Ordinary Differential Equations: A Constructive Approach

M. Gameiro, J.-P. Lessard, J. Mireles James, K. Mischaikow

January 8, 2017

Contents

Chapter 1

Motivation

1.1 Exercises

Exercise 1.1.1. Recalling that Hook's law for the force exerted on a mass by a stretched spring is

$$F(x) = -Kx,$$

where K is the "stiffness" constant of the spring, and x is the signed displacement from equilibrium. Derive the equation of motion for a mass connected to a spring on a frictionless table (perhaps a surface of ice or maybe the mass has well oiled wheels).

Suppose that the mass is

Exercise 1.1.2.

$$\vdots$$

Chapter 2

Existence and Uniqueness, Flows and Definitions

In this chapter we provide fundamental results concerning the existence, uniqueness, and continuity of solutions to ODEs. The results of this section are classical and can be found in any graduate level text on ordinary differential equations. We have chosen not to present the most general results, but rather the minimal results necessary for this book. This is done because we can directly use the contraction mapping theorem which is suggestive of many of the proofs in this class. See [1, Section 1.12] or [4, Chapter 1] for alternative and/or more general proofs of existence and uniqueness. Our presentation follows that of [6, Section V.5.3].

2.1 Contraction Mapping Theorem

Consider a function $T\colon X \to X$ where X is a topological space. An element $\tilde{x} \in X$ is a *fixed point* of T, if $T(\tilde{x}) = \tilde{x}$. A fixed point is *globally attracting* if $\lim_{n\to\infty} T^n(x) = \tilde{x}$ for all $x \in X$.

Definition 2.1.1. Let $(X, \mathbf{d})$ denote a metric space. A function $T\colon X \to X$ is a *contraction* if there is a number $\kappa \in [0, 1)$, called a *contraction constant*, such that

$$\mathbf{d}(T(x), T(y)) \leq \kappa \mathbf{d}(x, y)$$

for all $x, y \in X$.

Theorem 2.1.2 (Contraction Mapping Theorem). *Let $(X, \mathbf{d})$ be a complete metric space. Assume that $T\colon X \to X$ is a contraction with contraction constant κ. Then there exists a unique, globally attracting fixed point $\tilde{x} \in X$. Furthermore, for any $x \in X$,*

$$\mathbf{d}(T^n(x), \tilde{x}) \leq \frac{\kappa^n}{1-\kappa} \mathbf{d}(T(x), x). \tag{2.1}$$

Proof. Choose $x_0 \in X$. Recursively define

$$x_{n+1} := T(x_n).$$

By assumption

$$\mathbf{d}(x_{n+1}, x_n) = \mathbf{d}(T(x_n), T(x_{n-1})) \leq \kappa \mathbf{d}(x_n, x_{n-1}).$$

Thus, by induction

$$\mathbf{d}(x_{n+1}, x_n) \leq \kappa^n \mathbf{d}(x_1, x_0).$$

By the triangle inequality for $n < m$,

$$\begin{aligned}
\mathbf{d}(x_n, x_m) &\leq \sum_{j=n}^{m-1} \mathbf{d}(x_{j+1}, x_j) \\
&\leq \sum_{j=n}^{m-1} \kappa^j \mathbf{d}(x_1, x_0) \\
&\leq \kappa^n \left(\sum_{k=0}^{\infty} \kappa^k \right) \mathbf{d}(x_1, x_0) \\
&\leq \kappa^n \frac{1}{1-\kappa} \mathbf{d}(x_1, x_0).
\end{aligned}$$

This implies that $\{x_n\}$ is a Cauchy sequence. Since X is complete there exists $\tilde{x} \in X$ such that

$$\lim_{n\to\infty} x_n = \tilde{x}.$$

Hence, by continuity of T,

$$\tilde{x} = \lim_{n\to\infty} x_n = \lim_{n\to\infty} T(x_{n-1}) = T\left(\lim_{n\to\infty} x_{n-1} \right) = T(\tilde{x}).$$

This establishes the existence of a fixed point.

To prove (2.1), observe that again by the triangle inequality

$$\begin{aligned}
\mathbf{d}(x_n, \tilde{x}) &\leq \mathbf{d}(x_n, x_m) + \mathbf{d}(x_m, \tilde{x}) \\
&\leq \frac{\kappa^n}{1-\kappa} \mathbf{d}(x_1, x_0) + \mathbf{d}(x_m, \tilde{x})
\end{aligned}$$

for all $m > n$. Taking the limit as $m \to \infty$ gives the desired inequality.

Assume now that there is another fixed point $\tilde{y}$ of T, that is $T(\tilde{y}) = \tilde{y}$. Setting $x = \tilde{y}$ in (2.1) and using the fact that $T^n(\tilde{y}) = \tilde{y}$ for all n, implies that

$$\mathbf{d}(\tilde{y}, \tilde{x}) \leq \frac{\kappa^n}{1-\kappa} \mathbf{d}(\tilde{y}, \tilde{y}) = 0.$$

Hence, $\tilde{x} = \tilde{y}$ and therefore the fixed point is unique.

Finally, (2.1) and the fact that $\kappa \in [0, 1)$ proves that $\tilde{x}$ is a globally attracting fixed point. □

2.2 Existence and Uniqueness of Solutions to ODEs

Definition 2.2.1. Let $f\colon U \to \mathbb{R}^n$ be a continuous function defined on an open set $U \subset \mathbb{R}^n$. A *solution* to the differential equation

$$\dot{x} := \frac{dx}{dt} = f(x) \tag{2.2}$$

on an interval $J \subset \mathbb{R}$ is a differentiable function $\varphi\colon J \to U$ such that $\varphi(t) \in U$ and

$$\frac{d\varphi}{dt}(t) = f(\varphi(t))$$

for all $t \in J$.

In this section we focus on solutions to the *initial value problem* (IVP)

$$\dot{x} = f(x), \quad x(t_0) = x_0,$$

that is, the existence of a solution $x\colon J \to U$ such that $t_0 \in J$ and $x(t_0) = x_0$.

Definition 2.2.2. Consider metric spaces $(X, \mathbf{d}_X)$ and $(Y, \mathbf{d}_Y)$. A function $f\colon X \to Y$ is *Lipschitz* if there exists a real constant $K \geq 0$ such that, for all $x_1, x_2 \in X$,

$$\mathbf{d}_Y(f(x_1), f(x_2)) \leq K\mathbf{d}_X(x_1, x_2).$$

The smallest K satisfying this inequality is denoted by $Lip(f) := K$ and is called the *Lipschitz constant of* f.

More generally, f is *locally Lipschitz* if every point in X has a neighborhood such that f restricted to that neighborhood is Lipschitz.

The first goal of this section is the proof of the following theorem which guarantees local existence and uniqueness of solutions.

Theorem 2.2.3. *Assume $f\colon U \to \mathbb{R}^n$ is a locally Lipschitz continuous function defined on an open set $U \subset \mathbb{R}^n$. If $x_0 \in U$, then there exists an open interval $J \subset \mathbb{R}$, containing t_0, over which a solution to the initial value problem*

$$\dot{x} = f(x), \quad x(t_0) = x_0 \tag{2.3}$$

is defined. Furthermore, any two solutions to the initial value problem agree on the intersection of their domains of definition.

The proof of Theorem 2.2.3 is obtained via a series of propositions and lemmas. The first step in the proof is the observation that by the fundamental theorem of calculus a solution to an ODE can be recast as a solution to an integral equation.

Lemma 2.2.4. *Assume $f\colon U \to \mathbb{R}^n$ is a continuous function defined on an open set $U \subset \mathbb{R}^n$. Let $x_0 \in U$. A continuous function $\varphi\colon J \to \mathbb{R}^n$ is a solution to the initial value problem $\dot{x} = f(x)$, $x(t_0) = x_0$ if and only if*

$$\varphi(t) = x_0 + \int_{t_0}^{t} f(\varphi(s))\, ds.$$

Observe that since f is independent of t, there is no loss of generality in assuming that $t_0 = 0$, which for the sake of simplicity of expression will be done henceforth. The following proposition provides for existence of solutions.

Proposition 2.2.5. *If $f\colon U \to \mathbb{R}^n$ is a locally Lipschitz continuous function defined on an open set $U \subset \mathbb{R}^n$ and $x_0 \in U$, then there exists a solution $\varphi\colon (-a, a) \to \mathbb{R}^n$ for some $a > 0$ to the initial value problem*

$$\dot{x} = f(x), \quad x(0) = x_0. \tag{2.4}$$

Proof. Choose $\epsilon > 0$ such that

$$\overline{B_\epsilon(x_0)} = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq \epsilon\} \subset U,$$

where $\|\cdot\|$ is a norm for $\mathbb{R}^n$.

For ϵ sufficiently small f is Lipschitz on $\overline{B_\epsilon(x_0)}$, and hence there exist positive constants K and M such that

$$\|f(x) - f(y)\| \leq K\|x - y\| \quad \text{and} \quad \|f(x)\| \leq M$$

for all $x, y \in \overline{B_\epsilon(x_0)}$.

By Lemma 2.2.4 it is sufficient to prove the existence of φ satisfying

$$\varphi(t) = x_0 + \int_{0}^{t} f(\varphi(s))\, ds. \tag{2.5}$$

The strategy is to define a function space X within which we expect to find a solution and define a contraction $T\colon X \to X$ such that the fixed point of T is a solution to (2.5).

Since solutions are functions of time the domain of elements of the function space should be an interval $J \subset \mathbb{R}$. Choose $a > 0$ such that

$$a < \min\left\{\frac{\epsilon}{M}, \frac{1}{K}\right\} \tag{2.6}$$

and set $J := (-a, a)$. Define

$$X := \left\{\alpha\colon J \to \overline{B_\epsilon(x_0)} \;\middle|\; \alpha \in C^0(J), \text{ and } Lip(\alpha) \leq M\right\}.$$

We endow X with the C^0 norm

$$\|\alpha\|_{C^0(J)} := \sup\{\|\alpha(t)\| \mid t \in J\}$$

and leave it as an exercise to show that this implies that X is a complete metric space.

