SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

A Fairly Complete Qualitative Analysis of a Discrete SIR Epidemiological Model

av

Johan Hallberg Szabadváry

2019 - No K21

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET, 106 91 STOCKHOLM

A Fairly Complete Qualitative Analysis of a Discrete SIR Epidemiological Model

Johan Hallberg Szabadváry

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Yishao Zhou

2019

A Fairly Complete Qualitative Analysis of a Discrete SIR Epidemiological Model

Johan Hallberg Szabadváry

June 2019

Abstract

The main purpose of this paper is to study the local dynamics and bifurcations of a discrete-time SIR epidemiological model. The existence and stability of disease-free and endemic fixed points are investigated along with a fairly complete classification of the systems bifurcations. In the preliminaries we present two proofs of the classical Routh test in order to give conditions for stability in terms of the coefficients of the characteristic polynomial of the Jacobian matrix. We also show the existence of a 3-cycle, which implies the existence of cycles of arbitrary length by the celebrated Sharkovskii's theorem, which we prove using directed graphs.

Genericity of some bifurcations is examined both analytically and through numerical computations. Bifurcation diagrams along with numerical simulations are presented. The system turns out to have both rich and interesting dynamics.

A possibly more biologically realistic, generalized system is suggested in the conclusions together with some bifurcation diagrams of this new system.

Sammanfattning

Det huvudsakliga syftet med denna uppsats är att studera lokal dynamik och bifurkationer hos en diskret SIR epidemiologisk modell. Existensen och stabiliteten av den sjukdomsfria och den endemiska fixpunkten undersöks tillsammans med en rätt så komplett klassifikation av systemets bifurkationer. För att kunna formulera villkor för stabilitet i termer av koefficienter till det karakteristiska polynomet till Jacobimatrisen, presenteras två olika bevis för det klassiska Routh-testet. Vi påvisar också existensen av en 3-cykel. Det implicerar att det finns cykler av godtycklig längd enligt den hyllade Sharkovskii's sats, som vi bevisar medelst riktade grafer.

"Genericity" (ung. allmängiltighet) hos vissa bifurkationer undersöks såväl analytiskt som genom numeriska beräkningar. Bifurkationsdiagram, tillsammans med numeriska simuleringar presenteras. Systemet visar sig ha både rik och intressant dynamik.

Ett möjligen mer bilogiskt realistiskt system föreslås i slutsatsdelen tillsammans med några bifurkationsdiagram för detta nya system.

Acknowledgements

For being a source of inspiration, guidance and encouragement, I would like to thank my supervisor Yishao Zhou. Her help and patience, along with her faith in my ability has been invaluable in the process of writing this paper.

For never failing support in my studies, and for forcing me to get out of the house and take a walk when needed, I thank my wife, Kinga Szabadváry. I could not have done it without you.

Contents

1 Introduction

This paper aims to give a complete analysis of a discrete SIR-model with logistic growth of the susceptible population. In recent times, many interesting papers have appeared in the literature that discuss the stability, bifurcation and chaos phenomena in discrete-time systems, for instance [1, 2]. Discrete-time systems described by difference equations are particularly well suited for efficient numerical simulations, and in general display richer dynamics than their continuous counterparts.

In this paper we study stability and bifurcation of a particular discrete-time SIR-model from the paper [3], in which the authors partially analyse the system and give some numerical examples. However, their analysis is far from complete. Here we give a fairly complete analysis of the system dynamics.

Stability of fixed points is investigated using the standard technique of linearization together with the well-known Routh test for localizing polynomial zeros for which we give two separate proofs; one using Cauchy indices and Sturm chains and one more simple relying on complex analysis.

We derive the conditions for fold, flip, Neimark-sacker and some co-dimension 2 bifurcations using bifurcation theory, mainly from [4].

The existence of a period 3 cycle is shown which by the celebrated Sharkovskii's theorem implies the existence of cycles of any length. We find some cycles by computation and give a full proof of Sharkovskii's theorem using directed graphs.

The paper consists of two distinct parts and is organized as follows: In Section 2 we give some preliminaries including the definition of a dynamical system and bifurcations. We also briefly introduce the SIR model. Section 3 is devoted to the statement and proof of the Routh test in order to formulate sufficient conditions for stability of fixed points. Analysis of bifurcations along with classifications in co-dimension 1 and 2 are discussed in Section 4. In Section 5 we state and prove Sharkovskii's theorem.

The second part of the paper uses this theory to give a fairly complete analysis of a discrete time SIR model in Section 6. We analyse local stability of the three distinct fixed points and propose a candidate for the basic reproduction number R_0. Some analysis of the second iterate is also given. We classify all the bifurcations and investigate non-degeneracy of some of them. Bifurcation diagrams are presented along with some numerical simulations. We find a 3-cycle and some other n-cycles using similarity with the logistic map. Lastly Section 7 contains a short conclusion and a possible generalization of the system.

2 Preliminaries

2.1 Definition of a dynamical system

A dynamical system is the mathematical formalization of a deterministic process. The future behaviour of many systems in nature can sometimes be accurately predicted given knowledge of their present state and some law that govern their evolution in time. Such systems can be physical or chemical, but also biological or even social or economic. Provided that the governing law does not change over time, the future states of such a system is essentially completely determined by its initial state.

Thus, the notion of a dynamical system consists of a *state space*, the set of all possible states of the system, and a law that determines the *evolution* of the state in *time*. The following discussion and subsequent definition is closely modelled on Kuznetsovs book [4].

2.1.1 State space

Every possible state of a dynamical system can be thought of as a point x in some set X. This set is called the *state space* of the system. Typically, the state space has some natural structure which allows us to compare different states. In particular most state spaces of interest allow for the definition of a *distance* d, making X a metric space. Depending on the dimension of the state space, the dynamical system is called either finite- or infinite-dimensional.

2.1.2 Time

By the evolution of a dynamical system we mean a change in the state of the system with time $t \in \mathbb{T}$, where $\mathbb{T}$ is the time set. Essentially there are two distinct types of dynamical systems; *continuous-time* dynamical systems have $\mathbb{T} = \mathbb{R}$, and *discrete-time* dynamical systems where $\mathbb{T} = \mathbb{Z}$. Discrete-time systems appear naturally in ecology and economics when the state of a system at time t completely determines its state after, say a year at time $t + 1$.

2.1.3 Evolution operator

The main component of a dynamical system is the law that determines the state x_t of the system at time t given an *initial state* x_0 at time $t = 0$. This law can be specified in many different ways; for example, in the continuous-time case by means of differential equations. However, the most general way to specify the evolution is to assume that for a given time $t \in \mathbb{T}$ a map φ^t is defined in the state space X,

$$\varphi^t : X \to X,$$

which maps an initial state x_0 in X to some state x_t X at time t, so that

$$x_t = \varphi^t x_0.$$

The map φ^t is called the *evolution operator*. It might be known explicitly, but in most cases, it is defined indirectly, for instance by differential equations, and can only be approximated. Dynamical systems with evolution operator defined for both $t \geq 0$ and $t < 0$ are called *invertible*. In this case the initial state determine not only the future behaviour of the system, but also the past.

The evolution operator must have two properties that naturally reflects the deterministic character of dynamical systems, namely

$$\varphi^0 = id, \tag{DS.0}$$

where id is the identity map on X. Secondly

$$\varphi^{t+s} = \varphi^t \circ \varphi^s, \tag{DS.1}$$

for all x in X and t, s in $\mathbb{T}$ such that both sides are defined.

The first property implies that the system does not change its state spontaneously, while the second property states that starting at some state x and letting the system evolve in $t + s$ time units, yields the same result as starting at the same state x, letting the system first evolve over only s units of time, and then let it evolve over the next t units of time from the resulting state $\varphi^s x$. Essentially this means that the governing law does not change in time.

Note that a discrete-time dynamical system is fully specified by defining just one map, $f = \varphi^1$, since knowing this map we obtain

$$\varphi^2 = \varphi^1 \circ \varphi^1 = f \circ f = f^2$$

where f^2 is the *second iterate* if the map f. Similarly, one finds that $\varphi^k = f^k$ for all $k > 0$.

Many dynamical systems defined on $\mathbb{R}^n$ are such that φ^t is smooth as a function of (x, t). Such systems are called *smooth dynamical systems.*

We are now able to give a formal definition of a dynamical system:

Definition 1. *A dynamical system is a triple $\{\mathbb{T}, X, \varphi^t\}$, where $\mathbb{T}$ is a time set, X is a state space, and $\varphi^t : X \to X$ is a family of evolution operators parametrized by $t \in \mathbb{T}$ and satisfying the properties* (DS.0) *and* (DS.1)*.*

In the present work we shall only be concerned with discrete-time dynamical systems.

2.2 Orbits and invariant sets

Associated with a dynamical system $\{\mathbb{T}, X, \varphi^t\}$ are its *orbits* and the *phase portrait* composed of these orbits.

Definition 2. *An orbit starting at x_0 is an ordered subset of the state space X,*

$$Or(x_0) = \{x \in X : x = \varphi^t x_0, \text{ for all } t \in \mathbb{T} \text{ such that } \varphi^t x_0 \text{ is defined}\}.$$

Orbits are also called *trajectories.* If $y_0 = \varphi^{t_0} x_0$ for some t_0, then the sets $Or(x_0)$ and $Or(y_0)$ are equal. The simplest orbits are *equilibria* or *fixed points.*

Definition 3. *A point $x^* \in X$ is called an equilibrium or fixed point if $\varphi^t x^* = x^*$ for all $t \in \mathbb{T}$.*

Evidently a system placed at an equilibrium remains there forever. Thus, equilibria represent the simplest behaviour of the system. In discrete-time systems, an equilibrium is usually called a *fixed point.* Another relatively simple type of orbit is a *cycle.*

Definition 4. *A cycle is a periodic orbit. More precisely a nonequilibrium orbit L_0, such that each point x_0 in L_0 satisfies $\varphi^{t+T_0} x_0 = \varphi^t x_0$ with some $T_0 > 0$, for all $t \in \mathbb{T}$.*

The smallest T_0 with this property is called the period of the cycle L_0. If the system starts its evolution at some point x_0 on the cycle, it will return exactly to the same point after T_0 units of time.

Definition 5. *The phase portrait is a partitioning of the state space into its orbits.*

To geometrically represent the phase portrait in a figure is of course not possible. In practice only some particularly interesting orbits that are somehow representative of the system dynamics are shown.

Definition 6. *An invariant set of a dynamical system $\{\mathbb{T}, X, \varphi^t\}$ is a subset $S \subset X$ such that $x_0 \in S$ implies $\varphi^t x_0 \in S$ for all $t \in \mathbb{T}$.*

It should be clear that an invariant set S consists of orbits of the dynamical system, and that any individual orbit in itself is an invariant set.

2.2.1 Stability of invariant sets

An invariant set S_0 is called stable if it attracts nearby orbits. More formally, suppose we have a dynamical system $\{\mathbb{T}, X, \varphi^t\}$ with a complete metric space X. Let S_0 be a closed invariant set.

Definition 7. *An invariant set S_0 is called stable if*

(i) for any sufficiently small neighbourhood $U \supset S_0$, there exists a neighbourhood $V \supset S_0$ such that $\varphi^t x \in U$ for all $x \in V$ and all $t > 0$;

(ii) there exists a neighbourhood $U_0 \supset S_0$ such that $\varphi^t x \to S_0$ for all $x \in U_0$, as $t \to \infty$.

If S_0 is an equilibrium, this definition turns into the standard definition of stable equilibria. The first property is called *Lyapunov stability* and implies that orbits close to S_0 do not leave its neighbourhood. The second property is called *asymptotic stability.* It is possible for a system to be Lyapunov stable without being asymptotically stable. On the other hand, there are invariant sets that are asymptotically stable, but not stable in the Lyapunov sense, since some orbits starting near S_0 eventually approach S_0, but only after excursion outside some small but fixed neighbourhood of S_0.

If x^* is a fixed point of a finite-dimensional, smooth, discrete-time dynamical system, then sufficient conditions for its stability can be given in terms of the *Jacobian matrix* evaluated at x^*.

Definition 8. *Given a discrete-time dynamical system*

$$x \mapsto f(x), \ x \in \mathbb{R}^n,$$

where $f = (f_1, f_2, \ldots, f_n)$ is a C^1-map, its Jacobian matrix, denoted by $J(x)$ is the matrix of all partial derivatives of the map f, arranged as follows:

$$J(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}. \tag{2.1}$$

The Jacobian matrix can be defined more generally, but for our purposes this is enough. Recall that $f \in C^1$ if all partial derivatives exist and are continuous.

Theorem 1. *Consider a discrete-time dynamical system*

$$x \mapsto f(x), \ x \in \mathbb{R}^n,$$

where f is smooth. Suppose it has a fixed point x^*, so that $f(x^*) = x^*$, and denote by A the Jacobian matrix of $f(x)$ evaluated at x^*. Then the fixed point is locally asymptotically stable if all eigenvalues $\mu_1, \mu_2, \ldots, \mu_n$ of A satisfy $|\mu| < 1$.

The proof of this theorem is beyond the scope of this paper, but in [5] a proof in two dimensions is given.

There is another case where one can assure the stability of a fixed point, namely if the map f is a *contraction*:

Theorem 2. *Let X be a complete metric space with distance d. Assume that there is a map $f : X \to X$ that is continuous and satisfies for all x, y in X,*

$$d(f(x), f(y)) \le \lambda d(x, y)$$

for some $0 < \lambda < 1$. Then the discrete-time dynamical system $\{\mathbb{Z}_+, X, f^k\}$ has a stable fixed point x^* in X. Moreover, $f^k(x) \to x^*$ as $k \to \infty$, starting from any point $x \in X$.

The proof of this fundamental theorem can be found in most textbooks on mathematical analysis, for example [6].

Recall that the eigenvalues of a $n \times n$-matrix A are the roots of the *characteristic equation*

$$\det(A - \mu I_n) = 0,$$

where I_n is the $n \times n$ identity matrix. It is a well-known fact, and central to the analysis to come that this is a polynomial equation.

From Theorem 1 it is clear that a large part in determining the stability of a fixed point, is to determine whether the eigenvalues of the Jacobian matrix lie inside the unit circle. Since the eigenvalues are roots of the characteristic polynomial equation, this turns into the problem of locating zeros of a polynomial. We would like to find conditions on the coefficients of the characteristic polynomial that guarantee that all zeros lie inside the unit circle. Fortunately, this problem can be solved by using the rather convenient *Routh test* together with a certain *Möbius transformation*.

2.3 Topological equivalence and bifurcations

To make comparison between two different dynamical systems, we need some notion of when two dynamical systems are "qualitatively similar". Intuitively such a definition must meet some natural criteria. For instance, two equivalent systems should have the same number of equilibria, and cycles of the same stability type.

Definition 9. *A dynamical system $\{\mathbb{T}, \mathbb{R}^n, \varphi^t\}$ is called topologically equivalent to a dynamical system $\{\mathbb{T}, \mathbb{R}^n, \psi^t\}$ if there is a homeomorphism $h : \mathbb{R}^n \to \mathbb{R}^n$ mapping orbits of the first system into orbits of the second system, preserving the direction of time.*

A *homeomorphism* is an invertible map such that both the map and its inverse are continuous. The definition of topological equivalence can be extended to more general state spaces, but for our purposes it is enough to consider $\mathbb{R}^n$. It should be clear that the relation "is topologically equivalent to" is an equivalence relation. Clearly a system is topologically equivalent to itself, just take $h = id$, were id is the identity map on $\mathbb{R}^n$. Also, if system a is topologically equivalent to system b by the existence of some homeomorphism h, then system b is also topologically equivalent to system a, by the homeomorphism h^{-1}. Lastly if system a is topologically equivalent to system b and system b is topologically equivalent to system c, then there are homeomorphisms h_a and h_b that relate them. Then the map $h_b \circ h_a$ is a homeomorphism, since it and its inverse, $h_a^{-1} \circ h_b^{-1}$ are compositions of continuous maps, and it maps orbits of system a into orbits of system c. Hence system a is topologically equivalent to system c.

In the case of discrete dynamical systems, an explicit relation between the corresponding maps of the equivalent systems can be obtained. Let

$$x \mapsto f(x), \ x \in \mathbb{R}^n, \tag{2.2a}$$

and

$$y \mapsto g(y), \ y \in \mathbb{R}^n, \tag{2.2b}$$

be two topologically equivalent, discrete-time invertible dynamical systems, that is $f = \varphi^1$ and $g = \psi^1$ are smooth invertible maps. Consider an orbit of system (2.2a) starting at some point x:

$$\ldots, f^{-1}(x), x, f(x), f^2(x), \ldots$$

and an orbit of system (2.2b) staring at some point y:

$$\ldots, g^{-1}(y), y, g(y), f^2(y), \ldots$$

Topological equivalence implies that if x and y are related by the homeomorphism h so that $y = h(x)$, then the first orbit is mapped onto the second one by h. Symbolically we present this as

$$\begin{array}{ccc} x & \xrightarrow{f} & f(x) \\ \downarrow{\scriptstyle h} & & \downarrow{\scriptstyle h} \\ y & \xrightarrow{g} & g(y). \end{array}$$

Therefore $g(y) = h(f(x)$ or $g(h(x)) = h(f(x))$ for all points x in $\mathbb{R}^n$, which can be written as $f(x) = h^{-1}(g(h(x)))$, or more compactly

$$f = h^{-1} \circ g \circ h. \tag{2.4}$$

Definition 10. *Two maps f and g satisfying* (2.4) *for some homeomorphism h are called conjugate.*

We are often interested in the system dynamics, not in the whole state space $\mathbb{R}^n$, but *locally* in some region $U \subset \mathbb{R}^n$. Usually such a region is a neighbourhood of a fixed point or a cycle. The above definition can easily be localized by the introduction of appropriate regions. For example, in the topological classification of the phase portraits near a fixed point, the following definition is useful.

Definition 11. *A dynamical system $\{\mathbb{T}, \mathbb{R}^n, \varphi^t\}$ is called topologically equivalent near an equilibrium x^* to a dynamical system $\{\mathbb{T}, \mathbb{R}^n, \psi^t\}$ near an equilibrium y^* if there is a homeomorphism $h : \mathbb{R}^n \to \mathbb{R}^n$ that is*

(i) defined in a small neighbourhood $U \subset \mathbb{R}^n$ of x^*;

(ii) satisfies $y^* = h(x^*)$;

(iii) maps orbits of the first system in U onto orbits of the second system in $V = h(U) \subset \mathbb{R}^n$, preserving the direction of time.

Here $V = h(U)$ is the *image* of U under h, that is

$$V = \{y \in \mathbb{R}^n : y = h(x), x \in U\}.$$

If U is an open neighbourhood of x^* then V is an open neighbourhood of y^*. This is true since both h and h^{-1} are continuous functions on $\mathbb{R}^n$. One should also note that x^* and y^* as well as U and V may coincide.

2.3.1 Hyperbolic fixed points in discrete-time systems

Let the map f and its inverse be smooth and consider the discrete-time dynamical system

$$x \mapsto f(x),\ x \in \mathbb{R}^n \tag{2.5}$$

with a fixed point $x^* = 0$. Let A denote the Jacobian matrix evaluated at x^*. The eigenvalues $\mu_1, \ldots, \mu_n$ of A are sometimes called *multipliers* of the fixed point. Let n_-, n_0, and n_+ be the numbers of multipliers lying inside, on and outside the unit circle respectively.

Definition 12. *A fixed point is called hyperbolic if $n_0 = 0$, so there are no multipliers on the unit circle. A hyperbolic fixed point is called a hyperbolic saddle if $n_- n_+ \neq 0$.*

Theorem 3. *The phase portraits of* (2.5) *near two hyperbolic fixed points* x^* *and* y^* *are locally topologically equivalent if and only if these fixed points have the same number* n_- *and* n_+ *of multipliers with* $|\mu| < 1$ *and* $|\mu| > 1$ *respectively, and the sign of the products of all multipliers with* $|\mu| < 1$ *and with* $|\mu| > 1$ *are the same for both* x^* *and* y^**.*

Often the fixed points x^* and y^* are also called topologically equivalent. The proof of the theorem is not given here but is based on the fact that near a hyperbolic fixed point, the system (2.5) is locally topologically equivalent to its *linearization* $x \mapsto Ax$. This is the discrete version of the Grobman-Hartman theorem. A proof of the continuous version can be found in [7]. This then has to be applied both near x^* and y^*. Next, one has to prove that two *linear* systems with the same types of multipliers are locally topologically equivalent.

2.4 The SIR model

The SIR model, or rather models, developed by Ronald Ross[1], William Hamer and others in the early twentieth century, are *compartmental models* used primarily in the mathematical modeling of infectious disease. In its original formulation the model consists of a system of three coupled nonlinear differential equations which does not possess an explicit formula solution. In [8, 9] a brief history and some applications to public health along with some testing against data are given.

In short, the model in its original formulation is this: A population is partitioned into three groups or compartments; susceptible individuals, infected individuals, and removed individuals. The susceptible and infected groups are self-explanatory, but it should be noted that the removed group includes in principle anyone that is not susceptible or infected whether immune, dead or launched into space. The sizes of these compartments at time t are denoted by $S(t), I(t)$, and $R(t)$ respectively.

The original model makes several severe assumptions, including a large and closed population in which no natural deaths or births occur. The infection is also assumed to have no incubation period, and upon recovery from the infection, an individual is assumed to gain lifetime immunity. Age, sex, social status or ethnicity is not assumed to have any effect on the probability of being infected. Also, it is assumed that there is mass action mixing of individuals, which ensures that the rate of encounter between susceptible and infected individuals is proportional to the product $S(t)I(t)$. This last assumption requires that the members of both the infected and susceptible parts of the population are homogeneously distributed in space rather than mainly mixing in smaller subgroups.

There are hundreds of papers where this basic model is extended in many directions to suit particular applications. For instance, one may add natural death rates, death due to illness, natural population growth, effects of vaccination just to name a few. Some infections such as the common cold and influenza (unfortunately) do not confer any long-term immunity, nor do they readily kill their host. When recovered, an infected individual is therefore susceptible again. This observation motivates the SIS model in which the population is partitioned into just two groups of susceptible and infected.

In the present work we shall attend to a particular discrete-time version of the SIR model in which the growth of the susceptible population, some inhibitory effects and death rates have been accounted for.

A central number in epidemiology is the so-called basic reproduction number, denoted R_0. This is defined as the number of cases one case generates on average over the course of its infectious period in a totally susceptible population. The use of this quantity is not without complications,

[1]Sir Ronald Ross received the second Noble Prize in Medicine and Physiology for his discovery of the transmission of malaria by the mosquito.

but as a rule of thumb, one says that if $R_0 < 1$ the infection dies out in the long run, and if $R_0 > 1$ the infection will spread in the population and will require intervention to eradicate.

3 Routh-Hurwitz stability criterion and sufficient conditions for stability of nonlinear systems

The Routh test is a convenient way to determine the number k of distinct zeros of a real polynomial $p(x)$ in the open right half-plane $\{z \in \mathbb{C} : Re(z) > 0\}$. This is not immediately useful since our problem is to determine the number of zeros in the open unit disc. However, we shall see that the same theory can be applied to our problem by transforming the unit disc to the left half plane via a certain *Möbius transformation.* This section aims to prove the Routh test, which has many proofs. First, we follow Gantmacher [10] to give the standard proof using *Sturm chains* to compute *Cauchy indices.* Next a somewhat simpler proof follows.

3.1 Cauchy indices

Definition 13. *Given a real rational function $R(x)$, the Cauchy index between the limits a and b, $a < b$ denoted $I_a^b R(x)$ is the difference between the number of jumps of $R(x)$ from $-\infty$ to $+\infty$, and the number of jumps from $+\infty$ to $-\infty$ as x goes from a to b. Here a and b are real numbers or $\pm\infty$.*

Using the definition, if $R(x)$ is a real rational function,

$$R(x) = \sum_{i=1}^{p} \frac{A_i}{x - \alpha_i} + R_1(x),$$

where A_i, α_i are real numbers, and $R_1(x)$ is a rational function without real poles (zeros of the denominator), we have

$$I_a^b R(x) = \sum_{a<\alpha_i<b} sign(A_i), \tag{3.1}$$

where $a < b$. This holds true if $a = -\infty$ and $b = +\infty$.

In particular, if $p(x) = \alpha_0(x - \alpha_1)^{n_1} \dots (x - \alpha_m)^{n_m}$ is a real polynomial where $\alpha_i \neq \alpha_k$ for $i \neq k$, $i, k = 1, 2 \dots, m$, and if only the first p of its zeros are real then

$$\frac{p'(x)}{p(x)} = \sum_{i=1}^{m} \frac{n_i}{x - \alpha_i} = \sum_{i=1}^{p} \frac{n_i}{x - \alpha_i} + R_1(x),$$

where $R_1(x)$ is a real rational function with no real poles.

