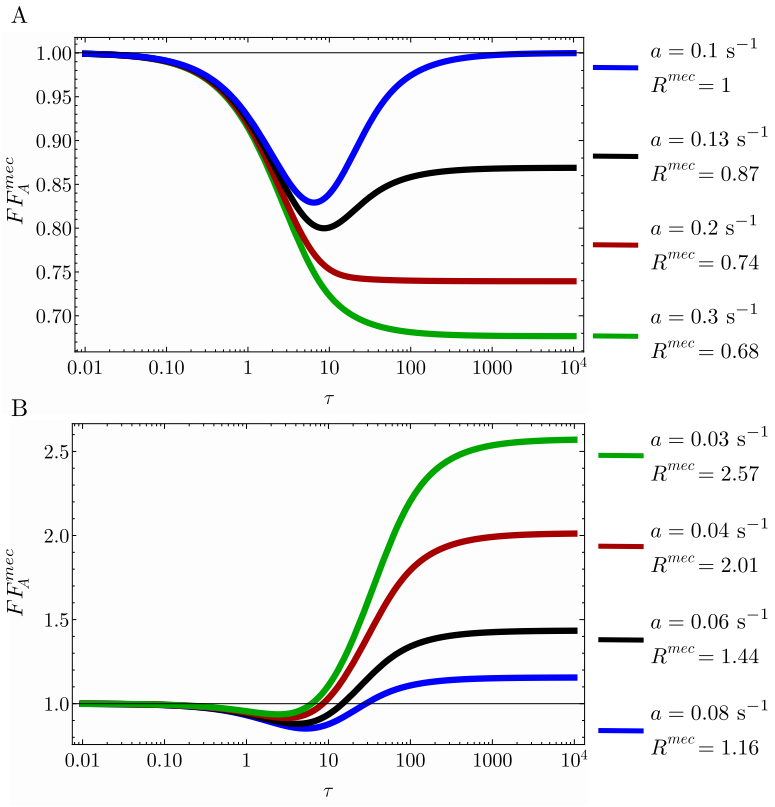


Figure 5.5: Fano factor of the active Pol II number distribution for the mechanistic model as a function of the elongation time τ and the randomness parameter R^{mec}. The Fano factor is evaluated using Eq. (5.45). Note that the large τ limit of the Fano factor is equal to the randomness parameter R^{mec} which is given by Eq. (5.36); R^{mec} is here varied via the parameter a whilst keeping the rest of parameters constant: $b' = 0.0125$ s^{-1}, $a' = 0.032$ s^{-1}, $b = 0.16$ s^{-1}, $c = 0.4$ s^{-1}. (A) shows that if $R^{mec} \leq 1$ then the Fano factor is less than 1 for all τ. (B) shows that if $R^{mec} \geq 1$ then the Fano factor is less than 1 for a small enough value of τ.

where A_0, A_1 and A_2 can be positive or negative and are complicated functions of a, a', b, b', c, and where we have defined

$$S = a + a' + b + b' + c,$$
$$\gamma = \frac{abc}{a'(b'+c) + a(b'+b+c)}. \tag{5.46}$$

Note that the arguments of the exponential functions are negative for all positive values of the parameters. The Fano factor of the mature mRNA statistics is derived in Appendix C.4 and is given by:

$$\mathrm{FF}_M^{mec} = 1 + bc\left(a'(b'+c) - a(a'+d) - a^2\right)/\chi, \tag{5.47}$$

with,

$$\chi = \left(a'(b'+c) + a(b'+b+c)\right)\left(a'(b'+c+d) + a(b'+b+c+d) + d(b'+b+c+d)\right). \tag{5.48}$$

Since $\mathrm{FF}_A^{dtele} \geq 1$ *and* $\mathrm{FF}_M^{tele} \geq 1$*, clearly model reduction using molecule number moments will only be possible if the parameters of the mechanistic model are such that* $\mathrm{FF}_A^{mec} \geq 1$ *and* $\mathrm{FF}_M^{mec} \geq 1$. For the mature mRNA, this analysis is straightforward. Similar to the derivation of the condition (5.37), from the numerator of the second term in Eq. (5.47), it can be deduced that $\mathrm{FF}_M^{mec} \geq 1$ provided the following condition holds:

$$b' + c \geq \frac{a}{a'}(a + a' + d). \tag{5.49}$$

When condition (5.49) is satisfied, one can find a mapping between the standard telegraph model describing mature mRNA and the mechanistic model. We note that this condition is not the same as that derived from model reduction using waiting time statistics, namely Eq. (5.37). In fact, if Eq. (5.37) is not satisfied then neither is Eq. (5.49) i.e., for all points in parameter space in which it is not possible to match the moments of waiting time distributions of the two-state and mechanistic models, it is also not possible to match the moments of the mature mRNA numbers. However, it also follows that there is a region of parameter space of the mechanistic model where moment matching of the two-state model using waiting time statistics is possible (Eq. (5.37) is satisfied) whereas moment matching using number statistics is not (Eq. (5.49) is not satisfied). This region of parameter space where the two methods give different results is very small whenever $a + a' \gg d$ where the rates of transcriptional factor binding/unbinding to the promoter are much larger than the rate of mature mRNA degradation. This seems

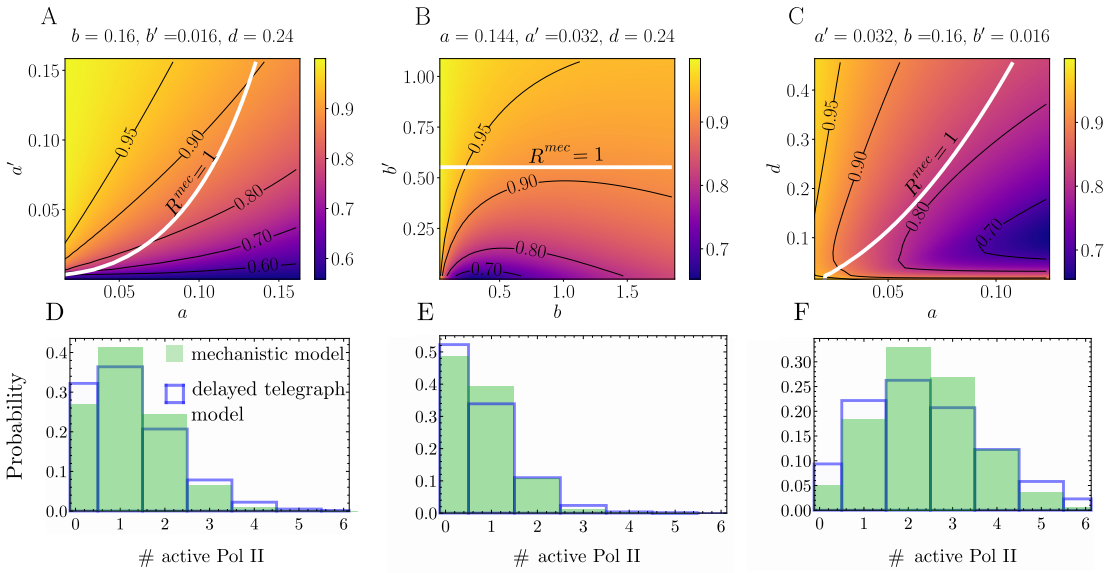


Figure 5.6: Accuracy of the number distributions of active Pol II constructed using the delay telegraph model for parameters close to the boundary $R^{mec} = 1$. (A-C) The minimum Fano factor of active Pol II numbers in the mechanistic model as a function of the parameters $\theta^{mec} = \{a, a', b, b', c\}$. For a given set of θ^{mec}, the minimum Fano factor is found by varying the elongation time τ in Eq. (5.45). The region above the white line $R^{mec} = 1$ is where model reduction using waiting time statistics is possible. Note that inside this region, the minimum Fano factor of active Pol II numbers is smaller than 1 meaning that for small enough values of τ, model reduction using number statistics is not possible. However, in (D-F) we show that even when this is the case, the number distributions constructed using the delay telegraph model with effective parameters given by Eq. (5.38) provide a reasonably good approximation to the mechanistic model distribution of active Pol II. Note that these distributions represent a worst case scenario—inside the boundary, for the vast majority of points, the two-state model distributions provide an almost perfect fit to the mechanistic model distributions as shown in Fig. 2. The parameters are as follows: (D) $a = 0.0336$ s^{-1}, $a' = 0.006$ s^{-1}, $b = 0.16$ s^{-1}, $b' = 0.016$ s^{-1}, $c = 0.24$ s^{-1}, $\tau = 13.65$ s, $\text{FF}_A^{mec} = 0.74$, $R^{mec} = 1.034$. (E) $a = 0.072$ s^{-1}, $a' = 0.0288$ s^{-1}, $b = 0.16$ s^{-1}, $b' = 0.016$ s^{-1}, $c = 0.24$ s^{-1}, $\tau = 8.76$ s, $\text{FF}_A^{mec} = 0.81$, $R^{mec} = 1.004$. (F) $a = 0.08$ s^{-1}, $a' = 0.005$ s^{-1}, $b = 2$ s^{-1}, $b' = 0.0001$ s^{-1}, $c = 1.36$ s^{-1}, $\tau = 3$ s, $\text{FF}_A^{mec} = 0.62$, $R^{mec} = 1.0001$.

to be generally the case since degradation timescales of mature mRNA are generally many hours in eukaryotic cells [44]. Incidentally, this offers an explanation why the Fano factor of mature mRNA is invariably measured to be greater than 1 in the literature of eukaryotic gene expression [244, 45, 46].

Due to the complicated nature of the expression in Eq. (5.45), the derivation of an analytic condition for which the Fano factor of active Pol II is greater than 1 appears to be difficult to obtain. However, in the limit of $\tau \to \infty$ it is clear that $\text{FF}_A^{mec} \to R^{mec}$. Hence, in the limit of long elongation times, the condition necessary for model reduction using active Pol II moment number statistics i.e., $\text{FF}_A^{mec} \geq 1$, is equivalent to the condition necessary for model reduction using waiting time statistics given by Eq. (5.37) (which is the same as $R^{mec} \geq 1$). This is intuitive since the waiting time calculation

does not consider the removal of active Pol II via elongation but only their production time statistics. It can also be proved from Eq. (5.45) that in the limit $\tau \to 0$ we have $\mathrm{FF}_A^{mec} \to 1$. What happens for finite $\tau > 0$ is difficult to deduce from Eq. (5.45) and hence we investigate this numerically.

In Fig. 5.5 we evaluate Eq. (5.45) as a function of τ for a number of parameter sets with different R^{mec}. Several notable features can be seen: (i) if $R^{mec} < 1$ then $\mathrm{FF}_A^{mec} < 1$ i.e., if model reduction using waiting time statistics is not possible then it is also impossible using number statistics; (ii) for $R^{mec} \geq 1$, as we increase τ, FF_A^{mec} decreases from 1 to a value below 1, reaches a minimum and then increases up to the value R^{mec}. *Consequently, if the condition for model reduction using waiting time statistics is satisfied, it is not necessarily true that it is possible to achieve model reduction according to number statistics.* In Fig. 5.6 (A-C), we show a heat map of the minimum Fano factor (achieved at intermediate τ) in the parameter space of the mechanistic model. Note that the minimum achieved inside the region where $R^{mec} > 1$ (the region above the white line) is not far below 1. As a consequence, while here there is no model reduction from a number statistics point of view, model reduction using waiting time statistics is possible, and the distribution computed using the effective parameters given by Eqs. (5.38) while not perfect, it is acceptable—see Fig. 5.6 (D-F).

Thus far, we have looked at model reduction via number statistics from the perspective of when the Fano factor numbers of the mechanistic and two-state models are both greater than one. In Appendix C.5 we extend this analysis further by considering two other types of model reduction via number statistics: (1) matching of the molecule number moments and (2) of the number distributions of the mechanistic and the two-state models for active Pol II and mature mRNA numbers. In particular, we found the following: (i) within the region of parameter space of the mechanistic model described by the condition Eq. (5.37), it was possible to numerically find parameters of the two-state models such that the first three moments of the active Pol II and mature mRNA number distributions of the two-state models agreed with those of the mechanistic model—see Fig. 5.7 (A-C) and (G-I); (ii) the Hellinger distance between the molecule number distributions predicted by the mechanistic model and the molecule number distributions of the two-state models that provides the best approximate distribution of the mechanistic model, is very small within the region defined by Eq. (5.37)—see Fig. 5.7 (D-F) and (J-L). The analysis shows there is a close relationship between model reduction using waiting time and number statistics, and supports the conclusions reached in Sections 5.5 and 5.6 using waiting time statistics.

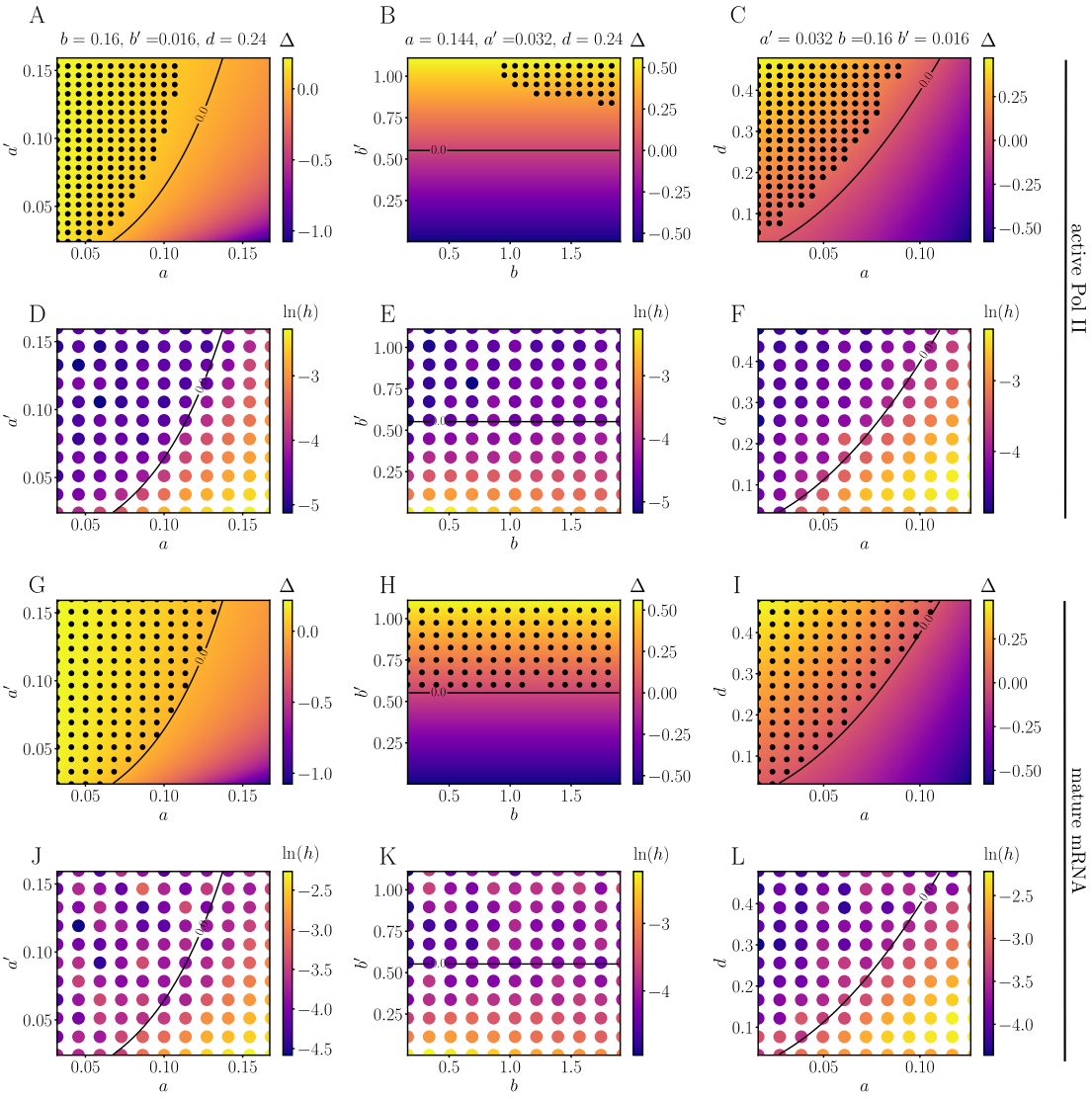


Figure 5.7: Comparison of model reduction of the mechanistic model to two-state models using two different types of number statistics and comparison with reduction from waiting time statistics. In (A-C) black dots show the points in parameter space where the first 3 moments of the active Pol II number distributions of the mechanistic and the delayed telegraph model match numerically using the Newton-Raphson method; in (G-I) we show the same for the distributions of mature mRNA of the mechanistic and telegraph models. The heat map shows the the value of $\Delta = b' + c - a - (a^2/a')$ and the solid black lines divides areas where $\Delta > 0$ (waiting time moment matching exists) and $\Delta < 0$ (waiting time moment matching does not exist). Note that the black dots in A-C do not fill the whole region $\Delta > 0$ because of numerical issues with the solver (See Appendix C.5 for a discussion). In (D-F) and (J-L) we show the Hellinger distance (h in log scale) between the molecule number distributions predicted by the mechanistic model and the molecule number distributions of the two-state models that provides the best approximate distribution of the mechanistic model; the parameters of the two-state models are those learnt after $O(10^5)$ iterations of an algorithm that maximises the likelihood. The mature mRNA decay rate $d = 0.0016$ s $^{-1}$ in all cases and the delay time is $\tau = 273.62$ s. See Appendix C.5 for details of the numerical procedures used.

5.7.2 Obtaining reduced models with three states using waiting time statistics

Thus far, we have considered the approximation of the mechanistic model by two-state models (telegraph and delay telegraph models). However, some papers have postulated the existence of two off states for some mammalian genes because the time spent by the gene in the off state is measured to be non-exponential [47]. This has led to a variation of the telegraph model, which we will refer to as the refractory model

$$G \xrightarrow{\sigma_u} G^\star \xrightarrow{\sigma_u^*} G^{\star\star} \xrightarrow{\sigma_b} G, \quad G \xrightarrow{\rho} G + M, \quad M \xrightarrow{d} \varnothing. \tag{5.50}$$

One can also postulate a modification (parallel to the delay telegraph model) that describes active Pol II rather than mature mRNA:

$$G \xrightarrow{\sigma_u} G^\star \xrightarrow{\sigma_u^*} G^{\star\star} \xrightarrow{\sigma_b} G, \quad G \xrightarrow{\rho} G + A, \quad A \underset{\tau}{\Longrightarrow} \varnothing. \tag{5.51}$$

An analysis akin to the one shown for the two-state models in Appendix C.1 shows that the Laplace transform of the waiting time distribution of the time between consecutive active Pol II or mature mRNA production events is given by:

$$\tilde{f}(s) = \frac{\rho\left(\sigma_b + s\right)\left(\sigma_u^* + s\right)}{\left(\rho + s\right)\left(\sigma_b + s\right)\sigma_u^* + s\sigma_u\sigma_u^* + s\left(\sigma_b + s\right)\left(\rho + s + \sigma_u\right)}, \tag{5.52}$$

where $\tilde{f}(s) = \int_0^\infty f(t)e^{-st}dt$. From the definition of the Laplace transform, we have that the moments are given by

$$\langle t^i \rangle = (-1)^i \partial_s^i \tilde{f}(0). \tag{5.53}$$

The randomness parameter is then given by the square of the coefficient of variation squared of the time between two consecutive production events:

$$R = \frac{\langle t^2 \rangle - \langle t \rangle^2}{\langle t \rangle^2} = 1 + \frac{2\rho\sigma_u\left(\sigma_b\sigma_u^* + \left(\sigma_u^*\right)^2 + \sigma_b^2\right)}{\left(\left(\sigma_b + \sigma_u\right)\sigma_u{}^* + \sigma_b\sigma_u\right)^2}. \tag{5.54}$$

Hence the randomness parameter of the reduced models (5.50) and (5.51) is always greater than 1. In contrast, we have already shown by Eq. (5.36) that for the mechanistic model the randomness parameter can be greater than or less than 1. Hence it follows that the condition given by Eq. (5.37) is necessary for both telegraph models and those with a refractory state to approximate the mechanistic model. Similar to what we have previously done for the two-state models, analytical expressions expressing the four parameters of the reduced refractory models in terms of the six parameters of the mechanistic model can be derived by matching the first four moments of the waiting time distribution of the two models. The steady-state distribution solutions of active Pol II and mature mRNA numbers of the reduced refractory models evaluated with

these effective parameters provide an excellent approximation to the distributions of the mechanistic model. However, since this was already achieved using two-state models and since the refractory models have the same limitations as the two-state models (randomness parameter cannot be less than 1), it follows that *two-state models provide the optimal choice for model reduction within the parameter space defined by* Eq. (5.37).

5.8 Discussion

In this chapter we have investigated to what extent can two-state models predict the active Pol II and mature mRNA dynamics of a more realistic mechanistic model that incorporates transcriptional factor binding and unbinding, Pol II dynamics (binding, pausing, release, elongation) and mature mRNA dynamics. We found that there is a region of parameter space where there exists a choice of parameters of two-state models in terms of the mechanistic model such that the first three moments of their waiting time distributions exactly match. The distributions of active Pol II and mature mRNA numbers predicted by two-state models with these effective parameters provide a very close match to the distributions predicted by the mechanistic model; nevertheless, the models can be distinguished by comparison of the shape of their waiting time distribution. The waiting time distribution for the two-state model has a non-zero value at $t = 0$ and decreases monotonically with time; whereas for the mechanistic model, the waiting time distribution is zero at $t = 0$ and has a peak at a non-zero value of time. We note that while in principle these two distributions are always distinguishable, in practice the differences will be small if the rate of Pol II binding and entry into the paused state is very large. We also showed that the necessary condition for the reduction of the mechanistic model to two-state models that was analytically derived using waiting time statistics i.e., Eq. (5.37), is compatible with the region of parameter space identified by model reduction using matching of moments and distributions of molecule numbers. We note that while our model description was framed in terms of an activator, it has alternative interpretations which increase its generality and applicability. One such alternative interpretation is in terms of a repressor that operates via competitive binding [245, 246]. In this interpretation U is a state that has a repressor bound to the promoter such that Pol II is blocked from being able to bind to the promoter. $U^\star$ then represents the state where the promoter is free and neither repressor nor Pol II is bound to the promoter, meaning that the binding site is accessible to both repressors and Pol II. Finally, the $U^{\star\star}$ state represents the state in which Pol II is recruited and proximally-paused.

A main distinction of this work from the analysis of a similar model studied in [170] is that the present mechanistic model has an explicit description of active Pol II that allows us to study the accuracy of the delay telegraph model. It is also noteworthy that while [170] showed that the telegraph model provided an excellent approximation to

the mature mRNA distribution of a similar mechanistic model under the assumption of rapid entry and exit from the paused state, in this study we showed using a variety of model reduction techniques that this assumption though sufficient is not necessary. We also note that while other papers have made use of waiting time statistics in the context of gene expression [47], our approach is distinctly different. The distribution of the off time in another three-state model of gene expression [47] is not the same as the distribution of the time between two consecutive active Pol II production events; this is because the former provides only information about the time between two successive bursts of gene expression which occurs on long timescales and reflects the accessibility of the promoter but has no information on the fast Pol II processes within each burst. To the best of our knowledge, the experimental measurement of the distribution of the waiting time as defined in this chapter has not been attempted yet. This is because with current labelling and imaging technology, it is not easy to directly visualise, track and quantify individual transcriptional initiation events. However, a set of recent papers report progress in this direction by estimating an approximate distribution between two consecutive initiation events in Drosophila using a machine-learning approach [247, 248].

We finish by a discussion of the validity and interpretation of Eqs. (5.38) which express the parameters of two-state models as a function of the parameters of the mechanistic model. We have shown from these expressions that different parameters of the two-state models can be effectively correlated due to their dependence on a common parameter/s of the mechanistic model. This may explain correlations found between parameters of two-state models estimated from single cell RNA sequencing for mammalian cells [46]. There is a region of parameter space where the effective parameters given by our theory become negative (when the inequality given by Eq. (5.37) is not satisfied), meaning that in this case there is no two-state model that can match the first three moments of the waiting time distribution of the mechanistic model; we also showed that if the elongation time and the mature mRNA degradation timescale are large enough, the aforementioned region is also characterised by Fano factors of active Pol II numbers and mature mRNA numbers that are less than one i.e., sub-Poissonian statistics. To see whether such a case is realistic we extensively searched through the experimental literature of gene expression, and found that for mature mRNA all papers report a Fano factor of greater than 1 which is consistent with constitutive or bursty expression; for nascent RNA, the majority of papers report Fano factors greater than 1 (see for example [45, 249, 217]) with the exception of one paper (see Supplementary Fig. 6 of [250]). However, it is to be borne in mind that while theoretically nascent RNA numbers should equal the active Pol II numbers in our model, in practice due to the intricacies of smFISH this is not the case, as we now explain. The number of nascent mRNA is most commonly calculated by dividing the total fluorescent signal from a transcription site by the fluorescence emitted by a mature transcript. In this technique, a fluorescent signal is emitted by

oligonucleotide probes bound to the nascent RNA tail. Since as an active Pol II travels along the gene, its nascent RNA tail grows, we expect the fluorescent signal intensity to increase as well [120]. Hence it follows that the total nascent mRNA n_N calculated using this method is generally a lower bound on the actual number of active Pol II n_A at a transcription site in the nucleus i.e., $n_N \sim f n_A$ where f is a fraction. From this it immediately follows that the Fano factor of nascent mRNA is always less than the Fano factor of active Pol II. Thus the measurement of Fano factors of nascent mRNA numbers slightly less than 1 in [250] likely implies Fano factors of active Pol II which are above 1. Hence we come to the conclusion that all available evidence to date for both nascent and mature mRNA seems consistent with Eq. (5.37), which implies that Eq. (5.38) provides a generally useful means to understand the parameters of two-state models in terms of underlying microscopic processes.

Chapter 6

Stochastic time-dependent enzyme kinetics: closed-form solution and transient bimodality

This chapter has been published as [4] entitled *Stochastic time-dependent enzyme kinetics: closed-form solution and transient bimodality* in the *Journal of Chemical Physics*. Slight modifications have been made for its inclusion in this thesis.

6.1 Abstract

We derive an approximate closed-form solution to the chemical master equation describing the Michaelis-Menten reaction mechanism of enzyme action. In particular, assuming that the probability of a complex dissociating into enzyme and substrate is significantly larger than the probability of a product formation event, we obtain expressions for the time-dependent marginal probability distributions of the number of substrate and enzyme molecules. For delta function initial conditions, we show that the substrate distribution is either unimodal at all times or else becomes bimodal at intermediate times. This transient bimodality, which has no deterministic counterpart, manifests when the initial number of substrate molecules is much larger than the total number of enzyme molecules and if the frequency of enzyme-substrate binding events is large enough. Furthermore, we show that our closed-form solution is different from the solution of the chemical master equation reduced by means of the widely used discrete stochastic Michaelis-Menten approximation, where the propensity for substrate decay has a hyperbolic dependence on the number of substrate molecules. The differences arise because the latter does not take into account enzyme number fluctuations while our approach includes them. We confirm by means of stochastic simulation of all the elementary reaction steps in the Michaelis-Menten mechanism that our closed-form solution is accurate over a larger region of parameter space than that obtained using the discrete stochastic Michaelis-Menten approximation.

6.2 Introduction

The mechanistic basis of the simplest single-enzyme, single-substrate reaction consists of a reversible step between an enzyme and a substrate, yielding the enzyme–substrate complex, which subsequently forms the product. This reaction is commonly called the Michaelis-Menten (MM) reaction [251, 51].

For over a century, the dynamics of this reaction have been extensively studied using deterministic rate equations. Because these equations do not admit an exact closed-form solution, various approximations have been devised to obtain insight into the underlying dynamics. Use of the quasi-equilibrium or quasi steady-state approximations lead to the famous Michaelis-Menten equation, an ordinary differential equation relating the rate of product formation and the substrate concentration (see [53] for a discussion of these approximations and their range of validity). This equation provides a simple means to extract the relevant kinetic parameters (the Michaelis-Menten constant and the maximum velocity) from experimental data. The Michaelis-Menten equation has also been solved exactly leading to explicit expressions for the time-evolution of the substrate (and product) concentration [52].

The stochastic formulation of enzyme kinetics, while not as much studied as its deterministic counterpart, has received increasing attention since the 1960s when the chemical master equation (CME) for the MM reaction mechanism was first derived and studied by Anthony F. Bartholomay [252]. The CME is a probabilistic discrete description of chemical reaction kinetics that is valid in well-mixed environments for point reacting particles [68, 74]. Its relevance lies in its ability to describe kinetics when the molecule numbers are low, conditions typical in intracellular environments, e.g., the median copy number per cell of most enzymes in *E. coli* is below a thousand [253]. Research efforts concerning the MM mechanism in the area of stochastic chemical kinetics can be, broadly speaking, categorised into three types: (i) The search for a solution of the CME for the MM reaction and its various extensions, i.e., obtaining a closed-form solution for the time-dependent or steady-state probability distribution of the molecule numbers of each species in the reaction system [58, 57, 56]. (ii) The reduction of the CME and the construction of the stochastic equivalent of deterministic approximations (such as the fast equilibrium, quasi steady-state and total quasi steady-state approximations) and understanding their regime of validity [143, 86, 254, 144, 154, 255, 84, 110, 156, 256, 145, 81, 257, 258]. (iii) The derivation of exact or approximate expressions for the mean of the stochastic rate of product formation and an investigation of the differences or similarities from the predictions of the deterministic Michaelis-Menten equation [259, 260, 261, 262, 263, 264, 265].

The majority of the literature has focused on (ii) and (iii). There are very few studies that focus on (i) principally because the CME is notoriously difficult to solve analytically [81]. In this chapter, we are interested in deriving new solutions of the CME for enzyme kinetic systems and hence next we briefly review the known solutions (see also [266] for a lengthier discussion). Arányi and Tóth [58] were the first to exactly solve the CME introduced by Bartholomay for the special case in which there is only one enzyme molecule with several substrate molecules in a closed compartment; in particular, they obtained an exact expression for the joint distribution of the number of substrate and enzyme molecules as a function of time (since the original paper is rather difficult to find, in Appendix D.1 we have reproduced the derivation in a concise manner). Another exact solution is reported in [57] by Schnoerr *et al.* who derive the exact steady-state solution for the CME describing the MM reaction system with one enzyme molecule and augmented with a substrate production reaction step (to model for example the production of substrate via translation). To our knowledge, there are no known exact solutions for the time-dependent probability distribution of the CME of the MM reaction system with multiple enzyme molecules. However, an approximate closed-form solution was derived by Dóka and Lente [56], using a so-called stochastic equivalent of the quasi steady-state approximation. Namely, they make an ansatz that the joint distribution of the number of substrate and enzyme molecules takes the form of a product of a time-dependent function and a constant value which characterises the state occupied by the system. Using this assumption and a number of heuristic arguments, the authors reduce the problem to a one-variable master equation which they then solve iteratively. However, one could argue that the derivation outlined in [56] lacks a certain degree of rigour and the analysis of the accuracy of the solution over time and parameter space is rather limited, which raises questions about the validity of the approximation.

In this chapter, our aims are to: (a) Derive an expression for the approximate time-dependent solution of the CME of the MM reaction system with multiple enzyme molecules under quasi-equilibrium conditions using an approach that is more rigorous and systematic than in previously published works. (b) Compare and contrast this solution with the solution of an often used reduced CME for the MM reaction in the literature. (c) Use the closed-form solution to identify interesting dynamical phenomena. We verify our approximate analytic results against the benchmark stochastic simulation algorithm (SSA) [68]. This chapter is divided as follows. In Section 6.3, we briefly review the main results known for deterministic enzyme kinetics, focusing in particular on the quasi-equilibrium approximation. In Sections 6.4.1 and 6.4.2, we introduce our method by first applying it to the MM reaction with a single enzyme molecule and subsequently to the case of multiple enzyme molecules. The method consists of three steps: (1) using a time scale separation method called averaging [113] to define groups of rapidly equilibrating states which then allows the derivation of a master equation describing

the Markovian dynamics of these groups on the slower time scale; (2) solving the resultant time-dependent, single variable master equation for the group dynamics using the method developed in [127] which has the advantage of bypassing the calculation of the eigenvectors of the transition matrix and hence considerably simplifies the analytical computations; (3) using the time-dependent solution describing the group dynamics to construct the marginal time-dependent distributions for both the numbers of substrate and enzyme molecules. We use the closed-form solution to find the regions of parameter space where transient bimodality of the distribution of substrate molecules occur. In Section 6.5, we show that our solution is accurate over a wider region of parameter space than the solution of a commonly used reduced master equation with a propensity that has the same hyperbolic dependence on the number of substrate molecules as the deterministic Michaelis-Menten equation (an approach popularised by Rao and Arkin [86]). In Section 6.6, we show that the same three-step method used in Sections 6.4.1 and 6.4.2, can be used to derive time-dependent distributions for multi-substrate enzyme reactions. We finish by discussing our results in Section 6.7.

6.3 Deterministic enzyme kinetics

Before progressing to stochastic enzyme kinetics we first briefly outline some of the main results known for deterministic enzyme kinetics. We consider the chemical reaction system:

$$S + E \underset{k_1}{\overset{k_0}{\rightleftharpoons}} C \xrightarrow{k_2} E + P, \tag{6.1}$$

where S denotes the substrate species, E denotes the enzyme species, C denotes the enzyme-substrate complex and P denotes the product. This system can be thought of as a reduction of the more biologically realistic set of reactions:

$$S + E \rightleftharpoons ES \rightarrow EP \rightarrow E + P, \tag{6.2}$$

where the unbinding of the product from the enzyme is very fast. For simplicity, we assume the initial condition for this system is that all enzymes are unbound to the substrate. There are two conservation laws for this system: $[E] + [C] = [E]_0$ and $[S] + [C] + [P] = [S]_0$, where $[i]$ denotes the concentration of species i and $[i]_0$ the initial concentration of species i. Assuming well-mixed conditions and the law of mass action, the deterministic dynamics of the reaction system in Eq. (6.1) are described by a set of coupled ordinary differential equations (commonly called the rate equations) describing

the time-evolution of the substrate and complex concentrations:

$$\begin{aligned}\frac{d[S(t)]}{dt} &= -k_0[S(t)]([E]_0 - [C(t)]) + k_1[C(t)],\\ \frac{d[C(t)]}{dt} &= -(k_1 + k_2)[C(t)] + k_0[S(t)]([E]_0 - [C(t)]).\end{aligned} \tag{6.3}$$

Note that the time-dependent concentrations of E and P can be straightforwardly obtained from the time-dependent solutions of C and S by means of the conservation laws previously stated. Although seemingly simple, the rate equations given by Eq. (6.3) are not easy to solve analytically for the time-dependent analytic solution, and as such one is limited to finding approximate solutions. Two of the most common approximations used in the literature are the (i) *quasi steady-state assumption* (QSSA) and (ii) the *quasi-equilibrium approximation* (QEA), also called the rapid equilibrium approximation or the reverse quasi steady-state assumption. The QSSA, derived by Briggs and Haldane [267], assumes that after a short transient, the concentration of the complex (and enzyme) is in a quasi steady-state (with regard to the substrate and product); thus under the QSSA, it is assumed that $d[C(t)]/dt \approx 0$. See [268] for a detailed discussion of this approximation and for its range of validity. On the other hand, the QEA assumes that substrate binding and dissociation occur much more rapidly than product formation such that the substrate, enzyme and complex are approximately in equilibrium. Thus under the QEA, it is assumed that $d[S(t)]/dt \approx 0$; this approximation, popularised by Michaelis and Menten [251], is commonly used in the analysis of various biochemical models [269].

Enforcing either the QSSA or QEA leads to the following effective rate equation describing the time-evolution of the substrate concentration:

$$\frac{d[S(t)]}{dt} = \frac{-V_{\max}[S(t)]}{k + [S(t)]}, \tag{6.4}$$

where $V_{\max} = k_2[E]_0$, $k = (k_1 + k_2)/k_0$ if the QSSA is used, $k = k_1/k_0$ if the QEA is used, and where the conservation law $[S] + [P] = [S]_0$ holds. Note that a necessary limitation of Eq. (6.4) is that we have assumed that $[S] + [C] \approx [S]$, which is true in the limit $[S]_0/[E]_0 \gg 1$. Eq. (6.4) has been solved perturbatively in a number of studies, all of which also assessed the validity of the QSSA [270, 268]. An exact solution was reported in [52] which is given by:

$$\langle n(t)\rangle_a = \Omega[S(t)] = \Omega k\, W\left(\frac{[S]_0}{k}\exp\left(\frac{-V_{\max}t + [S]_0}{k}\right)\right), \tag{6.5}$$

where $\langle n(t)\rangle_a$ gives the (deterministic) *number* of bound and unbound substrate molecules obtained in the limit $[S]_0/[E]_0 \gg 1$ at time t, Ω is the volume of the system, and $W(\cdot)$ is the principal branch of the Lambert W function (also known as the Omega function). Note that within van Kampen's *system size expansion* [8] for monostable systems, the rate equations are obtained as the macroscopic limit of the stochastic description of a well mixed chemical system; within this formalism, the concentration of a species i multiplied by the volume is the same as the mean number of molecules of species i. Hence in our case $\langle n(t)\rangle_a$ can also be interpreted as the *mean* number of substrate molecules in the macroscopic limit. In the rest of this chapter, we study the stochastic equivalent of the QEA and thus we shall use $k = k_1/k_0$.

6.4 Stochastic QEA analysis

6.4.1 Single enzyme

For simplicity, we first illustrate the method by solving the enzyme system described in Eq. (6.1) for the case of one enzyme molecule with initially N substrate molecules. Since there are no birth-death processes coupled to any species, the conservation equations $n_E + n_C = 1$ and $n + n_C + n_P = N$ hold, where n denotes the number of substrate molecules and all other n_i denote the number of species i.

We label the microstate of the reaction network in Eq. (6.1) as (n, n_E), which fully specifies the state of the system due to the conservation laws stated previously. The possible transitions between all of the discrete microstates of this system are illustrated in Fig. 6.1(i): the system starts from the state $(N, 1)$ and eventually ends up in the state $(0, 1)$. Our goal now will be to find the marginal probability distribution $P(n;t)$, i.e., the probability of observing n substrate molecules at a time t.

Assuming Markovian dynamics [81], it follows that the time-evolution of $P(n, n_E;t)$ (the probability of observing n substrate molecules and n_E enzyme molecules at a time t) is given by the CME:

$$\begin{aligned}\frac{\partial P(n,n_E;t)}{\partial t} = \quad & k_0(n+1)(n_E+1)P(n+1,n_E+1;t) \\ & + (2-n_E)\left(k_1 P(n-1,n_E-1;t) + k_2 P(n,n_E-1;t)\right) \\ & - \left(k_0\, n\, n_E + (1-n_E)(k_1+k_2)\right) P(n,n_E;t).\end{aligned} \tag{6.6}$$

Note that this CME is valid only for a single enzyme system, i.e., $n_E \in \{0,1\}$. Furthermore note that the bimolecular propensity is inversely proportional to the volume Ω but for simplicity we set $\Omega = 1$ (a convention throughout the manuscript). The standard approach involves introducing the time-dependent marginal generating

functions $G_{n_E}(z;t) = \sum_n z^n P(n, n_E; t)$ and attempting to solve the generating function partial differential equations, e.g., using eigenfunction methods (see Section 2.6.1 and [8, 74]). However, this standard method quickly leads one to mathematical difficulty. An analytic solution only presents itself in a non-cumbersome form where one assumes the initial state contains a single substrate molecule [58]. In Appendix D.1 we summarise the single enzyme solution provided by [58], and its complexity even in the single substrate molecule case motivates the analysis we present below.

