Department of Medical Physics and Bioengineering
Department of Mathematics
University College London

Stability and robust behaviour across classes of biological and chemical models

Peter Donnell

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
at University College London

April 2009

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis describes three applications of the theory of continuous autonomous dynamical systems. The focus of the thesis is on qualitative, as opposed to numerical, analysis. The applications examined are biological and chemical, and as such there are significant uncertainties in any mathematical representation of them. While the qualitative relationships that define a biological or chemical system may be well understood, it is often difficult to obtain accurate measurements of the parameters that govern each interaction, due to inherent variability and/or experimental constraints. For this reason, a model that avoids dependence on numerical values while still accurately reflecting the qualitative structure of the system it represents is potentially of use in gaining a greater understanding of how the system can behave. Conversely, if a purely qualitative model allows certain behaviour that is never experimentally observed, this may highlight the importance of certain parameter values for the system's real world behaviour.

The first application presented is a model of electron transport in mitochondria, the second is a model of an inter-cellular gap junction, and the third represents a set of reactions occurring in a continuous flow stirred tank reactor. For each application, a reasonable set of qualitative assumptions is found under which there is a unique steady state to which all initial conditions converge, regardless of precise numerical values. Uniqueness of steady states is proved using results on the injectivity of functions, and degree theory. The convergence criteria are constructed using two different areas of dynamical systems theory. The first of these is the theory of monotone flows, while the second is a group of results known as "autonomous convergence theorems". The theory of monotone flows is fairly well known, and relies on finding conditions under which trajectories of a dynamical system preserve a partial ordering, thereby limiting the possibly asymptotic behaviour of the system. The autonomous convergence theorems appear much less well known; they work by finding a norm under which trajectories approach each other, either in phase space or in a related exterior algebra space. Both theories are discussed in detail, along with some extensions.

In memory of Ed

Contents

List of Figures

Preface

The focus of this thesis is on constructing mathematical models based as far as possible only on the structure of a system and qualitative assumptions about the relationship between its components, rather than detailed numerical data, which may not exist or may be inaccurate. Each real world system considered in the thesis is represented by an autonomous dynamical system of the form $\dot{x} = f(x)$. Knowledge about the structure of the real world system is translated into restrictions on $f(x)$, e.g. $f(x)$ might be assumed to be a monotone function of some of the components of x, or one parameter used to construct f may take a larger value than another parameter, even if the actual numerical values of both parameters are unknown.

The thesis has three aims to this end. The first of these is to collect together various results relating to the stability of dynamical systems, some of which appear not to be widely known but are potentially of use in a wide range of applications. Stability is discussed in the context of both local and global asymptotic stability of fixed points, but also in a broader sense that could be termed "robustness": identifying behaviour that persists for a whole class of dynamical systems that have the same structure. The second aim is to briefly elucidate the process by which qualitative mathematical models are constructed, with an emphasis on models of biology. The third aim is to construct models of this kind for a variety of applications, and analyse their behaviour using the relevant theory.

The thesis is structured in two parts. The first three chapters describe mathematical background, beginning in chapter 1 with some definitions and well known results. Section 1.4 of this chapter attempts to explain in more depth the reasons for investigating qualitative models. Chapter 2 then goes on to describe the theory of monotone flows. The potential asymptotic behaviour of monotone dynamical systems is fairly restricted; the consequences of monotonicity, specifically with regard to global asymptotic stability, are discussed in some detail. While the theory of monotone flows is fairly widely known, some new results are derived which will be used in the applications that appear later in the thesis. The final theory chapter, chapter 3, presents a number of results relating to so-called autonomous convergence theorems. As with the theory of monotone flows, autonomous convergence theorems describe conditions which limit the asymptotic behaviour of dynamical systems; unlike monotonicity, autonomous convergence does not appear to be widely recognised in the literature. The relationship between different autonomous convergence theorems is discussed, and some new results are presented.

The second part of the thesis consists of three chapters of applications. For each application, a description of the real world system it represents is given, along with assumptions about the relationships between its constituents. These assumptions are used to construct a dynamical system, which is subsequently analysed.

In chapter 4, a model of electron transport in mitochondria is presented. Mitochondria are present in most animal cells, and play a key role in generating energy needed for all cellular processes. It is shown that, under very broad assumptions, the mitochondrial electron transport chain has a unique equilibrium. For short chains, autonomous convergence is used to derive conditions under which all initial states converge to the equilibrium, while for longer chains it is demonstrated that these conditions no longer suffice to guarantee convergence.

Chapter 5 presents a model of two cells joined by a gap junction, which is a type of junction between cells allowing intercellular communication. The model is somewhat unusual in that it is initially constructed as a three-dimensional dynamical system, but this formulation is closely related to a simpler two-dimensional system which shares much of the behaviour of the three-dimensional system. Conditions are derived under which the model has a unique equilibrium, including a simple graphical test to determine the number of equilibria and their stability. The theory of monotone flows is then used to find conditions under which all initial states converge to the equilibrium in the case where the equilibrium is unique. Some limited extensions to the model are then made, corresponding to increasing the number of cells, and generalising the assumptions made about the properties of the gap junction itself.

Chapter 6, the third chapter of applications, concerns itself with chemical reactions taking place in a fixed volume container with certain restrictions on inflows and outflows. Conditions guaranteeing a unique equilibrium are once again discussed, and then both the theory of monotone flows and autonomous convergence are used to derive conditions under which all initial states converge to the equilibrium. The flow conditions discussed in this thesis are not specifically aimed at biological applications, but there is scope for applying the results to biological systems. There are also structural similarities between chemical reaction models and models of immunology and gene networks, so further extensions might be used to apply the theory to these areas.

The final chapter highlights areas for further work, and attempts to draw attention to possible links between the fundamental areas of theory.

Throughout the thesis, all results that are believed to be new are followed by a proof. All results quoted from other sources are referenced, usually with a reference to a proof.

Chapter 1

Mathematical background

This chapter outlines some of the basic theory of continuous time dynamical systems, upon which the rest of the thesis is based. No new results are developed in this chapter, any results for which references or proofs are not explicitly given are assumed to be well known.

1.1 Preliminaries and definitions

This short section contains a few facts and definitions that are not specific to dynamical systems, but will be referred to later on.

The set of eigenvalues or **spectrum** of a matrix M will be denoted $\sigma(M)$ throughout. Similarly, the maximal real part of the eigenvalues of M, $\max\{\mathrm{Re}(\lambda) : \lambda \in \sigma(M)\}$ (elsewhere referred to as the **stability modulus** or **spectral abscissa** of M) will be denoted $\alpha(M)$.

In some places throughout the thesis, a distinction will need to be made between the set of positive real numbers and nonnegative real numbers. To avoid the ambiguity inherent in the commonly used notation $\mathbb{R}_+$, from this point on $\mathbb{R}_{\geq 0}$ will signify the nonnegative real numbers and $\mathbb{R}_{>0}$ will signify the positive real numbers.

The signum function will be used at various points in the thesis. For a real number r, define the function

$$\mathrm{sgn}(r) \equiv \begin{cases} 1 & (r > 0) \\ 0 & (r = 0) \\ -1 & (r < 0) \end{cases} \tag{1.1}$$

1.2 Continuous time dynamical systems

This thesis concerns itself with dynamical systems describing the evolution of some variable x belonging to a space $X \subseteq \mathbb{R}^n$, referred to as the **phase space** of the system. Only continuous time dynamical systems will be considered, as opposed to a dynamical system in which x evolves in discrete steps according to some iterative map.

Let f be a function $f : X \times \mathbb{R} \to \mathbb{R}^n$. Evolution of x is assumed to be deterministic and described by a differential equation of the form

$$\dot{x} = f(x, t) \tag{1.2}$$

where f is C^1 in both its arguments. While some of the results in this thesis apply to nonautonomous dynamical systems governed by a differential equation of the form stated in 1.2, the focus will be on the simpler case of autonomous dynamical systems, where f is independent of time as follows:

$$\dot{x} = f(x) \tag{1.3}$$

In this case, $f : X \to \mathbb{R}^n$, and f is still assumed to be a C^1 function.

The function f defines a vector field on X. This will sometimes be referred to hereafter as "the vector field [of the dynamical system]". The assumption that f is C^1 guarantees that it has a unique solution over some time period $[-\epsilon, \epsilon]$ ($\epsilon > 0$) at a given point x; existence and uniqueness of solutions of differential equations in the dynamical systems context are discussed in depth in [Hirsch and Smale, 1974] and [Glendinning, 1994]. References such as [Coddington and Levinson, 1955], [Agarwal and Lakshmikantham, 1993] and [Bellman, 1968] go further still into problems of existence and uniqueness in the theory of differential equations.

Solutions to equation (1.3) are mappings of the form $\Phi : [-\epsilon, \epsilon] \times X \to X : (t, x) \mapsto \Phi(t, x)$, sometimes also written as $\Phi_t(x)$ [Hirsch and Smale, 1974]. In general a solution will exist for a time interval defined by some finite $\epsilon > 0$, but the existence of a solution for all time will be assumed throughout this thesis. When Φ satisfies $\Phi(t_2, \Phi(t_1, x)) = \Phi(t_1 + t_2, x)$, such as when a solution exists for all $t \in \mathbb{R}$, Φ is commonly referred to as the **flow** of f. For fixed x, the image of $\Phi(t, x)$ is referred to as the trajectory or orbit of x. Trajectories can be split into **forwards** and **backwards** parts: the forward trajectory or forward semi-orbit of a point x is the image of $\Phi(t, x)$ for $t \geq 0$. Likewise, the backward trajectory or semi-orbit is the image of $\Phi(t, x)$ for $t \leq 0$. The results in this thesis are solely concerned with forward trajectories, so from this point onward the unqualified word "trajectory" or "orbit" will refer to a forward trajectory.

A trajectory $\Phi(t, x)$ is **forwardly bounded** if there exists some $t_0 > 0$ and $r \in \mathbb{R}_{>0}$ such that $|\Phi(t, x)| < r$ for all $t > t_0$. An analogous condition can be used to define backwardly bounded trajectories; however, only forward bounded trajectories are of interest in this thesis, so the unqualified term "bounded" should be taken to mean "forwardly bounded". All forwardly bounded trajectories have an **ω-limit set**, defined as follows: Let (t_n) be a sequence of times satisfying $t_n \to \infty$ as $n \to \infty$. The ω-limit set of a point $x \in X$ is $\omega(x) = \{y \in X \mid \exists\ (t_n)$ such that $\Phi(t_n, x) \to y$ as $n \to \infty\}$. Every point on a trajectory has the same ω-limit set. The ω-limit set of a trajectory may be empty; however any trajectory with compact closure necessarily has a non-empty ω-limit set. All forward trajectories that are bounded in X have compact closure since $X \subseteq \mathbb{R}^n$ and therefore have non-empty ω-limit sets.

Common examples of ω-limit sets are: fixed points (sometimes also referred to as steady states or equilibria[1]), periodic orbits, and sets on which the dynamics are chaotic. In the applications discussed later on, the main focus is on proving that a given dynamical system has a fixed point (or several fixed points), and that all initial conditions converge towards it under the forward flow of the system. For this reason, more background detail will be discussed for fixed points than for periodic orbits or chaotic sets.

If $\Phi(t, x_f) = x_f$ for all $t \in \mathbb{R}$ then x_f is known as a fixed point. Note that fixed points are orbits of the dynamical system. They correspond to zeros of the vector field used to define the dynamical system, i.e. any point x_f at which $f(x_f) = 0$ is a fixed point of the dynamical system.

Clearly $\Phi(0, x) = x$ for all x. When $\Phi(t_p, x_p) = x_p$ for some $t_p \neq 0$ then x_p is referred to as a **t_p-periodic point**. If $\nexists\ T \in (0, t_p)$ such that $x_p = \Phi(T, x_p)$ then t_p is the **prime period** of x_p. The trajectory of a periodic point is a closed loop in phase space, referred to as a periodic orbit. Every point on a periodic orbit has the same trajectory. Fixed points are by definition T-periodic for any period T, and as such are referred to as "trivially periodic". In general, when a periodic orbit is mentioned, it is assumed that the orbit is not trivially periodic. The following two well-known results regarding periodic orbits in planar dynamical systems will be used later:

Theorem 1 (Poincaré–Bendixson). *Suppose that $C \subset \mathbb{R}^2$ is a nonempty compact limit set of a C^1 two-dimensional dynamical system. If C does not contain a fixed point, then it is a periodic orbit.*

This statement of the Poincaré–Bendixson theorem appears, among other places, on p. 248 of [Hirsch and Smale, 1974] and in [Ciesielski, 2001].

Lemma 1 (Dulac criterion). *Let $X \subseteq \mathbb{R}^2$ and $f : X \to \mathbb{R}^2 : x \mapsto f(x)$ for $x \in X$. Suppose that there also exists some C^1 function $g : X \to \mathbb{R}$ such that $\nabla.(gf)$ is not identically zero and does not change sign on a simply connected domain $D \subseteq X$. Then D contains no periodic orbits of the dynamical system defined by f.*

[1] Throughout this thesis, "equilibrium" will always refer to a fixed point, not thermodynamic equilibrium.

See e.g. [Guckenheimer and Holmes, 1983] or p. 130 of [Glendinning, 1994] for more detailed discussion of the Dulac criterion.

Other possible examples of limit sets, such as quasiperiodic sets and sets with chaotic dynamics, are more complicated. Such limit sets will not be discussed in detail, as the aim of the results presented in this thesis is to rule out the existence of periodic orbits and chaotic sets. For the purposes required here, it suffices to say that dynamical systems with bounded chaotic sets and other exotic limit sets are closely related to dynamical systems with periodic orbits, in a way to be made clear later on.

1.3 Fixed point theorems

The results in this thesis are largely concerned with ruling out the existence of chaotic and periodic ω-limit sets, and proving that all points in phase space eventually converge to a fixed point or set of fixed points. The first step in doing this for a given dynamical system is proving that the system contains at least one fixed point. There are a number of results that guarantee the existence of fixed points of dynamical systems, known collectively as **fixed point theorems**. The following fixed point theorem by Brouwer is commonly used:

Theorem 2 (Brouwer)**.** *Let C be a nonempty compact convex set in $\mathbb{R}^n$, and $g : C \to C : x \mapsto g(x)$ be a continuous mapping. Then g has a fixed point $\hat{x}$ satisfying $\hat{x} = g(\hat{x})$.*

Proof. See, for example, p. 63 in [Nikaido, 1968]. □

It is fairly well known that the Brouwer fixed point theorem implies that an autonomous dynamical system of the form outlined in §1.2 defined on a compact convex set has a fixed point; see for example theorem 12 on p. 197 of [Spanier, 1981]. It is also well known that the theorem generalises to any compact simply connected set C^*: in this case there exists a homeomorphism $h : C^* \to C$ where C is convex. If there is a continuous function $\tilde{g} : C^* \to C^*$, then the function $g = h \circ \tilde{g} \circ h^{-1} : C \to C$ has a fixed point $\hat{x}$ by the Brouwer fixed point theorem. $h \circ \tilde{g} \circ h^{-1}(\hat{x}) = \hat{x} \Rightarrow \tilde{g}(h^{-1}(\hat{x})) = h^{-1}(\hat{x})$, so $h^{-1}(\hat{x})$ is a fixed point of $\tilde{g}$.

For the results presented in this thesis the above-mentioned corollary of theorem 2 suffices for proving the existence of fixed points, but for the sake of completeness, the following more general result is included, which may be of interest:

Theorem 3. *Every bounded dynamical system on $\mathbb{R}^n$ has a fixed point.*

Proof. A proof appears in [Richeson and Wiseman, 2002]. Earlier work by other mathematicians that proves the same result is mentioned in [Richeson and Wiseman, 2004]. □

The Brouwer fixed point theorem relies on **compactness** of the set X. Since it is assumed herein that $X \subseteq \mathbb{R}^n$, X is compact if and only if it is closed and bounded. For this reason, in the applications that appear later, compactness of a set will not usually be discussed directly; boundedness will be investigated instead.

1.4 Stability and robustness of dynamical systems

When attempting to construct a biological model, it is frequently the case that some of the processes to be modelled are qualitatively well understood, but their detailed mechanisms are unknown. There are a number of possible causes for this. Modern high-throughput experimental techniques, such as microarrays, identify entities that interact with each other, but not the details of how they interact. When more in depth experiments are performed, which would ideally provide detailed numerical data describing the relationship between entities, there is often difficulty in making measurements of certain processes. The work in this thesis was initially inspired by attempts to model blood flow in the human brain (see [Banaji et al., 2005]). Clearly there are serious practical and ethical issues in making invasive measurements of biological processes on a living human being, limiting the data available to construct a model. While similar mechanisms often exist across individuals or even species, there may be a wide range of variability in quantitative detail.

The results in this thesis attempt to address these sorts of issues in biological modelling by avoiding the choice of functional forms and numerical parameter values as far as possible, and instead constructing generic models based on qualitative assumptions derived from the biology and physics of the systems, such as assuming that some functions are monotone in their arguments, or that some parameters take larger values than others. It is sometimes possible to draw fairly strong conclusions about the long term behaviour of some models constructed in this way, and these conclusions then hold for a whole class of numerical models that share the same underlying structure. Degree theory and injectivity of a vector field can be used to rule out multistability for whole classes of models, while recent results on monotone flows and autonomous convergence can be used to place limits on the long-term behaviour of classes of dynamical systems (e.g. [De Leenheer et al., 2007], [Banaji and Baigent, 2008]).

It is worth pointing out that biologists nearly always have some kind of implicit model of the structure of a biological system in mind when planning experiments, but rarely, if ever, do such models appear to be formally constructed and written down. In this sense, the applications that appear later in this thesis are exercises in formalising this process of identifying the relationship between the important structures in a biological system, and then drawing conclusions about how the system can behave based on a set of qualitative assumptions. If a system is believed, for example, to have a unique equilibrium to which all initial conditions converge, then it is useful to be able to demonstrate this mathematically.

If such behaviour can be proved for a whole class of systems with the same structure, this means that even if experimental data turns out to be inaccurate, the conclusions drawn are still valid (unless, of course, new data reveals a misunderstanding of the fundamental structure of a system). Alternatively, if qualitative mathematical analysis reveals that unexpected behaviour is possible, such as the existence of a periodic orbit, then either finding conditions under which this behaviour might occur in the real world or identifying previously overlooked properties of the system that preclude the unexpected behaviour is potentially of biological interest.

It is in this sense that the idea of "robustness" of a dynamical system is presented: if the same behaviour persists across a class of models, then this behaviour could be described as "robust." This notion of robustness is related to the notion **structural stability**, i.e. identifying when the topology of the trajectories of a system is unaltered under small perturbations (see chapter 16 of [Hirsch and Smale, 1974] for a brief introduction), but it is somewhat broader. For example, for some classes of systems analysed later in this thesis it is demonstrated that all initial conditions converge to a single fixed point. For other classes of systems, it is only possible to show that all solutions remain bounded and there are no stable (in the standard dynamical systems sense described below) periodic orbits. While this second characterisation of possible behaviour is less restrictive than the first, the fact that it is valid across a whole class of systems means that it is potentially still of significant real-world interest.

In all of the applications in this thesis, stability in the standard dynamical systems sense is also considered. Biologically speaking, ω-limit sets that are not **asymptotically stable** (defined below) are not of any practical significance. Any real biological system will be subject to background noise, so even if the initial state of such a system was, say, an unstable equilibrium, the system would drift away from it.

1.4.1 Definitions of stability and attractivity

A fixed point $x \in X$ is called **Lyapunov stable** if, for any $\epsilon > 0$, there exists some $\delta > 0$ such that, for any $y \in X$ satisfying $|y - x| < \delta$, $|\Phi(t, y) - x| < \epsilon$ for all $t \geq 0$. Stronger than this is asymptotic stability: x is (locally) asymptotically stable if it is Lyapunov stable and there exists $\delta > 0$ such that all points y satisfying $|y - x| < \delta$ approach x according to $\lim_{t\to\infty} |\Phi(t, y) - x| = 0$. Lyapunov stability will not be discussed any further, so from here on the word "stable" should be taken to mean asymptotically stable. These definitions can be slightly reformulated to include stability of periodic orbits and chaotic sets.

A different but related notion is that of **attractivity**. A fixed point x is locally attractive if $|\Phi(t, y) - x| \to 0$ as $t \to \infty$, for all y in a neighbourhood U of x. If $|\Phi(t, y) - x| \to 0$ as $t \to \infty$ for all $y \in X$ then x is globally attractive. In rare cases it is possible for a fixed point to be attractive but not asymptotically stable, for example if the fixed point is a saddle node on a homoclinic orbit — see figure 1.1.

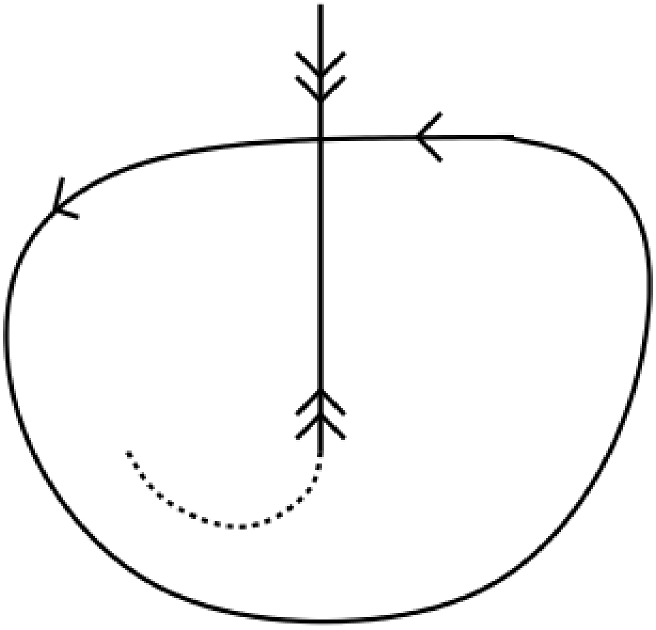

Figure 1.1: A diagram showing a two-dimensional section of a dynamical system containing a saddle node on a homoclinic orbit. The saddle node is a fixed point lying at the intersection of the trajectories marked. If all other trajectories approach the fixed point, it is globally attractive, but it is not asymptotically stable since one end of the homoclinic orbit travels away from it.

1.4.2 The Jacobian matrix

Typically, characterising the behaviour of a dynamical system relies on examining its **Jacobian**. Suppose a vector valued function $f : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto f(x)$ can be written componentwise as $f(x) = (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n))^T$ where x_i is the ith component of the vector x. Then the Jacobian of f is a matrix $Df(x)$ defined as follows:

$$Df(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix} \tag{1.4}$$

The Jacobian is a **local** approximation of the vector field. However, there are a number of results that link **global** properties of a system with conditions on its Jacobian. The remainder of this chapter is concerned with results connecting the Jacobian with properties of the dynamical system.

1.4.3 P matrices and related classes

The first result relates mainly to injectivity of the function f. This is particularly of interest with regard to fixed points; if the vector field of a dynamical system is injective then there can be at most one fixed point — though there may be none.

For some $n \times m$ matrix A, $A(\alpha|\gamma)$ will refer to the submatrix of A with rows indexed by the set $\alpha \subset \{1, \ldots, n\}$ and columns indexed by the set $\gamma \subset \{1, \ldots, m\}$. A **principal submatrix** of A is a submatrix containing columns and rows from the same index set, i.e. of the form $A(\alpha|\alpha)$. A **minor** is the determinant of any square submatrix of A. If $A(\alpha|\gamma)$ is a square submatrix of A (i.e. $|\alpha| = |\gamma|$), then $A[\alpha|\gamma]$ will refer to the corresponding minor, i.e. $A[\alpha|\gamma] = \det(A(\alpha|\gamma))$. A **principal minor** of A is the determinant of a principal submatrix of A, i.e. $A[\alpha|\alpha]$.

P matrices are real square matrices all of whose principal minors are positive. They are by definition nonsingular, since any square matrix is a principal submatrix of itself. If $-A$ is a P matrix, then define A to be a $P^{(-)}$ matrix. If A is a $P^{(-)}$ matrix, this means that each $k \times k$ principal minor of A has sign $(-1)^k$. Related to P matrices are P_0 matrices, the closure of the set of P matrices. The principal minors of a P_0 matrix are all nonnegative.

Functions for which the Jacobian is a nonsingular P_0 matrix are injective on rectangles according to the following theorem:

Theorem 4 (Gale and Nikaido)**.** *Let $f : C \to \mathbb{R}^n$ be a differentiable mapping where C is an open rectangular region of $\mathbb{R}^n$. If the Jacobian $Df(x)$ is a nonsingular P_0 matrix for all $x \in C$ then f is injective in C.*

Proof. See theorem 4w in [Gale and Nikaido, 1965]. □

As per theorem 4 of the same reference, if $Df(x)$ is a P matrix then C is a **closed** rectangle. Note that this result on injectivity also trivially applies to $P^{(-)}$ and $P_0^{(-)}$ matrices.

The eigenvalues of a P matrix are bounded away from a wedge in the complex plane centred on the negative real line, as shown in [Kellogg, 1972] and further discussed in [Fang, 1989].

Theorem 5 (Kellogg)**.** *If $\lambda = re^{i\theta}$ is an eigenvalue of an $m \times m$ P matrix, then*

$$|\theta - \pi| > \pi/m$$

Proof. See [Kellogg, 1972]. □

Since $\sigma(-M) = -\sigma(M)$, the eigenvalues of a $P^{(-)}$ matrix are bounded away from the positive real line according to

$$|\theta| > \pi/m$$

The inequality in theorem 5 becomes $|\theta - \pi| \geq \pi/m$ for the eigenvalues of P_0 matrices and $|\theta| \geq \pi/m$ for $P_0^{(-)}$ matrices.

1.4.4 Hurwitz stability of matrices

A square matrix A is defined to be Hurwitz stable if all its eigenvalues lie in the open left half of the complex plane — the real parts of all its eigenvalues are negative, i.e. $\alpha(A) < 0$. A fixed point of the vector field of a continuous time dynamical system is **locally asymptotically stable** if the Jacobian of the system is **Hurwitz stable** (sometimes abbreviated to **Hurwitz**) at the fixed point. The converse is not true; a fixed point at which the Jacobian has a zero eigenvalue can be asymptotically stable or unstable. See a textbook such as [Hirsch and Smale, 1974], [Glendinning, 1994] or [Wiggins, 1990] for more in-depth discussion of stability.

The Routh-Hurwitz theorem can be used to check whether a given matrix is Hurwitz stable. There are a number of equivalent statements of the Routh-Hurwitz theorem; the following statement appears on page 1076 of [Gradshteyn and Ryzhik, 2000].

Theorem 6 (Routh-Hurwitz). *Consider the characteristic polynomial of a matrix A:*

$$|\lambda I - A| = \lambda^n + b_1\lambda^{n-1} + \ldots + b_{n-1}\lambda + b_n \tag{1.5}$$

In this equation, I is the $n \times n$ identity matrix, and the coefficients b_i are the sums of all principal minors of $-A$ of dimension i. Now define $b_k \equiv 0$ for all $k > n$, and construct a set of numbers Δ_i as follows:

$$\Delta_i = \begin{vmatrix} b_1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ b_3 & b_2 & b_1 & 1 & 0 & 0 & \cdots & 0 \\ b_5 & b_4 & b_3 & b_2 & b_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{2i-1} & b_{2i-2} & b_{2i-3} & b_{2i-4} & b_{2i-5} & b_{2i-6} & \cdots & b_i \end{vmatrix} \tag{1.6}$$

A is Hurwitz if and only if $\Delta_i > 0$ for all $i \leq n$.

The Routh-Hurwitz theorem gives a set of necessary and sufficient conditions for a matrix to be Hurwitz stable, but these conditions are often difficult to check in practice, particularly when the matrix is only described algebraically without any numerical values. An equivalent set of conditions is later mentioned in theorem 22 (§3.2, p. 51). Additionally, there is a variety of alternative conditions implying Hurwitz stability that are stronger than necessary but often easier to check. One such example, involving a generalisation of **diagonal dominance** of a matrix, is as follows:

Theorem 7. *Let M be an $n \times n$ real matrix with M_{ij} being the element on the ith row and jth column. Suppose that the following conditions hold:*

1. $M_{ii} < 0 \ \forall \ i \in \{1, \dots, n\}$.
2. $\exists \ d_i > 0$ *for* $i = 1, \dots, n$ *such that* $d_i|M_{ii}| > \sum_{j \neq i} d_j |M_{ji}|$.

Then M is Hurwitz stable.

Proof. See theorem 2 of [McKenzie, 1960]. □

Remark. Since $\sigma(M) = \sigma(M^T)$, the diagonal dominance in columns required for theorem 7 can alternatively be replaced by diagonal dominance in rows.

Another useful result that can be used to check whether a matrix is Hurwitz is Lyapunov's second theorem, which uses **positive definite** matrices: a real symmetric matrix is positive (negative) definite if and only if its eigenvalues are positive (negative).

Theorem 8. *The eigenvalues of real square matrix A all have positive real part if and only if there exists a positive definite matrix G such that $GA + A^T G = H$, where H is positive definite.*

Proof. See theorem 2.2.1 on p. 96 of [Horn and Johnson, 1991]. □

Corollary 1 (Lyapunov's second theorem)**.** *An real square matrix A satisfies $\alpha(A) < 0$ if and only if there exists a positive definite matrix G such that $GA + A^T G = H$, where H is negative definite.*

Proof. Let $A' = -A$ and $H' = -H$. By theorem 8, the eigenvalues of A' have positive real part if and only if there exists positive definite G such that $GA' + A'^T G = H'$ and H' is positive definite. The result then follows directly since $\sigma(A) = -\sigma(A')$ and $\sigma(H) = -\sigma(H')$. □

P matrices, introduced in the previous section, also have some links to stability. For a $P^{(-)}$ matrix, $b_i > 0$ for all i in the Routh-Hurwitz conditions. A two-dimensional $P^{(-)}$ matrix is Hurwitz stable as both its eigenvalues lie in the left half plane. However for $m > 2$, $P^{(-)}$ matrices may be unstable. There exist further conditions that guarantee Hurwitz stability of a $P^{(-)}$ matrix: For example, if a $P^{(-)}$ matrix has real eigenvalues, then these eigenvalues must necessarily be negative.

1.4.5 Degree theory

Another piece of theory that will prove useful later is the **Brouwer degree** of a map. Let X be a compact manifold and Y be a connected manifold, with $\dim(X) = \dim(Y)$ and finite. Define the C^1 map $f : X \to Y : x \mapsto f(x)$ with Jacobian $Df(x)$, and assume that $f(x) \neq 0 \; \forall \; x \in \partial X$. Then the degree of f over $\text{int}(X)$ with respect to the regular value[2] $y \in Y$ is

$$\deg(f, \text{int}(X), y) = \sum_{\substack{x \in f^{-1}(y), \\ x \in \text{int}(X)}} \text{sgn}(|Df(x)|) \tag{1.7}$$

The discussion of degree theory given here is very brief; see a book on degree theory such as [Berger and Gostiaux, 1988] for a more complete discussion. One important property of the degree of a map is that it is **homotopy invariant**: any two maps that can be continuously deformed into each other have the same degree. This fact leads to the following result:

Theorem 9. *Suppose that $X \subset \mathbb{R}^n$ is a compact, convex set and let $f : X \to \mathbb{R}^n : x \mapsto f(x)$ be any C^1 vector field such that $f(x)$ points into the interior of X for all $x \in \partial X$. Then* $\deg(f, \text{int}(X), 0) = (-1)^n$.

Proof. The first part of the proof of lemma 2 in [De Leenheer et al., 2007] proves essentially the same result, so it is not repeated here. □

1.4.6 Determinants of block matrices

The final result for this chapter is a piece of general matrix theory: It is sometimes helpful to be able to calculate the determinant of a matrix by considering the determinants of block submatrices from which it is constructed. The following lemma will be used later:

Lemma 2. *Let A be any square matrix written in block form as follows:*

$$A = \left(\begin{array}{c:c} A_{11} & A_{12} \\ \hdashline A_{21} & A_{22} \end{array}\right)$$

Here A_{11} and A_{22} are square matrices. Assuming A_{11} is nonsingular, then:

$$|A| = |A_{22} - A_{21}A_{11}^{-1}A_{12}|\,|A_{11}|$$

Proof. See p. 46 of [Lancaster and Tismenetsky, 1985]. □

[2] $f(x) = y$ is a regular value of f if $|Df(x)| \neq 0$.

Chapter 2

Monotone convergence criteria

The purpose of this chapter, and the following chapter, is to outline certain sets of conditions under which trajectories of a dynamical system converge to fixed points. In this chapter **monotonicity** of a dynamical system will be discussed. A dynamical system on an ordered metric space is monotone if states that are initially ordered remain ordered as time increases. In the discussion that follows the Euclidean metric will always be assumed.

The description of monotonicity given here is largely based on that of [Smith, 1995], but there is a great deal of other literature on the subject; relevant sources include the work of Morris Hirsch and Hal Smith in [Hirsch and Smith, 2004], [Hirsch and Smith, 2005], [Hirsch and Smith, 2006]; also work by [Gilbert, 1956], [Ji-fa, 1994], [Mierczynski, 1995], [Rump, 1997], [Radjavi, 1999], [Kunze and Siegel, 1999] and [Kunze and Siegel, 2001]. Of the results presented in this chapter, most appear in one or more of the preceding references; the exceptions are lemma 3, corollary 2, lemma 4, lemma 5 and lemma 7, all of which are minor results developed for this thesis.

It is also worth mentioning in passing a related group of results known as "small–gain theorems," which work by breaking a dynamical system down into subsystems that are internally monotone, with monotone connections between subsystems. Small–gain theorems are not discussed in this thesis, but the interested reader is referred to [Angeli et al., 2004], [Angeli and Sontag, 2003], [Angeli and Sontag, 2004a], [Angeli and Sontag, 2004b], and [Enciso et al., 2006].

The discussion of monotonicity begins with a brief section on orderings in the context of monotonicity, followed by an outline of some possible conditions that a dynamical system must fulfil for its trajectories to preserve a suitable ordering.

2.1 Partial orders

Let X be an ordered metric space. The ordering is defined by a partial order relation $\leq$, which is reflexive, transitive and antisymmetric. These properties translate to the following conditions:

1. $x \leq x$ for any $x \in X$ (reflexivity),
2. $x \leq y$ and $y \leq z \Rightarrow x \leq z$ for $x, y, z \in X$ (transitivity),
3. $x \leq y$ and $y \leq x \Rightarrow x = y$ for $x, y \in X$ (antisymmetry).

Additionally, the partial order relation is assumed to be **closed** in X, i.e. whenever $x_n \leq y_n \, \forall n$ in some set $\{n\}$, it is necessary that if $x_n \to x$ and $y_n \to y$ as $n \to \infty$ then $x \leq y$. When this condition holds, the partial order and topology are called "compatible".

Perhaps the commonest example of such an order relation is the positive orthant ordering in $\mathbb{R}^n$. If $X = \mathbb{R}^n$ and x, y are vectors in X, then there exists a partial order relation $\leq$ on X satisfying $x \leq y \Leftrightarrow y_i - x_i \geq 0 \; \forall i \in \{1, \ldots, n\}$.

One way of defining partial orders is via **cones**. As explained in the next section, a cone is a geometric object, and can be used to generate an order on a pair of vectors by taking the difference between them; if the difference vector lies within the cone then the two vectors are ordered with respect to the cone. Other methods of defining partial orders are not discussed in this thesis.

2.1.1 Partial orders defined by cones

The definition of a cone given here is mainly based upon [Berman and Plemmons, 1994]. Let $K \subseteq \mathbb{R}^n$.[1] K is a cone if $\lambda x \in K$ for any real $\lambda > 0$ and all $x \in K$. A cone K is called **pointed** if $K \cap -K = \{0\}$. It is referred to as **solid** if it has non-empty interior. Following convention, any cone which is pointed, solid, closed and convex will be referred to as a **proper** cone. From this point onwards, unless explicitly stated otherwise, all cones are assumed to be proper. The nonnegative orthant in $\mathbb{R}^n$ is an example of a proper cone.

K defines a partial order relation on $\mathbb{R}^n$. Given a point $x \in \mathbb{R}^n$, imagine a copy of K with the apex at x. Then the set of all vectors that lie within the copied cone are ordered by K with respect to x; the copy of K is the set of all y such that $y \geq x$. This is demonstrated in figure 2.1. Similarly, a copy of $-K$ with the apex at x describes the set of all y such that $y \leq x$ (for simplicity, this is not shown in the diagram).

[1] A cone can be defined more generally on a real or complex vector space, but for simplicity the discussion here is restricted to $\mathbb{R}^n$.

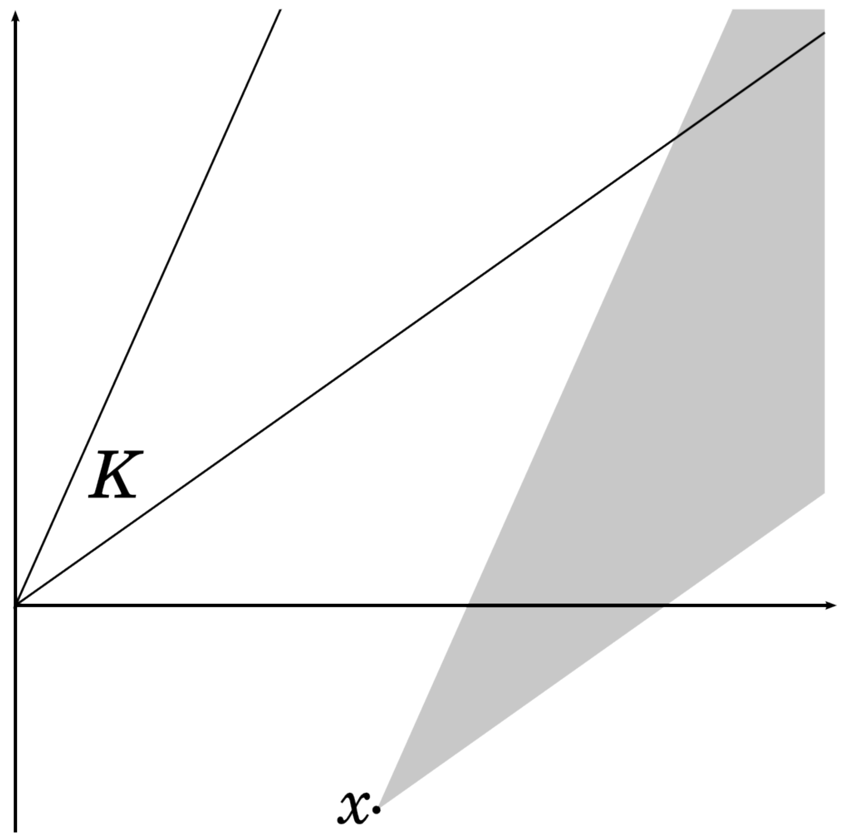

Figure 2.1: A diagram of a partial ordering defined by a cone in two dimensions. A proper cone, marked as K in the diagram, defines a partial ordering on $\mathbb{R}^2$. Given the point x, the shaded area is the set of all points y such that $y - x \in K$, denoted $y \underset{K}{\geq} x$.

The following notation is used for any pair of vectors $x, y \in \mathbb{R}^n$ from this point on:

1. $x \underset{K}{\leq} y$ if $y - x \in K$
2. $x \underset{K}{<} y$ if $x \underset{K}{\leq} y$ and $y - x \neq 0$
3. $x \underset{K}{\ll} y$ if $y - x \in \operatorname{int} K$

In cases where there is no ambiguity about which partial order is being used, the K subscript on the order relations will sometimes be omitted. A number of useful definitions related to cones follow.

For a cone K, its **dual cone** or **dual** is defined as $K^* = \{y \in \mathbb{R}^n : y^T x \geq 0 \; \forall \; x \in K\}$[2]. A vector $g \in K$ is called an **extremal** of K if $0 \leq_K x \leq_K g \Rightarrow x = \lambda g$ for some $\lambda \in \mathbb{R}_{\geq 0}$. Every point $x \in K$ can be written as a nonnegative sum of extremals, i.e. $y = \sum_i \alpha_i g_i$ with $\alpha_i \in \mathbb{R}_{\geq 0}$ and $\{g_i\}$ being the set of all extremals of K. For this reason, the set of extremals of K are also referred to as **generators** of K. Clearly in order for K to be solid it is necessary (but not sufficient) that $|\{g_i\}| \geq n$. If $|\{g_i\}|$ is finite then K is a **polyhedral** cone; if $|\{g_i\}| = n$ and the generators $\{g_i\}$ are linearly independent then K is called a **simplicial** cone (note that a simplicial cone is therefore a special case of a polyhedral cone). The only other possibility is that K is **infinitely generated**, such as an elliptic cone (also known as an "ice cream cone").

A **face** F of a cone K is a subset of K which is a cone such that $x \in F, y \in K, x - y \in K \Rightarrow y \in F$. Although F is itself a cone, unless $F = K$ its interior is empty so it is not proper. A **trivial face** of K is either $\{0\}$ or K itself.

2.2 Matrices that preserve a cone

The following definitions, all of which except the last appear in [Schneider and Tam, 2006], relate properties of vectors and matrices to a cone K:

1. A vector v is K-nonnegative if $v \in K$.
2. A vector v is K-semipositive if $v \in K$ and $v \neq 0$.
3. A vector v is K-positive if $v \in \text{int}\, K$.
4. A matrix M is K-nonnegative if $Mv \in K$ for any $v \in K$.
5. A matrix M is K-semipositive if $Mv \in K$ for any $v \in K$ and $M \neq 0$.
6. A matrix M is K-positive if $Mv \in \text{int}\, K$ for any $v \in K \setminus \{0\}$.
7. A matrix M is K-quasipositive if $\exists\, \lambda \in \mathbb{R}$ such that $M + \lambda I$ is K-nonnegative.

It is worth remembering that K-positivity $\Rightarrow$ K-semipositivity $\Rightarrow$ K-nonnegativity $\Rightarrow$ K-quasipositivity. In the case where $K = \mathbb{R}^n_{\geq 0}$, K-quasipositivity of a matrix M implies that the offdiagonal elements of M are nonnegative, which coincides with the standard definition of the Jacobian of a cooperative system (see §2.2.3 below). In order to simplify the notation later on, whenever $K = \mathbb{R}^n_{\geq 0}$ the K- prefix will be dropped, to coincide with

[2] More generally, the dual cone consists of the set of all linear forms y such that $y(x) \geq 0$. These linear forms can be associated with points in $\mathbb{R}^n$ via a scalar product, so for the purposes of this discussion the Euclidean scalar product suffices to define K^*.

the usual definition that a nonnegative matrix has all entries nonnegative and preserves the nonnegative orthant, and so on for semipositivity, positivity and quasipositivity.

Suppose, given a matrix M and a cone K, that for all $y \in K^*$ and $x \in K$ satisfying $y^T x = 0$, the relation $y^T Mx \geq 0$ holds. This will subsequently be referred to as **condition A**. Condition A is called "cross-positivity" of M in [Schneider and Vidyasagar, 1970]. It is demonstrated below that if M is a K-quasipositive matrix then it fulfils condition A. The converse does not hold, however: there exist some matrices that are not K-quasipositive but still satisfy condition A. The following theorem will be used:

Theorem 10 (Schneider and Vidyasagar)**.** *Given a cone K, a matrix M fulfils condition A if and only if* $\exp(tM)$ *is K-nonnegative for all $t \geq 0$.*

Proof. See theorem 3 in [Schneider and Vidyasagar, 1970]. □

The following simple result establishes that the exponential of a K-quasipositive matrix is K-nonnegative.

Lemma 3. *If M is a K-quasipositive matrix, then e^M is a K-nonnegative matrix.*

Proof. Start with M. To this, add a suitably large multiple of the identity matrix, $\lambda\mathbf{I}$ ($\lambda \in \mathbb{R}$), such that $M + \lambda\mathbf{I}$ is a K-nonnegative matrix. Since this new matrix $\lambda\mathbf{I} + M$ is K-nonnegative, $\exp(\lambda\mathbf{I} + M)$ is also K-nonnegative.

Basic matrix algebra gives:

$$\begin{aligned} e^{\lambda\mathbf{I}+M} &= e^{\lambda\mathbf{I}}e^M \\ &= e^{\lambda}\mathbf{I}e^M \\ &= e^{\lambda}e^M \end{aligned}$$

Since $e^{\lambda} > 0$, it follows that e^M is K-nonnegative. □

From the previous two results, it follows that

Corollary 2. *Any K-quasipositive matrix M fulfils condition A.*

Proof. For all $t \geq 0$, tM is trivially K-quasipositive and so $\exp(tM)$ is K-nonnegative by lemma 3. Therefore M fulfils condition A by theorem 10. □

Finally for this section, the following facts about **irreducibility** of matrices will be useful later (see e.g. [Schneider and Tam, 2006]):

1. If M is a K-nonnegative matrix, then a face F of K is called a M-invariant face if $MF \subseteq F$, i.e. vectors in F remain in F under the mapping described by M.
2. M is K-irreducible if the only M-invariant faces of K are the trivial faces.

Note that, in particular, all K-positive matrices are K-irreducible.

2.2.1 The Perron–Frobenius theorem and related results

The Perron–Frobenius theorem is a result relating nonnegativity of a matrix to its eigenvalues and eigenvectors. While it does not find any direct applications in this thesis, its generalisations are important in the context of identifying matrices that preserve cones, which is relevant to the theory of monotone flows. As such, the theorem is included for completeness.

Theorem 11 (Perron–Frobenius)**.** *Let A be an $n \times n$ nonnegative matrix with eigenvalues λ_i and corresponding eigenvectors v_i. Define the spectral radius $\rho(A) \equiv \max_i |\lambda_i|$. Then $\lambda_m = \rho(A)$ is an eigenvalue of A and there is a corresponding eigenvector $v_m > 0$. If, in addition, A is irreducible then $\lambda_m > 0$, $v_m \gg 0$, λ_m has algebraic multiplicity one and for any eigenvector v_i of A satisfying $v_i > 0$ there exists a scalar $s > 0$ such that $v_i = sv_m$.*

If B is a matrix satisfying $B > A$ (i.e. $B_{ij} \geq A_{ij}\ \forall\ i, j \in \{1, \dots, n\}, B \neq A$), then $\rho(B) > \rho(A)$. Finally, if A is a positive matrix then $|\lambda_i| < \lambda_m$ for all $i \neq m$.

Proof. See [Graham, 1987] or [Godsil and Royle, 2001]. □

The converse to the Perron–Frobenius theorem is false in general; not every real matrix that has a dominant positive real eigenvalue and corresponding positive eigenvector with all other eigenvectors lying outside the nonnegative orthant is a positive matrix, although such a matrix does preserve some cone, as stated below. It may however be possible to find a set of conditions on the eigenvalues and eigenvectors that guarantee nonnegativity or positivity in order to construct a partial converse to the theorem.

A similar theorem on matrices preserving some general cone, not necessarily the nonnegative orthant, is as follows:

Theorem 12. *If $\rho(A)$ is an eigenvalue of A, and if $\deg \rho(A) \geq \deg \lambda$ for every eigenvalue λ such that $|\lambda| = \rho(A)$, then A leaves a proper cone invariant.*

Proof. See theorem 3.5 in chapter 1 of [Berman and Plemmons, 1994]. □

In this theorem $\deg \lambda$ is the multiplicity of λ in the minimal polynomial of A (i.e. the algebraic multiplicity of λ), or equivalently the size of the Jordan block to which λ belongs. Following on from this, there are direct extensions to the Perron–Frobenius theorem.

Theorem 13. *Let P be a K-nonnegative matrix with spectral radius ρ. Then ρ is an eigenvalue of P, and there is a K-semipositive eigenvector of P corresponding to ρ.*

Theorem 14. *Let P be a K-nonnegative and K-irreducible matrix with spectral radius ρ. The following statements hold:*

1. *ρ is positive and is a simple eigenvalue of P (i.e. its algebraic multiplicity is 1).*
2. *There exists a unique (up to a scalar multiple) K-positive right eigenvector of P corresponding to ρ.*
3. *This K-positive eigenvector is the only K-semipositive eigenvector of P.*
4. *$K \cap (\rho I - P)\mathbb{R}^n = \{0\}$.*

Proof. Statements of the above theorems appear in [Schneider and Tam, 2006]; the underlying results can be found in [Krein and Rutman, 1962] and [Barker and Schneider, 1975]. □

Theorem 12 states that a matrix with a dominant positive real eigenvalue preserves a cone. Theorems 13 and 14 give information about where some or all of the eigenvectors lie in relation to the cone that is preserved. However, the theorems say nothing about the extent of the cone itself. For example, if a matrix M has a dominant positive eigenvalue and a corresponding positive eigenvector, then it preserves a cone K such that $K \cap \mathbb{R}^n_{\geq 0} \neq \{0\}$, but there is no guarantee that $\mathbb{R}^n_{\geq 0} \subseteq K$, or even if K does include the positive orthant, whether there are vectors that are mapped out of the positive orthant by M while still remaining inside K. As the next section shows, it is fairly straightforward to explicitly describe all cones that are preserved by a 2×2 real matrix with real positive eigenvalues.