To realize this, note that

$$\begin{aligned} p'(x) =& \alpha_0(n_1(x - \alpha_1)^{n_1 - 1}(x - \alpha_2)^{n_2} \dots (x - \alpha_m)^{n_m} + \\ & n_2(x - \alpha_1)^{n_1}(x - \alpha_2)^{n_2 - 1} \dots (x - \alpha_m)^{n_m} + \dots + \\ & n_m(x - \alpha_1)^{n_1}(x - \alpha_2)^{n_2} \dots (x - \alpha_m)^{n_m - 1}). \end{aligned}$$

Hence by (3.1) we note that

$$I_a^b \frac{p'(x)}{p(x)}$$

is equal to the number of distinct real zeros of $p(x)$ in the interval (a, b).

Using partial fraction decomposition, any real rational function $R(x)$ can be written in the form

$$R(x) = \sum_{i=1}^{p} \left(\frac{A_1^{(i)}}{x - \alpha_i} + \cdots + \frac{A_{n_i}^{(i)}}{(x - \alpha_i)^{n_i}} \right) + R_1(x),$$

where again $R_1(x)$ has no real poles, and all the α and A are real numbers with $A^{(i)} \neq 0$ for $i = 1, 2, \ldots, p$.

Then in general,

$$I_a^b R(x) = \sum_{a<\alpha_i<b,\ n_i \text{odd}} sign(A_{n_i}^{(i)})$$

for $a < b$, and in particular

$$I_{-\infty}^{+\infty} R(x) = \sum_{n_i \text{odd}} sign(A_{n_i}^{(i)}),$$

since for even n_i the leading term $\frac{A_{n_i}^{(i)}}{(x-\alpha_i)^{n_i}}$ is always positive, when it is well defined, so there is no jump.

3.2 Sturm's theorem

In order to compute the Cauchy index $I_a^b R(x)$, we shall make use of a certain sequence of polynomials, called a *Sturm chain.*

Definition 14. *A sequence of real polynomials*

$$p_1(x), p_2(x), \ldots, p_m(x) \tag{3.2}$$

is a Sturm chain in the interval (a, b) if it satisfies the following properties on (a, b), where $a < b$ and we may let $a = -\infty$ and $b = +\infty$:

(i) For any x such that $a < x < b$, if any $p_k(x)$ vanishes, the two polynomials next to it, $p_{k-1}(x)$ and $p_{k+1}(x)$ are nonzero, and of opposite signs. That is, if $p_k(x) = 0$ then

$$p_{k-1}(x)p_{k+1}(x) < 0.$$

(ii) The last polynomial, $p_m(x)$ has no zeros in the interval (a, b).

For a fixed value x we denote by $V(x)$ the number of sign changes in (3.2). The value of $V(x)$ as x passes from a to b can only change when one of the polynomials in (3.2) passes through a zero. By the first condition on a Sturm chain, when $p_k(x)$, $k = 2, \ldots, m-1$ passes through a zero, $V(x)$ does not change. More explicitly, if $p_k(x)$ passes through a zero at $x = \xi$, and we assume without loss of generality that before the zero we had

$$(\text{sign}(p_{k-1}(\xi - \epsilon)), \text{sign}(p_k(\xi - \epsilon)), \text{sign}(p_{k+1}(\xi - \epsilon))) = (+, +, -),$$

then after the zero the situation is

$$(\text{sign}(p_{k-1}(\xi + \epsilon)), \text{sign}(p_k(\xi + \epsilon)), \text{sign}(p_{k+1}(\xi + \epsilon))) = (+, -, -)$$

for some sufficiently small $\epsilon > 0$. Hence the number of sign changes does not change, that is $V(x)$ does not change.

However, when $p_1(x)$ passes through a zero, the value of $V(x)$ either increases or decreases by 1 depending on whether the ratio $p_2(x)/p_1(x)$ goes from $+\infty$ to $-\infty$ or vice versa. This result is known as:

Theorem 4 (Sturm)**.** *If $p_1(x), p_2(x), \ldots, p_m(x)$ is a Sturm chain in the interval (a, b), and $V(x)$ is the number of sign variations in the chain, then*

$$I_a^b \frac{p_2(x)}{p_1(x)} = V(a) - V(b). \tag{3.3}$$

3.2.1 Generalized Sturm chains

If we multiply all the polynomials in the Sturm chain (3.2) by an arbitrary polynomial $d(x)$, the result is called a *generalized Sturm chain*. Such a multiplication clearly does not change the left hand side of (3.3), since the quotient does not change, and neither does it change the right hand side, since when x passes through a zero of $d(x)$, all signs are flipped which does not alter the number of sign changes. However, one should note that consecutive polynomials in (3.2) may now vanish simultaneously for some values of x, but then every polynomial vanishes for such x. For this reason, Sturm's theorem is still valid for generalized Sturm chains.

Given two real polynomials $p(x), q(x)$ such that p has degree greater than or equal to the degree of q, we can always construct a generalized Sturm chain by letting $p_1(x) := p(x),\ p_2(x) := q(x)$. Next, we use the Euclidean algorithm to find a greatest common divisor of p and q. Denote by $-p_3(x)$ the remainder on dividing $p_1(x)$ by $p_2(x)$, and by $-p_4(x)$ the remainder on dividing $p_2(x)$ by $p_3(x)$ et cetera. This gives us a chain of identities

$$\begin{aligned}
&p_1(x) = q_1(x)p_2(x) - p_3(x),\\
&p_2(x) = q_2(x)p_3(x) - p_4(x),\\
&\ldots\\
&p_{k-1}(x) = q_{k-1}(x)p_k(x) - p_{k+1}(x)\\
&\ldots\\
&p_{m-1}(x) = q_{m-1}p_m(x),
\end{aligned} \tag{3.4}$$

where the last nonzero remainder $p_m(x)$ is a greatest common divisor of $p(x)$ and $q(x)$, and also of all the polynomials in the sequence (3.2) thus obtained.

If $p_m(x)$ has no zeros in the interval (a, b), the sequence satisfies both conditions in Definition 14 and is therefore a Sturm chain. Otherwise if $p_m(x)$ has zeros in (a, b), then (3.2) is a generalized Sturm chain for if divided by $p_m(x)$ it becomes a Sturm chain.

3.2.2 Computing Cauchy indices by Sturm's theorem

From the discussion above it follows that the Cauchy index of any real rational function $R(x)$ can be computed by Sturm's theorem. It suffices to write $R(x) = Q(x) + \frac{q(x)}{p(x)}$ where Q, p, g are polynomials with the degree of p greater than or equal to that of q. If we construct the generalized Sturm chain for $p(x), q(x)$, we find that

$$I_a^b R(x) = I_a^b \frac{q(x)}{p(x)} = V(a) - V(b).$$

As noted, before, the number of distinct real zeros of a real polynomial $p(x)$ in the interval (a, b) is $I_a^b \frac{p'(x)}{p(x)}$. Hence Sturm's theorem also gives us a way to determine the total number of distinct zeros of $p(x)$, namely

$$I_{-\infty}^{+\infty} \frac{p'(x)}{p(x)} = V(-\infty) - V(+\infty).$$

3.3 Routh's algorithm

Routh's algorithm lets us determine k, the number of distinct zeros of a real polynomial $p(x)$ in the open right half plane. First, we consider the case where $p(x)$ has no zeros on the imaginary axis. To do this, we consider a semicircle of radius R in the right half-plane, and denote by D the domain bounded by the semicircle and the imaginary axis, so that $D = \{z \in \mathbb{C} : |z| < R, Re(z) > 0\}$. If R is taken large enough, all zeros of $p(x)$ in the right half plane lie inside D.

Since polynomials are *analytic functions* in the whole complex plane, the *argument principle* holds, which in the case of an analytic function f can be summarized as

$$\frac{1}{2\pi}\Delta_C \mathrm{arg}\ f(z) = N_0(f)$$

where C is a simple closed positively oriented contour, $\Delta_C \mathrm{arg}\ f(z)$ is the net increase in argument as we go around C, and $N_0(f)$ is the number of zeros of f inside C counted with multiplicity. This well-known theorem is proved in most standard texts on complex analysis, for instance [11].

Then by the argument principle we have that arg $p(z)$ increases by $2\pi k$ on traversing the contour of D in the positive direction. For if

$$p(z) = a_0 \prod_{i=1}^{n} (z - z_i),$$

then

$$\Delta \mathrm{arg}\ f(z) = \sum_{i=1}^{n} \Delta arg\ (z - z_i),$$

and if z_i lies inside the domain, then $\Delta\mathrm{arg}\ (z - z_i) = 2\pi$, otherwise if z_i lies outside the domain we have that $\Delta\mathrm{arg}\ (z - z_i) = 0$. Here Δ denotes the change in the argument. On the other hand, as R tends to infinity, the increase of the argument of $p(z)$ along the semicircle of radius R is determined by the increase of the dominating term, $a_n z^n$, so it is $n\pi$. Piecing this together we find that the increase of arg $p(z)$ along the imaginary axis is

$$\Delta_{-\infty}^{+\infty} \mathrm{arg}\ p(i\omega) = (n - 2k)\pi \tag{3.5}$$

because going around the boundary of D as $R \to \infty$, the increase is

$$n\pi + \Delta_{\infty}^{-\infty} \mathrm{arg}\ p(i\omega) = 2\pi k.$$

Changing direction on the imaginary axis, we just flip the sign of the infinities, and the result follows.

To study $p(z)$ on the imaginary axis, we separate it into its real and imaginary part, that is

$$p(i\omega) = U(\omega) + iV(\omega), \tag{3.6}$$

where if we introduce the rather odd-looking notation

$$p(z) = a_0 z^n + b_0 z^{n-1} + a_1 z^{n-2} + b_0 z^{n-3} + \dots$$

we have for even n

$$\begin{aligned} U(\omega) &= (-1)^{\frac{n}{2}} (a_0 \omega^n - a_1 \omega^{n-2} + a_2 \omega^{n-4} - \dots), \\ V(\omega) &= (-1)^{\frac{n}{2}-1} (b_0 \omega^{n-1} - b_1 \omega^{n-3} + b_2 \omega^{n-5} - \dots) \end{aligned} \tag{3.7a}$$

and for odd n

$$U(\omega) = (-1)^{\frac{n-1}{2}}(b_0\omega^{n-1} - b_1\omega^{n-3} + b_2\omega^{n-5} - \dots),$$
$$V(\omega) = (-1)^{\frac{n-1}{2}}(a_0\omega^n - a_1\omega^{n-2} + a_2\omega^{n-4} - \dots). \tag{3.7b}$$

We note that

$$\arg p(i\omega) = \arctan\frac{V(\omega)}{U(\omega)} = \operatorname{arccot}\frac{U(\omega)}{V(\omega)}. \tag{3.8}$$

For even n we have $\lim_{\omega\to\pm\infty}\frac{V(\omega)}{U(\omega)} = 0$ by (3.7a). Since $\arctan\frac{V(\omega)}{U(\omega)}$ jumps from $\pi/2$ to $-\pi/2$ when $\frac{V(\omega)}{U(\omega)}$ jumps from $+\infty$ to $-\infty$, and vice versa, and this happens twice for every time $p(i\omega)$ winds around the origin, we conclude that

$$\frac{1}{\pi}\Delta_{-\infty}^{+\infty}\arg p(i\omega) = -I_{-\infty}^{+\infty}\frac{V(\omega)}{U(\omega)}$$

for even n.

On the other hand, (3.7b) tells us that $\lim_{\omega\to\pm\infty}\frac{U(\omega)}{V(\omega)} = 0$ for odd n. This means that $\operatorname{arccot}\frac{U(\omega)}{V(\omega)}$ passes through zero from the negative direction when $\frac{U(\omega)}{V(\omega)}$ jumps from $-\infty$ to ∞ and vice versa. Since this also happens twice for every time $p(i\omega)$ winds around the origin, we get

$$\frac{1}{\pi}\Delta_{-\infty}^{+\infty}\arg p(i\omega) = I_{-\infty}^{+\infty}\frac{U(\omega)}{V(\omega)}$$

for odd n. For clarity we summarize this as

$$\frac{1}{\pi}\Delta_{-\infty}^{+\infty}\arg p(i\omega) = \begin{cases} I_{-\infty}^{+\infty}\frac{U(\omega)}{V(\omega)} & \text{for odd } n \\ -I_{-\infty}^{+\infty}\frac{V(\omega)}{U(\omega)} & \text{for even } n. \end{cases} \tag{3.9}$$

From, (3.5), (3.7a), (3.7b) and (3.9) we get the nice result that for every n, even or odd

$$I_{-\infty}^{+\infty}\frac{b_0\omega^{n-1} - b_1\omega^{n-3} + b_2\omega^{n-5} - \dots}{a_0\omega^n - a_1\omega^{n-2} + a_2\omega^{n-4} - \dots} = n - 2k. \tag{3.10}$$

However, we should recall that this formula was derived under the assumption that $p(z)$ had no zeros on the imaginary axis.

3.3.1 The Routh table

In order to compute the index (3.10) we use Sturm's theorem. To this end we set

$$p_1(\omega) = a_0\omega^n - a_1\omega^{n-2} + a_2\omega^{n-4} - \dots$$
$$p_2(\omega) = b_0\omega^{n-1} - b_1\omega^{n-3} + b_2\omega^{n-5} - \dots \tag{3.11}$$

and construct a generalized Sturm chain

$$p_1(\omega), p_2(\omega), \dots, p_m(\omega) \tag{3.12}$$

by the Euclidean algorithm as described in (3.4).

We consider first the case when $m = n+1$, the *regular* case. Then the degree of each polynomial in the chain is one less than the previous one, and $p_m(\omega)$ has degree zero. Note that in the regular case, (3.12) is the ordinary, not the generalized Sturm chain. Then by (3.4) we have

$$p_3(\omega) = \frac{a_0}{b_0}\omega p_2(\omega) - p_1(\omega) = c_0\omega^{n-2} - c_1\omega^{n-4} + c_2\omega^{n-6} - \dots,$$

where

$$c_0 = a_1 - \frac{a_0}{b_0} b_1 = \frac{b_0 a_1 - a_0 b_1}{b_0}, \; c_1 = a_2 - \frac{a_0}{b_0} b_2 = \frac{b_0 a_2 - a_0 b_2}{b_0}, \ldots \tag{3.13}$$

Similarly

$$p_4(\omega) = \frac{a_0}{b_0} \omega p_2(\omega) - p_1(\omega) = d_0 \omega^{n-3} - d_1 \omega^{n-5} + d_2 \omega^{n-7} - \ldots,$$

where

$$d_0 = b_1 - \frac{b_0}{c_0} c_1 = \frac{c_0 b_1 - b_0 c_1}{b_0}, \; d_1 = b_2 - \frac{b_0}{c_0} c_2 = \frac{c_0 b_2 - b_0 c_2}{c_0}, \ldots \tag{3.13$'$}$$

The remaining polynomials $p_5(\omega), \ldots, p_{n+1}(\omega)$ are determined similarly.

Note that each polynomial

$$p_1(\omega), p_2(\omega), \ldots, p_{n+1}$$

is either an even or an odd function since all powers of ω are either even or odd, so clearly $p_k(\omega) = p_k(-\omega)$ if k is even, and $p_k(-\omega) = -p_k(\omega)$ if k is odd. Furthermore, two adjacent polynomials have the opposite parity, that is if p_k is an even function, then p_{k-1} and p_{k+1} are odd functions and vice versa.

From the coefficients of the polynomials in (3.12) we form the *Routh table*

$$\begin{array}{cccc} a_0, & a_1, & a_2, & \ldots, \\ b_0, & b_1, & b_2, & \ldots, \\ c_0, & c_1, & c_2, & \ldots, \\ d_0, & d_1, & d_2, & \ldots, \\ \vdots & \vdots & \vdots & \ddots \end{array} \tag{3.14}$$

By (3.13) and (3.13$'$) we conclude that every row can be determined from the preceding two by the following procedure: Multiply the lower row by the quotient of the first entry in the upper row and the first entry in the lower row. Then subtract this from the upper row. This eliminates the first entry. Now the next row is obtained by shifting the result one step to the left.

Given that the first entry in row k is the leading coefficient of $p_k(\omega)$, it is clear that in the regular case, this procedure never yields a zero in the sequence $a_0, b_0, c_0, d_0, \ldots$

In the regular case, the polynomials $p_1(\omega)$ and $p_2(\omega)$ have greatest common divisor $p_{n+1} = C \neq 0$ where C is a real constant. Then, by the factor theorem, these polynomials, and hence (by (3.7a) and (3.7b)) $U(\omega)$ and $V(\omega)$ cannot both vanish at the same time. This in turn means that

$$p(i\omega) = U(\omega) + iV(\omega) \neq 0$$

for real ω, so p has no zeros on the imaginary axis and hence the formula (3.10) holds in the regular case.

Applying Sturm's theorem in the interval $(-\infty, +\infty)$ to (3.10) we get

$$V(-\infty) - V(+\infty) = n - 2k. \tag{3.15}$$

Now, the sign of $p_k(\omega)$ at $\omega = +\infty$ is defined to be the sign of the leading coefficient. Likewise, the sign at $\omega = -\infty$ is equal to the sign of the leading coefficient if p_k has even degree and the opposite if p_k has odd degree. Hence

$$V(+\infty) = V(a_0, b_0, c_0, d_0, \ldots)$$

where the right-hand side should be interpreted as the number of sign changes in the sequence $a_0, b_0, c_0, d_0, \ldots$, and

$$V(-\infty) = V(a_0, -b_0, c_0, -d_0, \ldots).$$

Hence

$$V(-\infty) + V(+\infty) = n \tag{3.16}$$

because whenever there is a sign change in the sequence $a_0, b_0, c_0, d_0, \ldots$, the corresponding position in the sequence $a_0, -b_0, c_0, -d_0, \ldots$ does not have a sign change and vice versa. Since both sequences has length $n+1$, the total number of sign changes is n.

Then by (3.15) and (3.16) we have

$$k = V(+\infty) = V(a_0, b_0, c_0, d_0, \ldots), \tag{3.17}$$

and we have proved

Theorem 5 (Routh). *The number of distinct zeros of the real polynomial $p(z)$ in the open right half plane, $Re(z) > 0$ is equal to the number of sign variations in the first column of the Routh table.*

3.3.2 Case of stability

Usually we are interested in the special case when all zeros of $p(z)$ lie in the open left half plane, i.e. have negative real parts. In this case, if we form the generalized Strurm chain (3.12) for the polynomials (3.11), since $k = 0$, the formula (3.15) reduces to

$$V(-\infty) - V(+\infty) = n.$$

But since $0 \leq V(-\infty), V(+\infty) \leq m - 1 \leq n$, this is possible only when $m = n + 1$, i.e. the regular case, and $V(+\infty) = 0, V(-\infty) = n$. Then (3.17) implies

Routh's criterion. *All the zeros of the real polynomial $p(z)$ have negative real parts if and only if all the elements in the first column of the Routh table are nonzero and of like sign.*

We have proved Roth's theorem in the regular case, and for our purposes this is enough since we shall only require Routh's criterion. However, for completeness the rest of the proof is given in Appendix A.

Before moving on, we shall give a second, simpler proof due to Matsumoto that does not rely on Cauchy indices and Sturm chains. This proof is in some sense simpler but the price one pays is insight into why the test works.

3.4 Simple proof of the Routh stability criterion

In [12], Matsumoto aims to give a simple proof of the Routh stability criterion using order reduction of polynomials together with the argument principle.

Let $p_n(z)$ be a real polynomial of complex variable z and of order n:

$$p_n(z) = a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1} s + a_n.$$

Without loss of generality we shall assume that $a_0 > 0$.

Definition 15. *The polynomial $p_n(z)$ is an n:th order Hurwitz polynomial if all zeros of $p_n(z)$ lie in the open left half plane.*

We define the order reduction formula which will be used to generate the rows of the Routh table. Let $p_k(z)$ be a real polynomial of degree k of complex variable z:

$$p_k(z) = \alpha_0 s^k + \alpha_1 s^{k-1} + \cdots + \alpha_{k-1} s + \alpha_k. \tag{3.18}$$

Then the order reduction formula is

$$p_{k-1}(z) = p_k(z) - \mu_k z\big(p_k(z) - (-1)^k p_k(-z)\big) \tag{3.19a}$$

$$\mu_k = \frac{\alpha_0}{2\alpha_1}. \tag{3.19b}$$

If $p_k(z)$ in (3.18) is a Hurwitz polynomial, then every coefficient of $p_k(z)$ is positive since by the factor theorem we can write

$$p_k(z) = (z+\zeta_1)\dots(z+\zeta_s)(z+\xi_1+\sigma_1 i)(z+\xi_1-\sigma_1 i)\dots(z+\xi_t+\sigma_t i)(z+\xi_t-\sigma_t i)$$

where ζ_j are the s real zeros of $p_k(z)$, and $\xi_j \pm \sigma_j i$ are the $2t$ nonreal ones. Since

$$(z+\xi_j+\sigma_j i)(z+\xi_j-\sigma_j i) = z^2 + \xi_j\sigma_j z + \xi_j^2 + \sigma_j^2,$$

it is clear that all coefficients are positive. We will assume that μ_k in (3.19b) is finite and $\mu_k \neq 0$. Substituting (3.18) into (3.19) we obtain the reduced order polynomial as

$$p_{k-1}(z) = \beta_0 z^{k-1} + \beta_1 z^{k-2} + \dots + \beta_{k-2} z + \beta_k \tag{3.20a}$$

where

$$\beta_{2i} = \alpha_{2i+1},\ i = 0,1,2,\dots \tag{3.20b}$$

$$\beta_{2i+1} = -\frac{1}{\alpha_1}\begin{vmatrix}\alpha_0 & \alpha_{2i+2}\\ \alpha_1 & \alpha_{2i+3}\end{vmatrix}, i = 0,1,2\dots \tag{3.20c}$$

We should note that (3.20b) and (3.20c) are just the Routh algorithm previously discussed. The order reduction formula (3.19) is a polynomial representation of the Routh algorithm. From (3.20b) and (3.20c) we get that $p_k(z)$ and $p_{k-1}(z)$ have the same even polynomial part when k is odd and the same odd polynomial part when k is even. Furthermore, the last coefficient of $p_k(z)$ and $p_{k-1}(z)$ is always the same, namely $\alpha_k = \beta_{k-1}$.

Repeated use of the order reduction (3.19) on the polynomial $p_n(z)$ yields a sequence of reduced order polynomials

$$p_n(z), p_{n-1}(z), \dots, p_2(z), p_1(z),$$

and a sequence of constants

$$\mu_n, \mu_{n-1}, \dots, \mu_2, \mu_1.$$

Each row of the Routh table consists of the coefficients of either the even polynomial part or the odd polynomial part of $p_k(z)$. The last polynomial $p_1(z)$ is of the form $p_1(z) = (2\mu_1 z + 1)a_n$. We can represent the order reduction formula by a matrix operation

$$\begin{pmatrix} p_{k-1}(z)\\ p_{k-1}(-z)\end{pmatrix} = \begin{pmatrix} 1-\mu_k z & (-1)^k\mu_k z\\ -(-1)^k\mu_k z & 1+\mu_k z\end{pmatrix}\begin{pmatrix} p_k(z)\\ p_k(-z)\end{pmatrix}. \tag{3.21}$$

The reverse of (3.21) is

$$\begin{pmatrix} p_k(z)\\ p_k(-z)\end{pmatrix} = \begin{pmatrix} 1+\mu_k z & -(-1)^k\mu_k z\\ (-1)^k\mu_k z & 1-\mu_k z\end{pmatrix}\begin{pmatrix} p_{k-1}(z)\\ p_{k-1}(-z)\end{pmatrix}$$

which gives us the order augmentation formula which reconstructs $p_k(z)$ from $p_{k-1}(z)$ and $p_{k-1}(-z)$, namely

$$p_k(z) = (1+\mu_k z)p_{k-1}\left(1 - (-1)^k\frac{\mu_k z}{1+\mu_k z}\frac{p_{k-1}(-z)}{p_{k-1}(z)}\right) = (1+\mu_k z)p_{k-1}(z)g_{k-1}(z) \tag{3.22}$$

where

$$\mu_k = \frac{\alpha_0}{2\alpha_1} = \frac{\alpha_0}{2\beta_0} \tag{3.23a}$$

and

$$g_{k-1}(z) = 1 - (-1)^k\frac{\mu_k z}{1+\mu_k z}\frac{p_{k-1}(-z)}{p_{k-1}(z)}. \tag{3.23b}$$

3.4.1 The argument principle

From (3.23) two properties of $g_{k-1}(z)$ are apparent: If $\mu_k \neq 0$, we have

Property 1.

$$Re\Big(g_{k-1}(i\omega)\Big) > 0 \tag{3.24a}$$

for all real values of ω*, and*

$$\lim_{\omega\to\pm\infty} g_{k-1}(i\omega) = 1 - (-1)^k(-1)^{k-1} = 2. \tag{3.24b}$$

To show the first part, we return to the convenient split of a polynomial on the imaginary axis into its real and imaginary part and write $p_{k-1}(i\omega) = U(\omega) + iV(\omega)$. From (3.20a) it is clear that $p_{k-1}(-i\omega) = U(\omega) - iV(\omega)$, so we find that

$$\text{Re}\Big(g_{k-1}(i\omega)\Big) = \text{Re}\Big(1 - (-1)^k \frac{\mu_k i\omega}{1+\mu_k i\omega}\frac{U(\omega) - iV(\omega)}{U(\omega) + iV(\omega)}\Big).$$

Since $\text{Re}(z) \leq |z|$ for complex numbers z, it suffices to note that

$$\left|\frac{\mu_k i\omega}{1+\mu_k i\omega}\frac{U(\omega) - iV(\omega)}{U(\omega) + iV(\omega)}\right| = \left|\frac{\mu_k i\omega}{1+\mu_k i\omega}\right|\left|\frac{U(\omega) - iV(\omega)}{U(\omega) + iV(\omega)}\right| < 1.$$

Next, following Matsumoto we denote by $\Delta arg\ p_k(z)$ the net increment of the argument of $p_k(z)$ on the imaginary axis. We define $\Delta arg(1+\mu_k z)$ and $\Delta arg\ g_{k-1}(z)$ in the same way. We use the argument principle in essentially the same manner as we did in the last proof to obtain

Property 2.

$$\Delta arg(1+\mu_k z) = \begin{cases} \pi & \text{if } \mu_k > 0 \\ -\pi & \text{if } \mu_k < 0 \end{cases} \tag{3.25a}$$

$$\Delta arg\ g_{k-1}(z) = 0 \tag{3.25b}$$

If $\mu_k \neq 0$*, then*

$$\Delta arg\ p_k(z) = sign(\mu_k)\pi + \Delta\ p_{k-1}(z) \tag{3.25c}$$

and, a polynomial of degree k *satisfies*

$$\Delta arg\ p_k(z) = (k - 2R_k)\pi \tag{3.25d}$$

if $p_k(z)$ *has* R_k *zeros on the open right half-plane and* $(k - R_k)$ *zeros in the open left half-plane. Further,* $p_k(z)$ *is a Hurwitz polynomial of degree* k *if and only if*

$$\Delta arg\ p_k(z) = k\pi. \tag{3.25e}$$

Note that (3.25c) follows from (3.25a), (3.25b) and (3.23b).