We take a different approach. We first simplify the problem through the use of averaging [113, 115, 271]. Specifically the procedure lumps together microstates equilibrating on a fast timescale in groups which then allows one to write a master equation describing the dynamics of the groups on the slow timescale. We shall assume that the slow timescale is that associated with product formation, i.e., k_2 is sufficiently small (we will be more precise what this really means later) and hence the averaging procedure is in the same spirit as the QEA discussed in Section 6.3. Note that this time scale separation is justified in the literature, see [272].

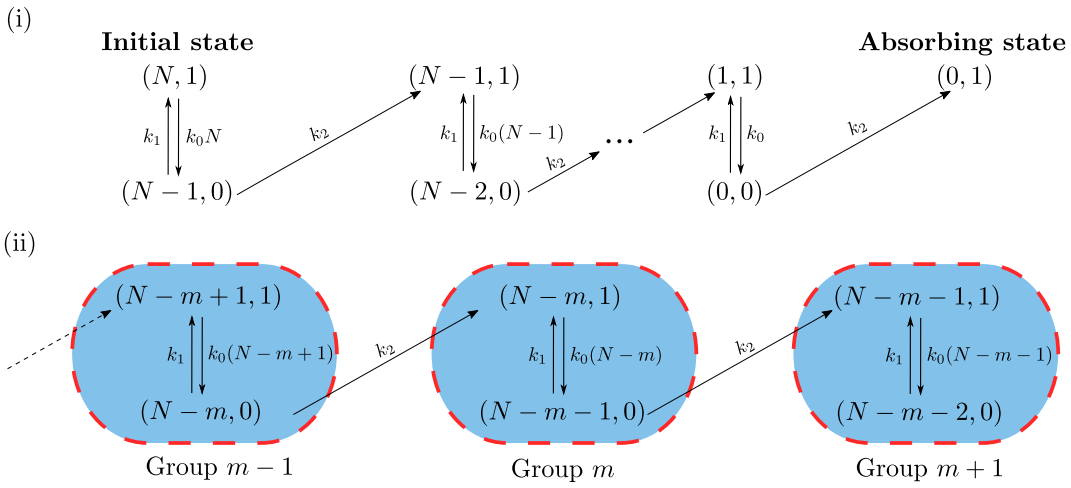


Figure 6.1: Illustration of the enzymatic system described by a single enzyme and N initial substrate molecules. (i) Markovian dynamics of the enzyme kinetic system described by a single enzyme. The initial condition for the system is $(N, 1)$, and as $t \to \infty$ the microstate of the system is guaranteed to be that of the absorbing state $(0, 1)$, with no remaining substrate and one free enzyme. (ii) Markovian dynamics in the reduced model, where processes occurring in a group are assumed to be much faster than the interactions between the groups themselves. The label 'group m' denotes the set of microstates that exist when m product molecules have been formed, given that there are no product molecules initially; hence, it is easily seen that there are $N + 1$ groups in total with labels $m = \{0, 1, 2, ..., N-1, N\}$.

Since k_2 is small, it follows that we can group all microstates that are in rapid equilibrium with each other (due to the fast processes of binding and unbinding of substrate from the enzyme) as shown in Fig. 6.1(ii); group m is then the set of microstates of the system accessible when m product molecules have been produced. We define $p^g_m(t)$ as the probability to be in group m at a time t, and $p^{qe}_{i,m}$ as the probability of having i free enzymes for the *reduced* system given by considering only reactions among microstates

in group m. Once these probabilities are found, we can construct $P(n;t)$, based on the fact that there are two microstates that contain n substrate molecules: $(n,0)$ and $(n,1)$ associated with groups $N-(n+1)$ and $N-n$ respectively. This means that under the stochastic QEA:

$$P(n;t) = p^g_{N-n}(t)p^{qe}_{1,N-n} + p^g_{N-(n+1)}(t)p^{qe}_{0,N-(n+1)}. \tag{6.7}$$

In the case of the single enzyme system studied in this section, the quasi-equilibrium probabilities are trivial (since there are only two microstates in each group) and are given by:

$$p^{qe}_{1,N-n} = \frac{k_1}{k_1 + k_0 n} \quad \text{and} \quad p^{qe}_{0,N-(n+1)} = \frac{k_0(n+1)}{k_1 + k_0(n+1)}. \tag{6.8}$$

All that remains is the task of finding $p^g_m(t)$. To do this we first write the master equation for the transitions between groups. Rescaling time as $t' = k_2 t$ and making use of the previous definition, $k = k_1/k_0$, the master equation for the groups is:

$$\partial_{t'} p^g_m(t') = a_m p^g_{m-1}(t') - a_{m+1} p^g_m(t'), \tag{6.9}$$

where:

$$a_m = \frac{N-(m-1)}{k+N-(m-1)},\ 1 \leq m \leq N+1, \tag{6.10}$$

and $a_{i\leq 0} = 0$. Note that a_m is the probability of the jump from group $m-1$ to group m in a unit interval of rescaled time. From Fig. 6.1 the probability of the jump from group $m-1$ to group m in a unit interval of normal time is equal to k_2 multiplied by the probability of being in the microstate $(N-m,0)$ which under the rapid equilibrium assumption is $k_0(N-m+1)/(k_1+k_0(N-(m-1)))$. Due to time rescaling, the factor of k_2 disappears and hence follows Eq. ((6.10)).

Since there are $N+1$ groups in total, Eq. (6.9) corresponds to a system of $N+1$ ODEs which can be concisely written as the matrix equation:

$$\partial_{t'} \underline{p}^g(t') = \mathcal{Q} \cdot \underline{p}^g(t'), \tag{6.11}$$

where $\underline{p}^g(t')$ is a $N+1$ element column vector defined by $\underline{p}^g(t') = (p_0^g(t'), p_1^g(t'), ..., p_N^g(t'))$ and $\mathcal{Q}$ is a $(N+1) \times (N+1)$ lower-bidiagonal square matrix defined by:

$$\mathcal{Q} = \begin{pmatrix} -a_1 & & & & \\ a_1 & -a_2 & & & \\ & a_2 & -a_3 & & \\ & & \ddots & \ddots & \\ & & & a_N & -a_{N+1} \end{pmatrix}. \tag{6.12}$$

As we will describe below, we solve the set of ODEs given by Eq. (6.11) using the method described in [127] which provides an *exact* time-dependent solution for any one-variable one-step master equation with finite number of microstates *as long as one can find the eigenvalues of the transition rate matrix exactly.* In our case, the eigenvalues of $\mathcal{Q}$ are trivial, since $\mathcal{Q}$ is lower-bidiagonal, and they are given by the diagonal elements. Hence the eigenvalues of $\mathcal{Q}$ are given by $\lambda_i = -a_i$, $1 \leq i \leq N+1$. Note that $\lambda_{N+1} = 0$ and is the largest eigenvalue, with all $\lambda_{1 \leq i \leq N} < 0$.

We now proceed to use these eigenvalues to find the time-dependent solution to Eq. (6.11). The solution to this set of ODEs is formally given by:

$$\underline{p}^g(t') = \exp\left(\mathcal{Q}t'\right) \cdot \underline{p}^g(0), \tag{6.13}$$

where $\exp(\mathcal{Q}t')$ is defined as a matrix exponential. For a general master equation, this matrix exponential is typically difficult to deal with, however in our case $\mathcal{Q}$ is lower-bidiagonal and hence we can proceed via the method of [127]. We first consider Cauchy's integral formula for matrices, explicitly given by [130]:

$$\underline{f}(\mathcal{Q}) = \frac{1}{2\pi i} \oint_C (zI - \mathcal{Q})^{-1} \cdot \underline{f}(z) dz, \tag{6.14}$$

where C is a closed contour in the complex plane that encloses all the eigenvalues of $\mathcal{Q}$ and I is the identity matrix. Taking $\underline{f}(z) = e^{zt'}\underline{p}^g(0)$ we then arrive at:

$$\underline{p}^g(t') = \frac{1}{2\pi i} \oint_C (zI - \mathcal{Q})^{-1} \cdot \underline{p}^g(0) e^{zt'} dz. \tag{6.15}$$

A typical initial condition is $p_m^g(0) = \delta_{m,0}$, meaning that we always start in group 0 which contains the microstates $(N, 1)$ and $(N-1, 0)$, as is shown in Fig. 6.1(i). Note that $\delta_{i,j}$ is the Kronecker delta. Using this initial condition, Eq. (6.15) becomes:

$$p_m^g(t') = \frac{1}{2\pi i} \oint_C (zI - \mathcal{Q})^{-1}_{m+1,1} e^{zt'} dz. \tag{6.16}$$

We show at the end of this section how to extend the time-dependent solution for a general initial distribution. Since it is bidiagonal, the inverse of $zI - \mathcal{Q}$ can easily be found via Cramer's rule [273]:

$$(zI - \mathcal{Q})^{-1}_{ij} = \begin{cases} 0, & i < j, \\ \frac{1}{a_i + z}, & i = j, \\ \frac{1}{a_j + z} \prod_{k=j+1}^{i} \frac{a_{k-1}}{a_k + z}, & i > j. \end{cases} \tag{6.17}$$

Substituting this into Eq. (6.16) then gives us:

$$p_m^g(t') = \begin{cases} 0, & m < 0, \\ \frac{1}{2\pi i} \oint_C \frac{e^{zt'}}{z - \lambda_1} dz, & m = 0, \\ \frac{1}{2\pi i} \left\{(-1)^m \prod_{k=1}^{m} \lambda_k\right\} \times \left\{ \oint_C \frac{e^{zt'}}{\prod_{k=1}^{m+1} (z - \lambda_k)} dz \right\}, & m > 0, \end{cases} \tag{6.18}$$

where we have utilised the relation $\lambda_i = -a_i$. These integrals can then be evaluated using Cauchy's residue theorem [274], explicitly stated as:

$$\oint_C f(z) dz = 2\pi i \sum_k \mathrm{Res}(f(z), z_k), \tag{6.19}$$

where the values $z = z_k$ are poles of $f(z)$ within C and the residues are $\mathrm{Res}(f(z), z_k) = \lim_{z \to z_k} (z - z_k) f(z)$ for the simple poles in Eq. (6.18). Note that the poles of the complex integrals in Eq. (6.18) are the eigenvalues of $\mathcal{Q}$. Therefore, from Eq. (6.18) we finally get an expression for $p_m^g(t')$ as:

$$p_m^g(t') = \begin{cases} 0, & m < 0, \\ e^{\lambda_1 t'}, & m = 0, \\ \left\{(-1)^m \prod_{k=1}^{m} \lambda_k\right\} \times \left\{ \sum_{k=1}^{m+1} \frac{e^{\lambda_k t'}}{\prod_{j=1, j \neq k}^{m+1} (\lambda_k - \lambda_j)} \right\}, & m > 0. \end{cases} \tag{6.20}$$

Hence the time-dependent probability distribution $P(n; t)$ is given by Eq. (6.7) together with Eqs. (6.8) and (6.20). The extension to a more general initial distribution is then relatively simple. Consider some initial distribution $\underline{p}^g(0) = \underline{q}$, where $\underline{q}$ is an $N + 1$ element vector; the time-dependent group probability $p^g_{m|\underline{q}}(t')$ is then given by the

weighted sum:

$$p^g_{m|\underline{q}}(t') = \sum_{j=0}^{N} p_{m|q_j}(t') q_j. \tag{6.21}$$

This initial condition could be useful to model variation in the initial number of substrate molecules due to uncertainty introduced by experimental error or else due to the intrinsic noise in the reaction mechanism generating the substrate. Note that if $q_m = \delta_{m,0}$, one clearly recovers the analysis shown above. For the rest of this chapter we only consider the initial condition $p^g_m(0) = \delta_{m,0}$, *specifically where all enzymes are initially unbound to the substrate and where there are initially zero product molecules*, but note that the analysis that follows can be easily extended for more general initial distributions.

In the beginning of this derivation, we stated that the main assumption is that k_2 is sufficiently small. This statement can be made more precise as follows. From Fig. 6.1(ii) it is clear that the exit from group m can only occur when the enzyme is bound to substrate, i.e., from state $(N-m-1, 0)$. Now given that we are in this state, it follows that only two reactions can occur: either a reaction which causes a group change, i.e., $(N-m-1,0) \to (N-m-1,1)$ which occurs with rate k_2 or a reaction that leads to no group change, i.e., $(N-m-1,0) \to (N-m,1)$ which occurs with rate k_1. Hence the probability of leaving the group is $k_2/(k_1+k_2)$, from which follows that the microstates in each group will achieve quasi-equilibrium if $k_2 \ll k_1$. Therefore, this is the condition under which our method provides a good approximation to the distribution of substrate molecules at all times.

We test the distributions predicted by Eq. (6.7) against the SSA in Fig. 6.2A(i-iii) and Fig. 6.2B(i-iii). In Fig. 6.2A(i-iii) we show that the solution is accurate for small $N = 8$, over a time range from $t' = 1$ near the initial condition, to $t' = 12$ close to the absorbing state, where the validity criterion $k_1 \gg k_2$ holds. In Fig. 6.2B(i-iii) we observe that our solution agrees similarly well to the SSA for larger values of N. For a more general comparison of the exact solution to SSA through time, we can compute the mean and standard deviation from Eq. (6.7):

$$\langle n(t') \rangle = \sum_{n=0}^{N} n P(n; t'), \tag{6.22}$$

$$\sigma(t') = \sqrt{\left(\sum_{n=0}^{N} n^2 P(n; t') \right) - \langle n(t') \rangle^2}. \tag{6.23}$$

The stochastic mean number of substrate $\langle n \rangle$ can then be compared to the deterministic mean number $\langle n \rangle_d$ obtained from the rate equations. That is, we numerically solve Eq. (6.3) for $[S(t)]$ with $k_2 = 1$, noting that $\langle n \rangle_d = [S(t)]$ as we have previously set $\Omega = 1$.

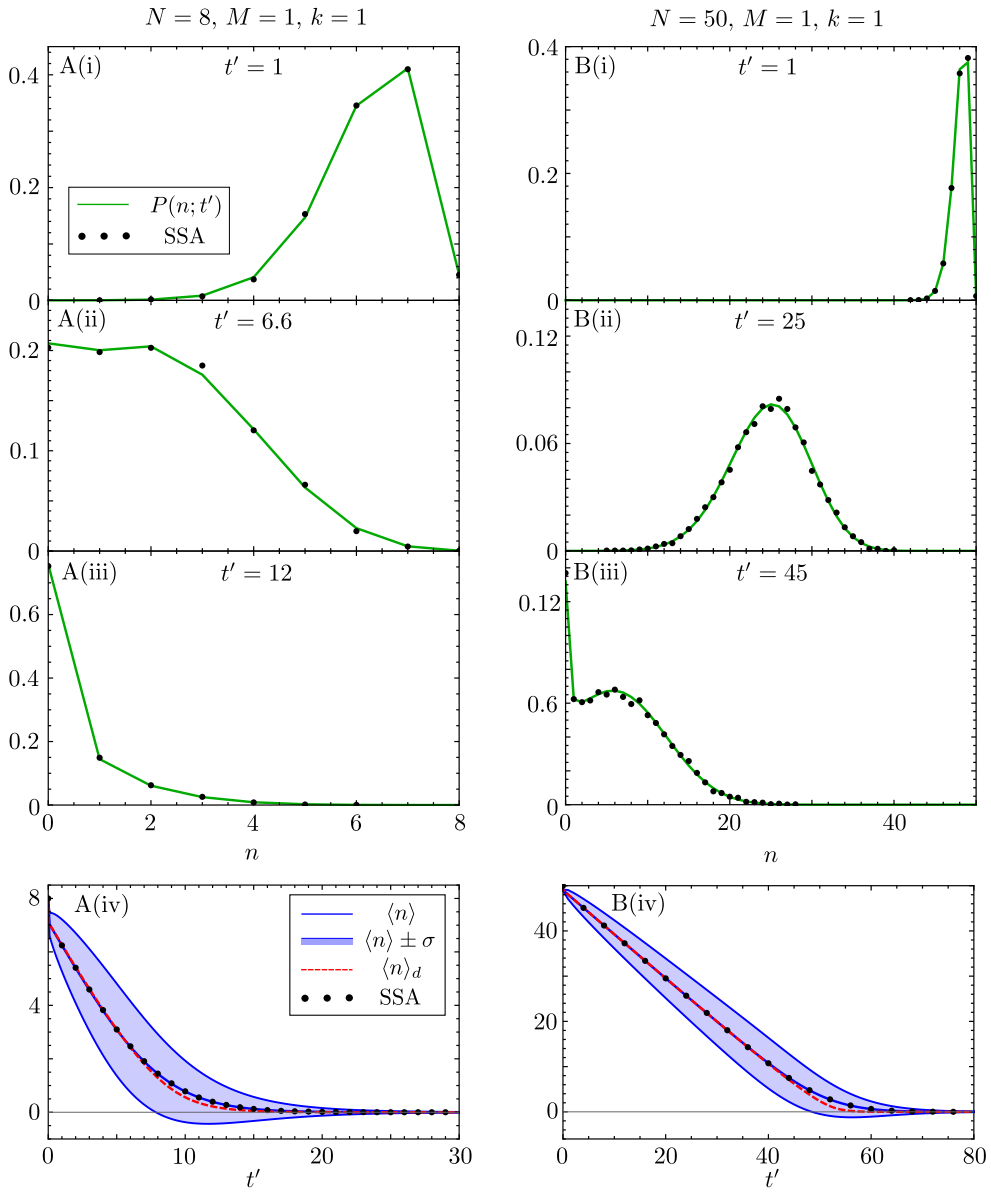


Figure 6.2: Comparison of the analytic time-dependent probability distribution of substrate molecules for the enzyme reaction in (6.1) with one enzyme molecule, i.e., $M = 1$, and N initial substrate molecules to the distribution obtained from the SSA [68]. Note that the analytic solution is given by Eq. (6.7) together with Eqs. (6.8) and (6.20). In all cases we enforce $k_1/k_2 \gg 1$ such that the quasi-equilibrium assumption behind the QEA is justified. We show the time-evolution of the distribution for substrate numbers, from near the initial condition to near the absorbing state, in two cases: A(i-iii) is for $N = 8$, $k_0 = k_1 = 10^3, k_2 = 1$, meaning that $k = k_1/k_0 = 1$. B(i-iii) is for $N = 50$, and all rate parameters as in the previous case. Note that the analytical solution (green lines) matches the SSA (black dots) for all times, for both a small and large initial number of substrate molecules. In A(iv) and B(iv) we show the corresponding plots of the time-evolution of the mean $\langle n \rangle$ and of the standard deviation σ of the distributions of substrate molecules, as predicted by our theory; these are compared with the mean calculated from the SSA and the mean $\langle n \rangle_d$ obtained from the numerical solution of the deterministic rate equations given by Eq. (6.3). Note that the deterministic mean is a better approximation to the stochastic mean for larger N. As shown in B(iii), and mildly in A(ii), the distribution can be bimodal at intermediate times. Each SSA probability distribution is constructed from 10^5 individual reaction trajectories.

In Figs. 6.2A(iv) and 6.2B(iv) we plot the evolution of the stochastic and deterministic mean substrate numbers in time, and compare them to the SSA for parameters sets $N = 8$, $k = 1$ and $N = 50$, $k = 1$ respectively. We also show the standard deviation about the mean, i.e., $\langle n \rangle \pm \sigma$, where we have dropped the time-dependence for brevity, given in the blue envelope. Clearly, $\langle n \rangle$ from Eq. (6.22) matches the mean predicted by the SSA for most times, aside from the initial condition at $t' = 0$, where a step-like drop is observed in the mean predicted by the SSA to the value predicted by the quasi-equilibrium analysis. This step-like drop to the quasi-equilibrium value of the mean is known as the *initial transient*, and is seen in more detail in Appendix D.2. The explanation of the initial transient follows by considering the system after quasi-equilibrium has been reached between the two microstates in group 0, $(N, 1)$ and $(N - 1, 0)$, after a time $t'_c \approx 1/\min\{k_0 N, k_1\}$ which is small under the rapid equilibrium assumption. Because of the discreteness of the substrate molecules, $\langle n \rangle$ after a time $t'_c \ll 1$ becomes an average over $n = N$ and $n = N - 1$ weighted by the quasi steady-state probabilities $p^{qe}_{1,0}$ and $p^{qe}_{0,0}$ respectively, hence the step-like drop in the mean predicted by the SSA at $t' = t'_c$. The method of averaging in the stochastic QEA assumes the immediate occurrence of the equilibrium in group 0 at $t' = 0$, hence the dispatch of $\langle n(t' = 0) \rangle$ from the exact initial condition. This also explains why the standard deviation predicted by the stochastic QEA (notably in Fig. 6.2A(iv)) appears to be non-zero at $t' = 0$: *since the system is modelled to be in quasi-equilibrium at initiation, equilibrium fluctuations are present even at* $t' = 0$. However, so long as the SSA parameters are chosen such that $k_1/k_2 \gg 1$ the stochastic QEA provides a very good approximation even for small times so long as $t' > t'_c$. Additionally, we compare $\langle n \rangle$ to the deterministic mean number of free substrate, $\langle n \rangle_d$, predicted the numerical solution of Eq. (6.3). Overall, the deterministic solution is found to be in good agreement with the mean predicted by the SSA and the stochastic QEA, however there does exist a small disagreement where the mean number of substrate molecules is small (seen more explicitly in Fig. 6.2B(iv)). This disagreement occurs since molecular discreteness is very important where $\langle n \rangle$ is very small, and properly accounting for it leads to differing dynamics for $\langle n \rangle$ in this region, whereas the behaviour of $\langle n \rangle_d$ does not change compared to $\langle n \rangle_d \gtrapprox 1$, since the deterministic analysis considers molecule number to be continuous. As we shall see later, increasing the number of enzyme molecules removes this discrepancy between the stochastic and deterministic means, highlighting that the discrepancy seen here is because we do not consider enzyme molecules to be discrete in the deterministic analysis.

From Fig. 6.2B(iii) we observe that the distribution of substrate molecule numbers can be bimodal at intermediate times (there are two peaks at $n = 0$ and $n = 6$ at $t' = 45$). This bimodality, though less conspicuous, can in fact be also observed in Fig. 6.2A(ii) with peaks at $n = 0$ and $n = 2$. From Fig. 6.2A(iv) and B(iv), we can see that in both cases the bimodality occurs at a time t' when $\langle n \rangle - \sigma \approx 0$, i.e., when the

fluctuations are large enough to cause frequent transitions to the absorbing state. This type of dynamical phase transition (which we shall refer to as transient bimodality), from a unimodal distribution to a bimodal one and then back to a unimodal one, as time progresses, has also been recently observed in genetic feedback loops [115] and is known in non-biological systems [275, 276]. We will discuss this phenomenon more extensively in later sections.

6.4.2 Multiple enzymes

We now extend the solution to the enzyme system (6.1) to the case where initially there are N free substrate molecules and M free enzyme molecules with the constraint of substrate abundance, i.e., $N \geq M$. Note that the solution to the system with $M \geq N$ follows as a special case of the $N \geq M$ system, discussed at the end of this section.

We proceed in solving this system as we did in the single enzyme case: assuming k_2 is sufficiently small, we group the microstates governed by the fast processes together to form $N+1$ groups between which the transitions are significantly slower than those between the fast internal states of an individual group. The Markov chain describing the system split into groups is shown in Fig. 6.3. Our task is then to find (i) the equilibrium probabilities $p^{qe}_{i,m}$ of being in each fast internal state i (considering only the reactions between the internal states in group m) and (ii) to find the time-dependent probability $p^{g}_{m}(t)$ of being in group m. Knowledge of both (i) and (ii) will allow us to approximate the distribution of interest, $P(n;t)$.

We begin by finding the probabilities $p^{qe}_{i,m}$ and redefine it for the case of multiple enzyme: $p^{qe}_{i,m}$ is the equilibrium probability of having $M-i$ free enzymes in the case of a reduced system involving only the reactions among the fast internal states contained in group m. Now, finding $p^{qe}_{i,m}$ for any group $0 \leq m < (N-1)$ is more complicated than was the case for a single enzyme system, since there we had only two fast internal states in each group. To proceed we consider the following Markovian dynamics of a system with $L+1$ possible microstates:

$$0 \underset{k_{1,0}}{\overset{k_{0,1}}{\rightleftharpoons}} 1 \underset{k_{2,1}}{\overset{k_{1,2}}{\rightleftharpoons}} ... \underset{k_{L,L-1}}{\overset{k_{L-1,L}}{\rightleftharpoons}} L. \tag{6.24}$$

One can then write the master equation for this dynamical system in matrix form:

$$\partial_t \underline{P}_t = \mathcal{M} \cdot \underline{P}_t, \tag{6.25}$$

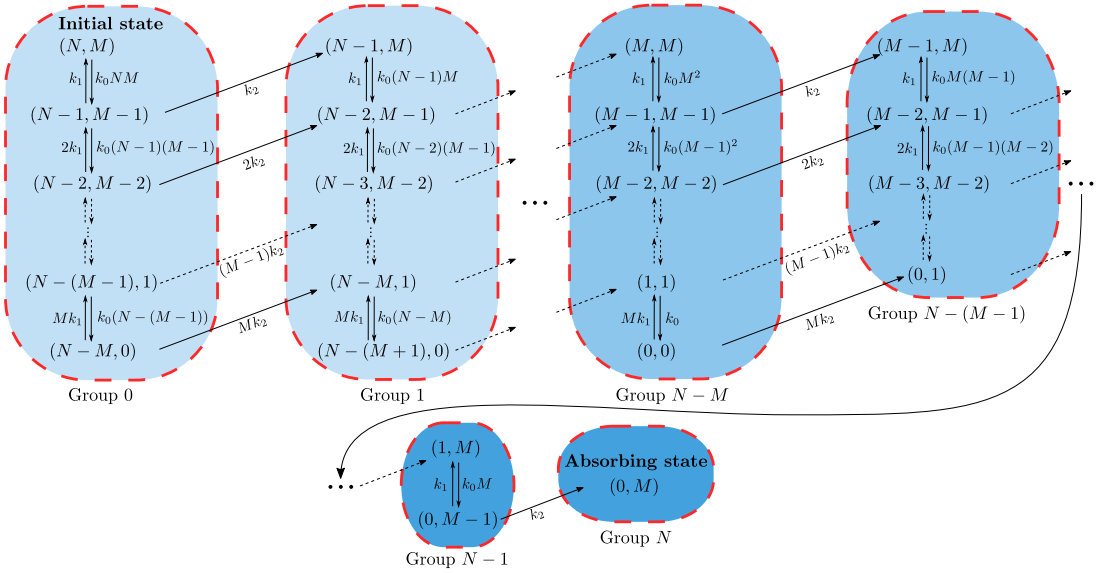


Figure 6.3: Illustration showing the transitions between the discrete microstates of the enzyme system (6.1) with initially M enzymes and N substrate molecules where $N \geq M$. Microstates that are in rapid equilibrium with each other are aggregated together, with each set of such fast internal states corresponding to a *group*. Namely, the label 'group m' denotes the set of microstates that exist when there are m product molecules, given that there are no product molecules initially. Groups $0 \leq m \leq N - M$ have $M + 1$ fast internal states, whereas groups $N - M < m \leq N$ have $N - m + 1$ fast internal states. Note that as $t \to \infty$ we are guaranteed to be in the absorbing state $(0, M)$.

where $\underline{P}_t = (P_t(0), P_t(1), ..., P_t(L))$, $P_t(i)$ is the probability of being in microstate i at time t and

$$\mathcal{M} = \begin{pmatrix} -k_{0,1} & k_{1,0} & & & & \\ k_{0,1} & -(k_{1,0} + k_{1,2}) & k_{2,1} & & & \\ & k_{1,2} & -(k_{2,1} + k_{2,3}) & k_{3,2} & & \\ & & \ddots & \ddots & \ddots & \\ & & & & k_{L-1,L} & -k_{L,L-1} \end{pmatrix}. \tag{6.26}$$

Enforcing the quasi-equilibrium condition, $\partial_t(\cdot) = 0$, converts the system of $L+1$ ODEs in Eq. (6.25) into a system of $L+1$ simultaneous equations in the equilibrium microstate probabilities $P(i)$, given by $\mathcal{M} \cdot \underline{P} = 0$. One can explicitly solve this set of simultaneous equations under the constraint $\sum_i P(i) = 1$, yielding the probabilities:

$$P(i) = \frac{\left(\prod_{j=1}^{i} k_{j-1,j}\right) \times \left(\prod_{j=i+1}^{L} k_{j,j-1}\right)}{\sum_{i=0}^{L} \left[\left(\prod_{j=1}^{i} k_{j-1,j}\right) \times \left(\prod_{j=i+1}^{L} k_{j,j-1}\right)\right]}. \tag{6.27}$$

Note that due to the definition of the empty product being equal to 1, when we have either $i = 1$ or $i = L$ the numerator of Eq. (6.27) is equal to 1. Further note that one could also utilise the King-Altman method [277, 278] to arrive at Eq. (6.27). Using this result we can find the quasi-equilibrium probabilities for each group shown in Fig. 6.3. First, we consider the groups $0 \leq m \leq N - M$, each with $M + 1$ fast internal states as these groups contain more (or the same number) free substrate molecules than enzymes. Taking the specific example of group $m = 0$, we see that we have a total of $M + 1$ microstates, i.e., $L = M$, $k_{j-1,j} = k_0(N - (j-1))(M - (j-1))$ and $k_{j,j-1} = jk_1$, with $1 \leq j \leq M$. Identifying $p^{qe}_{i,0}$ with $P(i)$ in Eq. (6.27), we find that:

$$p^{qe}_{i,0} = \frac{k_0^i k_1^{M-i} \left\{\prod_{j=1}^{i}(N-(j-1))(M-(j-1))\right\} \times \left\{\prod_{j=i+1}^{M} j\right\}}{\sum_{i=0}^{M}\left[k_0^i k_1^{M-i} \left\{\prod_{j=1}^{i}(N-(j-1))(M-(j-1))\right\} \times \left\{\prod_{j=i+1}^{M} j\right\}\right]} \,. \tag{6.28}$$

The result can be easily generalised for groups $0 \leq m \leq N - M$ and $0 \leq i \leq M$:

$$p^{qe}_{i,m} = \frac{k^{-i} \left\{\prod_{j=1}^{i}((N-m)-(j-1))(M-(j-1))\right\} \times \left\{\prod_{j=i+1}^{M} j\right\}}{\sum_{i=0}^{M}\left[k^{-i} \left\{\prod_{j=1}^{i}((N-m)-(j-1))(M-(j-1))\right\} \times \left\{\prod_{j=i+1}^{M} j\right\}\right]} \,, \tag{6.29}$$

where we have re-introduced $k = k_1/k_0$. The dynamics of groups $N - M < m \leq N$ are slightly different as they contain fewer substrate molecules than enzymes. These groups correspondingly have $N - m + 1$ fast internal states, i.e., $0 \leq i \leq N - m$. This leads to quasi-equilibrium probabilities of the form:

$$p^{qe}_{i,m} = \frac{k^{-i} \left\{\prod_{j=1}^{i}((N-m)-(j-1))(M-(j-1))\right\} \times \left\{\prod_{j=i+1}^{N-m} j\right\}}{\sum_{i=0}^{N-m}\left[k^{-i} \left\{\prod_{j=1}^{i}((N-m)-(j-1))(M-(j-1))\right\} \times \left\{\prod_{j=i+1}^{N-m} j\right\}\right]} \,, \tag{6.30}$$

Finally, by defining

$$g(m) = \Theta(m - (N - M)) \times (m - (N - M)) \,, \tag{6.31}$$

where $\Theta(m - (N - M))$ is the Heaviside step function, we can write down a joint expression for all groups $0 \leq m \leq N$ and $0 \leq i \leq M - g(m)$:

$$p^{qe}_{i,m} = \frac{z_{i,m}}{\mathcal{Z}_m}, \tag{6.32}$$

with

$$z_{i,m} = k^{-i}\left\{\prod_{j=1}^{i}((N-m)-(j-1))(M-(j-1))\right\} \times \left\{\prod_{j=i+1}^{M-g(m)} j\right\}, \tag{6.33}$$

$$\mathcal{Z}_m = \sum_{i=0}^{M-g(m)} z_{i,m}. \tag{6.34}$$

We now proceed to calculate $p_m^g(t)$. From Fig. 6.3, we observe that the transitions between the groups are described by the master equation identical in form to Eq. (6.9). However, the transition rates a_m in this case are different, as the group m can be reached from any of the $M - g(m-1)$ microstates in the group $m-1$ (excluding only the microstate with M free enzymes) and we must also take into account the quasi-equilibrium probabilities of being in the corresponding microstate. It follows that the transition rates can be defined as:

$$a_m = \sum_{n=1}^{M-g(m-1)} np_{n,m-1}^{qe} = -k\partial_k(\ln(\mathcal{Z}_{m-1})),\ 1 \le m \le N+1, \tag{6.35}$$

$$= \begin{cases} -M \times \left(\frac{k\ {}_1F_1(1-M,-m-M+N+3;-k)}{(-m-M+N+2)\ {}_1F_1(-M,-m-M+N+2;-k)} - 1\right), & m \le N-M+1, \\ -(N-m+1) \times \left(\frac{k\ {}_1F_1(m-N,m+M-N+1;-k)}{(m+M-N)\ {}_1F_1(m-N-1,m+M-N;-k)} - 1\right), & m > N-M+1, \end{cases}$$

where ${}_1F_1(a,b;c)$ is the confluent hypergeometric function, a result which we prove in Appendix D.3. As the dynamics between the groups are identical to the single enzyme case, $p_m^g(t')$ has exactly the same form as Eq. (6.20) but with the eigenvalues of $\mathcal{Q}$ being given by $\lambda_i = -a_i$, where the a_i are now defined in Eq. (6.35).

We can now obtain the probability distribution $P(n;t)$, which requires us to find all microstates in the system containing n free substrate molecules. From Fig. 6.3 we see that for substrate numbers n, where $0 \le n \le N-M$, there are $M+1$ corresponding microstates given by $(n,0),(n,1),...,(n,M)$ which respectively belong to groups $(N-M)-n,(N-M)-n+1,...,N-n$. Therefore, the distribution has the form:

$$P(n;t') = \sum_{j=0}^{M} p_{j,N-(n+j)}^{qe}\, p_{N-(n+j)}^{g}(t')\ , \quad \text{where} \quad 0 \le n \le N-M. \tag{6.36}$$

In the case of $N-M < n \le N$, there are $N-(n-1)$ microstates containing n substrate molecules, explicitly defined as $(n,M-(N-n)),(n,M-(N-n)+1),...,(n,M)$ and associated with groups $0,1,...,N-n$ respectively. Hence we have:

$$P(n;t') = \sum_{j=0}^{N-n} p_{j,N-(n+j)}^{qe}\, p_{N-(n+j)}^{g}(t')\ , \quad \text{where} \quad N-M < n \le N. \tag{6.37}$$

Finally, using the function $g(m)$ previously defined in Eq. (6.31), we obtain:

$$P(n;t') = \sum_{j=0}^{M-g(n)} p^{qe}_{j,N-(n+j)}\, p^{g}_{N-(n+j)}(t') \;, \quad \text{where} \quad 0 \leq n \leq N, \tag{6.38}$$

which fully describes the time-dependent solution for the multiple enzyme system $N \geq M$ with the initial condition $p^g_m(0) = \delta_{m,0}$. Note that the solution can also be extended to a more general initial distribution in the same way as was done for the single enzyme system in Section 6.4.1. The equations for mean number of substrate, $\langle n(t')\rangle$, and standard deviation, $\sigma(t')$, at rescaled time t' are the same as in Eqs. (6.22)–(6.23), but where $P(n;t')$ is now given by Eq. (6.38).

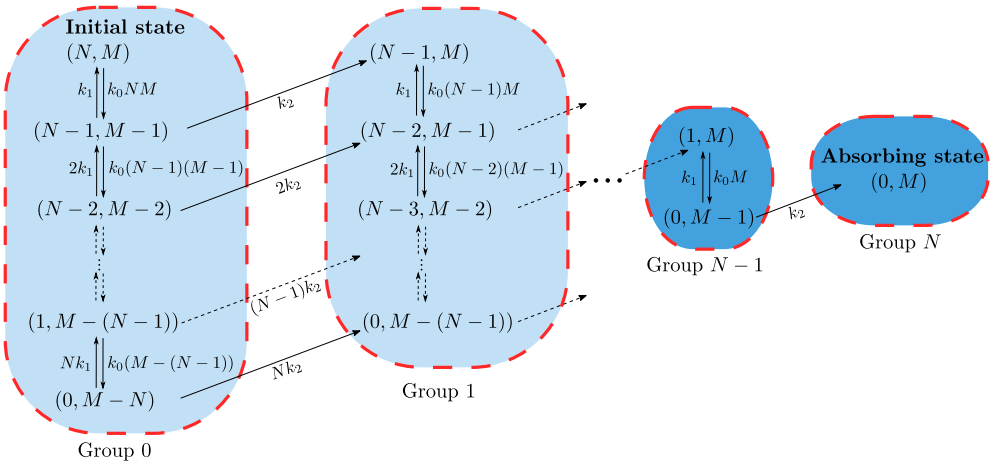


Figure 6.4: Illustration showing the transitions between the discrete microstates of the enzyme system (6.1) with initially M enzymes and N substrate molecules where $M \geq N$. As before, fast internal states are aggregated together into groups. The dynamics of the groups 0 to N can be mapped onto the dynamics of groups $N-M$ to N in the system with $N \geq M$ (shown in Fig. 6.3). See text for discussion.

Now consider a multiple enzyme system which initially contains fewer free substrate molecules than enzymes, i.e., $M \geq N$. The Markov chain describing the transitions between the microstates of this system, shown in Fig. 6.4, has similarities to that for the system with $N \geq M$. Specifically, if we replace N by M in groups 0 to N in the $M \geq N$ case of Fig. 6.4 then we exactly recover groups $N-M$ to N in the $N \geq M$ case of Fig. 6.3. This mapping implies that the dynamics of the system with $M \geq N$ are correctly described by Eq. (6.38) due to the utility of $g(m)$. Therefore, Eq. (6.38) is a valid solution for any positive integer values of N and M.