Given $\alpha \in X$ define an operator $T\colon X \to C^0(J)$ by

$$T(\alpha)(t) := x_0 + \int_0^t f(\alpha(s))\, ds.$$

Observe that by definition $T(\alpha) = \alpha$ if and only if $\varphi = \alpha$ satisfies (2.5). To apply the contraction mapping theorem (Theorem 2.1.2) we need to prove two things: (i) $T\colon X \to X$ and (ii) T is a contraction.

We first prove (i). Observe that

$$T(\alpha)(0) = x_0 + \int_0^0 f(\alpha(s))\, ds = x_0. \tag{2.7}$$

Furthermore, given $t_1, t_2 \in J$,

$$\begin{aligned}
\|T(\alpha)(t_2) - T(\alpha)(t_1)\| &= \left\| \int_0^{t_2} f(\alpha(s))\, ds - \int_0^{t_1} f(\alpha(s))\, ds \right\| \\
&= \left\| \int_{t_1}^{t_2} f(\alpha(s))\, ds \right\| \\
&\leq \left| \int_{t_1}^{t_2} M\, ds \right| \\
&= M|t_2 - t_1|
\end{aligned} \tag{2.8}$$

and hence $Lip(T(\alpha)) \leq M$. Setting $t_2 = t$, $t_1 = 0$ and applying (2.7) and(2.8) we obtain

$$\|T(\alpha)(t) - x_0\| \leq M|t| < Ma < \epsilon$$

for all $t \in J$ and hence $T(\alpha)\colon J \to \overline{B_\epsilon(x_0)}$. Therefore, $T(\alpha) \in X$.

We now prove (ii). Observe that given $\alpha, \beta \in X$

$$\begin{aligned}
\|T(\alpha) - T(\beta)\|_{C^0(J)} &= \sup_{t\in J} \|T(\alpha)(t) - T(\beta)(t)\| \\
&= \sup_{t\in J} \left\| \int_0^t f(\alpha(s))\, ds - \int_0^t f(\beta(s))\, ds \right\| \\
&\leq \sup_{t\in J} \int_0^t \|f(\alpha(s)) - f(\beta(s))\|\, ds \\
&\leq \sup_{t\in J} \int_0^t K\|\alpha(s) - \beta(s)\|\, ds \\
&\leq K\|\alpha - \beta\|_{C^0(J)} \sup_{t\in J} \int_0^t ds \\
&\leq Ka\|\alpha - \beta\|_{C^0(J)}.
\end{aligned}$$

By (2.6), $Ka < 1$, and therefore $T\colon X \to X$ is a contraction with contraction constant $\kappa = Ka < 1$. Denote by $\varphi\colon J \to \mathbb{R}^n$ the unique fixed point of T within X and observe that this implies that φ is a solution of the initial value problem (2.4). □

Returning to the details of the proof of Theorem 2.2.5 observe that if $\delta = \epsilon/2$, then for any $x \in B_\delta(x_0)$,

$$\overline{B_\delta(x)} \subset U, \quad \|f(z) - f(y)\| \leq K\|z - y\|, \quad \text{and} \quad \|f(z)\| \leq M$$

for all $z, y \in \overline{B_\delta(x)}$. This leads to the following corollary.

Corollary 2.2.6. *Let $f\colon U \to \mathbb{R}^n$ be a locally Lipschitz continuous function defined on an open set $U \subset \mathbb{R}^n$. For every $x_0 \in U$ there exists a neighborhood V of x_0 and a constant $a = a(V)$, such that for every $y \in V$ there exists a solution $\varphi(\cdot, y)\colon (-a, a) \to \mathbb{R}^n$ to the initial value problem*

$$\dot{x} = f(x), \quad x(0) = y.$$

Remark 2.2.7. It is worth noting that while we have proven the existence of a solution, it is limited to a time interval of length $2a$ where $a < \min\left\{\frac{\epsilon}{M}, \frac{1}{K}\right\}$, which could be very small. Theorem 2.2.18 provides information concerning existence over longer intervals of time.

Observe that we have only proven uniqueness of existence of solutions over the family of functions $Lip(\alpha) \leq M$, as opposed to all differentiable functions. Thus the proof of Theorem 2.2.3 remains to be completed. For the moment we turn our attention to a different question and demonstrate that solutions to an IVP are Lipschitz continuous as a function of the initial value. The following inequality will be used to prove the previous statement and is fundamental to the study of differential equations.

Theorem 2.2.8 (Gronwall's Inequality). *Let $\alpha, \beta \colon (a,b) \to [0,\infty)$ be continuous functions. Assume*

$$\alpha(t) \leq C + \left| \int_{t_0}^{t} \alpha(s)\beta(s)\, ds \right|, \quad t_0, t \in (a,b)$$

for some constant $C \geq 0$. Then,

$$\alpha(t) \leq C \exp\left(\left| \int_{t_0}^{t} \beta(s)\, ds \right| \right).$$

Proof. The proof is done in several cases.

Assume $a < t_0 \leq t < b$. Define

$$G(t) := C + \int_{t_0}^{t} \alpha(s)\beta(s)\, ds.$$

Then

$$G'(t) = \frac{dG}{dt}(t) = \alpha(t)\beta(t) \leq G(t)\beta(t).$$

Now assume $C > 0$. Then $G(t) > 0$ and hence

$$\begin{aligned}
\frac{G'(t)}{G(t)} &\leq \beta(t) \\
\int_{t_0}^{t} \frac{G'(s)}{G(s)}\, ds &\leq \int_{t_0}^{t} \beta(s)\, ds \\
\log\left(\frac{G(t)}{G(t_0)} \right) &\leq \int_{t_0}^{t} \beta(s)\, ds \\
G(t) &\leq G(t_0) \exp\left(\int_{t_0}^{t} \beta(s)\, ds \right) \\
G(t) &\leq C \exp\left(\int_{t_0}^{t} \beta(s)\, ds \right) \\
\alpha(t) &\leq C \exp\left(\int_{t_0}^{t} \beta(s)\, ds \right).
\end{aligned}$$

Now assume $C > 0$ and $a < t \leq t_0 < b$. Define

$$G(t) := C + \int_{t}^{t_0} \alpha(s)\beta(s)\, ds$$

and repeat the argument.

Finally, assume that $C = 0$ and consider a sequence $C_n > 0$ converging to 0. By the previous argument and the fact that α is non-negative,

$$0 \leq \alpha(t) \leq C_n \exp\left(\left| \int_{t_0}^{t} \beta(s)\, ds \right| \right)$$

for all C_n and hence $\alpha(t) \equiv 0$. □

We now use Gronwall's Inequality to show that solutions to an IVP are Lipschitz continuous with respect to initial conditions.

Proposition 2.2.9. *Let $U \subset \mathbb{R}^n$ be an open set and assume $f\colon U \to \mathbb{R}^n$ is a Lipschitz continuous function with $Lip(f) = K$. If $\varphi(\cdot, x_0)\colon J_{x_0} \to \mathbb{R}^n$ and $\psi(\cdot, y_0)\colon J_{y_0} \to \mathbb{R}^n$ are solutions to the initial value problem $\dot{x} = f(x)$ with $x(0) = x_0$ and $x(0) = y_0$, respectively, then*

$$\|\varphi(t, x_0) - \psi(t, y_0)\| \leq \|x_0 - y_0\| e^{K|t|} \tag{2.9}$$

for all $t \in J_{x_0} \cap J_{y_0}$. Furthermore, with respect to the C^0 norm, solutions to the IVP are locally Lipschitz continuous with respect to the initial value.

Proof. Let $\varphi(\cdot, x_0)\colon J_{x_0} \to \mathbb{R}^n$ and $\psi(\cdot, y_0)\colon J_{y_0} \to \mathbb{R}^n$ be solutions to the initial value problems $x(0) = x_0$ and $x(0) = y_0$, respectively. By Proposition 2.2.5 there exists $a > 0$ such that $J = (-a, a) \subset J_{x_0} \cap J_{y_0}$. By Lemma 2.2.4,

$$\varphi(t, x_0) - \psi(t, y_0) = x_0 - y_0 + \int_0^t f(\varphi(s, x_0)) - f(\psi(s, y_0))\, ds.$$

Let $\alpha(t) := \|\varphi(t, x_0) - \psi(t, y_0)\|$. Then

$$\begin{aligned}
\alpha(t) &\leq \|x_0 - y_0\| + \left\| \int_0^t f(\varphi(s, x_0)) - f(\psi(s, y_0))\, ds \right\| \\
&\leq \|x_0 - y_0\| + \left| \int_0^t \|f(\varphi(s, x_0)) - f(\psi(s, y_0))\|\, ds \right| \\
&\leq \|x_0 - y_0\| + \left| \int_0^t K \|\varphi(s, x_0) - \psi(s, y_0)\|\, ds \right| \\
&\leq \|x_0 - y_0\| + \left| \int_0^t K \alpha(s)\, ds \right|
\end{aligned}$$

since $Lip(f) = K$. Applying Gronwall's Inequality with $C = \|x_0 - y_0\|$ and $\beta(t) = K$, we obtain

$$\begin{aligned}
\alpha(t) &\leq \|x_0 - y_0\| \exp\left(\left| \int_0^t K\, ds \right| \right) \\
&\leq \|x_0 - y_0\| e^{K|t|} \\
\|\varphi(t, x_0) - \psi(t, y_0)\| &\leq \|x_0 - y_0\| e^{K|t|}.
\end{aligned}$$

To see that solutions are locally Lipschitz continuous with respect to initial conditions, observe that given initial values x_0 and y_0 and the interval $J = (-a, a)$ as defined above

$$\begin{aligned}
\|\varphi(\cdot, x_0) - \psi(\cdot, y_0)\|_{C^0(J)} &= \sup_{t \in J} \|\varphi(t, x_0) - \psi(t, y_0)\| \\
&\leq \|x_0 - y_0\| \sup_{t \in J} e^{K|t|} \\
&= \|x_0 - y_0\| e^{Ka}.
\end{aligned}$$

□

Proof of Theorem 2.2.3. As is remarked above Proposition 2.2.5 guarantees the existence of solutions. To show that two solutions to the same IVP agree on the intersection of their domains of definition let $\varphi\colon J_0 \to \mathbb{R}^n$ and $\psi\colon J_1 \to \mathbb{R}^n$ denote two solutions to the initial value problem $\dot{x} = f(x)$, $x(0) = x_0$. By (2.9) for all $t \in J_0 \cap J_1$

$$\|\varphi(t) - \psi(t)\| = 0. \qquad \square$$

The following Corollary provides a summary of the results derived in the proof of Theorem 2.2.3.