2.2.2 Identifying cones preserved by a matrix

In §2.2.3, the applications of cone-preservation to dynamical systems will be discussed. However, this will raise an important question: given a set of matrices, what cones (if any) do all of the matrices preserve? This is a very difficult question to answer in full generality; a logical starting point is, given an individual matrix, to find all the cones preserved by this single matrix. For a 2D matrix with positive eigenvalues this problem is tractable; while this particular problem does not find an application in the later results in this thesis, it is analysed here as an academic exercise.

Let M be a 2×2 matrix with unit eigenvectors α and β and eigenvalues λ_1 and λ_2 ($\lambda_1 \neq \lambda_2$) as follows:

$$\left.\begin{array}{r} M\alpha = \lambda_1\alpha \\ M\beta = \lambda_2\beta \end{array}\right\} M \in \mathbb{R}^2 \times \mathbb{R}^2; \alpha, \beta \in \mathbb{R}^2; \lambda_1, \lambda_2 \in \mathbb{R}_{>0}$$

Given a vector v, since α and β are linearly independent v can be decomposed into $v = v_1\alpha + v_2\beta$. Moreover, since α and β are only defined up to a change in sign, it can be assumed that v_1 and v_2 are positive (ignoring the trivial cases where one or both are zero). Suppose, without loss of generality, that $\lambda_1 > \lambda_2 > 0$.

Lemma 4. *The angle between v and α decreases under the mapping described by M.*

Proof. Let θ be the angle between α and v, and let ϕ be the angle between α and Mv. The following relations hold:

$$\begin{aligned} \frac{\lambda_1 v_1}{\lambda_2 v_2} &> \frac{v_1}{v_2} && (2.1) \\ \langle v, \alpha\rangle &= |v||\alpha|\cos\theta && (2.2) \\ \langle Mv, \alpha\rangle &= |Mv||\alpha|\cos\phi && (2.3) \end{aligned}$$

Since $|\alpha| = |\beta| = 1$, the following can be shown:

$$\begin{aligned} |v|^2 &= v_1^2 + 2v_1v_2\langle\alpha,\beta\rangle + v_2^2 && (2.4) \\ |Mv|^2 &= \lambda_1^2 v_1^2 + 2\lambda_1\lambda_2 v_1 v_2\langle\alpha,\beta\rangle + \lambda_2^2 v_2^2 && (2.5) \\ \langle v,\alpha\rangle &= v_1 + v_2\langle\alpha,\beta\rangle && (2.6) \\ \langle Mv,\alpha\rangle &= \lambda_1 v_1 + \lambda_2 v_2\langle\alpha,\beta\rangle && (2.7) \end{aligned}$$

Then there are four cases to consider:

1. $\langle v,\alpha\rangle \geq 0$ and $\langle Mv,\alpha\rangle \geq 0$
2. $\langle v,\alpha\rangle \geq 0$ and $\langle Mv,\alpha\rangle < 0$
3. $\langle v,\alpha\rangle < 0$ and $\langle Mv,\alpha\rangle \geq 0$
4. $\langle v,\alpha\rangle < 0$ and $\langle Mv,\alpha\rangle < 0$

All the cases except the first imply that $\langle\alpha,\beta\rangle < 0$, due to the assumption that v_1 and v_2 are positive.

Case 1; $\langle v, \alpha\rangle \geq 0$ and $\langle Mv, \alpha\rangle \geq 0$: By equations (2.2) and (2.3), the claim that the angle between v and α decreases under M is equivalent to the following:

$$\frac{\langle Mv, \alpha\rangle}{|Mv|} > \frac{\langle v, \alpha\rangle}{|v|} \Rightarrow \frac{\langle Mv, \alpha\rangle^2}{|Mv|^2} > \frac{\langle v, \alpha\rangle^2}{|v|^2}$$

From equations (2.4-2.7) it follows that

$$\begin{aligned} \frac{\langle v, \alpha\rangle^2}{|v|^2} &= 1 - \frac{v_2^2(1-\langle \alpha, \beta\rangle^2)}{|v|^2} \\ \frac{\langle Mv, \alpha\rangle^2}{|Mv|^2} &= 1 - \frac{\lambda_2^2 v_2^2(1-\langle \alpha, \beta\rangle^2)}{|Mv|^2} \end{aligned}$$

By a little comparison and rearrangement of the above equations, the condition that needs to be shown is that $|Mv|^2 > \lambda_2^2|v|^2$.

From equation (2.4) it is straightforward to see that

$$\lambda_2^2|v|^2 = \lambda_2^2 v_1^2 + 2\lambda_2^2 v_1 v_2 \langle \alpha, \beta\rangle + \lambda_2^2 v_2^2$$

By direct comparison with equation (2.5) it is clear that the above expression is less than $|Mv|^2$, since $0 < \lambda_2 < \lambda_1$. Therefore in case 1, M reduces the angle between α and v, as claimed.

Case 2; $\langle v, \alpha\rangle \geq 0$ and $\langle Mv, \alpha\rangle < 0$: In this case, the angle between v and α is acute, while the angle between Mv and α is obtuse, which is not consistent with the claim made that M reduces the angle. Consequently, this case should result in a contradiction.

From equations (2.6) and (2.7) it follows that

$$-\langle \alpha, \beta\rangle \leq \frac{v_1}{v_2} \quad \text{and} \quad -\langle \alpha, \beta\rangle > \frac{\lambda_1 v_1}{\lambda_2 v_2} \tag{2.8}$$

This is a contradiction (see equation (2.1)), as expected.

Case 3; $\langle v, \alpha\rangle < 0$ and $\langle Mv, \alpha\rangle \geq 0$: In this case, the angle between α and v is decreased by M from an obtuse angle to an acute angle, the converse of case 2. This is consistent with the claim that M reduces the angle and unlike case 2 there is no contradiction: case 3 implies that $\frac{v_1}{v_2} < -\langle \alpha, \beta\rangle \leq \frac{\lambda_1 v_1}{\lambda_2 v_2}$, which is consistent with equation (2.1).

Case 4; $\langle v, \alpha\rangle$ and $\langle Mv, \alpha\rangle < 0$. In a similar way to case 1, this condition implies that $\lambda_2^2|v|^2 > |Mv|^2$, or equivalently $\lambda_2^2|v|^2 - |Mv|^2 > 0$. Thus, the validity of this case can be checked by verifying that the RHS of the following expression is positive:

$$\begin{aligned}\lambda_2^2|v|^2 - |Mv|^2 &= \left(\lambda_2^2 - \lambda_1^2\right) v_1^2 + 2\left(\lambda_2 - \lambda_1\right)\lambda_2 v_1 v_2\langle\alpha,\beta\rangle &(2.9)\\ &= \left(\lambda_2 - \lambda_1\right)\left(\lambda_2 + \lambda_1\right) v_1^2 + 2\left(\lambda_2 - \lambda_1\right)\lambda_2 v_1 v_2\langle\alpha,\beta\rangle &(2.10)\end{aligned}$$

By dividing equation (2.10) through by $v_1\left(\lambda_2 - \lambda_1\right)$ the expression becomes

$$\frac{\lambda^2|v|^2 - |Mv|^2}{v_1\left(\lambda_2 - \lambda_1\right)} = \left(\lambda_2 + \lambda_1\right) v_1 + 2\lambda_2 v_2\langle\alpha,\beta\rangle$$

This expression will be negative as required if the following inequality holds true:

$$-\langle\alpha,\beta\rangle > \frac{\left(\lambda_1 + \lambda_2\right) v_1}{2\lambda_2 v_2}\left(\equiv \frac{\left(\frac{\lambda_1+\lambda_2}{2}\right) v_1}{\lambda_2 v_2}\right) \qquad (2.11)$$

From the fact that $\langle Mv,\alpha\rangle < 0$ and equation (2.7) it follows that $-\langle\alpha,\beta\rangle > \frac{\lambda_1 v_1}{\lambda_2 v_2}$. Since trivially $\lambda_1 > \frac{\lambda_1+\lambda_2}{2}$ it then follows that expression (2.11) is true, and therefore M reduces the angle between α and v in this case.

Since it has been shown that of the four cases listed, one leads to a contradiction and is therefore invalid, while in each of the other three M decreases the angle between α and v as claimed, and since there are no other possible cases, the proof is complete. □

From this, the following result can be obtained:

Lemma 5. *Let M be a 2×2 real matrix with positive real eigenvalues, λ_1 and λ_2, such that $\lambda_1 > \lambda_2$. Let α be the unit eigenvector corresponding to λ_1 and β be the unit eigenvector corresponding to λ_2. Then M preserves any cone K that satisfies both the conditions*

1. *α or $-\alpha \in K$*
2. *β and $-\beta \notin \operatorname{int} K$*

Proof. For simplicity, assume (by redefining the sign of α if necessary) that $\alpha \in K$. Take any $\gamma \in K$, and write it as $\gamma = \gamma_1\alpha + \gamma_2\beta$. If $\gamma_2 = 0$, then $M\gamma = \lambda_1\gamma_1\alpha$, which is trivially in K. Likewise, if $\beta \in \partial K$, it is possible that $\gamma_1 = 0$, in which case $M\gamma = \lambda_2\gamma_2\beta$, which is also trivially in K.

Now consider the nontrivial cases, where γ is not a multiple of an eigenvector, and hence $M\gamma = \lambda_1\gamma_1\alpha + \lambda_2\gamma_2\beta$ with $\gamma_1, \gamma_2 \neq 0$. Suppose initially that $\alpha \in \partial K$. The sign of β can be chosen such that γ_1 and γ_2 are both greater than zero for all $\gamma \in K$. Therefore $M\gamma$ lies between γ and α, in the sense that the coefficient of β has the same sign in both γ and $M\gamma$, while the angle between $M\gamma$ and α is smaller than the angle between γ and α (by lemma 4). Since α and γ lie in K by assumption, and K is solid, this implies $M\gamma \in K$.

Now suppose instead that $\alpha \in \operatorname{int} K$. Other than the trivial cases where $\gamma_1 = 0$ or $\gamma_2 = 0$, there are two possibilities: either $\gamma_1 > 0$, $\gamma_2 > 0$, or $\gamma_1 > 0$, $\gamma_2 < 0$. In the first of these cases, lemma 4 guarantees directly that $M\gamma$ lies between γ and α as above, and therefore $M\gamma \in K$ since K is solid. For the second case, define $\beta' = -\beta$ and $\gamma_2' = -\gamma_2$. Then β' is also a unit eigenvector of M, and $\gamma = \gamma_1\alpha + \gamma_2'\beta'$. Since $\gamma_2' > 0$, lemma 4 once again guarantees that $M\gamma$ lies between α and γ and therefore $M\gamma \in K$. This completes the proof. □

Lemma 5 is too restrictive to be particularly useful in itself; while it is fairly straightforward to see how the result could be extended for matrices with positive real eigenvalues in higher dimensions, the situation becomes more difficult for negative and complex eigenvalues. The problem of identifying cones that are preserved by a set of matrices with certain structural properties is developed in much greater detail in [Banaji, 2008] (note, however, that lemma 5 is independent of the results in Banaji). The results presented there arise from chemical reaction networks (see chapter 6); they are of much greater practical use than the above result, but also serve to illustrate the difficulties in answering this type of question in full generality.

2.2.3 Monotone dynamical systems and convergence of trajectories

So far, the discussion in this chapter has been solely of cones and matrices, without explaining how the ideas can be used to analyse dynamical systems. In this section, the links between cone-preserving matrices and dynamical systems will be outlined. A **monotone dynamical system** is a continuous semiflow $\Phi(t, x)$ (see the discussion of flows in §1.2) on a metric space X with a partial order $\leq_P$ that is preserved by the flow, i.e.

$$\forall x, y \in X : x \leq_P y \Rightarrow \Phi(t, x) \leq_P \Phi(t, y) \;\forall\; t \in \mathbb{R}_{\geq 0}$$

Following [Walcher, 2001], a differential equation $\dot{x} = f(x)$ will be called **K-cooperative** with respect to a cone K if its Jacobian $Df(x)$ fulfils condition A. When the differential equation defining a dynamical system is K-cooperative and the phase space upon which it is defined is convex, trajectories of the dynamical system are monotone with respect to the partial ordering defined by K. The Muller-Kamke theorem states this formally:

Theorem 15. *Let $X \subseteq \mathbb{R}^n$ be open and nonempty, and let K be a cone. Assume that given any $x, y \in X$ satisfying $x \underset{K}{\leq} y$, the line segment between x and y lies wholly within X. Let points in X evolve according to the differential equation $\dot{x} = f(x)$ with solutions*

$\Phi(t,x) : \mathbb{R} \times X \to X$, and let $f(x)$ be K-cooperative. Then for all $x, y \in X$ satisfying $x \underset{K}{\leq} y$, solutions of the system preserve the ordering $\Phi(t,x) \underset{K}{\leq} \Phi(t,y)$ for all $t \geq 0$.

Proof. See proposition (1.5) of [Walcher, 2001]. □

A well-known example of a class of monotone dynamical systems is the set of cooperative systems, which were analysed in [Hirsch, 1982], [Hirsch, 1985] and many subsequent papers. In a cooperative system, variables are mutually activatory, which is to say that increasing the value of any variable can only increase or leave unchanged the rate of production of the other variables. Trajectories of such systems obey the positive orthant ordering, and the Jacobian of a cooperative system is quasipositive.

Condition A is fulfilled by any K-quasipositive matrix, as pointed out earlier, hence K-quasipositivity of the Jacobian matrix of a function suffices to guarantee monotonicity of the function's solutions on a convex set by theorem 15. A monotone dynamical system cannot have stable non-trivial periodic orbits, as noted in [Hirsch and Smith, 2005] (see section 1.1). More restrictive results on monotone systems are required to obtain conditions for global convergence. The following definitions and the resulting theorem appear in [Smith, 1995].

Let X be an ordered metric space. A semiflow Φ on X is **strongly order preserving** if it is monotone, and when $x < y$ for $x, y \in X$, $\exists\ U, V$ with $x \in U$, $y \in V$ and $t_0 \geq 0$ such that $\Phi_{t_0}(U) \leq \Phi_{t_0}(V)$, where U and V are open subsets of X. By monotonicity of Φ it follows that $\Phi_t(U) \leq \Phi_t(V)\ \forall\ t \geq t_0$.

A point $x \in X$ can be **approximated from below** in X if $\exists$ a sequence $\{x_n\} \in X$ with the property $x_n < x_{n+1} < x$ for $n \geq 1$ and $x_n \to x$ as $n \to \infty$. Similarly, x can be **approximated from above** in X if $\exists$ a sequence $\{x_n\} \in X$ with the property $x < x_{n+1} < x_n$ for $n \geq 1$ and $x_n \to x$ as $n \to \infty$.

Theorem 16. *Let $X \subseteq \mathbb{R}^n$ be an ordered metric space, with all points $x \in X$ evolving according to some C^1 function f, the solutions of which are the set of semiflows $\Phi(t,x)$. Suppose that*

1. *X contains a unique equilibrium.*
2. *Every point in X (possibly excluding the equilibrium) can be approximated from above and below.*
3. *All semiflows Φ are strongly order preserving.*

Then the equilibrium is globally attractive.

Proof. See theorem 2.3.1 on p. 18 of [Smith, 1995]. □

Theorem 16 relies on being able to identify when the flows of a dynamical system are strongly order preserving. In general, this is a difficult problem, but the following result helps in some cases:

Theorem 17. *Let $X \subseteq \mathbb{R}^n$ be convex, $K \subset \mathbb{R}^n$ be a proper cone, and $f : X \to \mathbb{R}^n : x \mapsto f(x)$ be a C^1 function. For each point $x \in X$ assume that there exists some $\lambda \in \mathbb{R}$ such that $Df(x) + \lambda I$ is K-nonnegative and K-irreducible. Then the flows defined by f are strongly order preserving with respect to the cone K.*

Proof. This claim is an amalgamation of several results from the theory of monotone flows. The way that these results fit together is only briefly outlined here; the interested reader is referred to [Hirsch and Smith, 2005] for more details and technical discussion.

If $Df(x) + \lambda I$ is both K-nonnegative and K-irreducible, then for all $y_1 \in \partial K \setminus \{0\}$ there is some $y_2 \in K^*$ satisfying $\langle y_2, y_1 \rangle = 0$ and $\langle y_2, Df(x)y_1 \rangle > 0$ (henceforth referred to as condition B). This result appears as lemma 3.10 of [Hirsch and Smith, 2005]. Notice that condition B is closely related to condition A, but the inequality is strict in this case and does not necessarily hold for **all** $y_2 \in K^*$ satisfying $\langle y_2, y_1 \rangle = 0$.

By lemma 3.7 of [Hirsch and Smith, 2005], the flows of a function f defined on a convex set are strongly monotone if f fulfils condition B, and proposition 1.2 from the same reference states that flows of f are strongly order preserving if they are strongly monotone. This completes the proof. □

Given a set X with a partial order relation $\leq$, y is a lower (upper) bound of $C \subseteq X$ if $y \in X$ and $y \leq x \ \forall x \in C$ ($x \leq y \ \forall x \in C$). The **infimum** of C, which may or may not exist, is denoted $\inf C$ and is the greatest lower bound of C, i.e. $\inf C \geq y$ for every $y \in X$ that is a lower bound of C. Likewise, the **supremum** of C, denoted $\sup C$, is the least upper bound of C.

Using these definitions, a result similar to theorem 17 that guarantees global convergence via monotonicity is as follows:

Theorem 18 (De Leenheer et al.)**.** *Consider a metric space X with metric d and suppose that a partial order compatible with d has been defined on X. Let Φ be a continuous semiflow on X.*

Suppose the following conditions on X and Φ hold:

1. $\inf C, \sup C \in X$ *for every compact subset C of X.*

2. Φ *is monotone with respect to the ordering on X.*

3. *There is a unique equilibrium in X.*
4. *The orbit of every point in X has compact closure in X in forward time.*

Then the equilibrium in X is globally attractive for Φ.

Proof. See theorem 5 of [De Leenheer et al., 2007]. □

Note that these results only imply global **attractivity** of the equilibrium; for the fixed point to be globally **asymptotically stable** it must also be shown to be locally asymptotically stable.

Conditions 2 – 4 in theorem 18 are often relatively straightforward to check for a given dynamical system, but condition 1 is more problematic. [De Leenheer et al., 2007] includes a result that can be used in some cases to verify that this condition is met, for which the following definitions will be required:

1. A partially ordered set X is a **lattice** if there exists $\sup\{p, q\} \in X$ and $\inf\{p, q\} \in X$ for all $p, q \in X$.
2. A **closed order interval** of two points $p, q \in X$ is defined as $[p, q] = \{x \in X | p \leq x \leq q\}$.
3. A set S is **order bounded** in X if $\exists\, y, z \in X$ s.t. $S \subseteq [y, z]$.

Lemma 6 (De Leenheer et al)**.** *Let Y be a finite-dimensional normed vector space with cone K and let $X \subset Y$ be a lattice. Suppose that every bounded set in X is also order bounded in X. If C is a compact subset of X, then* $\inf C, \sup C \in X$.

Proof. See lemma 4 of [De Leenheer et al., 2007]. □

Note that when $Y = \mathbb{R}^n$ and $X = \mathbb{R}^n_{\geq 0}$, with $K = \mathbb{R}^n_{\geq 0}$ representing the standard ordering, then every bounded subset of X is a subset of an n-rectangle with faces parallel to the faces of K and vertices in X. To see why this is the case, consider that 0 is a lower bound of every point in X and every bounded subset $C \subset X$ must also have a finite upper bound $u(C) \in X$. The order interval $[0, u(C)]$ is a rectangle in n dimensions containing C. Thus X is a lattice, and every bounded set $C \subset X$ is also order bounded in X with respect to K since $C \subseteq [0, u(C)]$. Thus condition 1 of theorem 18 is met when $X = \mathbb{R}^n_{\geq 0}$ and $K = \mathbb{R}^n_{\geq 0}$.

By a similar argument, it is straightforward to see that condition 1 of theorem 18 is satisfied whenever $X = \mathbb{R}^n_{\geq 0}$ and K is **any** orthant. This can be further generalised to show that condition 1 of theorem 18 is satisfied for $X = \mathbb{R}^n_{\geq 0}$ and any simplicial cone K that is a superset of an orthant.

Lemma 7 below shows that simplicial cones that cover an orthant can be used to guarantee global convergence of trajectories. Figure 2.2 provides a simple two dimensional depiction of the idea behind the lemma. Given a bounded set B lying entirely within X, a rectangle N can be constructed that contains B and also lies entirely within X. Since K is a superset of an orthant P, the inf and sup of N with respect to K are equal to the inf and sup of N with respect to P, and hence lie within X. With respect to K, $\inf N \leq \inf B \leq \sup B \leq \sup N$, guaranteeing that $\inf B$ and $\sup B$ also lie within X.

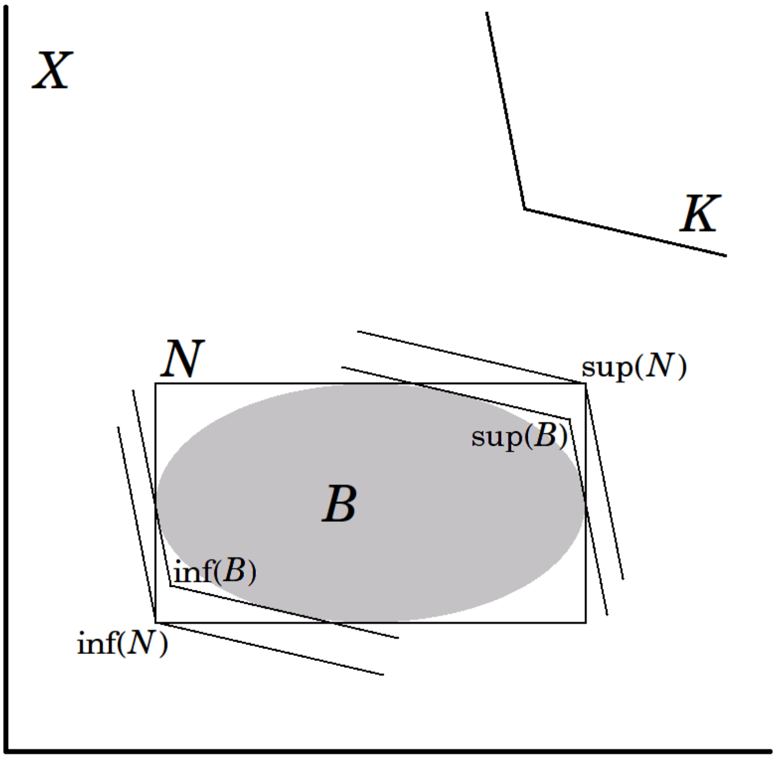

Figure 2.2: A two dimensional example of a cone that covers the standard orthant $\mathbb{R}^2_{\geq 0}$. The axes represent X, the nonnegative orthant. The cone K is shown at the top right. B is an arbitrary bounded set in X and N is the smallest rectangular region with sides parallel to the faces of the positive orthant that contains B.

Lemma 7. *When $X = \mathbb{R}^n_{\geq 0}$, then X is a lattice and every bounded set B in X is order bounded in X with respect to a simplicial cone K if $K \supseteq P$ where P is an orthant. Conversely, if $K \not\supseteq P$ for an orthant P then X is not a lattice.*

Proof. An orthant corresponds to an index set of coordinate components that are nonnegative, which will be labelled P_+, and an index set of coordinates that are nonpositive, which will be labelled P_-. By definition, $P_+ \cup P_- = \{1, \ldots, n\}$ and $P_+ \cap P_- = \emptyset$. Let $B \subset X$ be a bounded set, and let y_i be the ith component of a vector y. Define a set of half spaces as follows:

1. For each $i \in P_+$, define a half space that is the set of all points $x \in \mathbb{R}^n$ satisfying $x_i \leq \sup\limits_{y\in B} y_i$. Label the set of all such half spaces as H_A.
2. For each $i \in P_+$, define a half space that is the set of all points $x \in \mathbb{R}^n$ satisfying $x_i \geq \inf\limits_{y\in B} y_i$. Label the set of all such half spaces as H_B.
3. For each $i \in P_-$, define a half space that is the set of all points $x \in \mathbb{R}^n$ satisfying $x_i \geq \inf\limits_{y\in B} y_i$. Label the set of all such half spaces as H_C.
4. For each $i \in P_-$, define a half space that is the set of all points $x \in \mathbb{R}^n$ satisfying $x_i \leq \sup\limits_{y\in B} y_i$. Label the set of all such half spaces as H_D.

Note that each inf and sup above is guaranteed to exist since B is bounded. Let $N = H_A \cap H_B \cap H_C \cap H_D$. Thus N is the smallest n-rectangle with faces parallel to the faces of X that contains B. Clearly $N \subset X$. Let $l_N = \inf_P N$ and $u_N = \sup_P N$, taking $\inf_P$ and $\sup_P$ with respect to the ordering generated by the orthant P. Clearly l_N, u_N lie in X. For every $x \in N$, both $x - l_N \in P$ and $u_N - x \in P$, and since $K \supseteq P$ it follows that $x - l_N \in K$ and $u_N - x \in K$. Therefore N is order bounded in X with respect to K and consequently B is also order bounded in X with respect to K.

To complete the first part of the lemma, it must be demonstrated that the ordering generated by K makes $\mathbb{R}^n_{\geq 0}$ a lattice. Since K is simplicial, p and q can be uniquely written in terms of normalised generators of K: $p = \sum_{i=1}^n p_i \hat{g}_i, q = \sum_{i=1}^n q_i \hat{g}_i$. Let $m_i = \min\{p_i, q_i\}$ for $i = 1, \ldots, n$ and define $m = \sum_{i=1}^n m_i \hat{g}_i$. By this definition, $m \leq p, q$ since $p - m = \sum_{i=1}^n (p_i - m_i)\hat{g}_i \geq_K 0$ and similarly for $q - m$.

Now consider any $m' = \sum_{i=1}^n m'_i \hat{g}_i$. If there is some i such that $m'_i > p_i$ then $m' \not\leq_K p$ (equivalently, if for some i, $m'_i > q_i$, then $m \not\leq_K q$). Thus for any m' such that $m' \leq_K p, q$, it follows that $m'_i \leq p_i, q_i$, and therefore $m'_i \leq \min\{p_i, q_i\}$. Consequently $m' \leq_K m$, and so $m = \inf_K\{p, q\}$.

A similar argument proves the existence and uniqueness of $\sup_K\{p, q\}$. Suppose now that $N(p, q)$ is the minimum n-rectangle containing p and q. Since $l_{N(p,q)}$ is a lower bound of p, q and $l_{N(p,q)}$ lies within X, it follows that the set of lower bounds of $N(p, q)$ lying within X is nonempty and therefore $\inf_K\{p, q\}$ lies within X. The same is true of $u_{N(p,q)}$ and $\sup_K\{p, q\}$. This demonstrates that X is a lattice, concluding the first part of the lemma.

For the second part, suppose per contra that K is a cone that does not cover any orthant, but that $X = \mathbb{R}^n_{\geq 0}$ is a lattice, i.e. $\inf_K\{p, q\} \in \mathbb{R}^n_{\geq 0}$ and $\sup_K\{p, q\} \in \mathbb{R}^n_{\geq 0}$ for all $p, q \in \mathbb{R}^n_{\geq 0}$. Since K does not cover any orthant, there exists $i \in \{1, \ldots, n\}$ such that $x_i \neq 0$ for all $x \in K$, excluding $x = 0$. Assume without loss of generality that this means $x_i > 0$ for all nonzero $x \in K$ (if this is not the case, simply redefine K to be $-K$). The relation $x_i \geq 0$ defines a half space, which will be labelled $\mathcal{H}$. Clearly $K \subset \mathcal{H}$ and $\mathbb{R}^n_{\geq 0} \subset \mathcal{H}$.

Since $x_i < 0$ for all nonzero $x \in -K$, it follows that $-K \cap \mathcal{H} = \{0\}$ and therefore also $-K \cap \mathbb{R}^n_{\geq 0} = \{0\}$.

Since K does not cover an orthant, there exists $y \in \partial\mathbb{R}^n_{\geq 0}$ such that $y \notin K$ and $y \notin -K$ — this is true of any nonzero, nonnegative y for which $y_i = 0$. Such a y is by definition unordered (under the ordering defined by K) with respect to the origin, i.e. $0 \not\leq_K y$ and $y \not\leq_K 0$. Therefore $\inf_K\{0, y\} \neq 0$. Again by definition, $\inf_K\{0, y\} \in -K$, since $\inf_K\{0, y\} \leq 0$ and $-K = \{x \mid x \leq 0\}$. However, as $\inf_K\{0, y\} \neq 0$ and $-K \cap \mathbb{R}^n_{\geq 0} = \{0\}$, it follows that $\inf_K\{0, y\} \notin \mathbb{R}^n_{\geq 0}$. Therefore $\inf_K\{p, q\} \notin \mathbb{R}^n_{\geq 0}$ when $p = 0$ and $q = y$, even though $0, y \in \mathbb{R}^n_{\geq 0}$. This contradicts the assumption that X is a lattice, and therefore K must necessarily cover an orthant. □

The above result is rather technical; however it is potentially useful in demonstrating global attractivity of some dynamical systems by means of theorem 18. As noted after that theorem, the conditions required are relatively easy to check, with the exception of the first: if $X \subseteq \mathbb{R}^n$ is the phase space, $\inf C$ and $\sup C$ must lie in X for all compact $C \subseteq X$. It is not uncommon in applications for the phase space of a system to be the nonnegative orthant, for example in chemical reaction networks where reactant concentrations cannot be negative, or models of population dynamics, where negative populations are not allowed. In such systems where the phase space is $\mathbb{R}^n_{\geq 0}$, all trajectories are closed and bounded and there is a unique fixed point, lemma 7 means that if the flows of the system preserve an ordering defined by a simplicial cone that covers an orthant, then the fixed point is globally attracting.

Chapter 3

Autonomous convergence criteria

The second technique used to derive conditions for convergence of a dynamical system is based on a family of "autonomous convergence theorems." These theorems can loosely be thought of as a spiritual successor to the Markus–Yamabe theorem, which was originally proposed in [Markus and Yamabe, 1960]. Both the Markus–Yamabe theorem and the later autonomous convergence theorems relate to properties of the Jacobian matrix. Recall that the Jacobian matrix gives **local** information about a differential equation, but, as with the previously presented results on monotonicity of dynamical systems, properties of the Jacobian that hold over the whole of phase space are used to make claims about the **global** behaviour of the differential equation.

Theorem 19 (Markus–Yamabe)**.** *Let $f : \mathbb{R}^2 \to \mathbb{R}^2 : x \mapsto f(x)$ be a $\mathcal{C}^1$ map satisfying $f(0) = 0$. If the Jacobian matrix $Df(x)$ is Hurwitz stable $\forall\ x \in \mathbb{R}^2$ then the point $x = 0$ is globally asymptotically stable for the dynamical system described by $\dot{x} = f(x)$.*

Proof. A number of proofs have been published, see any of [Glutsyuk, 1994], [Feßler, 1995] or [Gutierrez, 1995]. □

It is important to note that the Markus–Yamabe theorem only holds for dynamical systems in two dimensions. It was conjectured that the result would also hold in higher dimensions, but counterexamples were subsequently discovered; see [Bernat and Llibre, 1996] and [Cima et al., 1997].

As in the previous chapter, most of the results presented are not new, but are quoted from other sources. The new results in this chapter are theorem 21, corollary 3, lemma 11, theorem 24, lemma 12 and lemma 13.

3.1 First autonomous convergence theorem

The first so-called autonomous convergence theorem was constructed by Russell Smith in [Smith, 1986]; it will be discussed later on as it is somewhat complicated. A simpler autonomous convergence theorem was subsequently constructed by Verbitskii and Gorban, and is presented here first due to requiring less background. Note that the presentation of Verbitskii and Gorban's result that follows is not based directly on their original paper [Verbitskii and Gorban, 1992], but instead on [Banaji and Baigent, 2008], which contains an independent derivation of the same result.

3.1.1 Logarithmic norms

Some definitions are required for the presentation of the theorems. Both of the autonomous convergence theorems stated in this chapter rely on the **logarithmic norm** or **Lozinskiĭ measure** of a matrix, which was independently introduced by Germund Dahlquist and Sergei Lozinskiĭ in 1958. For a more detailed overview of logarithmic norms than appears here, see [Ström, 1975] or [Söderlind, 2006].

Every normed vector space has an associated logarithmic norm. Let V be a vector space and $|\cdot|_n$ be a vector norm. By definition, $|\cdot|_n$ satisfies the three properties (a) $|v|_n \geq 0$ for all $v \in V$, $|v| = 0$ if and only if $v = 0$, (b) $|kv|_n = k|v|_n$ for all $v \in V$ and $k \in \mathbb{R}_{\geq 0}$, and (c) $|v + w|_n \leq |v|_n + |w|_n$. The vector norm induces a matrix norm $\|\cdot\|_n$, where for a square matrix M,

$$\|M\|_n = \sup_{|x|=1} |Mx|_n$$

A logarithmic norm μ_n is defined for a square complex matrix M as

$$\mu_n(M) = \lim_{h \to 0^+} \frac{\|\mathbb{I} + hM\|_n - 1}{h} \tag{3.1}$$

$\mathbb{I}$ is the identity matrix of the same dimension as M. It is known that the limit in h in equation (3.1) exists and convergence to it is monotonic [Ström, 1975]. The logarithmic norm is related to the matrix norm used to generate it, but unlike a matrix norm, a logarithmic norm can be negative. Note that a logarithmic norm is **only** defined in relation to a matrix norm; it does not exist independently. Intuitively speaking, if a matrix M has a negative logarithmic norm then the linear transform $I + hM$ maps the unit sphere (as defined by the corresponding vector norm) inside itself for small enough h. Equivalently, a negative logarithmic norm provides information about how much a linear mapping rotates vectors. For example, any linear map that rotates all vectors by over 90° will have negative Euclidean logarithmic norm. For norms other than the Euclidean norm, the limits on the angle by which vectors are rotated by a linear transform with negative logarithmic norm will vary depending on which direction the vector initially pointed. Figure 3.1 provides a simple illustration of this in two dimensions.

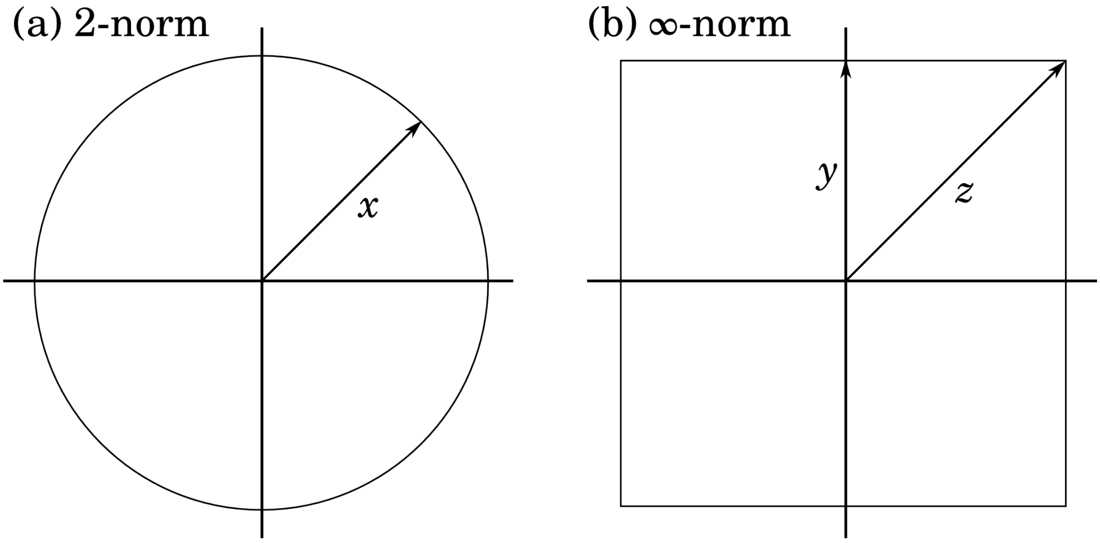

Figure 3.1: A diagram of the unit circle according to the 2-norm (Euclidean norm) and ∞-norm. A linear transform M has negative logarithmic norm if there exists some h such that $I + hM$ maps the unit circle inside itself. In the case of the Euclidean norm shown in (a), the vector x lies on the Euclidean unit circle, and there exists some h such that $(I + hM)x$ lies within this unit circle if and only if M rotates x by an angle greater than 90°. For this norm, it is relatively straightforward to see that every vector on the unit circle must be rotated by more than 90° by M in order for there to be some h such that the linear transform $I + hM$ maps the unit circle inside itself. The situation is more complicated for other norms, such as the ∞-norm shown in (b). The vector y lies on the ∞-norm unit circle, and in order for y to be mapped inside this unit circle by the linear transform $I + hM$ for some h, M must rotate y by an angle greater than 90°, as was the case for the Euclidean norm. However, the vector z, which also lies on the ∞-norm unit circle, must be rotated by an angle greater than 135° in order for there to exist some h such that $(I+hM)z$ lies within the unit circle. For other vectors that don't lie on an axis of symmetry of the ∞-norm unit circle, the minimum rotation required for M to have negative logarithmic ∞-norm is different depending on whether M rotates the vector clockwise or anticlockwise.

For a matrix M with eigenvalues $\sigma(M)$, let $\alpha(M) = \max(\mathrm{Re}(\sigma(M)))$, known as the **stability modulus** or **spectral abscissa** of a matrix. The following properties of the logarithmic norm will be used later:

$$\begin{aligned} \alpha(M) &\leq \mu(M) && (3.2)\\ \mu(M+N) &\leq \mu(M)+\mu(N) && (3.3)\\ \alpha(M) &= \inf \mu(M) && (3.4) \end{aligned}$$

The first two of these relations appear in lemma 1c of [Ström, 1975], and the third appears as proposition 2.3 of [Li and Wang, 1998]. The meaning of property (3.4) may not be obvious at first glance, but it is essentially a synthesis of two more straightforward statements: there does not exist any logarithmic norm μ such that $\mu(M) < \alpha(M)$ (which is a trivial corollary of property (3.2)), and given any $\epsilon > 0$, there exists some logarithmic norm μ such that $|\mu(M) - \alpha(M)| < \epsilon$ (see the proof of proposition 2.3 in [Li and Wang, 1998]).

In general, logarithmic norms are often difficult to construct explicitly; here are two of the better known ones. Recall that for a vector $v \in \mathbb{R}^n$, $|v|_1 = \sum_i |x_i|$ and $|v|_\infty = \max_i |x_i|$. These vector norms induce the matrix norms $\|M\|_1 = \sup_{|x|_1=1} |Mx|_1$ and $\|M\|_\infty = \sup_{|x|_\infty=1} |Mx|_\infty$. The corresponding logarithmic norms of a finite-dimensional matrix $M = (M_{ij})$ are [Li and Wang, 1998]

$$\mu_1(M) = \max_j \left(\mathrm{Re}(M_{jj}) + \sum_{i, i\neq j} |M_{ij}| \right) \tag{3.5}$$

$$\mu_\infty(M) = \max_i \left(\mathrm{Re}(M_{ii}) + \sum_{j, j\neq i} |M_{ij}| \right) \tag{3.6}$$

New logarithmic norms can also be constructed from a known logarithmic norm. Let μ_n be some logarithmic norm. If T is an invertible matrix, then a new logarithmic norm $\mu_{n,T}$ can be constructed from μ_n as follows:

Lemma 8 (Li and Wang)**.** *Suppose μ_n is a logarithmic norm corresponding to a vector norm $|\cdot|_n$. Define a new vector norm $|\cdot|_{n,T}$ according to $|x|_{n,T} = |Tx|_n$ for any x and some fixed invertible matrix T of suitable dimension. The corresponding logarithmic norm is then*

$$\mu_{n,T}(M) = \mu_n(TMT^{-1})$$

Proof. See [Li and Wang, 1998], lemma 2.2. □

By considering lemma 8 and the forms of equations (3.5) and (3.6), it is apparent that the infimum in property (3.4) is attained for any diagonalisable M.

3.1.2 Statement of theorem and interpretation

Now that the definitions are in place the first autonomous convergence theorem can be presented. The theorem gives conditions for global asymptotic stability of an autonomous differential equation:

Theorem 20 (Banaji and Baigent). *Let $\dot{x} = f(x)$ be an autonomous differential equation defined on a convex forward invariant subset of $\mathbb{R}^n$, with $Df(x)$ being the Jacobian matrix. If the system has an equilibrium and there exists a logarithmic norm μ such that $\mu(Df(x)) < 0$ for all x then the equilibrium is both unique and globally stable.*

Proof. See [Banaji and Baigent, 2008]. □

For an earlier statement of essentially the same result, see [Verbitskii and Gorban, 1992].

The key difference between this theorem and the Markus–Yamabe conjecture is that the **same** norm must be applied at all points in phase space. The property of logarithmic norms stated in equation (3.4) above implies that if the Jacobian J of a dynamical system is everywhere Hurwitz, then at each point in phase space there exists **some** logarithmic norm μ_p satisfying $\mu_p(J) < 0$. Clearly, since the Markus–Yamabe conjecture is known not to hold in more than two dimensions, this condition is insufficient to guarantee global stability of a unique fixed point. A simple illustration of why a decrease in distance between two trajectories according to different norms does not guarantee convergence appears in figure 3.2. However, theorem 20 shows that when the same logarithmic norm is negative at all points then global convergence does occur. As such, the condition that the same logarithmic norm be used at all points in phase space is probably unnecessarily strong. It is perhaps possible that a condition guaranteeing global convergence could be constructed allowing a different logarithmic norm to be used at different points in phase space, provided that the choice of norm used for neighbouring regions of phase space didn't change "too fast". However, such a condition appears very difficult to construct.

A negative logarithmic norm at a point in phase space means that the distance between neighbouring trajectories as measured by the associated vector norm is decreasing under flows of f. This can be explicitly demonstrated by a simple extension of the proof of theorem 3 in [Banaji and Baigent, 2008]. Showing this relies on lemma 2 from [Ström, 1975], which is as follows:

Lemma 9. *Let $A(t)$ be an $n \times n$ matrix, and let x be a vector evolving according to the equation $\dot{x} = A(t)x + \rho(t)$. Then, given a vector norm $|\cdot|_n$, $|x(t)|_n \leq e(t)$ where $e(t)$ is any solution to the equation $\dot{e} = \mu_n(A(t))e(t) + |\rho(t)|$ satisfying $e(0) \geq |x(0)|$.*

More advanced results similar to the above lemma exist; see for example [Neumaier, 1994].

Theorem 21. *Let $\dot{x} = f(x)$ be an autonomous differential equation defined on X, a forward invariant convex subset of $\mathbb{R}^n$, and let $Df(x)$ be the Jacobian matrix. If there exists a logarithmic norm μ such that $\mu(Df(x)) < 0$ for all $x \in X$ then, for any two solutions $y_1(t)$, $y_2(t)$ of f, $|y_1(t) - y_2(t)| \to 0$ as $t \to \infty$ for all $y_1, y_2 \in X$.*

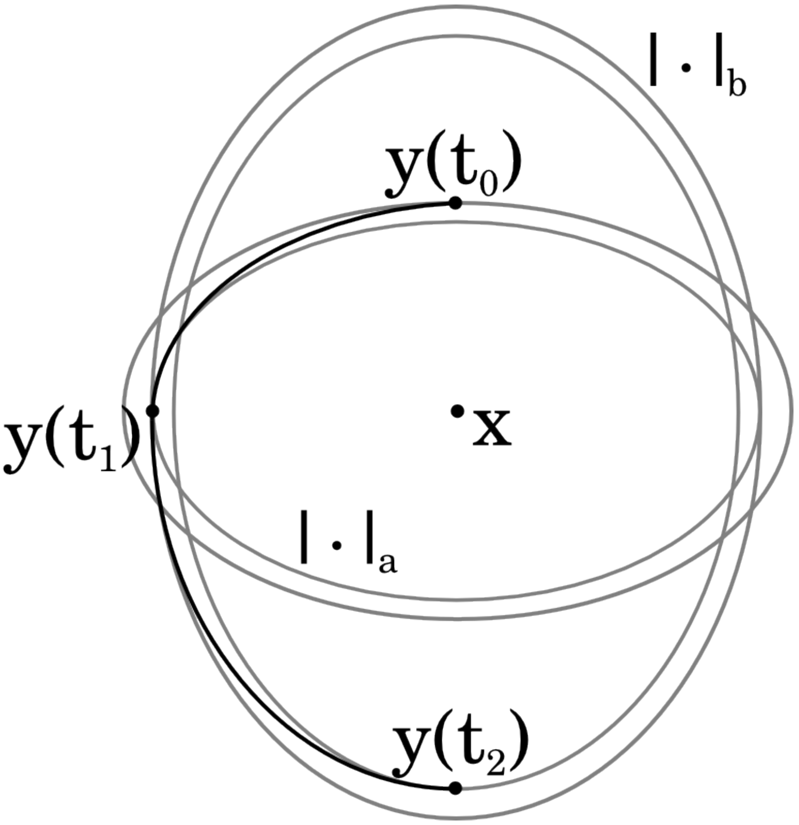

Figure 3.2: A diagram of a trajectory approaching a fixed point according to two different norms. x represents a fixed point and $y(t)$, the curve marked in black, is the trajectory of some point y. The grey ellipses aligned with the horizontal are level sets of distance from x with respect to some norm $|\cdot|_a$, and the ellipses aligned with the vertical are level sets of distance from x with respect to a different norm $|\cdot|_b$. Clearly $|y(t_1)-x|_a < |y(t_0)-x|_a$ and $|y(t_2)-x|_b < |y(t_1)-x|_b$. However, $|y(t_2)-x|_2 > |y(t_0)-x|_2$, where $|\cdot|_2$ is the Euclidean norm. Thus if $t_0 < t_1 < t_2$ it is apparent that y is not converging to x over the interval $[t_0, t_2]$, illustrating the fact that even if y approaches x according to first one and then another norm, this does not guarantee convergence in absolute terms.

Proof. The proof of theorem 3 in [Banaji and Baigent, 2008] is similar; the proof that appears here is a straightforward extension of Banaji and Baigent's result. The first step is to create a new variable $z(t) = y_1(t) - y_2(t)$. The time derivative of z is $\dot{z} = \dot{y_1} - \dot{y_2} = f(y_1) - f(y_2)$. In order to find an explicit expression for this statement, consider another new variable $w = \theta y_1 + (1-\theta)y_2$, where $\theta \in [0,1]$ is a dummy variable. By inspection the expression for w defines the straight line segment between y_1 and y_2. Consider the derivative of $f(w)$ with respect to θ:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} f(\theta y_1(t) + (1-\theta)y_2(t)) = Df(\theta y_1(t) + (1-\theta)y_2(t))(y_1 - y_2)$$

Integrating both sides with respect to θ over the interval $[0, 1]$ gives

$$f(y_1) - f(y_2) = \int_0^1 Df(\theta y_1(t) + (1-\theta)y_2(t))\,\mathrm{d}\theta\,(y_1 - y_2)$$

Define the matrix $A(t) = \int_0^1 Df(\theta x(t) + (1-\theta)y(t))\mathrm{d}\theta$. Therefore $\dot{z}$ can be simply restated as $\dot{z} = A(t)z$. By lemma 9, $|z(t)| \leq e(t)$, where $e(t)$ is any solution to the equation $\dot{e} = \mu(A(t))e(t)$ satisfying $e(0) \geq |z(0)|$. The solution to this equation takes the form

$$e(t) = e(0)\exp\left(\int_0^t \mu(A(s))\,\mathrm{d}s\right) \tag{3.7}$$

As demonstrated in the proof of theorem 3 in [Banaji and Baigent, 2008],

$$\mu\left(\int_0^1 Df(\theta y_1(t) + (1-\theta)y_2(t))\,\mathrm{d}\theta\right) \leq \int_0^1 \mu(Df(\theta y_1(t) + (1-\theta)y_2(t)))\,\mathrm{d}\theta < 0$$

The final inequality follows directly from the assumption that $\mu(Df) < 0$, and implies that $\mu(A(t)) < 0$. Consequently $\dot{e} < 0$ for all positive e: equation (3.7) can be rewritten $e(t) = e(0)\exp(-S(t))$ where $S(t)$ is an increasing function of time, meaning that $e(t) \to 0$ as $t \to \infty$. Since $|z(t)| \leq e(t)$ and $z(t) = y_1(t) - y_2(t)$, it follows that $|y_1(t) - y_2(t)| \to 0$ as $t \to \infty$, proving the result. □

This result essentially means that given a differential equation with a Jacobian that has a negative logarithmic norm, for every pair of points, the distance between the trajectories of the points is shrinking when measured using the vector norm associated with the logarithmic norm. When the set X contains a fixed point, the distance between every solution and the fixed point decreases, and solutions are therefore bounded. This is by contrast with the Markus–Yamabe conjecture (which assumes the existence of a fixed point), under which all trajectories nearby to a point will be converging according to at least one norm, but the distance between neighbouring trajectories may well increase with time when the same norm is used to measure distance at all times along the trajectories.