Using (3.22) and Property 2 we can formulate

Theorem 6. *A real polynomial* $p_n(z)$ *is a Hurwitz polynomial if and only if*

$$\mu_k > 0,\ \ k = n, n-1, \ldots, 2,$$

where μ_k *is the constant generated by the order reduction formula* (3.19).

Proof. By (3.25e), $p_n(z)$ is a Hurwitz polynomial if and only if $\Delta arg\ p_n(z) = n\pi$, and by (3.25c) we have

$$\begin{aligned}\Delta arg\ p_n(z) = sign(\mu_n)\pi + \Delta\ p_{n-1}(z) = \\ (sign(\mu_n) + sign(\mu_{n-1}) + \cdots + sign(\mu_2))\pi + \Delta\ p_1(z).\end{aligned}$$

We have seen before that $p_1(z) = (2\mu_1 z + 1)a_n$, so it is clear that $\Delta\ p_1(z) = sign(\mu_1)$, which means that

$$\Delta arg\ p_n(z) = \sum_{i=1}^{n} sign(\mu_i).$$

This sum can be equal to n only if every term is positive, or equivalently if $\mu_k > 0$ for all k. □

If the polynomials in question have no zeros on the imaginary axes, the following corollaries hold:

Corollary 6.1. *If $\mu_k > 0$, the number of zeros of $p_k(z)$ in the left half-plane, is one more than that of $p_{k-1}(z)$, and the number of zeros of $p_k(z)$ in the right half-plane is equal to that of $p_{k-1}(z)$.*

Corollary 6.2. *If $\mu_k < 0$, the number of zeros of $p_k(z)$ in the left right-plane, is one more than that of $p_{k-1}(z)$, and the number of zeros of $p_k(z)$ in the right left-plane is equal to that of $p_{k-1}(z)$.*

Corollary 6.3. *If every μ_k is nonzero, the number of negative μ_k:s among $\{\mu_n, \mu_{n-1}, \ldots, \mu_2, \mu_1\}$ coincides with the number of zeros of $p_n(z)$ in the right half-plane.*

The third corollary follows from the two preceding ones. We prove only the first one since the second is completely analogous. Suppose that $\mu_k > 0$. By (3.25) and (3.25c) we have

$$\Delta arg\ p_k(z) = sign(\mu_k)\pi + \Delta arg\ p_{k-1}(z) = \pi + \Delta arg\ p_{k-1}(z) = (k - 2R_k)\pi,$$

where R_k is the number of zeros of $p_k(z)$ in the right half-plane. It follows that

$$\Delta arg\ p_{k-1}(z) = \Delta arg\ p_k(z) - \pi = (k - 2R_k)\pi - \pi = ((k-1) - 2R_k)\pi,$$

so, the number of zeros in the right half-plane is the same. It follows from (3.25) that the number of zeros of $p_k(z)$ in the left half-plane is $(k - R_k)$ and therefore that the number of such zeros of $p_{k-1}(z)$ is $(k - 1 - R_k)$.

3.4.2 Sign changes in the Routh table

Using the coefficients of $p_k(z)$ and $p_{k-1}(z)$, the $(n-k+1)$:th row and the $(n-k+2)$:th row of the Routh table are

$$\begin{array}{llllll} (n-k+1)\text{:th row:} & \alpha_0, & \alpha_2, & \alpha_4, & \alpha_6, & \alpha_8, & \ldots \\ (n-k+2)\text{:th row:} & \beta_0, & \beta_2, & \beta_4, & \beta_6, & \beta_8, & \ldots \end{array}$$

Since the order reduction formula (3.19) propagates the coefficients from $p_k(z)$ to $p_{k-1}(z)$ as $\beta_{2i} = \alpha_{2i+1}$ as defined by (3.20a), the $(n-k+2)$:th row can also be expressed as

$$\begin{array}{llllll} (n-k+2)\text{:th row:} & \alpha_1, & \alpha_3, & \alpha_5, & \alpha_7, & \alpha_9, & \ldots \end{array}$$

We introduce the notation $R_{i,j}$ to denote the element in the i:th row and j:th column of the Routh table. Then μ_k can be expressed as

$$2\mu_k = \frac{\alpha_0 \text{ in } p_k(z)}{\beta_0 \text{ in } p_{k-1}(z)} = \frac{\alpha_0 \text{ in } p_k(z)}{\alpha_1 \text{ in } p_k(z)} = \frac{R_{(n-k+1),1}}{R_{(n-k+2),1}}, \tag{3.26}$$

where α_0 in $p_k(z)$ simply means the coefficient α_0 in the polynomial $p_k(z)$. Therefore, the number of sign changes in the first column of the Routh table coincides with the number of negative numbers among $\{\mu_n, \mu_{n-1}, \ldots, \mu_2, \mu_1\}$. Hence Corollaries 6.1-6.3 leads to

Theorem 7. *If every μ_k is nonzero, the number of sign changes in the first column of the Routh table for a polynomial $p_n(z)$ of degree n coincides with the number of zeros of $p_n(z)$ in the open right half-plane, provided that $p_n(z)$ has no zeros on the imaginary axis.*

Matsomuto does not discuss the singular case covered in [10], but we cover it in appendix A to which the interested reader is referred. In fact, we will only need the Routh criterion to give sufficient conditions on the characteristic polynomial for stability of a fixed point.

3.5 Sufficient conditions for stability

Given a fixed point x^* of a discrete-time dynamical system

$$x \mapsto f(x), x \in \mathbb{R}^n$$

with smooth f, theorem 1 tells us that x^* is stable the Jacobian matrix evaluated at x^*, denoted by A satisfy that all the eigenvalues of A lie inside the unit circle. This can be formulated it terms of the zeros of the characteristic polynomial

$$p(\mu) = \det(A - \mu I_n).$$

Using Routh's criterion, it is easy to give sufficient conditions for the zeros to lie in the open left-plane, but this is not immediately useful to us. Note however that the *Möbius transformation*

$$z \mapsto \frac{z+1}{z-1} \tag{3.27}$$

maps the unit disc onto the left half plane. Hence the zeros of $p(z)$ lie inside the unit circle if $p(\frac{z+1}{z-1})$ has all zeros in the open left half plane, i.e.

$$q(z) := p(\frac{z+1}{z-1})(z-1)^2$$

is a Hurwitz polynomial.

One can give explicit conditions on the coefficients of $p(z)$ for stability. We show this in the two-dimensional case. For a second-degree polynomial $p(z) = a_0 z^2 + b_0 z + a_1$, the Routh table is

$$\begin{array}{ll} a_0, & a_1 \\ b_0 & \\ a_1 & \end{array}$$

The Routh criterion then states that both zeros of $p(z)$ have negative real part if an only if a_0, b_0, a_1 are of like sign, and none of them are zero.

Now, the zeros of a polynomial do not change upon division by the leading coefficient, so without loss of generality we may assume that the characteristic polynomial is monic. Hence a generic characteristic polynomial in the two-dimensional case is

$$p(z) = z^2 + \alpha_1 z + \alpha_0.$$

Now we use the Möbius transformation (3.27) to define

$$q(z) := p(\frac{z+1}{z-1})(z-1)^2 = (1 + \alpha_0 + \alpha_1)z^2 + (2 - 2\alpha_0)z + (1 + \alpha_0 - \alpha_1).$$

By Routh's criterion, $q(z)$ has both its zeros in the open left half-plane if and only if the numbers

$$(1+\alpha_0+\alpha_1), (2-2\alpha_0), (1+\alpha_0-\alpha_1)$$

are nonzero and of like sign.

Assume first that all these numbers are positive, that is

$$\begin{cases} 1+\alpha_0+\alpha_1 > 0 \\ 2-2\alpha_0 > 0 \\ 1+\alpha_0-\alpha_1 > 0 \end{cases} \iff \begin{cases} 1+\alpha_0 > -\alpha_1 \\ 1+\alpha_0 > \alpha_1 \\ \alpha_0 < 1 \end{cases} \iff \begin{cases} |\alpha_1| < 1+\alpha_0 \\ |\alpha_0| < 1. \end{cases}$$

Assuming that they are all negative yields no solutions, so this is the only solution.

Now, note that the characteristic polynomial of a 2×2-matrix A is

$$p(\mu) = \det(\begin{pmatrix} a_{11}-\mu & a_{12} \\ a_{21} & a_{22}-\mu \end{pmatrix}) = (a_{11}-\mu)(a_{22}-\mu) - a_{21}a_{12} =$$
$$\mu^2 - (a_{11}+a_{22})\mu + (a_{11}a_{22} - a_{21}a_{12}) = \mu^2 - \operatorname{trace}(A)\mu + \det(A),$$

so, we have found that for a fixed point x^* of a two-dimensional discrete-time smooth dynamical system, with Jacobian matrix A evaluated at x^*, sufficient conditions for stability of x^* are

$$\begin{cases} |\operatorname{trace}(A)| < 1+\det(A) \\ |\det(A)| < 1. \end{cases} \tag{3.28}$$

4 Bifurcation analysis

Now consider a system that depends on parameters, which we write as

$$x \mapsto f(x, \alpha) \tag{4.1}$$

were $x \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}^m$. As the parameters vary, the phase portrait also varies, and there are two possibilities. Either the system remains topologically equivalent to the original one, or its topology changes.

Definition 16. *The appearance of a topologically non-equivalent phase portrait under variation of parameters is called a bifurcation.*

Thus, a bifurcation is a change of the topological type of the system as its parameters pass through a *bifurcation (critical) value.*

Definition 17. *The codimension of a bifurcation is the difference between the dimension of the parameter space and the dimension of the corresponding bifurcation boundary. Or equivalently, the codimension is the number of independent conditions determining the bifurcation.*

4.1 One-parameter bifurcation of fixed points

Now let

$$x \mapsto f(x, \alpha),\ x \in \mathbb{R}^n,\ \alpha \in \mathbb{R}^1, \tag{4.2}$$

where f is smooth with respect to both x and α. Let x^* be a hyperbolic fixed point of the system for $\alpha = \alpha_0$. Generically there are just three ways in which hyperbolicity may be violated. Either a simple multiplier approaches the unit circle, and we have $\mu_1 = 1$ or $\mu_1 = -1$, or a pair of complex conjugate multipliers approaches the unit circle, and we have $\mu_{1,2} = e^{\pm i\theta_0}, 0 < \theta_0 < \pi$.

Definition 18. *The bifurcation associated with the appearance of* $\mu_1 = 1$ *is called a fold bifurcation.*

Definition 19. *The bifurcation associated with the appearance of* $\mu_1 = -1$ *is called a flip- or period-doubling bifurcation.*

Definition 20. *The bifurcation associated with the appearance of* $\mu_{1,2} = e^{\pm i\theta_0}, 0 < \theta_0 < \pi$ *is called a Neimark-Sacker bifurcation.*

Note that flip and fold bifurcation may appear in one-dimensional systems, while Neimark-Sacker requires at least dimension two. However, for an n-dimensional system, these bifurcations occur in essentially the same way. As we shall see, there are certain one- or two-dimensional *invariant manifolds* on which the system exhibits the corresponding bifurcations, while the behaviour off the manifold is in some sense "trivial".

Theorem 8 (Generic flip)**.** *Suppose that a one-dimensional system*

$$x \mapsto f(x, \alpha), \ \ x \in \mathbb{R}, \ \ \alpha \in \mathbb{R},$$

with smooth map f*, has at* $\alpha = 0$ *the fixed point* $x^* = 0$*, and let* $\mu = f_x(0, 0) = -1$*, where* f_x *denotes derivative. Assume that the following nondegeneracy conditions are satisfied:*

$$\frac{1}{2}(f_{xx}(0,0))^2 + \frac{1}{3} f_{xxx}(0,0) \neq 0 \tag{B.1}$$

$$f_{x\alpha}(0,0) \neq 0. \tag{B.2}$$

Then there are smooth invertible coordinate and parameter changes transforming the system into

$$\eta \mapsto -(1+\beta)\eta \pm \eta^3 + O(\eta^4).$$

The proof which is given in in Chapter 4 in [4] is not difficult but we do not give it here. The system

$$\eta \mapsto -(1+\beta)\eta \pm \eta^3 \tag{4.3}$$

is called the *topological normal form* for the flip bifurcation. The sign of the cubic term depends on the sign of

$$c(0) = \frac{1}{4}(f_{xx}(0,0))^2 + \frac{1}{6} f_{xxx}(0,0).$$

Any generic, scalar, one-parameter system that satisfy the conditions in the theorem is locally topologically equivalent near the origin to (4.3). Depending on the sign of the cubic term, the flip is called stable or unstable. If the cubic term is positive, the flip is stable, which means that the 2-cycle thus appearing is stable.

Regarding the Neimark-Sacker bifurcation we refer to [4] for the relevant theorem and normal form. We just state the nondegeneracy conditions:

$$\rho'(0) \neq 0, \tag{C.1}$$

$$e^{ik\theta_0} \neq 1 \text{ for } k = 1, 2, 3, 4, \tag{C.2}$$

$$d(0) \neq 0, \tag{C.3}$$

where the system has smooth map $f(x, \alpha)$, $x \in \mathbb{R}^2$ with eigenvalues $\mu_{1,2}(\alpha) = \rho(\alpha)e^{i\varphi(\alpha)}$, where $\varphi(0) = \theta_0$. We will return to the third condition later.

4.2 Center manifolds

Consider a discrete-time dynamical system

$$x \mapsto f(x), x \in \mathbb{R}^n \tag{4.4}$$

where f is sufficiently smooth and $f(0) = 0$. Denote by n_-, n_0, n_+ the number of eigenvalues of the Jacobian matrix A evaluated at the fixed point $x^* = 0$ inside, on and outside the unit circle respectively. Assuming that the fixed point is non-hyperbolic, we have that $n_0 \neq 0$. Denote by T^c the linear invariant generalized eigenspace of A corresponding to the union of n_0 eigenvalues on the unit circle.

For clarity we define explicitly

Definition 21. *For a complex $n \times n$-matrix A with eigenvalues $\{\mu_1, \ldots, \mu_k\}$, the generalized eigenspace corresponding to μ_i is*

$$V_{\mu_i} = \{x \in \mathbb{C}^n : (A - \mu_i I_n)^n x = 0\}.$$

The following theorems from [4], which we give without proofs or detailed discussion, will be used to check genericity of bifurcations later on.

Theorem 9 (Center manifold theorem)**.** *There is a locally defined smooth n_0-dimensional invariant manifold $W^c_{loc}(0)$ of* (4.4) *that is tangent to T^c at $x = 0$. Moreover, there is a neighbourhood U of $x^* = 0$ such that if the k:th iterate of f, $f^k(x) \in U$ for all $k \geq 0$, then $f^k(x) \to W^c_{loc}(0)$ as $k \to \infty$. The manifold W^c_{loc} is called the center manifold.*

It is convenient to drop the subscript and just write W^c for the center manifold.

We may write the system (4.4) in an eigenbasis to get

$$\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} Bu + g(u,v) \\ Cv + h(u,v) \end{pmatrix}, \tag{4.5}$$

where $u \in \mathbb{R}^{n_0}, v \in \mathbb{R}^{n_+ + n_-}$, B is an $n_0 \times n_0$-matrix with all its n_0 eigenvalues on the unit circle while C is an $(n_+ + n_-) \times (n_+ + n_-)$-matrix with no eigenvalues on the unit circle. The functions $g(u,v), h(u,v)$ have Taylor expansions starting with at least quadratic terms. A center manifold W^c of (4.5) can be locally represented as the graph of a smooth function

$$W^c = \{(u,v) : v = V(u)\}$$

where $V : \mathbb{R}^{n_0} \to \mathbb{R}^{n_+ + n_-}$, and since W^c is tangent to T^c at $x^* = 0$ we have $V(u) = O(||u||^2)$.

Theorem 10 (Reduction principle)**.** *The system* (4.5) *is locally topologically equivalent near the origin to the system*

$$\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} Bu + g(u, V(u)) \\ Cv \end{pmatrix}. \tag{4.6}$$

If there is more than one center manifold, then all the resulting maps (4.6) *are locally smoothly conjugate.*

4.3 Computation of center manifolds

The following method for computing center manifolds is called the projection method. Only eigenvectors corresponding to the critical eigenvalues of A and its transpose A^T are used to "project" the system into the critical eigenspace and its complement. The method is based on the Fredholm alternative theorem and can be used both for continuous and discrete-time systems. What follows is a somewhat technical but straightforward computation, taken more or less directly from chapter 5 of [4].

We write the system (4.4) as

$$\tilde{x} = Ax + F(x), x \in \mathbb{R}^n \tag{4.7}$$

where $F(x) = O(||x||^2)$ is a smooth function with Taylor expansion near $x^* = 0$ as

$$F(x) = \frac{1}{2}B(x,x) + \frac{1}{6}C(x,x,x) + O(||x||^4), \tag{4.8}$$

where $B(x,y)$ and $C(x,y,z)$ are multilinear functions. In coordinates we have

$$B_i(x,y) = \sum_{j,k=1}^{n} \frac{\partial^2 F_i(\xi)}{\partial \xi_j \partial \xi_k}\Big|_{\xi=0} x_j y_k, \tag{4.9}$$

and

$$C_i(x,y) = \sum_{j,k,l=1}^{n} \frac{\partial^3 F_i(\xi)}{\partial \xi_j \partial \xi_k \partial \xi_l}\Big|_{\xi=0} x_j y_k z_l, \tag{4.10}$$

where $i = 1, 2, \ldots, n$.

4.3.1 Flip bifurcations

In the case of a flip bifurcation, A has a simple critical eigenvalue $\mu_1 = -1$, and the corresponding critical eigenspace T^c is one-dimensional and spanned by an eigenvector $q \in \mathbb{R}^n$ such that $Aq = \mu_1 q$. Let p be the *adjoint* eigenvector, that is $A^T p = \mu_1 p$. Normalize p with respect to q so that $\langle p, q\rangle = 1$, where $\langle ., .\rangle$ is the standard scalar product in $\mathbb{R}^n$. The following lemma follows from the Fredholm alternative theorem.

Lemma 11. *Let T^{su} denote an $(n-1)$-dimensional linear eigenspace of $(A - \mu_1 I_n)$ corresponding to all eigenvalues other than $\mu_1 = -1$. Then $y \in T^{su}$ if and only if $\langle p, y\rangle = 0$.*

Using the lemma, taking into account that the matrix $(A - \mu_1 I_n)$ has common invariant spaces with the matrix A, we can decompose any vector $x \in \mathbb{R}^n$ as

$$x = uq + y,$$

where $uq \in T^c$ and $y \in T^{su}$ and

$$\begin{cases} u = \langle p, x\rangle \\ y = x - \langle p, x\rangle q. \end{cases} \tag{4.11}$$

We can define two operators:

$$P_c x = \langle p, x\rangle q, \;\; P_{su} x = x - \langle p, x\rangle q.$$

These operators are projections onto T^c and T^{su} respectively, and

$$P_c^2 = p_c, \;\; P_{su}^2 = P_{su}, \;\; P_c P_{su} = P_{su} P_c = 0.$$

To realize this, we need only note that $\langle p, \langle p, x\rangle q\rangle q = \langle p, x\rangle q$ since we normalized p with respect to q. We show one of the identities explicitly. The rest are shown completely analogously:

$$\begin{aligned} P_{su}^2 x = P_{su}(P_{su}x) = P_{su}(x - \langle p, x\rangle q) &= x - \langle p, x\rangle q - \langle x, (p - \langle p, x\rangle q)\rangle q \\ &= x - \langle p, x\rangle q - \Big(\langle p, x\rangle q - \langle p, \langle p, x\rangle q\rangle q\Big) = x - \langle p, x\rangle q = P_{su}x. \end{aligned}$$

The scalar u and the vector y can be considered as new "coordinates" on $\mathbb{R}^n$. Although $y \in \mathbb{R}^n$, it always satisfies the orthogonality condition $\langle p, y\rangle = 0$ since

$$\langle p, y\rangle = \langle p, x - \langle p, x\rangle q\rangle = \langle p, x\rangle - \langle p, \langle p, x\rangle q\rangle = \langle p, x\rangle - \langle p, x\rangle\langle p, q\rangle = 0.$$

In the coordinates (u, y) the map (4.4) can be written as

$$\begin{cases} \tilde{u} = \mu_1 u + \langle p, F(uq + y)\rangle, \\ \tilde{y} = Ay + F(uq + y) - \langle p, F(uq + y)\rangle q. \end{cases} \tag{4.12}$$

Using Taylor expansion (4.8) we can write (4.12) in the form

$$\begin{cases} \tilde{u} = \mu_1 u + bu^2 + u\langle p, B(q, u)\rangle + ru^3 + \dots, \\ \tilde{y} = Ay + \frac{1}{2}au^2 + \dots, \end{cases} \tag{4.13}$$

where $u, b, r \in \mathbb{R}^n$ and $y, a \in \mathbb{R}^n$, and

$$b = \frac{1}{2}\langle p, B(q, q)\rangle, \tag{4.14}$$

$$r = \frac{1}{6}\langle p, C(q, q, q)\rangle, \;\; a = B(q, q) - \langle p, B(q, q)\rangle q. \tag{4.15}$$

We seek the second order term in the Taylor expansion for $y = V(u)$ representing the center manifold:

$$V(u) = \frac{1}{2}w_2 u^2 + O(u^3), \tag{4.16}$$

where $w_2 \in \mathbb{R}^n$ is an unknown vector. Since $V(u) \in T^{su}$ for small u, we have that $\omega_2 \in T^{su}$ implying that $\langle p, w_2\rangle = 0$. This vector w_2 satisfies the equation

$$(A - I_n)w_2 = -a. \tag{4.17}$$

This equation results from comparing the coefficients of the u^2-terms in the invariance condition for W^c,

$$\tilde{y} = V(\tilde{u}),$$

where $\tilde{u}$ and $\tilde{y}$ are given by (4.13).

The matrix $(A - I_n)$ is invertible in $\mathbb{R}^n$ since $\lambda = 1$ is not an eigenvalue of A in the flip case. Therefore, we can solve equation (4.17), giving $w_2 = -(A - I_n)^{-1}a$, and the restriction of (4.13) to the center manifold takes the form

$$\tilde{u} = -u + bu^2 + \left(r - \frac{1}{2}\langle p, B(q, (A - I_n)^{-1}a)\rangle\right)u^3 + O(u^4). \tag{4.18}$$

This restricted map can be simplified. Using (4.14) and the identity

$$(A - I_n)^{-1}q = -\frac{1}{2}q,$$

we can write the restricted map as

$$\tilde{u} = -u + a_0 u^2 + b_0 u^3 + O(u)^4, \tag{4.19}$$

where

$$a_0 = \frac{1}{2}\langle p, B(q,q)\rangle$$

and

$$b_0 = \frac{1}{6}\langle p, C(q,q,q)\rangle - \frac{1}{4}(\langle p, B(q,q)\rangle)^2 - \frac{1}{2}\langle p, B(q,(A-I_n)^{-1}B(q,q))\rangle.$$

It is shown in Chapter 4 in [4] that the map 4.19 can be transformed to the *normal form*

$$\tilde{\xi} = -\xi + c\xi^3 + O(\xi^4),$$

where

$$c = a_0^2 + b_0.$$

The normal form is such that any generic, scalar, one-parameter system with eigenvalue -1 at a fixed point is topologically equivalent to it. For a precise definition see section 4.5 in [4].