As for the single enzyme case, we can make the initial statement that k_2 must be sufficiently small for the derivation to hold, more precise. Suppose we are in the microstate (n, n_e). There are then 3 possible reactions which can occur: (i) $(n, n_e) \to (n, n_e + 1)$ with rate $k_2(M - n_e)$, (ii) $(n, n_e) \to (n+1, n_e + 1)$ with rate $k_1(M - n_e)$ and (iii) $(n, n_e) \to (n-1, n_e - 1)$ with rate $k_0\ n\ n_e$. Only the first reaction leads

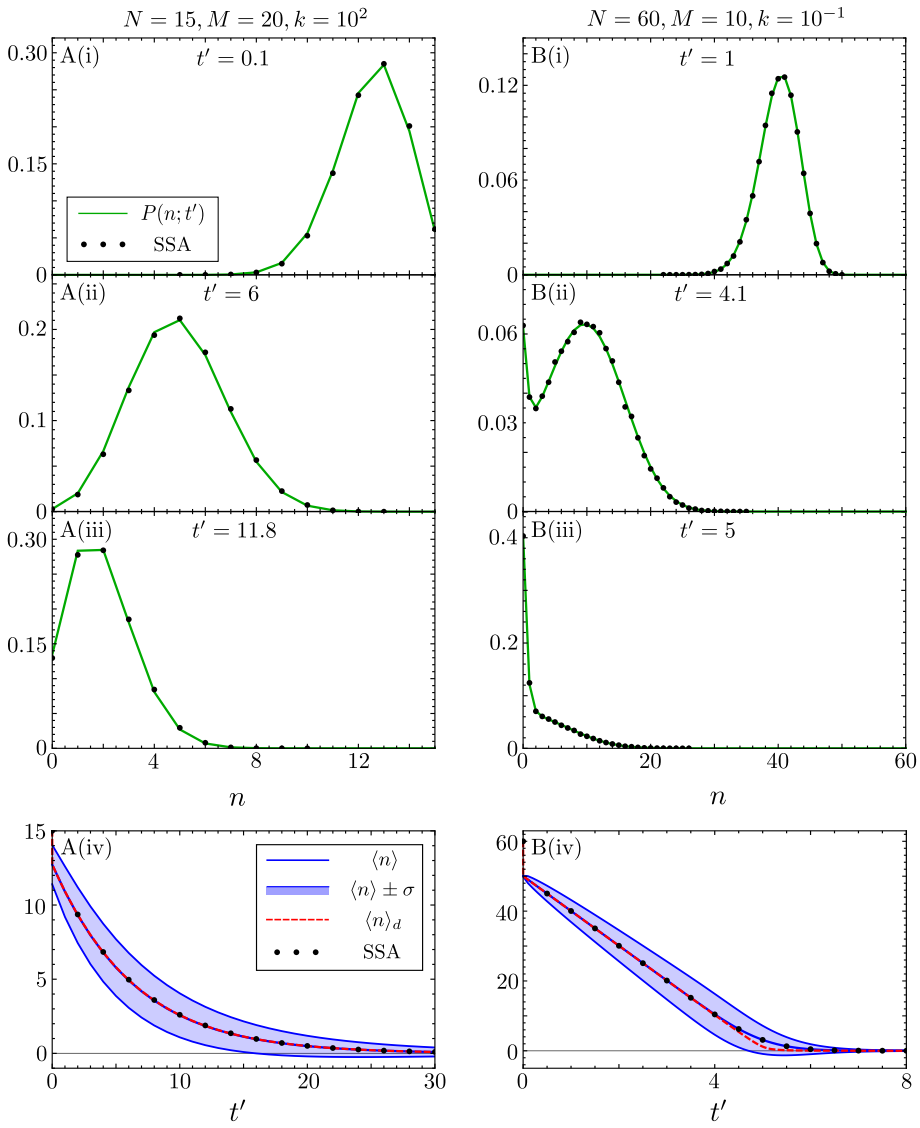


Figure 6.5: Comparison of the closed-form time-dependent probability distribution of substrate molecules, for the enzyme reaction (6.1) with multiple enzyme molecules M, and initial substrate molecules N, to the distribution obtained from the SSA. Note that the closed-form solution is given by Eq. (6.38). In A(i)–(iii), $N = 15, M = 20, k = 10^2$ and we simulate the SSA using $k_0 = 1$, $k_1 = 10^2$ and $k_2 = 1$; the theory (green lines) agrees with the SSA since the quasi-equilibrium assumption is justified, i.e., $k_1/k_2 \gg 1$. In B(i)–(iii), $N = 60, M = 10, k = 10^{-1}$ and we simulate the SSA using $k_0 = 10^3$, $k_1 = 10^2$ and $k_2 = 1$; again the theory is in agreement with the SSA since quasi-equilibrium is justified. Note that these results show that the theory accurately describes both the $N \geq M$ and the $M \geq N$ cases. In A(iv) and B(iv) we show the corresponding plots of the time-evolution of the mean $\langle n \rangle$ and of the standard deviation σ of the distributions of substrate molecules, as predicted by our theory; these are compared with the mean calculated from the SSA and the corresponding mean $\langle n \rangle_d$ obtained from the numerical solution of the deterministic rate equations given by Eq. (6.3). The parameter set in B is shown to be transiently bimodal in B(ii), whereas for the parameter set describing A transient bimodality is not observed. Each SSA probability distribution here is constructed from 10^5 individual reaction trajectories.

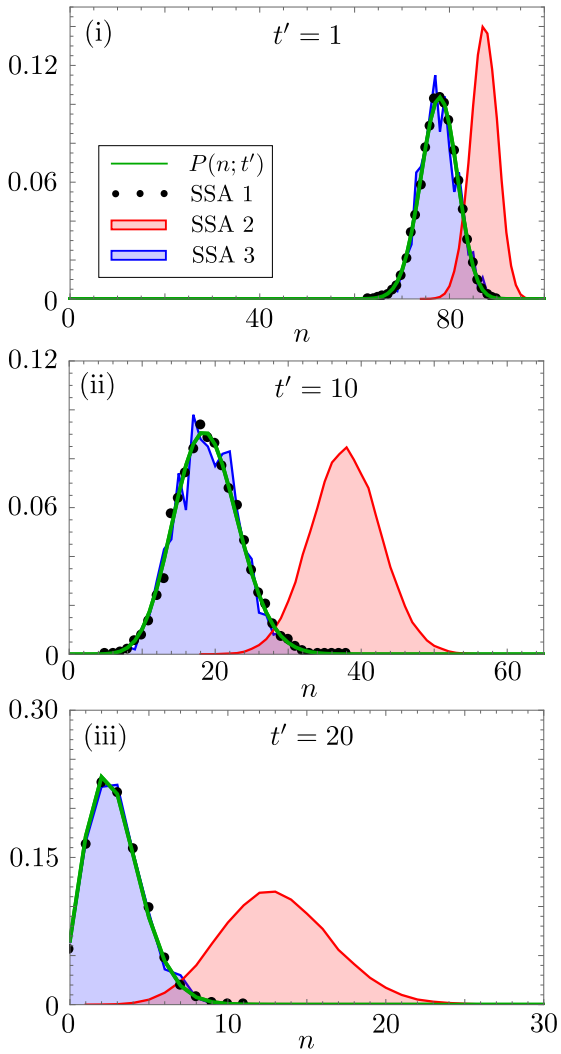


Figure 6.6: Testing the conditions necessary for the accuracy of the stochastic QEA. The three panels (i)–(iii) show how the accuracy of the closed-form time-dependent solution changes with time as we vary k_0/k_2 and k_1/k_2 whilst keeping $k = k_1/k_0$ fixed to 10^2 for the initial substrate number $N = 10^2$ and the total number of enzyme molecules equal to $M = 25$. The green line denotes the stochastic QEA solution from Eq. (6.38); SSA 1 (black dots) denotes the SSA prediction with parameters $k_0/k_2 = 1$, $k_1/k_2 = 10^2$ calculated over 10^4 trajectories; SSA 2 (blocked red region) denotes the SSA prediction with parameters $k_0/k_2 = 10^{-2}$, $k_1/k_2 = 1$ calculated over 10^5 trajectories; SSA 3 (blocked blue region) denotes the SSA prediction with parameters $k_0/k_2 = 10$, $k_1/k_2 = 10^3$ calculated over 10^3 trajectories. It is clear that SSA 2 is poorly predicted by $P(n;t)$, which is expected as $k_1 = \mathcal{O}(k_2)$. Since $P(n;t)$ is in equally good agreement with SSA 1 and SSA 3 it can be seen that the only requirement is $k_1 \gg k_2$, without requiring additional constraints on k_0.

to a transition out of the current group of microstates (since its associated with the product formation step) and hence the probability of exiting the current group is $k_2(M-n_e)/((k_1+k_2)(M-n_e)+k_0\ n\ n_e)$. It is easy to prove that the latter is always less than $k_2/(k_1+k_2)$. Hence quasi-equilibrium of microstates in each group is possible when $k_2/(k_1+k_2) \ll 1$. In other words, generally the closed-form solution for the distribution of substrate numbers will be accurate for all times provided $k_1 \gg k_2$.

In Fig. 6.5A(i-iii) and 6.5B(i-iii) we show agreement between $P(n;t')$ from Eq. (6.38) and the SSA where $k_1 \gg k_2$ is enforced, over times ranging between the initial time, when the number of substrate is $n=N$ and the absorbing state at $n=0$ for large times, for cases $M \geq N$ and $N \geq M$ respectively. In Fig. 6.5A(iv) and 6.5B(iv) we plot the mean and standard deviation of our analytical distribution ($\langle n \rangle$, σ), the deterministic mean $\langle n \rangle_d$ and the mean predicted by the SSA for $M \geq N$ and $N \geq M$ respectively. The SSA prediction of the mean is shown to be in exact correspondence with $\langle n \rangle$ when the QEA holds. The discrepancy previously seen in Fig. 6.2B(iv) between $\langle n \rangle$ and $\langle n \rangle_d$ at low molecule number is no longer observed in Fig. 6.5A(iv) where $M=\mathcal{O}(N)$, highlighting that the discrepancy seen in Fig. 6.2B(iv) originates from the molecular discreteness of the enzyme species. We additionally note the presence of transient bimodality in Fig. 6.5B(ii) similar to that seen in the single enzyme case from Section 6.4.1; note that the parameter set chosen for Figs. 6.5A(i-iii) does not exhibit transient bimodality. The parameter space of transient bimodality is explored later in more detail in Section 6.4.2. In Fig. 6.6 we demonstrate using stochastic simulations that, as predicted by our theory, the requirement for the stochastic QEA to be a good approximation relies only on satisfying the condition $k_1 \gg k_2$, and does not require any additional constraint on the value of k_0.

Time-dependent solution for the probability distribution of enzyme molecules

Having solved the master equation for the group dynamics, it is relatively straightforward to extract the time-dependent probability distribution for the number of free enzyme molecules, $P(n_E;t')$, and hence the distribution for the number of enzyme-substrate complexes, $P(n_C;t')$. As previously, we begin by considering the $N \geq M$ system depicted in Fig. 6.3. We observe that the groups $0 \leq m \leq N-M$ all contain a microstate with n_E free enzyme molecules, where $0 \leq n_E \leq M$, as enzymes are saturated with substrate. However, for groups $N-M < m \leq N$, free enzymes become more abundant than free substrate molecules, so that microstates containing $0 < n_E \leq M$ enzymes are found only in groups $N-M < m \leq N-(M-n_E)$. Note that the quasi-equilibrium probability of having n_E free enzymes in group m is $p^{qe}_{M-n_E,m}$, given by Eq. (6.32), and the group probabilities $p^g_m(t')$ are identical to the ones defined for the distribution of substrate

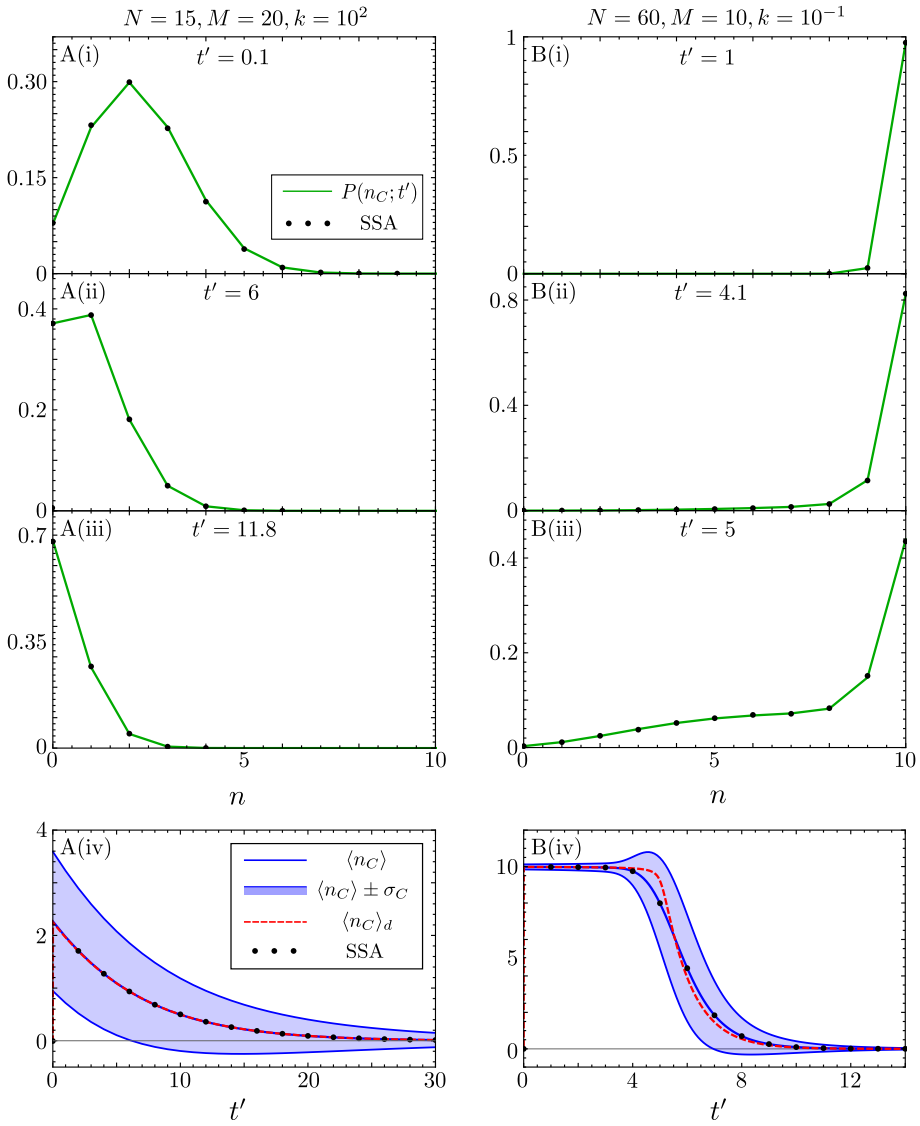


Figure 6.7: Comparison of the closed-form time-dependent probability distribution of enzyme-substrate complexes, for the enzyme reaction (6.1) with multiple enzyme molecules M, and initial substrate molecules N, to the distribution obtained from the SSA. Note that the closed-form solution is given by Eq. (6.40). In A(i)–(iii), $N = 15, M = 20, k = 10^2$ and we simulate the SSA using $k_0 = 1$, $k_1 = 10^2$ and $k_2 = 1$; In B(i)–(iii), $N = 60, M = 10, k = 10^{-1}$ and we simulate the SSA using $k_0 = 10^3$, $k_1 = 10^2$ and $k_2 = 1$ (parameters are the same as in Fig. 6.5). In both cases, the theory (green lines) agrees with the SSA since the quasi-equilibrium assumption is justified, i.e., $k_1/k_2 \gg 1$. In A(iv) and B(iv) we show the corresponding plots of the time-evolution of the mean $\langle n_C \rangle$ and of the standard deviation σ_C of the distributions of enzyme-substrate complex, as predicted by our theory; these are compared with the mean calculated from the SSA and the mean $\langle n_C \rangle_d$ obtained from the numerical solution of the deterministic rate equations given by Eq. (6.3). Each SSA probability distribution here is constructed from 10^5 individual reaction trajectories.

number in Eq. (6.38). Therefore, the distribution of free enzymes takes the form:

$$P(n_E;t') = \sum_{j=0}^{N-(M-n_E)} p^{qe}_{M-n_E,j}\, p^{g}_{j}(t')\ , \quad 0 \leq n_E \leq M. \tag{6.39}$$

This expression is valid for any positive integer values of N and M, again due to the mapping between the Markov chains of $N \geq M$ and $M \geq N$ systems, described above. Moreover, for the $N \leq M$ system, the definition of an empty sum as zero ensures that non-physical values of n_E are not allowed, i.e., the number of bound enzymes cannot be larger than N given the chosen initial conditions, so that $P(n_E;t') = 0$ for $n_E < M - N$. Finally, as $n_C = M - n_E$, the probability distribution of the enzyme-substrate complex follows trivially:

$$P(n_C;t') = \sum_{j=0}^{N-n_C} p^{qe}_{n_C,j}\, p^{g}_{j}(t')\ , \quad 0 \leq n_C \leq M. \tag{6.40}$$

In Fig. 6.7A(i-iii) and 6.7B(i-iii) we confirm that $P(n_C;t')$ from Eq. (6.40) and the SSA are in good agreement for enzyme systems with $M \geq N$ and $N \geq M$ respectively over the whole time-range from near the initial condition to the absorbing state, where again $k_1 \gg k_2$ is enforced (using the same parameters as in Fig. 6.5). Note that the transient bimodality is seemingly not manifest in $P(n_C;t')$ at the points in the parameter space where it is observed for the distribution of substrate number (c.f. Fig. 6.5B(ii) and 6.7B(ii)). In Fig. 6.7A(iv) and Fig. 6.7B(iv) we plot the mean and standard deviation of our analytical distribution for the enzyme-substrate complexes ($\langle n_C \rangle$ and σ_C), the mean predicted by the SSA and the mean number of complex molecules $\langle n_C \rangle_d$ obtained from the numerical solution of the deterministic rate equations given by Eq. (6.3) for $M \geq N$ and $N \geq M$ respectively. The SSA prediction of the mean matches $\langle n_C \rangle$ for all times further validating our solution, given that the QEA condition holds.

Bimodality

In Fig. 6.8A(i)–(iii) we explore further the transient bimodality observed in Figs. 6.2A(ii), 6.2B(iii) and 6.5B(ii). Namely, we investigate how the strength of the bimodality varies with the parameters N, M and k using the stochastic QEA solution from Eq. (6.38). Each point on the heatmap in Fig. 6.8A(i)–(iii) shows, for a particular parameter set, the maximum of the strength of bimodality calculated over the entire time course from $t' = 0$ to a time near the absorbing state of $n = 0$. We utilise the measure of bimodality

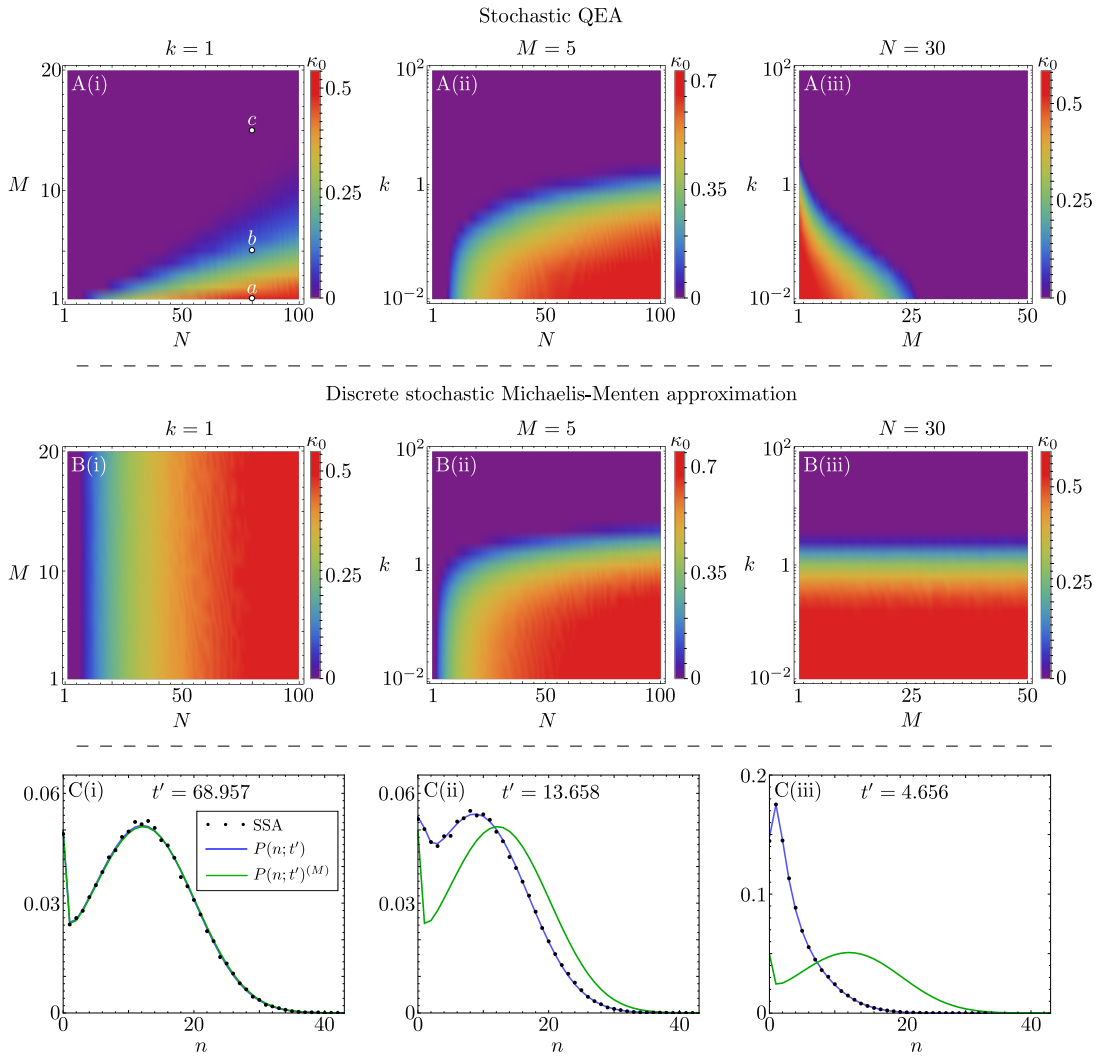


Figure 6.8: Heatmaps elucidating the regions of parameter space where transient bimodality is observed using the stochastic QEA solution (A(i)–(iii)) from Eq. (6.38) and the discrete stochastic MM approximation (B(i)–(iii)) given by Eq. (6.46). Note that κ_0 is a measure of how bimodal is the distribution of substrate molecules across the timecourse of the reaction (see text for details). Three parameter regimes are considered: N vs M with $k = 1$ (left), N vs k with $M = 5$ (middle) and M vs k with $N = 30$ (right). The plots C(i)–(iii) show the closed-form distributions of the stochastic QEA, $P(n;t')$, and the discrete stochastic MM approximation, $P(n;t')^{(M)}$, at the times when the stochastic QEA exhibits maximum bimodality, for cases with $k = 1$, $N = 80$ and (i) $M = 1$, (ii) $M = 5$ and (iii) $M = 15$ (highlighted on the heatmap A(i) as the points a, b and c respectively). The corresponding SSA predictions with $k_0/k_2 = 10^2$ and $k_1/k_2 = 10^2$ are also included (constructed from 10^5 individual reaction trajectories). Note that the two distributions (discrete stochastic MM approximation and stochastic QEA) are almost identical in C(i), but the difference becomes more pronounced in C(ii) and C(iii) with increasing M.

strength introduced in [115], which is explicitly given by:

$$\kappa = \frac{H_{\text{low}} - H_{\text{valley}}}{H_{\text{high}}}, \tag{6.41}$$

where H_{low} and H_{high} are the heights of the smallest and largest magnitude modes respectively, and H_{valley} is the height of the valley between the modes. For bimodal distributions κ has a value between 0 (no bimodality) and 1 (maximum bimodality), and for monomodal distributions is defined as zero. This definition of bimodality strength considers the 'most bimodal' distributions to have modes of equal height with a deep valley between them. In order to produce each heatmap we devised a simple algorithm, as follows. For each parameter set $\{N, M, k\}$:

1. Calculate the estimated time to reach the absorbing state which provides us with the time range, T_a, over which the transient bimodality search will be conducted. In order to avoid additional computational burdens of finding the absorption time using stochastic simulations, we use a much simpler but reasonably accurate estimate obtained from the deterministic QEA mean instead, given by solving Eq. (6.5) for $t' = k_2 t$:

$$T_a = \frac{N}{M} - \frac{k}{M} \log\left(\frac{\langle n\rangle_a e^{\frac{\langle n\rangle_a}{k}}}{N}\right), \tag{6.42}$$

   where we set $\langle n\rangle_a = 10^{-2}$, which was chosen small enough such that transient bimodality for all parameter sets was accounted for.
2. Choose the number of iterations, I, over which to check if the distribution is bimodal. In our case we chose $I = 400$. This gives the set of times over which we check for bimodality as $t_i = iT_a/I$ for $1 \leq i \leq I$.
3. Define a variable denoting the maximum bimodality measure κ_0 which is initially set to zero. For each t_i find the number of peaks in the distribution given by Eq. (6.38) for the stochastic QEA, and if two peaks are detected, calculate the bimodality strength κ from Eq. (6.41). If $\kappa > \kappa_0$ then set $\kappa_0 = \kappa$. Do for all t_n.
4. Once all iterations of this process are complete, the value of κ_0 will denote the largest value of the transient bimodality measure for all probability distributions at $t \in t_i$. We take κ_0 as the largest value of transient bimodality encountered on the time course.

The results obtained using this algorithm are summarised by the three heatmaps in Fig. 6.8A(i)–(iii). The distribution of substrate molecules corresponding to the time at which the maximal bimodality strength κ_0 occurs for points a, b, c in Fig. 6.8A(i) are shown by the solid blue lines in 6.8C(i)–(iii), respectively. Note that the bimodality is most pronounced in C(i), less in C(ii) and least in C(iii), in accordance with the value

of κ_0 in Fig. 6.8A(i); this validates the use of Eq. (6.41) as an accurate measure of the strength of bimodality. From Fig. 6.8A(i)–(iii), it is clear that bimodality is most pronounced when the initial number of substrate molecules N is significantly larger than the total enzyme number M and also when k is small, i.e., when the frequency of enzyme-substrate binding is much larger than the frequency of complex dissociation into enzyme and substrate. Note that generally the frequency of enzyme-substrate binding is inversely proportional to the volume of the compartment [68] in which the bimolecular reaction occurs and hence the transient bimodality is likely observable inside cells.

6.5 The discrete stochastic Michaelis-Menten approximation

We next consider how the analytical solution that we obtained for the reaction system (6.1) using a combination of averaging and linear algebra techniques in Section 6.4.2 compares with the solution of a commonly used reduced CME for enzyme kinetics.

The reduced CME for single substrate enzyme kinetics can be heuristically justified as follows (for a derivation see [86]). Under the QEA approximation, from the deterministic analysis in Section 6.3, it follows that the rate equation describing the time-evolution of the substrate concentration is given by:

$$\frac{d[S(t)]}{dt} = -\frac{V_{max}[S(t)]}{k + [S(t)]}. \tag{6.43}$$

Note that $V_{max} = k_2 M$, where M is the total number of enzyme molecules. Hence, species S can be seen as changing into P by means of an effective first-order decay reaction with rate given by the right hand side of Eq. (6.43). One common way to approximately describe the enzyme reaction stochastically consists of writing down an effective propensity describing the decay of substrate, i.e., we postulate that if there are m substrate molecules at time t then the probability that a reaction $S \to P$ occurs somewhere in a unit volume in the time interval $[t, t + dt)$ is approximately given by $a_m dt$ where $a_m = V_{max} m/(k + m)$. This is the discrete stochastic Michaelis-Menten (MM) approximation. Hence if we choose an initial condition of N substrate molecules, it follows that a corresponding effective CME is given by:

$$\partial_t P_{N-m}(t) = a_{m+1} P_{N-(m+1)}(t) - a_m P_{N-m}(t), \tag{6.44}$$

where $P_{N-m}(t)$ is the probability that there are m substrate molecules at time t $(0 \leq m \leq N)$. This CME can be conveniently written as:

$$\partial_t \underline{P}(t) = \mathcal{Q} \cdot \underline{P}(t) \tag{6.45}$$

where $\underline{P}(t) = (P_0(t), P_1(t), ..., P_N(t))$ and $\mathcal{Q}$ is a $(N+1)\times(N+1)$ lower bidiagonal matrix whose only non-zero elements are $\mathcal{Q}_{i,i} = -a_{N-(i-1)} = -\frac{(N-(i-1))V_{\max}}{k+(N-(i-1))}$ for $1 \leq i \leq N+1$, and $\mathcal{Q}_{i+1,i} = a_{N-(i-1)} = \frac{(N-(i-1))V_{\max}}{k+(N-(i-1))}$ for $1 \leq i \leq N$. Using the method in [127] that was used to solve the master equation for the group dynamics for the single enzyme, the solution is found to be given by Eq. (6.20), modified to take into account the fact that P_{N-n} is equivalent to the probability of being in the group $N-n$:

$$P_{N-n}(t')^{(M)} = \begin{cases} 0, & n > N, \\ e^{\lambda_1^{(M)} t'}, & n = N, \\ \left\{(-1)^{N-n}\prod_{k=1}^{N-n}\lambda_k^{(M)}\right\} \times \left\{\sum_{k=1}^{N-n+1} \frac{e^{\lambda_k^{(M)} t'}}{\prod_{j=1, j\neq k}^{N-n+1}(\lambda_k^{(M)} - \lambda_j^{(M)})}\right\}, & 0 \leq n < N. \end{cases} \tag{6.46}$$

Note the superscript (M) specifying that the solution is for the CME (6.45) resulting from the discrete stochastic MM approximation. Here, we have again rescaled the time $t' = k_2 t$, and $\lambda_m^{(M)}$ are the eigenvalues of $\mathcal{Q}$, which are simply given by the diagonal elements:

$$\lambda_m^{(M)} = -\frac{M(N-(m-1))}{k+N-(m-1)},\ 1 \leq m \leq N+1. \tag{6.47}$$

We shall denote the time-dependent mean and standard deviation of the distribution Eq. (6.46) by $\langle n(t')\rangle^{(M)}$ and $\sigma(t')^{(M)}$, respectively. Note that the distributions for the number of free enzymes/enzyme-substrate complexes cannot be obtained under the discrete stochastic MM approximation as the enzyme number fluctuations are not taken into account, in contrast to the Stochastic QEA from which enzyme/enzyme-substrate complex distributions can be obtained (see Section 6.4.2).

6.5.1 Comparison with the stochastic QEA

We used the algorithm described in Section 6.4.2 (with the difference that in step 3 we use Eq. (6.46) instead of Eq. (6.38)) to explore the regions of parameter space where the discrete stochastic MM approximation predicts the distribution of substrate molecules to be bimodal. The results are summarised by the three heatmaps in Fig. 6.8B(i)–(iii). By comparison to the heatmaps generated using the stochastic QEA in Fig. 6.8A(i)–(iii), it is clear that the discrete stochastic MM approximation tends to predict bimodality where in reality there is none. Notably, the bimodality predicted by the discrete stochastic MM approximation is independent of M (see Figs. 6.8B(i) and B(iii)) since M only acts to scale the eigenvalues representing the system's relaxation timescales in Eq. (6.47); in contrast, the stochastic QEA predicts bimodality which is strongly dependent on M (see Figs. 6.8A(i) and A(iii)). These issues with the discrete stochastic MM approximation are also clearly discernible in 6.8C(i)–(iii), where we compare the distribution of substrate molecule numbers predicted by this approximation (green line) with that predicted by the SSA (dots) and the stochastic QEA (blue line).

A different way to contrast the discrete stochastic MM approximation and the stochastic QEA involves comparing the eigenvalues of the transition matrix. In the single enzyme case where $M = 1$, one observes that the eigenvalues predicted by Eq. (6.47) exactly match the eigenvalues predicted by averaging for the group dynamics in the single enzyme case from Eq. (6.10). However, note that the group dynamics is not precisely the same as the substrate dynamics which is determined by two microstates in different groups. For example the averaging technique implies that there are two microstates that contain n substrate molecules: $(n, 0)$ and $(n, 1)$ associated with groups $N-(n+1)$ and $N-n$ respectively. However, this subtlety is not important if $N \gg 1$ and hence the CME resulting from the discrete stochastic MM approximation will practically lead to the same results as averaging for most cases of interest.

The comparison is more complicated in the case of multiple enzymes ($M > 1$) and abundant substrate $N \gg 1$, which we explore in Fig. 6.9 (for $N = 100$ and $k = 1$), showing how the discrete stochastic MM approximate solution differs to that from averaging as the ratio M/N is increased. We first consider the case where $M/N = 1/20$, and we see that $\langle n\rangle^{(M)}$ in Fig. 6.9A(i) is a good approximation of $\langle n\rangle$ for the time range of interest, i.e., from the initial state at $N = 100$ to a time $t' = 30$ where both $\langle n\rangle^{(M)}$ and $\langle n\rangle$ are small quantities. Note that the error in the standard deviation for this parameter set, shown in 6.9A(ii), is also small. The slight difference in the relaxation dynamics is corroborated by small differences in the eigenspectra of λ_m (given in Eq. (6.35) again noting that $\lambda_i = -a_i$) and $\lambda_m^{(M)}$ (given by Eq. (6.47)) which can be appreciated in Fig. (6.9)A(iii). We additionally plot the deterministic mean as predicted by Eq. (6.5) which clearly shows the relaxation dynamics of $\langle n\rangle_a$ accurately approximates $\langle n\rangle$ for short times only.

In Figs. 6.9B(i) and C(i) we see that as M/N increases to 1/5 and 1/2 respectively, $\langle n\rangle^{(M)}$ becomes a worse approximation of $\langle n\rangle$, with $\langle n\rangle^{(M)}$ tending more to $\langle n\rangle_a$ than $\langle n\rangle$. The corresponding error in the standard deviation, as shown in 6.9B(ii) and C(ii), also follows that of the mean, increasing with M/N. There are two main reasons for this disagreement:

1. If M is comparable to N then initially there will be large fluctuations in the number of enzyme molecules, which are taken into account by the averaging solution (since it allows for switching between microstates in each group) but not by the CME resulting from the discrete stochastic MM approximation (since the total number of enzymes only appears as a constant through V_{max}). This is most clearly seen in Fig. 6.9C(i) where we observe a large discrepancy between $\langle n\rangle$ and $\langle n\rangle^{(M)}$ at $t' = t'_c \ll 1$ (where t'_c is the time over which the initial transient occurs and is indistinguishable from $t' = 0$ in the figure).
2. Where $M/N \approx \mathcal{O}(1)$, the eigenspectra λ_m and $\lambda_m^{(M)}$ show a large disagreement (see Figs. 6.9B(iii) and C(iii)). This leads to the misprediction of the relaxation dynamics of $\langle n\rangle^{(M)}$, which better represents the dynamics predicted by $\langle n\rangle_a$ rather than of $\langle n\rangle$, for both small and large times. This is due to the fact that the effective Michaelis-Menten propensity in the reduced CME Eq. (6.45) is of the same form as the effective rate from the deterministic rate equation given by Eq. (6.43).

In summary, the solution of the CME obtained by the discrete stochastic MM approximation is a good approximation to the solution of the CME derived by averaging provided $N \gg 1$ and $N/M \gg 1$.

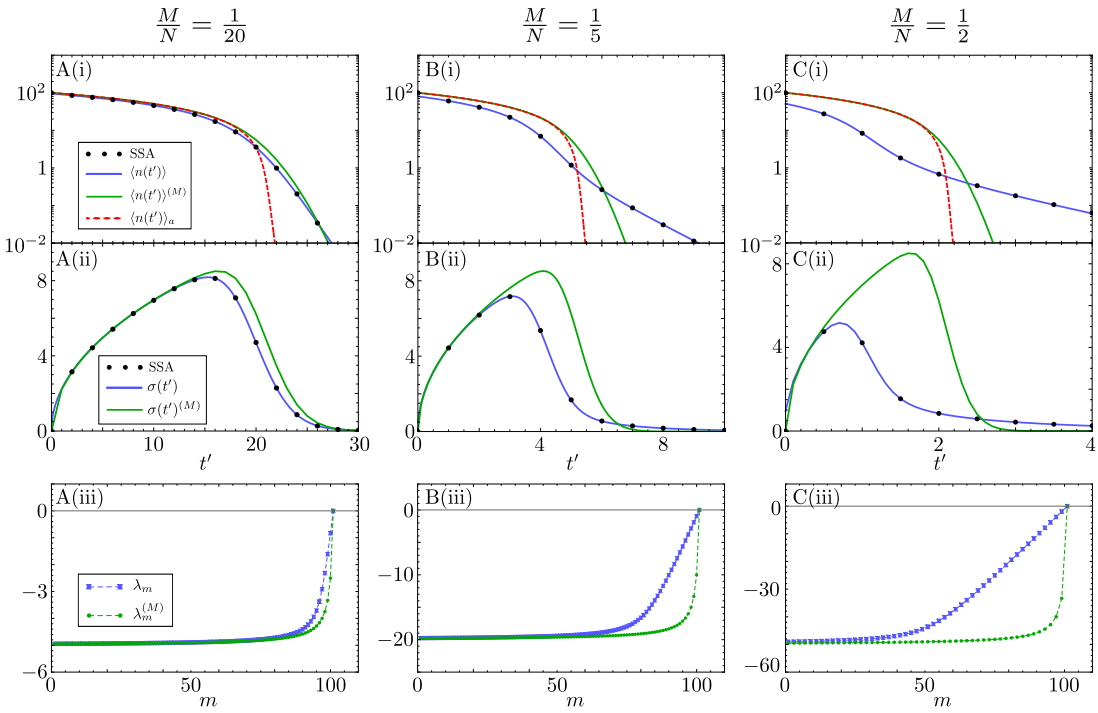


Figure 6.9: Comparison of the discrete stochastic MM approximation and the exact result from averaging in the quasi-equilibrium limit. A(i), B(i) and C(i) show log-scale plots of $\langle n\rangle$, $\langle n\rangle^{(M)}$ and $\langle n\rangle_a$ (from Eq. (6.5)) for $N = 100$, $k = 1$ and $M = 5$ (i.e, $M/N = 1/20$), $M = 20$ (i.e, $M/N = 1/5$) and $M = 50$ (i.e, $M/N = 1/2$) respectively. The corresponding SSA results with $k_0/k_2 = 10^2$ and $k_1/k_2 = 10^2$ are also included (constructed from 10^5 individual reaction trajectories). A(ii), B(ii) and C(ii) are the corresponding plots of the standard deviations $\sigma(t')$, $\sigma(t')^{(M)}$ and that of SSA. A(iii), B(iii) and C(iii) show the eigenspectra for each differing M/N; each symbol corresponds to an individual eigenvalue (since the spectra are discrete) and the dashed lines are only present to aid the reader.