Since the solutions of system that has a fixed point and a negative logarithmic norm are bounded by the argument above, and any system that is bounded has a fixed point (theorem 3, p. 17), any unbounded dynamical system that has a negative logarithmic norm necessarily has no fixed points (as noted in [Banaji and Baigent, 2008]).

A similar technique for demonstrating convergence of trajectories using the logarithmic norm, referred to by its authors as “contraction analysis,” was independently introduced in [Lohmiller and Slotine, 1998].

3.2 Second autonomous convergence theorem

The first version of an autonomous convergence theorem was published by Russell Smith in [Smith, 1986]. This result was later refined by Li and Muldowney in [Muldowney, 1990], [Li and Muldowney, 1993], [Li and Muldowney, 1995] and [Li and Muldowney, 1996]. It relies on some extra definitions and results.

Let $\mathcal{M}$ be a compact manifold, and let $\Phi : \mathbb{R} \times \mathcal{M} \to \mathcal{M} : (t, x) \mapsto \Phi(t, x)$ be a flow of a C^1 vector field on $\mathcal{M}$. A point $y \in \mathcal{M}$ is **non-wandering** if, for every neighbourhood Y of y in $\mathcal{M}$ and every time $t_0 > 0$, there exists some $z \in Y$ and $t > t_0$ such that $\Phi(z, t) \in Y$ [Pugh, 1967].

Suppose that $f : M \to M : x \mapsto f(x)$ and $g : M \to M : x \mapsto g(x)$ are functions defined on a metric space (M, d). The $\mathbf{C^1}$ **topology** is then defined by the metric $d(f, g) = \max(\sup_x d(f(x), g(x)),\ \sup_x d(Df(x), Dg(x)))$, where Df and Dg are the derivatives of f and g respectively [Hasselblatt and Katok, 2003].

Intuitively speaking, it seems reasonable that given a C^1 map f, since the trajectory of any non-wandering point of the flow of f passes arbitrarily close to itself, there is another map g arbitrarily close to f in the C^1 topology for which the trajectory of the non-wandering point is a periodic orbit. This idea is formally stated as Pugh's closing lemma:

Lemma 10 (Pugh). *Let $\mathcal{M}$ be a compact manifold, and let $f : \mathcal{M} \to \mathcal{M} : x \mapsto f(x)$ be a $\mathcal{C}^1$ function. Let $y \in \mathcal{M}$ be a non-wandering point of the flow defined by f. Then there exists a function g arbitrarily close to f in the $\mathcal{C}^1$ topology such that y is a periodic point of the flow defined by g.*

Proof. See [Pugh, 1967]. □

Note that in the event that y is a fixed point, it is both non-wandering and trivially periodic under the flow defined by f. The interesting part of the result is for points that are both non-periodic and non-wandering under the flow defined by f. Pugh's closing lemma means that for any continuous time real dynamical system with all orbits bounded, any non-periodic (e.g. chaotic, quasiperiodic) orbit is arbitrarily close to a periodic orbit. Russell Smith used this result to prove that certain conditions precluding the existence of periodic orbits in a given region of the phase space of a dynamical system also imply that the region cannot contain any non-wandering points other than fixed points. To do this, Smith used a spectral condition on the **second additive compound** of the Jacobian matrix, equivalent to the μ_2 norm being negative.

An overview of compound matrices as they relate to autonomous convergence follows in the next section, based partly on [Li and Wang, 1998], partly on [Muldowney, 1990], and partly on [Allen and Bridges, 2002]. There are several types of compound matrices, and for some types there is not a clear consensus on the naming convention. The results in this thesis concern themselves with two types of compound matrix, which will be referred to as additive compound matrices and multiplicative compound matrices, following [Muldowney, 1990]. The kth additive compound of a matrix M will be denoted $M^{[k]}$ and the kth multiplicative compound will be denoted $M^{(k)}$. Additive compound matrices satisfy $(A+B)^{[k]} = A^{[k]} + B^{[k]}$ and multiplicative compound matrices satisfy $(AB)^{(k)} = A^{(k)}B^{(k)}$, hence the names.

3.2.1 Structure of additive compound matrices

Additive and multiplicative compound matrices are closely associated with exterior algebras and the **exterior product** or **wedge product**. Exterior algebras have a number of applications and form a significant topic in their own right; out of necessity the discussion here is limited to an outline of their relationship with compound matrices. See a reference such as [Bishop and Goldberg, 1980] for a fuller discussion of exterior algebras and their relationship with more general tensors.

The exterior (or Grassmann) algebra of a vector space V is denoted $\Lambda(V)$, with V a subspace of $\Lambda(V)$. The wedge product, denoted $\wedge$, is a generalisation of the three-dimensional vector cross product, and is a bilinear form representing multiplication in the exterior algebra:

$$\wedge : \Lambda(V) \times \Lambda(V) \to \Lambda(V) : (x, y) \mapsto x \wedge y \ (x, y \in \Lambda(V))$$

The wedge product is associative and distributive, but not generally commutative. It has the property $x \wedge x = 0$ for all $x \in \Lambda(V)$. Wedge products are associated with areas, volumes and their higher-dimensional analogues: for example, if $dx, dy, dz \in V$ represent line elements (referred to as 1-forms), then $dx \wedge dy, dx \wedge dz$ and $dy \wedge dz$ represent area elements or 2-forms and $dx \wedge dy \wedge dz$ represents a volume element or 3-form. The set of all k-forms for a vector space V spans a space called the kth exterior power of V, denoted $\Lambda^k(V)$. Note in particular that $\Lambda^1(V) = V$ and that $\Lambda^k(V)$ is a subspace of $\Lambda(V)$ for all k.

Consider the real vector space $\mathbb{R}^n$ and suppose that a set of vectors $\{\xi_i\}, i = 1, \ldots, k$ span a k-dimensional subspace of $\mathbb{R}^n$. The wedge product $\xi_1 \wedge \ldots \wedge \xi_k$ is a k-form representing this k-dimensional subspace. The set of all such k-forms spans the kth exterior power of $\mathbb{R}^n$, $\Lambda^k(\mathbb{R}^n)$, which is a vector space with $\dim(\Lambda^k(\mathbb{R}^n)) = {}^nC_k$. Every real $n \times n$ matrix M has a corresponding real ${}^nC_k \times {}^nC_k$ additive compound matrix $M^{[k]}$, which is a linear transform on $\Lambda^k(\mathbb{R}^n)$. $M^{[k]}$ is formally defined by the relation

$$M^{[k]}(\xi_1 \wedge \ldots \wedge \xi_k) = \sum_{i=1}^{k} \xi_1 \wedge \ldots \wedge M\xi_i \wedge \ldots \wedge \xi_k \tag{3.8}$$

Note that $M^{[1]} = M$ and $M^{[n]} = \mathrm{Tr}(M)$.

Such an additive compound matrix can be constructed as follows: let $\{\hat{e}_j\}, j = 1, \ldots, n$ be an orthonormal basis set for $\mathbb{R}^n$. Define a set of vectors $\{\hat{e}_{i_1} \wedge \ldots \wedge \hat{e}_{i_k} \,|\, i_1, \ldots, i_k \in \{1, \ldots, n\}\}$. The lexicographically ordered set of distinct, non-zero vectors from this set form an orthonormal basis set for $\Lambda^k(\mathbb{R}^n)$, which will be labelled $\{\hat{\omega}_i\}$ for $i = 1, \ldots, {}^nC_k$. As noted in [Flanders, 1963] and [Allen and Bridges, 2002], given an inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^n$ it is possible to construct a corresponding inner product $\langle \cdot, \cdot \rangle_k$ on $\Lambda^k(\mathbb{R}^n)$ as follows: Define $u, v \in \Lambda^k(\mathbb{R}^n)$ to be $u = u_1 \wedge \ldots \wedge u_k$ and $v = v_1 \wedge \ldots \wedge v_k$ for $u_i, v_i \in \mathbb{R}^n$. The inner product on the exterior power $\Lambda^k(\mathbb{R}^n)$ is

$$\langle u, v \rangle_k = \begin{vmatrix} \langle u_1, v_1 \rangle & \cdots & \langle u_1, v_k \rangle \\ \vdots & \ddots & \vdots \\ \langle u_k, v_1 \rangle & \cdots & \langle u_k, v_k \rangle \end{vmatrix} \tag{3.9}$$

Noting that $M\hat{\omega}_j = \sum_i \hat{e}_{j_1} \wedge \ldots \wedge M\hat{e}_{j_i} \wedge \ldots \wedge \hat{e}_{j_k}$ for $\hat{\omega}_j = \hat{e}_{j_1} \wedge \ldots \wedge \hat{e}_{j_k}$, the elements of the kth additive compound of M are then

$$M^{[k]}_{ij} = \langle \hat{\omega}_i, M\hat{\omega}_j \rangle \text{ with } i, j \in \{1, \ldots, {}^nC_k\} \tag{3.10}$$

For the purposes of this thesis, the focus of interest is on the second additive compound. Suppose that $\hat{\omega}_i = \hat{e}_{i_1} \wedge \hat{e}_{i_2}$ and $\hat{\omega}_j = \hat{e}_{j_1} \wedge \hat{e}_{j_2}$. Then

$$M^{[2]}_{ij} = \langle \hat{\omega}_i, M\hat{\omega}_j \rangle = M_{i_1 j_1}\delta_{i_2 j_2} + M_{i_2 j_2}\delta_{i_1 j_1} - M_{i_1 j_2}\delta_{i_2 j_1} - M_{i_2 j_1}\delta_{i_1 j_2} \tag{3.11}$$

where δ_{ij} is the Kronecker delta. Notice that since (assuming the lexicographic ordering) $i_1 < i_2$ and $j_1 < j_2$, it is possible that $i_1 = j_2$ **or** $i_2 = j_1$, but not that $i_1 = j_2$ **and** $i_2 = j_1$. This structure means that, as stated in the appendix of [Li and Muldowney, 1995],

$$M^{[2]}_{ij} = \begin{array}{ll} M_{i_1 i_1} + M_{i_2 i_2}, & \text{if } i = j \text{ (i.e. } i_1 = j_1, i_2 = j_2). \\ (-1)^{r+s} M_{i_r j_s}, & \text{if } i_r \neq j_s, i_s = j_r \text{ (n.b. } r, s \in \{1, 2\}). \\ 0, & \text{in all other cases.} \end{array} \tag{3.12}$$

As noted in [Li and Wang, 1998], the eigenvalues of $M^{[k]}$ are simply the sums of k eigenvalues of M, i.e. if $\sigma(M) = \{\lambda_i\}, i = 1, \ldots, n$ then $\sigma(M^{[k]}) = \{\lambda_{i_1} + \ldots + \lambda_{i_k}\}, 1 \leq i_1 < \ldots < i_k \leq n$. This observation leads to the following necessary and sufficient condition for Hurwitz stability of a matrix:

Theorem 22 (Li and Wang)**.** *Let M be a real square matrix. $\alpha(M) < 0$ if and only if $\alpha(M^{[2]}) < 0$ and $(-1)^n|M| > 0$.*

Proof. See [Li and Wang, 1998], theorem 3.1. □

3.2.2 Structure of multiplicative compound matrices

The multiplicative compound matrix is also an ${}^nC_k \times {}^nC_k$ real matrix, and satisfies

$$M^{(k)}(\xi_1 \wedge \ldots \wedge \xi_k) = (M\xi_1) \wedge \ldots \wedge (M\xi_k) \tag{3.13}$$

It follows that $M^{(1)} = M$ and $M^{(n)} = |M|$.

As with additive compound matrices, for the purposes of this thesis only the second multiplicative compound is of interest. Let $\{\hat{e}_j\}$ be the standard orthonormal basis set for $\mathbb{R}^n$, and let $\{\hat{\omega}_i\}$ be the corresponding lexicographically ordered orthonormal basis set for the second exterior power $\Lambda^2(\mathbb{R}^n)$, with $\hat{\omega}_i = \hat{e}_{i_1} \wedge \hat{e}_{i_2}$.

By definition, given a pair of vectors $x, y \in \mathbb{R}^n$, $M^{(2)}(x \wedge y) = Mx \wedge My$. Therefore $M^{(2)}\hat{\omega}_i = M^{(2)}(\hat{e}_{i_1} \wedge \hat{e}_{i_2}) = M\hat{e}_{i_1} \wedge M\hat{e}_{i_2}$. Writing out the RHS of this expression in full gives

$$M\hat{e}_{i_1} \wedge M\hat{e}_{i_2} = \left(\sum_{h_1=1}^{n} M_{h_1,i_1}\hat{e}_{h_1} \right) \wedge \left(\sum_{h_2=1}^{n} M_{h_2,i_2}\hat{e}_{h_2} \right)$$

Since the elements M_{ij} are just numbers, and $\hat{e}_{h_1} \wedge \hat{e}_{h_2} = 0$ when $h_1 = h_2$, this expression can be rewritten as

$$M\hat{e}_{i_1} \wedge M\hat{e}_{i_2} = \left(\sum_{h_1=1}^{n-1} \sum_{h_2=h_1+1}^{n} M_{h_1,i_1}M_{h_2,i_2}(\hat{e}_{h_1} \wedge \hat{e}_{h_2}) \right) + \left(\sum_{h_1=2}^{n} \sum_{h_2=1}^{h_1-1} M_{h_1,i_1}M_{h_2,i_2}(\hat{e}_{h_1} \wedge \hat{e}_{h_2}) \right)$$

By switching the h_1 and h_2 labels in the second bracket, noting that $\hat{e}_{h_2} \wedge \hat{e}_{h_1} = -\hat{e}_{h_1} \wedge \hat{e}_{h_2}$ and writing the limits for the second bracket differently, this can be rewritten

$$M\hat{e}_{i_1} \wedge M\hat{e}_{i_2} = \sum_{h_1=1}^{n-1} \sum_{h_2=h_1+1}^{n} (M_{h_1,i_1}M_{h_2,i_2} - M_{h_2,i_1}M_{h_1,i_2})(\hat{e}_{h_1} \wedge \hat{e}_{h_2})$$

This corresponds to the ${}^nC_2$vector

$$M^{(2)}\hat{\omega}_i = M\hat{e}_{i_1} \wedge M\hat{e}_{i_2} = (M_{1,i_1}M_{2,i_2} - M_{2,i_1}M_{1,i_2}, \ldots, M_{n-1,i_1}M_{n,i_2} - M_{n,i_1}M_{n-1,i_2})^T$$

Since by definition

$$M^{(2)}\hat{\omega}_i = (M^{(2)}_{1,i}, \ldots, M^{(2)}_{{}^nC_2,i})^T$$

it follows that $M^{(2)}_{h,i} = M_{h_1,i_1}M_{h_2,i_2} - M_{h_2,i_1}M_{h_1,i_2}$.

As stated in [Muldowney, 1990], for the kth multiplicative compound this generalises to

$$M^{(k)}_{ij} = \left| M^{j_1,\ldots,j_k}_{i_1,\ldots,i_k} \right| \tag{3.14}$$

Here $M^{j_1,\ldots,j_k}_{i_1,\ldots,i_k}$ is the submatrix of M determined by the rows with indices $i_1,\ldots,i_k$ and columns with indices $j_1,\ldots,j_k$, where $1 \leq i_1 < \ldots < i_k \leq n$ and $1 \leq j_1 < \ldots < j_k \leq n$ are the sets of lexicographically ordered indices of M that correspond to i and j.

The autonomous convergence theorem that follows only requires the second additive compound, but the second multiplicative compound is useful when considering coordinate transforms on $\mathbb{R}^n$. Just as a linear coordinate transform on $\mathbb{R}^n$ induces a similarity transform on all other linear transforms on $\mathbb{R}^n$, it also induces a transform on the associated exterior powers of $\mathbb{R}^n$. Let T be a linear coordinate transform on $\mathbb{R}^n$ mapping $x \rightarrow Tx$ for all $x \in \mathbb{R}^n$. Let $\tilde{T}$ be the transform on $\Lambda^2(\mathbb{R}^n)$ induced by T, so for every $x \wedge y \in \Lambda^2(\mathbb{R}^n)$, $\tilde{T}$ maps $x \wedge y \rightarrow \tilde{T}(x \wedge y) = Tx \wedge Ty$. This is the definition of the second multiplicative compound of a matrix, and hence $\tilde{T} = T^{(2)}$.

3.2.3 Statement of theorem and interpretation

Now that the necessary preliminaries are in place, the statement of the second autonomous convergence theorem used in this thesis follows. For a more detailed outline of the theorem, including variants and generalisations, see any of the papers by Li and Muldowney listed at the beginning of this section (page 49).

Theorem 23 (Li and Muldowney)**.** *Let D be a simply connected open set in $\mathbb{R}^n$, and let $f : D \rightarrow \mathbb{R}^n : x \mapsto f(x)$ be a C^1 function. Suppose that the following conditions hold:*

1. *f has a unique zero $x_0 \in D$.*
2. *Solutions to f exist for all $t \geq 0$.*
3. *There exists a compact set $D_0 \subseteq D$ and for each bounded $D_1 \subseteq D$, $\Phi(t, D_1) \subset D_0$ for all large enough t.*
4. *The Jacobian $J(x)$ satisfies $\mu_m(J(x)^{[2]}) < 0$ for some fixed logarithmic norm μ_m at every point $x \in D$.*

Then x_0 is globally asymptotically stable.

Proof. The proof follows directly from corollary 2.6 of [Li and Muldowney, 1995]. □

Condition 3 essentially means that all trajectories enter a compact invariant set.

This theorem works in an analogous way to theorem 20 (p. 46). Notice that, unlike the Markus–Yamabe theorem (p. 42) and theorem 20, this result does not require that $|J| \neq 0$. Whereas theorem 20 relates directly to the distance between trajectories (1-forms in phase space), theorem 23 is concerned with the distance between 2-forms, which represent area elements. In this case, the distance between 2-forms is measured in the second exterior power of the vector space according to the norm used to generate μ_m. The existence of a negative logarithmic norm on the second exterior power of the vector space implies that areas are locally shrinking with time under the action of the flow defined by f. It is important to note that $\mu_m(J^{[2]}) < 0$ is an **open** condition, in other words the same inequality holds for every matrix that lies in some neighbourhood of $J^{[2]}$ in the space of ${}^nC_2 \times {}^nC_2$ real matrices. [Smith, 1986] showed that the condition $\mu_2(J^{[2]}) < 0$ on a region of phase space is strong enough to rule out the existence of periodic orbits. Smith also showed by means of Pugh's closing lemma that this condition is strong enough to rule out the existence of any other non-wandering points other than fixed points, since for each point that is non-wandering under the flow defined by f, a function g can be found arbitrarily close to f in the C^1 topology under the flow of which the non-wandering point is periodic; the open condition $\mu_2(J^{[2]}) < 0$ is strong enough to rule out the existence of periodic orbits for the flow of any g sufficiently close to f. Consequently, the ω-limit set within the specified region of phase space of any point is either empty or consists of a fixed point. As noted earlier, Smith did not present his result in terms of logarithmic norms, but rather as a spectral condition on the second additive compound of J which turned out to be equivalent to requiring $\mu_2(J^{[2]}) < 0$; the generalisation of the result to any logarithmic norm appeared in [Li and Muldowney, 1995].

A related autonomous convergence theorem has been constructed by Li and Muldowney for dynamical systems containing an embedded invariant manifold, using higher order additive compounds. See [Li and Muldowney, 2000] for more information.

As stated in lemma 8 (p. 45), the existence of an invertible ${}^nC_2 \times {}^nC_2$ matrix $\tilde{T}$ such that $\mu_m(\tilde{T}J^{[2]}\tilde{T}^{-1}) < 0$ for some logarithmic norm μ_m implies the existence of another logarithmic norm $\mu_{m,\tilde{T}}$ such that $\mu_{m,\tilde{T}}(J^{[2]}) < 0$. Although there are more possible transforms in the second exterior power of the vector space than in the vector space itself (since for $n > 3$, ${}^nC_2 > n$), in applications it is sometimes easier to find a transform in phase space prior to constructing the second additive compound in order to demonstrate the existence of a negative logarithmic norm for the second exterior power of the vector space. The following simple corollary will be referred to later:

Corollary 3. *Suppose there exists an invertible real $n \times n$ matrix T and a real $n \times n$ matrix M such that $\mu_m((TMT^{-1})^{[2]}) < 0$ for some logarithmic norm μ_m. Then there exists another logarithmic norm $\mu_{m,\tilde{T}}$ such that $\mu_{m,\tilde{T}}(M^{[2]}) < 0$.*

Proof. As noted at the end of §3.2.2, T induces a transform on $\Lambda^2(\mathbb{R}^n)$, which is $T^{(2)}$, i.e. $T^{(2)}M^{[2]}{T^{(2)}}^{-1} = (TMT^{-1})^{[2]}$. Since by assumption $\mu_m(T^{(2)}M^{[2]}{T^{(2)}}^{-1}) < 0$, the existence of a logarithmic norm $\mu_{m,T^{(2)}}$ such that $\mu_{m,T^{(2)}}(M^{[2]}) < 0$ follows from lemma 8. □

3.2.4 Links between the two autonomous convergence theorems

The remainder of this chapter is related to some small results relating the two autonomous convergence theorems. The first result follows directly from the structure of the second additive compound, as described in equation (3.12) on p. 51.

Lemma 11. *Let M be a real square matrix, and let $M^{[2]}$ be its second additive compound. If $\mu_1(M) < 0$ then $\mu_1(M^{[2]}) < 0$, and if $\mu_\infty(M) < 0$ then $\mu_\infty(M^{[2]}) < 0$.*

Proof. Recall the definitions of μ_1 and μ_∞ from equations (3.5) and (3.6) (p. 45):

$$\mu_1(M) = \max_j \left(\operatorname{Re}(M_{jj}) + \sum_{i,i\neq j} |M_{ij}| \right)$$

$$\mu_\infty(M) = \max_i \left(\operatorname{Re}(M_{ii}) + \sum_{j,j\neq i} |M_{ij}| \right)$$

Since $\mu_1(M) < 0$, for every $j_1 = 1, \ldots, n$, $M_{j_1 j_1} < 0$ and $|M_{j_1 j_1}| > \sum_{i_1, i_1 \neq j_1} |M_{i_1 j_1}|$. Likewise, for every $j_2 = 1, \ldots, n$, $M_{j_2 j_2} < 0$ and $|M_{j_2 j_2}| > \sum_{i_2, i_2 \neq j_2} |M_{i_2 j_2}|$.

From equation (3.12), $M^{[2]}_{jj} = M_{j_1 j_1} + M_{j_2 j_2}$, so $M^{[2]}_{jj} < 0$ and $|M^{[2]}_{jj}| = |M_{j_1 j_1}| + |M_{j_2 j_2}|$. Using the same equation,

$$\sum_{i,i\neq j} |M^{[2]}_{ij}| = \sum_{i_1, i_1 \neq j_1} \sum_{i_2, i_2 > i_1} \delta_{i_2 j_2} |M_{i_1 j_1}| + \sum_{i_2, i_2 \neq j_2} \sum_{i_1, i_1 < i_2} \delta_{i_1 j_1} |M_{i_2 j_2}|$$

Since

$$\sum_{i_1, i_1 \neq j_1} \sum_{i_2, i_2 > i_1} \delta_{i_2 j_2} |M_{i_1 j_1}| + \sum_{i_2, i_2 \neq j_2} \sum_{i_1, i_1 < i_2} \delta_{i_1 j_1} |M_{i_2 j_2}| \leq \sum_{i_1, i_1 \neq j_1} |M_{i_1 j_1}| + \sum_{i_2, i_2 \neq j_2} |M_{i_2 j_2}|$$

it follows that $|M^{[2]}_{jj}| > \sum_{i,i\neq j} |M^{[2]}_{ij}|$. This concludes the proof for μ_1. The proof for μ_∞ is very similar, so the argument is not repeated here. □

Lemma 11 is not of any great practical significance, but may be of interest in the name of completeness, since intuitively if the length of every 1-form is decreasing with time, it is to be expected that every 2-form will also be decreasing. The result can be extended to other norms, as shown below. Recall that $\mu(M) \geq \alpha(M)$ for any logarithmic norm μ, and so $\mu(M) < 0 \Rightarrow \alpha(M) < 0$. Since the eigenvalues of $M^{[2]}$ are simply sums of pairs of eigenvalues of M, it follows that $\alpha(M) < 0 \Rightarrow \alpha(M^{[2]}) < 0$. It is also known that $\alpha(M) = \inf \mu(M)$, which means that $\alpha(M^{[2]}) < 0 \Rightarrow \mu^*(M^{[2]}) < 0$ for some μ^*. The question is then whether for some matrix M and some given logarithmic norm μ,

$\mu(M) < 0 \Rightarrow \mu(M^{[2]}) < 0$, assuming that it makes sense to define μ on the exterior power $\Lambda^2(\mathbb{R}^n)$. In the case that the norm on $\Lambda^2(\mathbb{R}^n)$ used to construct μ satisfies $|x \wedge y| \leq |x||y|$ for each $x, y \in \mathbb{R}^n$ with corresponding $x \wedge y \in \Lambda^2(\mathbb{R}^n)$, this can be answered in the affirmative:

Theorem 24. *Suppose that $|\cdot|_l$ is a vector norm defined on $\mathbb{R}^n$, $\|\cdot\|_l$ is its induced matrix norm and $\mu_l(\cdot)$ is the corresponding logarithmic norm. Assume that $|\cdot|_l$ admits an extension $|\cdot|_m$ to the exterior power $\Lambda^2(\mathbb{R}^n)$ satisfying $|x \wedge y|_m \leq |x|_l|y|_l$ for all $x, y \in \mathbb{R}^n$. Let M be an $n \times n$ real matrix. Then $\mu_l(M) < 0 \Rightarrow \mu_m(M^{[2]}) < 0$.*

Proof. Going back to definition of the logarithmic norm, the statement requiring proof is

$$\lim_{h \to 0^+} \frac{\|I + hM^{[2]}\|_m - 1}{h} < 0 \tag{3.15}$$

Since it is known that this limit exists and is monotone, it suffices to show that there exists some $h_1 > 0$ such that for all $h < h_1$, $\|I + hM^{[2]}\|_m < 1$. Using the definition of the second additive compound and the matrix norm, this in turn implies that that there exists some $h_1 > 0$ such that for all $h < h_1$,

$$\sup_{|x \wedge y|_m = 1} |(x \wedge y) + h(Mx \wedge y) + h(x \wedge My)|_m < 1$$

Since x and y lie in the finite dimensional space $\mathbb{R}^n$, the supremum in this expression is over a compact set and can be replaced by a maximum. It therefore suffices to show that for every x, y satisfying $|x \wedge y|_m = 1$, there exists some $h_1(x, y) > 0$ such that for all $h < h_1$

$$|(x \wedge y) + h(Mx \wedge y) + h(x \wedge My)|_m < 1 \tag{3.16}$$

By assumption, $\mu_l(M) < 0$, and so for every x satisfying $|x|_l = 1$, there exists some $h_0(x) > 0$ such that for all $h < h_0$, $|x + hMx|_l < 1$. Thus for any pair of vectors x, y satisfying $|x|_l = |y|_l = 1$ there also exists $h_0(x, y) = \min(h_0(x), h_0(y))$ such that for all $h < h_0(x, y)$,

$$|x + hMx|_l |y + hMy|_l < 1 \tag{3.17}$$

It was also assumed that $|x \wedge y|_m \leq |x|_l|y|_l$, and consequently $|(x + hMx) \wedge (y + hMy)|_m \leq |x + hMx|_l|y + hMy|_l < 1$. Expanding the wedge product gives the relation

$$|(x \wedge y) + h(Mx \wedge y) + h(x \wedge My) + h^2(Mx \wedge My)|_m < 1 \tag{3.18}$$

So, to prove the theorem, it suffices to show that if there exists some $h_0(x, y) > 0$ such that for all $h < h_0(x, y)$ relation (3.18) is satisfied, then there also exists some $h_1 > 0$ such that for all $h < h_1$ relation (3.16) is satisfied. To demonstrate that this claim is true, let $(x \wedge y) + h(Mx \wedge y) + h(x \wedge My) + h^2(Mx \wedge My) = A(h)$ and let $(x \wedge y) + h(Mx \wedge$

$y) + h(x \wedge My) = B(h)$. Denote the difference between the norms of these two expression as $|A(h)|_m - |B(h)|_m = g(h)$. By a well-known vector inequality, $|A(h) - B(h)|_m \geq ||A(h)|_m - |B(h)|_m|$. Therefore $|g(h)| \leq h^2 |Mx \wedge My|_m$.

Clearly $g(h) \to 0$ as $h \to 0$. For all $h < h_0(x, y)$, $|A(h)|_m$ is bounded above by some $1 - \epsilon$, with $\epsilon > 0$ and $\epsilon \uparrow$ as $h \to 0$. Therefore there exists $h_1 < h_0(x, y)$ such that $g(h) < \epsilon$ for all $h < h_1$. Consequently $|B(h)|_m < 1$, as required. □

The condition $|x \wedge y|_m \leq |x|_l |y|_l$ is not satisfied for general norms, but it does hold for the case when both norms are the 2-norm, i.e. $l = m = 2$. The result relies on Lagrange's identity (stated on page 1049 of [Gradshteyn and Ryzhik, 2000] or page 41 of [Mitrinović, 1970], among many other places), which is

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |x_i y_j - x_j y_i|^2 = \left(\sum_{i=1}^{n} |x_i|^2 \right) \left(\sum_{j=1}^{n} |y_j|^2 \right) - \left| \sum_{i=1}^{n} x_i y_i \right|^2 \tag{3.19}$$

The result is as follows:

Lemma 12. $|x \wedge y|_2 \leq |x|_2 |y|_2$ *for all* $x, y \in \mathbb{R}^n$.

Proof. The 2-norm is defined as

$$|x|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$$

Let $\hat{e}_i$ be the ith vector in an orthonormal basis for $\mathbb{R}^n$. Given a pair of vectors $x = \sum_{i=1}^{n} x_i \hat{e}_i$ and $y = \sum_{j=1}^{n} y_j \hat{e}_j$, the corresponding vector in $\Lambda^2(\mathbb{R}^n)$ is

$$x \wedge y = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_j (\hat{e}_i \wedge \hat{e}_j)$$

Let (i, j) be a lexicographically ordered pair representing the index of a component of $x \wedge y$, i.e. the first component of $x \wedge y$ is indexed by $(1, 2)$, the second component is indexed by $(1, 3)$, and so on up to the final component, which is indexed by $(n - 1, n)$. The (i, j)th component of $x \wedge y$ is therefore $(x \wedge y)_{(i,j)} = (x_i y_j - x_j y_i) \hat{\omega}_{(i,j)}$, where $\hat{\omega}_{(i,j)} = \hat{e}_i \wedge \hat{e}_j$ is the (i, j)th vector in the induced orthonormal basis of $\Lambda^2(\mathbb{R}^n)$. The 2-norm of $x \wedge y$ is:

$$|x \wedge y|_2 = \left(\sum_{(i,j)=1}^{{}^nC_2} |(x \wedge y)_{(i,j)}|^2 \right)^{1/2}$$

This is equal to

$$|x \wedge y|_2 = \left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |x_i y_j - x_j y_i|^2 \right)^{1/2}$$

The relation $|x \wedge y|_2 \leq |x|_2 |y|_2$ can therefore be expanded to give

$$\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |x_i y_j - x_j y_i|^2 \right)^{1/2} \leq \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \left(\sum_{j=1}^{n} |y_j|^2 \right)^{1/2}$$

The power of 1/2 can be removed by squaring both sides, and thus the result is true if and only if

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |x_i y_j - x_j y_i|^2 \leq \left(\sum_{i=1}^{n} |x_i|^2 \right) \left(\sum_{j=1}^{n} |y_j|^2 \right) \tag{3.20}$$

Comparing inequality (3.20) with Lagrange's identity reveals that inequality (3.20) is true. This completes the proof. □

Lemma 12 is in fact a corollary of a more general result that arises from the relationship between norms and inner products. The relation $|x \wedge y|_m \leq |x|_l |y|_l$ can be fulfilled by generating a norm on $\Lambda^2(\mathbb{R}^n)$ from an inner product on $\mathbb{R}^n$. Let $\alpha = \alpha_1 \wedge \alpha_2$ and $\beta = \beta_1 \wedge \beta_2$. As per equation (3.9), the induced inner product $\langle \alpha, \beta \rangle$ is then

$$\langle \alpha, \beta \rangle = \begin{vmatrix} \langle \alpha_1, \beta_1 \rangle & \langle \alpha_1, \beta_2 \rangle \\ \langle \alpha_2, \beta_1 \rangle & \langle \alpha_2, \beta_2 \rangle \end{vmatrix} \tag{3.21}$$

This leads to the following result:

Lemma 13. *Let $\langle \cdot, \cdot \rangle$ be an inner product on $\mathbb{R}^n$, and let $|\cdot|_m$ be a norm on $\mathbb{R}^n$ defined by $|x|_m = \langle x, x \rangle^{1/2}$. The corresponding induced norm $|\cdot|_m$ on $\Lambda^2(\mathbb{R}^n)$ satisfies $|x \wedge y|_m \leq |x|_m |y|_m$ for all $x, y \in \mathbb{R}^n$.*

Proof. Choose any $x, y \in \mathbb{R}^n$. By definition,

$$|x \wedge y|_m = \begin{vmatrix} \langle x, x \rangle & \langle x, y \rangle \\ \langle y, x \rangle & \langle y, y \rangle \end{vmatrix}^{\frac{1}{2}}$$

Expanding the determinant gives $|x \wedge y|_m = (\langle x, x \rangle \langle y, y \rangle - \langle x, y \rangle^2)^{1/2}$. By contrast, $|x|_m |y|_m = \langle x, x \rangle^{1/2} \langle y, y \rangle^{1/2} = (\langle x, x \rangle \langle y, y \rangle)^{1/2}$. As $\langle x, y \rangle^2 \geq 0$, it follows that $|x \wedge y|_m \leq |x|_m |y|_m$. □

3.2.5 Discussion

Finding a suitable logarithmic norm in order to demonstrate that a given dynamical system displays autonomous convergence is a difficult problem in general. The forms of logarithmic norms are often difficult to construct explicitly; this is compounded by the fact that even when given a matrix M and the explicit form of a logarithmic norm μ_m, it is not always straightforward to check whether $\mu_m(M) < 0$. The difficulty is alleviated slightly by lemma 8, but further results simplifying the process of identifying useful logarithmic norms would be of great benefit in the application of autonomous convergence theorems.

Generally speaking, it is easier to work directly with the Jacobian matrix than with its second additive compound, and for this reason theorem 20 (p. 46) is the more straightforward of the two autonomous convergence theorems to apply in most cases. However, since the converse of theorem 24 (p. 56) is not true, i.e. $\mu(M^{[2]}) < 0 \not\Rightarrow \mu(M) < 0$, theorem 23 (p. 53) gives more general conditions on the Jacobian for global convergence of an autonomous system.

One final thing to note is that, by theorem 4.1 of [Muldowney, 1990], if for a continuous time dynamical system defined by $\dot{x} = f(x)$, there exists a logarithmic norm μ such that $\mu((Df(x))^{[2]}) < 0$ for all x in a region of phase space, that region of phase space contains no nontrivial periodic orbits. This result may be of interest as an alternative to the theory of monotone flows to rule out periodic behaviour in dynamical systems which have more than one fixed point. Since the applications that follow are focused on systems with only one fixed point, this line of investigation is not pursued further here, but may be worth looking into in the future.

Chapter 4

The mitochondrial electron transport chain

The generic model of electron transport processes presented in this chapter was originally motivated by work on a more complicated numerical model of the physiology of bloodflow to the brain, described in [Banaji et al., 2005]. A number of models of electron transport have been published previously, such as [Korzeniewski, 1996], [Korzeniewski and Zoladz, 2001], [Farmery and Whiteley, 2001] and [Beard, 2005]. However, these models were constructed with numerical data in mind, relying on complicated functional forms. The model presented here, in keeping with the rest of the thesis, is based as far as possible on qualitative assumptions; therefore the conclusions drawn are valid for a whole class of numerical models that have the same underlying structure.

The results developed in this chapter originally appeared (with less background detail) in [Donnell et al., 2008], which in turn answered open questions from [Banaji, 2006]. Banaji's paper presented a qualitative model of the electron transport chain, in which the chain was represented as a series of coupled redox reactions. Some of the redox reactions pump protons across an electrostatic gradient (details in the next section); for simplicity it was assumed that this electrostatic gradient be constant. In this context, it was demonstrated that the system has a unique globally asymptotically stable fixed point. However, in reality the electrostatic gradient varies as a direct effect of the protons pumped by the redox reactions, and [Banaji, 2006] went on introduce an extended version of the model including this feedback process. It was demonstrated that the extended model has a unique fixed point, but no further analysis of the model's behaviour was carried out.

[Donnell et al., 2008] picked up where [Banaji, 2006] left off, analysing the possible asymptotic behaviour of the system with a variable electrostatic gradient. It was anticipated that since the system with a fixed gradient is globally asymptotically stable and the variation of the electrostatic gradient is a negative feedback process (i.e. the redox reactions increase the gradient, but increasing the gradient inhibits their rates of reaction), that this extended system would also be globally asymptotically stable. As demonstrated in this chapter, it turned out that the extended system is not always globally asymptotically stable, which was an unexpected and new result.

The work in [Donnell et al., 2008] reproduced in this chapter comprises a series of results for increasing chain lengths. The first new result is a proof that for very short chains of two redox reactions coupled to a variable electrostatic gradient, the unique fixed point of the system is globally asymptotically stable. Since in this case the biological system can be represented by a two dimensional dynamical system, the proof is fairly straightforward, relying on well known theory presented in chapter 1. The chapter then goes on to present a more advanced proof that for longer chains of three redox reactions coupled to a varying gradient, the fixed point remains globally asymptotically stable. The proof uses some of the autonomous convergence results presented in chapter 3, which have not been applied to this area of biology before, and relies on making some extra (physically reasonable) assumptions about the relationship between the redox reaction rates and the number of protons pumped. Finally, it is shown that for chains of four or more redox reactions coupled to a gradient the fixed point is no longer guaranteed to be locally asymptotically stable, even with the extra assumptions made that guarantee global asymptotic stability in the three reaction case. There then follows some discussion of the results, including what they might imply in biological terms, and areas for further investigation.

The chapter proper begins with a brief description of the electron transport chain and its biological role, followed by construction of a model representing what are believed to be the key aspects of the electron transport chain as per [Banaji, 2006].

4.1 Biological description of the electron transport chain

An electron transport chain is a series of coupled oxidation-reduction (redox) reactions that produce energy for cellular respiration. Electron transport occurs in various systems in both prokaryotic cells (organisms without a cell nucleus) and eukaryotic cells. The focus in this chapter is on eukaryotic electron transport processes that take place in the mitochondria. The aim of the description given here is simply to provide enough background information to explain the structure of a qualitative model of mitochondrial electron transport, without going into detail. For a detailed description of electron transport, see a biochemistry textbook such as [Garrett and Grisham, 1995] or [Bhagavan, 2002].

Mitochondria are small ($\sim$1μm) organelles that are present in most eukaryotic cells. They consist of an outer membrane, the main function of which is thought to be purely structural, and an inner membrane, in and on which electron transport processes take place. Between the inner and outer membranes is a region known as the inter-membrane space, and the region enclosed within the inner membrane is referred to as the matrix. The inner membrane is a dynamic structure, containing many folds in order to maximise its surface area. A diagram appears in figure 4.1[1].

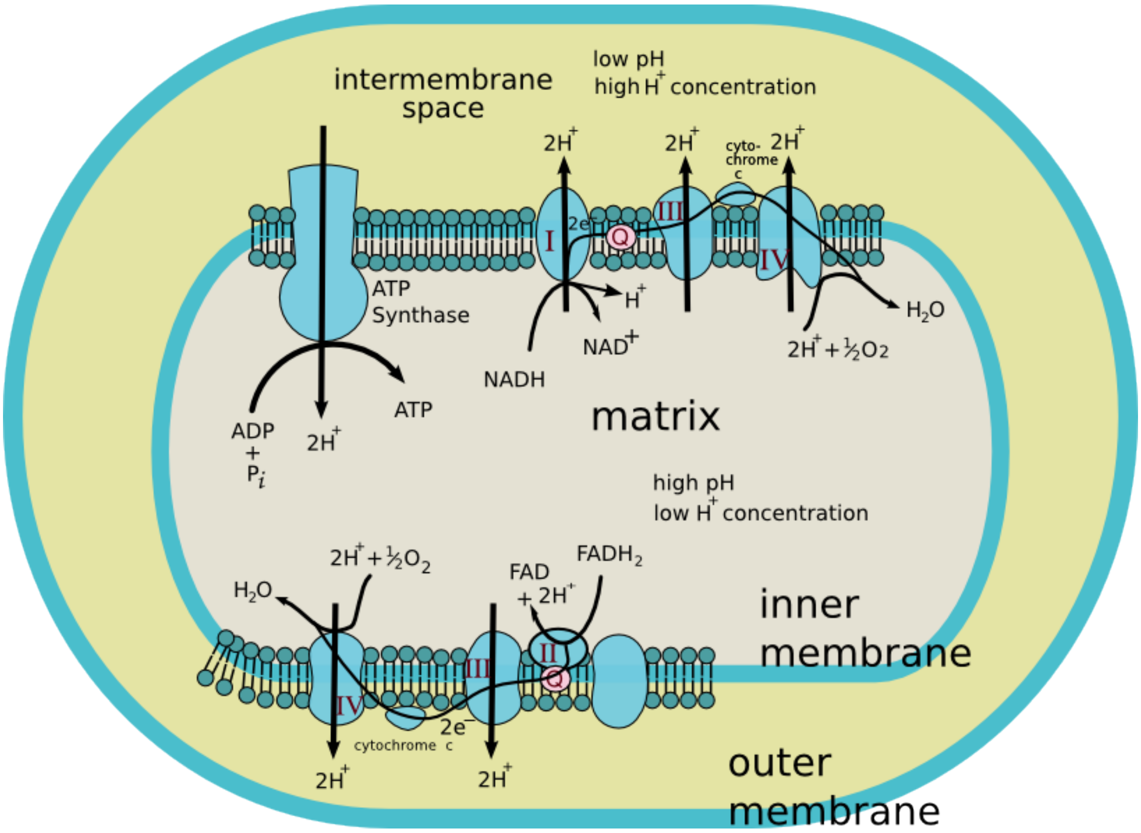

Figure 4.1: A diagrammatic representation of a mitochondrion, including the inner and outer membranes, and the complexes embedded in the inner membrane that relate to electron transport processes.

The reactions of the electron transport chain are centred around complexes I-IV, which are part of the inner membrane. The transport chain begins with the reduced form of nicotinamide adenine dinucleotide (NADH), which is oxidised in a reaction involving complex I. The electrons released by this reaction are then transferred in a series of steps between and within the various complexes, and are accepted by oxygen in the final reaction involving complex IV. It is perhaps worth pointing out that while the reactions are sequential, the complexes are not thought to be spatially related to one another. The energy released by several of the reactions in the electron transport chain is used to pump protons out

[1] Image source: Wikipedia, accessed on January 28th, 2008. The original file can be found online at `http://en.wikipedia.org/wiki/Image:Etc2.svg` and was released under the Creative Commons Attribution-ShareAlike 2.5 license by its creator, Rozzychan. The modified version that appears in this thesis is available online at `http://www.medphys.ucl.ac.uk/∼pdonnell/electron_transport_chain_diagram.svg` and may be used under the same license conditions. For details of the Creative Commons Attribution-ShareAlike 2.5 license see `http://creativecommons.org/licenses/by-sa/2.5/`.

of the matrix and into the inter-membrane space via an active transport process. In this way, an electrochemical gradient is built up by the electron transport chain. Some of the protons then leak back across the inner membrane, and others pass through ATP-synthase (also known as complex V) back into the matrix. The energy released by protons passing through is used by complex V to phosphorylate adenosine diphosphate (ADP) into adenosine 5'-triphosphate (ATP). The energy stored in the phosphate group bonds can then be used by many processes in cells, such as muscle contraction.

It is worth noting that while the general processes of the mitochondrial electron transport are fairly well understood, many of the mechanisms of each step are not yet known in detail [Belevich et al., 2006].

Generic models of electron transport chains were explored in [Banaji, 2006], where the main emphasis was on the input-output response of such models. In the simplest case, where the proton gradient across the membrane was ignored, these models were found to have very simple behaviour – at all physically meaningful parameter values there was a single, globally stable, equilibrium. In [Banaji and Baigent, 2008], this result was extended to the case of electron transfer networks with more general topology than a chain. However, in the more biologically realistic case where the build up of a proton gradient has an inhibitory effect on electron transport, analysis of the models is more difficult. This chapter analyses in more detail the behaviour in the case where the proton gradient is allowed to vary.

Before discussing generic models, it is worth mentioning that there are several detailed models of electron transport and oxidative phosphorylation such as [Korzeniewski, 1996], [Korzeniewski and Zoladz, 2001], [Farmery and Whiteley, 2001] and [Beard, 2005]. These ordinary differential equation models have been designed with numerical data in mind, and reflecting the complexity of the processes involved, the functional forms are quite involved. As mentioned in chapter 1, the approach here is quite different, and more akin to work in [Banaji, 2006], [Banaji and Baigent, 2008] and [De Leenheer et al., 2007]. The generic model constructed here could be instantiated in a great variety of numerical models, and the conclusions drawn are valid for all possible instances of the generic model.

4.1.1 The basic reaction scheme

The basic reaction scheme of interest here was described in some detail in [Banaji, 2006]; a briefer summary is given here. Assume that there are n substrates, each of which can exist in an oxidised state $\mathrm{A_i}$ and a reduced state $\mathrm{B_i}$ so that

$$\mathrm{A_i} + e^- \rightleftharpoons \mathrm{B_i}$$

Further, assume that protons can exist in two compartments – the mitochondrial matrix (where they are termed H_m^+), and the intermembrane space (where they are termed H_e^+) – with the possibility of transfers of the form

$$\mathrm{H}_m^+ \rightleftharpoons \mathrm{H}_e^+$$

The reactions of interest are in general the combination of three processes, a reduction, an oxidation, and the transport of some protons across the membrane. So for example, if substrate $\mathrm{A_i}$ is reduced to $\mathrm{B_i}$, $\mathrm{B_j}$ is oxidised to $\mathrm{A_j}$, and p protons are pumped across the mitochondrial membrane then the following half reactions are obtained:

$$\mathrm{A_i} + e^- \rightleftharpoons \mathrm{B_i}, \quad \mathrm{B_j} \rightleftharpoons \mathrm{A_j} + e^- \quad \text{and} \quad p\mathrm{H}_m^+ \rightleftharpoons p\mathrm{H}_e^+$$

These combine to give

$$\mathrm{A_i} + \mathrm{B_j} + p\mathrm{H}_m^+ \rightleftharpoons \mathrm{A_j} + \mathrm{B_i} + p\mathrm{H}_e^+$$

There is also the possibility that a reducing/oxidising agent may be external to the model, giving reactions such as

$$\mathrm{A_i} + p\mathrm{H}_m^+ \rightleftharpoons \mathrm{B_i} + p\mathrm{H}_e^+ \quad \text{or} \quad \mathrm{B_i} + p\mathrm{H}_m^+ \rightleftharpoons \mathrm{A_i} + p\mathrm{H}_e^+$$

A set of reactions of the kind just described can be combined into a network of reactions. A chain structure (as opposed to a more general network) derives from the assumption that each oxidised substrate accepts an electron from only one donor, and each reduced substrate transfers its electron to only one acceptor. This introduces a natural ordering on the substrates, so that for $i < n$, the ith substrate is able to donate electrons to the $(i+1)$th substrate, while for $i > 1$, the ith substrate is able to accept electrons from the $(i-1)$th substrate. The first substrate is able to accept electrons from outside the chain (reflecting the initial reduction of NADH), and the nth substrate is able to donate electrons to an acceptor outside the chain (reflecting the action of $\mathrm{O_2}$).