Thus, the critical normal form coefficient c, that determines the nondegeneracy of the flip bifurcation and allows us to predict the direction of bifurcation of the period-two cycle, is given by the *invariant formula*

$$c = \frac{1}{6}\langle p, C(q,q,q)\rangle - \frac{1}{2}\langle p, B(q,(A-I_n)^{-1}B(q,q))\rangle. \tag{4.20}$$

The Neimark-Sacker bifurcation is handled similarly. We do not give the details here but refer to section 5.4 in [4]. However, we state that the third nondegeneracy condition C.3 can be computed as

$$d = \frac{1}{2}\mathrm{Re}\Bigg(e^{-i\theta_0}\Bigg[\langle p, C(q,q,\bar{q})\rangle + 2\langle p, B(q,(A-I_n)^{-1}B(q,\bar{q})\rangle + \langle p, B(\bar{q},(e^{2i\theta_0}I_n - A)^{-1}B(q,q))\rangle\Bigg]\Bigg), \tag{4.21}$$

where q now is a *complex* eigenvector corresponding to $\mu_1 = e^{i\theta_0}$:

$$Aq = e^{i\theta_0}q, \;\; A\bar{q} = e^{-i\theta_0}\bar{q},$$

where $\bar{q}$ is the vector of complex conjugates of the elements in q.

4.4 List of codimension 2 bifurcations in $\mathbb{R}^2$

In our coming analysis we will consider a two-dimensional dynamical system, so we need only consider this case.

Consider a two-dimensional, two-parameter discrete-time dynamical system

$$x \mapsto f(x,\alpha) \tag{4.22}$$

with $x \in \mathbb{R}^2$ and $\alpha = (\alpha_1, \alpha_2)^T$ and f sufficiently smooth in (x,α) e.g. $f \in C^1$. Suppose that at $\alpha = \alpha_0$, the system (4.22) has a fixed point x^* for which the condition for fold, flip or Neimark-Sacker bifurcation is satisfied. Then there are eight degenerate cases that may occur.

(1) $\mu_1 = 1, b = 0$ *(cusp)*

(2) $\mu_1 = -1, c = 0$ *(generalized flip)*

(3) $\mu_{1,2} = e^{\pm i\theta_0}, d = 0$ *(Cheniciner bifurcation)*

(4) $\mu_1 = \mu_2 = 1$ *(1:1 resonance)*

(5) $\mu_1 = \mu_2 = -1$ *(1:2 resonance)*

(6) $\mu_{1,2} = e^{\pm i\theta_0}, \theta_0 = \frac{2\pi}{3}$ *(1:3 resonance)*

(7) $\mu_{1,2} = e^{\pm i\theta_0}, \theta_0 = \frac{\pi}{2}$ *(1:4 resonance)*

(8) $\mu_1 = 1, \mu_2 = -1$ *(fold-flip bifurcation)*

5 Cycles of period 3 and Sharkovskii's theorem

The next simplest type of orbit is a cycle. In discrete-time systems, a cycle of length k corresponds to a fixed point of the k:th iterate f^k. An interesting question to pose is whether one can draw any conclusions about the existence of cycles of other lengths from the presence of a cycle of length k.

In the paper "Period three implies chaos" [13], Li and Yorke were the first to introduce the word *chaos* in mathematics. In the paper, they show that if a continuous map has a cycle of period 3, then it must have cycles of any period k. This quite non-intuitive result is in fact a special case of a remarkable theorem of Sharkovskii. To state the theorem, we must first present a new ordering $\triangleright$ of the positive integers as follows:

$$\begin{aligned} &3 \triangleright 5 \triangleright 7 \triangleright \cdots \triangleright 2\cdot 3 \triangleright 2\cdot 5 \triangleright 2\cdot 7 \triangleright \cdots \triangleright 2^2\cdot 3 \triangleright 2^2\cdot 5 \triangleright 2^2\cdot 7 \cdots \triangleright \dots \\ &\triangleright 2^n\cdot 3 \triangleright 2^n\cdot 5 \triangleright 2^n\cdot 7 \cdots \triangleright \cdots \triangleright 2^n \triangleright 2^{n-1} \triangleright \cdots \triangleright 2^2 \triangleright 2 \triangleright 1. \end{aligned}$$

First the odd integers are listed, except 1, then 2 times the odd integers, followed by 2^2 times the odd integers, and in general 2^n times the odd integers for all positive integers n. Finally, one lists the powers of 2 in descending order. Clearly all positive integers are generated this way. The notation $m \triangleright n$ means that the positive integer m comes before n in the Sharkovskii ordering. In particular, this means that $3 \triangleright k$ for any positive integer k. We are now ready to state

Theorem 12. *Let $f : I \to I$ be a continuous map on the interval I, where I may be finite, infinite, or the whole real line. If f has a cycle of period k, then it has a cycle of period r for all r with $k \triangleright r$.*

5.1 Proof of Sharkovskii's theorem

In this section we give a proof of Sharkovskii's theorem. The proof is in essence the same as that given in [14]. Throughout the section, it will be convenient to refer to *periodic points* of the map f. We give a formal definition.

Definition 22. *A point c is said to be a periodic point of f with period m if $f^m(c) = c$ and $f^k(c) \neq c$ for $1 \leq k < m$. The orbit of c then consists of the m distinct points $c, f(c), f^2(c), \dots, f^{m-1}(c)$. By abuse of language, the orbit of c can also be said to be periodic. In particular, a fixed point is a periodic point with period 1.*

Note that f has a periodic point with period m if and only if f has a cycle of length m, and that any point c in this cycle is a periodic point with period m.

The proof of Sharkovskii's theorem will be quite involved and along the way, other results of independent interest will be derived. Throughout we will assume unless otherwise stated that $f : I \to \mathbb{R}$ is a continuous map of a compact (closed and bounded) interval into the real line. We will denote by $< a, b >$ the closed interval with endpoints a and b when we do not know or care whether $a < b$ or $b < a$. Since this notation is not standard, we shall take care to explicitly state the meaning whenever it is used.

Lemma 13. *If J is a compact subinterval such that $J \subseteq f(J)$, where $f(J)$ is the image of J under f, then f has a fixed point in J.*

Proof. If $I = [a, b]$, then for some $c, d \in J$ we have $f(c) = a$ and $f(d) = b$ by the intermediate value theorem. Thus $f(c) \leq c$ and $f(b) \geq b$. Form the continuous map $g(x) = f(x) - x$ and note that $g(c) \leq 0$ and $g(d) \geq 0$. One more application of the intermediate value theorem yields some point $\xi \in [a, b]$ such that $g(\xi) = 0$, but then $f(\xi) = \xi$ and we are done. $\square$

Lemma 14. *If J and K are compact subintervals such that $K \subseteq f(J)$, then there is a compact subinterval $L \subseteq J$ such that $f(L) = K$.*

Proof. Let $K = [a, b]$ and let c be the greatest point in J for which $f(c) = a$. If $f(x) = b$ for some $x \in J$ with $x > c$, let d be the least. Then we can take $L = [c, d]$.

Otherwise $f(x) = b$ for some $x \in J$ with $x < c$. Let c' be the greatest and let $d' \leq c$ be the least $x \in J$ with $x > c'$ for which $f(x) = a$. Then we can take $L = [c', d']$. $\square$

Lemma 15. *If $J_0, J_1, \ldots, J_m$ are compact subintervals such that $J_k \subseteq f(J_{k-1})$ for $1 \leq k \leq m$, then there is a compact subinterval $L \subseteq J_0$ such that $f^m(L) = J_m$ and $f^k(L) \subseteq J_k$ for $1 \leq k < m$.*

If in addition $J_0 \subseteq J_m$, then there exists a point y such that $f^m(y) = y$ and $f^k(y) \in J_k$ for $0 \leq k < m$.

Proof. We prove the first assertion by induction. The first assertion holds for $m = 1$ by Lemma 14. We assume that $m > 1$ and that it holds for all smaller values of m. Then we can choose $L' \subseteq J_1$ so that $f^{m-1}(L') = J_m$ and $f^k(L') \subseteq J_{k+1}$ for $1 \leq k < m - 1$. We now choose $L \subseteq J_0$ so that $f(L) = L'$.

The second assertion follows from the first by Lemma 13. $\square$

As a first independently interesting result, we prove

Proposition 16. *Between any two points of a periodic orbit with period $n > 1$ there is a point of a periodic orbit of period less than n.*

Proof. Let $a < b$ be two adjacent points of the orbit of period n. Since there is one more point of the orbit to the left of b than to the left of a we must have $f^m(a) > a$ and $f^m(b) < b$ for some m such that $1 \leq m < n$. It follows immediately that $f^m(c) = c$ for some c with $a < c < b$, if we assume that f^m is defined throughout $[a, b]$.

However, we do not need this assumption. For if we assume that f^m is not defined throughout $[a, b]$, let $J_k =< f^k(a), f^k(b) >$ be the closed interval with endpoints $f^k(a)$ and $f^k(b)$ for $1 \leq k \leq m$, then

$$J_k \subseteq f(J_{k-1}) \text{ for } 1 \leq k \leq m.$$

This is clear since if $J_{k-1} =< f^{k-1}(a), f^{k-1}(b) >$, then J_k has as endpoints the image under f of the endpoints of J_{k-1}. Hence the image of J_{k-1} contains at least J_k, for it contains its endpoints. But also $J_0 \subseteq J_m$ since $f^m(a) \geq b$ and $f^m(b) \leq a$. The result now follows from Lemma 15. $\square$

The method of argument can be refined. Suppose that f has a periodic orbit of period $n > 1$. Let $x_1 < x_2 < \cdots < x_n$ be the distinct points of this orbit. Note that f is a cyclical permutation of the orbit. Set $I_j = [x_j, x_{j+1}]$ for $1 \leq j < n$.

With the periodic orbit we associate a *directed graph* or *digraph* in the following way.

Definition 23. *Let $I_i = [x_i, x_{i+1}]$. The digraph of a periodic orbit is a directed graph with the subintervals $I_1, \ldots, I_{n-1}$ as vertices. There is a directed edge, which we will refer to as an arc $I_j \to I_k$ if I_k is contained in the closed interval $< f(x_j), f(x_{j+1}) >$ with endpoints $f(x_j)$ and $f(x_{j+1})$. That is $I_j \to I_k$ if $I_k \subseteq f(I_j)$*

For example, suppose c is a periodic point of period 3 with $f(c) < c < f^2(c)$. The corresponding digraph has two vertices, namely the intervals $I_1 = [f(c), c]$ and $I_2 = [c, f^2(c)]$, connected in the following way:

$$\begin{tikzcd} I_1 \arrow[loop left] \arrow[r, shift left] & I_2 \arrow[l, shift left] \end{tikzcd}$$

Some properties of our digraphs follow from the definition:

(i) For any vertex I_j there is always at least one vertex I_k for which $I_j \to I_k$. Moreover, it is always possible to choose $k \neq j$ unless $n = 2$. The proof i trivial; the endpoints of I_j cannot be mapped to the same point, so $f(I_j)$ contains at least one subinterval, and unless $n = 2$, both endpoints cannot be mapped to each other. Of course, when $n = 2$, the endpoints *must* be mapped to each other.

(ii) For any vertex I_k there is at least one vertex I_j for which $I_j \to I_k$. Moreover, it is always possible to choose $j \neq k$ unless n is even and $k = \frac{n}{2}$.

We prove this by contradiction. Suppose there is no $j \neq k$ for which $I_j \to I_k$. Then if $i \neq k$, $f(x_i) \leq x_k$ implies $f(x_{i+1}) \leq x_k$ and likewise $f(x_i) \geq x_{k+1}$ implies $f(x_{i+1}) \geq x_{k+1}$.

If $f(x_{k+1}) \geq x_{k+1}$ it follows that $f(x_i) \geq x_{k+1}$ for $k < i \leq n$. But this is impossible since no proper subset of the orbit can be mapped into itself by f. Hence $f(x_{k+1}) \leq x_k$ and similarly $f(x_k) \geq x_{k+1}$, and therefore $I_k \to I_k$.

Moreover $f(x_i) \leq x_k$ for $k < i \leq n$ implies $n - k \leq k$ which in turn implies $n < 2k$. Similarly, $f(x_i) \geq x_{k+1}$ for $1 \leq i \leq k$ implies that $k \leq n - k$, that is $n \geq 2k$. Hence $n = 2k$ and we are done.

(iii) The digraph always contains a loop. Since we must have $f(x_1) > x_1$ and $f(x_n) < x_n$, we have $f(x_j) > x_j$ and $f(x_{j+1}) < x_{j+1}$ for some j with $1 \leq j < n$. Then $f(x_j) \geq x_{j+1}$ and $f(x_{j+1}) \leq x_j$, and hence $I_j \to I_j$. To be explicit, choose

$$j = min\{1 \leq j < n : f(x_j) \geq x_{j+1}, f(x_{j+1}) \leq x_j\},$$

then $I_j \to I_j$ and we have a loop.

Definition 24. *A cycle $J_0 \to J_1 \to \cdots \to J_{n-1} \to J_0$ of length n in the digraph is called a fundamental cycle if J_0 contains an endpoint c such that $f^k(c)$ is an endpoint of J_k for $1 \leq k < n$.*

A fundamental cycle always exists and is unique, since without loss of generality take $c = x_1$, so that $J_0 = I_1$. Suppose $J_0, \ldots, J_{i-1}$ has been defined. If $J_{i-1} = [a, b]$, so that f^{i-1} is either a or b, we must take J_i to be the uniquely determined interval $I_k \subseteq < f(a), f(b) >$ which has $f^i(c)$ as one endpoint. It is clear that $J_n = J_0$ since n is the period of the orbit, and hence we obtain a cycle of length n.

In the fundamental cycle, some vertex must occur at least twice by the pigeonhole principle. There are $n-1$ vertices but the cycle has length n. On the other hand, any vertex occurs at most twice, since each interval I_k has only two endpoints.

Definition 25. *A cycle in a digraph is said to be primitive if it does not consist entirely of a cycle of smaller length described several times.*

If the fundamental cycle contains I_k twice, then it can be decomposed into two cycles of smaller length, each of which contains I_k only once, and therefore is primitive.

Straffin [15], who first showed the relevance of directed graphs in connection to Sharkovskii's theorem, observed that the existence of a primitive cycle of length m implies the existence of a periodic point of period m.

Lemma 17 (Straffin)**.** *Suppose f has a periodic point of period $n > 1$. If the associated digraph contains a primitive cycle*

$$J_0 \to J_1 \to \cdots \to J_{m-1} \to J_0$$

of length m, then f has a periodic point y of period m such that $f^k(y) \in J_k$ for $0 \leq k < m$.

Proof. The situation is that

$$J_1 \subseteq f(J_0), J_2 \subseteq f(J_1), \ldots, J_{m-1} \subseteq f(J_{m-1}), J_0 \subseteq f(J_{m-1}).$$

Then by Lemma 15 with $J_m = J_0$, there exists some point y such that $f^m(y) = y$ and $f^k(y) \in J_k$ for $0 \leq k < m$. Now, either m is the period of y, or the period of y is a divisor of m.

Since the cycle is primitive, and distinct intervals have at most one point in common, it follows that y has period m unless $y = x_i$ for some i and n is a divisor of m. However, this is possible only if the cycle is a multiple of the fundamental cycle since, given J_{k-1}, the requirements $f^k(y) \in J_k$ and $J_{k-1} \to J_k$ uniquely determine J_k. Hence, we must have $m = n$. □

Already we can prove, using Lemma 17 that the presence of a periodic point of period 3 implies the existence of any period. Consider the associated digraph

$$\begin{tikzcd} I_1 \arrow[loop left] \arrow[r, bend left] & I_2 \arrow[l, bend left] \end{tikzcd}.$$

Corresponding to the loop $I_1 \to I_1$ there is a fixed point of f and corresponding to the primitive cycle $I_1 \to I_2 \to I_1$ there is a point of period 2. Moreover, for any positive integer $m > 1$, there is a point of period m corresponding to the primitive cycle $I_1 \to I_2 \to I_1 \to I_1 \to \cdots \to I_1$ of length m.

Proposition 18. *If f has a periodic point of period > 1, then it has a fixed point and a periodic point of period 2.*

Proof. The first assertion follows since the digraph of any periodic orbit has a loop. More simply put, if f has no fixed point, then either $f(x) > x$ for all x or $f(x) < x$ for all x, and therefore f has no periodic point.

To prove the second assertion, let n be the least positive integer greater than 1 such that f has a periodic point of period n. If $n > 2$, decompose the fundamental cycle into two primitive cycles. This can always be done, since some vertex I_k appears twice in the fundamental cycle. Then at least one of the primitive cycles has length greater than 1, and obviously both of them has length less than n. By Lemma 17 we deduce that there is a periodic point with period strictly between 1 and n. □

$$
\begin{tikzcd}
J_1 \arrow[loop left] \arrow[r] & J_2 \arrow[r] & J_3 \arrow[r] & \cdots \arrow[r] & J_{n-3} \arrow[r] & J_{n-2} \arrow[r] & J_{n-1} \arrow[llllll, bend left] \arrow[llll, bend left] \arrow[l, bend left]
\end{tikzcd}
$$

Figure 1: The associated digraph

Proposition 18 was first proved in Coppel [16]. Next, we use Lemma 17 to prove a result due to Stefan [17].

Proposition 19. *Suppose that f has a periodic orbit of odd period $n > 1$, but no periodic orbit of odd period strictly between 1 and n. If c is the midpoint of the orbit of odd period n, then the points of this orbit have the order*

$$f^{n-1}(c) < f^{n-3}(c) < \cdots < f^2(c) < c < f(c) < \cdots < f^{n-2}(c)$$

or the inverse order

$$f^{n-2}(c) < \ldots f(c) < c < f^2(c) < \cdots < f^{n-3}(c) < f^{n-1}(c).$$

In either case, the associated digraph is given by Figure 1, where
$J_1 =< c, f(c) >$ and $J_k =< f^{k-2}(c), f^k(c) >$ for $1 < k < n$.

Proof. The fundamental cycle decomposes into two smaller primitive cycles, one of which must have odd length, for the sum of the lengths is the odd number n. This length must be 1, since f has no orbit of odd period strictly between 1 and n, and by Lemma 17, f must have an orbit of period equal to this length. Thus, the fundamental cycle is given by

$$J_1 \to J_1 \to J_2 \to J_3 \to \cdots \to J_{n-1} \to J_1,$$

where $J_i \neq J_1$ for $1 < i < n$. If we had $J_i = J_k$, where $1 < i < k < n$, then by omitting the intermediate vertices we would obtain a smaller primitive cycle. Moreover, by excluding the loop at J_1 if necessary, we can arrange that its length is odd. This contradicts the hypothesis, since Lemma 17 again lets us deduce an orbit of odd period strictly between 1 and n. So, we conclude that $J_1, \ldots, J_{n-1}$ are all distinct, and thus a permutation of $I_1, \ldots, I_{n-1}$. Similarly, we cannot have $J_i \to J_k$ if $k > i + 1$ or if $k = 1$ and $i \neq 1, n-1$.

Suppose $J_1 = I_h = [a, b]$. Since J_1 is directed only to J_1 and J_2, the interval J_2 is adjacent to J_1 on the real line, and f maps one endpoint of J_1 to an endpoint of J_1, and the other endpoint of J_1 to an endpoint of J_2. Since the endpoints are not fixed points, there are only two possibilities: either

$$x_h = a, \ x_{h+1} = f(a), \ x_{h-1} = f^2(a),$$

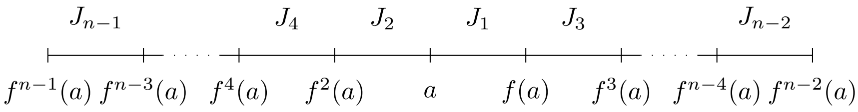
or

$$x_{h+1} = b, \ x_h = f(b), \ x_{h+2} = f^2(b).$$

We consider only the first case, the argument being similar in the second case.

For $n = 3$ the result follows immediately, so assume $n > 3$. If $f^3(a) < f^2(a)$ then $J_2 \to J_1$ which is forbidden. Hence $f^3(a) > f^2(a)$. Since J_2 is not directed to J_k for $k > 3$ it follows that $J_3 = [f(a), f^3(a)]$ is adjacent to J_1 on the right. If $f^4(a) > f^3(a)$ then $J_3 \to J_1$, which is forbidden.

Hence $f^4(a) < f^2(a)$ and, since J_3 is not directed to J_k for $k > 4$, $J_4 = [f^4(a), f^2(a)]$ is adjacent to J_2 on the left. Proceeding in this way we see that the order of the intervals J_i on the real line is given by



Since the endpoints of J_{n-1} are mapped into a and $f^{n-2}(a)$ we have that $J_{n-1} \to J_k$ if and only if k is odd. We found all the arcs in the digraph. $\square$

An orbit of odd period $n > 1$ with either one of the two configurations described in Proposition 19 is called a *Stefan orbit*, and the associated digraph will be called a *Stefan digraph*. The next result follows immediately using Lemma 17.

Proposition 20. *If f has a periodic orbit of odd period $n > 1$, then it has periodic points of arbitrary even period, and periodic points of arbitrary odd period $> n$.*

Proof. We may assume n is minimal, so the associated digraph is a Stefan digraph as in Proposition 19. If $m < n$ is even, then

$$J_{n-1} \to J_{n-m} \to J_{n-m+1} \to \cdots \to J_{n-1}$$

is a primitive cycle of length m. If $m > n$ is even or odd, then

$$J_1 \to J_2 \to \cdots \to J_{n-1} \to J_1 \to J_1 \to \cdots \to J_1$$

is a primitive cycle of length m. In either case, Lemma 17 lets us deduce the existence of a periodic orbit of period m. $\square$

Lemma 21. *If c is a periodic point of f with period n, then for any positive integer h, c is a periodic point of f^h with period $\frac{n}{(h,n)}$ where (h,n) denotes the greatest common divisor of h and n.*

Conversely, if c is a periodic point of f^h with period m, then c is a periodic point of f with period $\frac{mh}{d}$ where d divides h and is relatively prime to m.

Proof. Let h be an arbitrary positive integer. Suppose c has period n for f and let $m = \frac{n}{(h,n)}$. We have that

$$f^{mh}(c) = f^{\frac{nh}{(h,n)}}(c) = c,$$

since $\frac{nh}{(h,n)}$ is a multiple of n.

On the other hand, if $f^{kh}(c) = c$, then n must be a factor of kh, say $kh = dn$ for some integer d. This implies that m is a factor of k. Indeed

$$k = \frac{dn}{h} = \frac{n}{(h,n)} \frac{d(h,n)}{h} = m\frac{(dh,dn)}{h} = m\frac{(dh,kh)}{h} = m(d,k).$$

Hence c is a periodic point of f^h with period m, and the first assertion is proved.

Suppose now that c has period m for f^h. Then c has period n for f where n is a factor of mh, say $mh = nd$ for some integer d. From the first assertion of the lemma it follows that

$$m = \frac{n}{(h,n)} = \frac{nd}{h} \implies h = d(h,n) = de,$$

where e is the greatest common divisor of h and n. Then

$$(de, me) = (h, m(h,n) = (h,n) = e \implies (d,m) = 1.$$

$\square$

We are now able to prove Sharkovskii's theorem.

Proof of Sharkovskii's theorem. We give the proof initially for $f : I \to I$. Write $n = 2^d q$, where q is odd. First assume that $q = 1$ and $m = 2^e$ where $0 \leq e < d$. By Proposition 17 we may assume $e > 0$. Consider the map $g = f^{\frac{m}{2}}$ and apply the first assertion of Lemma 21 with $h = \frac{m}{2} = 2^{e-1}$ and $n = 2^d$. It follows that g has a periodic point c of period

$$\frac{n}{(h,n)} = \frac{2^d}{(2^{e-1}, 2^d)} = 2^{d-e+1},$$

and hence also a periodic point of period 2 by Proposition 18. Now we apply the second part of Lemma 21 to deduce that a periodic point of $g = f^{2^{e-1}}$ with period 2, is a periodic point of f with period

$$\frac{2 \cdot 2^{e-1}}{d} = \frac{2^e}{d},$$

where d divides 2^{e-1} *and* is relatively prime to 2. Hence $d = 1$, and f has a periodic point with period $m = 2^e$.

Now let $q > 1$. The remaining cases are $m = 2^d r$ where either (i) r is even, or (ii) r is odd and $r > q$. Consider now the map $g = f^{2^d}$. By the first part of Lemma 21 with $h = 2^d$ and $n = 2^d$, it has a periodic point of period

$$\frac{n}{(h,n)} = \frac{2^d q}{(2^d, 2^d)} = \frac{2^d q}{2^d} = q.$$

Hence by Proposition 20, $g = f^{2^d}$ has a periodic point of period r. Then by applying the second part if Lemma 21 once again with $h = 2^d$ and $m = r$, we find that this point is a periodic point of f with period

$$\frac{mh}{\bar{d}} = \frac{2^d r}{\bar{d}}$$

where $\bar{d}$ divides 2^d and is relatively prime to r. In case (i), $\bar{d} = 1$ and f therefore has a periodic point with period $2^d r$ as required. In case (ii), $\bar{d}$ is some power of 2, so f has a periodic point with period $2^e r$ for some $e \leq d$. If $e = d$ we are done. If $e < d$ we can replace n by $2^e r$. Since $m = 2^e(2^{d-e} r)$ it then follows from case (i) that f also has a periodic point of period m.