Corollary 2.2.10. *Assume $f\colon U \to \mathbb{R}^n$ is a locally Lipschitz continuous function defined on an open set $U \subset \mathbb{R}^n$. For every $x_0 \in U$ and $t_0 \in \mathbb{R}$ there exists a neighborhood $V_0 \subset U$ of x_0, an interval $J_0 \subset \mathbb{R}$ containing t_0, and a Lipschitz continuous function $\varphi\colon J_0 \times V_0 \to \mathbb{R}^n$ such that $\varphi(\cdot, x_0)$ is a solution to the IVP, $\dot{x} = f(x)$, $x(t_0) = x_0$.*

The following proposition guarantees existence and uniqueness of solutions for every C^1 vector field. The proof is left to the reader.

Proposition 2.2.11. *Let $U \subset \mathbb{R}^n$ be an open set and $f\colon U \to \mathbb{R}^n$. If $f \in C^1(U)$, then f is locally Lipschitz.*

Note that since $\varphi(\cdot, x_0)$ is a solution to the IVP it is differentiable in t. One can also prove differentiability with respect to the initial conditions [1, Theorem 1.261]. This question is considered in Exercise 2.5.9 of this Chapter.

Models that arise in applications typically depend on set of parameters Λ and often are time dependent. Thus, we are interested in solutions to differential equations that appear to take a more general form. Let $J \subset \mathbb{R}$, $U \subset \mathbb{R}^n$ and $\Lambda \subset \mathbb{R}^m$ be open sets and let $f\colon J \times U \times \Lambda \to \mathbb{R}^n$ be a continuous function. For fixed $\lambda \in \Lambda$ a *solution* to the differential equation

$$\dot{x} = f(t, x, \lambda) \tag{2.10}$$

is a differentiable function $\varphi\colon J_0 \to U$ defined on an open interval $J_0 \subset J$ such that

$$\frac{d\varphi}{dt}(t) = f(t, \varphi(t), \lambda)$$

for all $t \in J_0$. For $t_0 \in J$, $x_0 \in U$ and $\lambda_0 \in \Lambda$, the *initial value problem (IVP)* associated with (2.10) requires finding a solution $\varphi(t) = \varphi(t; t_0, x_0, \lambda_0)$ to $\dot{x} = f(t, x, \lambda_0)$ satisfying $\varphi(t_0) = x_0$.

The corresponding existence and uniqueness theorem is as follows.

Theorem 2.2.12. *Let $J \subset \mathbb{R}$, $U \subset \mathbb{R}^n$ and $\Lambda \subset \mathbb{R}^m$ be open sets, and assume $f\colon J \times U \times \Lambda \to \mathbb{R}^n$ is a Lipschitz function. If $(t_0, x_0, \lambda_0) \in J \times U \times \Lambda$, then there exists an open*

neighborhood of the form $J_0 \times U_0 \times \Lambda_0$ *of* (t_0, x_0, λ_0) *and a Lipschitz continuous function* $\varphi\colon J_0 \times J_0 \times U_0 \times \Lambda_0 \to \mathbb{R}^n$ *such that for every* $(t_1, x_1, \lambda_1) \in J_0 \times U_0 \times \Lambda_0$

$$\varphi(\cdot, t_1, x_1, \lambda_1)\colon J_0 \to \mathbb{R}^n$$

is a solution to the initial value problem

$$\dot{x} = f(t, x, \lambda_1), \quad x(t_1) = x_1. \tag{2.11}$$

Furthermore, if $\psi(\cdot, t_1, x_1, \lambda_1)$ *is another solution to the initial value problem, then* $\psi(t) = \phi(t)$ *on the intersection of their domains of definition.*

Proof. The proof follows from the realization that result can be viewed as a special case of Theorem 2.2.3. Define $F\colon U \times J \times \Lambda \to \mathbb{R}^{n+1+m}$ by

$$F(x, s, \lambda) = (f(s, x, \lambda), 1, 0).$$

By Theorem 2.2.3, there exists a function $\varphi\colon J_0 \to \mathbb{R}^{n+1+m}$ which satisfies the initial value problem

$$\begin{cases} \dot{x} = f(s, x, \lambda) \\ \dot{s} = 1 \\ \dot{\lambda} = 0 \\ (x(t_0), s(t_0), \lambda(t_0)) = (x_0, s_0\lambda_0) \in U \times J \times \Lambda. \end{cases}$$

Furthermore, if $\psi\colon J_1 \to \mathbb{R}^{n+1+m}$ is another solution to this initial value problem, then they agree on the domain $J_0 \cap J_1$. It is left to the reader to check that the first n components of φ is a solution to (2.11). □

The most significant restriction on the assumptions of Theorem 2.2.3 is that f is Lipschitz. As the following example indicates, existence is possible in more general settings, however uniqueness of the solution can no longer be assumed.

Example 2.2.13. Consider the initial value problem

$$\dot{x} = 3x^{\frac{2}{3}}, \quad x(0) = 0. \tag{2.12}$$

Observe that

$$\varphi(t) \equiv 0 \quad \text{and} \quad \psi(t) := \begin{cases} t^3 & \text{if } t \geq 0 \\ 0 & \text{if } t \leq 0 \end{cases}$$

are two distinct solutions to (2.12) on $\mathbb{R}$.

Having established existence and uniqueness we turn to the question of the maximal time interval on which a solution is defined.

Example 2.2.14. Consider the initial value problem

$$\dot{x} = x^2, \quad x(0) = x_0, \tag{2.13}$$

where we assume that $x_0 > 0$. By Theorem 2.2.3 we know that this IVP has a locally unique solution. It is straightforward to check that

$$\varphi(t, x_0) := \frac{x_0}{1 - tx_0}$$

is a solution and that the maximal interval in time over which φ is defined is $-\infty < t < 1/x_0$. As motivation for the next theorem it is also worth observing that

$$\lim_{t \to 1/x_0} \varphi(t, x_0) = \infty.$$

Definition 2.2.15. Let $\varphi\colon I \to \mathbb{R}^n$ and $\psi\colon J \to \mathbb{R}^n$ be solutions to $\dot{x} = f(x)$, where $f\colon U \to \mathbb{R}^n$ is a continuous function defined on an open set $U \subset \mathbb{R}^n$. We say that ψ is an *extension* of φ if $I \subset J$ and $\varphi(t) = \psi(t)$ for all $t \in I$. If we also have that I is a proper subset of J, then we say that ψ is a *proper extension* of φ.

Definition 2.2.16. A solution $\varphi\colon J \to \mathbb{R}^n$ to $\dot{x} = f(x)$ is called a *maximal solution* if it has no proper extensions. In this case the interval J is called the *maximal interval of existence* of φ.

Theorem 2.2.17. *If $f\colon U \to \mathbb{R}^n$ is a Lipschitz continuous function defined on an open set $U \subset \mathbb{R}^n$ and $x_0 \in U$, then the initial value problem*

$$\dot{x} = f(x), \quad x(0) = x_0$$

has a unique maximal solution $\varphi\colon J \to \mathbb{R}^n$, and the maximal interval of existence J is open.

Proof. Let J be the union of all intervals I for which there is a solution $\psi\colon I \to \mathbb{R}^n$ for the initial value problem $\dot{x} = f(x)$, $x(0) = x_0$. Since all the intervals I contain $t_0 = 0$ we have that J is an interval. Define $\varphi\colon J \to \mathbb{R}^n$ by $\varphi(t) := \psi(t)$ if $t \in I$ and $\psi\colon I \to \mathbb{R}^n$ is a solution. Since any two solutions agree on the intersection of their domains of definition, we have that φ is well defined. The function $\varphi\colon J \to \mathbb{R}^n$ just defined is clearly the unique maximal solution to $\dot{x} = f(x)$, $x(0) = x_0$. It remains to show that the interval J is open. Let t_- and t_+ be the left and right end points of J, respectively. Assume that J is closed at t_+, then by Theorem 2.2.3 we can extend the solution to an interval about t_+, which contradicts the fact that J is the maximal interval. Likewise if we assume that J is closed at t_-. Therefore J is the open interval (t_-, t_+). □

Theorem 2.2.18. *Let $f\colon U \to \mathbb{R}^n$ be a Lipschitz continuous function defined on an open set $U \subset \mathbb{R}^n$. Consider the initial value problem*

$$\dot{x} = f(x), \quad x(0) = x_0$$

with maximal solution $\varphi(t, x_0)$. Let $J = (t_-, t_+)$ be the maximal interval of existence of φ.

(i) If $t_+ < \infty$, then given any compact set $C \subset U$ there is a time $t_C^+ \in (0, t_+)$ such that $\varphi(t_C^+, x_0) \notin C$. Similarly, if $t_- > -\infty$, then given any compact set $C \subset U$ there is a time $t_C^- \in (t_-, 0)$ such that $\varphi(t_C^-, x_0) \notin C$.

(ii) If $U = \mathbb{R}^n$ and $\|f(x)\|$ is bounded, then $(t_-, t_+) = \mathbb{R}$.