Thus there are $n+1$ redox reactions and the ith reaction has forward rate f_i. No assumptions are made about the sign of the functions f_i, potentially allowing reactions to be reversible. For $i \leq n$, the ith reaction involves the reduction of $\mathrm{A_i}$, and for $i \geq 2$, the ith reaction involves the oxidation of $\mathrm{B_{i-1}}$. The number of protons pumped across the mitochondrial membrane by the ith reaction is defined to be p_i. In general, the number of protons pumped by a reaction is constant, although under certain conditions it appears that the number of protons pumped by a reaction may decrease, which is referred to as "redox slip" [Brand et al., 1994]. However, redox slip does not appear to be very important in normal circumstances [Canton et al., 1995], and therefore the quantities p_i are assumed constant.

A quantity ψ can be defined so that transfer of a single proton across the membrane creates one unit of ψ. ψ can take any real value and is a strictly increasing function of the electrical/chemical gradient against which protons are pumped across the membrane, generally termed the proton motive force. Finally, reflecting the combined effect of proton leak and ADP phosphorylation, there is a process with rate L representing the “decay” of ψ.

The model is not “complete” in the sense that it does not include a representation of the conversion of ATP into ADP, nor does it take into account phosphate transport processes.

The structure of the model is illustrated in figure 4.2[2].

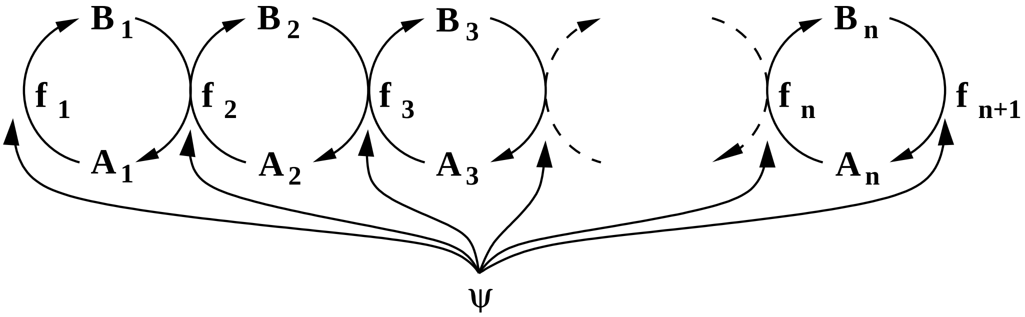

Figure 4.2: A schematic representation of the reaction network. The quantities $\mathrm{A_i}$ and $\mathrm{B_i}$ refer to oxidised and reduced states of the substrates. The functions f_i define the forward rates of reaction of the $n+1$ coupled redox reactions. The quantity ψ represents the electrical and chemical gradient across the mitochondrial membrane, which has an inhibitory effect on any redox reactions which involve proton pumping.

Since the total quantity (oxidised plus reduced) of any substrate in the chain is conserved, reduced forms of the substrates are not explicitly introduced. Instead, the concentration of $\mathrm{A_i}$ is referred to as x_i, and the total concentration of $\mathrm{A_i}+\mathrm{B_i}$ is assumed constant at m_i. This results in a model of the form:

$$\left.\begin{array}{rcl} \dot{x}_1 & = & -f_1(x_1,\psi)+f_2(x_1,x_2,\psi) \\ \dot{x}_i & = & -f_i(x_{i-1},x_i,\psi)+f_{i+1}(x_i,x_{i+1},\psi) \quad i=2,\ldots,n-1 \\ \dot{x}_n & = & -f_n(x_{n-1},x_n,\psi)+f_{n+1}(x_n,\psi) \\ \dot{\psi} & = & \sum\limits_{i=1}^{n+1} p_i f_i - L(\psi) \end{array}\right\} \tag{4.1}$$

The phase space of this system is defined by the equations:

$$\begin{array}{rcl} 0\leq & x_i & \leq m_i \quad i=1,\ldots,n \\ -\infty< & \psi & <\infty \end{array}$$

and is hence $n+1$ dimensional, being the product of a closed n-dimensional box and the real line.

[2] Image source: figure 3 in [Banaji, 2006].

4.1.2 Assumptions

In order to make the model as general as possible, minimal assumptions are made about the functions used. In particular, specific functional forms are not chosen.

All the functions f_i and L are assumed to be C^1 (once differentiable in all their arguments with continuous derivatives). The following notation is used for the derivatives of the functions:

$$f_{ij} \equiv \frac{\partial f_i}{\partial x_j}\,, \quad F_{ij} \equiv -f_{ij}\,, \quad f_{i\psi} \equiv \frac{\partial f_i}{\partial \psi}\,, \quad F_{i\psi} \equiv -f_{i\psi}\,, \quad L_\psi = \frac{\mathrm{d}L}{\mathrm{d}\psi} \tag{4.2}$$

At finite substrate concentrations, all reaction rates are finite, so that at any fixed ψ each f_i is bounded on its domain of definition.

When there is no gradient, no protons leak through the membrane. The rate at which protons either leak through the membrane or are used in ATP phosphorylation is assumed to be strictly increasing in ψ. Since ψ represents a gradient against which some of the reactions must do work, the following relations are obtained:

$$\left.\begin{array}{rcl} L(0)=0 & \text{and} & L_\psi > 0 \\ f_{i\psi} < 0 \text{ if } p_i \neq 0 & \text{and} & f_{i\psi} = 0 \text{ if } p_i = 0 \end{array}\right\} \tag{4.3}$$

If $p_i \neq 0$, then ψ inhibits the forward reaction and it is assumed that sufficiently large values of ψ make the reaction rate arbitrarily small or negative, i.e.

$$\begin{array}{ll} \lim\limits_{\psi\to\infty} f_i(\cdot,\psi) \leq 0 & \quad i = 1, n+1 \\ \lim\limits_{\psi\to\infty} f_i(\cdot,\cdot,\psi) \leq 0 & \quad i = 2,\ldots,n \end{array}$$

This reflects the fact that the energy required to pump a proton against a chemical and electrical gradient becomes large as the gradient increases. Similarly $-\psi$ inhibits the backward reaction so that:

$$\begin{array}{ll} \lim\limits_{\psi\to-\infty} f_i(\cdot,\psi) \geq 0 & \quad i = 1, n+1 \\ \lim\limits_{\psi\to-\infty} f_i(\cdot,\cdot,\psi) \geq 0 & \quad i = 2,\ldots,n \end{array}$$

The following equations imply that no reaction can proceed in the absence of any of its substrates:

$$\left.\begin{array}{rcll} f_1(0,\cdot) & = & 0 & \\ f_i(\cdot,0,\cdot) & = & 0 & i = 2,\cdots,n \\ f_i(m_{i-1},\cdot,\cdot) & = & 0 & i = 2,\cdots,n \\ f_{n+1}(m_n,\cdot) & = & 0 & \end{array}\right\} \tag{4.4}$$

The final set of conditions imply that increased substrate concentration increases the rate of reaction unless one of the substrates is entirely absent:

$$\left.\begin{array}{rcl} f_{11} & > & 0 \\ f_{ii} & \geq & 0 \text{ and } f_{ii} > 0 \text{ if } x_{i-1} < m_{i-1} \quad i = 2, \cdots, n \\ f_{i+1,i} & \leq & 0 \text{ and } f_{i+1,i} < 0 \text{ if } x_{i+1} > 0 \quad i = 1, \cdots, n-1 \\ f_{n+1,n} & < & 0 \end{array}\right\} \tag{4.5}$$

The fact that the first and final inequalities are always strict implies that there is always some electron donor to reduce the initial substrate, and some electron acceptor to oxidise the final substrate, and ensures nondegenerate behaviour. The assumptions from equations (4.3) and (4.5) mean that f_{ii}, F_{ij} and $F_{i\psi}$ as defined in (4.2) are all nonnegative. The definition of these nonnegative quantities is solely to simplify later arguments.

4.2 General behaviour of the system

Some properties of the model that hold regardless of the number n of redox pairs are outlined in this section.

4.2.1 Boundedness of solutions

It is convenient to define an $n \times (n+1)$ matrix which can be regarded as a stoichiometric matrix for the redox reactions:

$$S \equiv \begin{pmatrix} -1 & 1 & \cdots & 0 & 0 \\ 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

Defining the vector of reactant concentrations $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, the vector of reaction rates $\mathbf{v}(\mathbf{x}, \psi) = [f_1, f_2, \ldots f_{n+1}]^T$, and the nonnegative vector $P \equiv [p_1, \ldots, p_{n+1}]^T$, the system of equations (4.1) can be rewritten more briefly as

$$\begin{array}{rcl} \dot{\mathbf{x}} & = & S\mathbf{v}(\mathbf{x}, \psi) \\ \dot{\psi} & = & P^T\mathbf{v}(\mathbf{x}, \psi) - L(\psi) \end{array}$$

Since the phase space is bounded in $\mathbf{x}$, what needs to be shown is that all trajectories enter a bounded region in the ψ direction. This amounts to showing that $\dot{\psi} > 0$ for ψ sufficiently large and negative, and that $\dot{\psi} < 0$ for ψ sufficiently large and positive. By assumption, for any given i, either $p_i = 0$ or $f_{i\psi}$ is strictly negative and $\lim_{\psi\to\infty} f_i(\cdot, \cdot, \psi) \leq$

0, $\lim_{\psi \to -\infty} f_i(\cdot, \cdot, \psi) \geq 0$. This in turn implies that $\lim_{\psi \to \infty} P^T\mathbf{v}(\mathbf{x}, \psi) \leq 0$ and $\lim_{\psi \to -\infty} P^T\mathbf{v}(\mathbf{x}, \psi) \geq 0$. In addition, by equation (4.3), $L(0) = 0$ and $L_\psi > 0$. Thus for any fixed value of $\mathbf{x}$, $\lim_{\psi \to \infty} P^T\mathbf{v}(\mathbf{x}, \psi) - L(\psi) < 0$ and $\lim_{\psi \to -\infty} P^T\mathbf{v}(\mathbf{x}, \psi) - L(\psi) > 0$. Define $\psi_0(\mathbf{x})$ as the value of ψ at which $P^T\mathbf{v}(\mathbf{x}, \psi) - L(\psi) = 0$. $\psi_0(\mathbf{x})$ is uniquely defined since $P^T\mathbf{v}(\mathbf{x}, \psi) - L(\psi)$ is strictly decreasing. By the implicit function theorem, $\psi_0(\mathbf{x})$ is a differentiable function since $P^T\mathbf{v}(\mathbf{x}, \psi) - L(\psi)$ is a differentiable function of ψ. Since it has a compact domain, $\psi_0(\mathbf{x})$ achieves a maximum value which will be called ψ_{max}, and a minimum value which will be called ψ_{min}. By these definitions, $\dot{\psi}(\psi, \mathbf{x}) < 0$ for all $\psi > \psi_{max}$, and $\dot{\psi}(\psi, \mathbf{x}) > 0$ for all $\psi < \psi_{min}$.

Thus all trajectories enter a closed box, $\mathcal{B}$, bounded by the hyperplanes $x_i = 0$, $x_i = m_i$, $\psi = \psi_{min}$ and $\psi = \psi_{max}$, and this box forms a trapping region for the system in all dimensions.

4.2.2 The Jacobian

Direct calculation gives that the Jacobian, J, of the system is:

$$J = \begin{pmatrix} -f_{11} - F_{21} & f_{22} & \cdots & 0 & F_{1\psi} - F_{2\psi} \\ F_{21} & -f_{22} - F_{32} & \cdots & 0 & F_{2\psi} - F_{3\psi} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -f_{nn} - F_{n+1,n} & F_{n\psi} - F_{n+1,\psi} \\ p_1 f_{11} - p_2 F_{21} & p_2 f_{22} - p_3 F_{32} & \cdots & p_n f_{nn} - p_{n+1} F_{n+1,n} & -L_\psi - \sum_{i=1}^{n+1} p_i F_{i\psi} \end{pmatrix}$$

The structure of this Jacobian can be made clearer by defining two further quantities: A nonnegative vector in $\mathbb{R}^n$, $F \equiv [F_{1\psi}, \ldots, F_{n\psi}]^T$; and an $(n+1) \times n$ matrix

$$V \equiv \frac{\partial \mathbf{v}}{\partial \mathbf{x}} = \begin{pmatrix} f_{11} & 0 & 0 & \cdots & 0 \\ -F_{21} & f_{22} & 0 & \cdots & 0 \\ 0 & -F_{32} & f_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & f_{nn} \\ 0 & 0 & 0 & \cdots & -F_{n+1,n} \end{pmatrix}$$

Then the Jacobian can be written in the block form:

$$J = \begin{pmatrix} SV & SF \\ P^T V & -P^T F - L_\psi \end{pmatrix} \tag{4.6}$$

SV is the Jacobian of the system without feedback, which is tridiagonal, and can easily be shown to have real negative eigenvalues [Banaji, 2006] using a result from [Smillie, 1984].

Example 2 in [Banaji et al., 2007] shows that the structures of S and V along with the nonnegativity of P and F imply that J is a $P^{(-)}$ matrix. This result is independent of n, the length of the chain; its consequences are discussed further in the next section. It is also possible to show directly that J is $P^{(-)}$ matrix, but the calculation is extremely long and tedious.

4.2.3 A unique equilibrium

The existence of a unique equilibrium for this system was shown in [Banaji, 2006] by a direct method. It also follows from the arguments presented above: Since $\mathcal{B}$ is a compact, convex trapping region, there must be an equilibrium, as a corollary of the Brouwer fixed point theorem or by theorem 3 (p. 17). This equilibrium must be unique due to the fact that the Jacobian is a $P^{(-)}$ matrix, and hence the system is injective on rectangular regions of phase space (recall the discussion in §1.4.3). Thus the first result arising from the model of electron transport is that electron transport chains coupled to charge translocation across a membrane have exactly one equilibrium.

It is interesting that the possibility of multistability is immediately ruled out. However, this in itself does not guarantee that all trajectories must necessarily converge to the equilibrium; further analysis is required to determine whether periodic or chaotic behaviour is still possible.

4.3 Stability of the equilibrium

This section is concerned with analysing the stability of the equilibrium, starting with low dimensions (i.e. short chains). In [Banaji, 2006] it was proved that the system without feedback is globally asymptotically stable, using a result of [Smillie, 1984].

For the system with feedback in two dimensions it is proved that the equilibrium is also globally asymptotically stable. In three dimensions it is shown that the addition of an extra, reasonable, constraint implies that the equilibrium is locally stable, and further constraints ensure that it is globally stable. After this it is demonstrated that these constraints do not suffice to guarantee stability in four dimensions and higher.

4.3.1 The system in two dimensions

The system in 2D consists of a single redox pair subject to a reduction process and an oxidation process, both possibly coupled to proton translocation across the membrane. It takes the form

$$\begin{aligned}\dot{x}_1 &= -f_1(x_1,\psi)+f_2(x_1,\psi)\\ \dot{\psi} &= p_1f_1+p_2f_2-L(\psi)\end{aligned}$$

The Jacobian of the system in this case is:

$$J_2 = \begin{pmatrix} -f_{11}-F_{21} & F_{1\psi}-F_{2\psi} \\ p_1f_{11}-p_2F_{21} & -L_\psi-p_1F_{1\psi}-p_2F_{2\psi} \end{pmatrix} \tag{4.7}$$

It has already been pointed out in §1.4.4 that 2D $P^{(-)}$ matrices are Hurwitz stable, and it follows that the matrix J_2 is Hurwitz stable at all points in phase space. This can also be easily shown by a direct calculation: the Routh-Hurwitz conditions for a two-dimensional matrix M are that $\mathrm{Tr}(M) < 0$ and $|M| > 0$. Clearly $\mathrm{Tr}(J_2) < 0$ by inspection. $|J_2| = p_1(f_{11}F_{2\psi} + F_{21}F_{1\psi}) + p_2(f_{11}F_{2\psi} + F_{21}F_{1\psi}) + L_\psi(f_{11} + F_{21})$, which is positive by inspection. Therefore J_2 is Hurwitz stable.

Since J_2 is Hurwitz stable *everywhere*, not just at the equilibrium, the Markus-Yamabe theorem (theorem 19, p. 42) ensures that the equilibrium is globally stable.

An alternative proof of global asymptotic stability runs as follows: By the Poincaré–Bendixson theorem (theorem 1, p. 16), ω-limit sets of a flow on compact subsets of $\mathbb{R}^2$ must either contain equilibria or consist of a periodic orbit. In this case the possibility of periodic orbits can be ruled out: The divergence of the vector field is equal to

$$Tr(J) = -f_{11} - F_{21} - p_1F_{1\psi} - p_2F_{2\psi} - L_\psi$$

which is negative. Thus the vector field satisfies the Dulac criterion (theorem 1, p. 16) with $g = 1$ and there are no periodic orbits. There is only one equilibrium, which is locally stable, and therefore there are no heteroclinic or homoclinic orbits either. Since every forward trajectory enters the box $\mathcal{B}$, the equilibrium must be the ω-limit of every trajectory, and is hence globally stable.

4.3.2 The system in three dimensions

Slightly more complex than the two dimensional system is the system in three dimensions. This system has two substrates and three reactions, comprising an initial reduction reaction, an intermediary electron transfer reaction, and finally an oxidation. Any or all of these reactions might also pump protons across the membrane. The mathematical representation of this system takes the form

$$
\begin{aligned}
\dot{x}_1 &= -f_1(x_1,\psi) + f_2(x_1,x_2,\psi) \\
\dot{x}_2 &= -f_2(x_1,x_2,\psi) + f_3(x_2,\psi) \\
\dot{\psi} &= p_1 f_1 + p_2 f_2 + p_3 f_3 - L(\psi)
\end{aligned}
$$

with Jacobian

$$
J_3 = \begin{pmatrix} -f_{11} - F_{21} & f_{22} & F_{1\psi} - F_{2\psi} \\ F_{21} & -f_{22} - F_{32} & F_{2\psi} - F_{3\psi} \\ p_1 f_{11} - p_2 F_{21} & p_2 f_{22} - p_3 F_{32} & -L_\psi - p_1 F_{1\psi} - p_2 F_{2\psi} - p_3 F_{3\psi} \end{pmatrix} \tag{4.8}
$$

As it stands, J_3 is not always Hurwitz. For example, consider the Jacobian constructed using the following values: $p_1 = 3, p_2 = 0, p_3 = 88, F_{1\psi} = 33, F_{2\psi} = 4, F_{3\psi} = 0.6, f_{11} = 23, f_{22} = 3, F_{21} = 94, F_{32} = 76, L_\psi = 6$. Its eigenvalues are, to 2 d.p., $\lambda_1 = -357.50, \lambda_2 = 1.85 + 248.89i, \lambda_3 = 1.85 - 248.89i$.

J_3 can be shown to be Hurwitz everywhere in phase space provided one extra condition is met: p_1 and p_3 must have the same ordering as $F_{1\psi}$ and $F_{3\psi}$.

Then the ordering assumption translates to the following statement:

$$
\operatorname{sgn}(F_{3\psi} - F_{1\psi}) = \operatorname{sgn}(p_3 - p_1) \tag{4.9}
$$

With this assumption, the Jacobian is everywhere Hurwitz, and hence the equilibrium is locally asymptotically stable. The proof is simple but requires some lengthy evaluations using the Routh-Hurwitz theorem (theorem 6, p. 22). In three dimensions, the theorem requires that the three quantities

$$
\begin{aligned}
\Delta_1 &= b_1 && (4.10) \\
\Delta_2 &= b_1 b_2 - b_3 && (4.11) \\
\Delta_3 &= b_3(b_1 b_2 - b_3) = b_3 \Delta_2 && (4.12)
\end{aligned}
$$

are all positive. Since J_3 is a $P^{(-)}$ matrix and all the b_i are therefore positive (see the final paragraph of §1.4.4), all three Δ_i are positive if and only if $\Delta_2 > 0$. This in turn follows (condition 12 in [Kafri, 2002]) if, for a matrix (a_{ij}),

$$
0 < a_{12}a_{23}a_{31} + a_{21}a_{32}a_{13} - 2a_{11}a_{22}a_{33}
$$

Substituting a_{ij} for the elements of J_3 using equation (4.8) and expanding using the open source symbolic algebra program [Maxima, 2008] gives the following condition:

$$\begin{aligned} a_{12}a_{23}a_{31} + a_{21}a_{32}a_{13} - 2a_{11}a_{22}a_{33} &= F_{21}\,F_{32}\,(2p_3F_{3\psi} + 2p_1F_{1\psi} - p_3F_{1\psi}) \\ &\quad + f_{11}\,f_{22}\,(2p_3F_{3\psi} + 2p_1F_{1\psi} - p_1F_{3\psi}) \\ &\quad + \text{positive terms} \end{aligned}$$

With the ordering assumption given in equation (4.9), the following relations hold:

$$2p_3F_{3\psi} + 2p_1F_{1\psi} - p_3F_{1\psi} \geq 0 \tag{4.13}$$

$$2p_3F_{3\psi} + 2p_1F_{1\psi} - p_1F_{3\psi} \geq 0 \tag{4.14}$$

Thus the Jacobian is Hurwitz everywhere, and hence the unique equilibrium of the system must be locally asymptotically stable. Note that the restriction from equation (4.9) is stronger than necessary to ensure that J_3 is Hurwitz, but no other set of conditions with a clear physical meaning that make the Jacobian Hurwitz have been discovered. Finding a set of necessary and sufficient conditions with physical meaning for J_3 to be Hurwitz is a difficult problem.

Unlike the two-dimensional case it does not automatically follow that the equilibrium is globally stable, since the Markus-Yamabe conjecture does not hold in dimensions greater than two (see the discussion at the beginning of chapter 3). However global stability can be demonstrated in this case too, subject to a strengthened version of the ordering assumption on the quantities p_i and $F_{i\psi}$. The strengthened requirements are that

$$\operatorname{sgn}(F_{i\psi} - F_{j\psi}) = \operatorname{sgn}(p_i - p_j) \tag{4.15}$$

for $i, j \in \{1, 2, 3\}$.

With this assumption it is possible to use Li and Muldowney's autonomous convergence theorem (theorem 23, p. 53) to show that the unique equilibrium is globally stable.

Recall that for the purposes of this system, the autonomous convergence theorem states the following: Let $J(x)$ be the Jacobian of a dynamical system at a point x in phase space, and define $\mathbf{J}$ to be the set of all these Jacobians as x varies over phase space. If a logarithmic norm μ can be found such that

$$\mu(J^{[2]}) < 0 \;\forall\; J \in \mathbf{J} \tag{4.16}$$

and the dynamical system has a globally absorbing compact subset of phase space containing a unique fixed point, the fixed point is globally asymptotically stable.

Since all trajectories enter the compact trapping region $\mathcal{B}$ in the system, and since $\mathcal{B}$ contains a unique equilibrium, finding a suitable logarithmic norm satisfying equation (4.16) will suffice to prove global stability of the equilibrium.

The second additive compound in this case is:

$$J_3^{[2]} = \begin{pmatrix} -f_{11}-F_{21}-f_{22}-F_{32} & F_{2\psi} - F_{3\psi} & -(F_{1\psi} - F_{2\psi}) \\ p_2 f_{22}-p_3 F_{32} & -f_{11}-F_{21}-L_\psi-\sum\limits_{i=1}^{3} p_i F_{i\psi} & f_{22} \\ -(p_1 f_{11}-p_2 F_{21}) & F_{21} & -f_{22}-F_{32}-L_\psi-\sum\limits_{i=1}^{3} p_i F_{i\psi} \end{pmatrix}$$

It is possible to construct a logarithmic norm μ_T such that $\mu_T\left(J_3^{[2]}\right) < 0$. Recall from equation (3.5) (p. 45) that for a real $n \times n$ matrix (M_{ij}), the logarithmic norm corresponding to the usual $\| \cdot \|_1$ norm takes the form:

$$\mu_1(M) = \max_{i \in \{1,\dots,n\}} \left(M_{ii} + \sum_{k,k \neq i} |M_{ki}| \right)$$

From the definition it is clear that a matrix has negative logarithmic norm μ_1 if and only if every diagonal entry is negative and it is strictly diagonally dominant in every column. Define a constant diagonal coordinate transformation

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{p_{max}} & 0 \\ 0 & 0 & \frac{1}{p_{max}} \end{pmatrix}$$

where $p_{max} = \max\limits_{i \in \{1,2,3\}} (p_i)$.

This gives the transformed matrix

$$TJ^{[2]}T^{-1} = \begin{pmatrix} -f_{11}-F_{21}-f_{22}-F_{32} & p_{max}(F_{2\psi} - F_{3\psi}) & -p_{max}(F_{1\psi} - F_{2\psi}) \\ \frac{p_2 f_{22}-p_3 F_{32}}{p_{max}} & -f_{11}-F_{21}-L_\psi-\sum\limits_{i=1}^{3} p_i F_{i\psi} & f_{22} \\ -\frac{p_1 f_{11}-p_2 F_{21}}{p_{max}} & F_{21} & -f_{22}-F_{32}-L_\psi-\sum\limits_{i=1}^{3} p_i F_{i\psi} \end{pmatrix}$$

As stated in lemma 8 (p. 45), given any invertible matrix T of suitable dimension, $\mu_{1,T}(M) \equiv \mu_1(TMT^{-1})$ defines a new logarithmic norm. In this case, since T is a diagonal matrix, the diagonal entries of M are the same as those of TMT^{-1}. Thus in order to prove that $\mu_{1,T}(J_3^{[2]}) < 0$, it is necessary to show that each of the negative diagonal elements of $J' \equiv TJ_3^{[2]}T^{-1}$ dominates the sum of magnitudes of other elements in the same column.

For the first column the sum is

$$\begin{aligned} J'_{11} + \left|J'_{21}\right| + \left|J'_{31}\right| &= -f_{22} - F_{32} - f_{11} - F_{21} \\ &\quad + \left|\frac{p_2}{p_{max}} f_{22} - \frac{p_3}{p_{max}} F_{32}\right| + \left|\frac{p_2}{p_{max}} F_{21} - \frac{p_1}{p_{max}} f_{11}\right| \end{aligned}$$

It can be seen that the term on the right hand side is negative since for any two nonnegative scalars $|a-b| \leq \max\{|a|,|b|\}$, and $f_{22}, F_{21} \geq 0, f_{11}, F_{32} > 0$.

For the second column the sum is

$$J'_{22} + \left|J'_{12}\right| + \left|J'_{32}\right| = -\sum_{i=1}^{3} p_i F_{i\psi} - L_\psi - f_{11} + p_{max}\left|F_{2\psi} - F_{3\psi}\right|$$

For the final column the sum is

$$J'_{33} + \left|J'_{13}\right| + \left|J'_{23}\right| = -\sum_{i=1}^{3} p_i F_{i\psi} - L_\psi - F_{32} + p_{max}\left|F_{2\psi} - F_{1\psi}\right|$$

In order to show that the right hand sides of the last two expressions are negative it is necessary to show in each case that the ordering assumption given in equation (4.15) implies that the final (positive) term is dominated in magnitude by the other terms.

Note that $|F_{i\psi} - F_{j\psi}| \leq \max\{F_{i\psi}, F_{j\psi}\} \leq \max_{k\in\{1,2,3\}}(F_{k\psi})$. Then there are only three cases:

1. if $p_{max} = p_1$, then $p_{max}\left|F_{2\psi} - F_{3\psi}\right| \leq p_1 F_{1\psi}$, and $p_{max}\left|F_{2\psi} - F_{1\psi}\right| \leq p_1 F_{1\psi}$.
2. if $p_{max} = p_2$, then $p_{max}\left|F_{2\psi} - F_{3\psi}\right| \leq p_2 F_{2\psi}$, and $p_{max}\left|F_{2\psi} - F_{1\psi}\right| \leq p_2 F_{2\psi}$.
3. if $p_{max} = p_3$, then $p_{max}\left|F_{2\psi} - F_{3\psi}\right| \leq p_3 F_{3\psi}$, and $p_{max}\left|F_{2\psi} - F_{1\psi}\right| \leq p_3 F_{3\psi}$.

Each of these possibilities leads to the same conclusion – that $J'_{ii} + \sum_{k,k\neq i}|J'_{ki}| < 0$ for each i. Hence it follows that $\mu_T\left(J_3^{[2]}\right) < 0$.

This result means that if the ordering assumption from equation (4.15) holds, then the unique equilibrium is globally stable. The ordering assumption itself has the following reasonable physical meaning: Suppose redox reaction i is involved in pumping more protons across the membrane than redox reaction j. Then an increase in ψ will slow down reaction i at a greater rate than reaction j. It is interesting to note that this assumption is not necessary to prove global stability in the two-dimensional case. It is also unknown whether the weaker assumption given in equation (4.9), which guarantees that the Jacobian is Hurwitz everywhere, actually guarantees global stability in three dimensions.

4.3.3 Unstable examples in higher dimensions

The ordering assumption made in equation (4.15) does not guarantee global or even local stability of the equilibrium in dimensions greater than three. It is easy to construct counterexamples. For example, in four dimensions, the Jacobian constructed by choosing $p_1 = 2$, $p_2 = p_3 = 0$, $p_4 = 73$, $F_{1\psi} = 167$, $F_{2\psi} = F_{3\psi} = 0$, $F_{4\psi} = 176$, $f_{11} = 4$, $f_{22} = 7$, $f_{33} = 1$, $F_{21} = 32$, $F_{32} = 64$, $F_{43} = 174$, $L_{\psi} = 33$, satisfies all the constraints, including the ordering assumption on the values of p_i and $F_{i\psi}$. However its eigenvalues to 2 d.p. are $\lambda_1 = -113.61, \lambda_2 = -13384.34, \lambda_3 = 0.48 + 54.26i, \lambda_4 = 0.48 - 54.26i$, and so it is not Hurwitz stable. In this example, if F_{43} is decreased and all other parameters are unchanged, the real part of λ_3 and λ_4 passes through zero, corresponding to a Hopf bifurcation. This implies the existence of a periodic orbit for certain parameter values.

Note that:

1. By continuity, the fact that a non-Hurwitz Jacobian can be constructed in 4 dimensions guarantees that such examples also exist in all higher dimensions.
2. Systems with non-Hurwitz Jacobian satisfying the ordering assumption given by equation (4.15) seem to be rare. Through use of an automated computer script written for the open source numerical computation program [Scilab, 2008], counterexamples in dimension 4 were found by randomly choosing values for the different terms in the Jacobian, such that all the assumptions were satisfied. Out of hundreds of millions of sets of values, less than ten were non-Hurwitz.
3. The counterexamples found appear always to be close to breaking the ordering assumption. For instance, in the example shown, p_4 is much greater than p_1, whereas $F_{4\psi}$ is close in magnitude to $F_{1\psi}$.

4.4 Discussion and conclusions

In this chapter, the behaviour of electron transport chains coupled to a charge translocation process has been analysed in some detail, using a variety of mathematical techniques. In all cases trajectories are bounded, and there is a unique equilibrium, but questions about the stability of this equilibrium have proved harder to answer. Where the chain consists of a single redox pair, the unique equilibrium is globally stable. When there are two redox pairs the same conclusions can be reached subject to some extra conditions on the feedback process. In higher dimensions no such general conditions could easily be found. Thus the length of the electron transport chain is crucial in deciding on stability of the equilibrium.

It is somewhat surprising that the coupling of electron transfer to a membrane potential – a negative feedback loop – can serve to destabilise the unique equilibrium in these systems.

Interestingly, the system in any dimension can be proved to be Hurwitz at all points by making a further assumption about the reaction rates. The assumption takes two parts:

1. Associated with each half reaction is some "potential": In the case of a redox reaction of the form $A_i + e \rightleftharpoons B_i$, a potential means any strictly increasing scalar function of x_i (i.e. the concentration of A_i), such as a redox potential; in the case of a charge transfer across a membrane a potential means any strictly increasing scalar function of ψ.
2. The rate of any full reaction depends only on the sum of the potentials for the half reactions involved, and is a strictly decreasing function of this sum.

With this extra assumption, it can be shown that the Jacobian is D-stable [Kafri, 2002], and hence Hurwitz — see §4.4 of [Donnell et al., 2008] for the full analysis. Reaction rates cannot in general be seen in this way, but in the case of reactions which are primarily about charge transfer, the assumption could be reasonable. The choice of reaction rates in some existing numerical models, such as [Korzeniewski, 1996], satisfy this assumption.

There are some interesting open questions, both biological and mathematical. From a biological point of view, it is of interest to find out whether experiments on mitochondria with constant inputs ever display behaviour other than convergence to an equilibrium, such as periodic or chaotic behaviour. If this is never the case, then this suggests that the very general model presented here may be omitting certain important biological/thermodynamic restrictions on the reaction rates, which would tend to stabilise the system. For example, the ordering assumption of equation 4.15 used to guarantee stability in three dimensions may be too crude; a more realistic assumption might well relate to the free energy released by each reaction to the energy required to pump a proton across the membrane at a given gradient. Of course, such an assumption would be difficult to construct qualitatively (i.e. non-numerically). An assumption of this type could also prove difficult to justify given that the details of some of the reactions in the chain are currently not clear, although it might prove interesting to attempt to make predictions about the unknown details of the reactions, based on a known set of assumptions that guarantee stability. It would also be interesting to see how additional processes such as transport processes in the full numerical models ([Korzeniewski, 1996], [Beard, 2005] for example) affect the conclusions presented here.

An open mathematical question is whether there are equivalent conditions to the ordering condition in three dimensions which ensure that the Jacobian of the system is Hurwitz in arbitrary dimension, or better still that the second additive compound has negative logarithmic norm, and hence the unique equilibrium is globally stable. If such conditions exist, can they be given general biological meanings?

It would also be interesting to explore when the results presented here survive weakening of the assumption that electrons are transferred along a chain. Although electron transfers taking place in the mitochondrial membrane are often described via a “chain” it is likely that this description is a convenient simplification rather than the whole truth. As mentioned previously, general electron transfer networks in the absence of a gradient were analysed in [Banaji and Baigent, 2008] and found to have simple behaviour. Application of the theory presented in [Banaji et al., 2007] should allow determination of when these networks give rise to $P^{(-)}$ Jacobians when interacting with a membrane potential.

Finally, although the extra assumptions made above about the reaction rates imply that the Jacobian is everywhere Hurwitz, it is an open question as to whether this implies global stability of the unique equilibrium. Since the Markus-Yamabe conjecture does not hold in dimensions greater than two, as mentioned at the beginning of chapter 3, global stability does not follow automatically from local stability, and requires independent proof.

Chapter 5

Cellular gap junctions

This chapter focuses on a model of inter-cellular communication via gap junctions. It begins with a brief outline of the biology of gap junctions, and then translates the qualitative properties of a gap junction into a mathematical model. The remainder of the chapter is concerned with analysing the properties of this model.

Before the main discussion of the model, it is important to note its relationship with earlier work on the same problem. A similar model of a cellular gap junction was introduced in [Baigent et al., 1997], and further developed in [Baigent, 2003]. These earlier papers presented a model consisting of a pair of cells linked by a gap junction made up of a large number of conduction channels, each of which can exist in a finite number of conducting states[1]. All the channels were assumed to be identical, and the probability of a channel changing from one conducting state to another was given a specific functional form, depending exponentially on the trans-junctional voltage. The conclusion of [Baigent et al., 1997] was that, in some physically reasonable parameter ranges, the system could have two locally asymptotically stable steady states. [Baigent, 2003] examined conditions under which the system is monotone and globally convergent to a unique steady state.

The model analysed in this chapter is similar: it consists of a pair of cells linked by a gap junction made up of a large number of conduction channels, but it is assumed that the channels have only two possible states, “open” (high conductance) and “closed” (low conductance). The probabilities of transition between states are once again assumed to be functions of trans-junctional voltage, but are **not** given a functional form. From this structure, some results regarding the number and local stability of fixed points are constructed. Later on, mild constraints are put on the transition probabilities by assuming that they are monotone functions of trans-junctional voltage. The consequences of these extra assumptions are then explored: it is shown that if the signs on the functions are

[1] A more detailed description of the biology appears in §5.1; the essential features are mentioned briefly here to differentiate between previous work and the model described in this chapter.

one way round there can be only one fixed point. If the signs on the functions are the other way round, there may be more than one fixed point but the system is monotone and (under a strict monotonicity assumption) all trajectories of the system converge to one of these fixed points. Note that while the model discussed in this chapter is heavily based on that which appeared in [Baigent et al., 1997] and [Baigent, 2003], the analysis itself is independent.

The results presented here do not demonstrate any new behaviour of the gap junction system. However, they do elucidate some of the important features of the model. The monotonicity assumptions on the transition probabilities made in this chapter that guarantee monotonicity of the system as a whole are satisfied by the exponential functional forms used in [Baigent et al., 1997] and [Baigent, 2003], which are themselves based on experimental data. The results in this chapter demonstrate that the exponential functional forms are not necessary for the system to be monotone when the conduction channels have only two possible states, and in fact the system exhibits global convergence under much milder assumptions than those made in [Baigent, 2003], even when there is more than one fixed point. However, it is also apparent that finding conditions for monotonicity of the system when there are more than two possible conduction states is a difficult problem to solve in a general sense.

At the end of this chapter is a section with a few results pertaining to an extended model containing three or more cells joined by gap junctions, which was not done in the earlier papers. However, due to the extra complexity introduced by extending the model in this way, the conclusions drawn are more limited than those for the two cell model.

5.1 Background information

5.1.1 Biology of cellular gap junctions

The description of a cellular gap junction given here is simply an overview intended to highlight the key features required to construct a simple mathematical representation. For a more detailed discussion see a biology textbook such as [Alberts et al., 2002] or [Garrett and Grisham, 1995].

Gap junctions occur in most animal cells, and provide a means for neighbouring cells to share small ions and molecules. They allow cells without nerve connections, such as heart muscle cells, to act synchronously, and facilitate the passage of nutrients into cells that have no direct blood supply, such as cells in the lens of the eye. See figure 5.1 for a diagram of a gap junction[2]. As shown in the diagram, a gap junction is made up of a

[2] Image source: Wikipedia, accessed on January 25th, 2008. The original file is available at `http://commons.wikimedia.org/wiki/Image:Gap_cell_junction,_LangNeutral.svg`. The creator of the image

number of aqueous channels that connect the interiors of two cells. Each channel is made up of two **hemichannels**, one hemichannel being provided by each cell. The proteins that make up a hemichannel are known as **connexins**, and a group of six connexins forming a hemichannel is known as a **connexon**. The hemichannels of connexons on neighbouring cells line up with each other, forming a set of channels between the cells. The interaction of the hemichannels keeps the cell membranes at a distance of about 2–4 nanometres at the gap junction, which is considerably smaller than the intercellular gap in the absence of a junction.

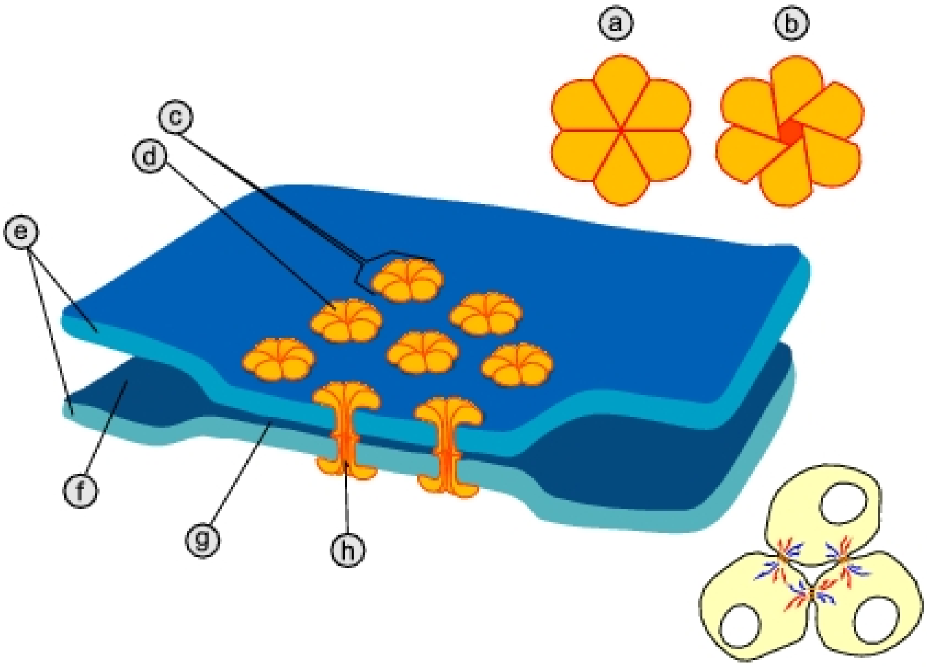

Figure 5.1: Diagram representing a cellular gap junction. In the diagram, **a** is an end-on view of a single conduction channel in its closed state, whereas **b** is a channel in an open state. Item **c** is called a **connexon**; it is a group of six **connexins** (**d**), which make up a hemichannel. The objects labelled **e** are cellular membranes, while **f** is the intercellular space. At a gap junction, the thickness of the intercellular space (**g**) is only a few nanometres. The final item, **h**, is a channel going between the two cells, made up of two aligned hemichannels. In the bottom right of the diagram is a representation of three cells connected by gap junctions. The red and blue arrows represent the flow of ions and molecules between the cells.

In order to protect cells and their neighbours from insult, the channels are able to close in response to environmental conditions. The model presented here only takes account of the potential difference between the two sides of a gap junction, but in reality channels are also sensitive to pH and Ca^{2+} differences. The permeability of each channel is generally non-

released it into the public domain.

linear with respect to the potential difference; the relationship between the trans-junctional voltage and conductance is discussed in [Baigent et al., 1997] and [Baigent, 2003].

As in the previous chapter, the aim is to construct a model based solely on qualitative information, and to examine its behaviour analytically. In this way, any conclusions drawn are valid are for a whole class of models that share the same underlying qualitative assumptions.

5.1.2 Mathematical representation of a gap junction

The model of a cellular gap junction outlined here is based on a model presented in [Baigent et al., 1997] and [Baigent, 2003]. Two cells, each with a capacitance, resting potential and membrane resistance are connected via a number of voltage-dependent conduction channels. The configuration of the cells can be considered analogous to an electronic circuit, as demonstrated in figure 5.2[3].

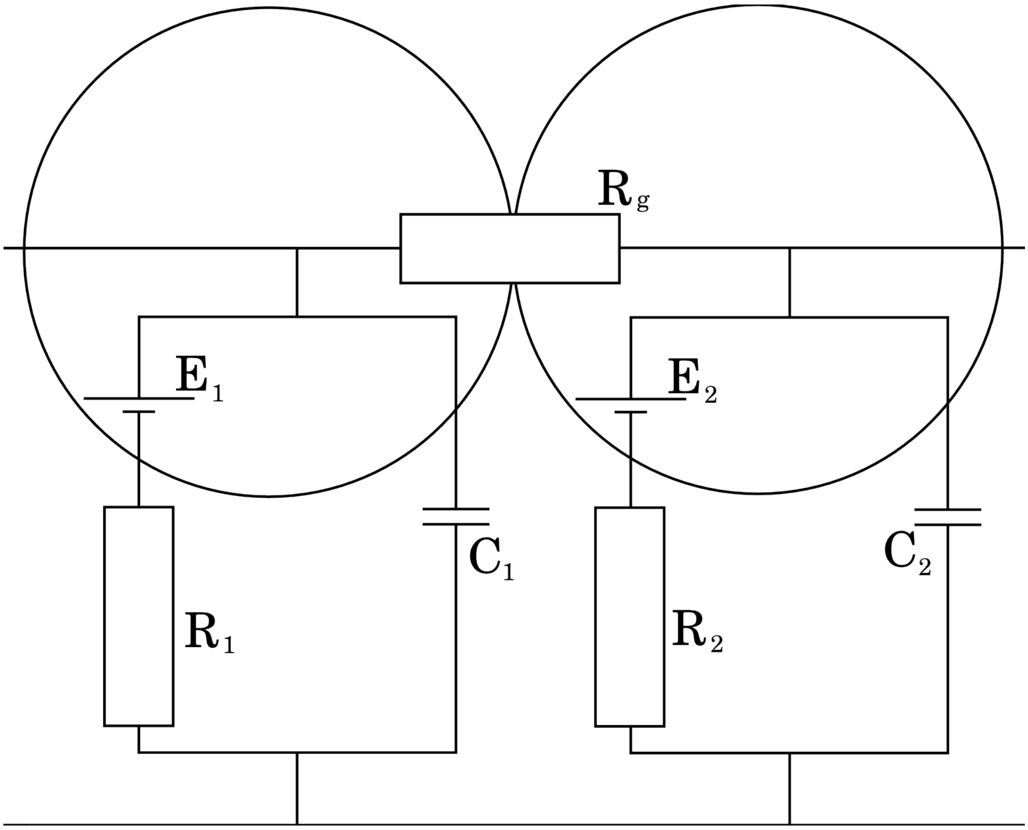

Figure 5.2: A circuit diagram representing a voltage dependent gap junction. C_i, E_i and R_i are the respective capacitance, internal E.M.F. and resistance of cell i for $i = 1, 2$. R_g is the resistance of the gap junction, and is equal to $1/g$.

[3] Image source: this diagram is essentially the same as figure 2 in [Baigent, 2003].

This system is modelled with three variables: the electrostatic potentials of the first and second cells are denoted $\phi_1, \phi_2 \in \mathbb{R}$. For simplicity the conduction channels are modelled as having only two states: "closed" and "open". The number of conduction channels in the gap junction is assumed to be large, and the fraction of the total number of channels in the closed state is denoted by the variable $x \in [0,1]$; accordingly, the fraction of channels in the open state is then equal to $(1-x)$.

The evolution of these variables is described by the following set of differential equations, which will be collectively referred to as system G:

$$\begin{aligned}
\dot{\phi}_1 &= -\frac{1}{R_1C_1}(\phi_1 - E_1) - \frac{g}{C_1}(\phi_1 - \phi_2) & (5.1)\\
\dot{\phi}_2 &= -\frac{1}{R_2C_2}(\phi_2 - E_2) + \frac{g}{C_2}(\phi_1 - \phi_2) & (5.2)\\
\dot{x} &= -\alpha x + \beta(1-x) & (5.3)
\end{aligned}$$

System G contains the following parameters: cell i has E.M.F. (or resting potential) E_i, membrane resistance R_i and capacitance C_i, all of which take positive values. To simplify some of the arguments in the following discussion, it is assumed (by switching labels on the cells if necessary) that $E_1 \geq E_2$.

The functions appearing in system G are defined as follows: $g : [0,1] \to \mathbb{R}_{>0} : x \mapsto g(x)$ is the conductance of the gap junction (i.e. the total conductance of all the conduction channels), and is given the functional form $g(x) = g_1 x + g_2(1-x)$. Since x is defined to represent the closed state, it follows that $g_1 < g_2$. It is further assumed that $g_1 > 0$, meaning that even when a channel is in its closed state its resistance remains finite, and consequently $g(x)$ is always strictly positive. The remaining functions are $\alpha : \mathbb{R} \to \mathbb{R}_{>0} : (\phi_1 - \phi_2) \mapsto \alpha(\phi_1 - \phi_2)$, which is the probability per unit time of a channel changing from a closed to an open state, and $\beta : \mathbb{R} \to \mathbb{R}_{>0} : (\phi_1 - \phi_2) \mapsto \beta(\phi_1 - \phi_2)$, which is the probability per unit time of a channel changing from an open to a closed state. The probability of a change in state is assumed to be always non-zero. Except when evaluated at a specified value of x, or $\phi_1 - \phi_2$, for notational convenience these functions will respectively be referred to as g, α and β. Both α and β are assumed to be C^1 in their arguments, but no other assumptions are initially made about their functional forms. Later on, the functional forms of α and β will be restricted in order to draw stronger conclusions about the behaviour of the system, but these extra assumptions will be made clear at the time they are introduced.

5.2 Properties of the model

The remainder of this chapter is dedicated to proving certain properties of the proposed model. It will be demonstrated that the system is closely linked to a simpler two dimensional system, and that consideration of this two dimensional system can be used to construct a simple graphical verification of the existence, number and stability of fixed points. Extra, physically reasonable restrictions on the system will then be presented that guarantee monotonicity of flows, and the consequences of this will be discussed.