Finally, we give the proof for $f : I \to \mathbb{R}$. Let x_1 and x_n denote the least and greatest points of a periodic orbit of f with period n. Then $K = [x_1, x_n] \cup f([x_1, x_n])$ is a compact interval. The map $g : K \to K$ defined by setting

$$g(x) = \begin{cases} f(x_1) & \text{if } x < x_1 \\ f(x) & \text{if } x \in [x_1, x_n] \\ f(x_n) & \text{if } x > x_n, \end{cases}$$

is then continuous. Since g has a periodic orbit of period n, g also has ha periodic point of period m by what we have already proved. Since this orbit of period m is contained in the interval K, it is also a periodic orbit of f. This concludes the proof. $\square$

6 Dynamical analysis of a discrete time SIR model with logistic growth

The following discrete time SIR epidemic model is the one of interest;

$$\begin{aligned} S_{n+1} &= rS_n(1-S_n) - \frac{\beta S_n I_n}{1+aS_n} \\ I_{n+1} &= (1-\mu-\gamma)I_n + \frac{\beta S_n I_n}{1+aS_n} \\ R_{n+1} &= \gamma I_n + (1-\lambda)R_n \end{aligned} \tag{6.1}$$

where r is the natural growth rate of the population; individuals are born susceptible and there is no inherited immunity. The force of infection is $\frac{\beta S_n I_n}{1+aS_n}$, and a measures the inhibitory effect, perhaps for example due to vaccination. We assume $a \neq 0$. Further parameters are γ, the recovery rate of the infected individuals, μ and λ that are death rate of infected and removed respectively. Hence clearly $\mu, \gamma, \lambda < 1$. The growth of the susceptible population is thus assumed to be *logistic* which essentially means that the population grows rapidly when it is small, and more slowly as it approaches some maximum value which in this case is 1.

We also have the obvious biological restriction that $S_0 + I_0 + R_0 = 1$, that is the sum of all the groups is equal to the total population when we start. Due to the logistic growth term, however, this is *not* invariant.

Since R_n does not appear in the other two equations, it can be ignored on analysis of the system since it will not affect the system dynamics. Hence our main concern is the reduced model

$$\begin{aligned} S_{n+1} &= rS_n(1-S_n) - \frac{\beta S_n I_n}{1+aS_n} \\ I_{n+1} &= (1-K)I_n + \frac{\beta S_n I_n}{1+aS_n} \end{aligned} \tag{6.2}$$

where we have put $K = \mu + \gamma$. In [3] the authors present some analysis and numerical simulations, indicating local stability of fixed points and bifurcation to periodic doubling but the analysis is short of rigorous, and far from complete. The rest of this paper is aimed at using the theory so far presented to give a more complete analysis of the dynamics of the system (6.2).

We shall assume throughout that the initial state (S_0, I_0) does not lie on the curve $rS(1-S) - \frac{\beta SI}{1+aS} = 0$, since in this case $S_n = 0$ for $n = 1, 2, \ldots$ and I converges to 0 by the contraction principle (Theorem 2). Just note that if $S = 0$, then $I_{n+1} = (1-K)I_n$, where $|1-K| < 1$.

6.1 Fixed points

Our first order of business is finding fixed points to the system. Fixed points are solutions to the system of equations

$$\begin{aligned} S = f(S, I) &= rS(1-S) - \frac{\beta SI}{1+aS} \\ I = g(S, I) &= (1-K)I + \frac{\beta SI}{1+aS}. \end{aligned}$$

We write such fixed points as (S^*, I^*).

Trivial equilibrium
The origin is obviously a trivial fixed point, and we denote it by $O = (0, 0)$.

Disease free equilibrium
Another, more interesting fixed point is the so-called disease-free equilibrium that is found by fixing $I^* = 0$, since then $g(S, 0) = 0$. Next, we solve

$$S^* = f(S^*, 0) = rS^*(1 - S^*) \iff 1 = r - rS^* \iff S^* = \frac{r-1}{r},$$

so $E_0 = (\frac{r-1}{r}, 0)$ is the disease-free equilibrium.

Endemic equilibrium
Next assume $I^* \neq 0$. Then the second equation becomes

$$I = (1 - K)I + \frac{\beta SI}{1 + aS} \iff 1 = 1 - K + \frac{\beta S}{1 + aS},$$

which simplifies to

$$K = \frac{\beta S}{1 + aS} \iff S = \frac{K}{\beta - aK} = \frac{K}{A},$$

where we put $A := \beta - aK$. So, we have $S^* = \frac{K}{A}$. Substituting this into the first equation we get

$$\begin{aligned} \frac{K}{A} &= r\frac{K}{\beta - aK}\left(1 - \frac{K}{\beta - aK}\right) - \frac{\beta\frac{K}{\beta - aK}I}{1 + a\frac{K}{\beta - aK}} \\ &= r\frac{K}{\beta - aK}\left(\frac{\beta - aK - K}{\beta - aK}\right) - \frac{\beta KI}{\beta - aK + aK} \\ &= \frac{rK(A - K)}{A^2} - KI. \end{aligned}$$

Solving for I we get

$$I = \frac{rA - rK}{A^2} - \frac{1}{A} = \frac{rA - rK - A}{A^2} = \frac{r-1}{A} - \frac{rK}{A^2},$$

and resubstituting we find another fixed point, $E_1 = (\frac{K}{\beta - aK}, \frac{r-1}{\beta - aK} - \frac{rK}{(\beta - aK)^2})$, the so-called endemic equilibrium.

6.2 Local stability

To study the local stability of the fixed points we shall use Theorem 1 to analyse the eigenvalues of the Jacobian matrix evaluated at the fixed points.

For our dynamical system (6.2), the Jacobian matrix is

$$J(S, I) = \begin{pmatrix} \frac{\partial f}{\partial S} & \frac{\partial f}{\partial I} \\ \frac{\partial g}{\partial S} & \frac{\partial g}{\partial I} \end{pmatrix} = \begin{pmatrix} r - 2rS - \left(\frac{\beta I}{1+aS} - \frac{a\beta SI}{(1+aS)^2}\right) & -\frac{\beta S}{1+aS} \\ \frac{\beta I}{1+aS} - \frac{a\beta SI}{(1+aS)^2} & 1 - K + \frac{\beta S}{1+aS} \end{pmatrix}$$

which simplifies to

$$J(S, I) = \begin{pmatrix} r - 2rS - \frac{\beta I}{(1+aS)^2} & -\frac{\beta S}{1+aS} \\ \frac{\beta I}{(1+aS)^2} & 1 - K + \frac{\beta S}{1+aS} \end{pmatrix}.$$

Stability of the trivial fixed point

We evaluate

$$J(E_0) = J(0,0) = \begin{pmatrix} r & 0 \\ 0 & 1-K \end{pmatrix}.$$

Since this is a triangular matrix we can just read the eigenvalues straight of the diagonal. Hence

$$\begin{aligned} \lambda_1 &= r \\ \lambda_2 &= 1-K, \end{aligned}$$

so the origin is stable if $r < 1$ and $0 < K < 2$, and due to biological restrictions, the second inequality is automatically fulfilled.

6.2.1 Stability of E_0

At the disease-free equilibrium, the Jacobian matrix is

$$J(E_0) = J(\frac{r-1}{r}, 0) = \begin{pmatrix} 2-r & -\frac{\beta(r-1)}{r+a(r-1)} \\ 0 & 1-K+\frac{\beta(r-1)}{r+a(r-1)} \end{pmatrix}.$$

Again we are lucky to get a triangular matrix, so the eigenvalues are

$$\begin{aligned} \lambda_1 &= 2-r \\ \lambda_2 &= 1-K+\frac{\beta(r-1)}{r+a(r-1)}. \end{aligned}$$

To find out when E_0 is stable, we must solve the system of inequalities

$$\begin{cases} |2-r| < 1 \\ |1-K+\frac{\beta(r-1)}{r+a(r-1)}| < 1 \end{cases} \iff \begin{cases} 1 < r < 3 \\ -2 < \frac{\beta(r-1)}{r+a(r-1)} - K < 0. \end{cases}$$

Then we have

$$\begin{cases} 1 < r < 3 \\ K-2 < \frac{\beta(r-1)}{r+a(r-1)} < K \end{cases} \iff \begin{cases} 1 < r < 3 \\ \frac{(K-2)(r+a(r-1))}{r-1} < \beta < \frac{K(r+a(r-1))}{r-1}. \end{cases}$$

Now, since β is the coefficient for the force of infection, it must be positive. It is clear, since $K = \mu + \gamma < 2$ that the lower bound for β is negative. So, to summarize, if $1 < r < 3$ and $0 < \beta < \beta_0$, where

$$\beta_0 = \frac{K(r+a(r-1))}{r-1},$$

then E_0 is locally asymptotically stable.

6.2.2 Stability of E_1

Now it gets a little more complicated. First, let us make sure that E_1 exists in a biologically sensible way, that is both S and I must be positive. For this to be true we must have, recalling that $K > 0$

$$\begin{cases} \frac{K}{\beta-aK} > 0 \\ \frac{r-1}{\beta-aK} - \frac{rK}{(\beta-aK)^2} > 0 \end{cases} \iff \begin{cases} \beta - aK > 0 \\ \beta - aK > \frac{rK}{r-1}. \end{cases}$$

Hence

$$\begin{cases} r > 1 \\ \beta > \beta_0, \end{cases} \tag{6.3}$$

guarantees that E_1 is positive.

The Jacobian matrix evaluated at E_1 is

$$J(E_1) = \begin{pmatrix} \frac{2Kr}{aK-\beta} + \frac{K(a(r-1)+r)}{\beta} + 1 & -K \\ r + \frac{K(a-(a+1)r)}{\beta} - 1 & 1 \end{pmatrix},$$

whose characteristic polynomial is

$$p(z) = z^2 - p_1 z + p_0$$

where $p_1 = -\text{trace}(J(E_1))$ and $p_0 = \det(J(E_1))$. We found sufficient conditions on the coefficients for stability in Section 3.5, and we have

$$\begin{aligned} p_1 &= -\text{trace}(J(E_1)) = -\frac{2Kr}{aK-\beta} - \frac{K(a(r-1)+r)}{\beta} - 2 \\ p_0 &= \det(E_1)) = 1 + K\left(\frac{2r}{aK-\beta} - \frac{(K-1)(a(r-1)+r)}{\beta} + r - 1\right). \end{aligned}$$

Then remembering the positivity constraint (6.3) we require

$$\begin{cases} |p_1| & < 1 + p_0 \\ |p_0| & < 1 \\ r & > 1 \\ \beta & > \frac{K(r+a(r-1))}{r-1}. \end{cases}$$

Using Mathematica, we get from this that

$$\begin{cases} 1 < r \leq 3 \\ \beta_0 < \beta < \beta_2 \end{cases} \quad \text{or} \quad \begin{cases} 3 < r < r_{max} \\ \beta_1 < \beta < \beta_2 \end{cases}$$

where

$$\begin{aligned} \beta_0 =& \frac{K(r+a(r-1))}{r-1} \\ \beta_1 =& \frac{1}{2}\left(\frac{K(2a(3+K(r-1)-r)+(K+1)r)}{4+K(r-1)} + \sqrt{\frac{K^2((K+2)^2r^2+4a^2(r+1)^2+4ar(14-5K-2r+3Kr))}{(4+K(r-1))^2}}\right) \\ \beta_2 =& \frac{1}{2}\left(a(2K-1) + \frac{r(K+1)}{r-1} + \sqrt{a^2 + \frac{2ar(3K-1)}{r-1} + \frac{r^2(K+1)}{(r-1)^2}}\right) \\ r_{max} =& \frac{1}{2}\sqrt{\frac{16a^2+88aK-32a+25K^2+40K+16}{K^2}} + \frac{4a+5K+4}{2K}. \end{aligned}$$

6.2.3 Conclusions local stability

We have found that if $r < 1$ the origin is locally asymptotically stable and if

$$\begin{cases} 1 < r < 3 \\ 0 < \beta < \beta_0 \end{cases}$$

then the disease-free equilibrium E_0 is locally asymptotically stable. Finally, if

$$\begin{cases} 1 < r \leq 3 \\ \beta_0 < \beta < \beta_2 \end{cases} \quad \text{or} \quad \begin{cases} 3 < r < r_{max} \\ \beta_1 < \beta < \beta_2 \end{cases}$$

then the endemic equilibrium E_1 is locally asymptotically stable.

6.3 Basic reproduction number R_0

We have found that the infection dies out and we reach a disease-free equilibrium if $1 < r < 3$ and $0 < \beta < \beta_0$.

Now

$$\beta < \beta_0 = \frac{K(r + a(r-1))}{r-1} \iff \frac{\beta(r-1)}{K(r + a(r-1))} < 1,$$

so we might expect that $R_0 = \frac{\beta(r-1)}{K(r+a(r-1))}$ is the basic reproduction number, that is I_n goes to zero if $\beta < \beta_0$.

To investigate this, note that by (6.2)

$$I_{n+1} = \left(1 - K + \frac{\beta S_n}{1 + aS_n}\right) I_n,$$

so we should examine the map

$$h(x) = 1 - K + \frac{\beta x}{1 + ax} \tag{6.4}$$

for $\beta < \beta_0$.

Proposition 22. *If $\beta < \beta_0, 1 < r$ and $0 < x \leq \frac{r-1}{r}$, then $|h(x)| < 1$.*

Proof. Since $0 < K < 2$, and the term $\frac{\beta x}{1+ax}$ is increasing, it is clear that $-1 < h(x)$ for all $x > 0$. To see that $h(x) < 1$, note that for fixed x, $h(x)$ increases as β increases. Hence

$$h(x) < 1 - K + \frac{\beta_0 x}{1 + ax} = 1 - K + \frac{K(r + a(r-1))}{r-1} \cdot \frac{x}{1 + ax} = h_{\beta_0}(x).$$

Now, $h_{\beta_0}(x)$ is monotonically increasing for $x > 0$ so it is clear that

$$h(x) < h_{\beta_0}(x) \leq h_{\beta_0}(\frac{r-1}{r}) = 1 - K + \frac{K(r + a(r-1))}{r-1} \cdot \frac{r-1}{r + a(r-1)} = 1,$$

which proves the proposition. $\square$

Notice that the map

$$y \mapsto h(x)y, \ 0 < x \leq \frac{r-1}{r} \tag{6.5}$$

is continuous and by Proposition 22 it satisfies the condition

$$|h(x)y_1 - h(x)y_2| = |h(x)(y_1 - y_2)| = |h(x)||y_1 - y_2| < |y_1 - y_2|$$

where $|h(x)| < 1$. With the Euclidean distance on the complete metric space $\mathbb{R}$, the map (6.5) is a contraction map. Further, the only fixed point of the map $y \mapsto h(x)y$ is $y = 0$. Hence by Theorem 2, $y \to 0$ as $k \to \infty$.

Now, if we could show that if $\beta < \beta_0$, there exists some k for which $n > k$ implies that $0 < S_n \leq \frac{r-1}{r}$, we would have shown that I goes to zero as suspected. Unfortunately, this does not seem to be true in general. Despite much effort, it could not be shown that $I \to 0$ if $\beta < \beta_0$ even tough extensive numerical simulations suggests that this is the case. Hence, we cannot prove that $R_0 = \frac{\beta(r-1)}{K(r+a(r-1))}$.

6.4 Second iterate

Although we cannot prove that $I \to 0$ if $\beta < \beta_0$, we can show that there is a stable 2-cycle with $I = 0$ for $r > 3$. This is further evidence, but of course not a proof for our hypothesis.

We study the *second iterate*

$$S_{n+1} = f^2(S_n, I_n)$$
$$I_{n+1} = g^2(S_n, I_n),$$

where

$$\begin{aligned} f^2(S,I) &= S(r - rS - \tfrac{\beta I}{1+aS})(r + r^2(S-1)S + \tfrac{r\beta SI}{1+aS} + \tfrac{\beta I(1-K(1+aS)+S(\beta+a))}{(1+aS)(S(S-1)ar-1)+\beta aSI}) \\ g^2(S,I), &= I(1-K+\tfrac{\beta S}{1+aS})(1-K+\tfrac{S\beta(r(S-1)(1+aS)+I\beta)}{(1+aS)(S(S-1)ar-1)+\beta aSI}). \end{aligned} \tag{6.6}$$

Assuming $I = 0$ we want to find the fixed point of this map. However, when $I = 0$, $f(S, 0)$ is just the well-known *logistic map*

$$x \mapsto rx(1-x) = f(x). \tag{6.7}$$

It is well known (see for example [5]) that the nontrivial fixed points of the second iterate of the logistic map are

$$x_{1,2} = \frac{1 + r \pm \sqrt{(r-3)(r+1)}}{2r}.$$

Of course, any fixed point of the first iterate is also a fixed point of the second, but we already know about them.

Hence the nontrivial disease free fixed points of the second iterate (6.4) are

$$(S, I) = (\frac{1 + r \pm \sqrt{(r-3)(r+1)}}{2r}, 0).$$

To analyse the stability of these points, that make up the 2-cycle in our original system, we need to compute the eigenvalues of the Jacobian matrix evaluated at the fixed points. Tedious but trivial calculations show that both fixed points yield upper triangular matrices, whose entries differ only at the off-diagonal element. Hence the eigenvalues of both points are equal, and we have

$$\begin{aligned} \mu_1 &= 4 - r(r-2) \\ \mu_2 &= \tfrac{(K-1)^2\left(a^2(r+1)+ar(r+1)+r^2\right)-\beta(K-1)(r+1)(2a+r)+\beta^2(r+1)}{a^2(r+1)+ar(r+1)+r^2}. \end{aligned} \tag{6.8}$$

As before, the fixed points are locally asymptotically stable if the eigenvalues lie inside the unit circle. Using Mathematica, we find that this is the case if $3 < r < 1 + \sqrt{6}$ and $0 < \beta < \beta_{max}$ where

$$\beta_{max} = \frac{1}{2}\left(\sqrt{4a^2 + 4ar + \tfrac{r^2((K-2)K(r-3)+r+1)}{r+1}} + 2a(K-1) + (K-1)r\right).$$

We would like to know whether this upper bound is less than β_0. First, it is easy to check that when $r = 3$, they intersect. Further

$$\frac{\mathrm{d}\beta_0}{\mathrm{d}r} = \frac{K(a+1)}{r-1} - \frac{K(r + a(r-1))}{(r-1)^2} = -\frac{K}{(r-1)^2} < 0$$

for all $r \neq 1$. So that β_0 is monotonically decreasing as a function of r. Now, we find using Mathematica that

$$\frac{\mathrm{d}\beta_{max}}{\mathrm{d}r} = \frac{1}{2}\left(\frac{2a(r+1)^2 + r\left(K^2(r^2-3) - 2K(r^2-3) + (r+1)^2\right)}{(r+1)^2\sqrt{4a^2+4ar+\frac{r^2((K-2)K(r-3)+r+1)}{r+1}}} + K - 1\right) > 0$$

for all positive r. Hence β_{max} is increasing as r increases. To summarize we have found that when $r = 3$, $\beta_0 = \beta_{max}$, after which β_0 decreases while β_{max} increases. Hence, we may draw the conclusion that $\beta_0 < \beta_{max}$ for all $r > 3$.

6.5 Bifurcation

We have found conditions on the parameters r and β for stability. Now we ask how the dynamics of the system (6.2) changes under variation of these parameters.

6.5.1 List of eigenvalues on bifurcation boundary

Here is the full list of eigenvalue types for each fixed point.

Trivial fixed point: $O = (0, 0)$

We found that the origin is stable if $r < 1$. When $r = 1$ we get eigenvalues $\lambda_1 = 1$ and $|\lambda_2| < 1$.

Disease free: $E_0 = (\frac{r-1}{r}, 0)$

The stability conditions were $1 < r < 3$ and $0 < \beta < \beta_0$. Recall however that the lower bound for β was derived under the biological constraint that β has to be non-negative. Hence 0 is not mathematically the lower bound for stability and can therefore be ignored here. The conditions, with this in mind, can be violated as follows:

1. $r = 1, 0 < \beta < \beta_0 \implies \lambda_1 = 1, |\lambda_2| < 1$
2. $r = 1, \beta = \beta_0 \implies \lambda_1 = 1, \lambda_2 = -1$
3. $r = 3, 0 < \beta < \beta_0 \implies \lambda_1 = 1, |\lambda_2| < 1$
4. $r = 3, \beta = \beta_0 \implies \lambda_1 = 1, \lambda_2 = -1$
5. $1 < r < 3, \beta = \beta_0 \implies \lambda_1 = 1, |\lambda_2| < 1$

Endemic: $E_1 = (\frac{K}{\beta - aK}, \frac{r-1}{\beta - aK} - \frac{rK}{(\beta - aK)^2})$

The stability conditions were $1 < r \leq 3$ and $\beta_0 < \beta < \beta_2$ or $3 < r < r_{max}$ and $\beta_1 < \beta < \beta_2$. Note that when $r = 3$, we get $\beta_0 = \beta_1$, and when $r = r_{max}$ we have $\beta_0 = \beta_2$. In fact, we can also have $\beta_0 = \beta_2$ but only when $r = 0$ or $r = \frac{a}{a+1} < 1$ so it has no effect here. The stability conditions can be violated as follows:

1. $1 < r < 3, \beta = \beta_0 \implies \lambda_1 = 1, |\lambda_2| < 1$
2. $1 < r < 3, \beta = \beta_2 \implies \lambda_{1,2} = e^{\pm i\theta_0}, 0 < \theta_0 < \pi$
3. $r = 3, \beta = \beta_0 \implies \lambda_1 = 1, \lambda_2 = -1$
4. $r = 3, \beta = \beta_2 \implies \lambda_{1,2} = e^{\pm i\theta_0}, 0 < \theta_0 < \pi$
5. $r = r_{max}, \beta = \beta_2 \implies \lambda_1 = -1, \lambda_2 = -1$
6. $3 < r < r_{max}, \beta = \beta_1 \implies \lambda_1 = -1, |\lambda_2| < 1$
7. $3 < r < r_{max}, \beta = \beta_2 \implies \lambda_{1,2} = e^{\pm i\theta_0}, 0 < \theta_0 < \pi$

6.5.2 Bifurcation types in codimension 1

We begin by considering the bifurcations that depend on just one parameter.

Bifurcations from O: Since $0 < K < 2$, we have that for $r > 1$, O is a saddle point. At $r = 1$ there is a fold bifurcation, and O loses stability to other fixed points.

Bifurcations from E_0: At $\beta = \beta_0$ for all $1 < r < 3$, or $r = 1$ for $\beta < \beta_0$ there is a fold bifurcation, and E_0 loses stability. There is a flip bifurcation at $r = 3$ for all $\beta < \beta_0$. We will show that it is generic and stable.

Bifurcations from E_1: For $1 < r < 3$ and $\beta = \beta_0$, there is a fold Bifurcation, and E_1 loses stability to E_0. When $\beta = \beta_1$ and $3 < r < r_{max}$, there is a flip. For $1 < r < r_{max}$ and $\beta = \beta_2$ there is a Neimark-Sacker bifurcation, except for some degenerate cases which we deal with later.

Now we turn to investigate the genericity conditions on some of these bifurcation points. This is somewhat technical, and include some rather long computations, some of which can be found in the appendices.

6.5.3 Stable flip bifurcation from E_0

At E_0, for $r = 3$ the Jacobian matrix is

$$A = J(E_0) = \begin{pmatrix} 2-r & -\frac{\beta(r-1)}{r+a(r-1)} \\ 0 & 1-K+\frac{\beta(r-1)}{r+a(r-1)} \end{pmatrix} = \begin{pmatrix} -1 & -\frac{2\beta}{3+2a} \\ 0 & 1-K+\frac{2\beta}{3+2a}. \end{pmatrix}$$

The eigenvalues of A are $\mu_1 = -1$ and $\mu_2 = 1 - K + \frac{2\beta}{3+2a}$. Now, $|\mu_2| < 1$ if and only if $0 < \beta < \frac{1}{2}(3K + 2aK) = \beta_0\big|_{r=3}$.

Following the procedure in Section 4.3 we compute an eigenvector q of A associated with $\mu_1 = -1$. We have

$$Aq = -q \iff (A + I_2)q = 0 \iff \begin{pmatrix} 0 & -\frac{2\beta}{3+2a} \\ 0 & 2-K+\frac{2\beta}{3+2a} \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = 0.$$

We may choose $q_1 = 1, q_2 = 0$ to get the eigenvector $q = (1\ 0)^T$. Next, we compute an adjoint eigenvector p, normalized with respect to q, so that $\langle p, q\rangle = 1$. Fortunately, we see that p must take the form $p = (1\ p_2)^T$. Then we can find p_2 by

$$A^T p = -p \iff (A^T + I_2)p = 0 \iff \begin{pmatrix} 0 & 0 \\ -\frac{2\beta}{3+2a} & 2-K+\frac{2\beta}{3+2a} \end{pmatrix} \begin{pmatrix} 1 \\ p_2 \end{pmatrix} = 0.$$

This implies that

$$p_2 = \frac{\frac{2\beta}{3+2a}}{2-K+\frac{2\beta}{3+2a}} = \frac{2\beta}{2\beta + (2a+3)(2-K)}.$$

Our goal is to compute

$$c = \frac{1}{6}\langle p, C(q,q,q)\rangle - \frac{1}{4}\langle p, B(q,(A-I_2)^{-1}B(q,q))\rangle,$$

which first requires the computation of $B(x, y)$ and $C(x, y, z)$. As this computation is quite tedious and of no immediate interest, we just move on to state that $c = 9 > 0$ which implies that the flip is stable. The interested reader is referred to appendix B for the details of the computation.