6.6 Multi-substrate mechanisms

Thus far we have considered the simple enzyme mechanism shown in (6.1) where an enzyme can catalyze a single type of substrate. However, in nature it is common for one enzyme species to be able to catalyze multiple substrates [279]. Multi-substrate reactions follow various mechanisms that describe how substrates bind and in what sequence. One such common mechanism is that of ternary complex formation, whereby two substrates bind sequentially to an enzyme to form a complex with three molecules. An example is the following mechanism involving two substrate species A and B and two corresponding reaction products, P and Q [279]:

$$\begin{aligned} &E + A \rightleftharpoons EA,\ EA + B \rightleftharpoons EAB,\ EAB \rightarrow EPQ, \\ &EPQ \rightarrow EQ + P,\ EQ \rightarrow E + Q. \end{aligned} \tag{6.48}$$

Note that here we have assumed an ordered binding mechanism, in the sense that binding of A must precede that of B. An alternative is a random binding mechanism, wherein either A or B could first bind the enzyme. We assume that both enzyme-substrate binding reactions and the steps subsequent to complex formation are fast such that we can consider the simpler reaction scheme:

$$E + A + B \underset{k_1}{\overset{k_0}{\rightleftharpoons}} C \xrightarrow{k_2} E + P + Q. \tag{6.49}$$

Note that ordered or random binding mechanisms cannot be distinguished within this reaction scheme. We assume that there are initially N_A molecules of substrate A, N_B molecules of substrate B, where $N_A \geq N_B$, and M free enzymes. There exists a relation between the number of species A and B, denoted n_A and n_B respectively, which we can write as $n_A - n_B = N_A - N_B \equiv \Delta_{AB}$. Hence each microstate of the system is fully specified by (n_B, n_E). Again the group dynamics where $k_1 \gg k_2$ are given by Eq. (6.20) but the eigenvalues λ_m specific to this mechanism are given by:

$$\lambda_m = -\sum_{n=1}^{M-g(m-1)} n p^{qe}_{n,m-1} = k\partial_k(\ln(\mathcal{Z}_{m-1})),\ 1 \leq m \leq N_B + 1, \tag{6.50}$$

where we have now defined

$$p^{qe}_{i,m} = \frac{z_{i,m}}{\mathcal{Z}_m}, \tag{6.51}$$

$$g(m) = \Theta(m - (N_B - M)) \times (m - (N_B - M)), \tag{6.52}$$

$$z_{i,m} = k^{-i} \left\{ \prod_{j=1}^{i} ((N_A - m) - (j-1))((N_B - m) - (j-1))(M - (j-1)) \right\} \times \left\{ \prod_{j=i+1}^{M-g(m)} j \right\}, \tag{6.53}$$

$$\mathcal{Z}_m = \sum_{i=0}^{M-g(m)} z_{i,m}. \tag{6.54}$$

Using the results for the group dynamics and quasi-equilibrium probabilities, we can then find the probability distribution for the substrate molecules:

$$P(n_A, n_B; t') = \delta_{n_A - \Delta_{AB}, n_B} \times \left(\sum_{j=0}^{M-g(n_B)} p^{qe}_{j,N_B-(n_B+j)} p^{g}_{N_B-(n_B+j)}(t') \right), \tag{6.55}$$

where $\delta_{i,j}$ is the Kronecker delta symbol. This allows us to find the marginal distributions:

$$P(n_B; t') = \sum_{n_A} P(n_A, n_B; t') = \sum_{j=0}^{M-g(n_B)} p^{qe}_{j,N_B-(n_B+j)} p^{g}_{N_B-(n_B+j)}(t'), \tag{6.56}$$

$$P(n_A; t') = P(n_B + \Delta_{AB}; t'). \tag{6.57}$$

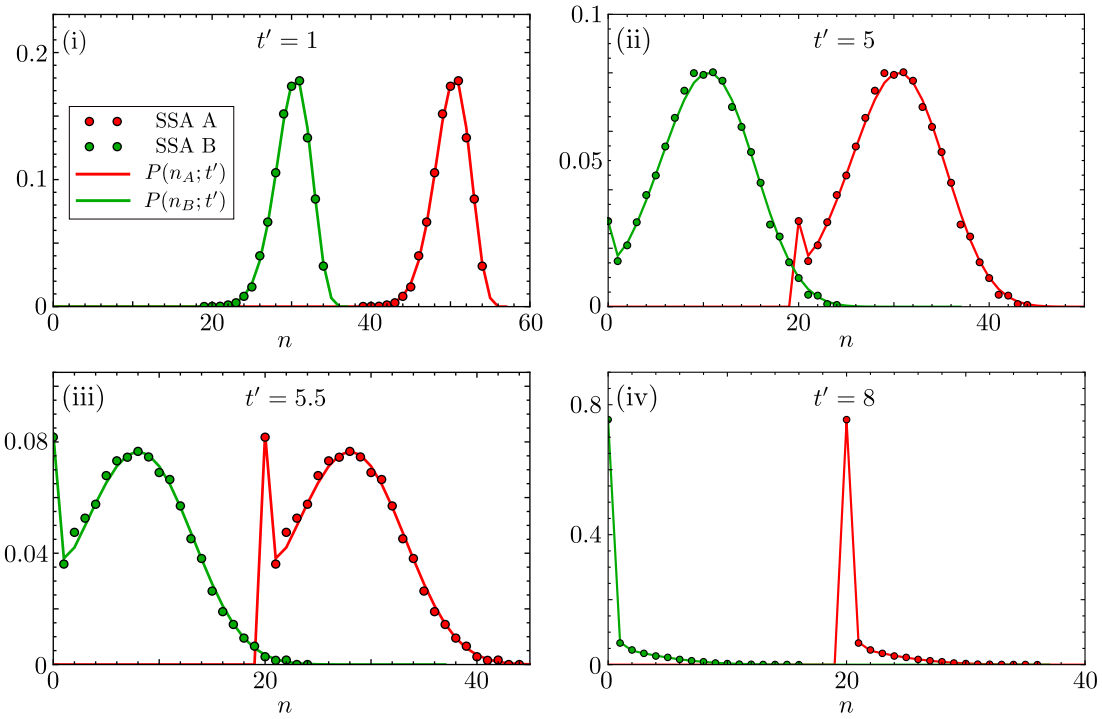


Figure 6.10: Comparison of the analytic distribution of two types of substrate species A and B, involved in the reaction mechanism (6.49), against the distributions obtained using the SSA. Note that SSA A and SSA B denote the SSA predictions for species A (red dots) and species B (green dots), respectively. In the panels (i)–(iv) we plot the probability distributions $P(n_A; t')$ (red line; from Eq. (6.57)) and $P(n_B; t')$ (green line; from Eq. (6.56)) for four different time points from near the initial condition (i) to near the absorbing state (iv) (time is non-dimensional as in previous figures). The initial number of substrate molecules are $N_A = 60$, $N_B = 40$ and the number of enzyme molecules is $M = 5$; the rates are $k_0/k_2 = k_1/k_2 = 10^3$ which enforce the QEA. The analytic distributions are in good agreement with the respective SSA distributions. Note that the absorbing point of A is $n_A = 20$ while that of B is $n_B = 0$; this is dictated by the difference between the initial number of substrate molecules $N_A - N_B = 20$. Each SSA probability distribution is constructed from 10^5 individual reaction trajectories.

In Fig. 6.10 we compare the analytic marginal distributions against the SSA and as expected we find very good agreement when the rate parameters are consistent with the QEA. As previously for the single substrate mechanism, the distributions of A and B molecules display bimodality at intermediate times.

6.7 Discussion

In summary, we have shown using averaging that in the limit of quasi-equilibrium between substrate and the enzyme, it is possible to reduce the two variable stochastic description of the MM reaction to that of an effective one variable master equation describing the slow transitions between groups of microstates. This master equation is subsequently solved exactly, using methods from linear algebra and complex analysis, to obtain closed-form solutions for the time-dependent marginal distributions of substrate and enzyme numbers. We have shown theoretically, and verified by means of stochastic simulations, that the solutions for the time-dependent marginal distributions are accurate for all times, provided the probability of complex decay into substrate and enzyme is much larger than the probability of complex decay into product and enzyme. To our knowledge, this is the first systematically derived approximate closed-form solution for the MM reaction for an arbitrary initial number of substrate and enzyme molecules; previous work treated a similar problem but using a heuristic approach [56] or derived closed-form solutions for the case of a single enzyme molecule [58, 57] or else considered reactions with multiple enzyme molecules focusing on deriving expressions for the turnover rate [259, 264, 113]. We have also shown how the same procedure can be used to obtain the solution of more complex enzyme mechanisms such as those involving the catalysis of multiple types of substrate by the same enzyme species.

For the MM reaction, we have compared our closed-form solution with that obtained by the solution of the CME reduced by means of the widely used discrete stochastic MM approximation [86], where the propensity for substrate decay has a hyperbolic dependence on the number of substrate molecules. If the initial substrate number N is not much larger than the total enzyme number M, but the rate constants satisfy the inequality $k_1 \gg k_2$, then the enzyme numbers fluctuations can be large, even though the rapid equilibrium approximation is valid. In this case, we show that the distribution predicted by the CME reduced using the discrete stochastic MM approximation is significantly different than the one obtained from stochastic simulations, whereas the solution provided by our theory accurately matches the simulations.

Using the closed-form solution for the time-dependent marginal probability distribution for substrate number, we have found that unexpectedly for a delta function (unimodal) initial condition, the distribution of substrate numbers can be bimodal at intermediate times, if the initial number of substrate molecules is significantly larger than the total number of enzyme molecules and provided the rate of complex decay into substrate and enzyme is much less than the rate of substrate and enzyme binding. We note that the latter rate in the CME formulation is inversely proportional to the compartment volume (since the encounter rate of two molecules decreases with increasing volume [68]), and hence our results imply that in the limit of small volumes (taken at constant initial number of substrate and enzyme molecules), bimodality of

the distribution of substrate molecules is observable. This result is of particular relevance to understanding enzyme dynamics inside cells where the volume is very small. Our system with the initial conditions used, can then be interpreted as modelling the enzyme-mediated decay of substrate molecules, following the production (via translation) of a short burst of substrate molecules N at time $t = 0$, provided there is not another burst of substrate expression before the substrate decays; these conditions are common for many cells where protein production occurs sporadically in bursts of short duration [30, 47]. We emphasise that the presence of transient bimodality in the MM reaction system is particularly interesting since it has no deterministic counterpart.

Chapter 7

Exact time-dependent dynamics of discrete binary choice models

This chapter has been published as [5] entitled *Exact time-dependent dynamics of discrete binary choice models* in the *Journal of Physics: Complexity*. Slight modifications have been made for its inclusion in this thesis.

7.1 Abstract

We provide a generic method to find full dynamical solutions to binary decision models with interactions. In these models, agents follow a stochastic evolution where they must choose between two possible choices by taking into account the choices of their peers. We illustrate our method by solving Kirman and Föllmer's ant recruitment model for any number N of *discrete* agents and for any choice of parameters, recovering past results found in the limit $N \to \infty$. We then solve extensions of the ant recruitment model for increasing asymmetry between the two choices. Finally, we provide an analytical time-dependent solution to the standard voter model and a semi-analytical solution to the vacillating voter model. Our results show that *exact* analytical time-dependent solutions can be achieved for discrete choice models without invoking that the number of agents N are continuous or that both choices are symmetric, and additionally show how to practically use the analytics for fast evaluation of the resulting probability distributions.

7.2 Introduction

That individuals take into account the choices made by others when making their own is evident to anyone who has witnessed fashion fads, trends and events of mass panic like bank runs. These collective phenomena have dramatic social consequences, as they are authentic "collective delusions" [280], of which economic bubbles and the subsequent crashes they produce are eloquent examples. The mechanism through which they appear is intuitive: sociable individuals tend to imitate the choices made by their peers, choosing to go to the same restaurant, dress the same way or buy/sell the same asset as their group of friends or the collective zeitgeist dictates. For any of these, when a choice becomes that of the majority its dominance and attractiveness tends to increase, as more and more individuals are persuaded to make it.

To the physicist, this is reminiscent of the mechanism governing certain phase transitions, and in particular that of the ferromagnetic transition, where the magnetic dipoles in a material all suddenly point in the same direction when cooled below a critical temperature. Owing to the common points between these mechanisms, similar behaviour is observed in the abrupt opinion swings seen in certain social systems (see [281] and references therein).

Thus it is no surprise that one of the strongest criticisms to the old paradigm of the *rational representative agent* used in textbook economics is that it does not sufficiently take into account interactions between agents. In that framework, agents make the choice that maximises a certain *utility function*, quantifying the level of satisfaction procured by said choice, by taking into account the different constraints they face—such as a limited budget.

Because there are no interactions, these models fail to capture the rich collective phenomena, or even the crises, that appear in real social systems [282]. For example, in a system made of non-interacting rational agents the only explanation for a large opinion swing is an *exogenous* event, such as the publication of new information that influences the agents. Therefore it is necessary to go further to understand the link between the *micromotives* that guide agents and their collective *macrobehaviour* [283].

A number of efforts have been made to alleviate this issue, notably by considering models where agents' decisions are influenced by interactions with their peers [284, 285, 59, 282, 286]. These models often study cases where agents face only two possible choices—reducing the problem to that of making a binary decision. In this way, one can study *toy models* describing social systems where agents interact, in the hope of gaining a better understanding of collective social phenomena much like the Ising model set a precedent for the understanding of emergent phenomena in condensed matter physics.

In spite of their simplicity, these models show a very rich phenomenology characterised by the appearance of crises, hysteresis and other emergent phenomena [287, 282]. In particular, the ant recruitment model has been of physical interest due to the occurrence of stochastic bimodality below a critical population, although deterministic analyses show monostability [62]. However, these dynamical models have often been studied only once their *stationary* state is reached, focusing in how their statistical description can change radically through subtle variations of the parameters that define it. But more insight can be gained by studying the full dynamics of how said stationary state is actually reached.

Indeed, one can consider Kirman and Föllmer's seminal ant recruitment model [59]. In its origin, it focused in explaining the results of an entomological experiment where an ant colony had access to two identical food sources. Instead of spreading evenly between the two sources, the ants were observed to concentrate in one of the two sources before randomly switching collectively to the other [288]. Similar models have arisen in independent parts of the literature—indeed the ant rationality model is very similar to the Bass diffusion model [289, 290], which describes the uptake of a new product or practice in a population, and if modelled stochastically would lead one to a model isomorphic to the model of Kirman and Föllmer.

The authors of [59] showed that this could be understood through a model were the ants had a certain propensity to imitate their peers, and another propensity to switch randomly between the two sources. When the effect of imitation is strong, the distribution of the number of ants in the food sources is bi-modal, and so one is more likely to find a majority of ants in either of the two sources, while when the random switching dominates one finds a regime with an unimodal distribution, with a rough half-and-half split between the two sources.

Although inspired by an example coming from behavioural biology, this model, and others that are very similar, has been used to explain behaviour in financial markets [291, 292, 63], firm agglomeration [64], the dynamics of fishing boats [67] and even wealth inequality [293]. The model is in fact also identical to the Moran model in genetics [60], and is also closely related to the Pólya urn model reviewed in [294]. Importantly, this chapter is distinguished from previous studies exploring time-dependent solutions to the ant recruitment model [66, 65] since here we consider the number of ants to be *discrete*, and then consider further applications of the methods herein to solve further binary choice models. We further show that our discrete, finite N results agree with those found in the thermodynamic limit upon taking $N \to \infty$ [65].

An interesting aspect of this model was found in [65], where a full dynamical solution to the model was provided in the limit of an infinitely large number of ants. Indeed, a key finding is that the time it takes for the ant colony to switch collectively from one food source to the other depends *exclusively* on the rate at which ants switch randomly. This can then be interpreted as implying that collective switches are driven by a single ant going to the other source and attracting all the others through an imitative avalanche. It is therefore clear that a precise dynamical description of such models is key in understanding the collective behaviours they display.

In this article, we solve this model in the case of a finite number of ants and show how the results from [65] can be recovered. We also show how our methods can be extended to solve a large class of similar models, such as the voter model [286, 295].

The chapter is structured as follows. In the first section we describe the ant recruitment model fully, and show how to map it onto a birth/death process. We solve it analytically using generating functions, and also obtain semi-analytical results in a computationally efficient way using the methods described in [127]. We then apply these methods to solving a more general version of the model, taking into account all possible asymmetries. Finally, we show applications of these techniques to the voter and vacillating voter models.

7.3 Setup

We first illustrate our setup by solving the stochastic dynamics of Kirman and Föllmer's ant rationality model. Consider a system of N ants where there are two different sources of food, *left* L and *right* R. Each ant is associated with a single food source, and we denote n as the number of ants at the right-hand food source. Since we do not track the spatial position of the ants, n completely specifies the state of the system.

The ants are subject to two separate influences: (i) a random influence whereby *each ant switches to the opposite food source at rate* ε, and (ii) a collective influence whereby when two ants meet—at rate ν—*if they are associated with opposing food sources, then one of the ants recruits the other to its food source.* Given that any two ants meet at rate ν *regardless of their current food source*, it is straightforward to show that the propensity at which two ants at opposing food sources meet is $\tilde{\nu}(n) = n(N-n)\nu/(N-1)$. This form of propensity comes from mass-action kinetics—the number of possible interactions between ants is the product of the number of ants on the right-hand food source (n) multiplied by the number of ants on the left-hand food source ($N-n$). The factor ν is essentially a 'reaction rate', and the scaling with respect to $(N-1)$ means that this propensity scales with the ant population in the same way of the random influence ϵ scales. We can now write a dynamical effective reaction scheme describing the number of ants on the right hand food source,

$$L \underset{n\varepsilon+\tilde{\nu}(n)}{\overset{(N-n)\varepsilon+\tilde{\nu}(n)}{\rightleftharpoons}} R. \tag{7.1}$$

Note that unlike effective reaction schemes often written in chemical reaction networks [81] the expressions labelling the arrows denote the full propensity for the event to occur given the state of the system n. From this effective reaction scheme one can then describe the dynamical evolution of the probability distribution $P(n,t)$ for reaction scheme (7.1) via the following master equation,

$$\begin{aligned}\partial_t P(n,t) =& \left[(N-(n-1))\varepsilon + \tilde{\nu}(n-1)\right] P(n-1,t) + \left[(n+1)\varepsilon + \tilde{\nu}(n+1)\right] P(n+1,t) \\ &- \left[(N-n)\varepsilon + n\varepsilon + 2\tilde{\nu}(n)\right] P(n,t),\end{aligned} \tag{7.2}$$

with a given initial condition that $n(t=0) = n_0$ ants are initially at the right-hand food source, represented by $P(n,0) = \delta_{n,n_0}$ where $\delta_{i,k}$ is the Kronecker delta symbol. Defining the $(N+1)\times(N+1)$-dimensional real matrix $\mathbf{M}$ as

$$\begin{aligned}(\mathbf{M})_{n,m} =& \delta_{n-1,m}\left((N-(n-1))\varepsilon + \tilde{\nu}(n-1)\right) \\ &+ \delta_{n+1,m}\left((n+1)\varepsilon + \tilde{\nu}(n+1)\right) \\ &- \delta_{n,m}\left(N\varepsilon + 2\tilde{\nu}(n)\right),\end{aligned} \tag{7.3}$$

it is straightforward to see that the master equation can be re-cast as $\partial_t \vec{P}(t) = \mathbf{M}\vec{P}$, where the n-th element of $\vec{P}(t)$ is $P(n,t)$. The matrix $\mathbf{M}$ corresponds to the Liouville or master operator and completely describes the dynamics of our system as it contains all the information on the transition rates.

The steady state distribution $\vec{P}_0$ can be derived by solving the equation $\mathbf{M}\vec{P}_0 = 0$. This can be solved by recursion, taking afterwards the $N \to \infty$ limit, with $n/N = x$ fixed to that the stationary distribution is given by a symmetric Beta distribution [59, 65]. That is,

$$P_s(n) \equiv P(n, t \to \infty) \underset{N \gg 1}{\propto} \left(\frac{n}{N}\right)^{\varepsilon/\mu - 1} \left(1 - \frac{n}{N}\right)^{\varepsilon/\mu - 1}, \tag{7.4}$$

where we introduce $\mu = \frac{\nu}{N-1}$ so that $\tilde{\nu}(n) = n(N-n)\mu$ to match the notation of [65].

This is the main point of interest of this model, as stressed by Kirman [59]: when imitation is strong, with $\varepsilon < \mu$, the most probable state is to have all of the ants in a single food source, as shown by the divergence of the probability distribution in Eq. (7.4), while the same probability is ≈ 0 in the high-noise regime $\varepsilon > \mu$ where the most probable state is to have a 50/50 split between the two sources. We exhibit this behaviour in Fig. 7.1, and show that even for $N = 50$ the Beta distribution is a good approximation to the exact distribution. Note that where $\varepsilon/\mu = 1$ we get uniform distributions for both $P_0(n)$ and $P_s(n)$

Because the tri-diagonal coefficients $(n, n+1)$ and $(n+1, n)$ are positive, and because the rank of the matrix is clearly $N+1$, it is straightforward to show that $\mathbf{M}$ has $N+1$ distinct *real* eigenvalues that we label $-\lambda_m$ for $m = 0, \ldots, N$. Indeed, write $\mathbf{M} = -\mathbf{P}\mathbf{\Delta}\mathbf{P}^{-1}$ with $\Delta = \mathrm{diag}(\lambda_m)$, then $(\vec{U}_m)_i = P_{im}$ and $(\vec{V}_m)_i = (\mathbf{P}^{-1})_{im}$. These vectors are known respectively as the right- and left-eigenvectors of $\mathbf{M}$. Further, direct application of the Perron-Frobenius theorem shows that $\lambda_m \geq 0$ and that 0 is an eigenvalue of $\mathbf{M}$. We choose therefore to label these eigenvalues as $0 = \lambda_0 < \lambda_2 < \ldots < \lambda_N$.

The model may now be formally solved as $\vec{P}(t) = e^{t\mathbf{M}}\vec{P}(0)$. We may then write $\mathbf{M} = -\sum_m \lambda_m \vec{U}_m \vec{V}_m^T$, which leads to $\vec{P}(t) = \sum_m e^{-\lambda_m t}\left(\vec{V}_m \cdot \vec{P}(0)\right)\vec{U}_m$. Denoting finally $c_m = (\vec{V}_m \cdot \vec{P}(0))$ and $\vec{U}_m = (f_m(0), \ldots, f_m(N))^T$ we reach the following formula for the full solution:

$$P(n,t) = \sum_m c_m f_m(n) e^{-\lambda_m t}, \tag{7.5}$$

where the different terms $c_m, f_m(n)$ and λ_m remain to be determined.

This is a formal solution of a discrete master equation. Master equations are notorious for being difficult to solve, *especially in time*. Common methods include the Poisson representation [74], Fokker-Planck (or Langevin) approximations [8, 74, 65, 100], field theory [296], the linear-mapping approximation [91] and the system-size expansion [8, 88, 89]. Below we utilise a combination of other methods, notably, the method of generating functions [8, 74], eigenfunction methods [8, 74] and the time-dependent solution to the 1D master equation [127].

In particular, the method used in [65] reached a solution of the same form as Eq. (7.5) by properly taking the limit $N \to \infty$ as to transform the matrix $\mathbf{M}$ into a Fokker-Planck partial-differential operator. The eigenfunctions (equivalent to the eigenvectors of $\mathbf{M}$) and eigenvalues of that operator were then found by two successive changes of variables mapping the problem onto a solvable quantum-mechanical problem. We claim to reach equivalent results using simpler methods that can be reused for other, similar models transparently.

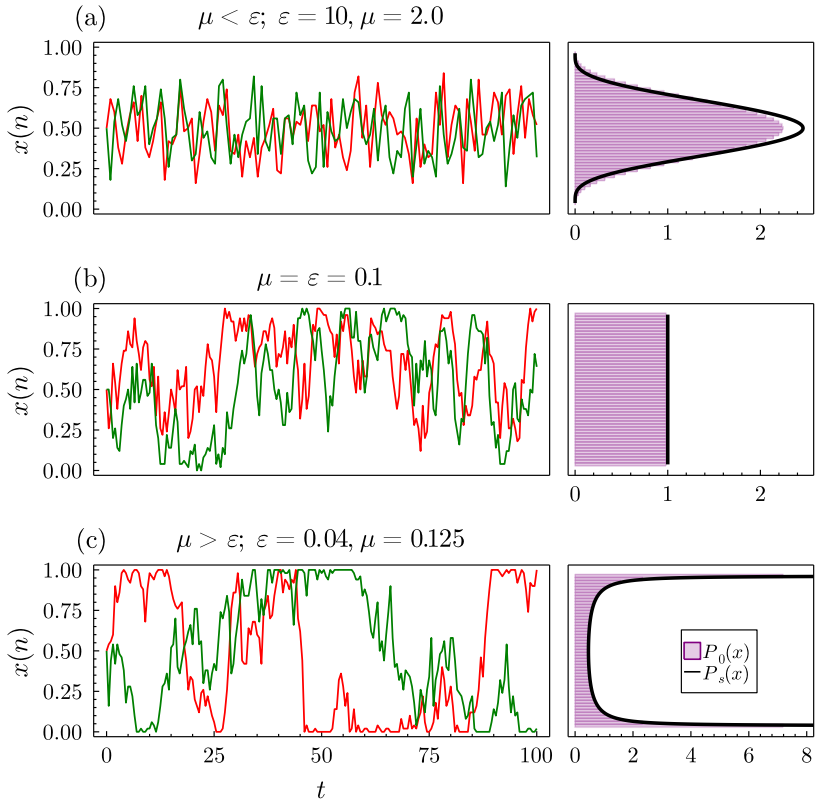


Figure 7.1: Sample of three different trajectories (left) of the fraction of ants $x(n) = n/N$ on the right-hand source, with the red and green lines showing two different realisations for $N = 50$ ants, along with the corresponding stationary densities (right). Purple bars are the exact stationary distribution for finite N, $P_0(x) = NP_0(n)$, and the black lines correspond to the symmetric Beta distribution, $P_s(x) = NP_s(n)$, given in Eq. (7.4). The stochastic simulations are done using the stochastic simulation algorithm [68]. Notice that in the high imitation regime $\varepsilon < \mu$, corresponding to plot (c) the ants tend to concentrate in one of the food sources for a time of order $1/\varepsilon$ before switching collectively to the other source. The case $\varepsilon = \mu$ in (b) corresponds to the situation where the Beta distribution is a uniform distribution over $[0, 1]$.

7.3.1 Explicit solution

The one-dimensional nature of the problem, along with the form of Eq. (7.5), invites us to introduce the generating function $G(z,t) = \sum_n z^n P(n,t)$, defined for $|z| \leq 1$. Plugging this definition into the ordinary differential equation system of Eq. (7.2) we obtain the following partial differential equation,

$$\begin{aligned}\frac{\partial_t G(z,t)}{z-1} =& \varepsilon N G(z,t) + (\mu(N-1)(z-1) \\ &- \varepsilon(z+1))\partial_z G(z,t) \\ &- \mu z(z-1)\partial_z^2 G(z,t),\end{aligned} \tag{7.6}$$

where we denote again $\mu = \nu/(N-1)$. This generating function PDE is subject to a boundary condition and an initial condition; the boundary condition relates to the normalisation of probability and is $G(1,t) = 1$, while the initial condition at $t = 0$ is found to be $G(z,0) = \sum_n \delta_{n,n_0} z^n = z^{n_0}$. Note that probabilities and moments can be obtained directly from the generating function:

$$\begin{aligned} P(n,t) &= \frac{1}{n!}\partial_z^n G(z,t)|_{z=0}, \\ \mathbb{E}[(n)_r] &= \partial_z^n G(z,t)|_{z=1}, \end{aligned} \tag{7.7}$$

where $\mathbb{E}[(n)_r] = \mathbb{E}\left[\prod_{i=0}^{r-1}(n-i)\right]$ is the r^{th} factorial moment.

From Eq. (7.5) it is clear that this function can be written as $G(z,t) = \sum_m c_m \left(\sum_n f_m(n) z^n\right) e^{-\lambda_m t}$. Defining now $g_m(z) = \sum_n f_m(n) z^n$ we reach the same form we would have obtained had we used an exponential ansatz for the solution [8, 296], namely $G(z,t) = \sum_m c_m g_m(z) e^{-\lambda_m t}$. The interpretation of the $g_m(z)$ functions is transparent, as they are the "generating functions" associated to each $f_m(n)$.

This is the same ansatz used in time-dependent solutions to quantum mechanical problems [297], where it arises naturally from the separation of variables $G(z,t) = f_1(z) f_2(t)$. Note that $g_0(z)$ corresponds, up to a normalisation constant, to the generating function of the steady state distribution $P(n, t \to \infty) \propto f_0(n)$.

Plugging this into Eq. (7.6) we reach an ODE in terms of z alone,

$$\mu z(z-1) g_m''(z) - (\mu(N-1)(z-1) - \varepsilon(z+1)) g_m'(z) - \left(\frac{\lambda_m}{z-1} + \varepsilon N\right) g_m(z) = 0. \tag{7.8}$$

One finds the singularities of this ODE are at $z = 0, 1$ and ∞ and are regular, hence the solution for $g_m(z)$ is given by a sum of two linearly independent hypergeometric type basis functions,

$$\begin{aligned} g_m(z) = (z-1)^{\alpha_m} \Big\{ & c_1^{(m)}\ {}_2F_1\left(\alpha_m + \frac{\varepsilon}{\mu}, \alpha_m - N; 1 - N - \frac{\varepsilon}{\mu}, z\right) \\ & + c_2^{(m)}\ z^{N+\frac{\varepsilon}{\mu}}\, {}_2F_1\left(\alpha_m + \frac{\varepsilon}{\mu}, \alpha_m + N; 1 + N + \frac{\varepsilon}{\mu}, z\right) \Big\}, \end{aligned} \tag{7.9}$$

where

$$\alpha_m = \frac{\mu - 2\varepsilon + \sqrt{4\varepsilon^2 - 4\varepsilon\mu + 4\lambda_m\mu + \mu^2}}{2\mu}. \tag{7.10}$$

However, owing to the definition of the generating function $G(z,t)$, the functions $g_m(z)$ should be polynomials of degree N in z (which follows since the probability of having $n > N$ is 0). We recall the definition of the hypergeometric function,

$${}_2F_1(a,b;c,z) = \sum_{\ell=0}^{\infty} \frac{(a)_\ell (b)_\ell}{(c)_\ell} z^\ell \tag{7.11}$$

where $(a)_\ell = \prod_{j=0}^{\ell}(a+j)$ is the Pochhammer symbol or rising factorial. From this definition, one can check that this function is a polynomial only when either the first or second argument is a negative integer. This must hold for all possible values of ε/μ, which means that $\alpha_m - N$ or $\alpha_m + N$ should be negative integers.

Consider now the first term in Eq. (7.9): if $\alpha_m - N$ is a negative integer, i.e., $\alpha_m \in [\![0, N]\!]$ (where $[\![0, N]\!] = [0, 1, \ldots, N]$), we have a polynomial of degree $N - \alpha_m$ for the hypergeometric function which becomes a polynomial of degree N after multiplication with $(z-1)^{\alpha_m}$, as required. We now relabel $c_1^{(m)} = c_m$.

On the other hand, for the second term we should have that $\alpha_m + N$ is a negative integer, suggesting to take an integer $\alpha_m \leq -N$ and giving a polynomial of degree $(-\alpha_m) - N$ for the hypergeometric term. However this is multiplied afterwards by $(z-1)^{\alpha_m}$, and one does not obtain a polynomial but a rational function. Therefore the only admissible solutions have $c_2^{(m)} = 0$.

We can therefore only keep the first term in the right-hand side of Eq. (7.9) and identify α_m with the index $m \in [\![0; N]\!]$, allowing us to find the $N+1$ eigenvalues of our problem,

$$\lambda_m = m\,(2\varepsilon + (m-1)\mu)\,, \quad m \in [\![0; N]\!], \tag{7.12}$$

which are precisely those given in [65], with the caveat that here μ depends explicitly on N as $\mu = \nu/(N-1)$. Without loss of generality, we set $c_1^{(m)} = 1$ and absorb it into the definition of c_m.

The constant c_m can then be evaluated by projecting the initial condition $G(z,0) = z^{n_0}$ onto the eigenfunctions $g_m(z)$, which form an orthogonal eigenbasis for a certain scalar product that can be determined fully using Sturm-Liouville theory (see Appendix E.1). In other words, there exists a function $w(z)$ such that $\langle g_m, g_n \rangle = \int_{-1}^{1} z\; w(z) g_m(z) g_n(z)) = \delta_{m,n}$. It follows then that

$$c_m = \frac{\int_{-1}^{1} z\; w(z) z^{n_0} g_m(z)}{\int_{-1}^{1} z\; w(z) (g_m(z))^2}, \tag{7.13}$$

which is equivalent to the projection method on the orthogonal eigenfunctions of the imaginary-time Hamiltonian used in [65].

We attract the reader's attention to the fact that the second eigenvalue λ_1 is still independent of N and equal to 2ε: the convergence to the stationary state, and therefore the rate at which ants switch to another source, is proportional only to the random switching rate, as found in [65] in the large N limit. We note that the waiting time to switch between the two food sources was explored more in depth in [62], where they approximately found the mean time it takes an ant to switch food sources for a given (ε, μ) as $N \to \infty$ based on first passage time theory.

It is also possible to retrieve the result from [65] that $\mathbb{E}[n(t)] - \mathbb{E}[n(t \to \infty)] \propto e^{-2\varepsilon t}$. Starting from the second line of Eq. (7.7), we write $\mathbb{E}[n(t)] = \sum_m c_m g_m'(1) e^{-\lambda_m t}$. Owing to the term $(z-1)^m$ in $g_m(z)$, it is quite straightforward to show that $g_m'(1) = 0$ for $m \geq 2$. Therefore we obtain that $\mathbb{E}[n(t)] = c_0 g_0'(1) + c_1 g_1'(1) e^{-2\varepsilon t}$ as required, with $\mathbb{E}[n(t \to \infty)] = c_0 g_0'(1) = 1/2$ because of symmetry considerations.

Note also that the spectrum obtained in Eq. (7.12) matches that of the $\tan^2$ Pöschl-Teller potential [298] for the quantum problem solved in [299]. The corresponding Schrödinger's equation is solved by a trigonometric change of variables that puts the eigenvalue problem into the form of an Euler hypergeometric differential equation, similar to the one obtained in Eq. (7.8). The discrete eigenvalues are then found by imposing that the wave-function must be square-normalisable, much as we must impose that the generating function be a polynomial in z.

The method for the large N limit used in [65] mapped the ant model into the $\tan^2$-potential Schrödinger's equation by writing a Fokker-Planck equation describing the random dynamics of the variable $x = n/N$, changing variables into $\varphi = 2x - 1$ to obtain another Fokker-Planck equation with a diffusive term that did not depend on φ and finally by using another common technique, described in detail in [189], to map this equation into a Schrödinger's equation. The method shown above achieves the same result in a much more straightforward way that can be applied to other similar problems and that allows one to obtain a solution for any value of N.

7.3.2 Practical evaluation of $P(n,t)$

Using the polynomial expression expressed above, ${}_2F_1(-k,a;b,z) = \sum_{l=0}^{k}\binom{k}{\ell}(-1)^\ell \frac{\Gamma(a+\ell)}{\Gamma(a)}\frac{\Gamma(b)}{\Gamma(b+\ell)}z^\ell$, we now recast $g_m(z) = c_m \sum_n f_m(n) z^n$, which yields

$$f_m(n) = (-1)^{m-n}\sum_{\ell=0}^{n}\binom{N-m}{\ell}\binom{m}{n-\ell} \frac{\Gamma(a+\ell)}{\Gamma(a)}\frac{\Gamma(b)}{\Gamma(b+\ell)}, \tag{7.14}$$

with $a = m + \frac{\varepsilon}{\mu}$ and $b = 1 - N - \frac{\varepsilon}{\mu}$. This describes the time-dependent solution up to the determination of the c_m coefficients. We show our results for $m = 0, 1$ on Figure 7.2, and note that the agreement with the $N \to \infty$ results from [65] is remarkably good.

Noticing then that

$$c_0 = \frac{1}{{}_2F_1\left(\frac{\varepsilon}{\mu}, -N; 1 - N - \frac{\varepsilon}{\mu}, 1\right)} \underset{N\to\infty}{\approx} \frac{\Gamma\left(2\frac{\varepsilon}{\mu}\right)}{\Gamma\left(\frac{\varepsilon}{\mu}\right)} N^{-\frac{\varepsilon}{\mu}} \tag{7.15}$$

and that $f_0(0) = 1$, one has directly that the probability of having $n = 0$ ants in the right-hand side food-source in the asymptotic regime behaves as $N^{-\frac{\varepsilon}{\mu}}$. In particular, if one does as in [59, 65] and studies the asymptotic probability density corresponding of the *fraction* of ants n/N in the right-hand food source, multiplication by the Jacobian of the transformation means that the asymptotic density at $n/N = 0$ behaves as $N^{1-\frac{\varepsilon}{\mu}}$. This corresponds to the behaviour of the density of the symmetric Beta distribution of parameter $\frac{\varepsilon}{\mu}$ at 0, as given in Eq. (7.4).

Nonetheless, it is possible to obtain the full time-dependent solution for a one-dimensional master equation such as (7.2) using the alternative method described in [127], which is exact up to the determination of the eigenvalues of the transition rate matrix which we have already obtained above. Similar applications of this little-known method have been employed in several

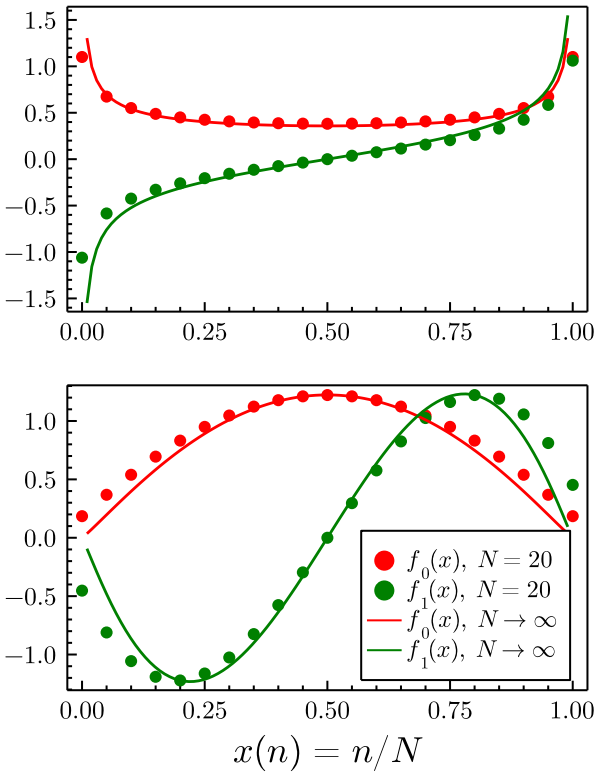


Figure 7.2: A figure showing the two first modes f_0 and f_1, represented with the reduced variable $x = n/N$. The solid lines correspond to the $N \to \infty$ results from [65] for $\varepsilon/\mu = 0.6$ (left) and $\varepsilon/\mu = 2.1$ (right), while the dots represent their discrete equivalent using Eq. (7.14) with $N = 20$. The fit is already remarkably good at only $N = 20$. In line with the quantum-mechanical interpretation given in [65], the mode f_1 can be interpreted as describing the hopping of ants from one source to the other, hence its asymmetric shape about $x = 0.5$. Note that this Figure reproduces Figure 2 from [65].

recent publications, for the solution of Brock and Durlauf's binary decision model [7], a solution to the Michaelis-Menten enzyme reaction [6], and in solving the fast-switching autoregulatory genetic feedback loop with bursty gene expression [115]. Note that this method is very similar to the one described in [129, 128], although these publications use Laplace transforms instead of Cauchy's integral formula.