5.2.1 Trapping region

Lemma 14. *All forward trajectories are bounded, and enter an invariant convex set in phase space.*

Proof. Since x lies within $[0,1]$, all trajectories are bounded in the x direction, and so the only concern is boundedness in the ϕ_1 and ϕ_2 directions. A ϕ_1-ϕ_2 phase plane for fixed x appears in figure 5.3.

Choose some $d > 0$ and use it to construct a square prism in the phase plane, with faces S_1, S_2, S_3 and S_4 such that each face is parallel to and perpendicular distance d from the planes defined by $\phi_1 = E_2, \phi_2 = E_2, \phi_1 = E_1$ and $\phi_2 = E_1$ respectively, as illustrated in figure 5.3. On face S_1, $\phi_1 = E_2 - d$ and $\phi_2 \geq E_2 - d$. Substituting these into the differential equation for ϕ_1 gives

$$\dot{\phi}_1 = -\frac{1}{R_1 C_1}(E_2 - d - E_1) - \frac{g}{C_1}(E_2 - d - \phi_2)$$

which is positive by inspection. Note that this holds for all values of x and also for all $d \geq 0$. Similar consideration of the vector field on face S_2 demonstrates that $\dot{\phi}_2 > 0$, since $\phi_2 = E_2 - d$ and $\phi_2 \leq \phi_1$. Likewise, along face S_3, $\dot{\phi}_1 < 0$ and along face S_4, $\dot{\phi}_2 < 0$.

Since the above inequalities hold for all values of $d \geq 0$, by varying d it is possible to construct an infinite family of square prisms that act as the level sets of a Lyapunov function, attracting all trajectories to the square prism defined by $d = 0$. This completes the lemma. □

It will be useful later on to demonstrate that all trajectories also enter a compact convex set in which $\phi_1 \geq \phi_2$. For fixed x, define a triangle $\mathcal{T}$ in phase space satisfying the inequalities $\phi_1 \leq E_1, \phi_2 \geq E_2, \phi_1 \geq \phi_2$, as illustrated in figure 5.4. Also define $\mathcal{T}' = \mathcal{T} \times [0,1]$, which is a triangular prism in phase space.

Lemma 15. *All trajectories enter the invariant, compact, convex set $\mathcal{T}'$.*

Proof. As shown in lemma 14, all trajectories are attracted to the square prism B'. It therefore suffices to show that the trajectory of every point in B' enters $\mathcal{T}'$.

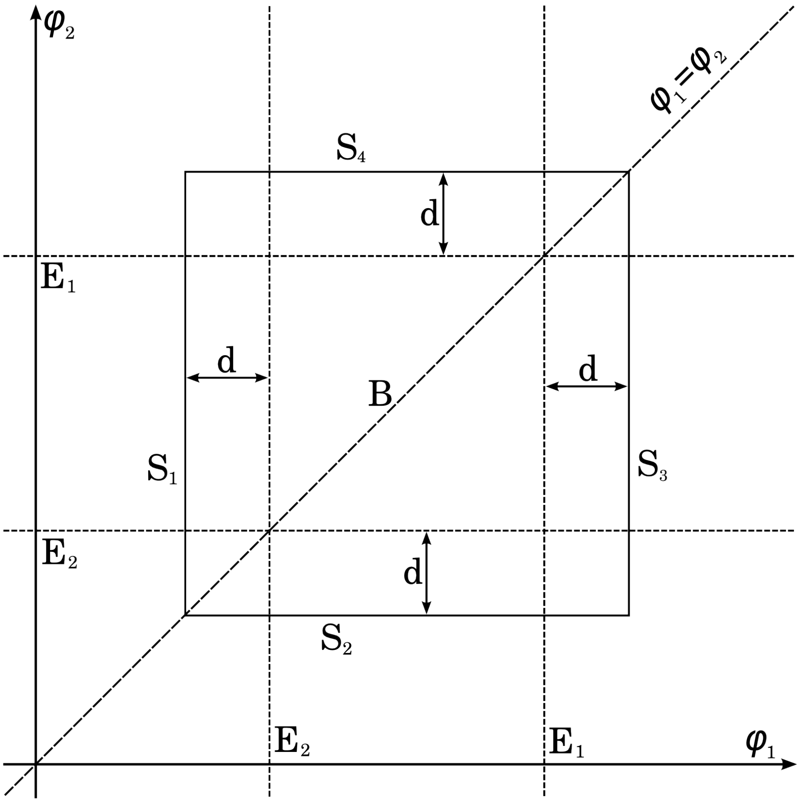

Figure 5.3: A two dimensional slice through phase space of the three dimensional cellular gap junction model. The x axis is not shown, but lies perpendicular to the page. At all points on the boundary of the square region B (bounded by dotted lines), the vector field of the system points into the interior of B, irrespective of the value of x. Therefore the square prism B' defined by $B' = B \times [0, 1]$ is a forward invariant set. Additionally, for all $d \geq 0$, the vector field points towards B' at all points on the square prism (bounded by solid lines) defined by the faces $S_1 \times [0, 1], S_2 \times [0, 1], S_3 \times [0, 1]$ and $S_4 \times [0, 1]$.

Consider the boundary of the triangular region marked $\mathcal{T}$ in figure 5.4. It is clearly convex and compact. Consequently, $\mathcal{T}'$ is also convex and compact. As the following argument shows, $\mathcal{T}'$ is also invariant. Along the section of the boundary of $\mathcal{T}$ where $\phi_1 = \phi_2$ (excluding the corners of $\mathcal{T}$ for the moment), the vector field components in the plane of fixed x are $\dot{\phi}_1 = -(\phi_1 - E_1)/(R_1C_1)$ and $\dot{\phi}_2 = -(\phi_2 - E_2)/(R_2C_2)$. Since in this region, $\phi_1, \phi_2 \in [E_2, E_1]$, it follows that $\dot{\phi}_1 > 0$ and $\dot{\phi}_2 < 0$. It is important to note that these inequalities are independent of the value of x, and therefore hold at all points along the face of $\mathcal{T}'$. Now consider the corner of $\mathcal{T}$ where $\phi_1 = \phi_2 = E_1$. For all values of x, the ϕ_i components of the vector field are $\dot{\phi}_1 = 0, \dot{\phi}_2 < 0$, so once again the vector field points into $\mathcal{T}'$ at all points that lie on the corner of $\mathcal{T}$. Likewise, when $\phi_1 = \phi_2 = E_2$, the vector field components are $\dot{\phi}_1 > 0, \dot{\phi}_2 = 0$. Thus the trajectories of all points along this part of the boundary remain within $\mathcal{T}'$ under the action of the flow.

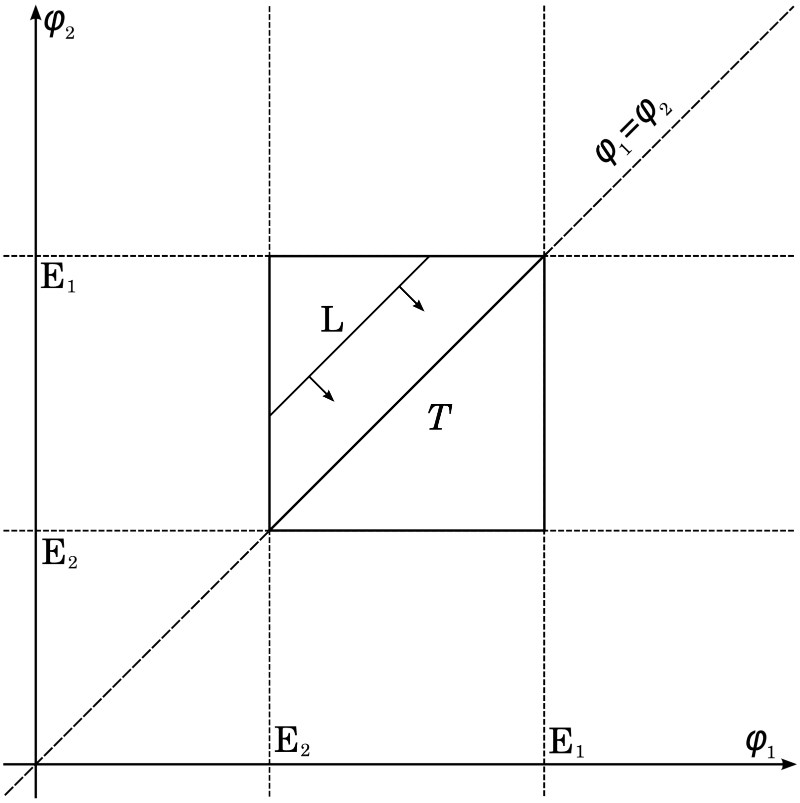

Figure 5.4: A two dimensional slice through phase space of the three dimensional cellular gap junction model. The x axis is not shown, but lies perpendicular to the page. The square region illustrated in the previous figure is divided into two triangles. All points along a given line segment L in the upper left half of the square are attracted to the lower right half, the triangular region $\mathcal{T}$.

Next, consider the section of the boundary of $\mathcal{T}$ where $\phi_1 = E_1$, ignoring the top corner since this has already been examined. At all points along this section, $\phi_1 > \phi_2$, and so the ϕ_1 component of the vector field is $\dot{\phi}_1 = -g(\phi_1 - \phi_2)/C_1$, which is strictly negative. Since g is always strictly positive, regardless of the value of x, this relation holds for all x when $\phi_1 = E_1$ along the boundary of $\mathcal{T}'$. Likewise, along the boundary section where $\phi_2 = E_2$, clearly $\phi_2 < \phi_1$ and therefore $\dot{\phi}_2 = g(\phi_1 - \phi_2)/C_2$ is strictly positive. Noting that both of these inequalities hold simultaneously in the bottom right corner of $\mathcal{T}$ where $\phi_1 = E_1$ and $\phi_2 = E_2$ for all values of x, the vector field therefore points inwards at all points along the boundary of $\mathcal{T}'$. Therefore $\mathcal{T}'$ is forward invariant.

To complete the proof, it remains to show that $\mathcal{T}'$ attracts all points in B'. In order to see this, consider a line segment L, parallel to the line $\phi_1 = \phi_2$. At every point on L, $\phi_1 < \phi_2$, $\phi_1 \in [E_2, E_1)$ and $\phi_2 \in (E_2, E_1]$. Consequently the vector field along L satisfies $\dot{\phi}_1 > 0$ and $\dot{\phi}_2 < 0$ for all values of x. In a similar way to the previous lemma, an infinite family of rectangles defined by $L \times [0, 1]$ lying at a perpendicular distance l from the line

$\phi_1 = \phi_2$, with $l \in (0, (E_1 - E_2)\sqrt{2}/2]$, form the level sets of a Lyapunov function attracting all points in $B' \setminus \mathcal{T}'$ into $\mathcal{T}'$. The proof is complete. □

Since all trajectories converge to the trapping region $\mathcal{T}'$, all further consideration of the dynamical system will be restricted to the interior of $\mathcal{T}'$. Note that the fact that $\phi_1 > \phi_2$ at all points in the interior of the trapping region came about due to the initial choice of labels that gave $E_1 > E_2$. In biological terms, this means that given a pair of cells that are connected by a gap junction, if the first cell has a higher resting potential than the second, then the potential in the first cell will eventually become higher than that in the second cell, and will remain higher.

5.2.2 Existence of fixed points

It is possible to argue that system G contains at least one fixed point by theorem 3 (p. 17), as was done for the electron transport chain model presented in the previous chapter. However, proving the existence of fixed points by a direct method turns out to have interesting results. The proof of the existence of fixed points presented in this section yields extra information about the fixed points, leading to a simple graphical check for local stability, which is outlined in §5.2.4.

The investigation of fixed points of system G that follows relies on the introduction of a new variable $V = \phi_1 - \phi_2, V \in \mathbb{R}$. V appears directly in equations (5.1–5.3), and as such is a natural coordinate of the system. Fixed points of system G in the original three variables are identified by examining a two-dimensional version of the system consisting only of V and x. The reduction of the system to two variables begins with the ϕ_1 and ϕ_2 null clines, which are, respectively:

$$-\frac{\phi_1 - E_1}{R_1 C_1} - \frac{g(x)}{C_1}(\phi_1 - \phi_2) = 0 \tag{5.4}$$

$$-\frac{\phi_2 - E_2}{R_2 C_2} + \frac{g(x)}{C_2}(\phi_1 - \phi_2) = 0 \tag{5.5}$$

The x null cline will also be used. It cannot be constructed explicitly, as the forms of α and β are not known, but it can be written down as

$$\dot{x} = 0 \Rightarrow x = \frac{\beta(V)}{\alpha(V) + \beta(V)} := \mathcal{X}(V) \tag{5.6}$$

Note that the above function describing the x null cline is defined for all $V \in \mathbb{R}$, since $\alpha(V)$ and $\beta(V)$ are always strictly positive.

In order to construct the two-dimensional version of system G, the ϕ_1 and ϕ_2 null clines will be used to generate a second equation relating x and V. Multiplying equation (5.4) by R_1C_1 and equation (5.5) by R_2C_2, then adding the two resulting equations together, gives the following equation relating x and V:

$$V = \frac{E_1 - E_2}{1 + (R_1 + R_2)g(x)} \tag{5.7}$$

Every fixed point of system G must satisfy both equation (5.6) and equation (5.7) (although the converse is not true since equation (5.7) is not a null cline). The proof of existence of fixed points of system G that follows relies on showing that there is at least one point in (V, x) space that simultaneously satisfies equations (5.6) and (5.7), and then that each such point in (V, x) space corresponds to a unique fixed point in (ϕ_1, ϕ_2, x) space. However, before proving these statements, some further information must be derived from equations (5.6) and (5.7).

Let $\hat{V}_0$ be the value of V at which $x = 0$, and let $\hat{V}_1$ be the value of V at which $x = 1$. Substituting $x = 0$ and $x = 1$ into equation (5.7) gives

$$\hat{V}_0 = \frac{E_1 - E_2}{1 + g_2(R_1 + R_2)} \tag{5.8}$$

$$\hat{V}_1 = \frac{E_1 - E_2}{1 + g_1(R_1 + R_2)} \tag{5.9}$$

Note that since $g_2 > g_1 > 0$, $R_1, R_2 > 0$ and $E_1 > E_2 > 0$, it follows that $0 < \hat{V}_0 < \hat{V}_1 < E_1 - E_2$. The derivative of equation (5.7) with respect to x is

$$\frac{\mathrm{d}V}{\mathrm{d}x} = -\frac{(E_1 - E_2)(R_1 + R_2)(g_1 - g_2)}{(1 + (R_1 + R_2)((g_1 - g_2)x + g_2))^2} > 0$$

Since this derivative is strictly positive and is defined for all x, it is clear that the function defining V in equation (5.7) is a bijection when its codomain is restricted to $[\hat{V}_0, \hat{V}_1]$. Consequently equation (5.7) can be inverted to give x as a function of V:

$$x = \frac{E_1 - E_2}{V(R_1 + R_2)(g_1 - g_2)} - \frac{1 + g_2(R_1 + R_2)}{(R_1 + R_2)(g_1 - g_2)} := \mathcal{V}(V) \tag{5.10}$$

The derivative of $\mathcal{V}(V)$ with respect to V is

$$\frac{\mathrm{d}}{\mathrm{d}V}\mathcal{V}(V) = -\frac{E_1 - E_2}{V^2(R_1 + R_2)(g_1 - g_2)} > 0$$

The first result now follows.

Lemma 16. *There exists at least one pair (V, x) that satisfies both equation (5.6) and equation (5.7).*

Proof. The set of pairs (V, x) defined by equation (5.6) and the set of pairs (V, x) defined by equation (5.7) intersect when $\mathcal{X}(V) = \mathcal{V}(V)$. Consider the value of $\mathcal{V} - \mathcal{X}$ at $V = \hat{V}_0$ and $V = \hat{V}_1$.

$$\mathcal{V}(\hat{V}_0) - \mathcal{X}(\hat{V}_0) \quad = \quad 0 - \frac{\beta(\hat{V}_0)}{\alpha(\hat{V}_0) + \beta(\hat{V}_0)} < 0 \tag{5.11}$$

$$\mathcal{V}(\hat{V}_1) - \mathcal{X}(\hat{V}_1) \quad = \quad 1 - \frac{\beta(\hat{V}_1)}{\alpha(\hat{V}_1) + \beta(\hat{V}_1)} > 0 \tag{5.12}$$

Since $\mathcal{V} - \mathcal{X}$ is continuous, by the intermediate value theorem it has at least one zero over the domain $V \in [\hat{V}_0, \hat{V}_1]$. Therefore there exists at least one point in (V, x) space that satisfies both equation (5.6) and equation (5.7). □

Lemma 17. *Pairs (V, x) satisfying equations (5.6) and (5.7) are in one to one correspondence with fixed points of system G.*

Proof. All fixed points of system G must satisfy equations (5.6) and (5.7), and hence corresponding to any fixed point of G is a pair (V, x) satisfying these equations. Therefore proving the result requires showing only that for each pair (V, x) satisfying equations (5.6) and (5.7) there is a unique point (ϕ_1, ϕ_2, x) in the phase space of system G which is a fixed point of G.

These fixed points can be constructed explicitly. In addition to satisfying equations (5.6) and (5.7), each fixed point must also satisfy equations (5.4) and (5.5). Let $(\overline{V}, \overline{x})$ be a pair of values satisfying both equations (5.6) and (5.7). Substituting these values into equations (5.4) and (5.5) gives

$$\overline{\phi}_1 = E_1 - g(\overline{x}) R_1 \overline{V} \tag{5.13}$$

$$\overline{\phi}_2 = E_2 + g(\overline{x}) R_2 \overline{V} \tag{5.14}$$

Thus corresponding to the point $(\overline{V}, \overline{x})$ in (V, x) space is a point $(\overline{\phi}_1, \overline{\phi}_2, \overline{x})$ in (ϕ_1, ϕ_2, x) space. For given $\overline{V}$ and $\overline{x}$, equations (5.13) and (5.14) have a unique solution, so there is only one point $(\overline{\phi}_1, \overline{\phi}_2, \overline{x})$ corresponding to the pair $(\overline{V}, \overline{x})$. This completes the proof. □

Corollary 4. *System G has at least one fixed point.*

Proof. The proof is immediate from lemmas 16 and 17. □

Corollary 5. *Suppose there exist two fixed points f_1, f_2 of system G, with the pair (V_1, x_1) corresponding to f_1 and the pair (V_2, x_2) corresponding to f_2. If $x_1 = x_2$ then $f_1 = f_2$. Likewise, if $V_1 = V_2$ then $f_1 = f_2$.*

Proof. The result follows from the fact that the function relating V and x in equation (5.7) is a bijection. □

It is possible to show that when the Jacobian matrix of system G fulfils certain conditions, there is only one fixed point.

Let J be the Jacobian matrix of the system G as described in equations (5.1 – 5.3) on p. 82. J takes the following form:

$$J = \begin{pmatrix} -\frac{1}{R_1C_1} - \frac{g}{C_1} & \frac{g}{C_1} & -\frac{\mathrm{d}g}{\mathrm{d}x}\frac{V}{C_1} \\ \frac{g}{C_2} & -\frac{1}{R_2C_2} - \frac{g}{C_2} & \frac{\mathrm{d}g}{\mathrm{d}x}\frac{V}{C_2} \\ -x\frac{\mathrm{d}\alpha}{\mathrm{d}V} + (1-x)\frac{\mathrm{d}\beta}{\mathrm{d}V} & x\frac{\mathrm{d}\alpha}{\mathrm{d}V} - (1-x)\frac{\mathrm{d}\beta}{\mathrm{d}V} & -\alpha - \beta \end{pmatrix} \tag{5.15}$$

The following lemma will be used:

Lemma 18. *If $|J| \leq 0$ then J is a $P_0^{(-)}$ matrix. If $|J| < 0$ then J is a $P^{(-)}$ matrix.*

Proof. For notational convenience, define the new variables $\mathcal{Z} = -x\alpha + (1-x)\beta$ and $\mathcal{Z}' = \frac{\partial \mathcal{Z}}{\partial V}$.

Explicit expansion and simplification of the determinant of the Jacobian gives

$$|J| = \frac{1}{C_1C_2}\left[-\mathcal{Z}'V\frac{\mathrm{d}g}{\mathrm{d}x}\left(\frac{1}{R_1} + \frac{1}{R_2}\right) - \frac{(\alpha+\beta)}{R_1R_2} - g(x)(\alpha+\beta)\left(\frac{1}{R_1} + \frac{1}{R_2}\right)\right] \tag{5.16}$$

For brevity, re-write the Jacobian as

$$J = \begin{pmatrix} -\frac{g}{C_1} - \frac{r}{C_1} & \frac{g}{C_1} & \frac{t}{C_1} \\ \frac{g}{C_2} & -\frac{g}{C_2} - \frac{u}{C_2} & -\frac{t}{C_2} \\ \mathcal{Z}' & -\mathcal{Z}' & -w \end{pmatrix} \tag{5.17}$$

Here, $r = \frac{1}{R_1}$, $u = \frac{1}{R_2}$, $t = -(g_1 - g_2)V$ and $w = \alpha + \beta$. Note that $g, r, u, w > 0$ (within the trapping region $\mathcal{T}'$, $t > 0$ also, although this is not important for the proof). The only variable with unknown sign is $\mathcal{Z}'$.

In this new notation, the determinant is

$$|J| = \frac{1}{C_1C_2}\left(rt\mathcal{Z}' + tu\mathcal{Z}' - grw - guw - ruw\right) \tag{5.18}$$

As stated in §1.4.3 of chapter 1, J is a $P^{(-)}$ matrix if and only if its principal minors of dimension k have sign $(-1)^k$. The diagonal elements of J are negative by inspection, and so the one-dimensional principal minors of J have the correct sign.

Let M_{ij} be the two-dimensional principal submatrices of J containing rows and columns from the set $\{i, j\}$. In the notation introduced in equation (5.17), the three two-dimensional principal minors of J are

$$\begin{aligned} |M_{12}| &= \frac{ru}{C_1C_2} + \frac{gr}{C_1C_2} + \frac{gu}{C_1C_2} && (5.19) \\ |M_{13}| &= \frac{gw}{C_1} + \frac{rw}{C_1} - \frac{t\mathcal{Z}'}{C_1} && (5.20) \\ |M_{23}| &= \frac{gw}{C_2} + \frac{uw}{C_2} - \frac{t\mathcal{Z}'}{C_2} && (5.21) \end{aligned}$$

The first of these is always positive, by inspection.

Substituting the expression for $|J|$ from equation (5.18) into $|M_{13}|$ and $|M_{23}|$ yields

$$\begin{aligned} |M_{13}| &= \frac{r^2w}{C_1(r+u)} - \frac{C_2\,|J|}{r+u} \\ |M_{23}| &= \frac{u^2w}{C_2(r+u)} - \frac{C_1\,|J|}{r+u} \end{aligned}$$

When $|J| \leq 0$ both of these expressions are > 0 by inspection. Hence when $|J| \leq 0$, all one dimensional principal minors of J are negative and all two dimensional principal minors of J are positive. Since $|J|$ is itself a principal minor of J, the only case where **all** principal minors of J are of the correct sign for J to be a $P^{(-)}$ matrix is when $|J| < 0$. However, all principal minors of J except $|J|$ still have the correct sign even when $|J| = 0$, in which case J is a $P_0^{(-)}$ matrix. □

A sufficient condition for system G to have a unique fixed point then follows:

Theorem 25. *If $|J| < 0$ at all points in the trapping region $\mathcal{T}'$, system G has exactly one fixed point.*

Proof. When J is a $P^{(-)}$ matrix, the vector field of system G is injective on rectangular regions of phase space by theorem 4 (p. 21), and hence there can be at most one fixed point. From corollary 4 (p. 88) it is known that there is at least one fixed point. Hence there is exactly one fixed point when $|J| < 0$ at all points in the trapping region. □

An alternative proof of theorem 25, using degree theory, runs as follows:

Proof. The trapping region $\mathcal{T}'$ constructed in §5.2.1 is compact and convex, and the vector field points inwards at all points along its boundary. Let v be the vector of $(\dot{\phi_1}, \dot{\phi_2}, \dot{x})^T$. By definition,

$$\deg(v, \operatorname{int}(\mathcal{T}'), 0) = \sum_{v(\phi_1,\phi_2,x)=0} \operatorname{sgn}(|J(\phi_1, \phi_2, x)|)$$

Since, by assumption, $|J| < 0$ at all points in $\operatorname{int}(\mathcal{T}')$, it follows that $\deg(v, \operatorname{int}(\mathcal{T}'), 0) = -p$, where $p \in \mathbb{N}$ is the number of fixed points. However, by theorem 9 (p. 24), $\deg(v, \operatorname{int}(\mathcal{T}'), 0) = -1$. Therefore the system has exactly one fixed point. □

5.2.3 Stability

The conditions required for lemma 18 (p. 89) also relate to the stability of fixed points of the system as follows:

Theorem 26. *A fixed point of the system is locally asymptotically stable if and only if $|J| < 0$ at the fixed point, where J is the Jacobian matrix.*

Proof. A fixed point is locally asymptotically stable if and only if all the eigenvalues of the Jacobian at the fixed point have negative real part (i.e. the Jacobian is Hurwitz). This condition can be checked using the Routh-Hurwitz conditions (theorem 6, p. 22), which in three dimensions translates to the following:

$$\operatorname{Tr}(J) < 0 \tag{5.22}$$

$$|J| < 0 \tag{5.23}$$

$$\operatorname{Tr}(J) \sum |M| - |J| < 0 \tag{5.24}$$

It is clear that $\operatorname{Tr}(J) < 0$ by inspection, and so condition (5.22) is always fulfilled. To complete the proof, it needs to be shown that condition (5.23) implies condition (5.24).

Expanding the LHS of (5.24) using the symbolic algebra program [Maxima, 2008] and the notation from lemma 18 gives

$$\operatorname{Tr}(J) \sum |M| - |J| = \begin{array}{l} -w(|M_{12}| + |M_{13}| + |M_{23}|) \\ -\frac{r+u}{C_2}(|M_{12}| + |M_{23}|) - \frac{g+r}{C_1}(|M_{12}| + |M_{13}|) \\ -\frac{g}{C_2}|M_{13}| - \frac{g}{C_1}|M_{23}| - \frac{ruw}{C_1 C_2} \end{array} \tag{5.25}$$

As shown in lemma 18, the two-dimensional principal minors $|M_{ij}|$ of J are positive if $|J| < 0$. In this case, the sum of all the terms in (5.25) is negative by inspection. Hence if $|J| < 0$ then $\operatorname{Tr}(J) \sum |M| - |J| < 0$ also. □

From this point onwards, let $\alpha' \equiv \frac{\mathrm{d}\alpha}{\mathrm{d}V}$ and $\beta' \equiv \frac{\mathrm{d}\beta}{\mathrm{d}V}$. The following result links conditions on α' and β' to the sign of $|J|$, and consequently to theorems 25 and 26. It will also be shown later that opposite conditions on α' and β' guarantee monotonicity of the system.

Corollary 6. *If $\alpha' \geq 0$ and $\beta' \leq 0$ for all $V \in [0, E_1 - E_2]$, system G has a unique locally asymptotically stable fixed point.*

Proof. Recall from equation (5.16) on page 89 that

$$|J| = \frac{1}{C_1 C_2}\left[-\mathcal{Z}' V \frac{\mathrm{d}g}{\mathrm{d}x}\left(\frac{1}{R_1} + \frac{1}{R_2}\right) - \frac{(\alpha+\beta)}{R_1 R_2} - g(x)(\alpha+\beta)\left(\frac{1}{R_1} + \frac{1}{R_2}\right)\right]$$

$\mathcal{Z}' = -x\alpha' + (1-x)\beta'$, which is nonpositive in the trapping region $\mathcal{T}'$ by assumption. $V \geq 0$ in $\mathcal{T}'$, $\frac{\mathrm{d}g}{\mathrm{d}x} < 0$, and all the other variables and parameters are known to be positive. Therefore $|J| < 0$ at all points in $\mathcal{T}'$; hence by theorem 25 there is a unique fixed point, and by theorem 26 this fixed point is locally asymptotically stable. □

Remark. In physical terms, the conditions on α' and β' in corollary 6 mean that an increase in the potential difference between the cells cannot decrease the probability per unit time of channels in the gap junction going from the closed to the open state; nor can it increase the probability per unit time of channels going from the open to the closed state. Consequently an increase in the trans-junctional voltage would increase the conductivity of the gap junction. This is opposite to the experimentally observed behaviour, but the result is interesting nonetheless.

5.2.4 Graphical interpretation of stability criteria

The following theorem shows how local stability can be directly determined by consideration of the relationship between x and V as described in equations (5.6) and (5.7). The conditions required for the theorem below follow from the result of theorem 26.

Theorem 27. *At a fixed point, the following statements hold:*

1. $\frac{\mathrm{d}\mathcal{X}}{\mathrm{d}V} < \frac{\mathrm{d}\mathcal{V}}{\mathrm{d}V} \Leftrightarrow |J| < 0$
2. $\frac{\mathrm{d}\mathcal{X}}{\mathrm{d}V} = \frac{\mathrm{d}\mathcal{V}}{\mathrm{d}V} \Leftrightarrow |J| = 0$
3. $\frac{\mathrm{d}\mathcal{X}}{\mathrm{d}V} > \frac{\mathrm{d}\mathcal{V}}{\mathrm{d}V} \Leftrightarrow |J| > 0$

Proof. To prove this, it suffices to show that $\frac{\mathrm{d}\mathcal{X}}{\mathrm{d}V} - \frac{\mathrm{d}\mathcal{V}}{\mathrm{d}V} = k\,|J|$, for some $k > 0$, at all fixed points.

$$\frac{\mathrm{d}\mathcal{X}}{\mathrm{d}V} = \left(\frac{\beta}{\alpha+\beta}\right)' = \frac{\alpha\beta' - \alpha'\beta}{(\alpha+\beta)^2} \tag{5.26}$$

$$\frac{\mathrm{d}\mathcal{V}}{\mathrm{d}V} = -\frac{E_1 - E_2}{V^2(R_1+R_2)(g_1-g_2)} \tag{5.27}$$

At a fixed point, from equations (5.6) and (5.26),

$$\begin{aligned} \left(\frac{\beta}{\alpha+\beta}\right)' &= \frac{1}{\alpha+\beta}\left[\frac{\alpha\beta'}{(\alpha+\beta)} - \frac{\alpha'\beta}{(\alpha+\beta)}\right] \\ &= \frac{1}{\alpha+\beta}[-\alpha' x + \beta'(1-x)] \\ &= \frac{\mathcal{Z}'}{\alpha+\beta} \end{aligned}$$

Consequently, again at a fixed point,

$$\frac{\mathrm{d}\mathcal{X}}{\mathrm{d}V} - \frac{\mathrm{d}\mathcal{V}}{\mathrm{d}V} = \frac{\mathcal{Z}'}{\alpha+\beta} + \frac{E_1 - E_2}{V^2(R_1+R_2)(g_1-g_2)} \tag{5.28}$$

By rearranging equation (5.16), the determinant of the Jacobian can be reformulated as

$$|J| = -\frac{(g_1-g_2)(\alpha+\beta)V(R_1+R_2)}{R_1C_1R_2C_2}\left[\frac{\mathcal{Z}'}{\alpha+\beta} + \frac{x}{V} + \frac{1+g_2(R_1+R_2)}{(g_1-g_2)V(R_1+R_2)}\right] \tag{5.29}$$

To simplify the notation, define

$$m = -\frac{(g_1-g_2)(\alpha+\beta)V(R_1+R_2)}{R_1C_1R_2C_2} \tag{5.30}$$

$$S = \frac{\mathcal{Z}'}{\alpha+\beta} + \frac{x}{V} + \frac{1+g_2(R_1+R_2)}{(g_1-g_2)V(R_1+R_2)} \tag{5.31}$$

so that $|J| = mS$.

Note that $m > 0$ within the trapping region. By substituting equation (5.10) into equation (5.31), an expression for S at fixed points can be derived:

$$S = \frac{\mathcal{Z}'}{\alpha+\beta} + \frac{E_1 - E_2}{V^2(R_1+R_2)(g_1-g_2)} \tag{5.32}$$

This final expression is identical to equation (5.28), and so $\frac{\mathrm{d}\mathcal{X}}{\mathrm{d}V} - \frac{\mathrm{d}\mathcal{V}}{\mathrm{d}V} = k\,|J|$, with $k = \frac{1}{m}$. □

The conditions in theorem 27 can be observed graphically in figure 5.5. At points p_2 and p_4, condition *1* holds, so these fixed points are locally asymptotically stable. At point p_3, condition *3* holds, and consequently this fixed point is unstable. At point p_1, condition *2* holds. By definition, this point is a non-hyperbolic fixed point and its stability cannot be determined from the Jacobian. Point p_1 is a bifurcation point; the dotted lines on either side of $\mathcal{X}$ demonstrate possible perturbations of a hypothetical bifurcation parameter, resulting in the destruction of p_1 on one side, and the splitting of p_1 into a stable and unstable fixed point on the other side.

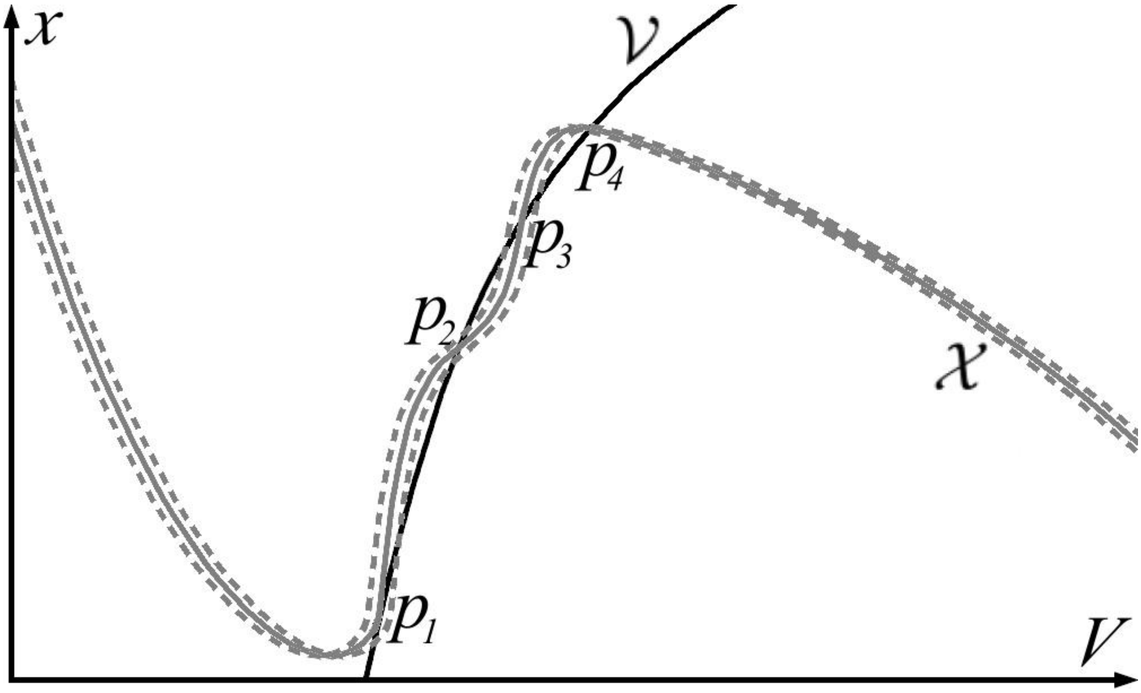

Figure 5.5: The three dimensional gap junction model represented in two dimensions at a saddle node bifurcation. The solid black line corresponds to equation (5.7), which is related to $\dot{\phi}_1 = \dot{\phi}_2 = 0$. The grey lines correspond to the x null cline, and points of intersection between the black line and each of the grey lines occur at fixed points of the dynamical system. The solid grey line corresponds to a hypothetical set of parameter values that result in a saddle node, labelled as p_1. Of the other points of intersection, p_2 and p_4 correspond to stable fixed points, and p_3 corresponds to an unstable fixed point. The dotted grey lines on either side represent perturbations of some hypothetical bifurcation parameter: on one side the non-hyperbolic fixed point p_1 disappears, and on the other side it splits into a stable and an unstable fixed point.

Notice that in the generic case, when there are no non-hyperbolic fixed points, there will be an odd number, say $2c+1$, of fixed points. Of these, c will be unstable fixed points and $c+1$ will be stable fixed points. This follows from the graphical conditions above, but can also be proved directly using degree theory:

Lemma 19. *If all the fixed points of system G are hyperbolic and there are c unstable hyperbolic fixed points, where $c \in \mathbb{Z}_{\geq 0}$, then there are $c+1$ stable hyperbolic fixed points.*

Proof. Recall that the zeros of v are in one-to-one correspondence with the fixed points of the flow defined by v. By theorem 9 (p. 24), $\deg(v, \mathrm{int}(\mathcal{T}'), 0) = -1$; consequently, if there are c fixed points at which $|J| > 0$, there must be $c + 1$ fixed points at which $|J| < 0$. As shown in theorem 26 (p. 91), when $|J| < 0$ at a fixed point, the fixed point is locally asymptotically stable. Conversely, at any unstable hyperbolic fixed point $|J| > 0$; otherwise the fact that $|J| = \prod_i \lambda_i$ for $\lambda_i \in \sigma(J)$ leads to a contradiction with theorem 26. Consequently there must be an odd number $2c+1$ of hyperbolic fixed points, of which $c + 1$ are stable and c are unstable. □

Remark. If system G has any non-hyperbolic fixed points then 0 is a critical value (i.e. not regular) of the vector field v, and therefore equation (1.7) cannot be used to calculate the degree.

5.2.5 Monotonicity

The theory of monotone flows presented in chapter 2 can be applied to system G. Interestingly, when the conditions on α' and β' given in corollary 6 on page 92 are reversed, flows of system G preserve an ordering.

Theorem 28. *Trajectories of the system within the half space defined by $\phi_1 \geq \phi_2$ are monotone with respect to a simplicial cone if $\alpha' \leq 0$ and $\beta' \geq 0$.*

Proof. To prove that the dynamical system preserves a simplicial cone, it suffices to show that there exists a similarity transform T such that the transformed Jacobian is quasipositive with respect to the standard orthant ordering, i.e. all of its offdiagonal elements are nonnegative.

When $R_1C_1 \geq R_2C_2$, define the transformation matrix T

$$T = \begin{pmatrix} C_1 & C_2 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5.33}$$

The system can then be re-coordinatised to give

$$J_{new} := TJT^{-1} \tag{5.34}$$

$$= \begin{pmatrix} \frac{1}{C_1+C_2}\left(\frac{1}{R_1}+\frac{1}{R_2}\right) & \frac{R_1C_1-R_2C_2}{(C_1+C_2)R_1R_2} & 0 \\ \frac{R_1C_1-R_2C_2}{C_1C_2(C_1+C_2)R_1R_2} & -\frac{1}{C_1+C_2}\left(\frac{C_1(gR_2+1)}{R_2C_2}+\frac{C_2(gR_1+1)}{R_1C_1}+2g\right) & (g_2-g_1)V\left(\frac{1}{C_1}+\frac{1}{C_2}\right) \\ 0 & -x\alpha' + (1-x)\beta' & -\alpha-\beta \end{pmatrix}$$

The offdiagonal elements other than $J_{new}[3,2]$ are nonnegative by inspection, since $V > 0$ inside the trapping region. $J_{new}[3,2]$ is nonnegative by assumption, as $\alpha' \leq 0$ and $\beta' \geq 0$.

If $R_1C_1 < R_2C_2$, instead define T to be

$$T = \begin{pmatrix} -C_1 & -C_2 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The same argument as for $R_1C_1 \geq R_2C_2$ applies. □

Theorem 28 leads to the final result on the two cell gap junction model described by system G, which describes a strengthening of the conditions required for theorem 28 that guarantee global asymptotic stability:

Theorem 29. *Suppose that system G satisfies $\alpha' < 0$, $\beta' > 0$, and that $R_1C_1 \neq R_2C_2$. Then every trajectory of the system converges to a fixed point.*

Proof. The proof follows from the main theorem presented in [Smillie, 1984]. This result states that for a differential equation f defined on a subset of $\mathbb{R}^n$ with tridiagonal Jacobian Df, if the elements on the sub- and super-diagonal of Df are strictly positive, the limit of every trajectory with compact closure exists and is a fixed point.

In the definition of system G it was assumed that $\alpha(V), \beta(V) \neq 0$. Combining this with lemma 15 (p. 83) guarantees that the vector field of system G points into the interior of $\mathcal{T}'$ at every point in $\partial\mathcal{T}'$. Since $\mathcal{T}'$ is globally attracting, every trajectory eventually enters $\text{int}(\mathcal{T}')$, and cannot subsequently leave. Therefore only the properties of the system in $\text{int}(\mathcal{T}')$ need be considered.

Since it was demonstrated in theorem 28 that there is a coordinate transform that makes J cooperative with respect to the standard orthant ordering, it suffices to show that every trajectory of this transformed system converges to a fixed point. As demonstrated by equation (5.34), the transformed Jacobian is tridiagonal. $V > 0$ in $\text{int}(\mathcal{T}')$, and by assumption $R_1C_1 \neq R_2C_2$ and $\alpha' < 0$, $\beta' > 0$, so all elements on the subdiagonal and superdiagonal of J_{new} are strictly positive. Every trajectory has compact closure in $\mathcal{T}'$, and therefore every trajectory converges to a fixed point. □

5.3 The model in higher dimensions

Some attempt has been made to extend the model, both by considering more than two cells joined by gap junctions, and by assuming that the conduction channels between cells have three or more possible states. However, these generalisations significantly increase the complexity of the model, and the results obtained are more limited.

Consider the following system of equations, representing n cells in a line, connected by 2-state gap junctions:

$$\left.\begin{array}{rcl}\dot{\phi}_1 & = & -\frac{\phi_1 - E_1}{R_1 C_1} - \frac{g_1(x_1)}{C_1}(\phi_1 - \phi_2) \\ \dot{\phi}_i & = & -\frac{\phi_i - E_i}{R_i C_i} + \frac{g_{i-1}(x_{i-1})}{C_i}(\phi_{i-1} - \phi_i) - \frac{g_i(x_i)}{C_i}(\phi_i - \phi_{i+1}), i = 2, \dots, n-1 \\ \dot{\phi}_n & = & -\frac{\phi_n - E_n}{R_n C_n} + \frac{g_{n-1}(x_{n-1})}{C_n}(\phi_{n-1} - \phi_n) \\ \dot{x}_j & = & -x_j \alpha_j(\phi_j - \phi_{j+1}) + (1 - x_j)\beta_j(\phi_j - \phi_{j+1}), j = 1, \dots, n-1 \end{array}\right\} \quad (5.35)$$

As in the two-cell case, $\phi_i \in \mathbb{R}$ for all $i \in \{1, \dots, n\}$ and $x_j \in [0, 1]$ for all $j \in \{1, \dots, n-1\}$. The functions g_i, α_i and β_i are equivalent to the functions g, α and β from system G for the ith gap junction. Equations (5.35) defined on the space $\mathbb{R}^n \times [0, 1]^{n-1}$ will collectively be referred to as "system H".

Lemma 20. *All trajectories in system H enter the closed compact invariant set* $B'_H = [\min_i E_i, \max_i E_i]^n \times [0, 1]^{n-1}$.

Proof. The proof is a simple extension of lemma 14 (p. 83). When $\phi_1 \leq \min E_i$ and $\phi_1 \leq \phi_2$, $\dot{\phi}_1 \geq 0$, with equality holding if and only if $\phi_1 = \phi_2 = E_1 = \min_i E_i$. Likewise, when $\phi_1 \geq \max_i E_i$ and $\phi_1 \geq \phi_2$, $\dot{\phi}_1 \leq 0$, with equality holding if and only if $\phi_1 = \phi_2 = E_1 = \max_i E_i$.

Similar conditions apply for ϕ_i ($i = 2, \dots, n-1$): When $\phi_i \leq \min_j E_j$, $\phi_i \leq \phi_{i-1}$ and $\phi_i \leq \phi_{i+1}$, $\dot{\phi}_i \geq 0$. Equality in this case holds if and only if $\phi_{i-1} = \phi_i = \phi_{i+1} = E_i = \min_j E_j$. When $\phi_i \geq \max_j E_j$, $\phi_i \geq \phi_{i-1}$ and $\phi_i \geq \phi_{i+1}$, $\dot{\phi}_i \leq 0$, with equality holding only for $\phi_{i-1} = \phi_i = \phi_{i+1} = E_i = \max_j E_j$.

When $\phi_{n-1} \leq \min_i E_i$ and $\phi_n \leq \phi_{n-1}$, $\dot{\phi}_n \geq 0$; equality holds if and only if $\phi_n = \phi_{n-1} = E_n = \min_i E_i$. When $\phi_n \geq \max_i E_i$ and $\phi_n \geq \phi_{n-1}$, $\dot{\phi}_n \leq 0$, with equality if and only if $\phi_n = \phi_{n-1} = E_n = \max_i E_i$.

In a similar way to the two cell model, an infinite family of cubes in $\phi_i \times \ldots \times \phi_n$ space which act as the level sets of a Lyapunov function can be constructed, attracting all trajectories in ϕ space into an n-dimensional cube $B_H = [\min_i E_i, \max_i E_i]^n$. As in the two cell case, the conditions are independent of the value of x_j. Since $x_j \in [0, 1]$ for all j, it follows that all trajectories enter the set $B'_H = [\min_i E_i, \max_i E_i]^n \times [0, 1]^{n-1}$, and this set is invariant.

B'_H is closed since it is the product of a finite number of closed sets, and it is also bounded; therefore it is compact. □

Remark. In order to keep the notation simple, system H was constructed under the assumption that the gap junctions between cells have only two possible states. However, it would be straightforward to construct a model with three or more states per gap junction. Since lemma 20 relies solely on the fact that $g_i > 0$, its conclusions would remain valid for a model with m conduction states per gap junction.

Unfortunately, the presence of both $\phi_{i-1} - \phi_i$ and $\phi_i - \phi_{i+1}$ terms in the equation(s) for $\dot{\phi}_i$ make it much more difficult to construct a trapping region in which $\phi_{i-1} - \phi_i \geq 0$. At present, no such result has been found, so there is no apparent generalisation of lemma 15 (p. 83) and consequently no conditions for monotonicity have been found in the general case. However, some limited results extending the two cell model can be constructed.

Lemma 21. *System H has at least one fixed point.*

Proof. Since all trajectories enter the invariant convex compact trapping region B'_H, system H has at least one fixed point as a corollary of the Brouwer fixed point theorem (see the paragraph following the statement of the theorem on p. 17), or by theorem 3 (p. 17). □

5.3.1 Three cells in a line

The remaining results presented in this chapter are focused on a model with three cells in a line, each joined by a two state gap junction. Following the set of equations (5.35), the evolution of this system is governed by the set of equations

$$\left.\begin{array}{rcl}
\dot{\phi}_1 & = & -\frac{\phi_1 - E_1}{R_1 C_1} - \frac{g_1(x_1)}{C_1}(\phi_1 - \phi_2) \\
\dot{\phi}_2 & = & -\frac{\phi_2 - E_2}{R_2 C_2} + \frac{g_1(x_1)}{C_2}(\phi_1 - \phi_2) - \frac{g_2(x_2)}{C_2}(\phi_2 - \phi_3) \\
\dot{\phi}_3 & = & -\frac{\phi_3 - E_3}{R_3 C_3} + \frac{g_2(x_2)}{C_3}(\phi_2 - \phi_3) \\
\dot{x}_1 & = & -x_1\alpha_1(\phi_1 - \phi_2) + (1 - x_1)\beta_1(\phi_1 - \phi_2) \\
\dot{x}_2 & = & -x_2\alpha_2(\phi_2 - \phi_3) + (1 - x_2)\beta_2(\phi_2 - \phi_3)
\end{array}\right\} \tag{5.36}$$

The state space of the model is $\mathbb{R}^3 \times [0,1]^2$, and the set of equations (5.36) defined on this state space will be referred to as system H_3. As demonstrated above, the hypercube $[\min_i E_i, \max_i E_i]^3 \times [0,1]^2$ is a globally absorbing set for this system, and it contains at least one fixed point. Under certain physically reasonable conditions, as in the two cell case, it is possible to show that there can only be one fixed point. To make the notation more readable, define the variables $V_1 = \phi_1 - \phi_2$, $V_2 = \phi_2 - \phi_3$, $\mathcal{Z}_1 = \dot{x}_1$ and $\mathcal{Z}_2 = \dot{x}_2$. Similarly, let $\alpha'_i = \frac{\mathrm{d}\alpha_i}{\mathrm{d}V_i}$, $\beta'_i = \frac{\mathrm{d}\beta_i}{\mathrm{d}V_i}$, $\mathcal{Z}'_i = \frac{\partial \mathcal{Z}_i}{\partial V_i}$ and $g'_i = \frac{\mathrm{d}g_i}{\mathrm{d}x_i}$. Note that $g'_i < 0$, so to make it easier to check the signs of terms later on the functions $G_i \equiv -g'_i$ are also defined.