Proof. (i) Let $C \subset U$ compact and assume $\varphi([0, t_+), x_0) \subset C$. Since C is compact and f is continuous, there are positive constants K_1 and K_2 such that

$$\|f(x) - f(y)\| \leq K_1 \|x - y\| \quad \text{and} \quad \|f(x)\| \leq K_2$$

for all $x, y \in C$. Thus the proof of Theorem 2.2.3 implies that the solution φ satisfies $Lip(\varphi) \leq K_2$ for all $t \in [0, t_+)$. Because φ is Lipschitz with respect to t, there exists $x_+ = \lim_{t \to t_+} \varphi(t, x_0)$ and the compactness of C implies that $x_+ \in C$. Theorem 2.2.3 guarantees the existence of $\delta > 0$ and a solution $\psi\colon (t_+ - \delta, t_+ + \delta)$ to the initial value problem $\dot{x} = f(x)$, $x(t_+) = x_+$. Furthermore, $\psi = \phi$ on $(t_+ - \delta, t_+)$. This contradicts the assumption that $J = (t_-, t_+)$ is the maximal interval of existence of φ.

The argument for t_- is similar.

(ii) Letting $K := \sup_{x \in \mathbb{R}^n} \|f(x)\| < \infty$, note that

$$\|\varphi(t, x_0) - x_0\| = \| \int_0^t f(\varphi(s, x_0)) ds\| \leq \left| \int_0^t \|f(\varphi(s, x_0))\| ds \right| \leq K|t|. \tag{2.14}$$

This implies that $\varphi(t, x_0) \in B_R(x_0)$ for all $|t| < R/K$. Suppose that $t_+ < \infty$. Consider the compact set $C := \overline{B_{R_0}(x_0)}$ with $R_0 := Kt_+$. From part (1), there is a time $t_C^+ \in (0, t_+)$ such that $\varphi(t_C^+, x_0) \notin C$. Now, $t_C^+ < t_+ = R_0/K$. By (2.14), we get that $\varphi(t_C^+, x_0) \in B_{R_0}(x_0) \subset C$. This is a contradiction. That implies that $t_+ = \infty$. Similarly, one can show that $t_- = -\infty$. That allows us to conclude that $J = \mathbb{R}$. □

2.3 Regularity of Solutions

Because we want to focus on a particular framework for developing constructive proofs for ODEs we restrict our attention in the first part of this book to analytic vector fields. As is demonstrated in this section an analytic vector field leads to analytic solutions. As a first step notice that if φ is a solution to $\dot{x} = f(x)$, then φ is diferentiable and $\dot{\varphi}(t) = f(\varphi(t))$. Hence if f is continuous, the solutions must be at least C^1. Therefore, it is easy to check

that if $f \in C^r$, then $\varphi \in C^{r+1}$. In particular, this shows that if $f \in C^\omega$, then $\varphi \in C^\infty$. However, to obtain the analyticity of φ requires a different line of reasoning. The argument we present here is based on the idea of majorants. It does not provide the most efficient proof, but has the advantage that it can be extended to obtain the Cauchy-Kovalevskaya theorem (also know as Cauchy-Kowalevski theorem) for partial differential equations [2, Chapter 4, Theorem 2], [3].

For the sake of presentation we first discuss the one-dimensional case before presenting the general case.

Theorem 2.3.1 (Cauchy-Kovalevskaya Theorem: ODE version I). *Let $f\colon U \to \mathbb{R}$ be an analytic function defined on an open interval $U \subset \mathbb{R}$ containing the origin. If $x\colon J \to \mathbb{R}$ is the solution to the IVP*

$$\dot{x} = f(x), \quad x(0) = 0, \tag{2.15}$$

then x is an analytic function in a neighborhood of 0.

As is made clear shortly the following example is of particular interest in this context.

Example 2.3.2. Direct substitution shows that the unique solution to the IVP

$$\dot{y} = g(y) := C\frac{r}{r-y} = \sum_{k=0}^{\infty} Cr^{-k}y^k, \quad y(0) = 0,$$

where $C, r > 0$, is given by

$$y(t) = r - \sqrt{r^2 - 2Crt} \tag{2.16}$$

which is analytic for $|t| < r/(2C)$.

Proof of Theorem 2.3.1. Let $x\colon J \to \mathbb{R}$ be the solution to (2.15). Our goal is to show that x is analytic at $t = 0$, which is equivalent to showing that x is given by a convergent Taylor series at $t = 0$. We know that x is C^∞ and thus we can compute all of its derivatives, i.e.,

$$\begin{aligned}
\dot{x}(t) &= f(x(t)) \\
\ddot{x}(t) &= f'(x(t))\dot{x}(t) = f'(x(t))f(x(t)) \\
x^{(3)}(t) &= f''(x(t))[f(x(t))]^2 + [f'(x(t))]^2 f(x(t)) \\
&\vdots \\
x^{(k)}(t) &= p_k\left(f(x(t)), f'(x(t)), \ldots, f^{(k-1)}(x(t))\right),
\end{aligned}$$

where p_k is a polynomial in k-variables with all the coefficients being positive integers. By definition x is analytic at $t = 0$ if and only if there exists $\rho > 0$ such that (using $x(0) = 0$)

$$\sum_{k=0}^{\infty} \frac{1}{k!} p_k\left(f(0), \ldots, f^{(k-1)}(0)\right) t^k$$

converges to $x(t)$ for $|t| < \rho$. Since the coefficients of p_k are positive, to show the convergence it is sufficient to show that

$$\sum_{k=0}^{\infty} \frac{1}{k!} p_k \left(|f(0)|, \ldots, |f^{(k-1)}(0)| \right) t^k \tag{2.17}$$

has a positive radius of convergence.

Observe that the form of the polynomial p_k is independent of f, i.e., for the ODE $\dot{y} = g(y)$, we end up with the same polynomial expression $y^{(k)}(t) = p_k \left(g(y(t)), \ldots, g^{(k-1)}(y(t)) \right)$. Thus, again making use of the fact that p_k has positive coefficients, to prove the convergence of (2.17) it suffices to prove that there is an analytic differential equation $\dot{y} = g(y)$ with an analytic solution $y(t)$, satisfying $y(0) = 0$, with the property that

$$|f^{(k)}(0)| \leq g^{(k)}(0), \quad \text{for all} \quad k \geq 0. \tag{2.18}$$

We now employ the assumption that f is analytic at 0 and choose $r > 0$ such that $[-r, r] \subset U$ and

$$\sum_{k=0}^{\infty} \frac{|f^{(k)}(0)|}{k!} r^k$$

converges to conclude that there exists $C > 0$ such that

$$|f^{(k)}(0)| \leq C k! r^{-k}, \quad \text{for all} \quad k \geq 0. \tag{2.19}$$

Observe that if we choose $g(y) = Cr/(r - y)$, then by Example 2.3.2, we have $g^{(k)}(0) = Ck!r^{-k}$ and from (2.19), we conclude that (2.18) is satisfied. In this case, the solution (2.16) of $\dot{y} = g(y)$, $y(0) = 0$ has Taylor expansion

$$\sum_{k=0}^{\infty} \frac{1}{k!} p_k \left(g(0), \ldots, g^{(k-1)}(0) \right) t^k$$

which converges for $|t| < \rho$, for some $\rho > 0$. We conclude from (2.18) that the series (2.17), and hence Taylor series of x about 0, has a positive radius of convergence. To finish the proof we need to show that the solution $x(t)$ is given by this power series. This is done by showing that the power series is a solution to (2.15). To this end let

$$\varphi(t) := \sum_{k=0}^{\infty} \frac{x^{(k)}(0)}{k!} t^k,$$

and consider the functions $\dot{\varphi}(t)$ and $f(\varphi(t))$. They are both analytic at $t = 0$ and $\varphi^{(k)}(0) = x^{(k)}(0)$ for all $k \geq 0$. From this, differentiating $f(\varphi(t))$ we get

$$(f \circ \varphi)^{(k)}(0) = x^{(k+1)}(0), \quad \text{for all} \quad k \geq 0.$$

Therefore $\dot{\varphi}(t)$ and $f(\varphi(t))$ are given by the same power series about $t = 0$, and hence they are equal, since all of their derivatives agree at $t = 0$. We now have that both $x(t)$ and $\varphi(t)$ are solutions to (2.15), and so by the uniqueness of solutions we conclude that

$$x(t) = \sum_{k=0}^{\infty} \frac{x^{(k)}(0)}{k!} t^k,$$

and then that the solution to (2.15) is analytic. □

Theorem 2.3.3 (Cauchy-Kovalevskaya Theorem: ODE version II). *Let $f\colon U \to \mathbb{R}^n$ be an analytic function defined on an open set $U \subset \mathbb{R}^n$ containing the origin. If $x\colon J \to \mathbb{R}^n$ is the solution to the IVP*

$$\dot{x} = f(x), \quad x(0) = 0,$$

then x is an analytic function in a neighborhood of 0.

The proof of this theorem is analogous to that of the one-dimensional case, hence we only present the main steps. We use multi-index notation (see Appendix **??**). First we prove the following lemma.

Lemma 2.3.4. *If $h\colon U \to \mathbb{R}^n$ is an analytic function defined on an open set $U \subset \mathbb{R}^n$ containing the origin, then there exist constants C, r, $\rho > 0$ such that*

$$h_k(z) \leq \frac{Cr}{r - \sum_{j=1}^n z_j}, \quad k = 1, \ldots, n,$$

for all $\|z\| < \rho$.