We shall now detail the essential steps from the method of [127] in a generalised form that allows for multi-step reactions/events. For more rigorous details, see [127]. We start again from the formal solution $\vec{P}(t) = e^{t\mathbf{M}}\vec{P}(0)$, which after using Cauchy's integral formula reads

$$\vec{P}(t) = \frac{1}{2\pi}\oint_\gamma \mathrm{d}z\; e^{zt}(z\mathbf{I} - \mathbf{M})^{-1}\vec{P}(0). \tag{7.16}$$

where γ is a contour containing all the eigenvalues of $\mathbf{M}$. However, because $P(n,0) = \delta_{n,n_0}$ one can verify that $P(n,t) = [(z\mathbf{I}-\mathbf{M})^{-1}\vec{P}(0)]_n = [(z\mathbf{I}-\mathbf{M})^{-1}]_{n,n_0}$ (where $\mathbf{M}_{0,0}$ is the top-left-hand element of $\mathbf{M}$). We next use that for any invertible matrix $\mathbf{A}$, $\mathbf{A}^{-1} = \mathrm{adj}(\mathbf{A})/\det(\mathbf{A})$, where $\mathrm{adj}(\mathbf{A})$ is the adjugate matrix of $\mathbf{A}$, or equivalently the transpose of the cofactor matrix. Defining $\mathbf{B}(z) = \mathrm{adj}\,(z\mathbf{I} - \mathbf{M})$ we therefore reach the following expression,

$$P(n,t) = \frac{1}{2\pi i}\oint_\gamma \mathrm{d}z\ \frac{e^{zt}}{\prod_{i=0}^{N}(z+\lambda_i)}\mathbf{B}(z)_{n,n_0}, \tag{7.17}$$

where $\mathbf{B}(z)_{n,n_0}$ is a polynomial in z, as expected, and can be determined using standard methods [130], including a simple iterative formula for the case of tridiagonal $\mathbf{M}$, i.e., for a one-step birth death process [131], as is the case here.

Evaluating the integral using Cauchy's residue theorem leads to the following expression,

$$P(n,t) = \sum_{m=0}^{N}\left\{e^{-\lambda_m t}\frac{\mathbf{B}(-\lambda_m)_{n,n_0}}{\prod_{j\neq m}(\lambda_j-\lambda_m)}\right\}. \tag{7.18}$$

where we now recognise the equivalence with the result obtained using generating functions,

$$c_m f_m(n) = \frac{\mathbf{B}(-\lambda_m)_{n,n_0}}{\prod_{j\neq m}(\lambda_j-\lambda_m)}. \tag{7.19}$$

To summarise our results, the generating function approach allowed us to obtain the eigenvalues $-\lambda_m$ and the functions f_m describing the solution, up to the constants c_m that depend on the initial state. The last approach, using Cauchy's integral formula, allowed us to obtain a more amenable expression that is easy to evaluate numerically provided we have the eigenvalues obtained previously. We apply these methods to simulate the time-evolution of the distribution $P(n,t)$ in the case of the symmetric model and the asymmetric generalisations considered below on Figure 7.3. Note that we validate our analytical results in Fig. 7.3 against the stochastic simulation algorithm (SSA, [68]), a Monte Carlo method from which one can simulate exact stochastic trajectories describing master equations, for example Eq. (7.2) (Fig. 7.3, top plot).

7.3.3 Extension to asymmetric sources

Asymmetric noise only

The same analysis can be extended to the asymmetric ant model, studied in [67] to model the dynamics of fishing boats and in [292] to model agents trading in a financial market. This version of the model amounts to saying that the noise level ε depends on whether an ant is in the left- or right-hand food source. The equivalent reaction scheme to Eq. (7.1) describing this asymmetry in the noise level is,

$$L \underset{n\varepsilon_2+n(N-n)\mu}{\overset{(N-n)\varepsilon_1+n(N-n)\mu}{\rightleftharpoons}} R, \tag{7.20}$$

where ε_1 and ε_2 represent the random influence at the left and right food sources respectively.

The same analysis as above may be carried out in exactly the same way. After solving the eigenvalue problem using the characteristic function and imposing that it be a polynomial we find the following expression for the eigenvalues,

$$\lambda_m = m\left(\varepsilon_1+\varepsilon_2+(m-1)\mu\right), \quad m\in[\![0;N]\!], \tag{7.21}$$

which is the same expression obtained in the continuum $N \to \infty$ version obtained in [67].

Similarly, the modes $g_m(z)$ read

$$g_m(z) = (z-1)^m \, {}_2F_1\left(m + \frac{\varepsilon_1}{\mu}, m - N; 1 - N - \frac{\varepsilon_2}{\mu}, z\right), \tag{7.22}$$

and the expressions for $f_m(n)$ are given by Eq. (7.14) but with $a = m + \frac{\varepsilon_1}{\mu}$ and $b = 1 - N - \frac{\varepsilon_2}{\mu}$.

Again in this case we find that the convergence rate is given by $\varepsilon_1 + \varepsilon_2$ and therefore does not depend on the imitation rate μ for any value of N. We verify our analytic solution in Fig. 7.3 (middle plot) against the SSA.

Full asymmetry

The fully asymmetric case corresponds to a situation where the ants have a different imitation propensity depending on the food source they are currently in. Thus, Eq. (7.1) now reads

$$L \underset{n\varepsilon_2 + n(N-n)\mu_2}{\overset{(N-n)\varepsilon_1 + n(N-n)\mu_1}{\rightleftharpoons}} R. \tag{7.23}$$

The eigenvalue problem is now an ordinary differential equation with four regular singularities, and can therefore be solved via the Heun function [242, Sec. 31],

$$g_m(z) = H(a, q(\lambda_m); \alpha, \beta, \gamma, 0; z), \tag{7.24}$$

where we define

$$\begin{aligned} a &= \mu_2/\mu_1, \\ q(\lambda_m) &= \frac{(\lambda_m - N\varepsilon_1)(N-1)}{\mu_1}, \\ \alpha &= -N, \\ \beta &= \frac{(N-1)\varepsilon_1}{\mu_1}, \\ \gamma &= -(N-1)\left(1 + \frac{\varepsilon_2}{\mu_2}\right). \end{aligned} \tag{7.25}$$

We require again that this function be a polynomial of order N. We therefore write $H(a, q(\lambda_m); \alpha, \beta, \gamma, 0; z) = \sum_{j=0}^{\infty} C_j z^j$, with the following recurrence relation (see [242, Sec. 31.3]):

$$\begin{aligned} &C_0 = 1, \quad \alpha\gamma C_1 - q(\lambda_m(t)) C_0 = 0, \\ &R_j C_{j+1} - (Q_j + q(\lambda_m(t))) C_j + P_j C_{j-1} = 0, \end{aligned} \tag{7.26}$$

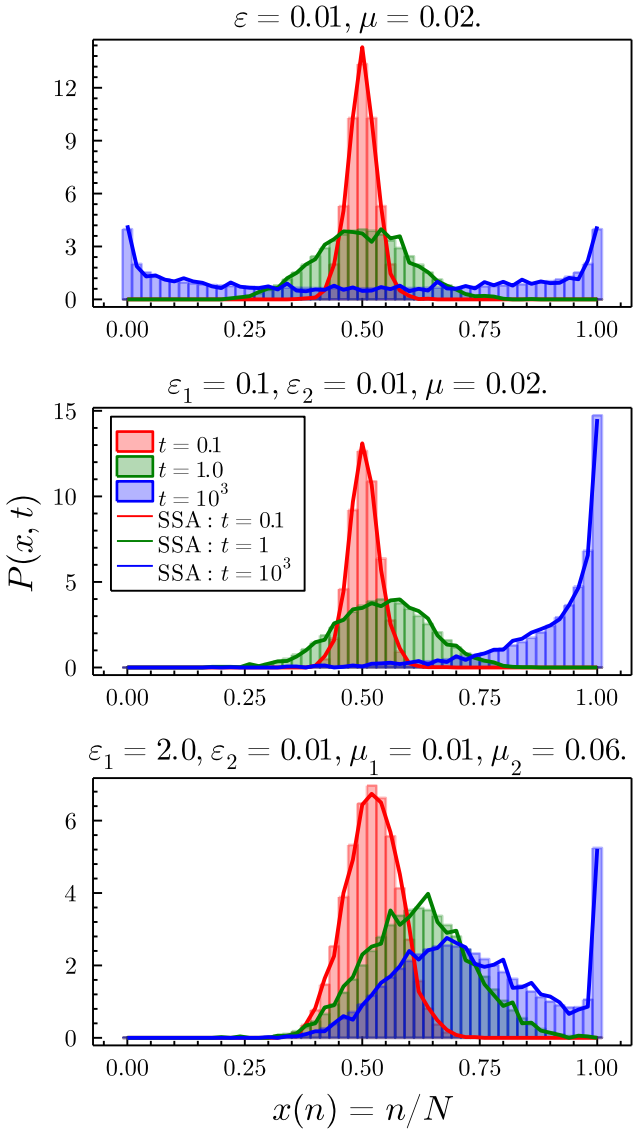


Figure 7.3: Plots showing the time evolution of ant rationality models under varying levels of asymmetry. In all plots the distributions are shown for $N = 50$ agents and initial condition $x = 0.5$, with the histograms showing the analytic solution (from Eq. (7.18)) and solid lines showing ensemble distributions from 2500 simulations of the stochastic simulation algorithm (SSA) [68]. The top plot shows a time evolution for the completely symmetric ant model; the middle plot shows a time evolution for the asymmetric ε model; and the bottom plot shows a time evolution for the entirely asymmetric case of $\varepsilon_1 \neq \varepsilon_2$ and $\mu_1 \neq \mu_2$. Clearly, as the model becomes more asymmetric more complex behaviours are possible.

with

$$\begin{aligned} R_j &= a(j+1)(j+\gamma), \\ Q_j &= j((j-1+\gamma)(1+a)+1+\alpha+\beta-\gamma), \\ P_j &= (j-1+\alpha)(j-1+\beta), \end{aligned} \tag{7.27}$$

and naturally $C_j = 0$ for $j > N$.

Setting $C_{N+1} = 0$, this recurrence leads to an equation for $q(\lambda_m)$ using continued fractions. Writing $\frac{a_1}{b_1+\frac{a_2}{b_2+\ldots}} = \frac{a_1}{b_1+}\frac{a_2}{b_2+}\ldots$, we find

$$q(\lambda_m) = \frac{R_0 P_1}{Q_1 + q(\lambda_m)-}\frac{R_1 P_2}{Q_2 + q(\lambda_m)-}\cdots\frac{R_{N-1}P_N}{Q_N + q(\lambda_m)}, \tag{7.28}$$

which then leads to a polynomial of order $N+1$ in λ_m and therefore to the $N+1$ distinct eigenvalues. This case, therefore, does not lead to a situation where we can improve on, say, a direct diagonalisation of the transition rate matrix.

It is nonetheless possible to study the time evolution of all instances of the model numerically, as shown on Figure 7.3 (bottom plot).

7.4 Applications to other models

A large class of binary decision models with interactions can be mapped onto birth/death processes. Indeed, if the dynamics is such that at every time step one or more agents change their mind from choice A to B, then this can be rewritten as removing an agent of class A from the population and replacing them with B. Thus it is possible to write a reaction scheme as we have done previously, write the master equation, find the corresponding differential equation for the generating function and solve using the methods we have shown.

One of the methods we used was already used in [7] to solve the Brock and Durlauf model [285]. We further illustrate this by giving solutions to the voter and vacillating voter models.

7.4.1 The voter model

In the voter model [300] one is interested in the opinion dynamics of individuals who can vote for two distinct choices—voting for a left- or right-wing political party, say. We can again chose to label those choices by L and R.

The model imagines that the agents are embedded in a social network, and they only communicate with nearest neighbours. In the dynamics, with probability p_d an agent is picked at random and their opinion becomes L or R with equal probability, or with probability $1-p_d$ a pair of neighbouring agents with opposite opinions LR is chosen, and then one of the agents persuades the other into adopting their opinion, so that the new pair becomes LL or RR with equal probability.

In this case, the model bears strong similarities with models for catalytic reactions between two different chemical species L and R [301, 302, 303] that are embedded in a substrate onto which they have adsorbed. The interpretation is now that (i) with probability p_d per unit time L or R desorb and are *immediately* replaced (at equal probability) with L or R, and (ii) with probability $1-p_d$ per unit time a nearest neighbour LR pair react and desorb, and are *immediately* replaced with $2L$s or $2R$s. The reaction scheme now reads

$$L \underset{\frac{p_d x}{2}+\frac{(1-p_d)x(1-x)}{2}}{\overset{\frac{p_d(1-x)}{2}+\frac{(1-p_d)x(1-x)}{2}}{\rightleftharpoons}} R, \tag{7.29}$$

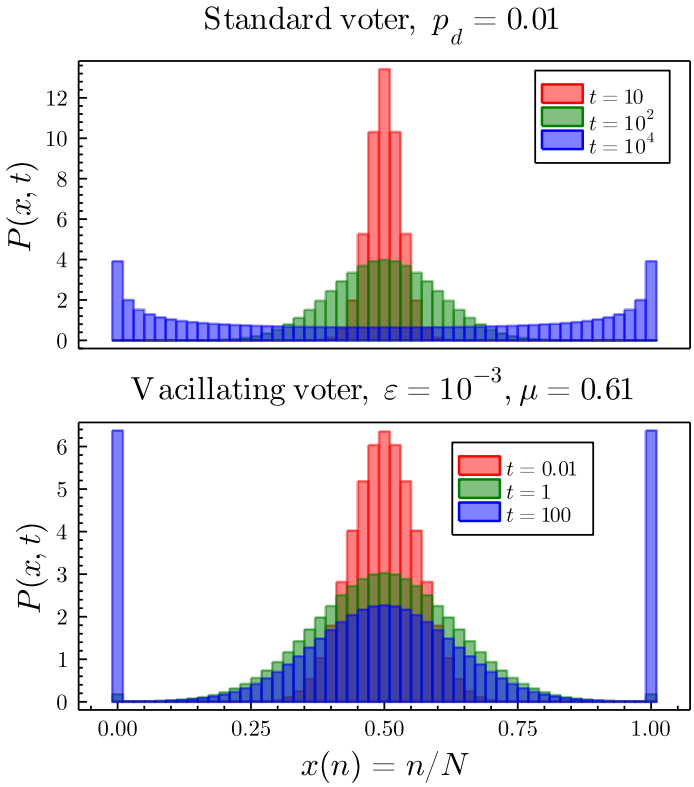


Figure 7.4: Plots showing the time evolution of the voter model (top) and the vacillating voter model (bottom) for $N = 50$ agents and initial condition $x = 0.5$. Unlike the voter model, the vacillating voter model is capable of exhibiting steady state trimodality (seen here for $t = 100$), brought on by the unsure nature of the voters.

where $x = n/N$ is the concentration of species R in the substrate.

Clearly, this is just a special case of the ant recruitment model of Eq. (7.1) in the special case where $\mu = (N-1)(1-2\varepsilon N)/2N^2$ and $\varepsilon < 1/2N$. Hence, its time-dependent dynamics is solved by the analyses above. We note that although it is a special case of the symmetric ant model it largely does share the models' phenomenology—showing the transition from monomodality to bimodality in the transient and steady state dynamics, albeit in a restricted section of the parameter space. A benefit of and Föllmer's ant model is that extensions towards higher degrees of asymmetry between the food sources are more easily implemented.

7.4.2 The vacillating voter model

Another version of said model is that of the *vacillating* voter model [295]. This model extends the voter model to the case where agents are unsure of their opinion. The dynamics is as follows: every time-step, an agent i with an opinion $S_i \in \{L, R\}$ is selected at random. With a probability $\propto \varepsilon$ the agent changes their mind randomly to the opposite choice, and with a probability $\propto \nu$ the agent then selects another agent $j \neq i$ at random. If $S_j \neq S_i$ then i's opinion is updated as $S_i \leftarrow S_j$, but if instead $S_j = S_i$ and the agents already agree, then i picks yet another random agent k. If $S_k \neq S_i$ then $S_i \leftarrow S_k$, and i retains their original opinion if $S_k = S_i$.

Again we may write the reaction scheme for this model,

$$L \underset{\varepsilon n+\nu\frac{(N-n)n}{N-1}\left(1+\frac{n}{N-1}\right)}{\overset{\varepsilon(N-n)+\nu\frac{(N-n)n}{N-1}\left(1+\frac{N-n}{N-1}\right)}{\rightleftharpoons}} R, \tag{7.30}$$

and apply the same reasoning as before.

This now yields third-order ODEs for the N generating function $g_m(z)$ (see Appendix E.2). Using the method of Frobenius, we write each $g_m(z)$ as a series and find the conditions for which it is a polynomial of degree $N+1$. This then allows to find a continued fraction expression that is satisfied by the eigenvalues λ_m, which again allows one to compute a full time-dependent solution using the resolvent relationship of Eq. (7.18). Our numerical results are shown in Fig. 7.4.

7.5 Conclusion

In this chapter, we have provided an exact solution to the ant recruitment model. We have proved that we can recover the $N \to \infty$ results found through other methods in [65], finding in particular that the stationary state is reached at an exponential rate of 2ε, independently of N.

We have also shown how our method can be extended to any binary decision model that can be mapped onto a one-step birth/death process. We have illustrated this with applications to the voter and the vacillating voter models.

More interesting lines of research are however possible in the context of decision theory. For example, our method works very well for models that display microscopic reversibility, as the process of one or more agents changing their mind from A to B can also be reversed by the process. However, as highlighted in [282], more complicated interactions between agents can break microscopic reversibility (or break detailed balance, in physics parlance) and possibly lead to more interesting phenomena.

A promising way to introduce this is through explicit path-dependency in the agents' decision-making process. Such effects have been studied through the inclusion of memory effects in utility functions [304, 305, 306], showing interesting effects such as ageing or memory induced condensation. There remains, however, to see how such memory effects could arise naturally from interactions. It should be noted, as highlighted in [293], that under a timescale of order $1/2\varepsilon$ the model we have described here is not ergodic: in the high imitation regime one may think that one of the two choices is optimal because it has been made all the time so far, but this may be only because under that timescale the ants are self-consistently "trapped" in one given choice, and one has not had the time to observe a full collective switch.

Other exciting results in decision theory can be obtained with random interactions. Agents interacting through random games are already known to produce very rich dynamics, in particular because multiple Nash equilibria emerge as the games become more complex [307] and because minute details such as the order in which players update their actions has an impact on the existence of an equilibrium [308].

Similarly, if the agents influence each other via a network with random topology and random weights then one can expect the dynamics to be similar to that of a spin glass, and therefore to display a rich phase diagram and non-intuitive dynamical behaviour. Progress in this direction has been made e.g. in [309], whose dynamics strongly resembles those describing glassy population dynamics in large ecosystems [310, 311] and where we can expect path dependency to emerge naturally.

By further studying the dynamics of economic toy models, such as those considered in this chapter, we can further expand the library of qualitative transient behaviours one expects to see in more complex models. Such qualitative behaviours are highly valuable, since they allow us to reframe the behaviours of real economic data, and parse them with respect to well understood behaviours from simpler models.

Chapter 8

Future Directions and Conclusions

In this final chapter, we discuss the most interesting future directions that have arisen from the work conducted in this thesis, and then conclude the work herein.

8.1 Future Directions

8.1.1 Inference of mechanistic models from single molecule data

Chapter 5 indicated that for gene expression, including nascent and mature mRNA, that the telegraph model is generally sufficient model for mechanistic gene expression. This is because there are few experimental data showing that the Fano factors of copy number for either nascent or mature mRNA are < 1 (which is not possible within the telegraph model). However, new data has come to our attention for mRNA expression in fission yeast (unpublished, from the lab of S. Hauf) where *simultaneous* measurements of nuclear and cytoplasmic mRNA are possible. Preliminary analysis shows that the Fano factor of the expression profiles can actually be < 1 for both nuclear and cytoplasmic mRNA, meaning that the telegraph model may not be the best model of gene expression for the data. We aim to map this data to mechanistic and telegraph models using moment-based maximum likelihood and Bayesian inference (informed by previous studies such as [312, 313, 314, 315]) and determine using information criteria which is the optimal minimal model for gene expression in fission yeast.

8.1.2 Inferring volume scaling laws in *E. coli*

In the work presented in this thesis, no emphasis has been placed on the relationship between gene expression and the dynamics of the cell cycle including gene replication, gene product partitioning and gene dosage compensation (among other effects, see [22, 23, 24, 25, 26, 27, 28, 316]). These studies manage to connect gene expression and cell cycle dynamics analytically, giving rise to great insights, but also increased complexity that is difficult to interpret. We propose a simpler approach to studying these dynamics by positing simple volume scaling laws on protein burst sizes and production rates, whose scaling exponents can be determined from maximum likelihood or Bayesian inference from mother machine data in *E. coli* [317]. Such minimalistic approaches aim to give quantitative results that allow for greater clarity in the coarse grained effect that

complex cell cycle dynamics exhibit on gene expression (similar to scaling law approaches seen in [318, 319]). We also hope to investigate the effect of perturbation experiments (unpublished data from the lab of M. El Karoui) on the volume scaling in an aim to answer whether such experiments lead to quantitative changes in the dynamics of gene expression.

8.1.3 Origins of transient bimodality in enzyme kinetics

The dynamics of Michaelis-Menten enzyme kinetics seen in Chapter 6 realised the rare phenomenon of transient bimodality, an effect that as of yet has little explanation. The phenomenon is bizarre in that it does not require the process to admit two distinct behavioural modes as is generally the case with steady state bimodality—e.g., for autoregulation bimodality can arise from being in one gene state or the other at any one time. Notably, transient bimodality has also been observed in the dynamics of lasers, see [275], where it has been claimed that "transient bimodality does not have a trivial origin because it arises from a delicate combination of critical slowing down and noise". Although this may give a qualitative explanation for transient bimodality it has not been confirmed, nor has it been analysed in a quantitative framework. We propose to further study the origins of transient bimodality. Work conducted in [320], where enzyme kinetics is studied from a reaction-path perspective, may suggest a way forward.

8.1.4 Time-dependent analytics for N source ant recruitment

The model of ant rationality studied in Chapter 7 considers two food sources—in the analogy to economical decision making this is clearly the most simplified case, since generally one does not have to choose between two options, but indeed a multitude of different ones [321]. Although the N food source case of Kirman's model has been solved at steady state [322], it has not been solved in time. Knowledge of its time-dependency may provide insight into studies of multiple choice in firm localisation [323, 324] or future technology transformation models [325, 326, 327]. We propose a study on N source ant rationality models using the linear noise approximation (LNA), in a similar vein to studies conducted on gene expression in cells and tissues [166] and large-scale reaction networks [328]. These approximate calculations can then be compared against simulations to determine the validity of the LNA on multiple choice models.

8.2 Conclusions

This thesis has consisted of three main themes: model reduction, mechanistic modelling and transience in models of stochastic kinetics. These themes have also been considered in three different fields of study, the main one being stochastic gene expression, but also in enzyme kinetics and models of social choice. In all three of these fields very similar questions are asked, and the aim is to accurately model such systems in such a way that we understand the limits of our approximations, but also that we can capture the key emergent phenomena in our models. For example, the Hill function is often the preferred choice to model the regulatory function of an auto-regulatory gene, but it is now found to be a very poor approximation in certain regimes of positive auto-regulation due to finite molecule number effects—in a very similar

way to how the Michaelis-Menten enzyme reaction can be approximated by a Hill function too. One suspects that the disagreement between the models of Michaelis-Menten kinetics modelled using mass-action kinetics versus the Hill function could also be explained from a finite molecule number perspective. Of particular note is that once we understand the limits of our approximations we can use them much more confidently—as is clear from this thesis, since in Chapter 3 we showed the limits of the Hill function to model auto-regulation and then in Chapter 4 we used that same approximation to make our analytics tractable.

In terms of capturing the key emergent phenomena there are interesting parallels to draw between the telegraph model, explaining mRNA gene expression, and Kirman's ant recruitment model. Both models are designed to be minimal, and both clearly ignore very important aspects of the complex systems that define them, e.g., complex gene regulatory interactions in the telegraph model and similar social network ignorance in the model of ant recruitment. However, in a sense they are optimally minimal—the telegraph model provides an easy and intuitive narrative to explain the 'bursty' behaviour underlying mRNA expression, while the ant recruitment model provides a way of understanding the coalescent nature of social choice from a perspective that relies only on the interactions between agents and not changes in the environment that they inhabit. Both models can clearly be studied to a higher degree, including more complex and realistic aspects, but the reasons for their success are also contained in their simplicity. As unfortunate we may find it as modellers, who are often looking to add complexity and realism, it is often the simplest explanations that stick around the longest.

A final key comparison between the variety of models seen in this thesis comes from the importance of transience in modelling across different fields. Transience does not have to mean that a system never settles to a steady state, but is also very important for small perturbations away from the steady state. The key question then becomes: What are the key elements of the model that determine the relaxation back to steady state. In all but the simplest cases, answering this question is non-trivial. In biology, it is often attributed that biological systems exist at the so-called "biological steady state", and in social choice and economics that equilibrium is reached on very fast time scales (instantaneously in neoclassical economics [329]). However, in reality this will rarely be the case. In biology, cell-cycle dynamics, circadian rhythms and differentiation all contribute to the changing state of the cell, and interactions between cells further act to perturb the state of a cell from its equilibrium. In economics and social choice, information is not transmitted instantaneously and agents are not perfect in the game theoretic sense, and changes in the external environment can mean a system takes very long time to settle back to a steady state [282].

There are however several key modelling features that have been neglected in this thesis, and that provide other key directions to conduct research in beyond the projects stated above. The first is that, when modelling gene expression or enzyme kinetics *in vivo* one should take into account aspects of cell-cycle dynamics including: partitioning of gene products at cell division, DNA replication, gene-dosage compensation and dilution [22, 23, 24, 25, 26, 27, 28]. One should also take care to properly design models that replicate the biological data one is comparing to, since whether one is using lineage or population measurements of molecule numbers can have a large impact on molecule number distributions [27, 23]. Another important direction not explored in this thesis is in the design of computational packages that allow other researchers

to easily conduct analytics or inference. For example, the scientific programming language Julia [94] now has packages that easily allow one to model reaction networks [330], produce and analyse non-linear moment equations [331] and perform Bayesian inference [332]. A final aspect that has not been considered in this thesis is the inclusion of spatial aspects of stochastic modelling to include the effects of macro-molecular crowding, compartmentalisation and the breakdown of the CME [333, 334, 335]. This could provide several new directions for the research conducted in this thesis, particularly combining transience or mechanistic modelling alongside the reaction-diffusion master equation as this has rarely been considered.

Appendix A

Chapter 2 Appendices

A.1 Relationship between the deterministic and stochastic models

For the non-bursty feedback loop (Eq. (3.1)), the stochastic description is given by the chemical master equation which can be formulated as a set of two coupled equations:

$$\begin{aligned}\frac{dP_0(n,t)}{dt} =&\rho_u(P_0(n-1,t)-P_0(n,t))+((n+1)P_0(n+1,t)-nP_0(n,t)) \\ &+\sigma_u P_1(n-1,t)-\sigma_b n P_0(n,t),\end{aligned} \tag{A.1}$$

$$\begin{aligned}\frac{dP_1(n,t)}{dt} =&\rho_b(P_1(n-1,t)-P_1(n,t))+((n+1)P_1(n+1,t)-nP_1(n,t)) \\ &-\sigma_u P_1(n,t)+\sigma_b(n+1)P_0(n+1,t),\end{aligned} \tag{A.2}$$

where $P_0(n,t)$ is the probability that at time t there are n proteins and the gene is in state G while $P_1(n,t)$ is the probability that at time t there are n proteins and the gene is in state G^*. Note that time t is non-dimensional and equal to the actual time multiplied by the protein degradation rate. The probability of n proteins is then given by $P(n,t)=P_0(n,t)+P_1(n,t)$. Using these equations it is straightforward to show that the time-evolution equations for the first moments:

$$\frac{d\langle g\rangle}{dt} = -\sigma_b\langle ng\rangle+\sigma_u(1-\langle g\rangle), \tag{A.3}$$

$$\frac{d\langle n\rangle}{dt} = -\sigma_b\langle ng\rangle+\sigma_u(1-\langle g\rangle)+\rho_u\langle g\rangle+\rho_b(1-\langle g\rangle)-\langle n\rangle, \tag{A.4}$$

where $\langle n\rangle=\sum_n nP(n)$ is the mean numbers of proteins, $\langle g\rangle=\sum_n P_0(n)$ is the fraction of time the gene is in the ON state (or equivalently the average number of gene in ON state) and $\langle ng\rangle=\sum_n nP_0(n)$. Note that we have here suppressed the time dependence for convenience. A comparison of Eqs. (3.2)–(3.3) with Eqs. (A.3)–(A.4) shows that the two are the same if $\langle ng\rangle=\langle n\rangle\langle g\rangle$, i.e., the deterministic and exact stochastic models agree in the means if the fluctuations in the gene and protein numbers are independent of each other.

We next show that in the limit of fast promoter switching and when there is a non-zero correlation between the fluctuations of protein and gene, the exact stochastic model gives a time-evolution equation for the protein number mean which is different than that given by the deterministic analysis. The limit of fast promoter switching implies that $d\langle g\rangle/dt\approx 0$ and hence using Eq. (A.3) it follows that the mean number of proteins conditional on the gene being in state G can

be written as:

$$\langle n|G\rangle = \frac{\sum_{n=0}^{\infty} nP_0(n)}{\sum_{n=0}^{\infty} P_0(n)} = \frac{\langle ng\rangle}{\langle g\rangle} \approx L\frac{(1-\langle g\rangle)}{\langle g\rangle}, \tag{A.5}$$

from which we obtain after rearrangement:

$$\langle g\rangle \approx \frac{L}{L+\langle n|G\rangle}. \tag{A.6}$$

Substituting in Eq. (A.4) we obtain an effective equation for the time-evolution of the protein numbers under the condition of fast promoter switching:

$$\frac{d\langle n\rangle}{dt} \approx \frac{L\rho_u + \rho_b\langle n|G\rangle}{L+\langle n|G\rangle} - \langle n\rangle. \tag{A.7}$$

Contrasting this equation with the effective equation obtained through the deterministic approach, Eq. (3.4) we see that the two are generally different. They are only the same when $\langle n|G\rangle = \langle n\rangle$, i.e., the mean number of proteins conditional on the gene being in state G is equal to the mean number of proteins which occurs when gene and protein number fluctuations are independent.

A.2 Exact steady state solution of non-bursty feedback loop with fast gene switching

The exact steady state solution of Eq. (A.1) has been previously reported in the literature [42, 336] and is given by:

$$P_e(n) = \frac{1}{n!}\frac{d(G_0(z)+G_1(z))^n}{dz^n}\bigg|_{z=0}, \tag{A.8}$$

$$G_0(z) = A^{-1}\exp\left(\rho_b(z-1)\right)\left(\frac{1+\sigma_b}{\sigma_b}\frac{\alpha}{\rho_u}M(1+\alpha,\beta,w(z)) - \frac{\alpha}{\rho_u-\rho_b}M(\alpha,\beta,w(z))\right), \tag{A.9}$$

$$G_1(z) = A^{-1}\exp\left(\rho_b(z-1)\right)M(\alpha,\beta,w(z)), \tag{A.10}$$

with the definitions:

$$A = \frac{1+\sigma_b}{\sigma_b}\frac{\alpha}{\rho_u}M(1+\alpha,\beta,w(1)) + M(\alpha,\beta,w(1))(1-\frac{\alpha}{\rho_u-\rho_b}), \tag{A.11}$$

$$\alpha = \frac{\sigma_u(\rho_u-\rho_b)}{\rho_u-\rho_b(1+\sigma_b)}, \tag{A.12}$$

$$\beta = 1 + \frac{\sigma_u + \sigma_b\frac{\rho_u}{1+\sigma_b}}{1+\sigma_b}, \tag{A.13}$$

$$w(z) = (\rho_u-\rho_b(1+\sigma_b))\frac{(1+\sigma_b)z-1}{(1+\sigma_b)^2}. \tag{A.14}$$

Note that $M(x,y,z)$ is the Kummer confluent hypergeometric function and $G_i(z)$ is the generating function $\sum_n z^n P_i(n)$. Replacing σ_u by σ_u/ϵ and σ_b by σ_b/ϵ and taking the limit of $\epsilon \to 0$ (the fast switching limit), we find that $G(z)$ becomes a function of only three non-dimensional parameters $L = \sigma_u/\sigma_b$, $N = \rho_u/\rho_b, \rho_b$ and the corresponding steady state distribution of protein numbers (to leading-order in ϵ) has the form:

$$P(n) = \frac{(1+L)N\rho_b^n(n\rho_b + L(L+n+N\rho_b))[1+LN]_n}{AM(1+LN,1+L,\rho_b) + BM(2+LN,2+L,\rho_b))}, \tag{A.15}$$

where

$$A = (LN+n)n!(1+L)(L+(N-1)\rho_b)[1+L]_n, \tag{A.16}$$

$$B = (LN+n)n!(1+LN)\rho_b[1+L]_n. \tag{A.17}$$

A.3 Limits of small and large L from exact steady state solutions

A.3.1 Interchanging the sum and the limit

Here we prove a result which will be used in Sections A.3.2 and A.3.3, for the purpose of interchanging limits of L with the infinite sum that defines the Kummer function (for example that in Eq. (3.11) and Eq. (3.30)). For reference, the Kummer function is defined through the sum:

$$M(\alpha,\beta,x) = \sum_{n=0}^{\infty} \frac{[\alpha]_n}{[\beta]_n}\frac{x^n}{n!}. \tag{A.18}$$

We will prove that:

$$\lim_{L\to A}\sum_{n=0}^{\infty}\frac{[a+LN]_n}{[a+L]_n}\frac{\rho_b^n}{n!} = \sum_{n=0}^{\infty}\lim_{L\to A}\frac{[a+LN]_n}{[a+L]_n}\frac{\rho_b^n}{n!}, \tag{A.19}$$

where A is some real number.

Let $f_n = [a+LN]_n/[a+L]_n$ which implies $f_{n+1} = f_n(a+n+LN)/(a+n+L)$. Consider first the case $N \geq 1$. Since $(a+n+LN)/(a+n+L) \leq N$ then if $f_n \leq N^n$ this implies that $f_{n+1} \leq N^{n+1}$. Also it is easy to check that $f_1 \leq N$. Hence by induction it follows that if $N \geq 1$ then $f_n \leq N^n$. Similarly it is straightforward to prove by induction that if $N < 1$ then $f_n < 1$. It then follows that:

$$\sum_{n=0}^{\infty}\frac{[a+LN]_n}{[a+L]_n}\frac{\rho_b^n}{n!} \leq \exp(N\rho_b), \quad N \geq 1, \tag{A.20}$$

$$\sum_{n=0}^{\infty}\frac{[a+LN]_n}{[a+L]_n}\frac{\rho_b^n}{n!} \leq \exp(\rho_b), \quad N < 1 \tag{A.21}$$

Since the sums are bounded by a finite value, it follows by the dominated convergence theorem that the limit and sum can be switched, i.e., Eq. (A.19) holds.

A.3.2 The limit of large L

Making use of a standard result [242]:

$$\lim_{x\to\infty}\frac{\Gamma(x+n)}{\Gamma(x)x^n}=1,\quad \forall n\in\mathbb{Z}, \tag{A.22}$$

and $[x]_n=\Gamma[x+n]/\Gamma[x]$, we have that for any $a\in\mathbb{R}$:

$$\lim_{L\to\infty}\frac{[a+LN]_n}{[a+L]_n}=N^n,\quad \lim_{L\to\infty}M(a+LN,a+L,\rho_b)=\sum_{n=0}^{\infty}\lim_{L\to\infty}\frac{[a+LN]_n}{[a+L]_n}\frac{\rho_b^n}{n!}=\exp\left(N\rho_b\right). \tag{A.23}$$

Note that here we have used the interchange of sum and limit as given by Eq. (A.19). Using Eq. (A.23) and $N=\rho_u/\rho_b$ it follows that in the limit of large L, the steady state distribution of molecule numbers as predicted by the reduced master equation Eq. (3.11) and by the full master equation in the limit of fast promoter switching Eq. (3.30) is a Poissonian with mean ρ_u:

$$\lim_{L\to\infty}P(n)=\lim_{L\to\infty}P_a(n)=\frac{\rho_u^n\exp\left(-\rho_u\right)}{n!} \tag{A.24}$$

A.3.3 The limit of small L

Using the definition of the Pochhammer symbol and the standard result $\Gamma(x)\sim 1/x$ as $x\to 0$, one can show that:

$$\lim_{L\to 0}\frac{[a+LN]_n}{[a+L]_n}=\lim_{L\to 0}\frac{\Gamma(a+LN+n)\Gamma(a+L)}{\Gamma(a+LN)\Gamma(a+L+n)}=1,\ \forall n\quad if\ a\neq 0,$$

and

$$\lim_{L\to 0}\frac{[LN]_n}{[L]_n}=\lim_{L\to 0}\frac{\Gamma(LN+n)\Gamma(L)}{\Gamma(LN)\Gamma(L+n)}=\begin{cases}1, & \text{if } n=0,\\ N, & \text{if } n\geq 1\end{cases}$$

Using these two results, it then follows that:

$$\lim_{L\to 0}M(a+LN,a+L,\rho_b)=\sum_{n=0}^{\infty}\lim_{L\to 0}\frac{[a+LN]_n}{[a+L]_n}\frac{\rho_b^n}{n!}=\exp\left(\rho_b\right)\quad if\ a\neq 0$$

and

$$\lim_{L\to 0}M(LN,L,\rho_b)=\sum_{n=0}^{\infty}\lim_{L\to 0}\frac{[LN]_n}{[L]_n}\frac{\rho_b^n}{n!}=1+N(\exp\left(\rho_b\right)-1)\quad if\ a=0$$

Note that here we have used the interchange of sum and limit as given by Eq. (A.19). Using these results it is straightforward to show that in the limit of small L, the steady state distribution of molecule numbers predicted by the heuristic master equation Eq. (3.11) reduces to:

$$\lim_{L\to 0}P_a(n)=\begin{cases}\frac{1}{1+N(\exp\left(\rho_b\right)-1)}, & \text{if } n=0,\\ \frac{\exp\left(-\rho_b\right)\rho_b^n}{n!}\left(1+\frac{N-1}{1+N(\exp\left(\rho_b\right)-1)}\right), & \text{if } n\geq 1\end{cases} \tag{A.25}$$

while the steady state distribution of molecule numbers predicted by the master equation in the limit of fast promoter switching Eq. (3.30) reduces to:

$$\lim_{L\to 0} P(n) = \frac{\rho_b^n \exp\left(-\rho_b\right)}{n!}. \tag{A.26}$$

Clearly $P(n) \neq P_a(n)$ since the latter is not-Poissonian and hence shows that in the small L limit, the approximate heuristic master equation gives the incorrect answer.