The Jacobian of system H_3 can then be written as follows:

$$J_{H_3} = \begin{pmatrix}
-\frac{1}{R_1 C_1} - \frac{g_1}{C_1} & \frac{g_1}{C_1} & 0 & \frac{G_1 V1}{C_1} & 0 \\
\frac{g_1}{C_2} & -\frac{1}{R_2 C_2} - \frac{g_1}{C_2} - \frac{g_2}{C_2} & \frac{g_2}{C_2} & -\frac{G_1 V_1}{C_2} & \frac{G_2 V_2}{C_2} \\
0 & \frac{g_2}{C_3} & -\frac{1}{R_3 C_3} - \frac{g_2}{C_3} & 0 & -\frac{G_2 V_2}{C_3} \\
\mathcal{Z}'_1 & -\mathcal{Z}'_1 & 0 & -\alpha_1 - \beta_1 & 0 \\
0 & \mathcal{Z}'_2 & -\mathcal{Z}'_2 & 0 & -\alpha_2 - \beta_2
\end{pmatrix} \tag{5.37}$$

The lemma that follows assumes that increasing the **magnitude** of the potential difference between cells cannot (a) decrease the probability per unit time of a conduction channel between the cells going from a closed to an open state or (b) increase the probability per unit time of a channel going from an open to a closed state. As with corollary 6 this is the opposite of what would be expected from experimental data, so the result is more a curiosity than an accurate picture of how the real system might behave.

Lemma 22. *If $\alpha_i' V_i \geq 0$ and $\beta_i' V_i \leq 0$ then $|J_{H_3}| < 0$ at every point and system H_3 has a unique equilibrium.*

Proof. The conditions $\alpha_i' V_i \geq 0$ and $\beta_i' V_i \leq 0$ guarantee that $V_i \mathcal{Z}_i' < 0$. Using a symbolic algebra program such as [Maxima, 2008] and setting $r_i \equiv 1/R_i$,

$$\begin{aligned}
C_1C_2C_3|J_{H_3}| = \quad & -(r_3G_1G_2V_1V_2\mathcal{Z}_1'\mathcal{Z}_2' + r_2G_1G_2V_1V_2\mathcal{Z}_1'\mathcal{Z}_2' + r_1G_1G_2V_1V_2\mathcal{Z}_1'\mathcal{Z}_2' \\
& -r_1r_3G_2\beta_1V_2\mathcal{Z}_2' - g_1r_3G_2\beta_1V_2\mathcal{Z}_2' - r_1r_2G_2\beta_1V_2\mathcal{Z}_2' - g_1r_2G_2\beta_1V_2\mathcal{Z}_2' \\
& -g_1r_1G_2\beta_1V_2\mathcal{Z}_2' - \alpha_1r_1r_3G_2V_2\mathcal{Z}_2' - \alpha_1g_1r_3G_2V_2\mathcal{Z}_2' - \alpha_1r_1r_2G_2V_2\mathcal{Z}_2' \\
& -\alpha_1g_1r_2G_2V_2\mathcal{Z}_2' - \alpha_1g_1r_1G_2V_2\mathcal{Z}_2' - r_2r_3G_1\beta_2V_1\mathcal{Z}_1' - r_1r_3G_1\beta_2V_1\mathcal{Z}_1' \\
& -g_2r_3G_1\beta_2V_1\mathcal{Z}_1' - g_2r_2G_1\beta_2V_1\mathcal{Z}_1' - g_2r_1G_1\beta_2V_1\mathcal{Z}_1' - \alpha_2r_2r_3G_1V_1\mathcal{Z}_1' \\
& -\alpha_2r_1r_3G_1V_1\mathcal{Z}_1' - \alpha_2g_2r_3G_1V_1\mathcal{Z}_1' - \alpha_2g_2r_2G_1V_1\mathcal{Z}_1' - \alpha_2g_2r_1G_1V_1\mathcal{Z}_1' \\
& +r_1r_2r_3\beta_1\beta_2 + g_1r_2r_3\beta_1\beta_2 + g_2r_1r_3\beta_1\beta_2 + g_1r_1r_3\beta_1\beta_2 + g_1g_2r_3\beta_1\beta_2 \\
& +g_2r_1r_2\beta_1\beta_2 + g_1g_2r_2\beta_1\beta_2 + g_1g_2r_1\beta_1\beta_2 + \alpha_1r_1r_2r_3\beta_2 + \alpha_1g_1r_2r_3\beta_2 \\
& +\alpha_1g_2r_1r_3\beta_2 + \alpha_1g_1r_1r_3\beta_2 + \alpha_1g_1g_2r_3\beta_2 + \alpha_1g_2r_1r_2\beta_2 + \alpha_1g_1g_2r_2\beta_2 \\
& +\alpha_1g_1g_2r_1\beta_2 + \alpha_2r_1r_2r_3\beta_1 + \alpha_2g_1r_2r_3\beta_1 + \alpha_2g_2r_1r_3\beta_1 + \alpha_2g_1r_1r_3\beta_1 \\
& +\alpha_2g_1g_2r_3\beta_1 + \alpha_2g_2r_1r_2\beta_1 + \alpha_2g_1g_2r_2\beta_1 + \alpha_2g_1g_2r_1\beta_1 + \alpha_1\alpha_2r_1r_2r_3 \\
& +\alpha_1\alpha_2g_1r_2r_3 + \alpha_1\alpha_2g_2r_1r_3 + \alpha_1\alpha_2g_1r_1r_3 + \alpha_1\alpha_2g_1g_2r_3 + \alpha_1\alpha_2g_2r_1r_2 \\
& +\alpha_1\alpha_2g_1g_2r_2 + \alpha_1\alpha_2g_1g_2r_1)
\end{aligned}$$

By looking at this expression, it is straightforward to verify that $V_i\mathcal{Z}_i' < 0$ is sufficient to guarantee $|J_{H_3}| < 0$: every variable other than V_i and $\mathcal{Z}_i'$ is positive, $|J_{H_3}|$ contains no terms with V_i or $\mathcal{Z}_i'$ by themselves, only terms in $V_i\mathcal{Z}_i'$. Thus every term inside the brackets is positive, and hence $C_1C_2C_3|J_{H_3}| < 0$. This concludes the first part of the lemma.

Degree theory is used to prove the second part of the result. Let $v(\phi_1, \phi_2, \phi_3, x_1, x_2) = (\dot\phi_1, \dot\phi_2, \dot\phi_3, \dot x_1, \dot x_2)^t$. Then, by theorem 9 (p. 24), $d(v, \text{int}(B_{H_3}), 0) = -1$. Since by definition

$$d(v, \text{int}(B_{H_3}), 0) = \sum_{v=0} \text{sgn}|J_{H_3}|$$

and $\text{sgn}|J_{H_3}| = -1$ at all points in B_{H_3}, it follows that B_{H_3} contains one fixed point. □

The conditions for lemma 22 can be strengthened to guarantee that the fixed point is locally asymptotically stable.

Theorem 30. *Suppose that the following conditions hold:*

1. $\alpha_i' V_i \geq 0$, $\beta_i' V_i \leq 0$.
2. $\alpha_i' \neq 0$, $\beta_i' \neq 0$ *when* $V_i \neq 0$.

Then J_{H_3} *is everywhere Hurwitz.*

Proof. Before beginning the main part of the proof, it needs to be shown that $\alpha_i' = \beta_i' = 0$ when $V_i = 0$. Recall that $\alpha_i = \alpha_i(V_i)$ and $\beta_i = \beta_i(V_i)$ are C^1 functions. Condition 1 implies that $\alpha_i'(V_i) \leq 0$ when $V_i < 0$, which when combined with condition 2 means that $\alpha_i'(V_i) < 0$ when $V_i < 0$. Likewise, $\alpha_i'(V_i) > 0$ when $V_i > 0$. Since $\alpha_i(V_i)$ is C^1, the derivative $\alpha_i'(V_i)$ is continuous in V_i. Therefore $\alpha_i'(0) = 0$ by continuity. A similar argument shows that $\beta_i'(0) = 0$. This completes the preliminary result.

There are then four possible cases, for which the proof of Hurwitz stability is slightly different. The first and simplest case is when $V_1 = V_2 = 0$.

Since $\alpha_i'(0) = \beta_i'(0) = 0$, it follows that $\mathcal{Z}_i = 0$ when $V_i = 0$. Therefore J_{H_3} reduces to a tridiagonal matrix at all points where $V_1 = V_2 = 0$, and can easily be shown to be Hurwitz via theorem 7 (p. 23).

The remaining cases use Lyapunov's second theorem (stated earlier as corollary 1 on p. 23): J_{H_3} is Hurwitz at a point if and only if there exists a positive definite matrix Q such that $QJ_{H_3} + J_{H_3}^T Q$ is negative definite. Start by looking at the points where $V_1 \neq 0$ and $V_2 \neq 0$. Consider the matrix

$$Q = \begin{pmatrix} C_1 & 0 & 0 & 0 & 0 \\ 0 & C_2 & 0 & 0 & 0 \\ 0 & 0 & C_3 & 0 & 0 \\ 0 & 0 & 0 & -\frac{G_1 V_1}{\mathcal{Z}_1'} & 0 \\ 0 & 0 & 0 & 0 & -\frac{G_2 V_2}{\mathcal{Z}_2'} \end{pmatrix}$$

Combining condition 1 with condition 2 implies that when $V_i \neq 0$, $\frac{\alpha_i'}{V_i} > 0$ and $\frac{\beta_i'}{V_i} < 0$. This in turn means that $-\frac{G_i V_i}{\mathcal{Z}_i'} > 0$, and therefore Q is positive definite.

It is straightforward to show that

$$\begin{matrix} QJ_{H_3} \\ +J_{H_3}^T Q \end{matrix} = 2 \begin{pmatrix} -\frac{1}{R_1} - g_1 & g_1 & 0 & 0 & 0 \\ g_1 & -\frac{1}{R_2} - g_1 - g_2 & g_2 & 0 & 0 \\ 0 & g_2 & -\frac{1}{R_3} - g_2 & 0 & 0 \\ 0 & 0 & 0 & \frac{G_1(\alpha_1+\beta_1)V_1}{\mathcal{Z}_1'} & 0 \\ 0 & 0 & 0 & 0 & \frac{G_2(\alpha_2+\beta_2)V_2}{\mathcal{Z}_2'} \end{pmatrix}$$

Since $QJ_{H_3}+J_{H_3}^TQ$ is symmetric, its eigenvalues are real. That they are also negative can easily be verified using theorem 7 (p. 23). Therefore $QJ_{H_3}+J_{H_3}^TQ$ is negative definite and hence J_{H_3} is Hurwitz at every point where $V_1, V_2 \neq 0$.

The remaining two cases are slight variants of the preceding case where $V_1 \neq 0$ and $V_2 \neq 0$. The first of these is when $V_1 = 0$ but $V_2 \neq 0$. As in the first case, $\mathcal{Z}_1' = 0$ since $V_1 = 0$. In this case, let Q be the positive definite matrix

$$Q = \begin{pmatrix} C_1 & 0 & 0 & 0 & 0 \\ 0 & C_2 & 0 & 0 & 0 \\ 0 & 0 & C_3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -\frac{G_2V_2}{\mathcal{Z}_2'} \end{pmatrix}$$

Therefore

$$QJ_{H_3}+J_{H_3}^TQ = 2\begin{pmatrix} -\frac{1}{R_1}-g_1 & g_1 & 0 & 0 & 0 \\ g_1 & -\frac{1}{R_2}-g_1-g_2 & g_2 & 0 & 0 \\ 0 & g_2 & -\frac{1}{R_3}-g_2 & 0 & 0 \\ 0 & 0 & 0 & -\alpha_1-\beta_1 & 0 \\ 0 & 0 & 0 & 0 & \frac{G_2(\alpha_2+\beta_2)V_2}{\mathcal{Z}_2'} \end{pmatrix}$$

and this matrix is negative definite as in the previous case where $V_1 \neq 0, V_2 \neq 0$. Hence J_{H_3} is Hurwitz.

The final case is where $V_1 \neq 0$ but $V_2 = 0$. Then define Q to be

$$Q = \begin{pmatrix} C_1 & 0 & 0 & 0 & 0 \\ 0 & C_2 & 0 & 0 & 0 \\ 0 & 0 & C_3 & 0 & 0 \\ 0 & 0 & 0 & -\frac{G_1V_1}{\mathcal{Z}_1'} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

This gives

$$QJ_{H_3}+J_{H_3}^TQ = 2\begin{pmatrix} -\frac{1}{R_1}-g_1 & g_1 & 0 & 0 & 0 \\ g_1 & -\frac{1}{R_2}-g_1-g_2 & g_2 & 0 & 0 \\ 0 & g_2 & -\frac{1}{R_3}-g_2 & 0 & 0 \\ 0 & 0 & 0 & \frac{G_1(\alpha_1+\beta_1)V_1}{\mathcal{Z}_1'} & \\ 0 & 0 & 0 & 0 & -\alpha_2-\beta_2 \end{pmatrix}$$

Once again, this is negative definite, and so J_{H_3} is Hurwitz. Combining all these results together, it follows that J_{H_3} is Hurwitz at every point in phase space, including the fixed point. This completes the proof. □

Remark. Note that $V_1 = V_2 = 0$ at the fixed point implies that $E_1 = E_2 = E_3$. This is because $V_1 = V_2 = 0$ means $\phi_1 = \phi_2 = \phi_3$. When this relation is substituted into the first three equations of set (5.36), it is apparent that at the fixed point $\phi_1 = E_1$, $\phi_2 = E_2$ and $\phi_3 = E_3$, and therefore $E_1 = E_2 = E_3$.

The final result for this chapter is a simple corollary of the previous two results, and is similar to corollary 6 on page 92.

Corollary 7. *Suppose that system H_3 satisfies the conditions required for theorem 30. Then the system has a unique locally asymptotically stable fixed point.*

Proof. System H_3 has a unique fixed point under condition 1 of theorem 30, by lemma 22. By theorem 30, conditions 1 and 2 together guarantee that the Jacobian is Hurwitz at every point in phase space, and the fixed point is therefore locally asymptotically stable. □

Chapter 6

Chemical reaction networks

Unlike the previous two chapters, the application in this chapter is not concerned with a specific biological process. Instead, it relates to the dynamics of a network of unspecified chemical reactions. There is a significant amount of existing literature on chemical reaction networks. [Craciun and Feinberg, 2005] considers the problem of identifying conditions guaranteeing that a network of chemical reactions obeying so-called mass action kinetics can have no more than one equilibrium. [Banaji et al., 2007] is in a similar vein, but is not restricted to mass action reactions, and also considers conditions that ensure the Jacobian of a dynamical system representing a set of reactions is Hurwitz. By contrast, [Kunze and Siegel, 2002b] and [Banaji, 2008] examine chemical reaction structures that generate monotone dynamical systems, but do not do any direct analysis of the systems' asymptotic behaviour. [De Leenheer et al., 2007] also describes a set of reaction networks which are monotone, and then goes on to analyse the asymptotic behaviour of such reaction networks, including generalising the model to include reaction-diffusion systems. Reaction-diffusion systems, while closely related to chemical reaction networks, lie outside the scope of the results presented in this chapter. See e.g. [Mincheva and Siegel, 2003] and [Mincheva and Siegel, 2007] for other work regarding reaction-diffusion systems.

In a similar way to the above references, the work in this chapter is concerned with predicting the possible behaviour of a chemical reaction or set of reactions by examining the reaction **structure**, while making minimal assumptions about the **kinetics** of the reactions (see the next paragraph for clarification of the meaning of these words in the context of chemical reactions). As in the previous two chapters, the assumptions made in setting up a model of a reaction are qualitative in nature. As such, the results can be applied to any system of chemical reaction(s) with the same underlying structure for a variety of different kinetics, and are potentially of use when constructing a more complicated system that incorporates such reactions.

For an idea of what is meant by the "structure" and "kinetics" of a reaction network, consider the following two reactions:

$$\begin{aligned} 2H_2 + O_2 &\rightarrow 2H_2O \\ 2C + O_2 &\rightarrow 2CO \end{aligned}$$

The first reaction represents the burning of hydrogen in an excess of oxygen to create water, while the second reaction represents the burning of carbon in an environment where there is a shortage of oxygen, resulting in the production of carbon monoxide. The reactions are obviously not the same, but they have the same structure in the sense that two molecules of some compound "X" combine with one molecule of a compound "Y" (in this example "Y" is the same in both reactions, namely O_2) to produce two molecules of another compound "Z". The structure of a reaction in this sense is a specification of information about how many molecules of each reactant take part in each reaction.

The kinetics of a reaction describes how the reaction rate varies with reactant concentration, which is also important when constructing a model of a set of chemical reactions. The simplest type is mass action kinetics: For the reaction structure above, the rate of reaction under mass action kinetics would be $R = k(c(X))^2c(Y)$, where $k > 0$ is a constant specific to the reaction, $c(X)$ is the concentration of compound "X" and $c(Y)$ is the concentration of compound "Y". Other more complicated kinetics can be included in a model; however, the aim in this chapter is to assume as little as possible about the kinetics of each reaction, so alternative reaction kinetics will not be discussed in detail.

The focus of investigation, as in the previous two chapters, is on finding a set of conditions that guarantee global asymptotic stability. The only reference listed above that does this explicitly is [De Leenheer et al., 2007], the results of which can be summed up as follows: Suppose that there is a set of distinct chemical reactants $\{X_i\}$ in some closed chamber. Suppose that there is also a set of complexes $\{C_j\}$ such that $C_j = \sum_{i_j=1}^{n_j} S_{i_j}X_{i_j}$, where $S_{i_j} \in \mathbb{N}$ and each X_i appears in only one complex. Then the chain of chemical reactions $C_1 \rightleftharpoons \ldots \rightleftharpoons C_j \rightleftharpoons \ldots \rightleftharpoons C_n$ has a unique globally asymptotically stable equilibrium, provided that the reaction rates are monotone C^1 functions of reactant concentration, with the rate of a reaction being zero when the concentration of any reactant taking part in the reaction is zero.

The results in this chapter make different but related assumptions. The assumption that all reaction rates be monotone C^1 functions of reactant concentration, with zero rate when any reactant concentration is zero, is kept in this chapter. However, rather than a closed reaction chamber, it is assumed here that the reactions take place in a reaction chamber that allows some inflow and outflow of reactants. Additionally, more general network topologies are allowed than the chain of complexes assumed in [De Leenheer et al., 2007].

[De Leenheer et al., 2007] and all the other papers listed above that directly or indirectly make claims about the asymptotic behaviour of a set of chemical reactions do so via monotonicity (see chapter 2). This chapter likewise includes some results using monotonicity for a single reaction, but also presents results based on autonomous convergence (as described in chapter 3) for both a single reaction and networks comprised of multiple reactions. Autonomous convergence theory does not appear to have been applied to chemical reaction networks before. It is hoped that the process of analysing a model of chemical reactions using multiple techniques is of some academic interest.

6.1 Overview

The chapter begins by describing the dynamics of a chemical reaction, including assumptions about inflow and outflow of reactants and the kinetics of the reaction (i.e. how the reaction rate depends on the concentrations of the different reactants). A differential equation model of the reaction is constructed, outlining in particular what the assumptions made about the reaction mean in mathematical terms.

Once the mathematical model has been set up, it is shown that the concentrations of the reactants are globally bounded, and have a unique equilibrium value. Following from this, it is then demonstrated that under certain conditions on the outflows of the reactants, all initial conditions converge to the equilibrium. Convergence criteria are constructed using both the theory of monotone flows discussed in chapter 2 and the autonomous convergence theorems given in chapter 3.

Following the treatment of a single reaction, further results relating to models of multiple reactions with some reactants in common are presented in §6.5.

6.2 Description of a chemical reaction

The aim of this section is to outline a mathematical representation of a chemical reaction and the physical setting in which it takes place. As in the previous applications in this thesis, the aim is to make the mathematical model of the reaction as general as possible. Rather than choosing specific functional forms or numerical values for terms, assumptions are made about monotonicity of the functions used to construct the model, with the aim of proving results for a whole class of mathematical models that all have the same underlying qualitative structure. The reaction process is modelled by the equation

$$\dot{x} = I + SR(x) - Q(x) \tag{6.1}$$

The reaction is assumed to take place in a container of fixed volume, with reactants being fed in and flowing out. In simple terms, x represents the concentration of the reactants, I represents the rate at which reactants are fed in, $Q(x)$ represents the rate at which the reactants flow out, and $SR(x)$ represents the dynamics of the reaction itself. The meaning of each of these terms will be explained in more detail throughout this section.

Let $\{X_i\}, i = 1, \ldots, n$ represent the set of chemical species taking part in the reaction, the concentrations of which are denoted by corresponding variables x_i. Spatial effects are ignored, meaning that the concentration of each reactant is assumed to be homogeneous throughout the container. Reactant concentrations are assumed to be initially nonnegative, and further assumptions (to be detailed later) on the inflow, outflow and reaction rate guarantee that every reactant concentration remains nonnegative. Therefore $x \in \mathbb{R}^n_{\geq 0}$.

It is assumed that the inflow and outflow of the reaction container are both constant at rate q, and that the concentration of reactant i in the inflow is a constant c_i. Thus the rate of increase of concentration of reactant i due to inflow is $I_i = qc_i$. For brevity, this will often be referred to as "the inflow rate of reactant i". I in equation (6.1) is the vector of inflow rates. It is assumed that $I_i \geq 0$ for all i, with $I_i > 0$ for at least one value of i.

The rate of decrease of concentration of reactant i due to outflow ("outflow rate of reactant i") is dependent on its concentration in the tank, x_i. The outflow rate will be assumed to take the general form $q_i(x_i)$, where $q_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a C^1 surjection satisfying the relations $q_i(0) = 0$ and $\frac{\mathrm{d}q_i}{\mathrm{d}x_i} > 0$. These assumptions correspond to the physically reasonable requirements that the concentration of a reactant cannot become negative due to outflow, and that the rate of outflow of a reactant strictly increases with its concentration in the container. The assumption that the function q_i is surjective means that the outflow rate of reactant i cannot saturate. $Q(x)$ in equation (6.1) is the vector of outflow rates. For notational convenience, the derivative of $q_i(x_i)$ with respect to x_i will be written $q_i'(x_i)$, which will usually be further abbreviated to q_i'. Write the derivative of Q with respect to x as Q', defined by

$$Q' = \begin{pmatrix} \frac{\partial q_1}{\partial x_1} & \cdots & \frac{\partial q_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial q_n}{\partial x_1} & \cdots & \frac{\partial q_n}{\partial x_n} \end{pmatrix}$$

Since $\frac{\partial q_i}{\partial x_j} = 0 \; \forall \; i \neq j$, Q' is a diagonal matrix with q_i' along the diagonal.

$SR(x)$, the final term in equation (6.1), represents the reaction itself. It is assumed that every reactant in the model takes part in the reaction. The temperature and pressure of the system are assumed constant, so they are not included in the model of the reaction. The rate of reaction is represented by the C^1 function $R : \mathbb{R}^n_{\geq 0} \to \mathbb{R} : x \mapsto R(x)$, and in general it is assumed that $R(x)$ can be positive for some values of x and negative for other values of x. This implies that the reaction is reversible. The reaction structure is described by the stoichiometric vector S. All the entries of S are integers. Let the two sets α and

β be subsets of $\{1,\ldots,n\}$ representing the two sides of the reaction. For each $i \in \alpha$ and $j \in \beta$, $\mathrm{sgn}(S_i) = -\mathrm{sgn}(S_j)$. The reaction can thus be written

$$\sum_{i\in\alpha} |S_i| X_i \overset{R}{\rightleftharpoons} \sum_{j\in\beta} |S_j| X_j \tag{6.2}$$

It is assumed that the single reaction under consideration is a "true" reaction, in that there is at least one reactant on each side of the reaction, and hence that it does not simply represent some inflow or outflow process. This means that both α and β are nonempty. It is also assumed that each reactant appears on one side of the reaction only, which rules out catalytic reactions, and consequently $\alpha \cap \beta = \emptyset$. Since it was assumed that all the reactants take part in the reaction, $\alpha \cup \beta = \{1,\ldots,n\}$ and none of the entries of S are zero. α and β therefore form a partition of $\{1,\ldots,n\}$. The sign of S is arbitrary: the important feature is that all entries on one side of the reaction have one sign, and all entries on the other have a different sign. Changing the sign of S preserves the reaction structure as described in equation (6.2), and though changing the sign generates a different dynamical system, its behaviour is exactly the same as the system before the switch. For convenience and without loss of generality it will be assumed that $S_i/|S_i| = 1 \;\forall\; i \in \alpha$ and $S_i/|S_i| = -1 \;\forall\; i \in \beta$. The order in which reactants are listed on each side of the reaction is also arbitrary; for each $i, j \in \alpha$ the labels i and j can be switched and likewise for each $i, j \in \beta$.

Denote the partial derivatives of the reaction rate R as $V_i \equiv \frac{\partial R}{\partial x_i}$. The following assumptions are made:

1. $\mathrm{sgn}(V_i) = \mathrm{sgn}(V_j)$ whenever $i, j \in \alpha$ or $i, j \in \beta$.
2. $\mathrm{sgn}(V_i) = -\mathrm{sgn}(V_j)$ whenever $i \in \alpha$ and $j \in \beta$.
3. $S_i V_i \leq 0$.
4. For every i, $S_i R(x) \geq 0$ when $x_i = 0$.
5. $V_i < 0$ for all $i \in \alpha$ whenever $x_j > 0$ for all $j \in \alpha$.
6. $V_i > 0$ for all $i \in \beta$ whenever $x_j > 0$ for all $j \in \beta$.

Assumptions 1–3 limit the kinetics of the reaction. They mean that R is a monotone function of the concentration of each of the reactants, and that the partial derivatives of R for each reactant concentration are assumed to have opposite sign to the corresponding stoichiometry. In chemical terms, this implies that if a reactant is "used up" by the reaction, increasing its concentration cannot decrease the rate of reaction and, conversely, if a reactant is "produced" by the reaction then increasing its concentration cannot increase the rate of reaction. Of course, the definitions of "used up" and "produced" are dependent on the sign convention chosen for R. Note that this set of assumptions is satisfied by a reaction with mass action or Michaelis-Menten kinetics, among others.

Assumption 4 means that the concentration of any reactant on one side of the reaction is zero, the reaction cannot continue to use up that reactant. Assumptions 5 and 6 mean that when the concentrations of all the reactants on one side of the reaction are non-zero, increasing the concentration of a reactant on that side increases the rate at which the reactant is used up. Therefore $S_i V_i < 0$ for all $x \in \mathbb{R}^n_{>0}$.

One particular context of interest is a continuous flow stirred tank reactor (CFSTR), as described in [Craciun and Feinberg, 2005]. A diagram appears in figure 6.1. A CFSTR is a tank of fixed volume, with an inflow and outflow both at rate q. For a CFSTR the outflow rate of a reactant is assumed to be directly proportional to its concentration, and so equation (6.1) becomes

$$\dot{x} = I + SR(x) - qx \tag{6.3}$$

In the CFSTR case, $Q' = q\mathbb{I}$, where $\mathbb{I}$ is the matrix identity. Note that a CFSTR is a specialised case of the system described in equation (6.1), and that it fulfils all of the assumptions made above. Some of the results in this chapter will be proved in the context of a CFSTR rather than the more general reaction system described in equation (6.1), but it will be made clear where this is the case.

The following result will be useful later:

Lemma 23. *The nonnegative orthant of the dynamical system described in equation (6.1) with the assumptions made above is forward invariant.*

Proof. Each component of x evolves according to the equation

$$\dot{x}_i = I_i + S_i R(x) - q_i(x_i)$$

$I_i \geq 0$ at all points, since it is a constant. When $x_i = 0$, $S_i R(x) \geq 0$ and $q_i(0) = 0$, by assumption. Therefore $\dot{x}_i \geq 0$ whenever $x_i = 0$. This concludes the proof. □

Following [Banaji et al., 2007], let V be a row vector $(V_1, \ldots, V_n)$. Using the definitions above, the Jacobian can be written

$$J = SV - Q' \tag{6.4}$$

The next section is dedicated to proving that all forward trajectories are bounded, entering a convex compact set, and that there is a unique fixed point. These conclusions hold without making any further assumptions. In later sections, convergence to the fixed point will be discussed, but the proofs of convergence presented rely on extra conditions being applied to the model.

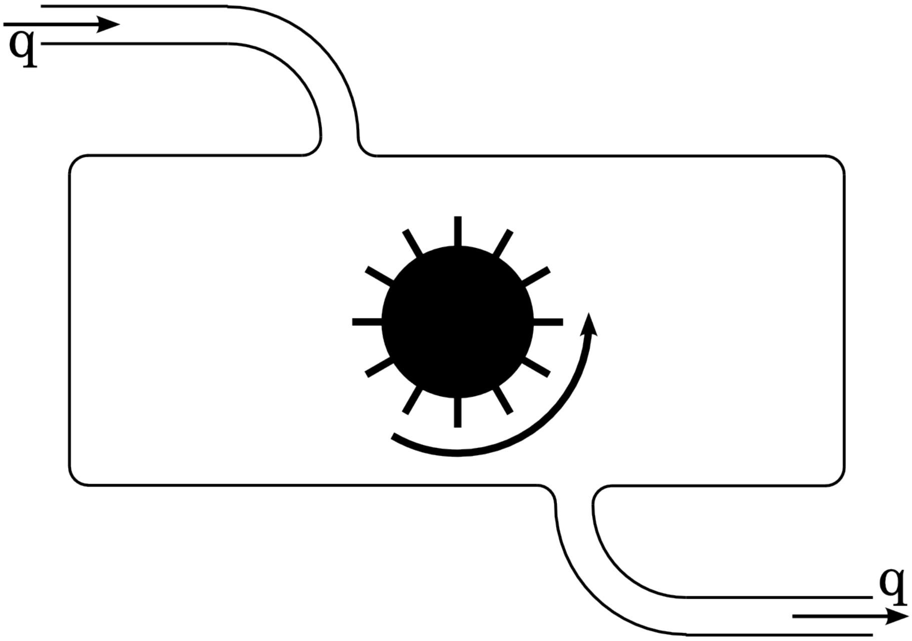

Figure 6.1: Representation of a continuous flow stirred tank reactor (CFSTR). Reactants are fed in through the pipe at the top at flow rate q. Therefore the increase in concentration of reactant i due to inflow will be $I_i = qc_i$, where c_i is the concentration of reactant i in the input feed. To ensure spatial homogeneity, the reactants are mixed inside the central tank while they react with each other. Outflow then occurs through the pipe at the bottom at rate q. The concentration of reactants in the input feed is assumed to be fixed, while their concentration in the outflow is the same as the concentration in the tank, which will generally change over time. If x_i is the concentration of reactant i in the tank, the rate at which its concentration will decrease is qx_i.

6.3 Boundedness of solutions and existence of a unique fixed point

In this section, it will be shown that all trajectories of equation (6.1) with the assumptions made in §6.2 enter a compact, convex, forward invariant set that contains a unique fixed point. While in the applications presented in previous chapters such a claim was relatively straightforward to verify, in the case of the chemical reaction model no simple proof has been found; consequently this section is rather long and technical.

The following result will be used:

Lemma 24. *There exists a vector $k \gg 0$ such that $\langle k, S \rangle = 0$.*

k is non-unique; this proof gives a constructive method of finding one possible value of k.

Proof. There exists a set of n rational numbers $\{k_i\}$ such that

$$k_i = \begin{cases} \frac{1}{|S_i||\alpha|} & \text{for} \quad i \in \alpha \\ \frac{1}{|S_i||\beta|} & \text{for} \quad i \in \beta \end{cases} \tag{6.5}$$

Choose $k = (k_1, \ldots, k_n)^T$. The inner product with S is

$$\langle k, S\rangle = \sum_{i\in\alpha} \frac{S_i}{|S_i||\alpha|} + \sum_{i\in\beta} \frac{S_i}{|S_i||\beta|} = \frac{|\alpha|}{|\alpha|} - \frac{|\beta|}{|\beta|} = 0$$

□

It will also be helpful to define a scalar function $\mathcal{X}(x)$ using k:

$$\mathcal{X}(x) = \sum_{i=1}^{n} k_i x_i = \langle k, x\rangle \tag{6.6}$$

The set of points $x \in \mathbb{R}^n$ satisfying $\mathcal{X}(x) = 0$ forms a hyperplane normal to k and containing the origin, which will be labelled $\mathcal{P}_0$. This hyperplane divides $\mathbb{R}^n$ into two half spaces, $H_+ = \{x \,|\langle x, k\rangle \geq 0\}$ and $H_- = \{x \,|\langle x, k\rangle \leq 0\}$. Let c be a positive real number, and $\mathcal{P}_c = \mathcal{P}_0 + ck$ be a coset of $\mathcal{P}_0$. The set of all such cosets of $\mathcal{P}_0$ forms an infinite family of hyperplanes normal to k. Define the set $\mathcal{C}(\mathcal{P}_c) = \mathbb{R}^n_{\geq 0} \cap (H_- + ck)$, which is a simplex with one vertex at the origin and n vertices at the points of intersection between $\mathcal{P}_c$ and the coordinate axes. An illustrative example of $\mathcal{C}(\mathcal{P}_c)$ in two dimensions appears in figure 6.2. The claim made at the start of this section will be verified by proving that there exists some $c > 0$ such that $\mathcal{C}(\mathcal{P}_c)$ is globally absorbing.

The first result required to show that all trajectories enter a compact convex set is this:

Lemma 25. *For all $c \in (0, \infty)$, $\mathcal{C}(\mathcal{P}_c)$ is bounded.*

Proof. Consider any vector $\theta \in \mathcal{C}(\mathcal{P}_c)$. The vector θ can be written $\theta = z + ck$, where $\langle z, k\rangle \leq 0$, and so $c|k|^2 \geq \langle \theta, k\rangle = \sum_j \theta_j k_j \geq \theta_i \min_j k_j$, where $i \in \{1, \ldots, n\}$. Therefore $\theta_i \leq \frac{c|k|^2}{\min_j k_j}$ for all $i \in \{1, \ldots, n\}$ and $\theta \in \mathcal{C}(\mathcal{P}_c)$, completing the proof. □

The following result, when combined with lemma 25, shows that all trajectories are forwardly bounded:

Theorem 31. *There exists $c \in (0, \infty)$ such that $\mathcal{C}(\mathcal{P}_c)$ is forward invariant and globally absorbing.*

Note that for such a value of c, $\mathcal{C}(\mathcal{P}_c)$ is closed and convex, by lemma 25 it is bounded; therefore it is also compact. As $\mathcal{C}(\mathcal{P}_c)$ is globally absorbing, it contains the ω-limit sets of every trajectory, which must be non-empty since $\mathcal{C}(\mathcal{P}_c)$ is bounded.

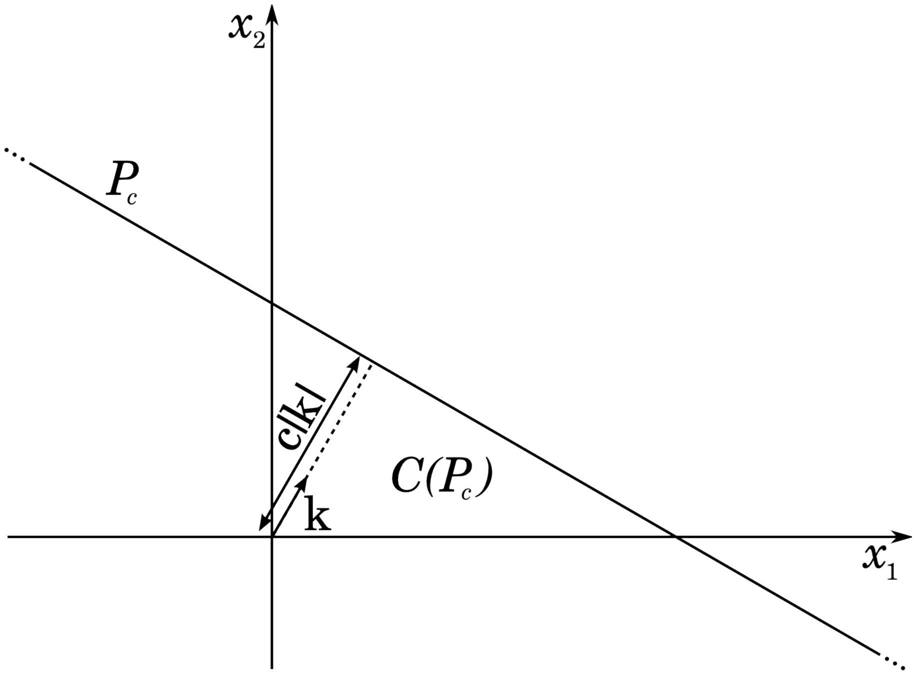

Figure 6.2: An example of a simplex in two dimensions. The hyperplane P_c, which in the two-dimensional case is simply a line, divides phase space into two half spaces. Since P_c is orthogonal to the vector k and passes through the point ck for some $c > 0$, the perpendicular distance between P_c and the origin is $c|k|$. As k is strictly positive, the intersection of the positive orthant and the lower half space is bounded and forms a simplex, the triangle $\mathcal{C}(P_c)$.

The proof of theorem 31 presented here is made up of two stages. The first stage is to find a value of c such that the vector field points into $\mathcal{C}(\mathcal{P}_c)$ for every $x \in \mathcal{P}_c$ satisfying $x > 0$. This establishes that $\mathcal{C}(\mathcal{P}_c)$ is both forward invariant and bounded. The second stage is to show that $\mathcal{C}(\mathcal{P}_c)$ is globally absorbing. In an attempt to make the proof more readable, each stage will be proved as a separate result.

The existence of $c \in (0, \infty)$ required for the first stage of theorem 31 can be demonstrated by examining the time derivative of $\mathcal{X}(x)$. Noting that $\nabla\mathcal{X}(x) = k$, the time derivative of $\mathcal{X}(x)$ is $\dot{\mathcal{X}}(x) = \nabla\mathcal{X} \cdot \dot{x} = \langle k, \dot{x}\rangle = \langle k, I\rangle + \langle k, S\rangle R(x) - \langle k, Q(x)\rangle$. The second term disappears from this equation since $\langle k, S\rangle = 0$, leaving

$$\dot{\mathcal{X}}(x) = \langle k, I\rangle - \langle k, Q(x)\rangle \tag{6.7}$$

This can also be written as

$$\dot{\mathcal{X}}(x) = \sum_{i=1}^{n} k_i(I_i - q_i(x_i)) \tag{6.8}$$

Hence, for $x \in \mathcal{P}_c$ where $x > 0$, $\dot{\mathcal{X}}(x) < 0$ means that the flow of x defined by $\dot{x}$ points into the interior of $\mathcal{C}(\mathcal{P}_c)$, and $\dot{\mathcal{X}}(x) > 0$ means that the flow of x points out of $\mathcal{C}(\mathcal{P}_c)$. Note also that since $q_i(x_i)$ is (by assumption) strictly increasing in x_i, $\dot{\mathcal{X}}(x)$ is strictly decreasing in each component x_i.

It will be of use to find out what happens to $\dot{\mathcal{X}}(x)$ along each half line in the nonnegative orthant that starts on the origin. Consider a unit vector $\hat{u} \in S^{n-1} \cap \mathbb{R}^n_{\geq 0}$, where S^{n-1} is the unit sphere in n dimensions. Any point $x \in \mathbb{R}^n_{\geq 0}$ other than 0 can be uniquely written $x = \lambda \hat{u}$ for some $\hat{u} \in S^{n-1}$ and $\lambda \in \mathbb{R}_{\geq 0}$. It is then possible to define a two parameter scalar function $\mathcal{Z}(\hat{u}, \lambda) = \dot{\mathcal{X}}(x)$. Examining $\mathcal{Z}(\hat{u}, \lambda)$ for fixed $\hat{u}$ and varying λ reveals the behaviour of $\dot{\mathcal{X}}(x)$ along the half line defined by $\hat{u}$. An important property of $\mathcal{Z}(\hat{u}, \lambda)$ is that:

Lemma 26. *$\mathcal{Z}(\hat{u}, \lambda)$ is C^1 and strictly decreasing in λ (for fixed $\hat{u}$).*

Proof. Let $\hat{u}_i$ be the ith component of $\hat{u}$. Differentiating $\mathcal{Z}(\hat{u}, \lambda)$ with respect to λ yields

$$\frac{\partial}{\partial \lambda}(\mathcal{Z}(\hat{u}, \lambda)) = -\sum_{i=1}^{n} k_i \frac{\mathrm{d}q_i}{\mathrm{d}x_i}\frac{\partial x_i}{\partial \lambda} = -\sum_{i=1}^{n} k_i q_i' \hat{u}_i$$

The fact that this derivative exists and is continuous means that $\mathcal{Z}(\hat{u}, \lambda)$ is C^1 in λ. Since $k_i > 0$, $q_i' > 0$ (by assumption) and $\hat{u}_i \geq 0$, with $\hat{u}_i > 0$ for at least one value of i, it follows that $\mathcal{Z}(\hat{u}, \lambda)$ is strictly decreasing in λ. □

Now consider the following result:

Lemma 27. *For each $\hat{u} \in S^{n-1} \cap \mathbb{R}^n_{\geq 0}$, there exists a unique $\lambda(\hat{u})$ such that $\mathcal{Z}(\hat{u}, \lambda(\hat{u})) = 0$.*

Proof. When $\lambda = 0$, $x = 0$. Substituting these values into equation (6.8) gives $\mathcal{Z}(\hat{u}, 0) = \sum_{i=1}^{n} k_i(I_i - q_i(0))$. Since $q_i(0) = 0$, this reduces to $\mathcal{Z}(\hat{u}, 0) = \sum_{i=1}^{n} k_i I_i > 0$.

For sufficiently large λ, $\mathcal{Z}(\hat{u}, \lambda) < 0$: Since $q_i(\lambda \hat{u}_i)$ is surjective and strictly increasing in λ, for each $\hat{u}_i > 0$ there exists λ_i such that $q_i(\lambda_i \hat{u}_i) > 1/k_i \sum_{j=1}^{n} k_j I_j$. Let $\lambda_m = \min_i \lambda_i$, which is guaranteed to exist as $\hat{u}_i > 0$ for at least one value of i. Substituting λ_m into equation (6.8) gives

$$\mathcal{Z}(\hat{u}, \lambda_m) = \sum_{j=1}^{n} k_j I_j - \sum_{i=1}^{n} k_i q_i(\lambda_m \hat{u}_i)$$

Since $k_m q_m(\lambda_m \hat{u}_m) > \sum\limits_{j=1}^{n} k_j I_j$, where m is the value of i satisfying $\min_i \lambda_i$, this expression is negative.

By lemma 26 $\mathcal{Z}(\hat{u},\lambda)$ is continuous and strictly decreasing in λ. Combining this with the fact that for fixed $\hat{u}$, $\mathcal{Z}(\hat{u},0) > 0$, $\mathcal{Z}(\hat{u},\lambda_m) < 0$, means that $\mathcal{Z}(\hat{u},\lambda) = 0$ has a unique solution $\lambda(\hat{u})$ for fixed $\hat{u}$. □

In a slight abuse of notation, for a given $\hat{u}$, label the solution of $\mathcal{Z}(\hat{u},\lambda) = 0$ as $(\hat{u},\lambda(\hat{u}))$. $\lambda(\hat{u})$ is a function with domain $S^{n-1} \cap \mathbb{R}^n_{\geq 0}$ and codomain $\mathbb{R}_{>0}$.

Lemma 28. *There exists some $\lambda_{\max}$ such that for all $\lambda > \lambda_{\max}$, $\mathcal{Z}(\hat{u},\lambda) < 0$.*

Proof. First suppose that $\lambda_{\max} = \max_{\hat{u}} \lambda(\hat{u})$ exists. By definition, $\mathcal{Z}(\hat{u},\lambda_{\max}) \leq 0$ and since it was shown in lemma 26 that $\mathcal{Z}(\hat{u},\lambda)$ is continuous and strictly decreasing in λ, $\mathcal{Z}(\hat{u},\lambda) < 0$ for all $\lambda > \lambda_{\max}$. Thus, to prove the claim it suffices to show that $\max_{\hat{u}} \lambda(\hat{u})$ exists.

Choose any $\hat{u}_0 \in S^{n-1} \cap \mathbb{R}^n_{\geq 0}$. By lemma 26, $\mathcal{Z}(\hat{u}_0,\lambda)$ is differentiable with respect to λ and $\frac{\partial}{\partial\lambda}(\mathcal{Z}(\hat{u}_0,\lambda)) \neq 0$, so $\lambda(\hat{u}_0)$ is locally C^1 by the implicit function theorem. Moreover, since the choice of $\hat{u}_0$ was arbitrary, this holds for every $\hat{u} \in S^{n-1} \cap \mathbb{R}^n_{\geq 0}$ and so $\lambda(\hat{u})$ is C^1 over its whole domain. Combining this with the fact that the domain $S^{n-1} \cap \mathbb{R}^n_{\geq 0}$ is closed and compact proves that $\lambda_{\max}$ is attained. □

With these results in place, it is now possible to give the proof to theorem 31.

Proof. All trajectories begin within the nonnegative orthant, which is forward invariant by lemma 23 (p. 108). Choose an arbitrary $\delta > 0$. By lemma 28, $\dot{\mathcal{X}}(x) < 0$ for every $x > 0$, $x \in \mathcal{P}_0 + (\lambda_{\max} + \delta)k$, so $\mathcal{C}(\mathcal{P}_{\lambda_{\max}+\delta})$ is forward invariant, proving the first part of the theorem.

By lemma 28, for any $\epsilon > \delta$, $\dot{\mathcal{X}}(x) < 0$ for every $x > 0$, $x \in \mathcal{P}_0 + (\lambda_{\max} + \epsilon)k$. Therefore $\mathcal{C}(\mathcal{P}_{\lambda_{\max}+\delta})$ is also globally attracting. □

With the existence of a bounded globally absorbing set having been established, all that remains for this section is to show that the system has a unique equilibrium.

Lemma 29. *The forward invariant globally absorbing set $\mathcal{C}(\mathcal{P}_{\lambda_{\max}+\delta})$ contains exactly one fixed point.*

Proof. Since the trapping region $\mathcal{C}(\mathcal{P}_{\lambda_{\max}+\delta})$ is a simplex it forms a compact convex set, and it therefore contains at least one fixed point as a corollary of the Brouwer fixed point theorem or by theorem 3 (p. 17).

Corollary 3.5 of [Banaji et al., 2007] states that if the stoichiometric matrix of a set of reactions in a CFSTR is **strongly sign determined** (SSD) and none of the reactions is a

one-step catalytic reaction[1] (i.e. no reactant appears on both sides of any one reaction, or affects the rate of a reaction that it does not take part in) then there cannot be more than one fixed point. Note that this result was proved in [Banaji et al., 2007] in the context of a CFSTR, but the result remains valid for the more general outflow conditions described in §6.2. The result was proved in three steps: first it was shown that, if the stoichiometric matrix of set of reactions is SSD and there are no one-step catalytic reactions, then the matrix SV is a $P_0^{(-)}$ matrix at all points in the phase space $\mathbb{R}^n_{\geq 0}$. It was then shown that, if SV is a $P_0^{(-)}$ matrix and Q' is a diagonal matrix with strictly positive diagonal entries, then the Jacobian $SV - Q'$ is a $P^{(-)}$ matrix, again at all points in $\mathbb{R}^n_{\geq 0}$. Non-existence of multiple fixed points then follows from injectivity via theorem 4 (p. 21) — note from the comment directly after theorem 4 that if the Jacobian of a differential mapping is a P (or $P^{(-)}$) matrix, then the mapping is injective on **closed** rectangular regions of $\mathbb{R}^n$.

Putting this result back in the context of lemma 29, since the forward invariant globally absorbing set $\mathcal{C}(\mathcal{P}_{\lambda_{\max}+\delta})$ is a subset of a closed rectangular region of $\mathbb{R}^n$, it contains only one fixed point if the stoichiometric matrix is SSD and there are no one-step catalytic reactions[2]. The conditions for "no one-step catalysis" are fulfilled if $S_i V_i \leq 0$ for all i and $S_i = 0 \Rightarrow V_i = 0$. A matrix B is defined as SSD when every square submatrix of B is either singular or "sign nonsingular". A square matrix M is called sign nonsingular when every matrix with the same sign structure (the qualitative class of M, $\mathcal{Q}(M)$, see [Brualdi and Shader, 1995]) is nonsingular. By a continuity argument it follows that every matrix $N \in \mathcal{Q}(M)$ has determinant of the same sign.