Proof. Throughout the proof α denotes a multi-index. Let

$$h_{k,\alpha} := \frac{1}{\alpha!} \partial^\alpha h_k(0).$$

The analyticity of h implies that there exists $\rho > 0$ such that for all $\|z\| < \rho$

$$h_k(z) = \sum_{|\alpha| \geq 0} h_{k,\alpha} z^\alpha, \quad k = 1, \ldots, n.$$

This implies that for fixed $0 < r < \rho$ there exists a positive constant C such that $|h_{k,\alpha}| r^{|\alpha|} \leq C$ for all α and k, and in particular,

$$|h_{k,\alpha}| \leq C r^{-|\alpha|} \leq C \frac{|\alpha|!}{\alpha!} r^{-|\alpha|}.$$

Therefore, for $\|z\| < \rho$, applying multinomial theorem (see Appendix **??**) leads to

$$\begin{aligned} h_k(z) &\leq \sum_{|\alpha|\geq 0} C\frac{|\alpha|!}{\alpha!} r^{-|\alpha|} z^\alpha \\ &= C\sum_{k=0}^{\infty}\sum_{|\alpha|=k} \frac{|\alpha|!}{\alpha!}\frac{z^\alpha}{r^{|\alpha|}} \\ &= C\sum_{k=0}^{\infty}\left(\frac{z_1+\cdots+z_n}{r}\right)^k \\ &= C\frac{1}{1-\left(\frac{z_1+\cdots+z_n}{r}\right)} \\ &= \frac{Cr}{r-\sum_{j=1}^n z_j}. \end{aligned}$$

□

Example 2.3.5. Direct substitution shows that the unique solution to the IVP

$$\dot{y}_k = g_k(y) = \frac{Cr}{r-\sum_{j=1}^n y_j}, \quad y(0) = 0, \quad k = 1,\ldots,n,$$

where C, $r > 0$, is given by

$$y_k(t) = \frac{r}{n} - \sqrt{\left(\frac{r}{n}\right)^2 - \frac{2Cr}{n}t},$$

and this is an analytic function for $|t| < \frac{r}{2Cn}$.

Proof of Theorem 2.3.3. Since f is analytic, we know that the solution x to the IVP is C^∞. Repeated differentiation produces

$$\begin{aligned} \dot{x}_j(t) &= f_j(x(t)) \\ \ddot{x}_j(t) &= \sum_{m=1}^n \frac{\partial f_j}{\partial x_m}(x(t))\dot{x}_m(t) \\ &\vdots \\ x_j^{(k)}(t) &= q_k\left(\left\{\partial^\alpha f_j(x(t))\right\}_{|\alpha|<k}, \left\{x_m^{(\ell)}(t)\right\}_{\ell<k,1\leq m\leq n}\right) \end{aligned}$$

where q_k is a polynomial with positive integer coefficients. By assumption $x(0) = 0$, thus

$$x_j^{(k)}(0) = q_k\left(\left\{\partial^\alpha f_j(0)\right\}_{|\alpha|<k}, \left\{x_m^{(\ell)}(0)\right\}_{\ell<k,1\leq m\leq n}\right).$$

Applying this formula by induction on k, we can eliminate all the lower derivatives $x_m^{(\ell)}(0)$ from the right hand side to get

$$x_j^{(k)}(0) = p_k\left(\{\partial^\alpha f_j(0)\}_{|\alpha|<k}\right),$$

where p_k is again a polynomial with positive integer coefficients.

Using the fact that p_k has positive coefficients, and then Lemma 2.3.4, we obtain

$$\begin{aligned} x_j^{(k)}(0) &= p_k\left(\{\partial^\alpha f_j(0)\}_{|\alpha|<k}\right) \\ &\leq p_k\left(\{|\partial^\alpha f_j(0)|\}_{|\alpha|<k}\right) \\ &\leq p_k\left(\{\partial^\alpha g_j(0)\}_{|\alpha|<k}\right) \\ &= y_j^{(k)}(0), \end{aligned}$$

if g and y are chosen as in Example 2.3.5. Since y is analytic it follows, as in the proof of Theorem 2.3.1, that x is analytic. □

2.4 Dynamical Aspects of ODEs

For the most part the focus of the previous sections is on the local existence of solutions to differential equations. In this section we focus on the asymptotic behavior, which leads us to a more geometric perspective.

Definition 2.4.1. Let $f\colon U \to \mathbb{R}^n$ be a Lipschitz continuous function defined on an open set $U \subset \mathbb{R}^n$. Let $\varphi(\cdot, x_0)\colon J \to \mathbb{R}^n$ be the maximal solution to the IVP

$$\dot{x} = f(x), \quad x(0) = x_0. \tag{2.20}$$

The *orbit* through x_0 is the set $\gamma(x_0) := \varphi(J, x_0) \subset \mathbb{R}^n$.

The dynamical systems perspective on ODEs is to focus on the geometry of orbits and/or sets of orbits. To simplify the discussion we restrict our attention to autonomous differential equations and assume that solutions exist for all time. From a geometric perspective this latter assumption is not a constraint as it can be achieved via a simple rescaling of time. To be more specific, in Definition 2.4.1 it is possible that $J \neq \mathbb{R}$. However, if we consider for $f\colon \mathbb{R}^n \to \mathbb{R}^n$ the following modified IVP,

$$\dot{x} = \frac{f(x)}{1 + \|f(x)\|^2}, \quad x(0) = x_0,$$

then by Theorem 2.2.18 the maximal solution takes the form $\psi(\cdot, x_0)\colon \mathbb{R} \to \mathbb{R}^n$. Furthermore, the reader can check that both differential equations produce the same orbits, that is,

$$\varphi(J, x_0) = \psi(\mathbb{R}, x_0).$$

More generally we have the following theorem.

Theorem 2.4.2 (Reparameterization of time). *Let $U \subset \mathbb{R}^n$ be an open set. Assume $f\colon U \to \mathbb{R}^n$ is C^k (with $k = \infty$ allowed). Given $x_0 \in U$, let $\varphi(\cdot, x_0)\colon J_{x_0} \to \mathbb{R}^n$ be the maximal solution for the IVP (2.20). Denote by $\gamma(x_0)$ the orbit through x_0. Then there exists a C^k function $g\colon U \to (0,1]$ such that $\mathbb{R}$ is the maximal domain of existence for any solution to*

$$\dot{x} = g(x)f(x).$$

Moreover, for each $x_0 \in U$, denote by $\psi(\cdot, x_0)\colon \mathbb{R} \to \mathbb{R}^n$ the solution for the initial value problem $\dot{x} = f(x)g(x)$, $x(0) = x_0$. Then,

$$\psi(t, x_0) = \varphi(\tau(t, x_0), x_0),$$

where $\tau(\cdot, x_0)\colon \mathbb{R} \to J_{x_0}$ satisfies $\tau(0, x_0) = 0$, and solves the differential equation

$$\dot{\tau}(t, x_0) = g\big(\varphi(\tau(t, x_0), x_0)\big) > 0.$$

Hence, $\psi(\cdot, x_0)$ is a reparameterization of $\varphi(\cdot, x_0)$ with the same oriented solution curves, yielding the same orbit $\gamma(x_0) = \varphi(J_{x_0}, x_0) = \psi(\mathbb{R}, x_0)$.

Proof. If $U = \mathbb{R}^n$, let $g(x) := \frac{1}{1+\|f(x)\|^2}$, with $\|\cdot\|$ the Euclidean norm (hence $\|\cdot\|^2$ is a C^∞ function from $\mathbb{R}^n$ to $[0, \infty)$). Then $g\colon U \to (0,1]$ is a C^k function, and $\|g(x)f(x)\| \leq 1$ for all $x \in U$. From Theorem 2.2.18 part 2, any solution to $\dot{x} = g(x)f(x)$ is defined on $\mathbb{R}$.

In case $U \neq \mathbb{R}^n$, then consider a function $G\colon U \to (0,1]$ to be a C^∞ function such that $\sup_{x\in U} \|DG(x)\|_M \leq 1$ (where $\|\cdot\|_M$ denotes the matrix norm induced by $\|\cdot\|$) and such that $\|G(x)\|$ approaches 0 as x goes to the boundary of U or as $\|x\|$ goes to infinity.

In this case, let $g(x) := \frac{G(x)^2}{1+\|f(x)\|^2}$, which is a C^k function, and let $F(x) := g(x)f(x)$. Then, $\|F(x)\| \leq |G(x)|^2 \leq 1$ for all $x \in U$. Now, consider $x_0 \in U$ and denote by $\psi(t, x_0)$ the unique solution of the IVP $\dot{x} = F(x)$, $x(0) = x_0$. From Theorem 2.2.18 part 1, to show that $\psi(t, x_0)$ is defined on $\mathbb{R}$, it is enough to show that $G(\psi(t, x_0))$ does not go to 0 in finite time, or equivalently that $\frac{1}{G(\psi(t,x_0))}$ does not go to infinity in finite time. Now,

$$\frac{d}{dt}\frac{1}{G(\psi(t, x_0))} = -\frac{1}{G(\psi(t, x_0))^2} DG(\psi(t, x_0))F(\psi(t, x_0))$$

and therefore,

$$\begin{aligned}\frac{1}{G(\psi(t, x_0))} - \frac{1}{G(\psi(0, x_0))} &= -\int_0^t \frac{1}{G(\psi(s, x_0))^2} DG(\psi(s, x_0))F(\psi(s, x_0))\, ds \\ &= -\int_0^t DG(\psi(s, x_0))\frac{f(\psi(s, x_0))}{1 + \|f(\psi(s, x_0))\|^2}\, ds.\end{aligned}$$

This implies that

$$\left|\frac{1}{G(\psi(t, x_0))}\right| \leq \left|\frac{1}{G(x_0)}\right| + \int_0^{|t|} ds = \left|\frac{1}{G(x_0)}\right| + |t|,$$

and therefore $\frac{1}{G(\psi(t,x_0))}$ does not go to infinity in finite time. Hence, $\psi(t,x_0)$ is defined for all $t \in \mathbb{R}$.