Appendix B

Chapter 3 Appendices

B.1 Stochastic simulations of autoregulation with extrinsic noise

In this paper extrinsic noise is accounted for in the SSA through the introduction of a new ghost species Y and some new ghost reactions. For example, consider the case where we want to model a fluctuating degradation rate $d = d_0(1 + \eta(t))$, where $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(t') \rangle = (D/\tau) \exp(-|t - t'|/\tau)$. We will then replace the degradation reaction, $P \xrightarrow{d} \varnothing$, by the following set of reactions:

$$\varnothing \underset{1/\tau}{\overset{1/(D\Omega)}{\rightleftharpoons}} Y, \; P + Y \xrightarrow{d_0 D\Omega/\tau} Y. \tag{B.1}$$

We now show why this equivalence exists. The *propensity* for the degradation reaction in Eq. (B.1) is $(n_P n_Y d_0 D\Omega/\tau)/\Omega$, meaning that the *effective degradation rate* is $d = (n_Y d_0 D\Omega/\tau)/\Omega$. Assuming there are large numbers of Y, we apply the van Kampen ansatz that fluctuations in Y occur around its deterministic steady state mean [8]:

$$\frac{n_Y}{\Omega} = \frac{\tau}{D\Omega} + \Omega^{-1/2}\epsilon(t). \tag{B.2}$$

Then, employing the system size expansion, and enforcing the linear noise approximation (LNA), we obtain a linear FPE for the probability of having a fluctuation of size $\epsilon(t)$ at a time t, denoted $\Pi(\epsilon, t)$ [8, 87]:

$$\frac{\partial \Pi(\epsilon, t)}{\partial t} = \frac{1}{\tau}\frac{\partial}{\partial \epsilon}(\epsilon \Pi(\epsilon, t)) + \frac{1}{2}\frac{2}{D\Omega}\frac{\partial^2 \Pi(\epsilon, t)}{\partial \epsilon^2}. \tag{B.3}$$

This FPE then admits an equivalent Langevin equation given by:

$$\frac{d\epsilon(t)}{dt} = -\frac{1}{\tau}\epsilon(t) + \sqrt{\frac{2}{D\Omega}}\beta(t), \tag{B.4}$$

where $\beta(t)$ is Gaussian white noise with zero mean and correlator $\langle \beta(t)\beta(t') \rangle = \delta(t - t')$. Hence, from Eq. (B.2) it follows that d goes as:

$$d = d_0 D\frac{\Omega}{\tau}\frac{n_Y}{\Omega} = d_0(1 + \eta(t)), \tag{B.5}$$

$$\frac{d\eta(t)}{dt} = -\frac{1}{\tau}\eta(t) + \frac{\sqrt{2D}}{\tau}\beta(t), \tag{B.6}$$

where $\eta(t) = \Omega^{1/2} D\epsilon(t)/\tau$. Eqs. (B.5) and (B.6) are consistent with the definition of colored noise described at the beginning of this section. This modified SSA requires that where τ and D are both individually large, that $\tau \gg D$ such that slow switching is not enforced between differing numbers of the ghost species.

B.2 Detailed explanation of condition 2

In order to explain the origin of condition 2—a condition on the length scale of colored noise fluctuation compared to the rate of variation of the drift term in Eq. (4.47)—we will first consider a more intuitive example. Consider a Brownian particle subject to a force $F(x)$, whose state is specified by both its position x, as well as its velocity v. The set of SDEs governing the state of this particle is then [8]:

$$\frac{dx}{dt} = v, \tag{B.7}$$

$$\frac{dv}{dt} = \frac{F(x)}{m} - \gamma v + \sqrt{\frac{k_b T \gamma}{m}}\Gamma(t), \tag{B.8}$$

where m is the mass of the particle, γ is the damping coefficient of the frictional force surrounding the particle (frictional force is $-\gamma m v$), $k_b T$ is the thermal energy of the particle, and $\Gamma(t)$ is Gaussian white noise with zero mean and correlator $\langle \Gamma(t)\Gamma(t')\rangle = \delta(t - t')$. The equivalent multivariate FPE for this set of SDEs is [74]:

$$\frac{\partial P(x, v; t)}{\partial t} = \gamma \left[\frac{\partial (vP)}{\partial v} + \frac{k_b T}{m} \frac{\partial^2 P}{\partial v^2} \right] - v \frac{\partial P}{\partial x} - \frac{F(x)}{m} \frac{\partial P}{\partial v}. \tag{B.9}$$

Now following van Kampen p. 216–218 [8], one can utilise singular perturbation theory assuming that the damping coefficient γ is small (although the same procedure could be done for γ large) in order to reduce the above FPE in two variables to a FPE in the position variable x alone. The result after having done this procedure is:

$$\frac{\partial P(x; t)}{\partial t} = -\frac{\partial}{\partial x} \left(\frac{F(x)}{m\gamma} P \right) + \frac{k_b T}{m\gamma} \frac{\partial^2 P}{\partial x^2}. \tag{B.10}$$

Aside from the requirement that γ must be small, there is another condition required of Eq. (B.10) such that it reasonably approximates Eq. (B.9). This condition arises *physically* since we realise that if we are to approximate Eq. (B.9) by Eq. (B.10), then the drift term $F(x)/m\gamma$ must be approximately constant over the distance that the velocity is damped. One finds that the associated 'length scale' L over which the velocity is damped is simply the pre-factor of diffusion term in Eq. (B.10), i.e., $L = \frac{k_b T}{m\gamma}$ [337]. Enforcing the requirement that $F(x)$ is slowly varying over this length scale we find the inequality $L|F'(x)| \ll |F(x)|$, explicitly:

$$\frac{m\gamma}{k_b T} \gg \left| \frac{F'(x)}{F(x)} \right|, \tag{B.11}$$

which must be satisfied for our one variable FPE to be a good approximation. We now return to our colored noise problem and recall the set of SDEs that define our system:

$$\frac{dn}{dt} = h(n) + F(n)\eta + g_2(n)\Gamma(t), \tag{B.12}$$

$$\frac{d\eta}{dt} = -\frac{1}{\tau}\eta + \frac{1}{\tau}\theta(t), \tag{B.13}$$

where all functions of n and t are defined in Section 4.4.2. This set of SDEs has an equivalent bi-variate (Stratonovich) FPE given by:

$$\frac{\partial P(n,\eta;t)}{\partial t} = -\frac{\partial}{\partial n}\Big(\big(h(n) + F(n)\eta - \frac{1}{2}g_2(n)g_2'(n)\big)P\Big) + \frac{1}{\tau}\frac{\partial}{\partial \eta}(\eta P) + \frac{1}{2}\frac{\partial^2}{\partial n^2}\left(g_2(n)^2 P\right) + \frac{1}{\tau}\frac{\partial^2}{\partial \eta^2}P. \tag{B.14}$$

We now recall Eq. (4.47), i.e., our UCNA approximated one variable FPE, where η was adiabatically eliminated:

$$\frac{\partial P(n;t)}{\partial t} = -\frac{\partial}{\partial n}\left[\big(\tilde{h}(n) + \tilde{g}(n)\tilde{g}'(n)\big)\,P(n,t)\right] + \frac{\partial^2}{\partial n^2}\left[\tilde{g}(n)^2 P(n,t)\right]. \tag{B.15}$$

Analogously to the case of the Brownian particle, this one variable FPE can only be approximately correct where the variation of the drift term with respect to the length scale of colored noise fluctuations is small. From Eq. (4.43), we identify our length scale as the pre-factor of the noise term whose origin is the adiabatic elimination of η, i.e.,

$$L = \frac{F(n)}{C(n,\tau)}. \tag{B.16}$$

Hence, $C(n,\tau)$ must satisfy the following length scale condition

$$C(n,\tau) \gg F(n)\left|\frac{\partial_n\left(\tilde{h}(n) + \tilde{g}(n)\tilde{g}'(n)\right)}{\tilde{h}(n) + \tilde{g}(n)\tilde{g}'(n)}\right| \tag{B.17}$$

if Eq. (B.15) is to be a good approximation of Eq. (B.14), as seen in Eq. (4.66) from the main text. Note that one can also make the argument that the diffusion term $\tilde{g}(n)^2$ should also slowly vary with respect to L. However, we generally find that this is satisfied if Eq. (B.17) is satisfied, and hence we do not include this as an additional condition on the validity of the UCNA applied to cooperative genetic auto-regulation. In application to other reaction networks one should again test whether this heuristic holds.

Appendix C

Chapter 4 Appendices

C.1 Waiting time calculations for two-state models

C.1.1 Derivation of the waiting time distribution and its moments

Consider the delayed telegraph model describing active Pol II dynamics:

$$G \underset{\sigma_b}{\overset{\sigma_u}{\rightleftharpoons}} G^\star, \quad G \xrightarrow{\rho} G + A, \quad A \underset{\tau}{\Rightarrow} \varnothing. \tag{C.1}$$

We want to calculate the distribution of the waiting time between the production of two consecutive active Pol II molecules (A) along the gene. In other words, given that a paused Pol II has just been released and become active, what is the distribution of the time before the next Pol II becomes active? Note that the mechanism of removal of active Pol II does not influence the statistics of the production events. Hence the calculation that proceeds remains the same if instead of the delayed telegraph model, we had to use the telegraph model to calculate the waiting time distribution between two consecutive mature mRNAs.

We define three states: state X where the gene is in state G and the number of active Pol II is n; state Y where the gene is in state $G^\star$ and the number of active Pol II is n; state Z where the gene is in state G and the number of active Pol II is $n+1$. Hence the effective reaction scheme describing these three states is

$$X \underset{\sigma_b}{\overset{\sigma_u}{\rightleftharpoons}} Y, \quad X \xrightarrow{\rho} Z. \tag{C.2}$$

Immediately after an active molecule of Pol II is produced, the gene is in state G and hence our initial condition is state X. The absorbing state is state Z. The master equations describing the effective reaction scheme are:

$$\begin{aligned} \partial_t P_X(t) &= -(\sigma_u + \rho) P_X(t) + \sigma_b P_Y(t), \\ \partial_t P_Y(t) &= \sigma_u P_X(t) - \sigma_b P_Y(t), \end{aligned} \tag{C.3}$$

with initial condition $P_X(0) = 1, P_Y(0) = 0$ and $P_Z(0) = 0$. The distribution $f(t)$ of the time t at which the system enters the absorbing state Z is given by the probability that the system is in state X at time t multiplied by the rate of switching from state X to Z, i.e., $f(t) = \rho P_X(t)$. Solving the differential equations Eq. (C.3) using the Laplace transform we obtain:

$$\tilde{f}(s) = \frac{\rho(s + \sigma_b)}{(\rho + s)(s + \sigma_b) + s\sigma_u}, \tag{C.4}$$

where $\tilde{f}(s) = \int_0^\infty f(t)e^{-st}dt$. It then follows that the first three moments of the time between two consecutive active Pol II production events are given by:

$$\begin{aligned}
\langle t \rangle &= -\partial_s \tilde{f}(0) = \frac{\sigma_b + \sigma_u}{\rho\sigma_b}, \\
\langle t^2 \rangle &= \partial_s^2 \tilde{f}(0) = \frac{2\left((\sigma_b + \sigma_u)^2 + \rho\sigma_u\right)}{\rho^2\sigma_b^2}, \\
\langle t^3 \rangle &= -\partial_s^3 \tilde{f}(0) = \frac{6\left((\sigma_b + \sigma_u)\left((\sigma_b + \sigma_u)^2 + 2\rho\sigma_u\right) + \rho^2\sigma_u\right)}{\rho^3\sigma_b^3}.
\end{aligned} \tag{C.5}$$

The square of the coefficient of variation of the waiting time (the randomness parameter) is:

$$R^{tele} = \frac{\langle t^2 \rangle - \langle t \rangle^2}{\langle t \rangle^2} = 1 + \frac{2\rho\sigma_u}{(\sigma_b + \sigma_u)^2}. \tag{C.6}$$

Note that $R^{tele} > 1$ for all parameter values. For reference, the exponential distribution is characterized by a coefficient of variation squared equal to 1.

C.1.2 Proof of the monotonicity of the waiting time distribution

Here we prove that the waiting time distribution of the delayed telegraph model (and of the telegraph model) is a monotonically decreasing function. We start by rewriting Eq. (C.4) in the form:

$$\tilde{f}(s) = \sum_{k=1}^{2} \beta_k (1 + sa_k)^{-1}, \tag{C.7}$$

where

$$\beta_1 + \beta_2 = 1, \tag{C.8}$$

$$a_2\beta_1 + a_1\beta_2 = \frac{1}{\sigma_b}, \tag{C.9}$$

$$a_1 + a_2 = \frac{\rho + \sigma_u + \sigma_b}{\rho\sigma_b}, \tag{C.10}$$

$$a_1 a_2 = \frac{1}{\rho\sigma_b}. \tag{C.11}$$

Taking the inverse Laplace transform of Eq. (C.7) one can show that

$$f(t) = \beta_1 \frac{e^{-t/a_1}}{a_1} + \beta_2 \frac{e^{-t/a_2}}{a_2}, \tag{C.12}$$

$$\partial_t f(t) = -\frac{\beta_1}{a_1^2} e^{-t/a_1} - \frac{\beta_2}{a_2^2} e^{-t/a_2}. \tag{C.13}$$

To determine if $f(t)$ is monotonically decreasing in t, we need to know what is the sign of $a_1, a_2, \beta_1, \beta_2$. From Eqs. (C.10) and (C.11), since the right hand sides of both equations are positive then a_1, a_2 must also be positive (if one or both are negative then the sign of the left hand side will not match the sign on the right hand side of one of the two equations). Also by solving Eqs. (C.8) and (C.9) simultaneously for $\beta_{1,2}$ one finds that these are positive. Because $a_1, a_2, \beta_1, \beta_2 > 0$, it follows from Eq. (C.13) that $\partial_t f(t) < 0$ for all times and hence $f(t)$ is a monotonic decreasing function of time t. Furthermore, by the initial value theorem [338] and Eq. (C.4), we have $f(0) = \lim_{s\to\infty} s\tilde{f}(s) = \rho$.

C.2 Waiting time calculations for the mechanistic model

C.2.1 Derivation of the waiting time distribution and its moments

In this section, we extend the analysis of Appendix C.1 to study the mechanistic model, which is given by:

$$U \underset{a'}{\overset{a}{\rightleftharpoons}} U^\star \underset{b'}{\overset{b}{\rightleftharpoons}} U^{\star\star}, \quad U^{\star\star} \xrightarrow{c} U^\star + A, \quad A \underset{\tau}{\Rightarrow} M \underset{d}{\rightarrow} \varnothing. \tag{C.14}$$

We now derive the distribution of the time between two consecutive active Pol II production events and also the same but for mature mRNA M. We first consider the active Pol II case. We define four states: state W where the gene is in state U and the number of active Pol II is n; state X where the gene is in state $U^\star$ and the number of active Pol II is n; state Y where the gene is in state $U^{\star\star}$ and the number of active Pol II is n; state Z where the gene is in state $U^\star$ and the number of active Pol II is $n+1$. Hence the effective reaction scheme describing these four states is

$$W \underset{a'}{\overset{a}{\rightleftharpoons}} X \underset{b'}{\overset{b}{\rightleftharpoons}} Y, \quad Y \xrightarrow{c} Z. \tag{C.15}$$

Just after an active Pol II is produced, the gene is in state $U^\star$ and hence our initial condition is state X. The absorbing state is state Z. The master equations describing the effective reaction scheme are:

$$\begin{aligned}
\partial_t P_W(t) &= -aP_W(t) + a'P_X(t), \\
\partial_t P_X(t) &= aP_W(t) + b'P_Y(t) - (a'+b)P_X(t), \\
\partial_t P_Y(t) &= bP_X(t) - (b'+c)P_Y(t),
\end{aligned} \tag{C.16}$$

with initial condition $P_X(0) = 1, P_W(0) = P_Y(0) = P_Z(0) = 0$. The distribution $f(t)$ of the time t at which the system enters the absorbing state Z is given by the probability that the system is in state Y at time t multiplied by the rate of switching from state Y to Z, i.e., $f(t) = cP_Y(t)$. Solving the differential equations Eq. (C.16) using the Laplace transform we obtain:

$$\tilde{f}(s) = \frac{bc(a+s)}{s\left(a'\left(b'+d+s\right)+s\left(b'+c+s\right)+b(c+s)\right)+a\left(s\left(b'+c+s\right)+b(c+s)\right)}. \tag{C.17}$$

From the definition of the Laplace transform, we have that the moments are given by

$$\langle t^i \rangle = (-1)^i \partial_s^i \tilde{f}(0). \tag{C.18}$$

The square of the coefficient of variation squared of the time between two consecutive production events (the randomness parameter) is:

$$R^{mec} = \frac{\langle t^2 \rangle - \langle t \rangle^2}{\langle t \rangle^2} = 1 + \frac{2bc\left(a'\left(-a+b'+c\right)-a^2\right)}{\left(a'\left(b'+c\right)+ab'+a(b+c)\right)^2}. \tag{C.19}$$

Note that depending on the parameter values, R^{mec} can be greater than or less than one (unlike for two-state models where it was shown in Appendix C.1 that the randomness parameter is always greater than one).

Suppose there is a fixed time τ between the production of an active Pol II and the production of a mature mRNA (via elongation and termination). It follows that the time between two consecutive mature mRNA production events is precisely the same as the time between two consecutive Pol II activation events, i.e., all the waiting time statistics that we have derived for active Pol II also hold for mature mRNA too.

C.2.2 Some properties of the waiting time distribution

We note that since $\tilde{f}(s)$ in Eq. (C.17) can be written in the form $\tilde{f}(s) = \sum_{k=1}^{3} \gamma_k (1 + sc_k)^{-1}$ (for particular values of the constants γ_k and c_k), it follows that

$$f^{mec}(t) = \gamma_1 \frac{e^{-t/c_1}}{c_1} + \gamma_2 \frac{e^{-t/c_2}}{c_2} + \gamma_3 \frac{e^{-t/c_2}}{c_3}. \tag{C.20}$$

This is unlike that for two-state models in Appendix A where the waiting time distribution was a sum of two exponentials.

Also by the initial value theorem and Eq. (C.17), we have that $f(0) = \lim_{s\to\infty} s\tilde{f}(s) = 0$. As well necessarily for any distribution we have that $\lim_{t\to\infty} f(t) = 0$. Hence it follows by the behavior of $f(t)$ at $t = 0$ and $t = \infty$, that the positive function $f(t)$ must achieve one or more maxima at intermediate times. Hence the waiting time distribution for the mechanistic model is non-monotonic in time t (unlike for two-state models, which have a monotonic waiting time distribution).

C.3 Steady state mean and variance of A in the mechanistic model

We first calculate the statistics of the accumulated active Pol II on the gene, i.e., ignoring its removal due to elongation. Hence we want to derive the time-dependent first and second moments of the reaction scheme:

$$U \underset{a'}{\overset{a}{\rightleftharpoons}} U^\star \underset{b'}{\overset{b}{\rightleftharpoons}} U^{\star\star}, \quad U^{\star\star} \xrightarrow{c} U^\star + A. \tag{C.21}$$

The easiest way to calculate these moments is using *the linear-noise approximation, which is exact up to second-order moments for any system with linear propensities (as in our case)*. The stoichiometric matrix and the propensity (column) vector are given by:

$$\mathbf{S} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{C.22}$$

$$\vec{f} = (a\langle U\rangle, a'\langle U^\star\rangle, b\langle U^\star\rangle, b'(1-\langle U\rangle-\langle U^\star\rangle), c(1-\langle U\rangle-\langle U^\star\rangle)), \tag{C.23}$$

where $\langle\psi\rangle$ denotes the average number of molecules of species ψ. The species are numbered in the order $U, U^\star, A$ and the reactions in the order $U \to U^\star, U^\star \to U, U^\star \to U^{\star\star}, U^{\star\star} \to U^\star, U^{\star\star} \to U^\star + A$. The matrix element $[\mathbf{S}]_{ij}$ is the net change in the number of molecules of species i when reaction j occurs, and the vector element f_j is the average propensity of the j^{th} reaction. Note that we have used the conservation law $\langle U^{\star\star}\rangle = 1 - \langle U\rangle - \langle U^\star\rangle$ to simplify the vector $\vec{f}$.

The equations for the first two moments are given by:

$$\frac{d}{dt}\langle\vec{n}\rangle = \mathbf{S}\cdot\vec{f}, \tag{C.24}$$

$$\frac{d}{dt}\mathbf{C} = \mathbf{J}\cdot\mathbf{C} + \mathbf{C}\cdot\mathbf{J}^T + \mathbf{D}, \tag{C.25}$$

where $\langle n_i\rangle$ is the average number of molecules of species i and $[\mathbf{C}]_{ij} = C_{ij}$ is the covariance between species i and j. Furthermore we have defined the matrix $\mathbf{J}$ as the Jacobian of the rate equations Eq. (C.24) and $\mathbf{D}$ as the diffusion matrix which equals $\mathbf{D} = \mathbf{S}\cdot\mathrm{Diag}(\vec{f})\cdot\mathbf{S}^T$. The matrix $\mathrm{Diag}(\vec{f})$ is a diagonal matrix with diagonal elements given by the elements of the vector $\vec{f}$.

The time-dependent solution of these equations is quite complex since we have three interacting species. However, the calculation is much simplified if one makes use of the fact that $U, U^\star, U^{\star\star}$ will reach a steady state after some time. This implies that

$$\frac{d\langle n_1\rangle}{dt} = \frac{d\langle n_2\rangle}{dt} = \frac{d\langle C_{11}\rangle}{dt} = \frac{d\langle C_{12}\rangle}{dt} = \frac{d\langle C_{13}\rangle}{dt} = \frac{d\langle C_{22}\rangle}{dt} = \frac{d\langle C_{23}\rangle}{dt} = 0,$$

which leads to the solutions:

$$\langle n_1 \rangle = \langle U \rangle = \frac{a'\left(b'+c\right)}{a'\left(b'+c\right)+a\left(b'+b+c\right)}, \tag{C.26}$$

$$\langle n_2 \rangle = \langle U^\star \rangle = \frac{a\left(b'+c\right)}{a'\left(b'+c\right)+a\left(b'+b+c\right)}, \tag{C.27}$$

$$C_{11} = \frac{aa'\left(b'+c\right)\left(b'+b+c\right)}{\left(a'\left(b'+c\right)+a\left(b'+b+c\right)\right)^2}, \tag{C.28}$$

$$C_{12} = -\frac{aa'\left(b'+c\right)^2}{\left(a'\left(b'+c\right)+a\left(b'+b+c\right)\right)^2}, \tag{C.29}$$

$$C_{13} = \frac{abca'\left(ab-\left(b'+c\right)\left(b'+b+c\right)\right)}{\left(a'\left(b'+c\right)+a\left(b'+b+c\right)\right)^3}, \tag{C.30}$$

$$C_{22} = \frac{a\left(b'+c\right)\left(a'\left(b'+c\right)+ab\right)}{\left(a'\left(b'+c\right)+a\left(b'+b+c\right)\right)^2}, \tag{C.31}$$

$$C_{23} = \frac{abc\left(a'\left(b'+c\right)^2+a^2b\right)}{\left(a'\left(b'+c\right)+a\left(b'+b+c\right)\right)^3}. \tag{C.32}$$

However, since active Pol II keeps accumulating with time, we have to solve the time-dependent equations for its mean and variance which from Eqs. (C.24) and (C.25) are given by:

$$\begin{aligned}
&\frac{d}{dt}\langle n_3 \rangle = c(1-\langle n_1 \rangle - \langle n_2 \rangle), \\
&\frac{d}{dt}C_{33} = c\left(\frac{ab}{a'\left(b'+c\right)+a\left(b'+b+c\right)} - 2C_{13} - 2C_{23}\right).
\end{aligned} \tag{C.33}$$

Substituting Eqs. (C.26), (C.27), (C.30) and (C.32) in Eq. (C.33) and solving the resulting differential equations with zero initial conditions, we finally obtain the time-dependent mean and variance of the accumulated active Pol II:

$$\langle n_3(t) \rangle = \frac{abc}{a'\left(b'+c\right)+a\left(b'+b+c\right)}t, \tag{C.34}$$

$$C_{33}(t) = \frac{abc\left(2cb'\left(ba'+\left(a'+a\right)^2\right)+\left(\left(a'+a\right)b'+ab\right)^2+c^2\left(2ba'+\left(a'+a\right)^2\right)\right)}{\left(a'\left(b'+c\right)+a\left(b'+b+c\right)\right)^3}t. \tag{C.35}$$

Hence the Fano factor of accumulated active Pol II is given by:

$$\mathrm{FF}_A^a = \frac{C_{33}}{\langle n_3 \rangle} = 1 + \frac{2bc\left(a'\left(-a+b'+c\right)-a^2\right)}{\left(a'\left(b'+c\right)+a\left(b'+b+c\right)\right)^2}. \tag{C.36}$$

Note that $\mathrm{FF}_A^a = R^{mec}$ given by Eq. (5.36). This equivalence between the Fano factor of accumulated products and the coefficient of variation of the waiting times has been previously reported in the single enzyme molecule literature [54].

Next we use these results to calculate the mean and variance of active Pol II in steady state conditions, i.e., the statistics of active Pol II due to both binding and unbinding reactions. Let the number of observed active Pol II at time t be $n(t)$; then it follows that if elongation happens after a deterministic time τ we can write:

$$n(t) = n_3(t) - n_3(t-\tau), \tag{C.37}$$

where $n_3(t)$ is the number of active Pol II accumulated up till time t. This relationship between the observed number of active Pol II and the number of accumulated active Pol II follows from the elongation dynamics: since all active Pol II molecules have a fixed lifetime of τ, it follows that molecules produced before time $t-\tau$ must have all died by time t and only those produced in the interval $(t-\tau, t]$ will contribute to the number of observed molecules at time t. Hence the first two moments of the observed active Pol II at time t are given by:

$$\langle n(t)\rangle = \langle n_3(t)\rangle - \langle n_3(t-\tau)\rangle, \tag{C.38}$$

$$\begin{aligned}\mathrm{Var}(n) =& \langle n(t)^2\rangle - \langle n(t)\rangle^2 = C_{33}(t) + C_{33}(t-\tau) - 2(\langle n_3(t)n_3(t-\tau)\rangle \\ &- \langle n_3(t)\rangle\langle n_3(t-\tau)\rangle).\end{aligned} \tag{C.39}$$

The equation for the steady state mean Eq. (C.38) can be easily evaluated by means of Eq. (C.34) leading to:

$$\langle n\rangle = \frac{abc}{a'(b'+c) + a(b'+b+c)}\tau. \tag{C.40}$$

To calculate the steady state variance of observed active Pol II, we need to first evaluate the correlator $\langle n_3(t)n_3(t-\tau)\rangle - \langle n_3(t)\rangle\langle n_3(t-\tau)\rangle$ which appears on the right-hand side of Eq. (C.39). Following Gardiner [74], for any linear system, the autocorrelation vector in steady state conditions $\vec{\epsilon}(t)$ with elements

$$\epsilon_i(t) = \langle n_i(t)n_i(t_0)\rangle - \langle n_i(t)\rangle\langle n_i(t_0)\rangle, \tag{C.41}$$

obeys the differential equation:

$$\frac{d}{dt}\vec{\epsilon} = \mathbf{J}\cdot\vec{\epsilon}, \tag{C.42}$$

with the initial condition given by $\mathbf{C}(t=t_0)$. Hence we have

$$\vec{\epsilon}(t) = \exp\left(-(t-t_0)\mathbf{J}\right)\cdot\mathbf{C}(t_0). \tag{C.43}$$

Choose $t_0 = t - \tau$, it follows that the correlator $\langle n_3(t)n_3(t-\tau)\rangle - \langle n_3(t)\rangle\langle n_3(t-\tau)\rangle$ is equal to $\epsilon_3(t)$. Note that $\mathbf{C}(t_0) = \mathbf{C}(t-\tau)$ has elements given Eqs. (C.28)-(C.32) and (C.35). Hence, we can finally evaluate Eq. (C.39):

$$\begin{aligned}\mathrm{Var}(n) =& \tau \frac{abc\left(2cb'\left(ba' + (a'+a)^2\right) + (a'b' + a(b'+b))^2 + c^2\left(2ba' + (a'+a)^2\right)\right)}{(a'(b'+c) + a(b'+b+c))^3} - A_0 + \\ & A_1 \exp\left(-\frac{1}{2}\tau\left(-\sqrt{(-a'+a-b'-b-c)^2 + 4a'(a-b'-c)} + a' + a + b' + b + c\right)\right) + \\ & A_2 \exp\left(-\frac{1}{2}\tau\left(\sqrt{(-a'+a-b'-b-c)^2 + 4a'(a-b'-c)} + a' + a + b' + b + c\right)\right),\end{aligned} \tag{C.44}$$

where

$$A_0 = \frac{2ab^2c^2\left(-a'(a-b'-c)(a'+2a+b'+b+c) - a^3\right)}{(a'(b'+c) + a(b'+b+c))^4}, \tag{C.45}$$

$$A_1 + A_2 = A_0, \tag{C.46}$$

$$\begin{aligned}A_1 - A_2 =& \frac{2ab^2c^2\left(a^3(-a+b'+b+c) + (-3a-2b)(a')^2(a-b'-c) + (a')^3(-a+b'+c)\right)}{(a'(b'+c) + a(b'+b+c))^4\sqrt{(a'-a+b'+b+c)^2 + 4a'(a-b'-c)}} + \\ & \frac{2ab^2c^2a'\left(b'\left(3a^2 - ab + b'(b'+2b+3c) + (b+c)(b+3c)\right) - a^2(b-3c) - 3a^3 - ab(b+c) + c(b+c)^2\right)}{(a'(b'+c) + a(b'+b+c))^4\sqrt{(a'-a+b'+b+c)^2 + 4a'(a-b'-c)}}.\end{aligned} \tag{C.47}$$

Note that A_1 and A_2 are the solution of the simultaneous equations Eqs. (C.46) and (C.47).

C.4 Derivation of the steady state mean and variance of mature mRNA numbers for the mechanistic model

The statistics of mature mRNA numbers can be derived much more straightforwardly than those of the active number of Pol II. In steady state, the flux across a system of species connected by irreversible reactions will be the same for each species and hence deletion of an intermediate species has no effect on the statistics of a downstream species. Hence, for the purpose of studying mature mRNA statistics in the steady state [201], instead of the full scheme (5.1), we can consider a reduced scheme where the active Pol II is not explicitly described:

$$U \underset{a'}{\overset{a}{\rightleftharpoons}} U^\star \underset{b'}{\overset{b}{\rightleftharpoons}} U^{\star\star}, \quad U^{\star\star} \xrightarrow{c} U^\star + M, \quad M \xrightarrow{d} \varnothing. \tag{C.48}$$

The stoichiometric matrix and the propensity vector are given by:

$$\mathbf{S}_M = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}, \tag{C.49}$$

$$\vec{f}_M = (a\langle U\rangle, a'\langle U^\star\rangle, b\langle U^\star\rangle, b'(1-\langle U\rangle) - \langle U^\star\rangle), c(1-\langle U\rangle - \langle U^\star\rangle), d\langle M\rangle)\,, \tag{C.50}$$

where $\langle X\rangle$ denotes the average number of molecules of species X. The species are numbered in the order $U, U^\star, M$ and the reactions in the order $U \to U^\star, U^\star \to U, U^\star \to U^{\star\star}, U^{\star\star} \to U^\star, U^{\star\star} \to U^\star + M, M \to \varnothing$. We have here also used the same conservation law as in the previous Appendix C.3.

The time-evolution equations for the mean numbers and covariance matrix are given by Eqs. (C.24) and (C.25) where we replace $\mathbf{S}$ by $\mathbf{S}_M$ and $\vec{f}$ by $\vec{f}_M$. Setting the time derivatives to zero and solving these equations simultaneously, we find the steady state mean and variance of mature mRNA given by $\langle n_3\rangle$ and C_{33}, respectively. The Fano factor of mature mRNA is then determined by their ratio:

$$\mathrm{FF}_M^{mec} = \frac{C_{33}}{\langle n_3\rangle} = 1 + bc\left(a'\left(b'+c\right) - a\left(a'+d\right) - a^2\right)/\chi, \tag{C.51}$$

using the definition of χ from the main text.

Table C.1: Comparison of the mean and variance of n active Pol II and m mature mRNA numbers in the mechanistic model evaluated from the exact theory (Appendices C.3 and C.4) and delay SSA (dSSA) with 10^5 samples. 6 different parameter sets are considered.

Method	$\langle n \rangle$	Var(n)	$\langle m \rangle$	Var(m)
1. $a = 0.016$ s^{-1}, $a' = 0.08$ s^{-1}, $b = 0.16$ s^{-1}, $b' = 0.016$ s^{-1}, $c = 0.24$ s^{-1}, $d = 0.002$ s^{-1}, $\tau = 273.62$ s				
Theory	6.195	17.554	14.151	27.701
dSSA	6.12	17.333	14.169	27.595
2. $a = 0.112$ s^{-1}, $a' = 0.032$ s^{-1}, $b = 0.16$ s^{-1}, $b' = 0.016$ s^{-1}, $c = 0.24$ s^{-1}, $d = 0.016$ s^{-1}, $\tau = 100$ s				
Theory	7.851	6.194	4.907	4.384
dSSA	7.664	6.08	4.908	4.38
3. $a = 0.144$ s^{-1}, $a' = 0.032$ s^{-1}, $b = 0.96$ s^{-1}, $b' = 0.16$ s^{-1}, $c = 0.24$ s^{-1}, $d = 0.002$ s^{-1}, $\tau = 273.62$ s				
Theory	43.511	37.646	99.387	92.745
dSSA	43.292	37.524	99.376	92.712
4. $a = 0.144$ s^{-1}, $a' = 0.032$ s^{-1}, $b = 1.12$ s^{-1}, $b' = 0.8$ s^{-1}, $c = 0.24$ s^{-1}, $d = 0.1$ s^{-1}, $\tau = 80$ s				
Theory	8.993	9.216	1.124	1.115
dSSA	8.904	9.127	1.124	1.114
5. $a = 0.032$ s^{-1}, $a' = 0.032$ s^{-1}, $b = 0.16$ s^{-1}, $b' = 0.016$ s^{-1}, $c = 0.32$ s^{-1}, $d = 0.002$ s^{-1}, $\tau = 273.62$ s				
Theory	16.838	36.098	38.462	61.715
dSSA	16.707	35.825	38.473	61.668
6. $a = 0.016$ s^{-1}, $a' = 0.032$ s^{-1}, $b = 0.16$ s^{-1}, $b' = 0.016$ s^{-1}, $c = 0.4$ s^{-1}, $d = 0.005$ s^{-1}, $\tau = 50$ s				
Theory	2.273	5.909	9.091	21.629
dSSA	2.185	5.6	9.096	21.611

C.5 Comparison to reduction methods using number statistics

Here we compare the waiting time moment matching approach to two well-known model reduction techniques, which are: (i) matching of the moments of the number distributions and (ii) matching of the number distributions.

The first method consists of matching the first three moments of transcript number distributions of the mechanistic and two-state models. We find the steady state mean, variance and skewness of the transcript numbers in two-state models as functions of ρ, σ_b and σ_u and then we equate them to the steady state mean, variance and skewness of the transcript numbers computed for the mechanistic model (the first two moments are in Appendices C and D while the third moments can be computed similarly by solving the moment equations). For a given set of parameters of the mechanistic model, we solve the resulting system of three equations (with three unknowns ρ,

σ_b and σ_u) numerically—this gives us the effective parameters of the two-state models. We have searched for numerical solutions throughout huge ranges of parameter space, however as shown in Fig. 5.7(A-C) and (G-I) we only find a physically meaningful solution (positive real numbers for the parameters of the two-state models which are shown by black dots in the figure) in the region of space given by Eq. (5.37) (where the mechanistic and two-state models can be matched using waiting time statistics; this is the region above the black solid line in the figure). Additionally note that the moment expressions for active Pol II for both the delayed telegraph model and mechanistic model are complicated and moment matching results in transcendental equations for the parameters ρ, σ_b and σ_u. The effective parameters for the delayed telegraph model, close to the contour lines where $\Delta = 0$, are relatively small. Thus, conventional numerical solvers struggle to find solutions close to the boundaries (see Fig. 5.7(A-C)). In contrast, Fig. 5.7(G-I) show that effective parameters for the telegraph model can be found for nearly all mechanistic model parameter sets within the region given by Eq. (5.37). This is thanks to the relative simplicity of the analytical expressions for the telegraph model (compared to the transcendental equations encountered for the delayed telegraph model). In both Fig. 5.7(A-C) and Fig. 5.7(G-I), we use the `FindRoot` function with the Newton-Raphson method in *Mathematica*, where we start very close to the parameter prediction of the waiting time moment matching method and do 2×10^4 iterations with a working precision of 200. Starting with points far from the theoretical predictions, no physical solutions are possible.