It has already been assumed that no reactant appears on both sides of the reaction, and that every reactant takes part in the reaction, i.e. $S_i \neq 0$. Thus the reaction is not a one-step catalytic reaction. The stoichiometric matrix S is a column vector and therefore trivially SSD. Thus the Jacobian of the system is a $P^{(-)}$ matrix, and the vector field is injective on every closed rectangular region of $\mathbb{R}^n$, by theorem 4 (see the comment following theorem 4). Consequently $\mathcal{C}(\mathcal{P}_{\lambda_{\max}+\delta})$ cannot contain multiple fixed points, which when combined with theorem 3 means there is a single fixed point. □

6.4 Conditions for convergence to the fixed point

Having established that the reaction system is globally bounded, and that there is a unique fixed point, conditions under which trajectories converge to the fixed point will be considered.

[1] [Banaji et al., 2007] uses the term "nonautocatalytic reaction," but this is something of a misnomer; "not a one-step catalytic reaction" is a more accurate description of the meaning.

[2] N.B. these conditions suffice but are by no means necessary to guarantee uniqueness of any fixed points.

Two approaches are used to derive conditions for convergence. The first of these is monotonicity, as discussed in chapter 2. The second, which is presented in §6.4.2, relies on autonomous convergence, as discussed in chapter 3. Although there is an overlap in parts of the proofs used to demonstrate convergence, it is hoped that the application of these two different techniques to the same problem is of some academic interest.

6.4.1 Monotonicity of a single reaction

Theorem 32. *Consider a model of a chemical reaction taking place in a closed container, as described in equation (6.1). Choose a set of $n-1$ vectors $\{\nu^i\}$, $i \in \{2,\ldots,n\}$, where the jth component ($j = 1,\ldots,n$) of ν^i is given by*

$$\nu^i_j = \begin{cases} -S_i, & i = j \\ 0, & i \neq j \end{cases}$$

If $\exists\ m \in \{1,\ldots,n\}$ such that $q'_m \leq q'_i\ \forall\ i$ at every value of x then J is K-quasipositive for the simplicial cone K generated by the set of vectors $S \cup \{\nu^i\}$.

Proof. Without loss of generality, it will be assumed that $m = 1$. This is possible since, as noted in §6.2, the order in which reactants appear in the reaction is arbitrary and therefore labels on the reactants can be switched without altering the dynamics.

K is a simplicial cone, due to the fact that $S \cup \{\nu^i\}$ is a set of n linearly independent vectors. Since K is simplicial, the original system can be transformed to a system with the generators of K as its basis vectors. Define a transform matrix

$$T = \begin{pmatrix} S_1 & & & \\ S_2 & -S_2 & & \\ \vdots & & \ddots & \\ S_n & & & -S_n \end{pmatrix} \tag{6.9}$$

The inverse of T is

$$T^{-1} = \begin{pmatrix} \frac{1}{S_1} & & & \\ \frac{1}{S_1} & -\frac{1}{S_2} & & \\ \vdots & & \ddots & \\ \frac{1}{S_1} & & & -\frac{1}{S_n} \end{pmatrix} \tag{6.10}$$

T represents a transform on phase space. Quasipositivity of the Jacobian in this transformed system is equivalent to K-quasipositivity in the original system. The columns of T are the new basis vectors. Define the transformed Jacobian as J_T:

$$J_T = T^{-1}JT \tag{6.11}$$

The transformed Jacobian can now be examined, using

$$J = SV - Q' \Rightarrow J_T = T^{-1}(SV - Q')T = T^{-1}SVT - T^{-1}Q'T$$

The two transformed parts will be considered separately.

$$SV = \begin{pmatrix} S_1V_1 & \cdots & S_1V_n \\ \vdots & \ddots & \vdots \\ S_nV_1 & \cdots & S_nV_n \end{pmatrix} \tag{6.12}$$

From this,

$$T^{-1}SVT = \begin{pmatrix} \sum\limits_{i=1}^{n} S_iV_i & -S_2V_2 & \cdots & -S_nV_n \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \tag{6.13}$$

The second part of the transformed Jacobian is

$$T^{-1}Q'T = \begin{pmatrix} q_1' & & & \\ q_1' - q_2' & q_2' & & \\ \vdots & & \ddots & \\ q_1' - q_n' & & & q_n' \end{pmatrix} \tag{6.14}$$

Thus the structure of the full transformed Jacobian is this:

$$J_T = \begin{pmatrix} \sum\limits_{i=1}^{n} S_iV_i - q_1' & -S_2V_2 & \cdots & -S_nV_n \\ q_2' - q_1' & -q_2' & & \\ \vdots & & \ddots & \\ q_n' - q_1' & & & -q_n' \end{pmatrix} \tag{6.15}$$

Every offdiagonal element of the first row is nonnegative, since $S_iV_i \leq 0$. Since, by assumption, $q_1' \leq q_i'$ for all $i \in \{1, \ldots, n\}$, every offdiagonal element of the first column is nonnegative. All other offdiagonal elements are zero, and hence J_T is quasipositive. □

The condition on the derivatives of the outflow functions is fairly restrictive, and in general will not hold true for every x. However, the condition is fulfilled in a CFSTR:

Corollary 8. *Consider a model of a chemical reaction taking place in a CFSTR, as defined in equation (6.3). For such a system, J is K-quasipositive for the simplicial cone K defined in theorem 32.*

Proof. By definition of a CFSTR, $q'_i = q'_j = \text{const}$ for all $i, j \in \{1, \ldots, n\}$. Therefore the subdiagonal elements in J_T are zero, and J_T is quasipositive by theorem 32. □

The above corollary also trivially generalises to any system where the outflow of each reactant is a linear function of concentration, even if the coefficient is different for each reactant.

The condition required for theorem 32 in fact guarantees global attractivity of the equilibrium, via theorem 18 (p. 38). Recall that theorem 18 requires that the following four conditions on the dynamical system defined in §6.1 be met:

1. For all compact $C \subset \mathbb{R}^n_{\geq 0}$, $\inf(C), \sup(C) \in \mathbb{R}^n_{\geq 0}$.
2. Flows of the dynamical system are monotone.
3. The system has a unique equilibrium.
4. The forward semi-orbit of every point has compact closure in $\mathbb{R}^n_{\geq 0}$.

As was shown in theorem 32, condition 2 of theorem 18 holds for the dynamical system defined in equation (6.1), provided that the requirements specified in theorem 32 are met. Conditions 3 and 4 were shown to be satisfied in section §6.3, since the system has an absorbing set which is closed and bounded. The only apparently problematic condition is 1. That this condition is also met can be demonstrated using the following lemma, which shows that K is a superset of an orthant:

Lemma 30. *Consider a model of a chemical reaction network taking place in a closed container, as described in equation (6.1). Assume that the dynamical system fulfils the conditions specified in theorem 32. The cone K preserved by the Jacobian of the dynamical system covers the orthant generated by $\{\operatorname{sgn}(S_i)\hat{e}_i\}$, where $i = 1, \ldots n$ and $\{\hat{e}_i\}$ is the standard orthonormal basis for $\mathbb{R}^n$.*

Proof. Recall that the generators of K are the columns of the transform matrix T. For notational convenience, let T_i be the ith column of T. From equation (6.9), T has the form

$$T = \begin{pmatrix} S_1 & & & \\ S_2 & -S_2 & & \\ \vdots & & \ddots & \\ S_n & & & -S_n \end{pmatrix}$$

Observe that

$$\sum_{i=1}^{n} \frac{T_i}{|S_1|} = \mathrm{sgn}(S_1)\hat{e}_1$$

and

$$\frac{T_i}{|S_i|} = \mathrm{sgn}(S_i)\hat{e}_i, i = 2, \ldots, n$$

Let the orthant generated by $\{\mathrm{sgn}(S_i)\hat{e}_i\}$ be denoted P. $P \subseteq K$ if and only if $x \in K$ for every $x \in P$. If x_i is the ith component of x, $x \in P$ can be written

$$x = \sum_{i=1}^{n} |x_i| \mathrm{sgn}(S_i)\hat{e}_i$$

Substituting in the expressions for $\hat{e}_i$ given above, this can be rewritten

$$x = \left(\sum_{i=1}^{n} \frac{T_i}{|S_1|} \right) |x_1| + \sum_{i=2}^{n} \frac{T_i}{|S_i|} |x_i|$$

Since the coefficients of T_i in this expression are all nonnegative, it follows that $x \in K$. The proof is complete. □

This leads directly to a proof of global attractivity of the fixed point:

Corollary 9. *The model of a chemical reaction network taking place in a closed container, as described in equation (6.1), has a unique globally attractive equilibrium if the dynamical system fulfils the conditions specified in theorem 32.*

Proof. That the dynamical system has a unique fixed point was established in §6.3. The proof of global attractivity follows from the results above: under the assumptions made for theorem 32, the system is monotone with respect to the partial ordering defined by a cone K. That this cone K covers an orthant is shown in lemma 30. Therefore, by lemma 7 (p. 39), the phase space $\mathbb{R}^n_{\geq 0}$ with the partial order defined by K is a lattice, and every bounded set in $\mathbb{R}^n_{\geq 0}$ is also order bounded in $\mathbb{R}^n_{\geq 0}$. In turn, this implies that $\inf(C), \sup(C) \in \mathbb{R}^n_{\geq 0}$ for all compact $C \in \mathbb{R}^n_{\geq}$ by lemma 6 (p. 38). Thus all four conditions required for theorem 18 are met, and the fixed point is globally attractive. □

That the fixed point is also globally asymptotically stable can be shown via the following lemma:

Lemma 31. *Consider a model of a chemical reaction network taking place in a closed container, as described in equation (6.1). For each value of x, if it is possible to choose $m \in \{1, \ldots, n\}$ such that $q'_m \leq q'_i$ then J_T is Hurwitz stable everywhere.*

Proof. According to theorem 7 (p. 23) and its accompanying remark, J_T is Hurwitz stable if each of its diagonal entries is negative and the magnitude of each of these entries dominates the sum of magnitudes of all other elements in the same row.

Theorem 33, below, proves essentially the same statement with stronger consequences, so the argument is not duplicated here. Note however that lemma 31 does not require m to be the same for each x, but theorem 33 only applies if m is the same for every value of x. □

In light of this lemma, it is apparent that for the dynamical system defined in equation (6.1), the conditions required for theorem 32 and lemma 30 guarantee that the steady state solution of the dynamical system is globally attracting and locally asymptotically stable, and hence that it is globally asymptotically stable.

6.4.2 Autonomous convergence and logarithmic norms

In this section, stability of the chemical reaction system is investigated using the autonomous convergence techniques outlined in chapter 3. The conditions derived that guarantee autonomous convergence of the system are similar to those found for monotone convergence in the previous section, but are slightly more general. The first result, regarding the existence of a negative logarithmic norm, is closely related to the proof of lemma 31.

All of the results in this section are expressed in terms of the chemical reaction model with general flow conditions as defined in equation (6.1). As such, they also apply to the CFSTR model appearing in equation (6.3) as a special case, but the results are not explicitly formulated for a CFSTR.

Theorem 33. *For J_T as defined in equation (6.15) on p. 116, $\mu_\infty(J_T)$ is negative provided that there exists $m \in \{1, \ldots, n\}$ such that $q'_m < 2q'_i \ \forall \ i \neq m$ at every point x.*

Proof. As in theorem 32, it will be assumed without loss of generality that $m = 1$. Recall that for a square matrix $M = (M_{pq})$, the μ_∞ logarithmic norm is

$$\mu_\infty = \max_p \left(\mathrm{Re}(M_{pp}) + \sum_{q \neq p} |M_{pq}| \right) \tag{6.16}$$

Hence for the logarithmic infinity norm of a matrix to be negative, every diagonal element must have negative real part, and it must dominate the sum of magnitudes of all the other elements on the same row.

The diagonal elements of J_T are negative by inspection. For the first row, the inequality required for the diagonal element to dominate is $|\sum_{i=1}^{n} S_iV_i - q_1'| > \sum_{i=2}^{n} |S_iV_i|$, which is true by inspection. For the ith row ($i = 2, \ldots, n$), $|q_i'| > |q_i' - q_1'|$ is required, which is equivalent to the condition $q_1' < 2q_i'$ since, by assumption, $q_j' > 0$ for $j = 1, \ldots, n$. □

Recall that theorem 20 on page 46, chapter 3, states that if the Jacobian of an autonomous differential equation defined on a convex forward invariant set $X \subset \mathbb{R}^n$ has at least one fixed point, the existence of a logarithmic norm which is negative at every point in X guarantees that the fixed point is unique and globally stable. Since the set $\mathcal{C}(\mathcal{P})$ is a simplex and therefore convex, and is known to contain a fixed point, this implies that the fixed point is globally stable when the conditions given in theorem 33 are fulfilled.

This result for global asymptotic stability is stronger than the one given at the end of section 6.4.1 in two ways. First, the condition required for the negative logarithmic norm of J_T, $q_1' < 2q_i'$, is a relaxation of the condition required for quasipositivity of J_T, $q_1' < q_i'$. Second, there is no need to explicitly verify that the fixed point is locally asymptotically stable (in fact this follows directly from $\mu_\infty(J_T) < 0$, cf. lemma 31).

It is possible to further strengthen the result, using the second autonomous convergence theorem presented as theorem 23 (p. 53) in chapter 3.

Theorem 34. *Let $J_T^{[2]}$ be the second additive compound of the transformed Jacobian. Then $\mu_\infty(J_T^{[2]}) < 0$ if for some fixed $m \in \{1, \ldots, n\}$, every pair of q_i, q_j satisfies $|q_i' - q_m'| + |q_j' - q_m'| < q_i' + q_j'$.*

Proof. As in previous results in this chapter, it can be assumed without loss of generality that $m = 1$. Recall from equation (3.6) that for a square matrix M

$$\mu_\infty(M) = \max_i \left(\mathrm{Re}(M_{ii}) + \sum_{j \neq i} |M_{ij}| \right)$$

Therefore $\mu_\infty(J_T^{[2]}) < 0$ follows if ${J_T^{[2]}}_{kk} < 0$ and $|{J_T^{[2]}}_{kk}| > \sum_{l, l \neq k} |{J_T^{[2]}}_{kl}|$ for all $k \in {}^nC_2$. The structure of J_T is given in equation (6.15). Using this, equation (3.12) can be used to work out the structure of $J_T^{[2]}$. Recall that each index i of $J_T^{[2]}$ corresponds to an ordered pair (i_1, i_2), with $i_1, i_2 \in \{1, \ldots, n\}$. In showing that $\mu_\infty(J_T^{[2]}) < 0$, the rows of $J_T^{[2]}$ with index i corresponding to an ordered pair $(1, i_2)$ will be considered first, followed by the rows with index corresponding to an ordered pair (i_1, i_2) with $i_1 > 1$.

The diagonal elements of $J_T^{[2]}$ are given by ${J_T^{[2]}}_{ii} = {J_T}_{i_1i_1} + {J_T}_{i_2i_2}$. Therefore, for each i where $i_1 = 1$, ${J_T^{[2]}}_{ii} = {J_T}_{11} + {J_T}_{i_2i_2} = \sum_{k=1}^{n} S_kV_k - q_1' - q_{i_2}' < 0$. Using equation (3.12),

$$\sum_{j, j \neq i} |{J_T^{[2]}}_{ij}| = \sum_{j_1, j_1 \neq i_1} \sum_{j_2, j_2 > j_1} \delta_{i_2j_2} |{J_T}_{i_1j_1}| + \sum_{j_2, j_2 \neq i_2} \sum_{j_1, j_1 < j_2} \delta_{i_1j_1} |{J_T}_{i_2j_2}| \qquad (6.17)$$

Since $i_2 \geq 2$ and ${J_T}_{i_2 j_2} = 0$ whenever $i_2 \geq 2$ and $j_2 \neq 1$ or i_2, the second sum on the RHS of this expression is zero. Therefore (when $i_1 = 1$)

$$\sum_{j,j\neq i} |{J_T^{[2]}}_{ij}| = \sum_{j_1,j_1\neq 1} \sum_{j_2,j_2>j_1} \delta_{i_2 j_2} |{J_T}_{1 j_1}| = \sum_{k=2,k\neq i_2}^{n} |S_k V_k|$$

Since $\sum_{k=2,k\neq i_2}^{n} |S_k V_k| < |\sum_{k=1}^{n} S_k V_k - q_1' - q_{i_2}'|$, the conditions required for $\mu_\infty(J_T^{[2]}) < 0$ are satisfied for all rows of $J_T^{[2]}$ with index i corresponding to an ordered pair $(1, i_2)$.

The rows of $J_T^{[2]}$ with index i corresponding to an ordered pair (i_1, i_2) where $i_1 > 1$ are simpler. The diagonal elements are given by ${J_T^{[2]}}_{ii} = {J_T}_{i_1 i_1} + {J_T}_{i_2 i_2} = -q_{i_1}' - q_{i_2}'$, each of which is negative. By equation (6.17), and once again using the fact that $i_2 \geq 2$ and ${J_T}_{i_2 j_2} = 0$ whenever $i_2 \geq 2$ and $j_2 \neq 1$ or i_2,

$$\sum_{j,j\neq i} |{J_T^{[2]}}_{ij}| = |q_{i_1}' - q_1'| + |q_{i_2}' - q_1'|$$

By assumption, $q_{i_1}' + q_{i_2}' > |q_{i_1}' - q_1'| + |q_{i_2}' - q_1'|$, completing the proof. □

Improved conditions for global asymptotic stability of the dynamical system can be constructed from this result, using the second autonomous convergence theorem in chapter 3 (theorem 23 on page 53).

Corollary 10. *Suppose that in the dynamical system described in equation (6.1), for some fixed $m \in \{1, \ldots, n\}$, every pair of $q_i, q_j (i \neq j \neq m)$ satisfies $|q_i' - q_m'| + |q_j' - q_m'| < q_i' + q_j'$. Then the dynamical system has a globally asymptotically stable fixed point.*

Proof. That the system has a unique fixed point was demonstrated in §6.3. That the fixed point is globally asymptotically stable follows from theorem 23. Recall that theorem 23 states that when a dynamical system has a globally absorbing set containing a unique fixed point, the fixed point is globally asymptotically stable if $\mu_k(J^{[2]}) < 0$ for some logarithmic norm μ_k. By assumption, the conditions required for theorem 34 are fulfilled, so $\mu_\infty(J_T^{[2]}) < 0$. As demonstrated in corollary 3 on page 54, this means that there exists a logarithmic norm $\mu_{\infty,\tilde{T}^{-1}}$ such that $\mu_{\infty,\tilde{T}^{-1}}(J^{[2]}) < 0$. Consequently, by theorem 23 the fixed point is globally asymptotically stable. □

Remark. This result is slightly stronger than theorem 33, in that the condition $q_i' + q_j' > |q_i' - q_1'| + |q_j' - q_1'| \ \forall\ i, j$ is a slight relaxation of the condition $q_i' > |q_i' - q_1'| \ \forall\ i$. Note however that $q_i' \leq |q_i' - q_1'|$ is possible for at most one value of i.

When the chemical reaction only involves three substrates, no assumptions about the outflow beyond those given in §6.2 are required to guarantee global convergence:

Lemma 32. *When the reaction modelled by equation (6.1) involves three substrates, there exists a transform matrix T such that $\mu_1((T^{-1}JT)^{[2]}) < 0$ at every point in phase space.*

Proof. Consider the matrix

$$T = \begin{pmatrix} S_1 & 0 & 0 \\ 0 & S_2 & 0 \\ 0 & 0 & S_3 \end{pmatrix}$$

A symbolic algebra package such as [Maxima, 2008] can be used to explicitly construct the second additive compound

$$(T^{-1}JT)^{[2]} =$$

$$\begin{pmatrix} V_1S_1 + V_2S_2 - q'_1 - q'_2 & V_3S_3 & -V_3S_3 \\ V_2S_2 & V_1S_1 + V_3S_3 - q'_1 - q'_3 & V_2S_2 \\ -V_1S_1 & V_1S_1 & V_2S_2 + V_3S_3 - q'_2 - q'_3 \end{pmatrix}$$

For a square matrix $M = (M_{pq})$, recall that the μ_1 logarithmic norm is

$$\mu_1 = \max_q \left(\mathrm{Re}(M_{qq}) + \sum_{p \neq q} |M_{pq}| \right) \tag{6.18}$$

Hence for the logarithmic 1-norm of a matrix to be negative, every diagonal element must have negative real part, and its magnitude must be greater that the sum of magnitudes of all the other elements in the same column.

Clearly the diagonal elements of $(T^{-1}JT)^{[2]}$ are always negative provided that $q'_i \geq 0 \,\forall\, i$, and the diagonal element of each column has greater magnitude than the sum of magnitudes of all the other elements provided that $q'_i > 0$ for at least one value of i. This requirement is satisfied since, by assumption, $q'_i > 0$ for all i. □

Corollary 11. *Suppose the reaction modelled by equation (6.1) involves only three substrates. Then all initial states converge to a unique equilibrium.*

Proof. As shown in §6.3, the dynamical system has a globally absorbing compact set containing a unique fixed point. By lemmas 32 and 8 (p. 45), there exists a similarity transform $\tilde{T}$ such that $\mu_{1,\tilde{T}^{-1}}(J^{[2]}) < 0$. Therefore, by theorem 23 (p. 53), the system is globally asymptotically stable. □

Unfortunately, lemma 32 does not generalise to reactions with more than three substrates in an obvious way. Consider a reaction involving four substrates: the analogous transform on $\mathbb{R}^4$ is

$$T = \begin{pmatrix} S_1 & 0 & 0 & 0 \\ 0 & S_2 & 0 & 0 \\ 0 & 0 & S_3 & 0 \\ 0 & 0 & 0 & S_4 \end{pmatrix}$$

This corresponds to a transform on $\Lambda^2(\mathbb{R}^4)$ of the form

$$T^{(2)} = \begin{pmatrix} S_1S_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & S_1S_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & S_1S_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & S_2S_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & S_2S_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & S_3S_4 \end{pmatrix}$$

The first two columns of the resulting transformed second additive compound of the Jacobian are

$$T^{(2)^{-1}} J^{[2]} T^{(2)} = \begin{pmatrix} V_1S_1 + V_2S_2 - q_1' - q_2' & V_3S_3 & \cdots \\ V_2S_2 & V_1S_1 + V_3S_3 - q_1' - q_3' & \cdots \\ V_2S_2 & V_3S_3 & \cdots \\ -V_1S_1 & V_1S_1 & \cdots \\ -V_1S_1 & 0 & \cdots \\ 0 & -V_1S_1 & \cdots \end{pmatrix}$$

The other columns are omitted for reasons of space; however, it is obvious that in each column, the magnitude of the diagonal element is not greater than the sum of the magnitudes of the off-diagonal elements, and hence $\mu_{1,T^{(2)-1}}(J^{[2]}) \nless 0$. In five dimensions, the first column contains three off-diagonal entries of $-V_1S_1$ and three off-diagonal entries of V_2S_2 but the same diagonal element as in three and four dimensions, so the situation appears to get worse as the number of dimensions increases. This is not to say that in general there is no ${}^nC_2 \times {}^nC_2$ similarity transform $\tilde{T}_n$ on the second exterior power of the phase space (with or without a corresponding $n \times n$ transform T_n on the phase space itself) that makes $\mu_s(\tilde{T}_n^{-1} J^{[2]} \tilde{T}_n) < 0$ for some logarithmic norm μ_s, but if such a transform exists it has not been found. It is possible that a more general set of conditions can be found under which the system is globally stable. In light of this, consider the following conjecture:

Conjecture 1. *Consider a dynamical system defined on $\mathbb{R}^n_{\geq 0}$ of the form*

$$\dot{x} = I + SR(x) - Q(x)$$

with the following properties:

1. *$I > 0$ and constant*
2. *$S \notin \pm\mathbb{R}^n_{\geq 0}$ and constant*
3. *$Q(x) = (q_1(x_1), \ldots, q_n(x_n))^T$*
4. *$q_i(0) = 0$, $q_i \in C^1$, $q_i' > 0$*
5. *Every $q_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} : x_i \mapsto q_i(x_i)$ is onto*
6. *$R(x) \in C^1$*
7. *$S_i V_i \leq 0$*

there is a unique globally stable equilibrium.

In [Banaji, 2008], it was demonstrated that any reaction taking place in a CFSTR and satisfying the assumptions made in §6.2 is monotone, and the set of cones that such a reaction preserves was characterised (the set of simplicial cones was completely characterised and the set of nonsimplicial cones was partially characterised). However, the paper made no direct claims about the asymptotic behaviour of such a reaction. By contrast, while the results in this section only demonstrate monotonicity of a single reaction with respect to one particular simplicial cone, they go on to prove that the reaction system is globally asymptotically stable using both monotonicity and autonomous convergence, under slightly more general outflows than those allowed for a CFSTR.

6.5 Multiple reactions

Some results regarding multiple reactions will now be considered. Unlike [Banaji, 2008], which characterises cones preserved by a set of reactions, this section of the thesis goes a different route, and does not deal with monotonicity at all. Instead the focus is on using autonomous convergence theory to find sufficient conditions on a reaction network to guarantee global asymptotic stability.

The reaction dynamics are once again governed by the equation

$$\dot{x} = I + SR(x) - Q(x) \tag{6.19}$$

By contrast to equation (6.1) on page 105, S is an $n \times r$ matrix rather than an n-vector and R is no longer a scalar but an r-vector, but this new equation is otherwise identical.

The assumptions made about the reactions are similar to those made for a single reaction in §6.2. Assume that there are n reactants taking part in r reactions in some fixed volume container, with the concentration of the ith reactant being represented by $x_i \in \mathbb{R}_{\geq 0}$. As in the single reaction case, let the inflow rate of each reactant be a nonnegative constant I_i, and let the outflow rate of each reactant be a function $q_i(x_i)$, satisfying the same assumptions as were made in §6.2. Define a stoichiometric $n \times r$ matrix S, with entry S_{ij} of S being the stoichiometry of reactant i in reaction j. Let R_j be the rate of reaction j, and define an $r \times n$ matrix V, with each entry $V_{ji} = \frac{\partial R_j}{\partial x_i}$. It is assumed that if $S_{ij} = 0$, $V_{ji} = 0$, i.e. no reactant affects the rate of a reaction it does not take part in. As with the single reaction, it is also assumed that every reaction is a "true" reaction, with at least one reactant on each side. It is once again assumed that each reaction rate is a monotone function of the concentrations of the reactants that participate in the reaction, with $S_{ij}V_{ji} < 0$ at all points in $\mathbb{R}^n_{>0}$ ($S_{ij}V_{ji} \leq 0$ at all points in $\mathbb{R}^n_{\geq 0}$).

One extra assumption will be added, namely that all reactions obey the law of atomic balance. In chemical terms, the law of atomic balance means that every chemical reactant is made of subcomponents, usually atoms (although other subcomponents such as electrons or ions can also be considered), and that these subcomponents are not created or destroyed in any reaction. Clearly any real reaction must obey the law of atomic balance since energy must be conserved, and reactant species cannot break down in chemical reactions (as opposed to nuclear reactions). However, in exceptional cases a model that doesn't obey atomic balance can be of use — one example of this is the Oregonator model of the Belousov-Zhabotinsky reaction, which is briefly mentioned in [Érdi and Tóth, 1989].

In mathematical terms, the law of atomic balance means that there exists an "atomic matrix" Z — see, for example, p. 22 of [Érdi and Tóth, 1989]. Z is an $m \times n$ nonnegative matrix, where m is the number of distinct subcomponents of the reactants in the reaction system. Z_{mn} is the number of particles of subcomponent m that are contained in reactant n. A Z matrix has the property that each of its rows is orthogonal to each column of S, since the same number of units of each subcomponent must appear on both sides of every reaction. Consider the simple reaction of hydrogen and oxygen forming water:

$$2\mathrm{H}_2 + \mathrm{O}_2 \rightarrow 2\mathrm{H}_2\mathrm{O}$$

This has atomic matrix

$$Z = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 2 & 1 \end{pmatrix}$$

The first row represents hydrogen, and the second row represents oxygen. The columns represent H_2, O_2 and $\mathrm{H}_2\mathrm{O}$ respectively. The stoichiometric matrix (or vector, in this single reaction case) is $S = (2, 1, -2)^T$. By inspection, $Z_1 S = Z_2 S = 0$, where Z_1 is the first row of Z and Z_2 is the second.

6.5.1 Boundedness and existence of fixed points

As with the single reaction case, the following simple but key result (cf. lemma 24 on page 109) will be useful in proving the existence of steady state solutions:

Lemma 33. *There exists some strictly positive vector k such that k lies in the left kernel of S, i.e. $k^T S = 0^T$.*

Proof. Each column of Z must contain at least one non-zero entry, as every reactant must be comprised of at least one subcomponent, and every row of Z is orthogonal to each column of S. Thus it is possible to choose the vector k to be the sum of rows of Z, in which case k is both strictly positive and orthogonal to every column of S. □

Remark. The columns of S span a subspace $\mathcal{S}$ of $\mathbb{R}^n$, with $\dim(\mathcal{S}) < n$ and $\mathcal{S} \cap \mathbb{R}^n_{\geq 0} = \{0\}$. $\dim(\mathcal{S}) < n$ is guaranteed by the existence of a left eigenvector of 0 (as shown in lemma 33), and $\mathcal{S} \cap \mathbb{R}^n_{\geq 0} = \{0\}$ follows directly from the law of atomic balance.

Define a new variable $\mathcal{X} = \langle k, x\rangle$, as in the single reaction case. From equation (6.19), the equation $\dot{\mathcal{X}} = \langle k, I\rangle + \langle k, S\rangle R - \langle k, Q\rangle$ can be obtained, where once again the second term vanishes, leaving $\dot{\mathcal{X}} = \langle k, I\rangle - \langle k, Q\rangle$.

Theorem 35. *There exists a compact convex forward invariant globally absorbing set $\mathcal{C}(\mathcal{P})$ containing at least one fixed point.*

Proof. The first part of this theorem, proving the existence of a compact convex forward invariant globally absorbing set, is an extension of theorem 31 (p. 110) for multiple reactions. Having established the existence of k in lemma 33, the argument proceeds in exactly the same way as in the single reaction case, so it is not repeated here.

Likewise, as in lemma 29 (p. 113), since $\mathcal{C}(\mathcal{P})$ is compact and convex it must contain at least one fixed point by theorem 3 (p. 17). □

The Jacobian of the dynamical system representing the evolution of the concentrations of the reactant is

$$J = SV - Q' \tag{6.20}$$

where Q' is the matrix derivative of outflow rates defined as for the single reaction.

It will now be shown that for a certain class of reaction systems there is a unique fixed point. The proof runs along the same lines as lemma 29 (p. 113), by showing that for certain reaction network structures, the stoichiometric matrix S is strongly sign determined, and that the mapping defined by the dynamical system is therefore injective.

The situation for multiple reactions is more complicated than for single reactions, as S is not always strongly sign determined in the multiple reaction case. Note that the result below gives just one example of a class of strongly sign determined matrices; there are other such classes, for which a similar result could be constructed. In proving the following theorem, the **qualitative class** of a matrix will be used, as defined in [Brualdi and Shader, 1995]. The qualitative class $\mathcal{Q}(M)$ of a matrix M is defined by $N \in \mathcal{Q}(M) \Leftrightarrow \mathrm{sgn}(N_{ij}) = \mathrm{sgn}(M_{ij}) \; \forall \; i, j$, i.e. the set of all matrices with the same sign pattern as M.

Lemma 34. *Suppose that S is an $n \times r$ constant real matrix, with each row of S containing a maximum of two non-zero elements, both of which have the same magnitude. Then S is strongly sign determined.*

Proof. Recall that if S is strongly sign determined, then every square submatrix of S is either singular or sign nonsingular. A square matrix is sign nonsingular if all matrices with the same sign structure are nonsingular, and therefore have determinants of the same sign.

The proof relies on an inductive method, beginning by asserting that every 2×2 submatrix of S is either singular or sign nonsingular. Any 2×2 matrix containing at least one zero entry is either singular or sign nonsingular by inspection. The situation where all four entries are non-zero can be divided into two cases. Any 2×2 matrix containing three non-zero entries of one sign and and one non-zero entry of the opposite sign is sign nonsingular. This is the first case. The second case is when all four entries have the same sign, or two entries have one sign and the other two entries have the opposite sign. In this case, since entries on the same row have the same magnitude, the matrix is singular. This establishes the result for 2×2 submatrices.

The inductive step now follows: suppose that all $k \times k$ square submatrices of S are either singular or sign nonsingular. Choose a $(k+1) \times (k+1)$ submatrix of S and call it P. If P contains a row or column of zeroes then it is singular. If P contains a row or column with only one non-zero entry, then its determinant is simply the non-zero entry multiplied by the determinant of a $k \times k$ submatrix, and P is therefore either singular or sign nonsingular by the induction hypothesis.

This leaves the case where no row or column of P contains less than two non-zero entries. Each row must therefore contain exactly two non-zero entries. Therefore P contains a total of $2(k + 1)$ non-zero entries. Since every column of P contains more than one non-zero entry, it follows that every column also contains exactly two non-zero entries.

Now suppose that P is **not** sign nonsingular. Therefore there exists some matrix R in the qualitative class of P such that R is singular. If R is singular, then there is some non-zero vector v lying in the (right) kernel of R, i.e. $Rv = 0$ for some $v \neq 0$. Consider this expression componentwise. Row i of R contains exactly two non-zero entries, the indices of which will be labelled j_1 and j_2. Therefore the ith component of Rv is

$$(Rv)_i = R_{ij_1}v_{j_1} + R_{ij_2}v_{j_2} = 0 \tag{6.21}$$

This leaves two possibilities. The first possibility is that $v_{j_1} = v_{j_2} = 0$, which trivially implies that $P_{ij_1}v_{j_1} + P_{ij_2}v_{j_2} = 0$. Since v is non-zero, there must be some component $v_{j_1} \neq 0$, which leads to the second possibility, that $R_{ij_1}v_{j_1} = -R_{ij_2}v_{j_2}$. Here $v_{j_2} \neq 0$ also, since $R_{ij_1}, R_{ij_2} \neq 0$. In this case, due to the relationship between R and P, and noting that $|P_{ij_1}| = |P_{ij_2}|$ by assumption, it can be seen that

$$P_{ij_1}\frac{v_{j_1}}{|v_{j_1}|} + P_{ij_2}\frac{v_{j_2}}{|v_{j_2}|} = 0$$

Therefore a new vector y can be defined as follows:

$$y_j = \left\{ \begin{array}{ll} 0, & v_j = 0 \\ \frac{v_j}{|v_j|}, & v_j \neq 0 \end{array} \right. \tag{6.22}$$

Under this definition, y is a non-zero vector lying in the (right) kernel of P. Therefore P is singular. It has just been shown that if P is not sign nonsingular then it is singular. This completes the proof. □

Remark. The mathematical constraint on S described in lemma 34 translates to the statement that each reactant can take part in no more than two reactions, and a reactant must have the same stoichiometry in every reaction. A trivial generalisation follows.

Lemma 35. *If S can be written $S'D$ where S' is an $n \times r$ real matrix containing at most two non-zero entries per row, both of which have the same magnitude, and D is a nonsingular diagonal $r \times r$ matrix, then S is strongly sign determined.*

Proof. As in §1.4.3, for a pair of sets $\alpha \subset \{1, \ldots, r\}$ and $\gamma \subset \{1, \ldots, r\}$ let $S_{(\alpha|\gamma)}$ be the submatrix of S with rows indexed by the elements of α and columns indexed by the elements of γ. It is easy to see that $S_{(\alpha|\gamma)} = S'_{(\alpha|\gamma)}D_{(\gamma|\gamma)}$. Any matrix $P \in \mathcal{Q}(S)$ with submatrices $P_{(\alpha|\gamma)} \in \mathcal{Q}(S_{(\alpha|\gamma)})$ can likewise be written $P = P'D$ with submatrices $P_{(\alpha|\gamma)} = P'_{(\alpha|\gamma)}D_{(\gamma|\gamma)}$. Since $D_{[\gamma|\gamma]}$ is of fixed sign, it follows that for $|\alpha| = |\gamma|$, $S_{(\alpha|\gamma)}$ is sign nonsingular if and only if $S'_{(\alpha|\gamma)}$ is. Thus S is strongly sign determined if and only if S' is, but it is known from lemma 34 that S' is strongly sign determined. This completes the proof. □

It is straightforward to see that lemma 35 can be generalised to include the case in which S can be written DS' where S' is an $n \times r$ real matrix containing at most two non-zero entries per **column**, both of the same magnitude, and D is a nonsingular diagonal $n \times n$ matrix.

Corollary 12. *Suppose a set of chemical reactions as described at the beginning of §6.5 has an $n \times r$ stoichiometric matrix S such that either*

(a) $S = S'D$, where S' is an $n \times r$ real matrix containing at most two non-zero entries per row, both of which have the same magnitude, and D is a nonsingular diagonal $r \times r$ matrix, or

(b) $S = DS'$, where S' is an $n \times r$ real matrix containing at most two non-zero entries per column, both of which have the same magnitude, and D is a nonsingular diagonal $n \times n$ matrix.

Then the reaction network has a unique equilibrium.

Proof. The assumptions made at the beginning of §6.5, that $S_{ij}V_{ji} \leq 0$ and $S_{ij} = 0 \rightarrow V_{ji} = 0$, mean that there is no one-step catalysis. It was shown in lemma 35 (and the comment directly after it) that conditions (a) and (b) imply that S is strongly sign determined. The Jacobian of the system is therefore a $P^{(-)}$ matrix, as in lemma 29 (p. 113). Hence the system is injective on closed rectangular subsets of $\mathbb{R}^n$, and consequently on the forward invariant globally absorbing set $\mathcal{C}(\mathcal{P})$.

The dynamical system representing the reaction network has at least one fixed point, which lies in $\mathcal{C}(\mathcal{P})$, as shown in theorem 35. Since it has been established that the system is injective on this region, the fixed point is necessarily unique. This completes the proof. □

There follows a trivial related lemma that may be of academic interest:

Lemma 36. *Let M be a $p \times q$ real matrix, with either every non-zero element on the same row of M having the same magnitude, or every non-zero element in the same column having the same magnitude. Define a matrix $\tilde{M}$ with elements $\tilde{M}_{ij} = \operatorname{sgn} M_{ij}$. Every minor of M has the same sign as the equivalent minor in $\tilde{M}$.*

Proof. When all non-zero elements on row (column) i of M have the same magnitude m_i, $M = D\tilde{M}$ ($M = \tilde{M}D$), where D is a diagonal matrix defined by $D_{ii} = m_i$. For every pair of index sets α, γ satisfying $|\alpha| = |\gamma|$, the corresponding square submatrix $M_{(\alpha|\gamma)}$ of M can be written $M_{(\alpha|\gamma)} = D_{(\alpha|\alpha)}\tilde{M}_{(\alpha|\gamma)}$ ($M_{(\alpha|\gamma)} = \tilde{M}_{(\alpha|\gamma)}D_{(\gamma|\gamma)}$). Since $M_{[\alpha|\gamma]} = D_{[\alpha|\alpha]}\tilde{M}_{[\alpha|\gamma]}$ ($M_{[\alpha|\gamma]} = \tilde{M}_{[\alpha|\gamma]}D_{[\gamma|\gamma]}$) and $D_{[\alpha|\alpha]}$ is strictly positive for any index set α, it follows immediately that $\operatorname{sgn} M_{[\alpha|\gamma]} = \operatorname{sgn} \tilde{M}_{[\alpha|\gamma]}$. □

6.5.2 Convergence to a unique fixed point

The focus now is on conditions that guarantee convergence of trajectories through use of the logarithmic norm method that was applied to single reactions in §6.4.2. For certain reaction network structures it is possible to find a similarity transform T such that $\mu_\infty(T^{-1}JT) < 0$. Unlike the single reaction case, monotonicity is not discussed in this section; for an extension of the monotonicity results that were applied to the single reaction case to reaction networks, see [Banaji, 2008] — note that this reference identifies monotone chemical reaction networks, but does not go on to explicitly prove global asymptotic stability.

From equation (3.3) (p. 44) it is known that $\mu(A+B) \leq \mu(A)+\mu(B)$, and since $J = SV-Q'$ it follows that $\mu_\infty(T^{-1}JT) \leq \mu_\infty(T^{-1}SVT) + \mu_\infty(T^{-1}(-Q')T)$. In a CFSTR, $Q' = qI$ (recall that q is the flow rate, which is assumed to be positive), so this simplifies to

$$\mu_\infty(T^{-1}(-Q')T) = \mu_\infty(T^{-1}(-qI)T) = q\,\mu_\infty(-I) = -q$$

for any choice of T. Therefore, in order to guarantee $\mu_\infty(T^{-1}JT) < 0$ for a set of reactions taking place in a CFSTR, it is only necessary to find T such that $\mu_\infty(T^{-1}SVT) \leq 0$, since q can be arbitrarily small but is always strictly positive. The case of more general Q' is not considered here, but is a potential area for future work.

Unlike in the single reaction case, J is not necessarily a $P^{(-)}$ matrix, so injectivity of the dynamical system is not guaranteed. However, in the event of there existing a negative logarithmic norm, uniqueness of the fixed point is guaranteed by theorem 20 (p. 46). It also turns out that the conditions that have been found on S that suffice to guarantee the existence of a suitable T are a subset of those found to suffice for J to be injective via lemma 34 (p. 127).

The restrictions on a reaction network described in the next two theorems are quite strong. It is important to note that they are **not** a set of necessary conditions for the existence of a negative logarithmic norm (and hence autonomous convergence of the system); it is highly likely that there are other possible sets of conditions that could be applied to a reaction network in order to guarantee the existence of a negative logarithmic norm. However, it is generally hard to identify suitable sets of conditions.

Theorem 36. *Suppose that there exists an $n \times n$ matrix T such that*

1. *T is equal to S with an extra $n - r$ columns added.*
2. *T is nonsingular.*
3. *Each row of T contains no more than two non-zero entries.*
4. *Every non-zero entry on a given row of T has the same magnitude.*

For this choice of T, $\mu_\infty(T^{-1}SVT) = 0$.

Proof. Consider a deconstruction of the transformed matrix $(T^{-1}SVT)$ into $(T^{-1}S)(VT)$. T and T^{-1} are $n \times n$ matrices, S is an $n \times r$ matrix, and V is an $r \times n$ matrix. Therefore $T^{-1}S$ is an $n \times r$ matrix, while VT is an $r \times n$ matrix. Since S is simply T with the last $n-r$ columns truncated, $T^{-1}S$ is the identity matrix in n dimensions, with the final $n-r$ columns removed. Consequently the first r rows of $T^{-1}SVT$ will simply be the rows of VT, while the remaining $n-r$ rows of $T^{-1}SVT$ will consist entirely of zeros.

Recall the definition of the μ_∞ norm of a matrix M:

$$\mu_\infty(M) = \max_i \left(\mathrm{Re}(M_{ii}) + \sum_{j \neq i} |M_{ij}| \right)$$

As the final $n-r$ rows of $T^{-1}SVT$ are made up of zeros,

$$\mathrm{Re}((T^{-1}SVT)_{ii}) + \sum_{j \neq i} |(T^{-1}SVT)_{ij}| = 0$$

for all $i > r$, and hence $\mu_\infty(T^{-1}SVT) \geq 0$. Since the rows of VT form the first r rows of $T^{-1}SVT$, the claim that $\mu_\infty(T^{-1}SVT) = 0$ holds true if and only if

$$\max_{i \in \{1,\ldots,r\}} \left(\mathrm{Re}((VT)_{ii}) + \sum_{j=1, j \neq i}^{n} |(VT)_{ij}| \right) \leq 0 \tag{6.23}$$

Recall that

$$V = \begin{pmatrix} V_{11} & \cdots & V_{1n} \\ \vdots & \ddots & \vdots \\ V_{r1} & \cdots & V_{rn} \end{pmatrix} \tag{6.24}$$

and T is defined to be

$$T = (\, S \,\vdots\, ? \,) = \left(\begin{array}{ccc:cc} S_{11} & \cdots & S_{1r} & ? & ? \\ \vdots & \ddots & \vdots & ? & ? \\ S_{n1} & \cdots & S_{nr} & ? & ? \end{array} \right) \tag{6.25}$$

By assumption, the elements in the $n \times (n-r)$ block marked with question marks are either zero or have the same magnitude as the (single) element on the same row of S. Let the elements of T be T_{ij}, with $T_{ij} = S_{ij} \, \forall \, j \leq r$. The ith element of the leading diagonal of VT is then $\sum_j V_{ij}T_{ji} = \sum_j S_{ji}V_{ij}$, which is nonpositive due to the assumption that $S_{ij}V_{ji} \leq 0$. The kth element on row i is $\sum_j V_{ij}T_{jk}$.

In order that the inequality in equation (6.23) be fulfilled, it is necessary and sufficient for the magnitude of element i on row i to be equal to or greater the sum of magnitudes of all the other elements on row i for $i = 1, \ldots, r$, i.e.:

$$\left|\sum_{j=1}^{n} S_{ji} V_{ij}\right| \geq \sum_{k=1, k\neq i}^{n} \left|\sum_{j=1}^{n} V_{ij} T_{jk}\right| \tag{6.26}$$

Consider this equation for fixed i. If $S_{ji} = 0$, then reactant j doesn't participate in reaction i. In this case, $V_{ij} = 0$ by assumption and therefore all terms corresponding to these values of i and j disappear from the inequality in equation (6.26). However, for given i, at least two distinct values of j must give non-zero S_{ji} since reaction i is assumed to be a true reaction. The corresponding reaction rate derivatives V_{ij} will also be non-zero in general, so the LHS of equation (6.26) contains between 2 and n non-zero terms for each i. Each of these non-zero terms in the sum on the LHS corresponding to an individual choice of values i, j has a corresponding set of terms on the RHS. By reversing the order of summation on the RHS, equation (6.26) the following relation can be constructed:

$$\left|\sum_{j=1}^{n} S_{ji} V_{ij}\right| \geq \sum_{j=1}^{n} \sum_{k=1, k\neq i}^{n} |V_{ij} T_{jk}| \tag{6.27}$$

Note that equation (6.27) implies equation (6.26), but not vice versa since the $V_{ij}T_{jk}$ terms are not necessarily all of the same sign. However, the $S_{ji}V_{ij}$ terms on the LHS **are** all of the same sign, so in order to fulfil the inequality in equation (6.27) it suffices to show that the following relation is satisfied for every $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, n\}$:

$$|S_{ji} V_{ij}| \geq \sum_{k=1, k\neq i}^{n} |V_{ij} T_{jk}| \tag{6.28}$$

Since, by assumption, T has no more than two non-zero entries per row, and one of these must be S_{ji} on the LHS of equation (6.28), there can be at most one value of k such that $T_{jk} \neq 0$ on the RHS of equation (6.28). If there are no non-zero terms on the RHS, i.e. $T_{jk} = 0$ for all $k \neq i$, then the inequality is satisfied since the RHS is zero. If there is one value of $k \neq i$ such that $T_{jk} \neq 0$, then since it was assumed that all non-zero entries on a row of T have the same magnitude it follows that $|S_{ji}| = |T_{jk}|$ for this value of k. In this case, both sides of equation (6.28) are equal and therefore the inequality is satisfied. This argument is valid for all i, j, and so the proof is complete. □

Having demonstrated a possible transform matrix T that can be used to construct a negative logarithmic norm for the Jacobian, the next step is to establish a set of necessary and sufficient conditions for a transform of this form to exist.

Theorem 37. *An $n \times n$ matrix T satisfying the requirements laid out in theorem 36 can be found if and only if the following conditions hold for S:*

1. *All non-zero elements on the same row of S have the same magnitude.*
2. *All rows of S contain a maximum of two non-zero elements.*
3. *The rows of S with two non-zero elements are linearly independent.*
4. *The columns of S are linearly independent.*

Proof. Note that conditions 3 and 4 together imply that $r \leq n$.

The proof begins by showing that the conditions on S are necessary. Clearly, if the first r columns of T are S, then T can be chosen with all non-zero elements on the same row having the same magnitude if and only if all non-zero elements on a row of S have the same magnitude. Likewise, any T with S forming its first r columns can only have two or less non-zero entries on each row if S contains two or less non-zero entries on every row. If any rows of S containing two non-zero entries are linearly dependent then the corresponding rows of T will also be linearly dependent since each row of T cannot contain more than two non-zero entries. All rows of S with two non-zero entries must therefore necessarily be linearly independent of one another. The final condition, that columns of S be linearly independent, is trivially necessary if T is to be nonsingular.