6.5.4 Generic investigation of flip from E_1

In a similar manner one can find eigenvectors and compute c for $3 < r < r_{max}$ and $\beta = \beta_1$. We denote by A the Jacobian matrix evaluated at E_1 when $\beta = \beta_1$. Then

$$A = \begin{pmatrix} a_{11} & -K \\ a_{21} & 1 \end{pmatrix}$$

where

$$a_{11} = 1+4(K(r-1)+4)(4a+Kr)\Big/\Bigg(4Kr-5K^2r+K^2r^2-(K(r-3)+8)(4a+Kr)-2(K(r-1)+4)\sqrt{\frac{K^2\left(4a^2(r+1)^2+4ar(3Kr-5K-2r+14)+(K+2)^2r^2\right)}{(K(r-1)+4)^2}}\Bigg),$$

and

$$a_{21} = \frac{2K(a-(a+1)r)}{\sqrt{\frac{K^2\left(4a^2(r+1)^2+4ar(3Kr-5K-2r+14)+(K+2)^2r^2\right)}{(K(r-1)+4)^2}}+\frac{K(2a(K(r-1)-r+3)+(K+2)r)}{K(r-1)+4}} + r - 1.$$

Our first order of business is to find an eigenvector of A associated with $\mu_1 = -1$. Hence, we solve the equation

$$(A + I_2)q = 0$$

where $q = (q_1\ q_2)^T$. This yield

$$\begin{cases} (a_{11}+1)q_1 - Kq_2 = 0 \\ a_{21}q_1 + 2q_2 = 0 \end{cases} \iff \begin{cases} q_1 = -\frac{2a_{11}}{2+a_{21}K} \\ q_2 = \frac{a_{11}a_{21}}{2+a_{21}K}. \end{cases}$$

For convenience we divide both q_1 and q_2 by q_1 to get the eigenvector

$$q = \begin{pmatrix} 1 \\ -\frac{a_{21}}{2} \end{pmatrix}.$$

Next, we determine the adjoint eigenvector $p = (p_1\ p_2)^T$:

$$(A^T + 1)\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} (a_{11}+1)p_1 + a_{21}p_2 \\ -Kp_1 + 2p_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Together with the constraint that $\langle p, q \rangle = 1$ this yields that

$$\begin{cases} (a_{11}+1)p_1 + a_{21}p_2 = 0 \\ Kp_1 = 2p_2 \\ p_1 - \frac{a_{21}}{2}p_2 = 1. \end{cases}$$

From the second and third equation we get that $p_1 = \frac{4}{4-Ka_{21}}$ and $p_2 = \frac{2K}{4-Ka_{21}}$, and one can check that this fulfils the first equation as well. This gives us the adjoint eigenvector

$$p = \frac{4}{4-Ka_{21}}\begin{pmatrix} 1 \\ \frac{K}{2} \end{pmatrix}.$$

Again, we wish to compute

$$c = \frac{1}{6}\langle p, C(q,q,q)\rangle - \frac{1}{4}\langle p, B(q, (A-I_2)^{-1}B(q,q))\rangle,$$

which first requires the computation of $B(x,y)$ and $C(x,y,z)$. We refer the interested reader to appendix B. Unfortunately, numerical simulations show that c can take on both positive and negative values depending on r.

6.5.5 Stable flip of the second iterate

We found that the second iterate has two nontrivial fixed points

$$(S, I) = (\tfrac{1+r\pm\sqrt{(r-3)(r+1)}}{2r}, 0),$$

that are stable for $3 < r < 1 + \sqrt{6}$. We also found that both fixed points yield the eigenvalues (6.8). At $r = 1 + \sqrt{6}$, we find that $\mu_1 = -1$, so there is a flip in both cases. We will now show that the flip is stable. We consider only the case with negative square root since the computations for the other one are almost exactly the same.

Again, we look for an eigenvector of the Jacobian matrix J_2 of the second iterate at $r = 1 + \sqrt{6}$ which is quite easy since then,

$$A = J_2(\tfrac{1+r\pm\sqrt{(r-3)(r+1)}}{2r}, 0) = \begin{pmatrix} -1 & a_{12} \\ 0 & a_{21} \end{pmatrix},$$

where, if we denote by $a_{12}^{\pm}$ the off-diagonal element in the case of positive and negative square roots respectively we have

$$\begin{aligned}
a_{12}^{-} &= \tfrac{\beta\left(\left(2\left(\sqrt{6}+2\right)a+\sqrt{2}+2\sqrt{3}+3\sqrt{6}+8\right)K+4\left(\sqrt{2}+\sqrt{3}\right)a-2\left(\left(\sqrt{6}+2\right)\beta-3\sqrt{2}-\sqrt{3}+\sqrt{6}+1\right)\right)}{2\left(a\left(\left(\sqrt{6}+2\right)a+3\sqrt{6}+8\right)+2\sqrt{6}+7\right)} \\
a_{12}^{+} &= -\tfrac{\beta\left(-2\left(\sqrt{6}+2\right)aK+4\left(\sqrt{2}+\sqrt{3}\right)a+2\left(\left(\sqrt{6}+2\right)\beta+3\sqrt{2}+\sqrt{3}+\sqrt{6}+1\right)+\left(\sqrt{2}+2\sqrt{3}-3\sqrt{6}-8\right)K\right)}{2\left(a\left(\left(\sqrt{6}+2\right)a+3\sqrt{6}+8\right)+2\sqrt{6}+7\right)} \\
a_{21} &= \tfrac{\beta\left(\left(\sqrt{6}+2\right)\beta-\left(2\left(\sqrt{6}+2\right)a+3\sqrt{6}+8\right)(K-1)\right)}{a\left(\left(\sqrt{6}+2\right)a+3\sqrt{6}+8\right)+2\sqrt{6}+7} + (K-1)^2.
\end{aligned}$$

We want to determine q so that

$$(A + I_2)q = \begin{pmatrix} 0 & a_{12} \\ 0 & 1 + a_{21} \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = 0.$$

Hence, we may take $q = (1\ 0)^T$. Since we require $\langle p, q \rangle = 1$, p must take the form $(1\ p_2)^T$. Then we can find p_2 by considering

$$(A^T + I_2)p = \begin{pmatrix} 0 & 0 \\ a_{12} & a_{21} \end{pmatrix} \begin{pmatrix} 1 \\ p_2 \end{pmatrix} = 0,$$

which tells us that

$$p_2 = -\frac{a_{12}}{1 + a_{21}}.$$

From here following the same procedure as before we can compute c. The computations are completely analogous to what has been shown in appendix B and we find in the case of the negative square root that

$$c = -10\left(\sqrt{2} - 2\right)\left(2\sqrt{6} + 7\right) \approx 69.7$$

and for the positive square root

$$c = 10\left(\sqrt{2} + 2\right)\left(2\sqrt{6} + 7\right) \approx 406.3.$$

Hence the flip is stable in both cases.

6.5.6 Nondegeneracy on $\beta = \beta_2$

In this section we will investigate the nondegeneracy conditions to see whether the Neimark-Sacker bifurcation is generic. First, for $\beta = \beta_2$, the Jacobian matrix is

$$A = \begin{pmatrix} a_{11} & -K \\ a_{21} & 1 \end{pmatrix}$$

where

$$\begin{aligned} a_{11} &= \frac{K\left(-r\sqrt{a^2+\frac{2a(3K-1)r}{r-1}+\frac{(K+1)^2r^2}{(r-1)^2}}+\sqrt{a^2+\frac{2a(3K-1)r}{r-1}+\frac{(K+1)^2r^2}{(r-1)^2}}+Kr+r\right)+a(Kr+K-2)}{2a(K-1)} \\ a_{21} &= \frac{2K(r+a(r-1))}{\sqrt{a^2+\frac{2a(3K-1)r}{r-1}+\frac{(K+1)^2r^2}{(r-1)^2}+a(2K-1)+\frac{(K+1)r}{r-1}}} + r - 1. \end{aligned}$$

The characteristic polynomial is

$$P_A(z) = z^2 - tr(A)z + det(A),$$

and using standard relations between coefficients and zeros of a degree two polynomial we get that

$$\begin{cases} \mu_1 + \mu_2 = a_{11} + 1 \\ \mu_1\mu_2 = det(A) = a_{11} + a_{21}K = 1. \end{cases} \tag{6.9}$$

We have used that the zeros sum to negative the coefficient of z, and that the product is equal to the constant term. It is a simple but tedious matter to check that $det(A) = 1$. Knowing that one eigenvalue lies on the unit circle, we immediately get that the other one must do so as well, for otherwise their product could not be 1. This also excludes the case $\mu_{1,2} = \pm 1$ so we must have complex conjugate eigenvalues

$$\mu_{1,2} = e^{\pm i\theta_0} = \sigma \pm i\omega.$$

From (6.9) it is clear that $\mu_1 + \mu_2 = 2\sigma = a_{11} + 1$, and specifically we get

$$\sigma = \frac{a_{11}+1}{2} = \frac{K\left(\sqrt{a^2+\frac{2a(3K-1)r}{r-1}+\frac{(K+1)^2r^2}{(r-1)^2}}+Kr+r-r\sqrt{a^2+\frac{2a(3K-1)r}{r-1}+\frac{(K+1)^2r^2}{(r-1)^2}}\right)}{4a(K-1)} + \frac{Kr+K-2}{4(K-1)} + \frac{1}{2}$$

The degenerate cases $e^{ik\theta_0} = 1$ for $k = 1, 2, 3$ or 4 correspond to $\sigma = 1, -1, -\frac{1}{2}, 0$, so we may determine for which values of r these nondegeneracy conditions are violated. We will solve the equations for r, with the constraint that $1 < r \leq r_{max}$.

Case 1: $\sigma = 1$. This corresponds to $\theta_0 = 0$, that is 1:1 resonance. There are however no solutions except $r = 0$. This means that there is no 1:1 resonance.

Case 2: $\sigma = -1$. Then $\theta_0 = \pi$, so this is 1:2 resonance. We find the solution $r = r_{max}$, which means that when $r = r_{max}, \beta = \beta_2 (= \beta_1)$, there is a 1:2 resonance.

Case 3: $\sigma = -\frac{1}{2}$. This is $\theta_0 = \frac{3\pi}{2}$, which means 1:3 resonance. We find a solution

$$\tilde{r} = \frac{3a + 4K + 3}{2K} + \frac{1}{2}\sqrt{\frac{9a^2 + 48aK - 18a + 16K^2 + 24K + 9}{K^2}}.$$

So, for $\beta = \beta_2, r = \tilde{r}$, there is a 1:3 resonance.

Case 4: $\sigma = 0$. Then $\theta_0 = \frac{\pi}{2}$, corresponding to 1:4 resonance. Here too, there is a solution

$$\bar{r} = \frac{2a + 3K + 2}{2K} + \frac{1}{2}\sqrt{\frac{4a^2 + 20aK - 8a + 9K^2 + 12K + 4}{K^2}},$$

which means that for $\beta = \beta_2, r = \bar{r}$ there is 1:4 resonance.

The expressions for $r_{max}, \bar{r}$ and $\tilde{r}$ are quite similar, and in fact one can write

$$\begin{aligned}\bar{r} &= R(2)\\ \tilde{r} &= R(3)\\ r_{max} &= R(4),\end{aligned}$$

where

$$R(x) = \frac{ax + K(x+1) + x}{2K} + \frac{1}{2}\sqrt{\frac{a^2x^2 + 2ax(K(3x-1) - x) + (K(x+1)+x)^2}{K^2}}, \tag{6.10}$$

which we define for $2 \leq x \leq 4$. In this interval, the derivative of R is

$$R'(x) = \frac{2K(a+K+1) + \frac{2\left(aK(6x-1)+(a-1)^2x+K^2(x+1)+2Kx+K\right)}{\sqrt{\frac{(ax-(Kx+K+x))^2+8aKx^2}{K^2}}}}{4K^2} > 0$$

for $2 \leq x \leq 4$, and in fact for all positive x, which is clear since every term is strictly positive for $x > 0$. So $R(x)$ is strictly monotonically increasing for $2 \leq x \leq 4$, which implies that we always have

$$\bar{r} < \tilde{r} < r_{max}.$$

We should also check that $d \neq 0$, where d is given by (4.21). This is quite involved, and in fact we are not able to solve it analytically. However, numerical experiments strongly suggest that $d < 0 forallparameters$. The computation of d can be found in appendix C.

Finally, we check that $\rho'(\beta_2) \neq 0$ where $\rho(\beta) = |\mu_{1,2}(\beta)|$. This is the genericity condition (C.1) given in section 4 in [4]. We have that (see appendix C)

$$\mu_{1,2} = \frac{a_{11} + 1 \pm i\sqrt{4(a_{11} + Ka_{21}) - (a_{11}+1)^2}}{2},$$

which means that

$$\rho(\beta) = |\mu_{1,2}(\beta)| = \frac{1}{2}\sqrt{(a_{11}+1)^2 + 4(a_{11} + Ka_{21}) - (a_{11}+1)^2} = \sqrt{a_{11} + Ka_{21}},$$

where a_{11} and a_{21} depends on β. Explicitly

$$\begin{aligned}a_{11} &= \frac{2Kr}{aK - \beta} + \frac{K(a(r-1)+r)}{\beta} + 1\\ a_{21} &= \frac{K(a-(a+1)r)}{\beta} + r - 1.\end{aligned}$$

Hence

$$\begin{aligned}\rho'(\beta) =& \frac{a'_{11}(\beta) + Ka'_{21}(\beta)}{2\sqrt{a_{11}(\beta) + Ka_{21}(\beta)}} = \\ & \frac{-\frac{K^2(a-(a+1)r)}{\beta^2} - \frac{K(a(r-1)+r)}{\beta^2} + \frac{2Kr}{(aK-\beta)^2}}{2\sqrt{\frac{2Kr}{aK-\beta} + K\left(\frac{K(a-(a+1)r)}{\beta} + r - 1\right) + \frac{K(a(r-1)+r)}{\beta} + 1}},\end{aligned}$$

which gives us

$$\rho'(\beta_2) = \frac{(r-1)^2(2a(r-1) + (K+1)r)}{2K\sqrt{2-r}r(a(r-1)+r)} \neq 0$$

for $r > 1$.

This investigation leads us in to codimension 2.

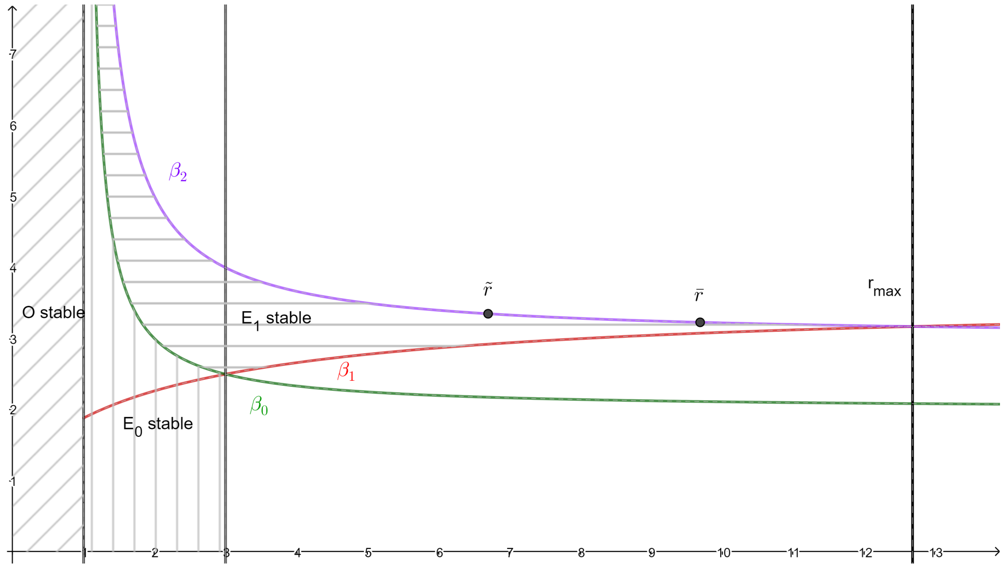


Figure 2: The curves β_0, β_1 and β_2 in the $r\beta$-plane, with $a = K = 1$ and the relevant regions for stability.

6.6 Bifurcation types in codimension 2

Bifurcations from E_0

In codimension 2 there are two fold-flips; one when $r = 1, \beta = \beta_0$ and the other one when $r = 3, \beta = \beta_0$.

Bifurcations from E_1

As noted above, we have 1:2, 1:3 and 1:4 resonance when $\beta = \beta_2$ and $r = r_{max}, \tilde{r}, \bar{r}$ respectively. Apart from that, there is a fold-flip bifurcation at $r = 3, \beta = \beta_0$.

6.7 Bifurcation diagrams and numerical simulations

We have found conditions for stability and bifurcation types on the parameters r and β. This analysis is summed up in Figure 2. In Figure 3 and 4 we see bifurcation diagrams as r varies for two different values of β; one less than β_0 and one greater than β_0. In Figure 5, the bifurcation parameter is β, and we have $r = 3.5$ In all the bifurcation diagrams we have $a = K = 1$.

When $\beta_0 < \beta < \beta_1$, our numerical experiments seem to show that for any initial values, the system diverges. As an example, in Figure 6 we set the initial values within a circle of radius 10^{-30} centred on the fixed point E_1. Initially one might expect it to converge, but after 75 iterations, S becomes negative, and then goes to $-\infty$.

When $\beta > \beta_0$ we can start within a circle of radius 10^{-2} centred on E_1, and after 5000 iterations the system seems to converge to E_1 as shown in Figure 7 with $r = 11.44$ and $\beta = 3.18$.

Perhaps the most interesting is the case when $r < 3$ and $\beta > \beta_2$. If, with $r = 2$ and $\beta = 6$ we start the system within a circle of radius 10^{-3} centred on E_1, the system first move from E_1 to settle into a limit cycle an shown in Figure 8.

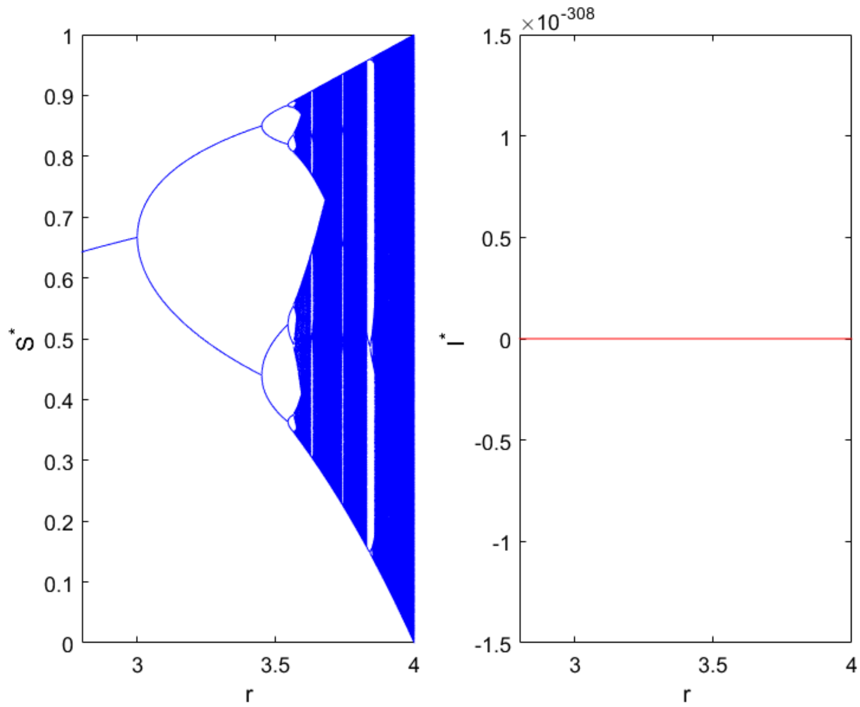


Figure 3: Bifurcation diagram with parameter r. We have $\beta = 1 < \beta_0$.

6.8 Finding a 3-cycle using the logistic map

As we know, the existence of a 3-cycle implies the existence of cycles of arbitrary length. Hence it is of interest whether the system has one.

Since the system (6.2) simply becomes the logistic map (6.7) when $I = 0$, we could hope that the logistic map has a 3-cycle, for then our system would also inherit this cycle when $I = 0$. Unfortunately, we have not been able to show that I tends to 0 for $\beta < \beta_0$, but numerical simulations strongly suggest that this is the case. Assuming that it is indeed true, we should expect to find a 3-cycle in our system exactly for the same values of r as the logistic map, since if our conjecture is true, we would always eventually reach values of I arbitrarily close to 0. At this stage, the dynamics should be strongly dominated by the logistic term. Hence, we expect the bifurcation diagram of (6.2) to look exactly the same as that of the logistic map (6.7).

The logistic map has a 3-cycle, and there are several ways to locate the onset of it. One can for instance approximate it numerically by graphical analysis, but it is also possible to find the relevant value of r analytically by solving a certain system of algebraic equation as described in [18]. In short, we have that the conditions for having a 3-cycle can be expressed in terms of the three points x, y, x in the cycle:

$$\begin{aligned} y =& rx(1-x) = f(x) \\ z =& ry(1-y) = f(y) = f^2(x) \\ x =& rz(1-z) = f(z) = f^2(y) = f^3(x). \end{aligned}$$

A fourth condition is given, since at the onset of the 3-cycle, the line $y = x$ must be tangent to the

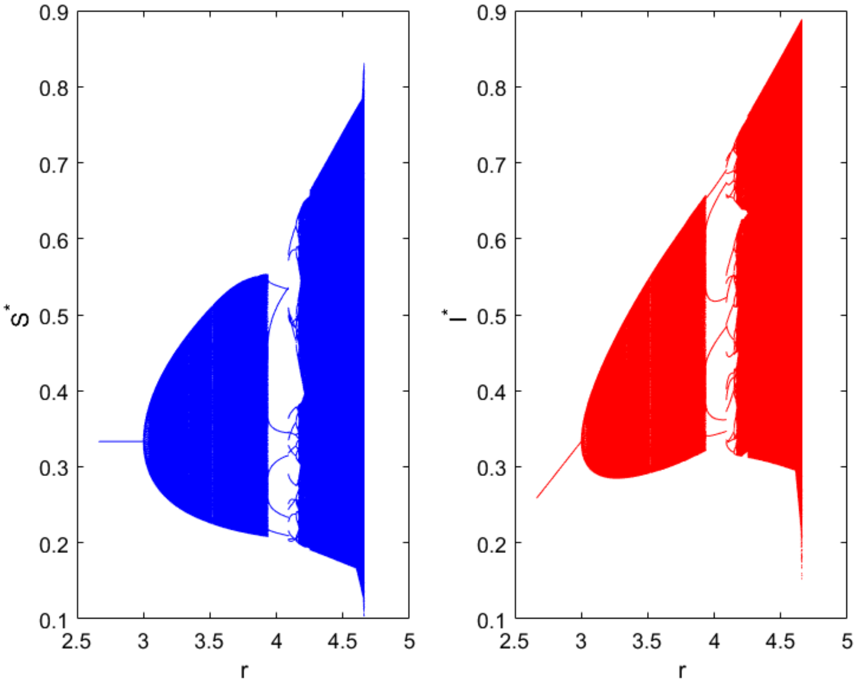


Figure 4: Bifurcation diagram with parameter r, and $\beta = 4 > \beta_0$.

graph of $f^3(x)$ as described in Figure 9. At x this yield

$$\begin{aligned}\frac{\mathrm{d}(f^3(x))}{\mathrm{d}x} &= \frac{\mathrm{d}(f^3(x))}{\mathrm{d}(f^2(x))} \cdot \frac{\mathrm{d}(f^2(x))}{\mathrm{d}(f(x))} \cdot \frac{\mathrm{d}(f(x))}{\mathrm{d}x} = \frac{\mathrm{d}f(z)}{\mathrm{d}z} \cdot \frac{\mathrm{d}f(y)}{\mathrm{d}y} \cdot \frac{\mathrm{d}f(x)}{\mathrm{d}x} \\ &= r^3(1-2z)(1-2y)(1-2x) = 1.\end{aligned}$$

Hence, we have four equations and four unknowns x, y, z, r, and in [18], the authors give a method for solving for r analytically. In fact, for any period n, the $n+1$ equations

$$\begin{cases} x_2 = rx_1(1-x_1) \\ x_3 = rx_2(1-x_2) \\ \vdots \\ x_n = rx_{n-1}(1-x_{n-1}) \\ x_1 = rx_n(1-x_n) \\ r^n \prod_{k=1}^{n}(1-2x_k) = 1 \end{cases} \tag{6.11}$$

can be solved for r to give the onset of the n-cycle. However, the complexity grows rapidly with n. When $r = 3$ the only non-negative solution for r is $r = 1 + 2\sqrt{2}$. Therefore we expect our system (6.2) to have a 3-cycle for this value of r, and indeed as shown in Figure 10, the orbit of (S, I) stabilizes after about 30 iterations to a 3-cycle.