Now, let $\tau(t,x_0) \in \mathbb{R}$ the unique solution of the IVP

$$\dot{\tau}(t,x_0) = g\big(\varphi(\tau(t,x_0),x_0)\big), \quad \tau(0,x_0) = 0.$$

Then, $\varphi(\tau(t,x_0),x_0)$ satisfies $\varphi(\tau(0,x_0),x_0) = \varphi(0,x_0) = x_0$ and

$$\begin{aligned}\frac{d}{dt}\varphi(\tau(t,x_0),x_0) &= f\big(\varphi(\tau(t,x_0),x_0)\big)\dot{\tau}(t,x_0)\\ &= f\big(\varphi(\tau(t,x_0),x_0)\big)g\big(\varphi(\tau(t,x_0),x_0)\big)\\ &= F\big(\varphi(\tau(t,x_0),x_0)\big).\end{aligned}$$

By unicity of the solutions of the IVP $\dot{x} = F(x)$, $x(0) = x_0$, we obtain that $\psi(t,x_0) = \varphi(\tau(t,x_0),x_0)$. □

With Theorem 2.4.2 as justification for the remainder of this chapter we work with autonomous differential equations for which the solutions exist for all time. This allows us to encode all the dynamics in the form of a continuous map.

Definition 2.4.3. Let X be a topological space. A continuous map $\varphi\colon \mathbb{R}\times X \to X$ is a *flow* if

(i) $\varphi(0,x) = x$

(ii) $\varphi(t,\varphi(s,x)) = \varphi(t+s,x)$

for all $x \in X$ and $t,s \in \mathbb{R}$. The space X is called the *phase space* for the flow.

For the sake of simplicity for the remainder of this book we will always assume that the phase space X of a flow is a subset of $\mathbb{R}^n$.

Observe that Theorem 2.2.18 in combination with Corollary 2.2.10 implies that if $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is bounded Lipschitz continuous function, then the solutions to the differential equation $\dot{x} = f(x)$ define a flow on $\mathbb{R}^n$. Using [1, Theorem 1.261] and extensions thereof one can prove that if $f \in C^r$, then the flow $\varphi\colon \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ is also C^r.

With the language of flows we can generalize the concept of an orbit.

Definition 2.4.4. A set $S \subset X$ is an *invariant set* for the flow $\varphi\colon \mathbb{R}\times X \to X$ if $\varphi(\mathbb{R},S) = S$.

Observe that every orbit is an invariant set and any invariant set is the union of orbits. Furthermore, if S is an invariant set for φ then $\varphi(t,S) = S$ for all $t \in \mathbb{R}$.

Slightly weaker, but useful notions of invariance include the following.

Definition 2.4.5. A set $S \subset X$ is *forward invariant* if $\varphi([0,\infty),S) = S$ and is *backward invariant* if $\varphi((-\infty,0],S) = S$.

Observe that an invariant set is both forward and backward invariant.

Definition 2.4.6. A point $x \in X$ is a *equilibrium point* if $\varphi(\mathbb{R}, x) = x$ or equivalently if $\varphi(t, x) = x$ for all $t \in \mathbb{R}$.

Observe that if the flow φ is generated by an ODE $\dot{x} = f(x)$, then x_0 is an equilibrium point if and only if $f(x_0) = 0$.

Example 2.4.7. The *logistic equation* is used in biology as a simple model for population growth that includes overcrowding effects. It takes the form

$$\dot{x} = rx(K - x) \tag{2.21}$$

where $r > 0$ represents the birth rate and $K > 0$ is the carrying capacity for the environment. Explicit solutions to this equation take the form

$$\varphi(t, x_0) = \frac{x_0 K e^{rKt}}{K - x_0 + x_0 e^{rKt}}. \tag{2.22}$$

Observe that φ is not a flow on the phase space $\mathbb{R}$. If $x_0 < 0$ or $x_0 > K$, then $|\varphi(t, x_0)| \to \infty$ as $t \to \frac{1}{rK} \ln\left(\frac{x_0 - K}{x_0}\right)$. However, from the biological perspective, the population levels of interest lie in the interval $[0, K]$ and the restriction $\varphi\colon \mathbb{R} \times [0, K] \to [0, K]$ is a flow.

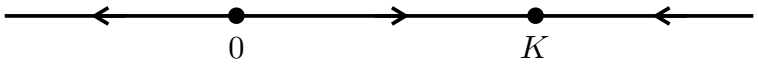

Figure 2.1: The phase portrait of the logistic equation (2.21) for $r, K > 0$.

The dynamics of the logistic equation restricted to $[0, K]$ is particularly simple (e.g. see Figure 2.1). There are three orbits: two equilibrium points and the orbit $(0, K)$ along which the dynamics moves from 0 to K. This suggests that it is worth naming the latter type of orbit.

Definition 2.4.8. A point $x_0 \in X$ is a *heteroclinic point* of a flow $\varphi\colon \mathbb{R} \times X \to X$ if

$$\lim_{t \to \infty} \varphi(t, x_0) = x_+ \quad \text{and} \quad \lim_{t \to -\infty} \varphi(t, x_0) = x_-$$

where $x_- \neq x_+$ are equilibria. The orbit $\varphi(\mathbb{R}, x_0)$ is called a *heteroclinic orbit* from x_- to x_+. In case $x_- = x_+$ we say that x_0 is a *homoclinic point* and $\varphi(\mathbb{R}, x_0)$ is called a *homoclinic orbit*

We leave the proof of the following proposition as an exercise.

Proposition 2.4.9. *If $f\colon \mathbb{R} \to \mathbb{R}$ is locally Lipschitz continuous, then every bounded solution to*

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

is either an equilibrium or a heteroclinic orbit.

Definition 2.4.10. A point $x \in X$ is a *periodic point* if there exists $T > 0$ such that $\varphi(T, x) = x$. The associate *periodic orbit* is $\varphi([0, T], x)$.

Observe that a periodic orbit is an invariant set, since by Definition 2.4.3.2 $\varphi(\mathbb{R}, x) = \varphi([0, T], x)$.

Example 2.4.11. Consider the system of differential equations

$$\begin{cases} \dot{x_1} = -x_2 + \lambda x_1(K^2 - x_1^2 - x_2^2) \\ \dot{x_2} = x_1 + \lambda x_2(K^2 - x_1^2 - x_2^2). \end{cases}$$

Changing to polar coordinates results in the system

$$\begin{cases} \dot{\theta} = 1 \\ \dot{r} = \lambda r(K^2 - r^2). \end{cases}$$

Observe that the circle $r = |K|$ or equivalently $\Gamma := \left\{x \in \mathbb{R}^2 \;\middle|\; \|x\| = |K|\right\}$ is a periodic orbit. Since the equation in r is a scalar differential equation, by Proposition 2.4.9 if $r_0 \in (0, K)$, then r_0 is a heteroclinic point. The associated heteroclinic orbit goes from 0 to K and thus solutions spiral away from the origin toward the periodic orbit.

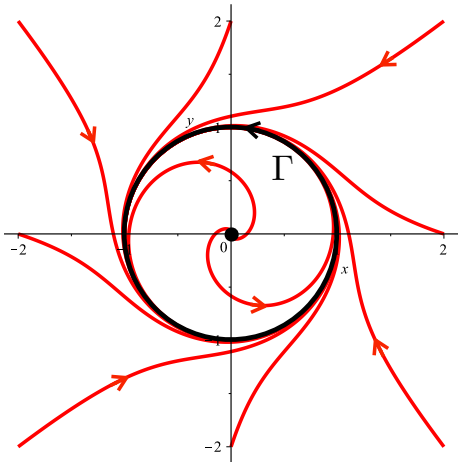

Figure 2.2: The phase portrait of the equation of Example 2.4.11 with $K = \lambda = 1$.

Example 2.4.11 suggests that we want to be able to discuss limits of trajectories that are not single points. We begin with a general definition.

Definition 2.4.12. Let $\varphi\colon \mathbb{R} \times X \to X$ be a flow and let $U \subset X$. The *alpha* and *omega limit sets* of U are defined by

$$\alpha(U) = \alpha(U, \varphi) := \bigcap_{T \leq 0} \overline{\varphi\big((-\infty, T], U\big)} \quad \text{and} \quad \omega(U) = \omega(U, \varphi) := \bigcap_{T \geq 0} \overline{\varphi\big([T, \infty), U\big)},$$

respectively.

In a slight abuse of notation, if $x \in X$ then we will let $\omega(x) = \omega(\{x\})$ and $\alpha(x) = \alpha(\{x\})$. As an exercise we let the reader check that if x_0 is a heteroclinic point from x_- to x_+, then $\alpha(x_0) = x_-$ and $\omega(x_0) = x_+$.

Proposition 2.4.13. *Let $\varphi\colon \mathbb{R} \times X \to X$ be a flow and $U, V \subset X$. We have:*

(i) If there exists $t \in \mathbb{R}$ such that $\varphi(t, U) = V$, then

$$\omega(U) = \omega(V) \quad \text{and} \quad \alpha(U) = \alpha(V).$$

(ii) A point $y \in \omega(U, \varphi)$ if and only if there exists a sequence of times $t_k \to \infty$ and a sequence of points $x_k \in U$ such that

$$\lim_{k \to \infty} \varphi(t_k, x_k) = y.$$

Furthermore, $\omega(U) = \omega(\varphi([0, \infty), U))$.

(iii) Let V be a closed, forward invariant set. If $U \subset V$, then $\omega(U) \subset V$.

(iv) If U is forward invariant, then

$$\omega(U) = \bigcap_{t \geq 0} \overline{\varphi(t, U)}. \tag{2.23}$$

(v) $\omega(U)$ and $\alpha(U)$ are closed invariant sets. Furthermore, if U is a closed invariant set, then

$$U = \omega(U) = \alpha(U).$$

(vi) If $\varphi([0, \infty), U) \subset K$ where $K \subset X$ is compact, then $\omega(U)$ is nonempty and compact. If in addition, U is connected, then $\omega(U)$ is connected.