To prove sufficiency, it is assumed (without loss of generality) that the reaction system is indecomposable, i.e. every reaction shares at least one reactant with at least one other reaction. If this is not the case, the system can be considered as two or more sets of reactions involving disjoint sets of reactants.

If a set of reactions is indecomposable then at least one non-zero entry in each column of S must appear at the same index as a non-zero entry in another column of S. Consider the first column of S as a matrix C_1 and add the other columns one at a time. For the second column, choose a column that has a non-zero element in common with C_1, and label the new $n \times 2$ matrix C_2. C_2 therefore contains at least one row with two non-zero entries. Add another column that has a non-zero entry in common with one of the rows of C_1 or C_2, and label this $n \times 3$ matrix C_3. C_3 has a minimum of two rows with two non-zero entries. Continuing in this way, it is apparent that since S itself consists of r columns it must contain at least $r - 1$ rows with two non-zero entries.

Since there are r columns in S it follows that there can be at most r rows in S with two non-zero entries, as it is not possible for a set of more than r such rows to be linearly independent. Thus there are either r rows in S with two non-zero entries and $n - r$ rows in S with only one non-zero entry, or there are $r - 1$ rows in S with two non-zero entries and $n - r + 1$ rows in S with only one non-zero entry.

The next stage in the proof is to construct a suitable T. The first step in constructing T is to re-order the rows of S. Since the order of rows in S is arbitrary, this makes no difference to the argument. Group all of the rows containing two non-zero elements together into a matrix E_2. The remaining rows with only one non-zero entry are then grouped together into another submatrix E_1. Now construct T out of four submatrices as follows: E_2 appears on the top left, and E_1 appears on the bottom left. This constitutes S in reordered form. Then add a square diagonal $(n-r)$-dimensional matrix M at the bottom right (entries defined below), and an $r \times (n-r)$ zero matrix, which will be called Z_0, at the top right.

Let element M_{ii} of M be equal to the non-zero element on the row of E_1 that lies on the same row of T as M_{ii}. In the event that E_1 has $n-r$ rows this is simply the non-zero element from row i of E_1, but when E_1 has $n-r+1$ rows then this corresponds to row $i+1$ of E_1. The matrix T therefore satisfies requirements 1, 3 and 4 of theorem 36 by inspection. All that remains is to verify it also satisfies requirement 2, namely $|T| \neq 0$.

For notational convenience, define an $(n-r) \times r$ matrix E_1^* and an $r \times r$ matrix E_2^* as follows: If E_2 contains $r-1$ rows then E_2^* is E_2 with the first row of E_1 added on the end, and E_1^* is E_1 with the first row removed. If, however, E_2 contains r rows then simply $E_1^* = E_1$ and $E_2^* = E_2$. This is a trivial redefinition in the sense that

$$S = \begin{pmatrix} E_2 \\ \hdashline E_1 \end{pmatrix} = \begin{pmatrix} E_2^* \\ \hdashline E_1^* \end{pmatrix}$$

However, it allows T to be written in block form as follows:

$$T = \left(\begin{array}{c:c} E_2^* & Z_0 \\ \hdashline E_1^* & M \end{array}\right) \tag{6.29}$$

Here the blocks on the diagonal are square. Since the determinant of a matrix is unchanged by switching rows and/or columns, $|T| = |T'|$, where

$$T' = \left(\begin{array}{c:c} M & E_1^* \\ \hdashline Z_0 & E_2^* \end{array}\right)$$

From lemma 2 (p. 24) it is known that

$$|T'| = |E_2^* - Z_0 M^{-1} E_1^*|\,|M|$$

Since Z_0 is a zero matrix, this simplifies to $|T| = |T'| = |E_2^*|\,|M|$. M is trivially nonsingular, so in order to show that T is nonsingular it suffices to show that E_2^* is nonsingular.

In the event that there were r rows in S with two non-zero entries, then $E_2^* = E_2$, which is nonsingular by the assumption that the rows of S containing two non-zero entries are linearly independent. This leaves the case where E_2^* is E_2 with an extra row containing only one non-zero entry added on the end.

To prove that E_2^* is nonsingular in this case, two families of sets will be used. γ_i will denote sets whose members are rows of E_2^*, and ι_i will denote subsets of the set $\{1, \ldots, r\}$. Let γ_1 be a set consisting of the final row of E_2^*, which by assumption has only one non-zero entry, and let ι_1 be a set consisting of the index of the non-zero entry of this row. ι_1 can then be used to induce another set: let γ_2 be the set of rows of E_2^* that have a non-zero entry at the index in ι_1, not including the row in γ_1. Then let ι_2 be the set of indices of the non-zero entries of rows in γ_2. Continuing in this way, γ_p can be defined as the set of rows that have a non-zero entry at an index in ι_{p-1}, but aren't in $\gamma_1 \cup \ldots \cup \gamma_{p-1}$, with ι_p being the set of indices of non-zero entries in the rows in γ_p. The indecomposability assumption means that there exists some q such that $\iota_1 \cup \ldots \cup \iota_q = \{1, \ldots, r\}$. Consequently $\gamma_1 \cup \ldots \cup \gamma_{q+1}$ contains every row of E_2^*.

Now suppose that E_2^* is singular. Consider a vector v such that $E_2^* v = 0$. Clearly v must have a zero entry at the index in ι_1, as otherwise $uv \neq 0$, where $u \in \gamma_1$. This in turn means that v must also have zeros at all of the indices in ι_2, since every row in γ_2 contains two non-zero entries, one of which is at the index in ι_1. By iterating this argument it is apparent that if v has zeros at the indices in ι_{p-1} it must also have zeros at the indices in ι_p. Since $\iota_1 \cup \ldots \cup \iota_q = \{1, \ldots, r\}$, v is therefore the zero vector, and so E_2^* is nonsingular. Consequently $|T| \neq 0$. □

An example may help to illustrate what the conditions on a reaction network stated in the above theorems look like in practical terms. Consider the pair of reactions

$$\begin{aligned} \mathrm{A} &\rightleftharpoons 2\mathrm{B} + \mathrm{C} \\ 2\mathrm{B} &\rightleftharpoons \mathrm{C} \end{aligned}$$

The stoichiometric matrix of this reaction pair is

$$S = \begin{pmatrix} 1 & 0 \\ -2 & 2 \\ -1 & -1 \end{pmatrix}$$

Thus S fulfils the conditions required in theorem 37, and a matrix T can be constructed fulfilling the requirements of theorem 36. One such possible T is

$$T = \begin{pmatrix} 1 & 0 & 1 \\ -2 & 2 & 0 \\ -1 & -1 & 0 \end{pmatrix}$$

In a slight abuse of notation, V for this system can be written as

$$V = \begin{pmatrix} -|V_{11}| & |V_{12}| & |V_{13}| \\ 0 & -|V_{22}| & |V_{23}| \end{pmatrix}$$

Note that $V_{21} = 0$ since $S_{12} = 0$, and the other elements of V have been written as signed absolute values in order to highlight the assumption that S and V have opposite sign structure, i.e. $S_{ij}V_{ji} \leq 0$. Putting the above matrices into the open source algebra program [Maxima, 2008] then yields

$$T^{-1}SVT = \begin{pmatrix} -(|V_{11}| + 2|V_{12}| + |V_{13}|) & 2|V_{12}| - |V_{13}| & -|V_{11}| \\ 2|V_{22}| - |V_{23}| & -(2|V_{22}| + |V_{23}|) & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

By putting this expression into the equation for the μ_∞ logarithmic norm (equation (3.6), p. 45), it is straightforward to verify that $\mu_\infty(T^{-1}SVT) = 0$, as claimed in theorem 36.

The final result of this section combines the previous results to show that when a set of reactions taking place in a CFSTR have a network structure as described above, all initial conditions converge to a unique equilibrium:

Corollary 13. *Suppose that a set of reactions taking place in a CFSTR satisfy the assumptions made at the beginning of §6.5, and that the stoichiometric matrix of the reactions obeys the following conditions:*

1. *All non-zero elements on the same row of S have the same magnitude.*
2. *All rows of S contain a maximum of two non-zero elements.*
3. *The rows of S with two non-zero elements are linearly independent.*
4. *The columns of S are linearly independent.*

Then the dynamical system representing the reaction network is globally asymptotically stable.

Proof. It was demonstrated in §6.5.1 that all trajectories of the system enter a convex compact forward invariant set, and that there must be at least one fixed point.

The system has Jacobian $J = SV - qI$. By theorems 36 and 37, there exists an invertible matrix T such that $\mu_\infty(T^{-1}SVT) = 0$. As noted at the beginning of §6.5.2, for any invertible T, $\mu_\infty(T^{-1}(-qI)T) = -q$. This means that by equation (3.3), $\mu_\infty(T^{-1}JT) \leq -q < 0$. This in turn implies by lemma 8 that there exists a logarithmic norm $\mu_{\infty,T^{-1}}$ such that $\mu_{\infty,T^{-1}}(J) < 0$. Therefore, by theorem 20 (p. 46), there is only one fixed point, and it is globally asymptotically stable. □

Remark. Notice that the conditions on S that guarantee global convergence to a unique fixed point via corollary 13 are a stricter subset of those required for uniqueness of the fixed point in lemma 34 (p. 127).

6.6 Conclusions

In this chapter, the dynamics of a set of chemical reactions were investigated. The possible behaviour of a single reaction was investigated using both monotonicity and autonomous convergence, partly because it is of interest to see how different areas of theory can be used to get similar results, and partly as a first step in generalising to systems of multiple reactions, where the two techniques may lead to different criteria for convergence. For a single reaction it was shown that, under fairly mild constraints on the reaction kinetics and structure, all initial conditions converge to a unique steady state solution if the reaction takes place in a CFSTR. The same behaviour was also shown to hold for slightly more general flow rates than those allowed by a CFSTR. For multiple reactions, it was demonstrated using autonomous convergence only that certain reaction network structures, when taking place in a CFSTR, globally converge to a unique steady state. Li and Muldowney's autonomous convergence theorem (theorem 23, p. 53), which involves the second additive compound of the Jacobian matrix, was not applied to the multiple reaction case, but it would be possible to develop results using this area of theory. As suggested at the end of chapter 3, a similar type of result might make it possible to rule out periodic behaviour in a reaction network even in the event that there is more than one steady state.

The results in this chapter suggest several areas for further work. One promising avenue is to investigate how monotonicity of a single reaction can be used to find networks of reactions that are collectively monotone with respect to some ordering, and then to analyse the asymptotic behaviour of such a reaction network. A substantial amount of work has been done along these lines in [Banaji, 2008], but there is room for further development. The monotonicity results also potentially lend themselves to a graph-theoretic formulation, along the lines of [Craciun and Feinberg, 2006b] and [Kunze and Siegel, 2002a].

In a similar vein, an attempt to identify families of norms that can be used to demonstrate convergence of a single chemical reaction could lead to the characterisation of more reaction network structures that are globally stable via autonomous convergence. It would also be worthwhile to find generalisations of the CFSTR flow conditions under which the reaction networks remain globally stable, as this would potentially increase the applicability of the results to chemical reaction networks that occur in biology. It is worth pointing out that these areas of further work all seem to lead to mathematical problems that would be difficult to solve in full generality; it is possible that if the reaction kinetics were more restricted, e.g. to mass-action only, it would be easier to generalise the existing results to more general network structures and flow conditions.

Chapter 7

Summary and discussion

In this thesis, a number of analytical techniques have been applied to broad classes of biological and chemical models. The key aspect of this process has been the construction and analysis of models using, as far as possible, qualitative information instead of quantitative information, such as assuming that functions are monotone in their arguments, or that one parameter takes a larger value than another. The implicit approach taken has been to identify systems where global asymptotic stability might reasonably be expected, and then prove that this behaviour persists across a whole class of models that might represent the system.

This is particularly relevant when modelling processes that occur in biology, as it is often hard to get accurate measurements of all the parameters involved in a given process. For many biological systems it is possible to get a good understanding of the structure, e.g. what interacts with what, and whether one element activates or inhibits another. Modern high-throughput experimental techniques, such as microarrays, identify entities that interact with each other without providing detailed information about the nature of the interaction. However, when more in depth experiments are performed to investigate a biological process, getting precise numerical measurements of the interactions, e.g. how fast a reaction occurs, is often problematic in practice. Some biological quantities are inherently variable between individuals, such as the radius of blood vessels. Other quantities may not exhibit much variability, but can still be difficult to measure for practical reasons.

The approach taken in this thesis was originally motivated by the type of numerical model exemplified in [Banaji et al., 2005]. Many such complex models contain subsystems that exhibit simple behaviour, such as global asymptotic stability. Identification and analysis of these subsystems in isolation can yield useful insight into their structure and behaviour, potentially simplifying the process of constructing the larger model. Additionally, due to the difficulties in getting precise measured data, models of this type are often constructed using a mixture of in vivo measurements from humans and in vivo/in vitro measurements made on equivalent processes in animals. In some cases, inaccuracies in the parameters chosen for the model, arising from difficulties in obtaining data, may make the predicted behaviour qualitatively different to the real world behaviour of the system. For this reason, qualitative models can be useful in gaining a deeper understanding of how the system they represent behaves, by identifying the behaviour of all possible instantiations of the model that satisfy a certain set of assumptions. In some cases it is possible to draw quite strong conclusions about the behaviour of a whole class of systems based on minimal assumptions.

7.1 Electron transport processes

Chapter 4 discussed a qualitative model of the mitochondrial electron transport chain, which has been numerically modelled in a number of papers, e.g. [Korzeniewski, 1996] and [Beard, 2005]. It was demonstrated that the electron transport chain must necessarily have one (and only one) equilibrium, due to the fact that its trajectories are forwardly bounded and its Jacobian is a $P^{(-)}$ matrix. For chains involving only two or three electron transfer reactions it was also demonstrated via an autonomous convergence theorem that the equilibrium must be globally asymptotically stable, but for longer chains this conclusion no longer holds.

There are a number of ways in which the qualitative model presented in chapter 4 could potentially be extended. The obvious next step is to attempt to further characterise the behaviour of the system when it is not globally asymptotically stable: it appears possible that the system may undergo a Hopf bifurcation, implying the existence of a periodic orbit at some parameter values, which is of potential experimental interest. Alternatively, as suggested in §4.4 of [Donnell et al., 2008], the reaction rates of the electron transport chain may satisfy extra conditions not included in the model presented here, which suffice to guarantee that the Jacobian remains Hurwitz at all points for a chain of any length. Some numerical models of the electron transport chain, such as [Korzeniewski, 1996], model the reaction rates in such a way that these extra assumptions are met and the Jacobian of the system is Hurwitz at all points. This rules out the possibility of a Hopf bifurcation, but it is an open question as to whether it guarantees global asymptotic stability of the electron transport chain model.

The other obvious area for extension of the electron transport chain model would be to include topologically more complicated electron transfer networks. General electron transfer networks were analysed in [Banaji and Baigent, 2008], under the assumption that the proton gradient across the mitochondrial membrane is fixed. As the discussion and differing conclusions in [Banaji, 2006] (electron transport chain with fixed proton gradient) and [Donnell et al., 2008] (electron transport chain with varying proton gradient) show, allowing the proton gradient to vary in electron transport models makes the behaviour much more complicated to analyse and potentially qualitatively different to the same model with a fixed proton gradient. For this reason, analysing generalised electron transfer networks with a varying proton gradient could yield some interesting results.

7.2 Cellular gap junctions

It is a little more difficult to see where useful extensions could be made to the model of the cellular gap junction in chapter 5. The model was based on work that appeared in [Baigent et al., 1997] and [Baigent, 2003]. It was demonstrated in chapter 5 that for any number of cells in a line joined by gap junctions, all trajectories are bounded and therefore there is at least one equilibrium.

When the system consists of only three cells in a line, separated by gap junctions in which the conduction channels have only two possible states, conditions regarding the relationship between the intercellular voltage and the probability of the conduction channels changing between the high and low conduction states were found under which there can be no more than one equilibrium. It was also demonstrated that the equilibrium is locally asymptotically stable when these conditions are slightly strengthened. However, no results pertaining to the global asymptotics of the system were found.

Similar conditions regarding the transfer between conducting states of the conduction channels, when applied to a system consisting of two cells joined by a two state gap junction, were used to show that the system has a unique, locally asymptotically stable equilibrium. When the conditions were reversed it was shown that the system's solutions preserve an ordering within a globally attracting forward invariant set, and the conditions for the two cell system were strengthened in a similar way to the three cell case, it was shown that the system is strongly monotone within the same invariant set and every initial condition converges to an equilibrium.

Since in real cellular networks the cells tend to form more complicated networks than just a straight line, it would be worthwhile to extend the model to more general topologies. However, while it should be relatively straightforward to demonstrate boundedness of solutions, making any stronger claims about the global behaviour of the system appears to be very difficult to do. The autonomous convergence techniques presented in chapter 3 do not seem to be applicable, and so far no linear transform has been found that makes the system monotone in an invariant trapping region for more than two cells joined by a two state gap junction. The transform used for the two cell, two state system guarantees monotonicity of the system provided that the intercellular voltage is of fixed sign. It happens that the system with two cells and two states has a globally attracting forward invariant set in which this condition holds, so the system is monotone within this invariant set. When extra cells are added, it is still possible to construct a similar transform that guarantees monotonicity when the intercellular voltages are of fixed sign, but unfortunately there no longer appears to be a globally attracting invariant set in phase space where this is the case. Whether the system with more than two cells is monotone (or its trajectories enter a forward invariant set in which the system is monotone) with respect to some as yet undiscovered ordering, such as an ordering defined by a non-simplicial cone, is an open question.

The problem when extra conduction channels are added is different, and relates to how the probability of transitions between the conduction states in a gap junction varies with the voltage across the junction. It seems likely that it would be possible to construct a set of conditions on the transitions between states that guarantee monotonicity of the system, but whether a physically reasonable set of such conditions can be found is much less certain. The probability of transition between each pair of states must be restricted, which in the case where there are only two states is fairly simple, but for three or more states becomes increasingly complicated. [Baigent, 2003] demonstrates monotonicity in a model of two cells joined by a gap junction in which the conduction channels have more than two possible states, but this result relies on choosing a functional form for the transition probabilities. It would be interesting to know whether constraining the system in this way is necessary to limit the dynamics when there are three or more conduction states, or if the solutions of the system are still monotone without this extra structure.

7.3 Chemical reaction networks

The work on chemical reactions in chapter 6 is much more open ended than the earlier applications. It was demonstrated that, subject to certain minimal assumptions, any single reaction in which each reactant appears on only one side of the reaction, occurring in a continuous flow stirred tank reactor (CFSTR), will globally converge to a unique equilibrium. This result was proved using both the theory of monotone flows and autonomous convergence theory. While the global convergence of a single reaction of this type is not unexpected and could almost certainly be proved using simpler methods, the exercise of applying a variety of techniques to the problem is potentially of interest.

Further to the single reaction case, it was demonstrated using autonomous convergence theory that a network of reactions with some reactants in common will also converge to a unique equilibrium for some network structures, when the same assumptions that were made for the single reaction are made for each individual reaction in the network. In particular, the reaction kinetics were not specified, it was simply assumed that the rate of a given reaction is a monotone function of the concentrations of the reactants that take part in it. [Banaji, 2008] addressed the problem of identifying networks of chemical reactions that correspond to monotone dynamical systems by initially considering a single reaction and then extending the results to cover various network structures; in a similar way it might be possible to extend the results in chapter 6 by attempting to identify all norms that can be used to demonstrate autonomous convergence for a single reaction, and then identifying networks of reactions in which the norms for each reaction overlap. Another possible approach to identifying networks of chemical reactions that globally converge would be to use a graph-theoretic technique to demonstrate monotonicity of flows, in a similar way to [Craciun and Feinberg, 2006b] and [Kunze and Siegel, 2002a].

The extensions suggested so far relate primarily to the structure of the reaction network, but there are other possible directions for generalisation, such as by relaxing the flow conditions. Most of the results in chapter 6 were proved in the context of a CFSTR, which is a concept from theoretical chemistry. The associated flow conditions are fairly strict, so generalising these conditions would increase the model's applicability to biological problems. There are a number of possible approaches to generalising the results in this way. The work that appears in [Craciun and Feinberg, 2006a] addresses a related problem, that of injectivity of solutions of a set of chemical reactions taking place in a reactor where some of the reactants cannot flow in or out. In [Craciun and Feinberg, 2005], a set of conditions were identified under which a set of reactions in a CFSTR could only have one equilibrium. In [Craciun and Feinberg, 2006a], the flow assumptions were weakened by assuming that some chemical species did not flow into or out of the tank, corresponding to the idea of enzymes being trapped by a membrane. It was then demonstrated that if the dynamical system representing a reaction network was injective when the reactions took place in a CFSTR, then the same system where some of the reactants could not flow in or out can have no more than one nondegenerate positive equilibrium within any so called compatibility class. Perhaps an analogous approach might be useful for identifying reaction networks that are monotone or exhibit autonomous convergence when some species do not flow in or out.

Another possible approach for generalising both the reaction network structure and the flow conditions is to try and identify invariant submanifolds to which all solutions converge. In [Banaji, 2008] it was demonstrated that a set of reactions taking place in a CFSTR converge to an invariant submanifold, and therefore only the dynamics on this submanifold need be analysed. [Banaji, 2008] examined this idea in terms of monotonicity of flows, but autonomous convergence theory might lead to useful results, either via explicit construction of a reduced system on the submanifold or by using results from [Li and Muldowney, 2000]. These same techniques might also be applicable if a globally attracting submanifold could be found for a set of reactions under more general outflow conditions than that of a CFSTR.

The identification of invariant submanifolds is also relevant to a closed system, i.e. a set of reactions with no inflow or outflow. Conditions for such a system to be monotone were examined in [Banaji, 2008] but the consequences of monotonicity were not investigated; these results could potentially be extended by further analysing the possible asymptotic behaviour via the theory of monotone flows, or by applying autonomous convergence theory.

A third area for investigation is that of reaction networks with different kinetics. In chapter 6, minimal assumptions were made about the kinetics of the reactions. Additional results might also be found by adding extra assumptions about the kinetics of the reactions. [Banaji et al., 2007] proved that certain systems of reactions can have at most one equilibrium, based only on their structure and an assumption that the kinetics are monotone. By further restricting the kinetics to mass action only, weaker conditions were found under which the set of reactions could only have one equilibrium. A similar approach might allow application of the theory of monotone flows or autonomous convergence to a broader class of reaction networks.

Another possible extension of the chemical reaction model relates back to the electron transport model in chapter 4. The electron transport chain without feedback is essentially a chain of interconversion reactions, as implied in [Banaji et al., 2007], [Donnell et al., 2008] and [Banaji and Baigent, 2008]. This opens up the possibility of examining networks of electron transfer reactions (or indeed other chemical reaction networks) coupled to non-chemical processes, such as the proton gradient that occurs in the electron transport chain.

The final apparent area for development of the chemical reaction model is in generalising it to other areas of biology that have similar structure. Models of immunology/epidemiology could be thought of as similar chemical reaction networks, for example consider the infection "reaction"

$$I + S \overset{P}{\rightarrow} 2I$$

Here I could represent an infected individual, S could represent a susceptible individual, and P could be the infection process. By incorporating other equations representing recovery, birth and death of infections, a pseudo chemical reaction network could be constructed, which might be analysed using a generalisation of the results presented here, although account would be need to be made of the fact that many of the "reactions" are irreversible, and may include one-step catalysis. Gene regulatory networks are also at their heart chemical reaction networks, although as with immunological models, the assumption that there can be no one-step catalysis may need to be weakened in order to treat these systems.

7.4 Theoretical concepts

In addition to possible extensions of the applications investigated in the thesis, some of the results also suggest interesting work relating to the underlying theory. While the theory of monotone flows has been fairly extensively developed and applied, it is still a fruitful area of research, both in terms of identifying systems that are monotone and then exploring, as far as possible, what the consequences of monotonicity are for a system. Knowing that a system is monotone is valuable information in itself, but making stronger claims about the dynamics of a system, e.g. identifying conditions for global asymptotic stability, is always desirable. There is certainly room for further discoveries in this area. Finding a set of conditions that guarantee stronger forms of monotonicity, such as the results presented in [Kunze and Siegel, 2002a], is often helpful when characterising the possible asymptotics of a system.

The situation with autonomous convergence theory appears more open. Whereas applications of the theory of monotone flows appear extensively in the literature, applications of the various autonomous convergence theorems are seemingly few and far between, examples being [Arino et al., 2003] and [Ballyk et al., 2005]. Identifying potential applications of autonomous convergence theory should prove a fruitful area of research, since the theory is seemingly not widely known.

It is unclear why autonomous convergence theory has not been more widely applied, considering how popular the theory of monotone flows has become. Part of the reason is probably that the theory of monotone flows has been established for longer, and is applicable to both autonomous and nonautonomous dynamical systems, whereas the autonomous convergence theorems discussed in this thesis (as the name suggests) are only applicable to autonomous systems. It may also be that monotone dynamical systems are generally easier to identify. It is usually trivial to check whether a given dynamical system is cooperative; finding a linear transform that makes a dynamical system cooperative (and thereby proving that its solutions preserve an ordering defined by a simplicial cone) is more difficult, but often still tractable, and it is in these two areas that most results regarding monotonicity have been published.

For the autonomous convergence theorems, the identification of convergent systems is more difficult. Verifying whether the Jacobian of a given dynamical system has a negative logarithmic norm is not a trivial task, since there are an infinite number of possible logarithmic norms and the forms of very few of them appear to be explicitly known, nor does it appear easy in general to check whether a logarithmic norm, the form of which is known, of a matrix is negative. Once linear transforms and second (or higher order) additive compounds are added in, the situation becomes even more complex. For this reason, any results that simplify the task of identifying useful logarithmic norms for classes of dynamical systems would be of great potential value in applying autonomous convergence theory.

The other area of particular theoretical interest that the applications in this thesis touch upon is the group of results guaranteeing uniqueness of any fixed points. The structures of the Jacobians of the electron transport chain in chapter 4 and the chemical reaction networks in chapter 6 guarantee that they are $P^{(-)}$ matrices; therefore the vector field in each application is injective on rectangular regions of $\mathbb{R}^n$, as shown in [Gale and Nikaido, 1965], and consequently there cannot be more than one fixed point. Boundedness of solutions in each system guarantees that there is a fixed point. Alternatively, the fact that the Jacobian of each system is nonsingular in some closed, compact, simply connected subset of phase space, combined with the fact that the vector field points inwards on the boundary of this subset, also guarantees the existence of a unique fixed point via degree theory. A similar pair of results involving $P^{(-)}$ matrices and degree theory appears in the gap junction model in chapter 5 when assumptions are made that guarantee the determinant of the Jacobian is of fixed sign. Whether these results are merely a curiosity arising from the structures of the systems under consideration, or hint at some deeper relationship between P matrices, degree theory and injectivity of a function, is unclear at the time of writing. There are many approaches to proving injectivity of a function that do not involve P matrices, e.g. [Smyth and Xavier, 1996] and [Fernandes et al., 2004], and some of these may be applicable to qualitative models.

Acknowledgements

The writing of this thesis and the research that went into owes a great deal to a number of people and organisations. I'd like to thank those who helped:

First, I must acknowledge the financial support of MIAS-IRC (with particular thanks to Professor Dave Delpy), without which this work would not have been possible.

Second, my supervisors: thanks to Steve and Murad for their patience, good advice, feedback on my work (particularly thesis drafts), and many peppermint teas! Having spoken to a number of other PhD students in recent years, I am (sometimes uncomfortably) aware of how lucky I've been to be supervised by people who, despite their other commitments, have made the time to see me regularly and discuss the work we've been doing. Often the highlight of my week has been meeting up and talking about some interesting maths. It's been a pleasure working with them both, they've taught me a lot, and I hope that we will stay in touch.

Third, without the support of my family I would never have had the chance to study for a PhD, and completing this thesis would have been much more difficult without them putting me up for (too) many months. Love and thanks to all of them. I'm sorry I wasn't around to see them more during my studies.

Fourth, thanks to my many housemates at Kitson Road over the years I spent in London, for all the good times! Of them, I'd particularly like to thank Joe Parker and James Kirkpatrick for the many interesting scientific discussions we had during the time that I lived with them.

Fifth, I would like to the thank my examiners, whose comments helped to make this thesis better.

Finally, both the preparation of the thesis and the mathematical work that went into it used a large number of free and open source computer programs. While I don't personally know any of the developers, I have benefited from the hard work of a whole community of people around the world, for which I am grateful. Too many programs were used to list them all here individually; the projects that I relied on the most were Ubuntu and its variants, Kile and the texlive version of the LaTeX typesetting system, Inkscape, Maxima and Scilab. Without these, my work would have been much more difficult.

Bibliography

[Agarwal and Lakshmikantham, 1993] Agarwal, R. P. and Lakshmikantham, V. (1993). *Uniqueness and nonuniqueness criteria for ordinary differential equations*, volume 6 of *Series in Real Analysis*. World Scientific.

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland Science, fourth edition.

[Allen and Bridges, 2002] Allen, L. and Bridges, T. J. (2002). Numerical exterior algebra and the compound matrix method. *Numerische Mathematik*, 92(2):197–232.

[Angeli et al., 2004] Angeli, D., Ferrell, Jr, J. E., and Sontag, E. D. (2004). Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *PNAS*, 101(7):1822–1827.

[Angeli and Sontag, 2003] Angeli, D. and Sontag, E. D. (2003). Monotone control systems. *IEEE Transactions of Automatic Control*, 48(10).

[Angeli and Sontag, 2004a] Angeli, D. and Sontag, E. D. (2004a). Interconnections of monotone systems with steady-state characteristics. In de Queiroz, M. S., Malisoff, M., and Wolenski, P., editors, *Optimal Control, Stabilisation and Nonsmooth Analysis*. Springer.

[Angeli and Sontag, 2004b] Angeli, D. and Sontag, E. D. (2004b). Multi-stability in monotone input/output systems. *Systems & Control Letters*, 51(3–4):185–202.

[Arino et al., 2003] Arino, J., McCluskey, C. C., and van den Driessche, P. (2003). Global results for an epidemic model with vaccination that exhibits backward bifurcation. *SIAM Journal on Applied Mathematics*, 64(1):260–276.

[Baigent, 2003] Baigent, S. (2003). Cells coupled by voltage-dependent gap junctions: the asymptotic dynamical limit. *BioSystems*, 68:213–222.

[Baigent et al., 1997] Baigent, S., Stark, J., and Warner, A. (1997). Modelling the effect of gap junction nonlinearities in systems of coupled cells. *Journal of Theoretical Biology*, 186:223–239.

[Ballyk et al., 2005] Ballyk, M. M., McCluskey, C. C., and Wolkowicz, G. S. K. (2005). Global analysis of competition for perfectly substitutable resources with linear response. *Journal of Mathematical Biology*, 51:458–490.

[Banaji, 2006] Banaji, M. (2006). A generic model of electron transport in mitochondria. *Journal of Theoretical Biology*, 243(4):501–516.

[Banaji, 2008] Banaji, M. (2008). Monotonicity in chemical reaction systems. *Dynamical Systems*. DOI:10.1080/14689360802243813.

[Banaji and Baigent, 2008] Banaji, M. and Baigent, S. (2008). Electron transfer networks. *Journal of Mathematical Chemistry*, 43(4):1355–1370.

[Banaji et al., 2007] Banaji, M., Donnell, P., and Baigent, S. (2007). P matrix properties, injectivity and stability in chemical reaction systems. *SIAM Journal on Applied Mathematics*, 67(6):1523–1547.

[Banaji et al., 2005] Banaji, M., Tachtsidis, I., Delpy, D., and Baigent, S. (2005). A physiological model of cerebral blood flow control. *Mathematical Biosciences*, 194:125–173.

[Barker and Schneider, 1975] Barker, G. P. and Schneider, H. (1975). Algebraic Perron-Frobenius theory. *Linear Algebra and its Applications*, 11(3):219–233.

[Beard, 2005] Beard, D. A. (2005). A biophysical model of the mitochondrial respiratory system and oxidative phosphorylation. *PLoS Computational Biology*, 1(4):e36.

[Belevich et al., 2006] Belevich, I., Verkhovsky, M., and Wikström, M. (2006). Proton-coupled electron transfer drives the proton pump of cytochrome *c* oxidase. *Nature*, 440(6):829–832.

[Bellman, 1968] Bellman, R. (1968). *Modern Elementary Differential Equations*. Addison–Wesley Series in Mathematics. Addison–Wesley.

[Berger and Gostiaux, 1988] Berger, M. and Gostiaux, B. (1988). *Differential Geometry: Manifolds, Curves and Surfaces*, volume 115 of *Graduate Texts in Mathematics*. Springer-Verlag. Translated into English by Silvio Levy.

[Berman and Plemmons, 1994] Berman, A. and Plemmons, R. J. (1994). *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial and Applied Mathematics.

[Bernat and Llibre, 1996] Bernat, J. and Llibre, J. (1996). Counterexample to Kalman and Markus-Yamabe conjectures in larger than 3. *Dynamics of Continuous, Discrete and Impulsive Systems*, 3(2):337–379.

[Bhagavan, 2002] Bhagavan, N. V. (2002). *Medical Biochemistry*. Harcourt/Academic Press.

[Bishop and Goldberg, 1980] Bishop, R. L. and Goldberg, S. I. (1980). *Tensor Analysis on Manifolds*. Courier Dover Publications.

[Brand et al., 1994] Brand, M. D., Chien, L., and Diolez, P. (1994). Experimental discrimination between proton leak and redox slip during mitochondrial electron transport. *Biochemical Journal*, 297(1):27–29.

[Brualdi and Shader, 1995] Brualdi, R. A. and Shader, B. L. (1995). *Matrices of sign-solvable linear systems*, volume 116 of *Cambridge tracts in mathematics*. Cambridge University Press.

[Canton et al., 1995] Canton, M., Luvisetto, S., Schmehl, I., and Azzone, G. (1995). The nature of mitochondrial respiration and discrimination between membrane and pump properties. *Biochemical Journal*, 310:477–81.

[Ciesielski, 2001] Ciesielski, K. (2001). On the Poincaré-Bendixson theorem. In Kryszewski, W. and Nowakowski, A., editors, *Lecture Notes in Nonlinear Analysis, Vol. 3. Proceedings of the 3rd Polish Symposium on Nonlinear Analysis*, pages 49–69.

[Cima et al., 1997] Cima, A., van den Essen, A., Gasull, A., Hubbers, E., and Manosas, F. (1997). A polynomial counterexample to the Markus-Yamabe conjecture. *Advances in Mathematics*, 131(2):453–457.

[Coddington and Levinson, 1955] Coddington, E. A. and Levinson, N. (1955). *Theory of Ordinary Differential Equations*. Mcgraw–Hill.

[Craciun and Feinberg, 2005] Craciun, G. and Feinberg, M. (2005). Multiple equilibria in complex chemical reaction networks: I. The injectivity property. *SIAM Journal on Applied Mathematics*, 65(5):1526–1546.

[Craciun and Feinberg, 2006a] Craciun, G. and Feinberg, M. (2006a). Multiple equilibria in complex chemical reaction networks: extensions to trapped species models. *IEE Proceedings — Systems Biology*, 153(4):179–186.

[Craciun and Feinberg, 2006b] Craciun, G. and Feinberg, M. (2006b). Multiple equilibria in complex chemical reaction networks: II. The species-reaction graph. *SIAM Journal on Applied Mathematics*, 66(4):1321–1338.

[De Leenheer et al., 2007] De Leenheer, P., Angeli, D., and Sontag, E. D. (2007). Monotone chemical reaction networks. *Journal of Mathematical Chemistry*, 41(3):295–314.

[Donnell et al., 2008] Donnell, P., Banaji, M., and Baigent, S. (2008). Stability in generic mitochondrial models. *Journal of Mathematical Chemistry*. DOI:10.1007/s10910-008-9464-6.

[Enciso et al., 2006] Enciso, G. A., Smith, H. L., and Sontag, E. D. (2006). Nonmonotone systems decomposable into monotone systems with negative feedback. *Journal of Differential Equations*, 224(1):205–227.

[Érdi and Tóth, 1989] Érdi, P. and Tóth, J. (1989). *Mathematical models of chemical reactions.* Manchester University Press.

[Fang, 1989] Fang, L. (1989). On the spectra of P- and P_0-matrices. *Linear Algebra and its Applications*, 119:1–25.

[Farmery and Whiteley, 2001] Farmery, A. D. and Whiteley, J. P. (2001). A mathematical model of electron transfer within the mitochondrial respiratory cytochromes. *Journal of Theoretical Biology*, 213:197–207.

[Fernandes et al., 2004] Fernandes, A., Gutierrez, C., and Rabanal, R. (2004). On local diffeomorphisms of $\mathbb{R}^n$ that are injective. *Qualitative Theory of Dynamical Systems*, 4(2):255–262.

[Feßler, 1995] Feßler, R. (1995). A proof of the two-dimensional Markus-Yamabe stability conjecture. *Annales Polonici Mathematici*, 62:45–75.

[Flanders, 1963] Flanders, H. (1963). *Differential Forms*, volume 11 of *Mathematics in Science and Engineering.* Academic Press.

[Gale and Nikaido, 1965] Gale, D. and Nikaido, H. (1965). The Jacobian matrix and univalence of mappings. *Mathematische Annalen*, 159:81–93.

[Garrett and Grisham, 1995] Garrett, R. H. and Grisham, C. M., editors (1995). *Biochemistry.* Saunders College Publishing.

[Gilbert, 1956] Gilbert, W. M. (1956). Completely monotonic functions on cones. *Pacific Journal of Mathematics*, 6(4):685–689.

[Glendinning, 1994] Glendinning, P. (1994). *Stability, instability and chaos.* Cambridge University Press.

[Glutsyuk, 1994] Glutsyuk, A. A. (1994). The complete solution of the Jacobian problem for vector fields on the plane. *Russian Mathematical Surveys*, 49(3):185–186.

[Godsil and Royle, 2001] Godsil, C. and Royle, G. (2001). *Algebraic graph theory*, volume 207 of *Graduate Texts in Mathematics.* Springer.

[Gradshteyn and Ryzhik, 2000] Gradshteyn, I. S. and Ryzhik, I. M. (2000). *Table of integrals, series and products.* Academic Press, 6th edition. Translated into English by Alan Jeffrey and Daniel Zwillinger.

[Graham, 1987] Graham, A. (1987). *Nonnegative matrices and applicable topics in linear algebra.* Mathematics and its Applications. Ellis Horwood.

[Guckenheimer and Holmes, 1983] Guckenheimer, J. and Holmes, P. (1983). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.* Springer-Verlag.

[Gutierrez, 1995] Gutierrez, C. (1995). A solution to the bidimensional global asymptotic stability conjecture. *Annales de l'Institute Henri Poincaré (C) Analyse Non Linéaire*, 12:627–671.

[Hasselblatt and Katok, 2003] Hasselblatt, B. and Katok, A. B. (2003). *A First Course in Dynamics*. Cambridge University Press.

[Hirsch, 1982] Hirsch, M. W. (1982). Systems of differential equations which are competitive or cooperative I: Limit sets. *SIAM Journal on Mathematical Analysis*, 13(2):167–179.

[Hirsch, 1985] Hirsch, M. W. (1985). Systems of differential equations which are competitive or cooperative II: Convergence almost everywhere. *SIAM Journal on Mathematical Analysis*, 16(3):423–439.

[Hirsch and Smale, 1974] Hirsch, M. W. and Smale, S. (1974). *Differential equations, dynamical systems, and linear algebra*, volume 60 of *Pure and Applied Mathematics*. Academic Press.

[Hirsch and Smith, 2004] Hirsch, M. W. and Smith, H. L. (2004). Generic quasiconvergence for strongly order preserving semiflows: a new approach. *Journal of Dynamics and Differential Equations*, 16(2):433–439.

[Hirsch and Smith, 2005] Hirsch, M. W. and Smith, H. L. (2005). Monotone dynamical systems. In Canada, A., Drabek, P., and Fonda, A., editors, *Handbook of Differential Equations: Ordinary Differential Equations*, volume 2, chapter 4. Elsevier.

[Hirsch and Smith, 2006] Hirsch, M. W. and Smith, H. L. (2006). Asymptotically stable equilibria for monotone semiflows. *Discrete and Continuous Dynamical Systems - Series A*, 14(3):385–398.

[Horn and Johnson, 1991] Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. Cambridge University Press.

[Ji-fa, 1994] Ji-fa, J. (1994). On the global stability of cooperative systems. *Bulletin of the London Mathematical Society*, 26:455–458.

[Kafri, 2002] Kafri, W. S. (2002). Robust D-stability. *Applied Mathematics Letters*, 15:7–10.

[Kellogg, 1972] Kellogg, R. B. (1972). On complex eigenvalues of M and P matrices. *Numerische Mathematik*, 19:170–175.

[Korzeniewski, 1996] Korzeniewski, B. (1996). Simulation of oxidative phosphorylation in hepatocytes. *Biophysical Chemistry*, 58:215–224.

[Korzeniewski and Zoladz, 2001] Korzeniewski, B. and Zoladz, J. A. (2001). A model of oxidative phosphorylation in mammalian skeletal muscle. *Biophysical Chemistry*, 92:17–34.

[Krein and Rutman, 1962] Krein, M. G. and Rutman, M. A. (1962). *Linear operators leaving invariant a cone in a Banach space*, volume 10 of *American Mathematical Society Translations Series 1*, pages 199–325. American Mathematical Society. Translation of article in Russian originally published in *Uspekhi Matematicheskikh Nauk*, 3:3–95 (1948).

[Kunze and Siegel, 1999] Kunze, H. and Siegel, D. (1999). Monotonicity with respect to closed convex cones I. *Dynamics of Continuous, Discrete and Impulsive Systems*, 5:433–449.

[Kunze and Siegel, 2001] Kunze, H. and Siegel, D. (2001). Monotonicity with respect to closed convex cones II. *Applicable Analysis*, 77(3–4):233–248.

[Kunze and Siegel, 2002a] Kunze, H. and Siegel, D. (2002a). A graph theoretic approach to strong monotonicity with respect to polyhedral cones. *Positivity*, 6:95–113.

[Kunze and Siegel, 2002b] Kunze, H. and Siegel, D. (2002b). Monotonicity properties of chemical reactions with a single initial bimolecular step. *Journal of Mathematical Chemistry*, 31(4):339–344.

[Lancaster and Tismenetsky, 1985] Lancaster, P. and Tismenetsky, M. (1985). *The Theory of Matrices.* Computer Science and Applied Mathematics. Academic Press, second edition.

[Li and Muldowney, 1995] Li, M. Y. and Muldowney, J. S. (1995). On R.A. Smith's autonomous convergence theorem. *Rocky Mountain Journal of Mathematics*, 25(1):365–379.

[Li and Muldowney, 1996] Li, M. Y. and Muldowney, J. S. (1996). A geometric approach to global-stability problems. *SIAM Journal on Mathematical Analysis*, 27(4):1070–1083.

[Li and Muldowney, 2000] Li, M. Y. and Muldowney, J. S. (2000). Dynamics of differential equations on invariant manifolds. *Journal of Differential Equations*, 168:295–320.

[Li and Wang, 1998] Li, M. Y. and Wang, L. (1998). A criterion for stability of matrices. *Journal of Mathematical Analysis and Applications*, 225:249–264.

[Li and Muldowney, 1993] Li, Y. and Muldowney, J. S. (1993). On Bendixson's criterion. *Journal of Differential Equations*, 106:27–39.

[Lohmiller and Slotine, 1998] Lohmiller, W. and Slotine, J.-J. E. (1998). On contraction analysis for nonlinear systems. *Automatica*, 34(6):683–696.

[Markus and Yamabe, 1960] Markus, L. and Yamabe, H. (1960). Global stability criteria for differential systems. *Osaka Journal of Mathematics*, 12:305–317.

[Maxima, 2008] Maxima (2008). A computer algebra system. Documentation and free download available at http://maxima.sourceforge.net/.

[McKenzie, 1960] McKenzie, L. (1960). Matrices with dominant diagonals and economic theory. In Arrow, K. J., Karlin, S., and Suppes, P., editors, *Mathematical Methods in the Social Sciences*, pages 47–62. Stanford University Press.

[Mierczynski, 1995] Mierczynski, J. (1995). Cooperative irreducible systems of ordinary differential equations with first integral. In *Proceedings of the Second Marrakesh International Conference of Differential Equations.*

[Mincheva and Siegel, 2003] Mincheva, M. and Siegel, D. (2003). Comparison principles for reaction–diffusion systems with respect to proper polyhedral cones. *Dynamics of Continuous, Discrete and Impulsive Systems Series A: Mathematical Analysis*, 10:477–490.

[Mincheva and Siegel, 2007] Mincheva, M. and Siegel, D. (2007). Nonnegativity and positiveness of solutions to mass action reaction–diffusion systems. *Journal of Mathematical Chemistry*, 42:1135–1145.

[Mitrinović, 1970] Mitrinović, D. S. (1970). *Analytic Inequalities.* Springer-Verlag.

[Muldowney, 1990] Muldowney, J. S. (1990). Compound matrices and ordinary differential equations. *Rocky Mountain Journal of Mathematics*, 20(4):857–872.

[Neumaier, 1994] Neumaier, A. (1994). Global, rigorous and realistic bounds for the solution of dissipative differential equations. *Computing*, 52(4):315–336.

[Nikaido, 1968] Nikaido, H. (1968). *Convex Structures and Economic Theory.* Academic Press.

[Pugh, 1967] Pugh, C. C. (1967). An improved closing lemma and a general density theorem. *American Journal of Mathematics*, 89(4):1010–1021.

[Radjavi, 1999] Radjavi, H. (1999). The Perron-Frobenius theorem revisited. *Positivity*, 3:317–331.

[Richeson and Wiseman, 2002] Richeson, D. and Wiseman, J. (2002). A fixed point theorem for bounded dynamical systems. *Illinois Journal of Mathematics*, 46(2):491–495.

[Richeson and Wiseman, 2004] Richeson, D. and Wiseman, J. (2004). Addendum to “A fixed point theorem for bounded dynamical systems”. *Illinois Journal of Mathematics*, 48(3):1079–1080.

[Rump, 1997] Rump, S. M. (1997). Theorems of Perron-Frobenius type for matrices without sign restrictions. *Linear Algebra and its Applications*, 266:1–42.

[Schneider and Tam, 2006] Schneider, H. and Tam, B.-S. (2006). Matrices leaving a cone invariant. In Hogben, L., editor, *Handbook for Linear Algebra*. Chapman & Hall.

[Schneider and Vidyasagar, 1970] Schneider, H. and Vidyasagar, M. (1970). Cross-positive matrices. *SIAM Journal on Numerical Analysis*, 7(4):508–519.

[Scilab, 2008] Scilab (2008). A platform for numerical computation. Documentation and free download available at http://www.scilab.org/.

[Smillie, 1984] Smillie, J. (1984). Competitive and cooperative tridiagonal systems of differential equations. *SIAM Journal on Mathematical Analysis*, 15(3):530–534.

[Smith, 1995] Smith, H. L. (1995). *Monotone Dynamical Systems*. American Mathematical Society.

[Smith, 1986] Smith, R. A. (1986). Some applications of Hausdorff dimension inequalities for ordinary differential equations. *Proceedings of the Royal Society of Edinburgh Section A*, 104A:235–259.

[Smyth and Xavier, 1996] Smyth, B. and Xavier, F. (1996). Injectivity of local diffeomorphisms from nearly spectral conditions. *Journal of Differential Equations*, 130(2):406–414.

[Söderlind, 2006] Söderlind, G. (2006). The logarithmic norm: history and modern theory. *BIT Numerical Mathematics*, 46(3):631–652.

[Spanier, 1981] Spanier, E. H. (1981). *Algebraic Topology*. Springer.

[Ström, 1975] Ström, T. (1975). On logarithmic norms. *SIAM Journal on Numerical Analysis*, 12(5):741–753.

[Verbitskii and Gorban, 1992] Verbitskii, V. I. and Gorban, A. N. (1992). Jointly dissipative operators and their applications. *Siberian Mathematical Journal*, 33(1):19–23. English translation of a Russian article, originally published in Sibirskii Matematicheskii Zhurnal.

[Walcher, 2001] Walcher, S. (2001). On cooperative systems with respect to arbitrary orderings. *Journal of Mathematical Analysis and Applications*, 263:543–554.

[Wiggins, 1990] Wiggins, S. (1990). *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, volume 2 of *Texts in Applied Mathematics*. Springer-Verlag.