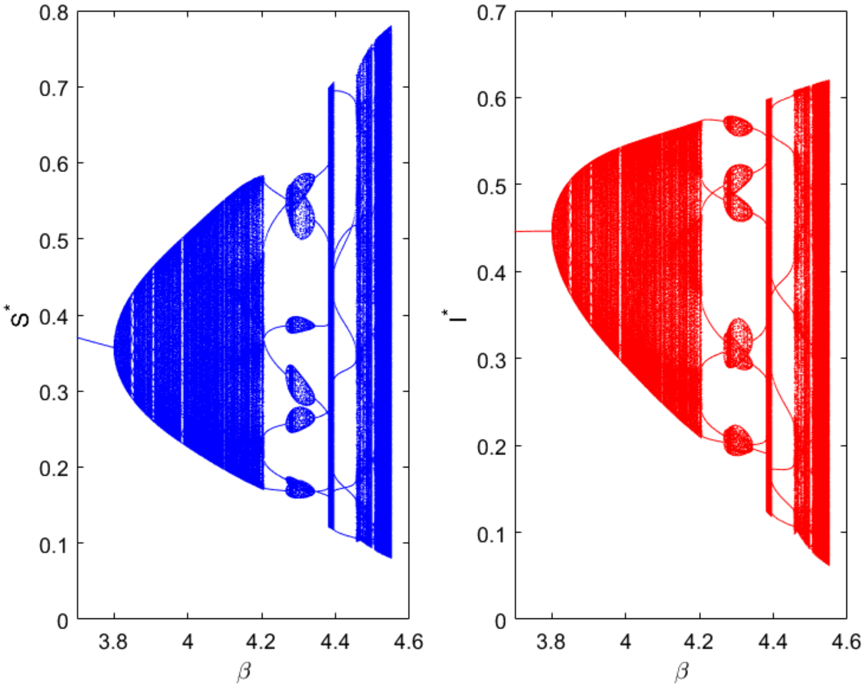


Figure 5: Bifurcation diagram with parameter β. Here $r = 3.5$.

6.8.1 Some n-cycles

Now that we know there is a 3-cycle, Sharkovskii's theorem tells us that there are cycles of arbitrary length. We can solve the system (6.11) for $n = 5$ numerically which yield three distinct solutions greater than 3, namely $r_1 = 3.73817, r_2 = 3.90557, r_3 = 3.99026$. We expect these values of r to yield 5-cycles in the bifurcation diagram when $\beta < \beta_0$, and indeed Figure 11 show all three of them.

For $n = 6$ we can solve the system of equations numerically, and find eight values of r that are greater than 3, namely

$$r = 3.21486, 3.63386, 3.83185, 3.83265, 3.85556, 3.93769, 3.97781, 3.99759,$$

and with patience one can numerically find all nine solutions greater than 3 when $n = 7$. For completeness these are

$$r = 3.71955, 3.78707, 3.88935, 3.92373, 3.95204, 3.96955, 3.98497, 3.99461, 3.99941.$$

For larger n it is no longer practical to solve the system of equations. We can however by simply looking at the bifurcation diagram find some more cycles. As an example, Figure 12 show a 7-cycle, a 10-cycle and an 18-cycle.

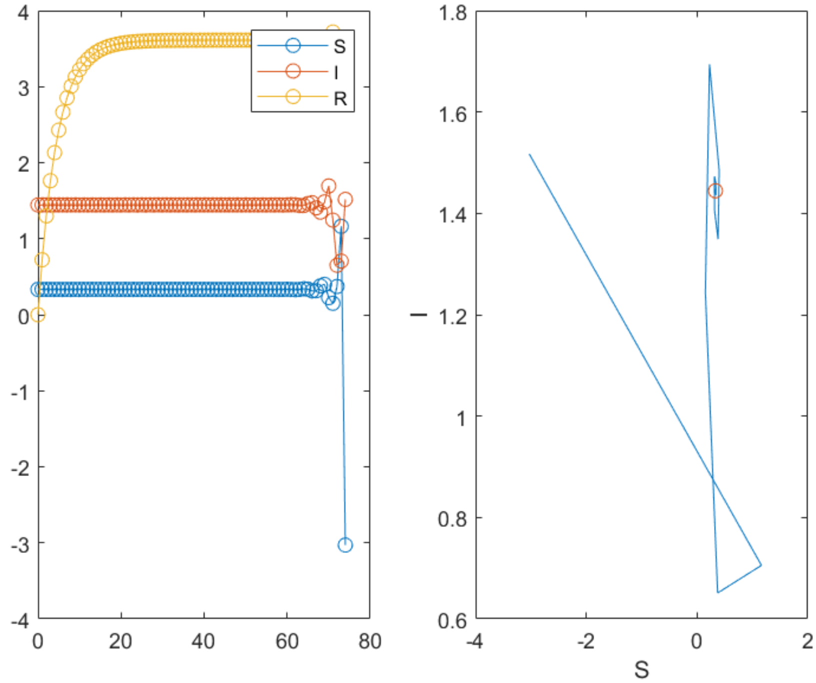


Figure 6: After 75 iterations the system diverges.

7 Conclusion

We have given a fairly complete analysis of the system in terms of stability and bifurcation, but have not been able to show that $I \to 0$ if $\beta < \beta_0$, which in turn means that we have not been able to prove that R_0 is as hypothesized. The existence of an invariant set that preserves non-negativity has also been sought in vain at this point and remains to be found. The system can diverge to $-\infty$, but numerical simulations suggests that it does not diverge to $+\infty$. Global stability is also an open question, as is the genericity of some of the co-dimension 2 bifurcations.

7.1 Generalized system

In our SIR-model, we have essentially assumed that susceptible individuals are immortal as long as they do not get infected. Assuming that we want to model a short outbreak of some illness, this is perhaps realistic, but for longer periods of time one would perhaps want to take into account the natural deaths that occur in any population. Another issue that appears with longer time periods is that recovered individuals typically are not immune forever, but eventually loose immunity.

To make the model more realistic, one can add some more parameters to account for natural death in the population, loss of immunity and perhaps a carrying capacity that limits the maximum size

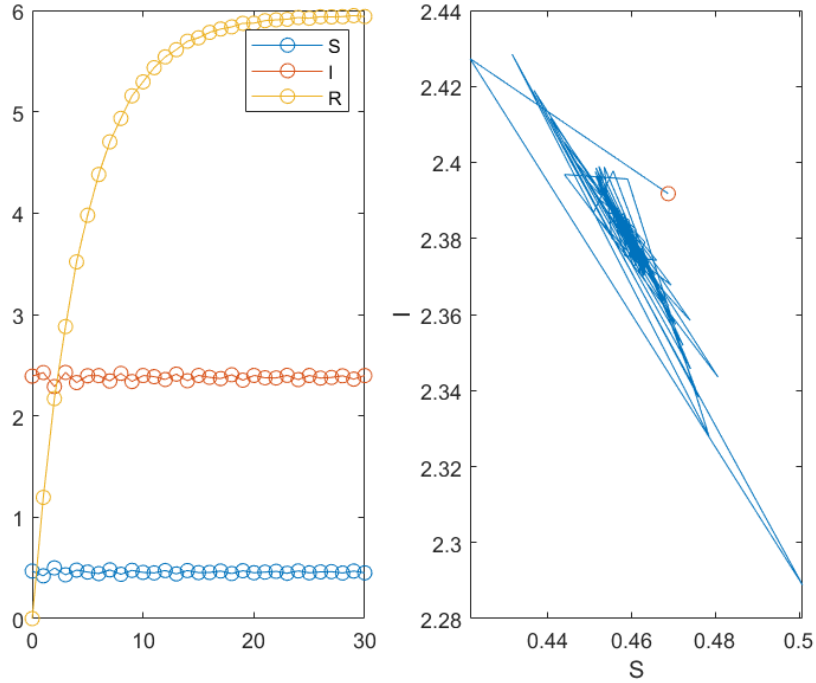


Figure 7: The system looks like it converges to E_1.

of the susceptible population. One possible generalization is the system

$$\begin{aligned} S_{n+1} &= rS_n(1 - \frac{S_n}{k}) - \frac{\beta S_n I_n}{1 + aS_n} - dS_n + \mu R_n \\ I_{n+1} &= \frac{\beta S_n I_n}{1 + aS_n} + (1 - d - \gamma)I_n \\ R_{n+1} &= \gamma I_n + (1 - d - \mu)R_n, \end{aligned} \tag{7.1}$$

where r is the intrinsic growth rate of S, that is the birth rate, k is the carrying capacity, β the force of infection, a is the inhibitory effect due to medication, isolation, vaccination et cetera. We have also introduced d, the natural death rate of the population as a whole, and now μ is the rate at which recovered individuals loses immunity, while γ is the recovery rate of infected individuals. Note that the disease does not affect the death rate of infected individuals.

Using the methods presented in this paper, it should be possible to analyse this system in detail although the analysis is more difficult since this system cannot be reduced to two dimensions. The system (7.1) also has three fixed points, one trivial, one disease free, and one endemic. The two non-trivial ones are

$$\begin{aligned} E_0 &= \left(\frac{k(-d + r - 1)}{r}, 0, 0\right) \\ E_1 &= \Bigg(- \tfrac{\gamma+d}{a(\gamma+d)-\beta}, \tfrac{(\gamma+d)(d+\mu)(ak(\gamma+d)(d-r+1)-d(\beta k+r)+\beta k(r-1)-\gamma r)}{dk(\gamma+d+\mu)(\beta-a(\gamma+d))^2}, \\ &\quad \tfrac{\gamma(\gamma+d)(ak(\gamma+d)(d-r+1)-d(\beta k+r)+\beta k(r-1)-\gamma r)}{dk(\gamma+d+\mu)(\beta-a(\gamma+d))^2}\Bigg). \end{aligned}$$

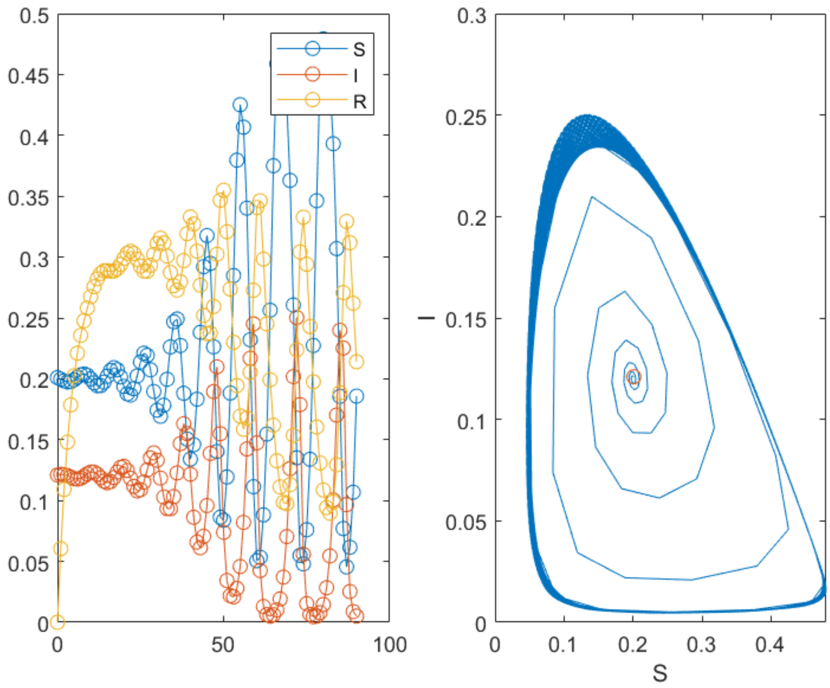
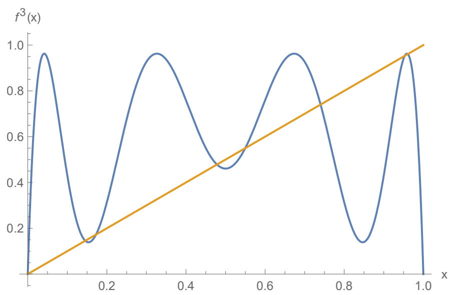
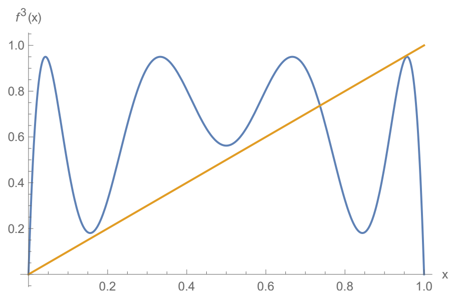


Figure 8: A stable limit cycle



(a) For $r = 3.85$ there are six nontrivial intersections.



(b) At $r = 3.8$, the 3-cycle has disappeared. Only trivial intersections.

Figure 9: The graph of $f^3(x)$ for two different values of r. The six nontrivial intersections in (a) correspond to two 3-cycles; one stable and one unstable. Somewhere between $r = 3.85$ and $r = 3.8$, the graph of $f^3(x)$ must have become tangent to the line $y = x$. This value corresponds to the onset of the 3-cycle.

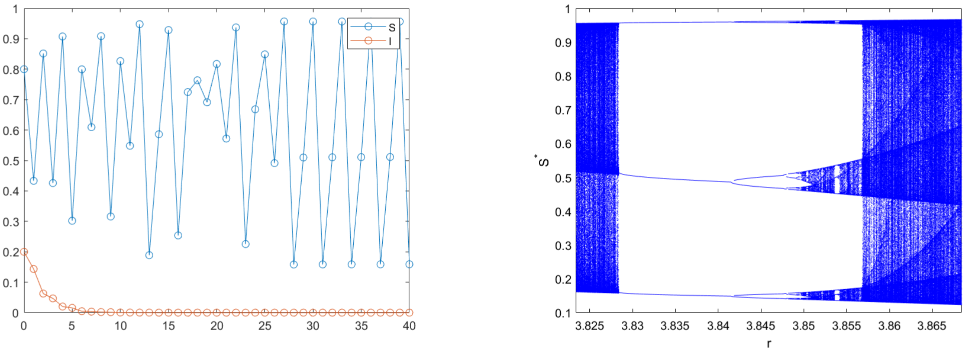


Figure 10: Time series for $r = 1 + 2\sqrt{2}, \beta = 2.09, \mu = 0.19, \gamma = 0.99, a = 1.08$, and $S_0 = 0.8, I_0 = 0.2$, and part of the bifurcation diagram for these parameters as r varies in a neighbourhood of the critical value.

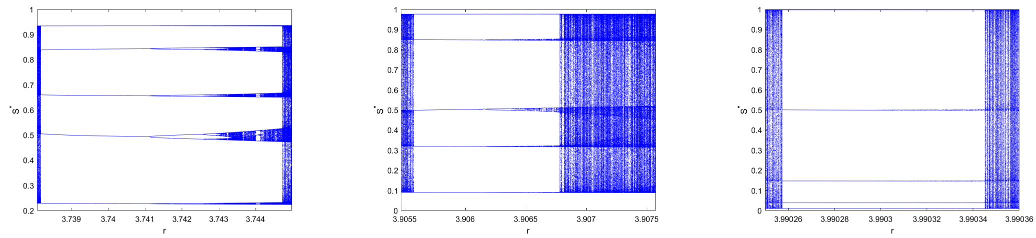


Figure 11: All the 5-cycles found in the bifurcation diagram as r varies, and with $\beta < \beta_0$.

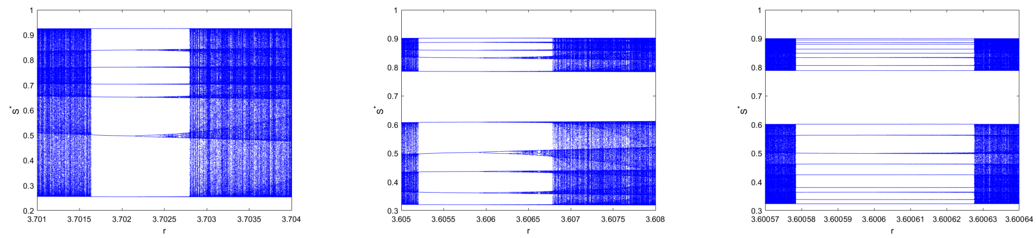


Figure 12: Cycles of length 7, 10 and 18, when $\beta < \beta_0$.

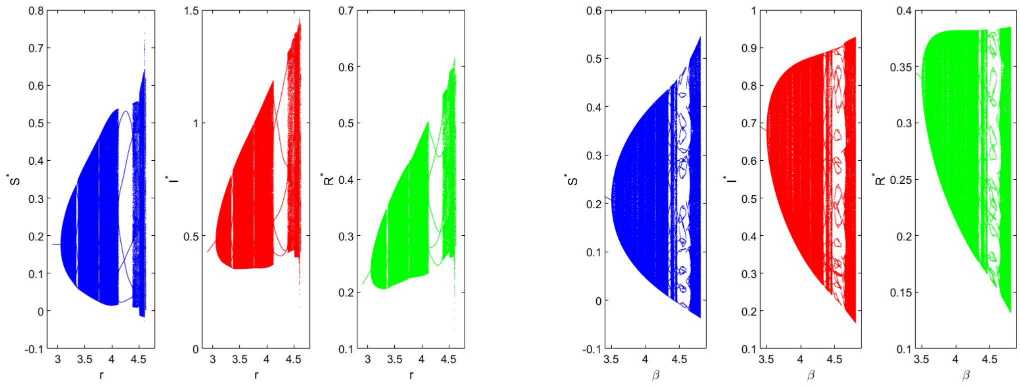


Figure 13: Bifurcation diagrams for system (7.1) with $a = k = 1, \mu = \gamma = d = 0.3$. In the left figure $\beta = 4$, and in the right figure $r = 3.5$.

To possibly arouse interest in further analysis of the system, we present some bifurcation diagrams in Figure 13 to hint at the dynamics.

Appendices

A End of the proof of Routh's theorem; the singular case

For completeness we give the rest of the proof of Routh's theorem, so that we can deal with cases other than the singular case.

A.1 Adjustment for zeros on the imaginary axis

In deriving Routh's theorem, we made heavy use of the formula (3.10), that was deduced under the assumption that $p(z)$ has no zeros on the imaginary axis. In the following discussion we shall have to generalize (3.10) to take into account the situation when $p(z)$ has k zeros in the open right half plane and s zeros on the imaginary axis. In this case, (3.10) is replaced by

$$I_{-\infty}^{+\infty} \frac{b_0\omega^{n-1} - b_1\omega^{n-3} + b_2\omega^{n-5} - \ldots}{a_0\omega^n - a_1\omega^{n-2} + a_2\omega^{n-4} - \ldots} = n - 2k - s, \tag{A.1}$$

for

$$p(z) = d(z)p^*(z),$$

where $d(z)$ is a real monic polynomial with s zeros on the imaginary axis and the polynomial $p^*(z)$ is of degree $n^* = n - s$ and has no such zeros.

First, we consider the case where s is even. Then

$$p(i\omega) = U(\omega) + iV(\omega) = d(i\omega)[U^*(\omega) + iV^*(\omega)].$$

Since $d(\omega)$ has even degree s, and all its zeros lie on the imaginary axis, it follows from the factor theorem and the complex conjugate root theorem that

$$d(\omega) = (\omega^2 + \omega_1\overline{\omega_1})(\omega^2 + \omega_2\overline{\omega_2})\cdots(\omega^2 + \omega_{s/2}\overline{\omega_{s/2}})$$

where $\overline{\omega_k}$ denotes complex conjugate.

But then $d(i\omega)$ is a real polynomial in ω, and we have

$$\frac{U(\omega)}{V(\omega)} = \frac{U^*(\omega)}{V^*(\omega)}.$$

Since n and n^* have the same parity, that is they are either both even or both odd, we get from (3.7a), (3.7b) and (3.11) that

$$\frac{p_2(\omega)}{p_1(\omega)} = \frac{p_2^*(\omega)}{p_1^*(\omega)}.$$

Applying (3.10) to $p^*(z)$, we find

$$I_{-\infty}^{+\infty}\frac{p_2(\omega)}{p_1(\omega)} = I_{-\infty}^{+\infty}\frac{p_2^*(\omega)}{p_1^*(\omega)} = n^* - 2k = n - 2k - s,$$

which is precisely what (A.1) states.

Next, if s is odd, we note that the only way this can occur is if $d(z)$ has a zero at the origin of odd multiplicity. The rest of the zeros must for the same reason as in the even case be pairs of complex conjugates. Then $d(i\omega) = i\hat{d}(i\omega)$ where $\hat{d}(i\omega)$ is a real polynomial in ω. So

$$p(i\omega) = U(\omega) + iV(\omega) = i\hat{d}(i\omega)[U^*(\omega) + iV^*(\omega)] = \hat{d}(i\omega)[-V^*(\omega) + iU(\omega)],$$

and therefore

$$\frac{U(\omega)}{V(\omega)} = -\frac{V^*(\omega)}{U^*(\omega)}.$$

Since s is odd, n and n^* now have opposite parity, so (3.7a), (3.7b) and (3.11) again tells us that

$$\frac{p_2(\omega)}{p_1(\omega)} = \frac{p_2^*(\omega)}{p_1^*(\omega)},$$

which again confirms (A.1).

A.2 The singular case

This far we have considered the regular case, where in the Routh table all the numbers $b_0, c_0, d_0, \ldots$ are nonzero. In this section we shall consider the *singular cases* where in the first column of the Routh table there occurs a zero, say $h_0 = 0$. at this point Routh's algorithm stops, because to obtain the next row we must divide by h_0. The singular case can be of two types:

1) In the row in which h_0 occurs there are numbers different from zero. This means that at some point, the degree drops by more than one in the generalized Sturm chain (3.12).

2) All the numbers of the row in which h_0 occurs are zero. Then this row is the $(m+1)$:st where m is the number of polynomials in (3.12). In this case the degree drops by one from any polynomial to the next, but the degree of the last one, $p_m(\omega)$ has degree larger than one. In both cases the number of polynomials in (3.12) is $m < n + 1$.

In these cases, the ordinary Routh's algorithm comes to an end. However, Routh gives a special rule for continuing the scheme in both cases.

In case 1), by this special rule, we have to substitute for $h_0 = 0$ a small value ϵ of definite but arbitrary sign and continue to fill in the table. The rest of the elements of the first column of the table is are then rational functions of ϵ. The signs of these elements are determined by the 'smallness' and sign of ϵ. If any one of these elements vanishes, i.e. is constantly zero as a function of ϵ, then we replace this element by another small value η and continue the algorithm. This is perhaps best illustrated by an example.

Consider the polynomial

$$p(z) = z^4 + z^3 + 2z^2 + 2z + 1.$$

Following the procedure given so far, we obtain the table

$$\begin{array}{ccc} 1, & 2, & 1 \\ 1, & 2 & \\ \epsilon, & 1 & \\ 2 - \frac{1}{\epsilon} & & \\ 1 & & \end{array}$$

Here we find that $k = V(1, 1, \epsilon, 2 - \frac{1}{\epsilon}, 1)$. For sufficiently small ϵ, in fact $|\epsilon| < \frac{1}{2}$ we find that $k = 2$. By numerical methods one can find that $p(z)$ has the zeros $z = -0.62 \pm 0.44i$ and $z = 0.12 \pm 1.31i$ confirming the conclusion.

This method is based on the following observation: Since we assume that there is no singularity of the second kind, the polynomials $p_1(\omega)$ and $p_2(\omega)$ are relatively prime. We have seen before that it follows that the polynomial $p(z)$ has no zeros on the imaginary axis.

In the Routh table all the elements are expressed rationally in terms of the elements of the first two rows, i.e. the coefficients of the given polynomial. However one can with some effort see from (3.13) and (3.13′) and the analogous formulas for the rest of the rows, that given any two adjacent rows of the Routh table, and the first element of the preceding rows, we can reconstruct the first two rows. That is, we can express the coefficients of the original polynomial in terms of these numbers. Thus, all the numbers a, b can be represented as polynomials in

$$a_0, b_0, c_0, \dots, g_0, g_1, g_2, \dots, h_0, h_1, h_2, \dots$$

For clarity we give a small example. Suppose that we know the third and fourth row, and the first entry the first and second row, and that these are $a_0, b_0, c_0, c_1, d_0, d_1$. Then we can get the first two rows as

$$\begin{array}{lcc} a_0, & c_0 + \frac{a_0}{b_0}(d_0 + \frac{b_0}{c_0}c_1), & c_1 + \frac{a_0}{b_0}d_1 \\ b_0, & d_0 + \frac{b_0}{c_0}c_1, & d_1 \\ c_0, & c_1 & \\ d_0, & d_1 & \end{array}$$

Therefore, on replacing $h_0 = 0$ by ϵ we in fact modify our original polynomial. In place of the Routh table for $p(z)$ we have the Routh table for a polynomial $P(z, \epsilon)$, where $P(z, \epsilon)$ is a polynomial in z and the parameter ϵ which reduces to $p(z)$ for $\epsilon = 0$. Since the zeros of $P(z, \epsilon)$ change continuously with a change in ϵ, and since there are no zeros on the imaginary axis for $\epsilon = 0$, the number k of roots in the right half-plane is the same for $P(z, \epsilon)$ and $P(z, 0) = p(z)$ for values of ϵ sufficiently small in modulus.