The second method consists of matching the transcript number distributions of the mechanistic and two-state models via maximum likelihood estimation (MLE). The results are presented in Fig. 5.7(D-F) and Fig. 5.7(J-L). The likelihood function is given by

$$\mathcal{L}_\theta(\{x_i\}) = \prod_{i=1}^{N_s} P(x_i|\theta), \tag{C.52}$$

where $\{x_i\}$ is a set of samples (with length N_s) generated using the delay SSA of the mechanistic model with specified parameters $\{a, a', b, b', c\}$ (each x_i represents the ith sample for the transcript number), θ is some candidate set of telegraph model parameters $\{\rho, \sigma_u, \sigma_b\}$, and $P(x_i|\theta)$ is the probability of measuring x_i given a telegraph model with parameters θ. We calculate the probability density function for a given parameter set θ using the exact solution for the delayed telegraph model [120] in Fig. 5.7(D-F) and the exact solution for the telegraph model [35] in Fig. 5.7(J-L). Next, we minimise the negative log-likelihood function

$$\theta^* = \arg\min_{\theta \in \Theta} \left(-\log \mathcal{L}_\theta(\{x_i\}) \right), \tag{C.53}$$

using the adaptive differential evolution algorithm in *Julia*'s `BlackBoxOptim` package [339] to find the optimal parameters θ^* of the two-state model, where Θ is the set of all possible two-state model parameters (essentially amounting to a choice of parameter space bounds in `BlackBoxOptim`). Using these optimal parameters we obtain the steady state distributions of active Pol II and of mature mRNA using the exact solutions of the two-state models. Finally, we compute the Hellinger distance (h) between these distributions and the corresponding ones from the mechanistic model—the distance is shown by the colour in Fig. 5.7(D-F) and Fig. 5.7(J-L). Note that in the regions where $\Delta > 0$, h is generally smaller than in the regions where $\Delta < 0$ i.e.,

the transcript number distributions found by MLE best approximate those of the mechanistic model in the region of parameter space given by Eq. (5.37) (where the mechanistic and two-state models can be matched using waiting time statistics). Therefore, we can conclude that waiting time moment matching agrees with this alternative model reduction method, albeit *the former is much more computationally efficient and accurate than the latter.*

Appendix D

Chapter 5 Appendices

D.1 Exact time-dependent solution of single enzyme system

The master equation for a single enzyme molecule (given by Eq. (6.6)) was first solved by Arányi and Tóth [58]. As the original paper is rather difficult to find, we present the solution here. The authors used marginal probability generating functions

$$G_{n_E}(z,t) = \sum_{n=0}^{N-1+n_E} z^n P(n, n_E, t) \qquad (n_E = 0, 1;\ t \geq 0) \tag{D.1}$$

to transform Eq. (6.6) into the following first-order partial differential equations:

$$\begin{cases} \dfrac{\partial G_0(z,t)}{\partial t} = -(k_1 + k_2) G_0(z,t) + k_0 \dfrac{\partial G_1(z,t)}{\partial z} \ , \\ \dfrac{\partial G_1(z,t)}{\partial t} = -k_0 z \dfrac{\partial G_1(z,t)}{\partial z} + k_1 z G_0(z,t) + k_2 G_0(z,t) \ . \end{cases} \tag{D.2}$$

By a simple substitution one can prove that the solutions have the form:

$$G_0(z,t) = \overline{\Gamma} e^{-\frac{k_1}{k_0}(z-1)} e^{-k_2 t} + \overline{\overline{\Gamma}} \frac{k_1 + k_2}{k_1 z + k_2} e^{-(k_0+k_2)t} + \sum_{i=1}^{2} \sum_{m=0}^{\infty} \Gamma_i^{(m)} \left[\frac{k_2 - (k_2 + \lambda_i^{(m)}) z}{-\lambda_i^{(m)}} \right]^{q_m} e^{\lambda_i^{(m)} t} \ , \tag{D.3}$$

$$G_1(z,t) = \Gamma^{(-1)} - \overline{\Gamma} e^{-\frac{k_1}{k_0}(z-1)} e^{-k_2 t} - \overline{\overline{\Gamma}} e^{-(k_1+k_2)t} - \sum_{i=1}^{2} \sum_{m=0}^{\infty} \Gamma_i^{(m)} \left[\frac{k_2 - (k_2 + \lambda_i^{(m)}) z}{-\lambda_i^{(m)}} \right]^{q_m+1} e^{\lambda_i^{(m)} t} \ , \tag{D.4}$$

where

$$\lambda_i^{(m)} \neq -k_2 \ , \quad q_m = -\frac{(\lambda_i^{(m)})^2 + (k_1 + k_0 + k_2)\lambda_i^{(m)} + k_0 k_2}{k_0 (k_2 + \lambda_i^{(m)})} \ , \quad i = 1, 2 \ . \tag{D.5}$$

Since G_0 and G_1 are generating functions of a system with a finite state space, i.e., the number of substrate and enzyme are bounded quantities ($n \in [0, N]$, $n_E \in [0, 1]$), they must be polynomials of a finite degree in z. Hence, the summations in Eqs. (D.3) and (D.4) must contain a finite number of terms only, meaning that $\overline{\Gamma} = \overline{\overline{\Gamma}} = 0$ (if $k_1 \neq 0$). By the same reasoning the q_m must be positive integers, i.e., $0 \leq q_m \leq N - 1$, ($q_m = m$), then the $\lambda^{(m)}$ are the roots of a quadratic

equation:

$$(\lambda^{(m)})^2 + [k_1 + k_0(m+1) + k_2]\,\lambda^{(m)} + k_0 k_2 (m+1) = 0, \quad (m = 0, 1, ..., N-1). \tag{D.6}$$

The constants Γ can be determined from the initial conditions:

$$\begin{aligned} &G_0(1,t) + G_1(1,t) = 1\ , \\ &G_0(z,0) = 0\ , \\ &G_1(z,0) = z^N\ . \end{aligned} \tag{D.7}$$

The first constraint implies that $\Gamma^{(-1)} = 1$, while the remaining two lead to a linear algebraic system for $\Gamma_i^{(m)}$ by enforcing the constraints explicitly on each coefficient of the polynomials G_0 and G_1 for each power of z. However, solving for $\Gamma_i^{(m)}$ becomes computationally expensive for larger values of N.

To summarise, the solution has the form:

$$\begin{aligned} G_0(z,t) &= \sum_{i=1}^{2} \sum_{m=0}^{N-1} \Gamma_i^{(m)} \left[\frac{k_2 - (k_2 + \lambda_i^{(m)})z}{-\lambda_i^{(m)}} \right]^m e^{\lambda_i^{(m)} t}\ ; \\ G_1(z,t) &= 1 - \sum_{i=1}^{2} \sum_{m=0}^{N-1} \Gamma_i^{(m)} \left[\frac{k_2 - (k_2 + \lambda_i^{(m)})z}{-\lambda_i^{(m)}} \right]^{m+1} e^{\lambda_i^{(m)} t}\ , \end{aligned} \tag{D.8}$$

where

$$\begin{aligned} \lambda_1^{(m)} &= -\frac{k_0(m+1) + k_1 + k_2}{2} + \frac{\sqrt{[k_0(m+1) + k_1 + k_2]^2 - 4k_0k_2(m+1)}}{2}\ ; \\ \lambda_2^{(m)} &= -\frac{k_0(m+1) + k_1 + k_2}{2} - \frac{\sqrt{[k_0(m+1) + k_1 + k_2]^2 - 4k_0k_2(m+1)}}{2}\ . \end{aligned} \tag{D.9}$$

Finally, the probabilities can be calculated from the generating functions according to

$$P(n, n_E, t) = \frac{1}{n!} \frac{\partial^n G_{n_E}(z,t)}{\partial z^n}\bigg|_{z=0}\ . \tag{D.10}$$

D.2 Figure showing the initial transient

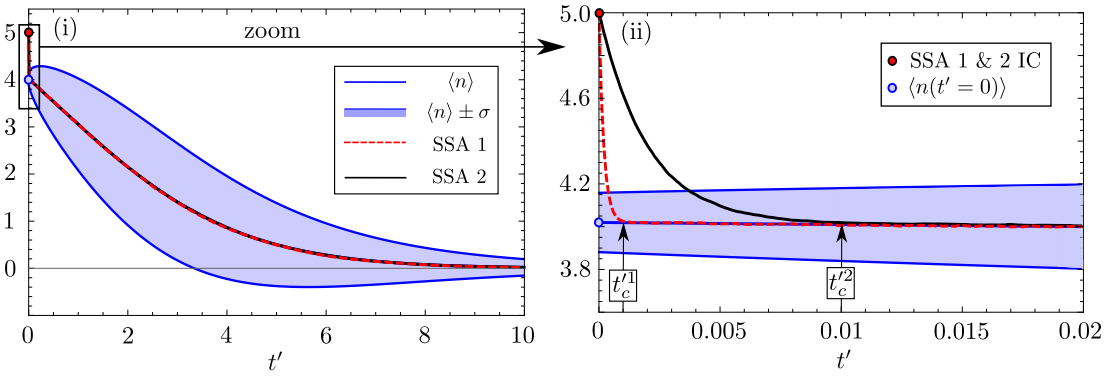


Figure D.1: Exhibition of the initial transient seen explicitly from the SSA. (i) Comparison of two differing SSA means for the same value of $k = 0.1$ against $\langle n \rangle$ from Eq. (6.22). SSA 1 was simulated with parameters $N = 5$, $M = 1$, $k_0 = 10^3$ and $k_1 = 10^2$, and SSA 2 $N = 5$, $M = 1$, $k_0 = 10^2$ and $k_1 = 10$. For most of the time course the SSA means agree with the mean predicted analytically from Eq. (6.22), aside from the initial transient very close to $t' = 0$. (ii) Zoomed in area around the initial transient. There exists some critical time t'_c for both SSA 1 and 2, denoted by t'^1_c and t'^2_c respectively, over which the mean predicted by the SSA relaxes to the mean value predicted by Eq. (6.22). In SSA 1, where k_0 and k_1 are a magnitude of 10 larger than the same parameters in SSA 2, one observes that the initial transient occurs over a much shorter time. *This relaxation of the SSA means to the mean predicted by the quasi-equilibrium analysis is known as the initial transient.* Dots of differing colour, seen in the legend, show the means of the stochastic QEA and SSAs at $t' = 0$ (IC in the legend refers to initial conditions). The mean predicted by the stochastic QEA in Eq. (6.22) reaches the quasi-equilibrium instantaneously at $t' = 0$, unlike that seen in the SSA. In both cases the SSA means were determined as an average over 10^5 individual reaction trajectories.

D.3 Derivation of Eq. (6.35)

In this appendix we prove the resulted stated in Eq. (6.35) of the main text. First consider the sum that defines $\mathcal{Z}_{m-1}$ explicitly:

$$\begin{aligned} \mathcal{Z}_{m-1} &= \sum_{i=0}^{M-g(m-1)} z_{i,m-1} \\ &= \sum_{i=0}^{M-g(m-1)} k^{-i} \left(\prod_{j=1}^{i} ((N-m+1) - -(j-1))(M-(j-1)) \right) \left(\prod_{j=i+1}^{M-g(m-1)} j \right). \end{aligned} \tag{D.11}$$

We now relabel $g(m-1) = Q$ for brevity and consider later the individual cases where $g(m-1) = 0$ for $m \leq N - M + 1$ and $g(m-1) = (m-1) - -(N-M)$ for $m > N - M + 1$. Using the definition of the Pochhammer function, $(x)_n = \prod_{j=0}^{n-1}(x+j)$, one can re-write Eq. (D.11) to give

$$\mathcal{Z}_{m-1} = \sum_{i=0}^{M-Q} k^{-i}(m-N-1)_i(-M)_i(i+1)_{M-Q-i}. \tag{D.12}$$

We now utilise the relation between the Pochhammer function and the Gamma function, namely $(x)_n = \Gamma(x+n)/\Gamma(x)$, which allows us to tactically write Eq. (D.12) as:

$$\mathcal{Z}_{m-1} = \frac{k^{-(M-Q)}\Gamma(-Q)\Gamma(m+M-N-Q-1)}{\Gamma(-M)\Gamma(m-N-1)} \times \mathcal{S}, \tag{D.13}$$

where $\mathcal{S}$ is defined by the sum,

$$\mathcal{S} = \sum_{i=0}^{M-Q} k^{M-Q-i} \frac{\Gamma(m-N-1+i)\Gamma(i-M)}{\Gamma(m+M-N-Q-1)\Gamma(-Q)} (i+1)_{M-Q-i}. \tag{D.14}$$

Our task is now to find an analytic function that is equal to the sum $\mathcal{S}$. Motivated by the Pochhammer and Gamma functions contained within the sum, we look to match this sum to the definition of a generalised hypergeometric function ${}_pF_q(\{\alpha_1, \alpha_2, ..., \alpha_{L_1}\}, \{\beta_1, \beta_2, ..., \beta_{L_2}\}, z)$ defined by:

$${}_pF_q(\{\alpha_1, \alpha_2, ..., \alpha_{L_1}\}, \{\beta_1, \beta_2, ..., \beta_{L_2}\}; z) = \sum_{n=0}^{\infty} \left(\frac{\prod_{l=1}^{L_1} (\alpha_l)_n}{\prod_{l=1}^{L_2} (\beta_l)_n} \times \frac{z^n}{n!} \right). \tag{D.15}$$

We begin by relabelling the summation index in Eq. (D.14) by $j = M - Q - i$ and again utilising the definition of the Pochhammer function in terms of Gamma functions, giving us

$$\mathcal{S} = \sum_{j=0}^{M-Q} k^j (M-Q+1-j)_j (m-N-1+M-Q)_{-j} (-Q)_{-j}. \tag{D.16}$$

Consider now the latter two Pochhammer functions in the summand of Eq. (D.16). Using the relation $(b)_{-n} = (-1)^n/(1-b)_n$ we find that:

$$(m-N-1+M-Q)_{-j} \times (-Q)_{-j} = \frac{1}{(Q+1)_j (Q+N+2-m-M)_j}. \tag{D.17}$$

Now consider the first Pochhammer function in the summand of Eq. (D.16). One can re-write this as:

$$(M-Q+1-j)_j = (-1)^j (Q-M)_j. \tag{D.18}$$

Note that $(Q-M)_j$ has the property $(Q-M)_{j>M-Q} = 0$, which is found trivially from the definition of the Pochhammer function. Using Eqs. (D.17) and (D.18), and the relation $j! = (1)_j$, one can then show that:

$$\begin{aligned} \mathcal{S} &= \sum_{j=0}^{\infty} \left(\frac{(1)_j (Q-M)_j}{(Q+1)_j (Q+N+2-m-M)_j} \times \frac{(-k)^j}{j!} \right), \\ &= {}_2F_2(\{1, Q-M\}, \{Q+1, Q+N+2-m-M\}; -k), \end{aligned} \tag{D.19}$$

using the definition of the generalised hypergeometric function in Eq. (D.15). Note, one is able to extend the upper limit of the sum defining $\mathcal{S}$ to infinity due to the property $(Q-M)_{j>M-Q} = 0$. One finds that $\mathcal{Z}_{m-1}$ is now fully specified by Eqs. (D.13) and (D.19), and we can now return to our original problem of finding the group transition rates a_m in Eq. (6.35).

In order to find a_m we must now compute $a_m = -k\partial_k(\ln(\mathcal{Z}_{m-1}))$, which using the chain rule and the differentiation rules for generalised hypergeometric functions gives:

$$\begin{aligned} a_m =&(M-Q) \\ &- \frac{k(Q-M)\,{}_2F_2(\{2, Q-M+1\},\{Q+2, Q+N+3-m-M\};-k)}{(Q+1)(Q+N+2-m-M)\,{}_2F_2(\{1, Q-M\},\{Q+1, Q+N+2-m-M\};-k)}. \end{aligned} \tag{D.20}$$

Where $m \leq N-M+1$, $Q=0$, and Eq. (D.20) becomes:

$$a_m = -M \times \left(\frac{k\,{}_1F_1(1-M, -m-M+N+3;-k)}{(-m-M+N+2)\,{}_1F_1(-M, -m-M+N+2;-k)} - 1 \right), \tag{D.21}$$

noting that for $Q=0$ the ${}_2F_2(...)$ general hypergeometric function reduces to the ${}_1F_1(...)$ confluent hypergeometric function. And finally, where $m > N-M+1$, $Q=(m-1)--(N-M)$, and Eq. (D.20) becomes:

$$a_m = -(N-m+1) \times \left(\frac{k\,{}_1F_1(m-N, m+M-N+1;-k)}{(m+M-N)\,{}_1F_1(m-N-1, m+M-N;-k)} - 1 \right), \tag{D.22}$$

where again the ${}_2F_2(...)$ general hypergeometric function reduces to the ${}_1F_1(...)$ confluent hypergeometric function. This completes our derivation of Eq. (6.35) from the main text.

Appendix E

Chapter 6 Appendices

E.1 Calculation of c_m from Sturm–Liouville theory

In this Appendix we complete the specification of the generating function from Eq. (7.9) in the main text, explaining how the coefficients c_m may be computed from an initial condition. Aside from the first coefficient that corresponds to the weight on the steady state generating function, given by $c_0 = 1/g_{1,0}(1)$, the other coefficients are determined from the initial condition. We now look to determine the non-zero coefficients $c_{m\geq 1}$ from Sturm–Liouville theory [340].

Consider a second-order linear ODE of the same type as Eq. (7.8) in the main text,

$$\left[\beta_1(z)\partial_z^2 + \beta_2(z)\partial_z + \beta_3(z)\right] f(z,t) = \partial_t f(z,t), \tag{E.1}$$

which can be solved using separation of variables to obtain the general solution

$$f(z,t) = \sum_m b_m F_m(z) e^{-\lambda_m t}, \tag{E.2}$$

where each $F_m(z)$ is a linearly-independent eigenfunction, i.e., a solution of,

$$\hat{O}F_m(z) \equiv \beta_1(z)F_m''(z) + \beta_2(z)F_m'(z) + \beta_3(z)F_m(z) = -\lambda_m F_m(z), \tag{E.3}$$

and $-\lambda_m$ are the eigenvalues of $\hat{O}$.

Sturm–Liouville theory states that the eigenfunctions will form an orthogonal basis under the w-weighted inner product in the Hilbert space $L^2([a,b], w(z)dz)$ denoted,

$$\langle F_n(z), F_m(z)\rangle \equiv \int_a^b F_n(z)F_m(z)w(z)\ \mathrm{d}z \propto \delta_{n,m}, \tag{E.4}$$

where $w(z)$ is given by,

$$w(z) = \frac{1}{\beta_1(z)} e^{\int \frac{\beta_2(z)}{\beta_1(z)}\ \mathrm{d}z}. \tag{E.5}$$

This orthogonality property then allows one to find the coefficient b_m with respect to projections onto the initial state,

$$b_m = \frac{\langle F_m(z), q(z)\rangle}{\langle F_m(z), F_m(z)\rangle}. \tag{E.6}$$

In our case, from Eq. (7.8) we find that

$$w(z) = (1-z)^{\frac{2\varepsilon}{\mu}-1} z^{-(N+\frac{\varepsilon}{\mu})} \tag{E.7}$$

and $[a,b] = [-1,1]$, as it is the region over which the generating function is defined.

One can then find the coefficients c_m as,

$$c_m = \frac{\langle g_m(z), z^{n_0}\rangle}{\langle g_m(z), g_m(z)\rangle}, \tag{E.8}$$

where the initial number of ants on the right-hand source is n_0. This completes the specification of the generating function solution which is now simply given by,

$$G(z,t) = \sum_{m=0}^{N} c_m (z-1)^m {}_2F_1\left(m+\varepsilon/\mu, m-N; 1-N-\varepsilon/\mu, z\right) e^{-m(2\varepsilon+(m-1)\mu)t}. \tag{E.9}$$

E.2 Solution to the vacillating voter model

Starting from the reaction scheme (7.30), one obtains the following ordinary differential equation for the functions $g_m(z)$:

$$\begin{aligned}
&\nu z^2(z+1)(z-1)g_m'''(z)\\
&+\nu z(z-1)(2+4z-3Nz)g_m''(z)\\
&-(N-1)(z-1)((N-1)(z+1)\varepsilon+\nu(N+2z(1-N)))g_m'(z)\\
&+(N-1)(\lambda_m+N(z-1)\varepsilon)g_m(z)=0.
\end{aligned} \tag{E.10}$$

We next obtain a recursion relation for the coefficients C_j, as

$$\begin{aligned}
&C_0 = 1, \quad (N-1)((N-1)\varepsilon+N\nu)C_1 - q(\lambda_m)C_0 = 0,\\
&R_jC_{j+1} - (Q_j + q(\lambda_m(t)))C_j + P_jC_{j-1} = 0,
\end{aligned} \tag{E.11}$$

with the condition that $C_{N+1} = 0$, and where we write

$$\begin{aligned}
q(\lambda_m) =&(N-1)(\varepsilon N - \lambda_m),\\
R_j =&(j+1)\Big(j\Big(-j^2+j-2\Big)\nu\\
&+\nu N(N-1)+(N-1)^2\varepsilon\Big),\\
Q_j =&-j\nu(3N-2)(N-j),\\
P_j =&(j-1)\nu(j-2N)(j-N-1)\\
&-(N-1)\varepsilon((j-2)N-j+1).
\end{aligned} \tag{E.12}$$

We obtain again a continued fraction relation that determines the eigenvalues λ_m,

$$q(\lambda_m(t)) = \frac{R_0 P_1}{Q_1 + q(\lambda_m(t)) -} \frac{R_1 P_2}{Q_2 + q(\lambda_m(t)) -} \cdots \frac{R_{N-1} P_N}{Q_N + q(\lambda_m(t))}. \tag{E.13}$$

Calculating λ_m from this relation, the full time-dependent solution is given using the resolvent relationship in Eq. (7.18).

Bibliography

[1] James Holehouse and Ramon Grima. Revisiting the reduction of stochastic models of genetic feedback loops with fast promoter switching. *Biophysical journal*, 117(7):1311–1330, 2019.

[2] James Holehouse, Abhishek Gupta, and Ramon Grima. Steady-state fluctuations of a genetic feedback loop with fluctuating rate parameters using the unified colored noise approximation. *Journal of Physics A: Mathematical and Theoretical*, 53(40):405601, 2020.

[3] Svitlana Braichenko, James Holehouse, and Ramon Grima. Distinguishing between models of mammalian gene expression: telegraph-like models versus mechanistic models. *Journal of the Royal Society Interface*, 18(183):20210510, 2021.

[4] James Holehouse, Augustinas Sukys, and Ramon Grima. Stochastic time-dependent enzyme kinetics: Closed-form solution and transient bimodality. *The Journal of Chemical Physics*, 153(16):164113, 2020.

[5] James Holehouse and José Moran. Exact time-dependent dynamics of discrete binary choice models. *Journal of Physics: Complexity*, 3(3):035005, 2022.

[6] James Holehouse, Zhixing Cao, and Ramon Grima. Stochastic modeling of autoregulatory genetic feedback loops: A review and comparative study. *Biophysical Journal*, 118(7):1517–1525, 2020.

[7] James Holehouse and Hector Pollitt. Non-equilibrium time-dependent solution to discrete choice with social interactions. *PloS one*, 17(5):e0267083, 2022.

[8] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*. Elsevier, Third edition, 2007.

[9] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[10] Peter S Swain, Michael B Elowitz, and Eric D Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.

[11] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature genetics*, 31(1):64, 2002.

[12] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[13] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2006.

[14] Nitzan Rosenfeld, Michael B Elowitz, and Uri Alon. Negative autoregulation speeds the response times of transcription networks. *Journal of molecular biology*, 323(5):785–793, 2002.

[15] Rutger Hermsen, David W Erickson, and Terence Hwa. Speed, sensitivity, and bistability in auto-activating signaling circuits. *PLoS Computational Biology*, 7(11):e1002265, 2011.

[16] Chen Jia, Hong Qian, Min Chen, and Michael Q Zhang. Relaxation rates of gene expression kinetics reveal the feedback signs of autoregulatory gene networks. *The Journal of Chemical Physics*, 148(9):095102, 2018.

[17] Peijiang Liu, Zhanjiang Yuan, Haohua Wang, and Tianshou Zhou. Decomposition and tunability of expression noise in the presence of coupled feedbacks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(4):043108, 2016.

[18] Chen Jia, Peng Xie, Min Chen, and Michael Q Zhang. Stochastic fluctuations can reveal the feedback signs of gene regulatory networks at the single-molecule level. *Scientific reports*, 7(1):1–9, 2017.

[19] Bhaswar Ghosh, Rajesh Karmakar, and Indrani Bose. Noise characteristics of feed forward loops. *Physical biology*, 2(1):36, 2005.

[20] Carolyn Zhang, Ryan Tsoi, Feilun Wu, and Lingchong You. Processing oscillatory signals by incoherent feedforward loops. *PLoS Computational Biology*, 12(9):e1005101, 2016.

[21] Ayan Biswas. Pathway-resolved decomposition demonstrates correlation and noise dependencies of redundant information processing in recurrent feed-forward topologies. *Physical Review E*, 105(3):034406, 2022.

[22] Zhixing Cao and Ramon Grima. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proceedings of the National Academy of Sciences*, 117(9):4682–4692, 2020.

[23] Casper Beentjes, Ruben Perez-Carrasco, and Ramon Grima. Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. *Physical Review E*, 101(3):032403, 2020.

[24] Ruben Perez-Carrasco, Casper Beentjes, and Ramon Grima. Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance. *Journal of the Royal Society Interface*, 17(168):20200360, 2020.

[25] Chen Jia, Abhyudai Singh, and Ramon Grima. Characterizing non-exponential growth and bimodal cell size distributions in fission yeast: An analytical approach. *PLoS Computational Biology*, 18(1):e1009793, 2022.

[26] Chen Jia and Ramon Grima. Frequency domain analysis of fluctuations of mRNA and protein copy numbers within a cell lineage: theory and experimental validation. *Physical Review X*, 11(2):021032, 2021.

[27] Philipp Thomas. Intrinsic and extrinsic noise of gene expression in lineage trees. *Scientific reports*, 9(1):1–16, 2019.

[28] Philipp Thomas and Vahid Shahrezaei. Coordination of gene expression noise with cell size: analytical results for agent-based models of growing cell populations. *Journal of the Royal Society Interface*, 18(178):20210274, 2021.

[29] Attila Becskei, Benjamin B Kaufmann, and Alexander van Oudenaarden. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature genetics*, 37(9):937–944, 2005.

[30] Ji Yu, Jie Xiao, Xiaojia Ren, Kaiqin Lao, and X Sunney Xie. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767):1600–1603, 2006.

[31] Samuel O Skinner, Leonardo A Sepúlveda, Heng Xu, and Ido Golding. Measuring mRNA copy number in individual *Escherichia coli* cells using single-molecule fluorescent *in situ* hybridization. *Nature protocols*, 8(6):1100–1113, 2013.

[32] Samuel O Skinner, Heng Xu, Sonal Nagarkar-Jaiswal, Pablo R Freire, Thomas P Zwaka, and Ido Golding. Single-cell analysis of transcription kinetics across the cell cycle. *Elife*, 5:e12175, 2016.

[33] Leonardo A Sepúlveda, Heng Xu, Jing Zhang, Mengyu Wang, and Ido Golding. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science*, 351(6278):1218–1222, 2016.

[34] Irina Kalita, Ira Alexandra Iosub, Sander Granneman, and Meriem El Karoui. Fine-tuning of RecBCD expression by post-transcriptional regulation is required for optimal DNA repair in *Escherichia coli*. *bioRxiv*, 2021.

[35] Jean Peccoud and Bernard Ycart. Markovian modeling of gene-product synthesis. *Theoretical population biology*, 48(2):222–234, 1995.

[36] Philipp Thomas, Guillaume Terradot, Vincent Danos, and Andrea Y Weiße. Sources, propagation and consequences of stochasticity in cellular growth. *Nature Communications*, 9(1):1–11, 2018.

[37] Ulysse Herbach, Arnaud Bonnaffoux, Thibault Espinasse, and Olivier Gandrillon. Inferring gene regulatory networks from single-cell data: a mechanistic approach. *BMC systems biology*, 11(1):105, 2017.

[38] Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[39] Mads Kaern, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, 2005.

[40] Giorgio Recordati and Tommaso Bellini. A definition of internal constancy and homeostasis in the context of non-equilibrium thermodynamics. *Experimental Physiology*, 89(1):27–38, 2004.

[41] Srividya Iyer-Biswas, Fernand Hayot, and Ciriyam Jayaprakash. Stochasticity of gene products from transcriptional pulsing. *Physical Review E*, 79(3):031911, 2009.

[42] Ramon Grima, Deena R Schmidt, and Timothy J Newman. Steady-state fluctuations of a genetic feedback loop: an exact solution. *The Journal of Chemical Physics*, 137(3):035104, 2012.

[43] José EM Hornos, Daniel Schultz, Guilherme CP Innocentini, J Wang, Aleksandra M Walczak, José N Onuchic, and Peter G Wolynes. Self-regulating gene: an exact solution. *Physical Review E*, 72(5):051907, 2005.

[44] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011.

[45] Keren B Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty gene expression in the intact mammalian liver. *Molecular cell*, 58(1):147–156, 2015.

[46] Anton JM Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, 2019.

[47] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, 2011.

[48] Koichi Takahashi, Sorin Tănase-Nicola, and Pieter R Ten Wolde. Spatio-temporal correlations can drastically change the response of a MAPK pathway. *Proceedings of the National Academy of Sciences*, 107(6):2473–2478, 2010.

[49] Guy Wiedermann, Robert A Bone, Joana Clara Silva, Mia Bjorklund, Philip J Murray, and Kim Dale. A balance of positive and negative regulators determines the pace of the segmentation clock. *Elife*, 4:e05842, 2015.

[50] Leonor Michaelis and Maud L Menten. *Die kinetik der invertinwirkung*. Universitätsbibliothek Johann Christian Senckenberg, 2007.

[51] Athel Cornish-Bowden. One hundred years of Michaelis–Menten kinetics. *Perspectives in Science*, 4:3–9, 2015.

[52] S Schnell and C Mendoza. Closed form solution for time-dependent enzyme kinetics. *Journal of Theoretical Biology*, 187(2):207–212, 1997.

[53] Santiago Schnell and Philip K Maini. A century of enzyme kinetics. Should we believe in the K_M and v_{max} estimates? *Comments on Theoretical Biology*, 8:169–187, 2003.

[54] Jeffrey R Moffitt and Carlos Bustamante. Extracting signal from noise: kinetic mechanisms from a Michaelis–Menten-like expression for enzymatic fluctuations. *The FEBS journal*, 281(2):498–517, 2014.

[55] Irina V Gopich and Attila Szabo. Theory of the statistics of kinetic transitions with application to single-molecule enzyme catalysis. *The Journal of Chemical Physics*, 124(15):154712, 2006.

[56] Éva Dóka and Gábor Lente. Stochastic mapping of the Michaelis-Menten mechanism. *The Journal of Chemical Physics*, 136(5):054111, 2012.

[57] David Schnoerr, Guido Sanguinetti, and Ramon Grima. The complex chemical Langevin equation. *The Journal of Chemical Physics*, 141:024103, 2014.

[58] P Arányi and J Tóth. A full stochastic description of the Michaelis-Menten reaction for small systems. *Acta biochimica et biophysica; Academiae Scientiarum Hungaricae*, 12(4):375–388, 1977.

[59] Alan Kirman. Ants, rationality, and recruitment. *The Quarterly Journal of Economics*, 108(1):137–156, 1993.

[60] Patrick Alfred Pierce Moran. Random processes in genetics. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 54, pages 60–71. Cambridge University Press, 1958.

[61] Alan Kirman. Whom or what does the representative individual represent? *Journal of Economic Perspectives*, 6(2):117–136, 1992.

[62] Tommaso Biancalani, Louise Dyson, and Alan J McKane. Noise-induced bistable states and their mean switching time in foraging colonies. *Physical Review Letters*, 112(3):038101, 2014.

[63] Kazuo Sano. Ants, Traders, and Fat Tails: An Application of the Kirman (1993) Model. *SSRN Electronic Journal*, 2014.

[64] Giulio Bottazzi and Pietro Dindo. An evolutionary model of firms' location with technological externalities. In *The Handbook of Evolutionary Economic Geography*, volume 1, chapter 24, pages 508–528. Edward Elgar Publishing, 2010.

[65] José Moran, Antoine Fosset, Michael Benzaquen, and Jean-Philippe Bouchaud. Schrödinger's ants: a continuous description of kirman's recruitment model. *Journal of Physics: Complexity*, 1(3):035002, 2020.

[66] Tommaso Biancalani, Louise Dyson, and Alan J McKane. The statistics of fixation times for systems with recruitment. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(1):P01013, 2015.

[67] José Moran, Antoine Fosset, Alan Kirman, and Michael Benzaquen. From ants to fishing vessels: a simple model for herding and exploitation of finite resources. *Journal of Economic Dynamics and Control*, 129:104169, 2021.

[68] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[69] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124(4):044104, 2006.

[70] Augustinas Sukys, Kaan Öcal, and Ramon Grima. Approximating solutions of the chemical master equation using neural networks. *bioRxiv*, 2022.

[71] Ankit Gupta, Christoph Schwab, and Mustafa Khammash. DeepCME: A deep learning framework for computing solution statistics of the chemical master equation. *PLoS Computational Biology*, 17(12):e1009623, 2021.

[72] Francesca Cairoli, Ginevra Carbone, and Luca Bortolussi. Abstraction of Markov Population Dynamics via Generative Adversarial Nets. In *International Conference on Computational Methods in Systems Biology*, volume 1, pages 19–35. Springer, 2021.

[73] Qingchao Jiang, Xiaoming Fu, Shifu Yan, Runlai Li, Wenli Du, Zhixing Cao, Feng Qian, and Ramon Grima. Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nature Communications*, 12(1):1–12, 2021.

[74] Crispin Gardiner. *Stochastic methods*, volume 4. Springer Berlin, 2009.

[75] Niraj Kumar, Thierry Platini, and Rahul V Kulkarni. Exact distributions for stochastic gene expression models with bursting and feedback. *Physical Review Letters*, 113(26):268105, 2014.

[76] Chen Jia and Ramon Grima. Small protein number effects in stochastic models of autoregulated bursty gene expression. *The Journal of Chemical Physics*, 152(8):084115, 2020.

[77] Max Delbrück. Statistical fluctuations in autocatalytic reactions. *The Journal of Chemical Physics*, 8(1):120–124, 1940.

[78] Kenji Ishida. Stochastic model for bimolecular reaction. *The Journal of Chemical Physics*, 41(8):2472–2478, 1964.

[79] Donald A McQuarrie. Stochastic approach to chemical kinetics. *Journal of Applied Probability*, 4(3):413–478, 1967.

[80] Ian J Laurenzi. An analytical solution of the stochastic master equation for reversible bimolecular reaction kinetics. *The Journal of Chemical Physics*, 113(8):3315–3322, 2000.

[81] David Schnoerr, Guido Sanguinetti, and Ramon Grima. Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *Journal of Physics A: Mathematical and Theoretical*, 50(9):093001, 2017.

[82] JAM Janssen. The elimination of fast variables in complex chemical reactions. II. Mesoscopic level (reducible case). *Journal of Statistical Physics*, 57(1):171–185, 1989.

[83] JAM Janssen. The elimination of fast variables in complex chemical reactions. III. Mesoscopic level (irreducible case). *Journal of Statistical Physics*, 57(1):187–198, 1989.

[84] Philipp Thomas, Arthur V Straube, and Ramon Grima. The slow-scale linear noise approximation: an accurate, reduced stochastic description of biochemical networks under timescale separation conditions. *BMC systems biology*, 6(1):39, 2012.

[85] Bence Mélykúti, Joao P Hespanha, and Mustafa Khammash. Equilibrium distributions of simple biochemical reaction systems for time-scale separation in stochastic reaction networks. *Journal of The Royal Society Interface*, 11(97):20140054, 2014.

[86] Christopher V Rao and Adam P Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *The Journal of Chemical Physics*, 118(11):4999–5010, 2003.

[87] Johan Elf and Måns Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome research*, 13(11):2475–2484, 2003.

[88] Philipp Thomas, Christian Fleck, Ramon Grima, and Nikola Popović. System size expansion using Feynman rules and diagrams. *Journal of Physics A: Mathematical and Theoretical*, 47(45):455007, 2014.

[89] Philipp Thomas and Ramon Grima. Approximate probability distributions of the master equation. *Physical Review E*, 92(1):012120, 2015.

[90] Ramon Grima, Philipp Thomas, and Arthur V Straube. How accurate are the nonlinear chemical Fokker-Planck and chemical Langevin equations? *The Journal of Chemical Physics*, 135(8):084103, 2011.

[91] Zhixing Cao and Ramon Grima. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nature Communications*, 9(1):3305, 2018.

[92] Daniel T Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425, 1992.

[93] Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, volume 1, pages 260–265, 1986.

[94] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[95] Christopher Rackauckas and Qing Nie. `Differentialequations.jl`–a performant and feature-rich ecosystem for solving differential equations in Julia. *Journal of Open Research Software*, 5(1), 2017.

[96] Milton Abramowitz, Irene A Stegun, and Robert H Romer. Handbook of mathematical functions with formulas, graphs, and mathematical tables, 1972.

[97] Francesco Tricomi. Sulle funzioni ipergeometriche confluenti. *Annali di Matematica Pura ed Applicata*, 26(1):141–175, 1947.

[98] Ankit Gupta, Jan Mikelson, and Mustafa Khammash. A finite state projection algorithm for the stationary solution of the chemical master equation. *The Journal of Chemical Physics*, 147(15):154101, 2017.

[99] Daniel T Gillespie. The multivariate Langevin and Fokker–Planck equations. *American Journal of Physics*, 64(10):1246–1257, 1996.

[100] Daniel T Gillespie. The chemical Langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[101] Ramon Grima. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. *Physical Review E*, 92(4):042124, 2015.

[102] Ramon Grima. An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in non-equilibrium steady-state conditions. *The Journal of Chemical Physics*, 133(3):07B604, 2010.

[103] Philipp Thomas, Hannes Matuschek, and Ramon Grima. Intrinsic noise analyzer: a software package for the exploration of stochastic biochemical kinetics using the system size expansion. *PloS one*, 7(6):e38518, 2012.

[104] Johan Paulsson and Måns Ehrenberg. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Physical Review Letters*, 84(23):5447, 2000.

[105] Tomás Aquino, Elsa Abranches, and Ana Nunes. Stochastic single-gene autoregulation. *Physical Review E*, 85(6):061913, 2012.

[106] Manuel Pájaro, Irene Otero-Muras, Carlos Vázquez, and Antonio A Alonso. Inherent stochasticity precludes hysteresis in gene regulatory networks. *arXiv preprint arXiv:1810.10409*, 2018.

[107] Michael Assaf, Elijah Roberts, and Zaida Luthey-Schulten. Determining the stability of genetic switches: explicitly accounting for mRNA noise. *Physical Review Letters*, 106(24):248102, 2011.

[108] JaeJun Lee and Julian Lee. Quantitative analysis of a transient dynamics of a gene regulatory network. *Physical Review E*, 98(6):062404, 2018.

[109] A Rami Tzafriri. Michaelis-Menten kinetics at high enzyme concentrations. *Bulletin of Mathematical Biology*, 65(6):1111–1129, 2003.

[110] Jae K Kim, Krešimir Josić, and Matthew R Bennett. The validity of quasi-steady-state approximations in discrete stochastic simulations. *Biophysical journal*, 107(3):783–793, 2014.

[111] Santiago Schnell and Philip K Maini. Enzyme kinetics at high enzyme concentration. *Bulletin of Mathematical Biology*, 62(3):483–499, 2000.

[112] Michael M Saint-Antoine, Ramon Grima, and Abhyudai Singh. A fluctuation-based approach to infer kinetics and topology of cell-state switching. *bioRxiv*, 2022.

[113] Stefano Bo and Antonio Celani. Multiple-scale stochastic processes: decimation, averaging and beyond. *Physics reports*, 670:1–59, 2017.

[114] Nikola Popović, Carsten Marr, and Peter S Swain. A geometric analysis of fast-slow models for stochastic gene expression. *Journal of mathematical biology*, 72(1):87–122, 2016.

[115] Chen Jia and Ramon Grima. Dynamical phase diagram of an auto-regulating gene in fast switching conditions. *The Journal of Chemical Physics*, 152(17):174110, 2020.

[116] Chen Jia and Youming Li. Analytical time-dependent distributions for gene expression models with complex promoter switching mechanisms. *bioRxiv*, 2022.

[117] Manuel Barrio, Kevin Burrage, André Leier, and Tianhai Tian. Oscillatory regulation of Hes1: discrete stochastic delay modelling and simulation. *PLoS Computational Biology*, 2(9):e117, 2006.

[118] Andre Leier and Tatiana T Marquez-Lago. Delay chemical master equation: direct and closed-form solutions. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150049, 2015.

[119] Gennady Gorin and Lior Pachter. Analytical solutions of the chemical master equation with bursty production and isomerization reactions. *bioRxiv*, 2021.

[120] Heng Xu, Samuel O Skinner, Anna M Sokac, and Ido Golding. Stochastic kinetics of nascent RNA. *Physical Review Letters*, 117(12):128101, 2016.

[121] Xiaodong Cai. Exact stochastic simulation of coupled chemical reactions with delays. *The Journal of Chemical Physics*, 126(12):124108, 2007.

[122] Xiaoming Fu, Xinyi Zhou, Dongyang Gu, Zhixing Cao, and Ramon Grima. `DelaySSAToolkit.jl`: stochastic simulation of reaction systems with time delays in Julia. *bioRxiv*, 2022.