(vii) If $V \subset \omega(U)$, then $\omega(V) \subset \omega(U)$ and $\alpha(V) \subset \omega(U)$.

Similar results apply to α limit sets with regard to negative time and backward invariance.

Proof. (i). This follows from the definition of a flow and the definition of the alpha and the omega limit sets.

(ii). Assume $y \in \omega(U)$. By definition $y \in \overline{\varphi\big([T,\infty),U\big)}$ for all $T \geq 0$. In particular, for each $k \in \mathbb{Z}_+$ there exist $s_k \geq 0$, $x_k \in U$ with $y_k := \varphi(k+s_k, x_k) \in \varphi\big([k,\infty),U\big)$ such that $\|y_k - y\| \leq \frac{1}{k}$. Let $t_k := k + s_k$. Then $\lim_{k\to\infty} \varphi(t_k, x_k) = y$.

Conversely, assume there exists a sequence $t_k \to \infty$ and points $x_k \in U$ such that $\lim_{k\to\infty} \varphi(t_k, x_k) = y$. Then $y \in \overline{\varphi([t,\infty),U)}$ for all $t \geq 0$ and hence $y \in \omega(U)$.

The final remark follows from the observation that $\varphi([t,\infty),U) = \varphi(t, \varphi([0,\infty),U))$.

(iii). Since $U \subset V$ and V is forward invariant

$$\varphi([0,\infty),U) \subset \varphi([0,\infty),V) \subset V.$$

Since V is closed $\overline{\varphi([0,\infty),U)} \subset V$ and hence $\omega(U) \subset V$.

(iv). Since U is forward invariant

$$\varphi(t+s,U) \subset \varphi(t,\varphi(s,U)) \subset \varphi(t,U)$$

for all $t, s \geq 0$. Therefore, $\varphi([t,\infty),U) = \varphi(t,U)$, from which (2.23) follows.

(v). By definition, alpha and omega limit sets are defined in terms of intersections of closed sets and hence are closed.

We prove that $\omega(U)$ is forward and backward invariant, which implies that it is invariant. As a preliminary step we show that if U is forward invariant, then $\omega(U)$ is forward invariant. Observe that by (iv) for all $t \geq 0$,

$$\begin{aligned}
\varphi(t,\omega(U)) &= \varphi\left(t, \bigcap_{s\geq 0} \overline{\varphi(s,U)}\right) \\
&\subset \bigcap_{s\geq 0} \varphi\left(t, \overline{\varphi(s,U)}\right) \\
&\subset \bigcap_{s\geq 0} \overline{\varphi\left(t,\varphi(s,U)\right)} \\
&\subset \bigcap_{s\geq 0} \overline{\varphi\left(s,\varphi(t,U)\right)} \\
&= \omega(\varphi(t,U)) \\
&= \omega(U)
\end{aligned}$$

where the inclusions follow from the fact that φ is continuous and the last equality follows from (i). To finish the proof we note that $\varphi([0,\infty),U)$ is forward invariant. Thus

$$\varphi(t,\omega(U)) = \varphi(t,\varphi([0,\infty),U)) = \omega(\varphi([0,\infty),U)) = \omega(U)$$

where the last equality follows from (ii). This shows that $\omega(U)$ is forward invariant. A similar argument shows that U is backward invariant.

If U is a closed invariant set, then U is forward invariant. Thus by (iv)

$$U = \bigcap_{t\geq 0} \varphi(t, U) \subset \bigcap_{t\geq 0} \overline{\varphi(t, U)} = \omega(U).$$

However, since U is a closed forward invariant set by (ii), $\omega(U) \subset U$.

(vi). Since for every $T \geq 0$, $\varphi([T, \infty), U) \subset \varphi([0, \infty), U) \subset K$ and K is compact, then $\overline{\varphi([T, \infty), U)} \subset K$ is compact. By definition this implies that $\omega(U)$ is the intersection of a nested collection of non empty compact sets and so is nonempty and compact.

Now assume that U is connected. Then $\varphi([T, \infty), U)$ is connected and thus $\overline{\varphi([T, \infty), U)}$ is connected. Again, since $\omega(U)$ is the intersection of a nested collection of compact connected sets, it is connected.

(vii). Since $\omega(U)$ is a closed invariant set, by (iii) $\omega(V) \subset \omega(U)$. The corresponding result of 3 for alpha limit sets shows $\alpha(V) \subset \alpha(U)$. □

It is important to observe that $\alpha(U)$ and $\omega(U)$ only describe the asymptotic dynamics of points in U and ignores that of nearby points. Returning to Example 2.4.11 observe that $\omega(0) = 0$, but $\omega(y) = \left\{x \in \mathbb{R}^2 \;\middle|\; \|x\| = |K|\right\}$ for all $y \in \mathbb{R}^2 \setminus \{0\}$. To discuss sets whose asymptotic dynamics is in agreement with that of its neighbors we introduce the following concepts.

Definition 2.4.14. Let $\varphi\colon \mathbb{R} \times X \to X$ be a flow. A forward invariant set U is a *trapping region* if there exists $T > 0$ such that

$$\overline{\varphi(T, U)} \subset \operatorname{int}(U).$$

Dually, a backward invariant set U is a *repelling region* if there exists $T < 0$ such that

$$\overline{\varphi(T, U)} \subset \operatorname{int}(U).$$

Definition 2.4.15. Let $\varphi\colon \mathbb{R} \times X \to X$ be a flow on a locally compact metric space. A set $A \subset X$ is an *attractor* if there exists a trapping region U such that $A = \omega(U)$. A set $R \subset X$ is a *repeller* if there exists a repelling region U such that $R = \alpha(U)$.

Definition 2.4.16. Let $\varphi\colon \mathbb{R} \times X \to X$ be a flow on a locally compact metric space. The *maximal invariant set* in U is defined by

$$\operatorname{Inv}(U, \varphi) := \{x \in U \;\mid\; \varphi(\mathbb{R}, x) \subset U\}.$$

We leave the proof of the following proposition as an exercise.

Proposition 2.4.17. *If U is a trapping region and $A = \omega(U)$, then A is the maximal invariant set in U.*

In applications it can be difficult to determine a trapping region, because they must be forward invariant. A seemingly weaker notion is the following.

Definition 2.4.18. Let $\varphi\colon \mathbb{R} \times X \to X$ be a flow on a locally compact metric space. A compact set N is called an *attracting neighborhood* if

$$\omega(N) \subset \operatorname{int}(N).$$

Dually, a compact set N is a *repelling neighborhood* if $\alpha(N) \subset \operatorname{int}(N)$.

In general attracting neighborhoods are easier to identify than trapping regions (see Exercise 4.7.3). However, as the following results indicate they are closely related.

Proposition 2.4.19. *If U is a trapping region, then $\overline{U}$ is an attracting neighborhood. If N is an attracting neighborhood, then there exists a trapping region U such that $\omega(N) \subset \overline{U} \subset N$.*

The proof of Proposition 2.4.19 follows from the continuity of the flow. However, the argument is somewhat technical thus we do not present them here.

Corollary 2.4.20. *A is an attractor for a flow if and only if there exists an attracting neighborhood N such that $A = \omega(N)$.*

Corresponding results hold for repellers. Attractors and repellers provide a powerful tool for decomposing dynamics.

Definition 2.4.21. Let A be an attractor for a flow $\varphi\colon \mathbb{R} \times X \to X$ on a compact metric space. The *dual repeller* to A is

$$A^* := \{x \in X \mid \omega(x) \cap A = \emptyset\}.$$

(A, A^*) is called an *attractor-repeller pair decomposition* for φ.

Lemma 2.4.22. *Let (A, A^*) be an attractor repeller pair decomposition, then*

$$A \cap A^* = \emptyset.$$

Proof. Since A is an attractor there exists a trapping region N such that $A = \omega(N) \subset \operatorname{int}(N)$. Observe that this implies that for any $y \in N$, $\omega(y) \subset A$. Hence $A^* \cap N = \emptyset$ and therefore $A \cap A^* \subset \operatorname{int}(N) \cap A^* \subset N \cap A^* = \emptyset$. □

Returning to the logistic equation in Example 2.4.7 and the induced flow $\varphi\colon \mathbb{R} \times [0, K] \to [0, K]$ observe that $\{K\}$ is an attractor. Its dual repeller is $\{0\}$, thus $(\{K\}, \{0\})$ forms an attractor-repeller pair decomposition for φ. Observe that contrary to the name, this is not a decomposition of the phase space $[0, K]$. The justification for calling it a decomposition is made clear in Theorem 2.4.24. It is also worth noting that $\{0\}$ is a repeller. As the following result, which is left as an exercise, indicates this is true in general.

Proposition 2.4.23. *If A is an attractor for a flow $\varphi\colon \mathbb{R}\times X\to X$ on a compact metric space, then its dual repeller A^* is a repeller for φ.*

Theorem 2.4.24. *Let $\varphi\colon \mathbb{R}\times X\to X$ be a flow on a compact metric space. Let (A, A^*) be an attractor-repeller pair decomposition for φ. If $x\in X\setminus(A\cup A^*)$, then*

$$\omega(x)\subset A \quad \textit{and} \quad \alpha(x)\subset A^*.$$

Proof. Since A is an attractor there exists a trapping region N such that $A=\omega(N)\subset \operatorname{int}(N)$. Observe that this implies that for any $y\in N$, $\omega(y)\subset A$.

We want to prove that $\omega(x)\subset A$. By definition if $x\notin A^*$, then $\omega(x)\cap A\neq\emptyset$. Therefore, by Proposition 2.4.13(i) there exists $t>0$ such that $\varphi(t,x)=y\in N$. By Proposition 2.4.13(ii), $\omega(x)=\omega(y)\subset A$.