We proceed to case 2). Suppose that

$$a_0 \neq 0, b_0 \neq 0, \dots, e_0 \neq 0, g_0 = 0, g_1 = 0, g_2 = 0, \dots$$

In this case the last polynomial in the generalized Sturm chain (3.12) is of the form

$$p_m(\omega) = e_0\omega^{n-m+1} - e_1\omega^{n-m-1} + \dots$$

Routh now proposes that we replace $p_{m+1}(\omega)$ which is the zero polynomial, by $p'_m(\omega)$. So, in place of $g_0, g_1, \dots$ we write instead the corresponding coefficients

$$(n - m + 1)e_0, (n - m - 1)e_1, \dots$$

and continue the algorithm.

By (A.1) we have that

$$I_{-\infty}^{+\infty} \frac{p_2(\omega)}{p_1(\omega)} = n - 2k - s$$

where the zeros of $p(z)$ on the imaginary axis coincide with the real zeros of $p_m(\omega)$. Thus if these real zeroes are simple, then

$$I_{-\infty}^{+\infty} \frac{p'_m(\omega)}{p_m(\omega)} = s$$

as noted before. Therefore

$$I_{-\infty}^{+\infty} \frac{p_2(\omega)}{p_1(\omega)} + I_{-\infty}^{+\infty} \frac{p'_m(\omega)}{p_m(\omega)} = n - 2k.$$

This formula shows that the missing part of the Routh table should be filled by the Routh table for the polynomials $p_m(\omega)$ and $p'_m(\omega)$. The coefficients of $p'_m(\omega)$ should therefore be used to replace the elements of the zero row.

If however the zeros of $f_m(\omega)$ are not simple, then we denote by $d(\omega)$ the greatest common divisor of $p_m(\omega)$ and $p'_m(\omega)$, by $e(\omega)$ the greatest common divisor of $d(\omega)$ and $d'(\omega)$ et cetera, and we get

$$I_{-\infty}^{+\infty}\frac{p'_m(\omega)}{p_m(\omega)}+I_{-\infty}^{+\infty}\frac{d'(\omega)}{d(\omega)}+I_{-\infty}^{+\infty}\frac{e'(\omega)}{e(\omega)}+\cdots=s.$$

Thus the number k can be found if the missing part of the Routh table is filled in by the Routh table for $p_m(\omega)$ and $p'_m(\omega)$, then the Routh table for $d(\omega)$ and $d'(\omega)$, then that for $e(\omega)$ and $e'(\omega)$ and so on. So if the zeros of $p_m(\omega)$ are not simple, the rule has to be applied several times to dispose of a singularity of the second type.

Using these two special rules one can determine k in most cases. However, the application of both rules does not enable us to determine the number k in all cases.

B Computing c

B.1 Flip from E_0

To show that the flip bifurcation from E_0, happening when $r = 3$ and $\beta < \beta_0$, is stable we had to determine the nondegeneracy coefficient

$$c=\frac{1}{6}\langle p, C(q,q,q)\rangle-\frac{1}{2}\langle p, B(q,(A-I_2)^{-1}B(q,q))\rangle,$$

where $B(x,y), C(x,y,z)$ are given by

$$B_i(x,y)=\sum_{j,k=1}^{n}\frac{\partial^2F_i(\xi)}{\partial\xi_j\partial\xi_k}\Big|_{\xi=0}x_jy_k,\tag{B.1}$$

and

$$C_i(x,y)=\sum_{j,k,l=1}^{n}\frac{\partial^3F_i(\xi)}{\partial\xi_j\partial\xi_k\partial\xi_l}\Big|_{\xi=0}x_jy_kz_l,\tag{B.2}$$

where $i = 1, 2$, and A is the Jacobian matrix evaluated at E_0.

We have

$$\begin{aligned}f(S,I)&=rS(1-S)-\frac{\beta SI}{1+aS}\\ g(S,I)&=(1-K)I+\frac{\beta SI}{1+aS}.\end{aligned}$$

To shift the fixed point to the origin, define

$$\begin{aligned}\xi_1&=S-S_0\\ \xi_2&=I-I_0=I\end{aligned}$$

and note that $\xi_1=\xi_2=0$ if and only if $S=S_0$ and $I=0$.

In these new coordinates the system becomes

$$\begin{aligned}\xi_1(n+1)&=f(\xi_1(n)+S_0,\xi_2(n))-S_0\\ \xi_2(n+1)&=g(\xi_1(n)+S_0,\xi_2(n)).\end{aligned}\tag{B.3}$$

We write the system (B.3) as

$$\begin{pmatrix}\xi_1(n+1)\\ \xi_2(n+1)\end{pmatrix} = J(E_0)\begin{pmatrix}\xi_1(n)\\ \xi_2(n)\end{pmatrix} + F(\xi_1(n),\xi_2(n)) \tag{B.4}$$

where as usual $J(E_0)$ is the Jacobian matrix evaluated at E_0. Then by definition

$$F = \begin{pmatrix}F_1\\ F_2\end{pmatrix} = \begin{pmatrix}f(\xi_1+S_0,\xi_2)-S_0\\ g(\xi_1+S_0,\xi_2)\end{pmatrix} - J(E_0)\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}, \tag{B.5}$$

and its Taylor expansion near the origin is given by

$$F(x) = \frac{1}{2}B(x,x) + \frac{1}{6}C(x,x,x) + O(||x||^4),$$

with $B(x,x), C(x,x,x)$ given by (B.1) and (B.2). Our system is two-dimensional, so we have

$$B(x,y) = \begin{pmatrix}B_1(x,y)\\ B_2(x,y)\end{pmatrix}.$$

Form (B.5) we find that

$$F_1(\xi_1,\xi_2) = r(\xi_1+\frac{r-1}{r})(1-\xi_1+\frac{r-1}{r}) - \frac{\beta(\xi_1+\frac{r-1}{r})\xi_2}{1+a(\xi_1+\frac{r-1}{r})} - \frac{r-1}{r} - (2-r)\xi_1 - \frac{\beta(r-1)\xi_2}{r+a(r-1)},$$

and

$$F_2(\xi_1,\xi_1) = (1-K)\xi_2 + \frac{\beta(\xi_1+\frac{r-1}{r})\xi_2}{1+a(\xi_1+\frac{r-1}{r})} - (1-K)\xi_2 - \frac{\beta(r-1)\xi_2}{r+a(r-1)} = \frac{\beta(\xi_1+\frac{r-1}{r})\xi_2}{1+a(\xi_1+\frac{r-1}{r})} - \frac{\beta(r-1)\xi_2}{r+a(r-1)}.$$

Now we can compute partial derivatives. As these computations are completely straight forward but somewhat tedious, we just state that

$$\begin{array}{lll}\frac{\partial^2 F_1}{\partial \xi_1^2}\Big|_{\xi=0} = -2r, & \frac{\partial^2 F_1}{\partial \xi_1 \partial \xi_2}\Big|_{\xi=0} = -\frac{\beta}{(1+\frac{a(r-1)}{r})^2}, & \frac{\partial^2 F_1}{\partial \xi_2^2}\Big|_{\xi=0} = 0,\\ \frac{\partial^2 F_2}{\partial \xi_1^2}\Big|_{\xi=0} = 0, & \frac{\partial^2 F_2}{\partial \xi_1 \partial \xi_2}\Big|_{\xi=0} = \frac{\beta}{(1+\frac{a(r-1)}{r})^2}, & \frac{\partial^2 F_2}{\partial \xi_2^2}\Big|_{\xi=0} = 0.\end{array}$$

Hence by (B.1) we get

$$B(x,y) = \begin{pmatrix}-2rx_1y_1 - \frac{\beta}{(1+\frac{a(r-1)}{r})^2}x_1y_2 - \frac{\beta}{(1+\frac{a(r-1)}{r})^2}x_2y_1\\ \frac{\beta}{(1+\frac{a(r-1)}{r})^2}x_1y_2 + \frac{\beta}{(1+\frac{a(r-1)}{r})^2}x_2y_1\end{pmatrix}.$$

Since $q = (1\ 0)^T$ we find that

$$B(q,q) = B(1,0,1,0) = \begin{pmatrix}-2r\\ 0\end{pmatrix},$$

which tells us that

$$B(q,q)\big|_{r=3} = -6\begin{pmatrix}1\\ 0\end{pmatrix}.$$

Finally, the matrix

$$(A-I_2)^{-1} = \begin{pmatrix}-2 & -\frac{2\beta}{2a+3}\\ 0 & \frac{2\beta}{2a+3}-K\end{pmatrix}^{-1} = \begin{pmatrix}-\frac{1}{2} & \frac{\beta}{(2a+3)K-2\beta}\\ 0 & -\frac{2}{2K-\frac{4\beta}{2a+3}}\end{pmatrix}$$

so that

$$(A - I_2)^{-1} B(q,q) = -6\Big(-\frac{1}{2}\Big) \begin{pmatrix}1\\0\end{pmatrix} = 3\begin{pmatrix}1\\0\end{pmatrix},$$

which implies that

$$B(q,(A - I_2)^{-1} B(q,q)) = B(1,0,3,0) = \begin{pmatrix}-6\cdot 3\\0\end{pmatrix} = -18\begin{pmatrix}1\\0\end{pmatrix}.$$

Now we can compute

$$-\frac{1}{2}\langle p, B(q,(A - I_2)^{-1} B(q,q))\rangle = -\frac{1}{2}\langle \begin{pmatrix}1\\p_2\end{pmatrix}, -18\begin{pmatrix}1\\0\end{pmatrix} = 9. \tag{B.6}$$

We are now well on the way. All that remains is to find $C(x,y,z)$ given by (B.2). Again, the computations are tedious but not very difficult. We just give the results:

$$\begin{aligned}
&\frac{\partial^3 F_1}{\partial \xi_1^3}\Big|_{\xi=0} = 0, && \frac{\partial^3 F_1}{\partial \xi_2^3}\Big|_{\xi=0} = 0,\\
&\frac{\partial^3 F_1}{\partial \xi_1^2 \partial \xi_2}\Big|_{\xi=0} = \frac{2a\beta}{(1+\frac{a(r-1)}{r})^3}, && \frac{\partial^3 F_1}{\partial \xi_1 \partial \xi_2^2}\Big|_{\xi=0} = 0,\\
&\frac{\partial^3 F_2}{\partial \xi_1^3}\Big|_{\xi=0} = 0, && \frac{\partial^3 F_2}{\partial \xi_2^3}\Big|_{\xi=0} = 0,\\
&\frac{\partial^3 F_2}{\partial \xi_1^2 \partial \xi_2}\Big|_{\xi=0} = -\frac{2a\beta}{(1+\frac{a(r-1)}{r})^3}, && \frac{\partial^3 F_2}{\partial \xi_1 \partial \xi_2^2}\Big|_{\xi=0} = 0.
\end{aligned}$$

Using this and (B.2) we get

$$C(x,y,z) = \begin{pmatrix} \frac{2a\beta}{(1+\frac{a(r-1)}{r})^3} x_1 y_1 z_2 + \frac{2a\beta}{(1+\frac{a(r-1)}{r})^3} x_1 y_2 z_1 + \frac{2a\beta}{(1+\frac{a(r-1)}{r})^3} x_2 y_1 z_1 \\ -\frac{2a\beta}{(1+\frac{a(r-1)}{r})^3} x_1 y_1 z_2 - \frac{2a\beta}{(1+\frac{a(r-1)}{r})^3} x_1 y_2 z_1 - \frac{2a\beta}{(1+\frac{a(r-1)}{r})^3} x_2 y_1 z_1 \end{pmatrix},$$

and we see that

$$C(q,q,q) = C(1,0,1,0,1,0) = \begin{pmatrix}0\\0\end{pmatrix}$$

which entails

$$\frac{1}{6}\langle p, C(q,q,q)\rangle = 0. \tag{B.7}$$

Now, using (B.6) and (B.7) we finally get

$$c = \frac{1}{6}\langle p, C(q,q,q)\rangle - \frac{1}{2}\langle p, B(q,(A - I_2)^{-1} B(q,q))\rangle = 0 + 9 = 9.$$

B.2 Flip from E_1

Again, our aim is to compute

$$c = \frac{1}{6}\langle p, C(q,q,q)\rangle - \frac{1}{2}\langle p, B(q,(A - I_2)^{-1} B(q,q))\rangle.$$

Again, we shift the fixed point to the origin by defining

$$\begin{aligned}
\xi_1 &= S - S_1 = S - \frac{K}{\beta - aK}\\
\xi_2 &= I - I_1 = I - \frac{r-1}{\beta - aK} - \frac{rK}{(\beta - aK)^2}.
\end{aligned}$$

Then $\xi_1 = \xi_2 = 0$ if and only if $S = S_1$ and $I = I_1$. Again, we write

$$\begin{pmatrix} \xi_1(n+1) \\ \xi_2(n+1) \end{pmatrix} = J(E_1) \begin{pmatrix} \xi_1(n) \\ \xi_2(n) \end{pmatrix} + F(\xi_1(n), \xi_2(n)), \tag{B.8}$$

so that, again

$$F = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = \begin{pmatrix} f(\xi_1 + S_1, \xi_2 + I_1) - S_1 \\ g(\xi_1 + S_1, \xi_2 + I_1) - I_1 \end{pmatrix} - J(E_1) \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \tag{B.9}$$

and its Taylor expansion near the origin is given by

$$F(x) = \frac{1}{2}B(x,x) + \frac{1}{6}C(x,x,x) + O(||x||^4),$$

with $B(x,x), C(x,x,x)$ given by (B.1) and (B.2). Our system is two-dimensional, so we have

$$B(x,y) = \begin{pmatrix} B_1(x,y) \\ B_2(x,y) \end{pmatrix}.$$

We see that

$$F_1(\xi_1, \xi_2) = r(\xi_1 + S_1)(1 - \xi_1 - S_1) - \frac{\beta(\xi_2 + I_1)(\xi_1 + S_1)}{1 + a(\xi_1 + S_1)} - S_1 - a_{11}\xi_1 + K\xi_2$$

and

$$F_2(\xi_1, \xi_2) = (1 - K)(\xi_2 + I_1) + \frac{\beta(\xi_2 + I_1)(\xi_1 + S_1)}{1 + a(\xi_1 + S_1)} - I_1 - a_{21}\xi_1 - \xi_2.$$

Again, the computation of partial derivatives is not particularly interesting, so we just state that

$$\begin{gathered}
\frac{\partial^2 F_1}{\partial \xi_1^2}\Big|_{\xi=0} = \frac{2a(aK-\beta)(aK(r-1)+\beta+Kr-\beta r)}{\beta^2} - 2r, \\
\frac{\partial^2 F_1}{\partial \xi_1 \partial \xi_2}\Big|_{\xi=0} = -\frac{(\beta - aK)^2}{\beta}, \ \frac{\partial^2 F_1}{\partial \xi_2^2}\Big|_{\xi=0} = 0, \\
\frac{\partial^2 F_2}{\partial \xi_1^2}\Big|_{\xi=0} = -\frac{2a(aK-\beta)(aK(r-1)+\beta+Kr-\beta r)}{\beta^2}, \\
\frac{\partial^2 F_2}{\partial \xi_1 \partial \xi_2}\Big|_{\xi=0} = \frac{(\beta - aK)^2}{\beta}, \ \frac{\partial^2 F_2}{\partial \xi_2^2}\Big|_{\xi=0} = 0,
\end{gathered}$$

which means that

$$B(x,y) = \begin{pmatrix} x_1 y_1 \left(\frac{2a(aK-\beta)(aK(r-1)+\beta+Kr-\beta r)}{\beta^2} - 2r \right) - \frac{x_2 y_1 (\beta - aK)^2}{\beta} - \frac{x_1 y_2 (\beta - aK)^2}{\beta} \\ \frac{x_1(aK-\beta)(\beta y_2(aK-\beta) - 2a y_1(K(a(r-1)+r)+\beta-\beta r)) + \beta x_2 y_1 (\beta - aK)^2}{\beta^2} \end{pmatrix}.$$

Next, we compute

$$\begin{aligned}
&\frac{\partial^3 F_1}{\partial \xi_1^3}\Big|_{\xi=0} = \frac{6a^2(\beta-aK)^2(aK(r-1)+\beta+Kr-\beta r)}{\beta^3}, && \frac{\partial^3 F_1}{\partial \xi_2^3}\Big|_{\xi=0} = -\frac{2a(aK-\beta)^3}{\beta^2}, \\
&\frac{\partial^3 F_1}{\partial \xi_1^2 \partial \xi_2}\Big|_{\xi=0} = 0, && \frac{\partial^3 F_1}{\partial \xi_1 \partial \xi_2^2}\Big|_{\xi=0} = 0, \\
&\frac{\partial^3 F_2}{\partial \xi_1^3}\Big|_{\xi=0} = -\frac{6a^2(\beta-aK)^2(aK(r-1)+\beta+Kr-\beta r)}{\beta^3}, && \frac{\partial^3 F_2}{\partial \xi_2^3}\Big|_{\xi=0} = \frac{2a(aK-\beta)^3}{\beta^2}, \\
&\frac{\partial^3 F_2}{\partial \xi_1^2 \partial \xi_2}\Big|_{\xi=0} = 0, && \frac{\partial^3 F_2}{\partial \xi_1 \partial \xi_2^2}\Big|_{\xi=0} = 0,
\end{aligned}$$

which allows us to determine

$$C(x,y,z) = \begin{pmatrix} C_1(x,y,z) \\ C_2(x,y,z) \end{pmatrix},$$

where

$$C_1(x,y,z) = \tfrac{6a^2x_1y_1z_1(\beta-aK)^2(aK(r-1)+\beta+Kr-\beta r)}{\beta^3} - \tfrac{2ax_2y_1z_1(aK-\beta)^3}{\beta^2} - \tfrac{2ax_1y_2z_1(aK-\beta)^3}{\beta^2} - \tfrac{2ax_1y_1z_2(aK-\beta)^3}{\beta^2}$$

and

$$C_2(x,y,z) = -\tfrac{6a^2x_1y_1z_1(\beta-aK)^2(aK(r-1)+\beta+Kr-\beta r)}{\beta^3} + \tfrac{2ax_2y_1z_1(aK-\beta)^3}{\beta^2} + \tfrac{2ax_1y_2z_1(aK-\beta)^3}{\beta^2} + \tfrac{2ax_1y_1z_2(aK-\beta)^3}{\beta^2}.$$

This then would in principle allow us to compute

$$c = \frac{1}{6}\langle p, C(q,q,q)\rangle - \frac{1}{2}\langle p, B(q,(A-I_2)^{-1}B(q,q))\rangle,$$

where we would have to replace β by β_1 everywhere. Unfortunately, even using Mathematica this is a very complicated expression. Numerical computations show that c can be both positive and negative, which means by continuity and the intermediate value theorem that it can also be zero.

C Computing d

We give briefly the steps one goes through to compute the nondegeneracy coefficient d. In appendix B we have computed the multilinear functions $B(x,y)$ and $C(x,y,z)$ for E_1. They remain the same here. First, we note that the characteristic polynomial is

$$P(z) = z^2 - (a_{11}+1)z + a_{11} + Ka_{21},$$

which yields the eigenvalues (that we know are complex)

$$\mu_{1,2} = \frac{a_{11}+1 \pm i\sqrt{4(a_{11}+Ka_{21}) - (a_{11}+1)^2}}{2},$$

and we discussed before that $\mu_{1,2} = e^{\pm i\theta_0} = \sigma \pm \omega$ where $2\sigma = a_{11}+1$. It follows from Euler's formula that $\sigma = \cos\theta_0$, and hence $\theta_0 = \arccos(\frac{a_{11}+1}{2})$.

Now, we wish to determine a generalized eigenvector q of A. Such a vector satisfies

$$Aq = e^{i\theta_0}q, \;\; A\bar{q} = e^{-i\theta_0}\bar{q}.$$

We get q by solving

$$\begin{pmatrix} a_{11} - e^{i\theta_0} & -K \\ a_{21} & 1 - e^{i\theta_0} \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

We may choose $q_1 = 1$ which yields $q_2 = \frac{a_{11}-e^{i\theta_0}}{K}$. Hence

$$q = \begin{pmatrix} 1 \\ \frac{a_{11}-e^{i\theta_0}}{K} \end{pmatrix}.$$

Next, we seek a generalized adjoint eigenvector p, which we normalize as before. Then p must satisfy

$$A^Tp = e^{i\theta_0}p, \;\; A^T\bar{p} = e^{-i\theta_0}\bar{p}, \;\; \langle p, q\rangle = 1,$$

which gives us three equations to solve:

$$\begin{cases} p_1 + \frac{a_{11}-e^{i\theta_0}}{K} p_2 = 1 \\ p_1(a_{11} - e^{i\theta_0}) + a_{21}p_2 = 0 \\ -Kp_1 + (1 - e^{i\theta_0})p_2 = 0. \end{cases}$$

This yields

$$p = \frac{1}{a_{11} - 2e^{i\theta_0} + 1} \begin{pmatrix} 1 - e^{i\theta_0} \\ K \end{pmatrix}.$$

Now, using Mathematica, replacing β everywhere by β_2, we can compute

$$d = \frac{1}{2} Re\Bigg(e^{-i\theta_0} \Bigg[\langle p, C(q, q, \bar{q}) \rangle + 2\langle p, B(q, (A - I_n)^{-1} B(q, \bar{q}) \rangle + \langle p, B(\bar{q}, (e^{2i\theta_0} I_n - A)^{-1} B(q, q)) \rangle \Bigg] \Bigg).$$

Unfortunately, this is a massively complicated expression, so we have to resort to numerical experimentation. This strongly suggests that $d < 0$ for all choices of a and K when $1 < r < r_{max}$. Further, as r approaches 1 from above, it seems very clear that $d \to -\infty$. If one plots d as a function of r, it reaches a local maximum for r between 1 and 3. Usually this maximum is attained quite close to $r = 1$. All this strongly suggests that $d < 0$ for $1 < r < r_{max}$.

References

[1] Abdul Qadeer Khan. Bifurcations of a two-dimensional discrete-time predator–prey model. *Advances in Difference Equations*, 2019(1):56, 2019.

[2] Ming Zhao, Zuxing Xuan, and Cuiping Li. Dynamics of a discrete-time predator-prey system. *Advances in Difference Equations*, 2016(1):191, 2016.

[3] Selvam A.G.M & Praveen D.J. Qualitative analysis of a discrete sir epidemic model. *International Journal of Computational Engineering Research*, 5(3):35–39, 2015.

[4] Yuri A Kuznetsov. *Elements of applied bifurcation theory*, volume 112. Springer Science & Business Media, 2013.

[5] Saber N Elaydi. *Discrete chaos: with applications in science and engineering*. Chapman and Hall/CRC, 2007.

[6] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[7] Scott Zimmerman. An undergraduate's guide to the hartman-grobman and poincaré-bendixon theorems. *Lecture notes*, 2008.

[8] Howard Howie Weiss. The sir model and the foundations of public health. *Materials matematics*, pages 0001–17, 2013.

[9] Teri Johnson and Barry McQuarrie. Mathematical modeling of diseases: Susceptible-infected-recovered (sir) model. In *University of Minnesota, Morris, Math 4901 Senior Seminar*, 2009.

[10] F.R. Gantmakher and K.A. Hirsch. *The Theory of Matrices, Volume 2*. AMS Chelsea Publishing Series. American Mathematical Society, 2000.

[11] Edward B Saff and Arthur David Snider. *Fundamentals of complex analysis for mathematics, science, and engineering*. Prentice-Hall, 1976.

[12] N Matsumoto. Simple proof of the routh stability criterion based on order reduction of polynomials and principle of argument. In *ISCAS 2001. The 2001 IEEE International Symposium on Circuits and Systems (Cat. No. 01CH37196)*, volume 1, pages 699–702. IEEE, 2001.

[13] Tien-Yien Li and James A Yorke. Period three implies chaos. *The American Mathematical Monthly*, 82(10):985–992, 1975.

[14] Louis S Block and William A Coppel. *Dynamics in one dimension*. Springer, 2006.

[15] Philip D Straffin. Periodic points of continuous functions. *Mathematics Magazine*, 51(2):99–105, 1978.

[16] William A Coppel. The solution of equations by iteration. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 41–43. Cambridge University Press, 1955.

[17] Peter Štefan. A theorem of šarkovskii on the existence of periodic orbits of continuous endomorphisms of the real line. *Communications in Mathematical Physics*, 54(3):237–248, 1977.

[18] Partha Saha and Steven H Strogatz. The birth of period three. *Mathematics Magazine*, 68(1):42–47, 1995.