[123] Jesper Tegner, MK Stephen Yeung, Jeff Hasty, and James J Collins. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Sciences*, 100(10):5944–5949, 2003.

[124] Mukesh Bansal, Giusy Della Gatta, and Diego Di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, 2006.

[125] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego Di Bernardo. How to infer gene networks from expression profiles. *Molecular systems biology*, 3(1):78, 2007.

[126] Abraham Berman and Robert J Plemmons. *Nonnegative matrices in the mathematical sciences*. SIAM, 1994.

[127] Stephen Smith and Vahid Shahrezaei. General transient solution of the one-step master equation in one dimension. *Physical Review E*, 91(6):062119, 2015.

[128] Peter Ashcroft, Arne Traulsen, and Tobias Galla. When the mean is not enough: Calculating fixation time distributions in birth-death processes. *Physical Review E*, 92(4):042154, 2015.

[129] Peter Ashcroft. Metastable states in a model of cancer initiation. In *The Statistical Physics of Fixation and Equilibration in Individual-Based Models*, volume 1, pages 91–126. Springer, 2016.

[130] Nicholas J Higham. *Functions of matrices: theory and computation*, volume 104. SIAM, 2008.

[131] Riaz A Usmani. Inversion of a tridiagonal Jacobi matrix. *Linear Algebra and its Applications*, 212(213):413–414, 1994.

[132] Jonathan M Raser and Erin K O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013, 2005.

[133] Arren Bar-Even, Johan Paulsson, Narendra Maheshri, Miri Carmi, Erin O'Shea, Yitzhak Pilpel, and Naama Barkai. Noise in protein expression scales with natural protein abundance. *Nature genetics*, 38(6):636, 2006.

[134] John RS Newman, Sina Ghaemmaghami, Jan Ihmels, David K Breslow, Matthew Noble, Joseph L DeRisi, and Jonathan S Weissman. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature*, 441(7095):840, 2006.

[135] Rajesh Ramaswamy, Nélido González-Segredo, Ivo F Sbalzarini, and Ramon Grima. Discreteness-induced concentration inversion in mesoscopic chemical systems. *Nature Communications*, 3:779, 2012.

[136] Tobias Jahnke and Wilhelm Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54(1):1–26, 2007.

[137] David F Anderson, Gheorghe Craciun, and Thomas G Kurtz. Product-form stationary distributions for deficiency zero chemical reaction networks. *Bulletin of Mathematical Biology*, 72(8):1947–1970, 2010.

[138] Daniele Cappelletti and Carsten Wiuf. Product-form poisson-like distributions and complex balanced reaction systems. *SIAM Journal on Applied Mathematics*, 76(1):411–432, 2016.

[139] Attila Becskei, Bertrand Séraphin, and Luis Serrano. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *The EMBO journal*, 20(10):2528–2535, 2001.

[140] Abhyudai Singh and Joao P Hespanha. Optimal feedback strength for noise suppression in autoregulatory gene networks. *Biophysical journal*, 96(10):4013–4023, 2009.

[141] Eduardo S Zeron and Moisés Santillán. Distributions for negative-feedback-regulated stochastic gene expression: Dimension reduction and numerical solution of the chemical master equation. *Journal of Theoretical Biology*, 264(2):377–385, 2010.

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

[143] Eric L Haseltine and James B Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *The Journal of Chemical Physics*, 117(15):6959–6969, 2002.

[144] John Goutsias. Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems. *The Journal of Chemical Physics*, 122(18):184102, 2005.

[145] Xingye Kan, Chang Hyeong Lee, and Hans G Othmer. A multi-time-scale analysis of chemical reaction networks: II. stochastic systems. *Journal of Mathematical Biology*, 73(5):1081–1129, 2016.

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

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

[148] Daniele Cappelletti and Carsten Wiuf. Elimination of intermediate species in multiscale stochastic reaction networks. *The Annals of Applied Probability*, 26(5):2915–2958, 2016.

[149] Jae K Kim, Grzegorz A Rempala, and Hye-Won Kang. Reduction for stochastic biochemical reaction networks with multiscale conservations. *Multiscale Modeling & Simulation*, 15(4):1376–1403, 2017.

[150] Jay Newby. Bistable switching asymptotics for the self regulating gene. *Journal of Physics A: Mathematical and Theoretical*, 48(18):185001, 2015.

[151] Nir Friedman, Long Cai, and X Sunney Xie. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Physical Review Letters*, 97(16):168302, 2006.

[152] Djomangan A Ouattara, Wassim Abou-Jaoudé, and Marcelle Kaufman. From structure to dynamics: Frequency tuning in the p53-Mdm2 network. II: Differential and stochastic approaches. *Journal of Theoretical Biology*, 264(4):1177–1189, 2010.

[153] Hao Ge, Hong Qian, and X Sunney Xie. Stochastic phenotype transition of a single cell in an intermediate region of gene state switching. *Physical Review Letters*, 114(7):078101, 2015.

[154] Philipp Thomas, Arthur V Straube, and Ramon Grima. Communication: limitations of the stochastic quasi-steady-state approximation in open biochemical reaction networks. *The Journal of chemical physics*, 135(18):181103, 2011.

[155] Philipp Thomas, Ramon Grima, and Arthur V Straube. Rigorous elimination of fast stochastic variables from the linear noise approximation using projection operators. *Physical Review E*, 86(4):041110, 2012.

[156] Jae K Kim, Krešimir Josić, and Matthew R Bennett. The relationship between stochastic and deterministic quasi-steady state approximations. *BMC systems biology*, 9(1):87, 2015.

[157] R Bundschuh, F Hayot, and C Jayaprakash. Fluctuations and slow variables in genetic networks. *Biophysical journal*, 84(3):1606–1615, 2003.

[158] Tomás Alarcón. Stochastic quasi-steady state approximations for asymptotic solutions of the chemical master equation. *The Journal of Chemical Physics*, 140(18):05B609_1, 2014.

[159] Yves Vandecan and Ralf Blossey. Self-regulatory gene: an exact solution for the gene gate model. *Physical Review E*, 87(4):042705, 2013.

[160] Frits Veerman, Carsten Marr, and Nikola Popović. Time-dependent propagators for stochastic models of gene expression: an analytical method. *Journal of Mathematical Biology*, 77(2):261–312, 2018.

[161] Philipp Thomas, Nikola Popović, and Ramon Grima. Phenotypic switching in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 111(19):6994–6999, 2014.

[162] Paul C Bressloff. Stochastic switching in biology: from genotype to phenotype. *Journal of Physics A: Mathematical and Theoretical*, 50(13):133001, 2017.

[163] David Toner and Ramon Grima. Effects of bursty protein production on the noisy oscillatory properties of downstream pathways. *Scientific reports*, 3:2438, 2013.

[164] Otto G Berg. A model for the statistical fluctuations of protein numbers in a microbial population. *Journal of Theoretical Biology*, 71(4):587–603, 1978.

[165] Marco J Morelli, Rosalind J Allen, Sorin Tănase-Nicola, and Pieter Rein ten Wolde. Eliminating fast reactions in stochastic simulations of biochemical networks: a bistable genetic switch. *The Journal of Chemical Physics*, 128(4):01B620, 2008.

[166] Stephen Smith, Claudia Cianci, and Ramon Grima. Analytical approximations for spatial stochastic gene expression in single cells and tissues. *Journal of The Royal Society Interface*, 13(118):20151051, 2016.

[167] Marc Sturrock, Andreas Hellander, Sahar Aldakheel, Linda Petzold, and Mark AJ Chaplain. The role of dimerisation and nuclear transport in the Hes1 gene regulatory network. *Bulletin of Mathematical Biology*, 76(4):766–798, 2014.

[168] Michael J Lawson, Linda Petzold, and Andreas Hellander. Accuracy of the Michaelis-Menten approximation when analysing effects of molecular noise. *Journal of The Royal Society Interface*, 12(106):20150054, 2015.

[169] Stephen Smith and Ramon Grima. Breakdown of the reaction-diffusion master equation with nonelementary rates. *Physical Review E*, 93(5):052135, 2016.

[170] Zhixing Cao, Tatiana Filatova, Diego A Oyarzún, and Ramon Grima. A stochastic model of gene expression with polymerase recruitment and pause release. *Biophysical Journal*, 119(5):1002–1014, 2020.

[171] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell*. Garland Science, sixth edition, 2015.

[172] Minoru SH Ko. A stochastic model for gene induction. *Journal of Theoretical Biology*, 153(2):181–194, 1991.

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

[174] Chen Jia, George G Yin, and Michael Q Zhang. Single-cell stochastic gene expression kinetics with coupled positive-plus-negative feedback. *Physical Review E*, 100(5):052406, 2019.

[175] Pavel Kurasov, Alexander Lück, Delio Mugnolo, and Verena Wolf. Stochastic hybrid models of gene regulatory networks–A PDE approach. *Mathematical Biosciences*, 305:170–177, 2018.

[176] Alexander Andreychenko, Luca Bortolussi, Ramon Grima, Philipp Thomas, and Verena Wolf. Distribution approximations for the chemical master equation: comparison of the method of moments and the system size expansion. In *Modeling Cellular Systems*, volume 1, pages 39–66. Springer, 2017.

[177] Anna Ochab-Marcinek and Marcin Tabaka. Transcriptional leakage versus noise: a simple mechanism of conversion between binary and graded response in autoregulated genes. *Physical Review E*, 91(1):012704, 2015.

[178] Jakub Jędrak and Anna Ochab-Marcinek. Influence of gene copy number on self-regulated gene expression. *Journal of Theoretical Biology*, 408:222–236, 2016.

[179] Matthew Scott, Brian Ingalls, and Mads Kærn. Estimations of intrinsic and extrinsic noise in models of nonlinear genetic networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 16(2):026107, 2006.

[180] Tina Toni and Bruce Tidor. Combined model of intrinsic and extrinsic variability for computational network design with application to synthetic biology. *PLoS computational biology*, 9(3):e1002960, 2013.

[181] Emma M Keizer, Björn Bastian, Robert W Smith, Ramon Grima, and Christian Fleck. Extending the linear-noise approximation to biochemical systems influenced by intrinsic noise and slow lognormally distributed extrinsic noise. *Physical Review E*, 99(5):052417, 2019.

[182] Elijah Roberts, Shay Be'er, Chris Bohrer, Rati Sharma, and Michael Assaf. Dynamics of simple gene-network motifs subject to extrinsic fluctuations. *Physical Review E*, 92(6):062717, 2015.

[183] Peter Jung and Peter Hänggi. Dynamical systems: a unified colored-noise approximation. *Physical review A*, 35(10):4464, 1987.

[184] Vahid Shahrezaei, Julien F Ollivier, and Peter S Swain. Colored extrinsic fluctuations and stochastic gene expression. *Molecular systems biology*, 4(1):196, 2008.

[185] Guilherme CP Innocentini, Michael Forger, Alexandre F Ramos, Ovidiu Radulescu, and José EM Hornos. Multimodality and flexibility of stochastic gene expression. *Bulletin of Mathematical Biology*, 75(12):2600–2630, 2013.

[186] Justine Dattani and Mauricio Barahona. Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization. *Journal of The Royal Society Interface*, 14(126):20160833, 2017.

[187] Ulysse Herbach. Stochastic gene expression with a multistate promoter: Breaking down exact distributions. *SIAM Journal on Applied Mathematics*, 79(3):1007–1029, 2019.

[188] Congxin Li, François Cesbron, Michael Oehler, Michael Brunner, and Thomas Höfer. Frequency modulation of transcriptional bursting enables sensitive and rapid gene regulation. *Cell systems*, 6(4):409–423, 2018.

[189] Hannes Risken. Fokker-planck equation. In *The Fokker-Planck Equation*, volume 1, pages 63–95. Springer, 1996.

[190] Cao Li, Wu Da-Jin, and Ke Sheng-Zhi. Bistable kinetic model driven by correlated noises: unified colored-noise approximation. *Physical Review E*, 52(3):3228, 1995.

[191] Ronald F Fox. Functional-calculus approach to stochastic differential equations. *Physical Review A*, 33(1):467, 1986.

[192] Ronald F Fox. Uniform convergence to an effective Fokker-Planck equation for weakly colored noise. *Physical Review A*, 34(5):4525, 1986.

[193] Ronald F Fox. Stochastic calculus in physics. *Journal of Statistical Physics*, 46(5-6):1145–1157, 1987.

[194] Ronald F Fox and Rajarshi Roy. Steady-state analysis of strongly colored multiplicative noise in a dye laser. *Physical Review A*, 35(4):1838, 1987.

[195] Paolo Grigolini, Luigi A Lugiato, Riccardo Mannella, Peter VE McClintock, M Merri, and M Pernigo. Fokker-Planck description of stochastic processes with colored noise. *Physical Review A*, 38(4):1966, 1988.

[196] Eugene Wong and Moshe Zakai. On the convergence of ordinary integrals to stochastic integrals. *The Annals of Mathematical Statistics*, 36(5):1560–1564, 1965.

[197] Giuseppe Pesce, Austin McDaniel, Scott Hottovy, Jan Wehr, and Giovanni Volpe. Stratonovich-to-Itô transition in noisy systems with multiplicative feedback. *Nature Communications*, 4:2733, 2013.

[198] Damien Nicolas, Nick E Phillips, and Felix Naef. What shapes eukaryotic transcriptional bursting? *Molecular BioSystems*, 13(7):1280–1290, 2017.

[199] Jiajun Zhang and Tianshou Zhou. Stationary moments, distribution conjugation and phenotypic regions in stochastic gene transcription. *Mathematical Biosciences and engineering: MBE*, 16(5):6134–6166, 2019.

[200] Sandeep Choubey, Jane Kondev, and Alvaro Sanchez. Deciphering transcriptional dynamics *in vivo* by counting nascent RNA molecules. *PLoS Computational Biology*, 11(11), 2015.

[201] Tatiana Filatova, Nikola Popović, and Ramon Grima. Statistics of nascent and mature RNA fluctuations in a stochastic model of transcriptional initiation, elongation, pausing, and termination. *Bulletin of Mathematical Biology*, 83(1):1–62, 2021.

[202] Tatiana Filatova, Nikola Popović, and Ramon Grima. Modulation of nuclear and cytoplasmic mRNA fluctuations by time-dependent stimuli: analytical distributions. *Mathematical Biosciences*, 347:108828, 2022.

[203] Shasha Chong, Chongyi Chen, Hao Ge, and X Sunney Xie. Mechanism of transcriptional bursting in bacteria. *Cell*, 158(2):314–326, 2014.

[204] Michael C Mackey, Marta Tyran-Kaminska, and Romain Yvinec. Dynamic behavior of stochastic gene expression models in the presence of bursting. *SIAM Journal on Applied Mathematics*, 73(5):1830–1852, 2013.

[205] Geoffrey M Cooper, Robert E Hausman, and Robert E Hausman. *The cell: a molecular approach*, volume 4. ASM press Washington, DC, 2007.

[206] Erik McShane, Celine Sin, Henrik Zauber, Jonathan N Wells, Neysan Donnelly, Xi Wang, Jingyi Hou, Wei Chen, Zuzana Storchova, Joseph A Marsh, Angelo Valleriani, and Matthias Selbach. Kinetic analysis of protein stability reveals age-dependent degradation. *Cell*, 167(3):803–815, 2016.

[207] Juan M Pedraza and Johan Paulsson. Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, 319(5861):339–343, 2008.

[208] Aleksandra M Walczak, Masaki Sasai, and Peter G Wolynes. Self-consistent proteomic field theory of stochastic gene switches. *Biophysical journal*, 88(2):828–850, 2005.

[209] Manuel Pájaro, Antonio A Alonso, Irene Otero-Muras, and Carlos Vázquez. Stochastic modeling and numerical simulation of gene regulatory networks with protein bursting. *Journal of Theoretical Biology*, 421:51–70, 2017.

[210] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309, 2006.

[211] Shawn C Little, Mikhail Tikhonov, and Thomas Gregor. Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell*, 154(4):789–800, 2013.

[212] Tineke L Lenstra, Joseph Rodriguez, Huimin Chen, and Daniel R Larson. Transcription dynamics in living cells. *Annual review of biophysics*, 45:25–47, 2016.

[213] Yihan Wan, Dimitrios G Anastasakis, Joseph Rodriguez, Murali Palangat, Prabhakar Gudla, George Zaki, Mayank Tandon, Gianluca Pegoraro, Carson C Chow, Markus Hafner, and Daniel R Larson. Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell*, 184(11):2878–2895, 2021.

[214] Evelina Tutucci, Nathan M Livingston, Robert H Singer, and Bin Wu. Imaging mRNA *in vivo*, from birth to death. *Annual review of biophysics*, 47:85–106, 2018.

[215] Yinfeng Zhang, Susan J Anderson, Sarah L French, Martha L Sikes, Olga V Viktorovskaya, Jacalyn Huband, Katherine Holcomb, John L Hartman IV, Ann L Beyer, and David A Schneider. The SWI/SNF chromatin remodeling complex influences transcription by RNA polymerase I in *Saccharomyces cerevisiae*. *PloS one*, 8(2):e56793, 2013.

[216] Leighton J Core, Joshua J Waterfall, and John T Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848, 2008.

[217] Heng Xu, Leonardo A Sepúlveda, Lauren Figard, Anna Marie Sokac, and Ido Golding. Combining protein and mRNA quantification to decipher transcriptional regulation. *Nature methods*, 12(8):739–742, 2015.

[218] Caroline R Bartman, Nicole Hamagami, Cheryl A Keller, Belinda Giardine, Ross C Hardison, Gerd A Blobel, and Arjun Raj. Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. *Molecular cell*, 73(3):519–532, 2019.

[219] Leighton Core and Karen Adelman. Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes & development*, 33(15-16):960–982, 2019.

[220] Iris Jonkers, Hojoong Kwak, and John T Lis. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife*, 3:e02407, 2014.

[221] Peter B Rahl, Charles Y Lin, Amy C Seila, Ryan A Flynn, Scott McCuine, Christopher B Burge, Phillip A Sharp, and Richard A Young. c-Myc regulates transcriptional pause release. *Cell*, 141(3):432–445, 2010.

[222] Rajesh Karmakar. Control of noise in gene expression by transcriptional reinitiation. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(6):063402, 2020.

[223] Rajesh Karmakar and Amit Kumar Das. Effect of transcription reinitiation in stochastic gene expression. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(3):033502, 2021.

[224] William J Blake, Mads Kærn, Charles R Cantor, and James J Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

[225] Nicholas J Fuda, M Behfar Ardehali, and John T Lis. Defining mechanisms that regulate RNA polymerase II transcription *in vivo*. *Nature*, 461(7261):186–192, 2009.

[226] Ty C Voss and Gordon L Hager. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics*, 15(2):69–81, 2014.

[227] Ibrahim I Cisse, Ignacio Izeddin, Sebastien Z Causse, Lydia Boudarene, Adrien Senecal, Leila Muresan, Claire Dugast-Darzacq, Bassam Hajj, Maxime Dahan, and Xavier Darzacq. Real-time dynamics of RNA polymerase II clustering in live human cells. *Science*, 341(6146):664–667, 2013.

[228] Won-Ki Cho, Jan-Hendrik Spille, Micca Hecht, Choongman Lee, Charles Li, Valentin Grube, and Ibrahim I Cisse. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, 361(6400):412–415, 2018.

[229] Patrick Cramer. Organization and regulation of gene transcription. *Nature*, 573(7772):45–54, 2019.

[230] Leighton J Core and John T Lis. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*, 319(5871):1791–1792, 2008.

[231] Wanqing Shao and Julia Zeitlinger. Paused RNA polymerase II inhibits new transcriptional initiation. *Nature genetics*, 49(7):1045, 2017.

[232] Arnaud R Krebs, Dilek Imanci, Leslie Hoerner, Dimos Gaidatzis, Lukas Burger, and Dirk Schübeler. Genome-wide single-molecule footprinting reveals high RNA polymerase II turnover at paused promoters. *Molecular cell*, 67(3):411–422, 2017.

[233] Kinga Kamieniarz-Gdula and Nick J Proudfoot. Transcriptional control by premature termination: a forgotten mechanism. *Trends in Genetics*, 35(8):553–564, 2019.

[234] Clare A Beelman and Roy Parker. Degradation of mRNA in eukaryotes. *Cell*, 81(2):179–183, 1995.

[235] Jonathan Houseley and David Tollervey. The many pathways of RNA degradation. *Cell*, 136(4):763–776, 2009.

[236] Veronika A Herzog, Brian Reichholf, Tobias Neumann, Philipp Rescheneder, Pooja Bhat, Thomas R Burkard, Wiebke Wlotzka, Arndt von Haeseler, Johannes Zuber, and Stefan L Ameres. Thiol-linked alkylation of RNA to assess expression dynamics. *Nature methods*, 14(12):1198–1204, 2017.

[237] Tianshou Zhou and Jiajun Zhang. Analytical results for a multistate gene model. *SIAM Journal on Applied Mathematics*, 72(3):789–818, 2012.

[238] Joseph Rodriguez, Gang Ren, Christopher R Day, Keji Zhao, Carson C Chow, and Daniel R Larson. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell*, 176(1-2):213–226, 2019.

[239] Tae H Kim, Leah O Barrera, Ming Zheng, Chunxu Qu, Michael A Singer, Todd A Richmond, Yingnian Wu, Roland D Green, and Bing Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, 2005.

[240] Matthew G Guenther, Stuart S Levine, Laurie A Boyer, Rudolf Jaenisch, and Richard A Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, 2007.

[241] Xiaoming Fu, Heta P Patel, Stefano Coppola, Libin Xu, Zhixing Cao, Tineke L Lenstra, and Ramon Grima. Accurate inference of stochastic gene expression from nascent transcript heterogeneity. *bioRxiv*, 2021.

[242] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.22 of 2019-03-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[243] Fei Chen, Xin Gao, and Ali Shilatifard. Stably paused genes revealed through inhibition of transcription initiation by the TFIIH inhibitor triptolide. *Genes & development*, 29(1):39–47, 2015.

[244] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature structural & molecular biology*, 15(12):1263, 2008.

[245] Douglas F Browning and Stephen JW Busby. Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, 14(10):638, 2016.

[246] Ryohei Sekido, Kasumi Murai, Yusuke Kamachi, and Hisato Kondoh. Two mechanisms in the action of repressor δef1: binding site competition with an activator and active repression. *Genes to Cells*, 2(12):771–783, 1997.

[247] Katjana Tantale, Encar Garcia-Oliver, Marie-Cécile Robert, Adèle L'hostis, Yueyuxiao Yang, Nikolay Tsanov, Rachel Topno, Thierry Gostan, Alja Kozulic-Pirher, Meenakshi Basu-Shrivastava, Kamalika Mukherjee, Vera Slaninova, Jean-Christophe Andrau, Florian Mueller, Eugenia Basyuk, Ovidiu Radulescu, and Edouard Bertrand. Stochastic pausing at latent HIV-1 promoters generates transcriptional bursting. *Nature Communications*, 12(1):1–20, 2021.

[248] Virginia L Pimmett, Matthieu Dejean, Carola Fernandez, Antonio Trullo, Edouard Bertrand, Ovidiu Radulescu, and Mounia Lagha. Quantitative imaging of transcription in living *Drosophila* embryos reveals the impact of core promoter motifs on promoter state dynamics. *Nature communications*, 12(1):1–16, 2021.

[249] Adrien Senecal, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. Transcription factors modulate c-Fos transcriptional bursts. *Cell reports*, 8(1):75–83, 2014.

[250] Saumil J Gandhi, Daniel Zenklusen, Timothée Lionnet, and Robert H Singer. Transcription of functionally related constitutive genes is not coordinated. *Nature structural & molecular biology*, 18(1):27, 2011.

[251] Leonor Michaelis and Maud L Menten. Die kinetik der invertinwirkung. *Biochem Z*, 49:333–369, 1913.

[252] Anthony F Bartholomay. A stochastic approach to statistical kinetics with application to enzyme kinetics. *Biochemistry*, 1(2):223–230, 1962.

[253] Yasushi Ishihama, Thorsten Schmidt, Juri Rappsilber, Matthias Mann, F Ulrich Hartl, Michael J Kerner, and Dmitrij Frishman. Protein abundance profiling of the *Escherichia coli* cytosol. *BMC genomics*, 9(1):102, 2008.

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

[255] Kevin R Sanft, Daniel T Gillespie, and Linda R Petzold. Legitimacy of the stochastic Michaelis-Menten approximation. *IET systems biology*, 5(1):58–69, 2011.

[256] Hong Qian and Lisa M Bishop. The chemical master equation approach to non-equilibrium steady-state of open biochemical systems: Linear single-molecule enzyme kinetics and nonlinear biochemical reaction networks. *International journal of molecular sciences*, 11(9):3472–3500, 2010.

[257] Narmada Herath and Domitilla Del Vecchio. Reduced linear noise approximation for biochemical reaction networks with time-scale separation: The stochastic tQSSA+. *The Journal of Chemical Physics*, 148(9):094108, 2018.

[258] Hye-Won Kang, Wasiur R KhudaBukhsh, Heinz Koeppl, and Grzegorz A Rempała. Quasi-steady-state approximations derived from the stochastic model of enzyme kinetics. *Bulletin of Mathematical Biology*, 81(5):1303–1336, 2019.

[259] Erel Levine and Terence Hwa. Stochastic fluctuations in metabolic pathways. *Proceedings of the National Academy of Sciences*, 104(22):9224–9229, 2007.

[260] Marianne O Stefanini, Alan J McKane, and Timothy J Newman. Single enzyme pathways and substrate fluctuations. *Nonlinearity*, 18(4):1575, 2005.

[261] Brian P English, Wei Min, Antoine M van Oijen, Kang T Lee, Guobin Luo, Hongye Sun, Binny J Cherayil, SC Kou, and X Sunney Xie. Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nature chemical biology*, 2(2):87–94, 2006.

[262] Ramon Grima. Noise-induced breakdown of the Michaelis-Menten equation in steady-state conditions. *Physical Review Letters*, 102(21):218103, 2009.

[263] Ramon Grima. Investigating the robustness of the classical enzyme kinetic equations in small intracellular compartments. *BMC systems biology*, 3(1):101, 2009.

[264] Ramon Grima and André Leier. Exact product formation rates for stochastic enzyme kinetics. *The Journal of Physical Chemistry B*, 121(1):13–23, 2017.

[265] Divya Singh and Srabanti Chaudhury. Single-molecule kinetics of an enzyme in the presence of multiple substrates. *ChemBioChem*, 19(8):842–850, 2018.

[266] Ramon Grima, Nils G Walter, and Santiago Schnell. Single-molecule enzymology à la Michaelis–Menten. *The FEBS journal*, 281(2):518–530, 2014.

[267] George E Briggs and John BS Haldane. A note on the kinetics of enzyme action. *Biochemical Journal*, 19(2):338, 1925.

[268] Lee A Segel and Marshall Slemrod. The quasi-steady-state assumption: a case study in perturbation. *SIAM review*, 31(3):446–477, 1989.

[269] John J Tyson. Biochemical oscillations. In *Computational Cell Biology*, volume 1, pages 230–260. Springer, 2002.

[270] Lee A Segel. On the validity of the steady state assumption of enzyme kinetics. *Bulletin of Mathematical Biology*, 50(6):579–593, 1988.

[271] Chen Jia. Model simplification and loss of irreversibility. *Physical Review E*, 93(5):052149, 2016.

[272] Mona K Tonn, Philipp Thomas, Mauricio Barahona, and Diego A Oyarzún. Computation of single-cell metabolite distributions using mixture models. *Frontiers in Cell and Developmental Biology*, 8:614832, 2020.

[273] Emrah Kılıç and Pantelimon Stanica. The inverse of banded matrices. *Journal of Computational and Applied Mathematics*, 237(1):126–135, 2013.

[274] James W Brown and RV Churchill. *Complex variables and applications*. McGraw-Hill Higher Education, 2009.

[275] G Broggi, Luigi A Lugiato, and A Colombo. Transient bimodality in optically bistable systems. *Physical Review A*, 32(5):2803, 1985.

[276] Marcin Mierzejewski, Jerzy Dajka, Jerzy Łuczka, Peter Talkner, and Peter Hänggi. Dynamical bimodality in equilibrium monostable systems. *Physical Review E*, 74(4):041102, 2006.

[277] Chan F Lam and David G Priest. Enzyme Kinetics: Systematic Generation of Valid King-Altman Patterns. *Biophysical journal*, 12(3):248–256, 1972.

[278] Paul A Sims. An "Aufbau" Approach To Understanding How the King–Altman Method of Deriving Rate Equations for Enzyme-Catalyzed Reactions Works. *Journal of chemical education*, 86(3):385, 2009.

[279] Athel Cornish-Bowden. *Fundamentals of enzyme kinetics*. John Wiley & Sons, 2013.

[280] Charles Mackay. *Extraordinary popular delusions and the madness of crowds*. Harriman House, 2003.

[281] Quentin Michard and Jean-Philippe Bouchaud. Theory of collective opinion shifts: from smooth trends to abrupt swings. *The European Physical Journal B*, 47(1):151–159, 2005.

[282] Jean-Philippe Bouchaud. Crises and collective socio-economic phenomena: Simple models and challenges. *Journal of Statistical Physics*, 151(3-4):567–606, 2013.

[283] Thomas Schelling. *Micromotives and macrobehavior*. Norton, 2006.

[284] Thomas C. Schelling. Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186, 1971.

[285] William A Brock and Steven N Durlauf. Discrete choice with social interactions. *The Review of Economic Studies*, 68(2):235–260, 2001.

[286] Sidney Redner. Reality-inspired voter models: A mini-review. *Comptes Rendus Physique*, 20(4):275–292, 2019.

[287] Ali Hosseiny, Mohammadreza Absalan, Mohammad Sherafati, and Mauro Gallegati. Hysteresis of economic networks in an XY model. *Physica A: Statistical Mechanics and its Applications*, 513:644–652, 2019.

[288] JL Deneubourg, Serge Aron, Simon Goss, and Jacques M Pasteels. The self-organizing exploratory pattern of the argentine ant. *Journal of insect behavior*, 3(2):159–168, 1990.

[289] Frank M Bass. A new product growth for model consumer durables. *Management science*, 15(5):215–227, 1969.

[290] H Peyton Young. Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American economic review*, 99(5):1899–1924, 2009.

[291] Simone Alfarano and Thomas Lux. A minimal noise trader model with realistic time series properties. In *Long Memory in Economics*, volume 1, pages 345–361. Springer Berlin Heidelberg, 2007.

[292] Simone Alfarano, Thomas Lux, and Friedrich Wagner. Estimation of Agent-Based Models: The Case of an Asymmetric Herding Model. *Computational Economics*, 26(1):19–49, 2005.

[293] Jean-Philippe Bouchaud and Roger Farmer. Self-fulfilling prophecies, quasi non-ergodicity and wealth inequality. *arXiv preprint arXiv:2012.09445*, 2021.

[294] Robin Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4, 2007.

[295] Renaud Lambiotte and Sidney Redner. Dynamics of vacillating voters. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(10):L10001, 2007.

[296] Uwe C Täuber. *Critical dynamics: a field theory approach to equilibrium and non-equilibrium scaling behavior*. Cambridge University Press, 2014.

[297] Steven Weinberg. *The quantum theory of fields*, volume 2. Cambridge University Press, 1995.

[298] Michael M Nieto and L. M. Simmons. Coherent states for general potentials. II. Confining one-dimensional examples. *Physical Review D*, 20:1332–1341, 1979.

[299] H. Taşeli. Exact Analytical Solutions of the Hamiltonian with a Squared Tangent Potential. *Journal of Mathematical Chemistry*, 34(3/4):243–251, 2003.

[300] Thomas M Liggett. *Stochastic interacting systems: contact, voter and exclusion processes*, volume 324. Springer Science & Business Media, 1999.

[301] Kristen Fichthorn, Erdogan Gulari, and Robert Ziff. Noise-induced bistability in a monte carlo surface-reaction model. *Physical Review Letters*, 63(14):1527–1530, 1989.

[302] David Considine, Sidney Redner, and Hideki Takayasu. Comment on "Noise-induced bistability in a Monte Carlo surface-reaction model". *Physical Review Letters*, 63(26):2857–2857, 1989.

[303] Pavel L Krapivsky. Kinetics of monomer-monomer surface catalytic reactions. *Physical Review A*, 45(2):1067–1072, 1992.

[304] José Moran, Antoine Fosset, Davide Luzzati, Jean-Philippe Bouchaud, and Michael Benzaquen. By force of habit: Self-trapping in a dynamical utility landscape. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(5):053123, 2020.

[305] Rosemary J Harris. Random walkers with extreme value memory: modelling the peak-end rule. *New Journal of Physics*, 17(5):053049, 2015.

[306] Evangelos Mitsokapas and Rosemary J Harris. Decision-making with distorted memory: Escaping the trap of past experience. *Physica A: Statistical Mechanics and its Applications*, 593:126762, 2022.

[307] Samuel C Wiese and Torsten Heinrich. The frequency of convergent games under best-response dynamics. *Dynamic Games and Applications*, 12(2):689–700, 2022.

[308] Torsten Heinrich, Yoojin Jang, Luca Mungo, Marco Pangallo, Alex Scott, Bassel Tarbush, and Samuel Wiese. Best-response dynamics, playing sequences, and convergence to equilibrium in random games. *arXiv preprint arXiv:2101.04222*, 2021.

[309] Joseph W Baron. Consensus, polarization, and coexistence in a continuous opinion dynamics model with quenched disorder. *Physical Review E*, 104(4):044309, 2021.

[310] Ada Altieri, Felix Roy, Chiara Cammarota, and Giulio Biroli. Properties of equilibria and glassy phases of the random lotka-volterra model with demographic noise. *Physical Review Letters*, 126(25):258301, 2021.

[311] Felix Roy, Matthieu Barbier, Giulio Biroli, and Guy Bunin. Can endogenous fluctuations persist in high-diversity ecosystems? *arXiv preprint arXiv:1908.03348*, 2019.

[312] Zhixing Cao and Ramon Grima. Accuracy of parameter estimation for auto-regulatory transcriptional feedback loops from noisy data. *Journal of The Royal Society Interface*, 16:20180967, 2019.

[313] Christoph Zechner, Jakob Ruess, Peter Krenn, Serge Pelet, Matthias Peter, John Lygeros, and Heinz Koeppl. Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences*, 109(21):8340–8345, 2012.

[314] Jakob Ruess and John Lygeros. Moment-based methods for parameter inference and experiment design for stochastic biochemical reaction networks. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 25(2):1–25, 2015.

[315] Joshua Burton, Cerys S Manning, Magnus Rattray, Nancy Papalopulu, and Jochen Kursawe. Inferring kinetic parameters of oscillatory gene regulation from single cell time-series data. *Journal of the Royal Society Interface*, 18(182):20210393, 2021.

[316] Chen Jia and Ramon Grima. Accuracy and limitations of extrinsic noise models to describe gene expression in growing cells. *bioRxiv*, 2022.

[317] Yu Tanouchi, Anand Pai, Heungwon Park, Shuqiang Huang, Nicolas E Buchler, and Lingchong You. Long-term growth data of *Escherichia coli* at a single-cell level. *Scientific data*, 4(1):1–5, 2017.

[318] Geoffrey B West, Van M Savage, James Gillooly, Brian J Enquist, William H Woodruff, and James H Brown. Why does metabolic rate scale with body size? *Nature*, 421(6924):713–713, 2003.

[319] Geoffrey B West. Scale: the universal laws of life and death in organisms. *Cities and Companies*, 2017.

[320] Hyuntae Lim and YounJoon Jung. Reaction-path statistical mechanics of enzymatic kinetics. *The Journal of Chemical Physics*, 156(13):134108, 2022.

[321] Christian Borghesi and Jean-Philippe Bouchaud. Of songs and men: a model for multiple choice with herding. *Quality & quantity*, 41(4):557–568, 2007.

[322] Giulio Bottazzi and Angelo Secchi. Repeated choices under dynamics externalities. Technical report, LEM Working Paper Series, 2011.

[323] Giulio Bottazzi and Ugo M Gragnolati. Cities and clusters: Economy-wide and sector-specific effects in corporate location. *Regional Studies*, 49(1):113–129, 2015.

[324] Giulio Bottazzi, Ugo M Gragnolati, and Fabio Vanni. Non-linear externalities in firm localization. *Regional Studies*, 51(8):1138–1150, 2017.

[325] Jean-François Mercure. Fashion, fads and the popularity of choices: micro-foundations for diffusion consumer theory. *Structural Change and Economic Dynamics*, 46:194–207, 2018.

[326] Jean-François Mercure. FTT: Power: A global model of the power sector with induced technological change and natural resource depletion. *Energy Policy*, 48:799–811, 2012.

[327] Jean-François Mercure, Pablo Salas, Pim Vercoulen, Gregor Semieniuk, Aileen Lam, Hector Pollitt, Philip B Holden, Negar Vakilifard, Unnada Chewpreecha, Neil R Edwards, and Jorge E Vinuales. Reframing incentives for climate policy action. *Nature Energy*, 6(12):1133–1143, 2021.

[328] Philipp Thomas, Arthur V Straube, and Ramon Grima. Stochastic theory of large-scale enzyme-reaction networks: Finite copy number corrections to rate equation models. *The Journal of Chemical Physics*, 133(19):11B607, 2010.

[329] Eric D Beinhocker. *The origin of wealth: Evolution, complexity, and the radical remaking of economics*. Harvard Business Press, 2006.

[330] T. Loman, Y. Ma, V. Ilin, S. Gowda, N. Korsbo, N. Yewale, C. V. Rackauckas, and S. A. Isaacson. Catalyst: Fast biochemical modeling with julia. *bioRxiv*, 2022.

[331] Augustinas Sukys and Ramon Grima. Momentclosure. jl: automated moment closure approximations in julia. *Bioinformatics*, 38(1):289–290, 2022.

[332] Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: a language for flexible probabilistic inference. In *International conference on artificial intelligence and statistics*, pages 1682–1690. PMLR, 2018.

[333] Stephen Smith and Ramon Grima. Spatial stochastic intracellular kinetics: A review of modelling approaches. *Bulletin of mathematical biology*, 81(8):2960–3009, 2019.

[334] Stephen Smith, Claudia Cianci, and Ramon Grima. Macromolecular crowding directs the motion of small molecules inside cells. *Journal of the Royal Society Interface*, 14(131):20170047, 2017.

[335] Adam J Ellery, Matthew J Simpson, Scott W McCue, and Ruth E Baker. Characterizing transport through a crowded environment with different obstacle sizes. *The Journal of chemical physics*, 140(5):02B601_1, 2014.

[336] Guilherme CP Innocentini, Alexandre F Ramos, and José Eduardo M Hornos. Comment on "Steady-state fluctuations of a genetic feedback loop: An exact solution". *The Journal of Chemical Physics*, 142(2):035104, 2015.

[337] Peter Hänggi and Peter Jung. Colored noise in dynamical systems. *Advances in Chemical Physics*, 89:239–326, 1995.

[338] Robert H Cannon. *Dynamics of physical systems*. Courier Corporation, 2003.

[339] Robert Feldt. Blackboxoptim.jl. `https://github.com/robertfeldt/BlackBoxOptim.jl`, 2018.

[340] Anton Zettl. *Sturm-liouville theory*. Number 121. American Mathematical Soc., 2012.