We want to prove that $\alpha(x)\subset A^*$. We first show that $\alpha(x)\cap A=\emptyset$. Assume not. Then, by the analogue of Proposition 2.4.13(i) for alpha limit sets, there exists a sequence of times $t_k\to-\infty$ such that

$$\lim_{k\to\infty}\varphi(t_k,x)=\lim_{k\to\infty}y_k=y\in A.$$

Thus without loss of generality we can assume that $y_k\in N$ for all k. Since N is a trapping region $\varphi([0,\infty),y_k)\subset N$. Since this is true for all t_k, we can conclude that $\varphi(\mathbb{R},x)\subset N$. This implies that x belongs to the maximal invariant set in N and hence, by Proposition 2.4.17 that $x\in A$. Therefore, $\alpha(x)\cap A=\emptyset$.

By the analogue of Proposition 2.4.13(iv) for alpha limit sets, $\alpha(x)\neq\emptyset$. Assume $y\in\alpha(x)$. Since $\alpha(x)$ is invariant, $\omega(y)\subset\alpha(x)$. This implies that $\omega(y)\cap A=\emptyset$ and hence by definition $y\in A^*$. □

Attractor-repeller pair decompositions for the dynamics of Examples 2.4.7 and 2.4.11 are reasonably easy to describe: $(\{K\},\{0\})$ and $\left(\left\{x\in\mathbb{R}^2 \;\middle|\; \|x\|=|K|\right\},\{0\}\right)$, respectively. A significant feature distinguishing the systems is that in the first case the attractor is an equilibrium, while in the second case it is a periodic orbit. It should also be observed that both systems depend on two parameters, r and K for the logistics equation and λ and K for Examples 2.4.11, and that as one changes these parameters the solutions to the differential equations change. This is made explicit by (2.22). This raises the question of what level of refinement do we want to use to distinguish between the dynamics of different systems of differential equations.

Definition 2.4.25. Two flows $\varphi\colon \mathbb{R}\times X\to X$ and $\psi\colon \mathbb{R}\times Y\to Y$ are *topologically equivalent* if there exists a homeomorphism $h\colon X\to Y$ such that orbits of φ are mapped onto orbits of ψ preserving the direction of time, that is, there exists a continuous and strictly increasing *time-rescaling map* $\tau\colon \mathbb{R}\times X\to\mathbb{R}$ such that

$$h(\varphi(t,x))=\psi(\tau(t,x),h(x)),$$

for all $(t,x)\in\mathbb{R}\times X$.

Observe that the dynamics of Example 2.4.7 and Example 2.4.11 are not topologically equivalent. Consider a flow $\varphi\colon \mathbb{R}\times[0,K]\to[0,K]$ (resp. $\psi\colon \mathbb{R}\times[0,K]\to[0,K]$) generated by Example 2.4.7 with $r, K > 0$ (resp. $r < 0 < K$). Then, as one can see in Figure 2.4, the flows φ and ψ are not topologically equivalent. However, for all $r > 0$ and $K > 0$ all flows generated by Example 2.4.7 are topologically equivalent. The same is true for all $\lambda > 0$ and $K \neq 0$ in Example 2.4.11. Demonstrating these last three statement can be done because of the extremely simple form of the equations.

Figure 2.3: Phase portraits of the logistic equation (2.21) for $r, K > 0$ (left) and for $r < 0 < K$ (right). The flows defined on the same phase space $X = [0, K]$ are not topologically equivalent.

In general showing that two ODEs generate topologically equivalent dynamics is extremely challenging. We leave it to the reader to check that given a differential equation $\dot{x} = f(x)$, if $\psi_i\colon \mathbb{R}\times\mathbb{R}^n\to\mathbb{R}^n$, $i = 0, 1$, are flows generated by $\dot{x} = g_i(x)f(x)$ for different rescaling functions g_i as in Theorem 2.4.2, then ψ_0 and ψ_1 are topologically equivalent. Thus under this equivalence class there is no ambiguity about discussing the flow generated by a differential equation.

We introduced the concept of an attractor by arguing that we should be interested in the asymptotic dynamics of neighborhoods about a set of initial conditions. In the context of fixed points and periodic orbits we will return to this question repeatedly and in greater generality than just that of an attractor. However, before leaving the subject we introduce some fundamental definitions. Consider the an autonomous differential equation

$$\dot{x} = f(x) \tag{2.24}$$

where $f\colon U\to\mathbb{R}^n$ is locally Lipschitz continuous on an open set U and let φ denote the associated flow.

Definition 2.4.26. An equilibrium $\tilde{x}$ of (2.24) is *stable* if for any $\epsilon > 0$, there exists $\delta > 0$ such that for all $t > 0$, if $\|x_0 - \tilde{x}\| < \delta$, then $\|\varphi(t, x_0) - \tilde{x}\| < \epsilon$. An equilibrium that is not stable is called *unstable*.

A stronger notion of stability is the following.

Definition 2.4.27. An equilibrium $\tilde{x}$ of (2.24) is *asymptotically stable* if it is stable and there exists $\rho > 0$ such that if $\|x_0 - \tilde{x}\| < \rho$, then $\lim_{t\to\infty}\varphi(t, x_0) = \tilde{x}$.

The relationship between the concepts of stability and attractors is subtle as the following examples and propositions indicate.

Example 2.4.28. Consider the harmonic oscillator

$$\begin{cases} \dot{x_1} = x_2 \\ \dot{x_2} = -x_1 \end{cases} \tag{2.25}$$

The unique equilibrium is the origin and the remaining orbits are periodic orbits that take the form of concentric circles. The origin is a stable equilibrium where one can choose $\delta = \epsilon$. Observe that the origin 0 is not an attractor; given any neighborhood N of the origin the maximal invariant set in N will contain periodic orbits and thus $\omega(N) \neq \{0\}$.

The proof of the following result is nontrivial.

Proposition 2.4.29. *Let $X \subset R^n$ be a compact neighborhood of $\tilde{x}$. An equilibrium $\tilde{x}$ is an attractor for the flow $\varphi\colon \mathbb{R}\times X \to X$ if and only if $\tilde{x}$ is an asymptotically stable equilibrium.*

2.5 Exercises

Exercise 2.5.1 (Existence via Picard's method). An alternative proof for the existence of solutions is given by Picard's method. In this exercise we use Picard's method to prove the existence of a solution to the *non autonomous* IVP

$$\dot{x} = f(x,t), \quad x(t_0) = x_0.$$

We begin with some hypotheses.

Consider an open set $U \subset \mathbb{R}^n$ and an open interval $I \subset \mathbb{R}$ which contains t_0. Denote $D := U \times I$. Assume that $f\colon D \to \mathbb{R}^n$ is continuous and that for all $t \in I$, $f(\cdot,t)\colon U \to \mathbb{R}^n$ is a locally Lipschitz continuous function.

The goal is to show that there exists a solution $\varphi\colon J = (t_0 - a, t_0 + a) \to \mathbb{R}^n$ for some $a > 0$ to the integral equation

$$\varphi(t) = x_0 + \int_{t_0}^{t} f(\varphi(s), s)\, ds, \quad \text{for all} \quad t \in J. \tag{2.26}$$

Let $\epsilon, \delta > 0$ small enough so that $D_{\epsilon,\delta} := \{(x,t) \mid \|x - x_0\| \leq \epsilon, |t - t_0| \leq \delta\} \subset D$ and so that

$$\|f(x,t) - f(y,t)\| \leq K\|x - y\|, \quad \text{for all} \quad (x,t),(y,t) \in D_{\epsilon,\delta}. \tag{2.27}$$

Let

$$M_{\epsilon,\delta} := \max_{(x,t)\in D_{\epsilon,\delta}} \|f(x,t)\|, \tag{2.28}$$

and let $a > 0$ such that

$$a \leq \min\left\{\delta, \frac{\epsilon}{M_{\epsilon,\delta}}\right\}. \tag{2.29}$$

Define a time interval by

$$J := (t_0 - a, t_0 + a). \tag{2.30}$$

For any $t \in J$, define the *Picard operator* by

$$T(x)(t) = x_0 + \int_{t_0}^{t} f(x(s), s)\, ds. \tag{2.31}$$

(i) Show that $(T(x)(t), t) \in D_{\epsilon,a}$, for every $(x, t) \in D_{\epsilon,a}$.

This allows us to define a sequence of functions $\{x^n\}_{n\geq 0}$ with $x^n \colon J \to \mathbb{R}^n$ by

$$x^0(t) \equiv x_0, \quad x^{n+1}(t) = x_0 + \int_{t_0}^{t} f(x^n(s), s)\, ds, \quad n \geq 0. \tag{2.32}$$

The iterations (2.32) define what is known as the *Picard iterative process.*

(ii) Show that for every $t \in J = (t_0 - a, t_0 + a)$,

$$\|x^n(t) - x^{n-1}(t)\| \leq M_{\epsilon,\delta} K^{n-1} \frac{|t - t_0|^n}{n!}.$$

(iii) Show that $\{x^n\}_{n\geq 0}$ is a Cauchy sequence in the space $C^0(J)$ endowed with the supremum norm.

(iv) Show that there exists a continuous function $\varphi \colon J \to \mathbb{R}^n$ that is solution of the integral equation (2.26).

Exercise 2.5.2. Prove Proposition 2.2.11.

Exercise 2.5.3. Prove Proposition 2.4.9.

Exercise 2.5.4. Prove that given a flow $\varphi \colon \mathbb{R} \times X \to X$ and a set $A \subset X$

$$\bigcup_{x\in A} \omega(x, \varphi) \subset \omega(A, \varphi).$$

Provide an example for which

$$\bigcup_{x\in A} \omega(x, \varphi) \neq \omega(A, \varphi).$$

Exercise 2.5.5. Prove Proposition 2.4.23.

Exercise 2.5.6. Let $f \colon \mathbb{R}^n \to \mathbb{R}^n$ be C^1. Show that the flows generated by

$$\dot{x} = f(x) \quad \text{and} \quad \dot{x} = \frac{f(x)}{1 + \|f(x)\|}$$

are topologically equivalent.

Exercise 2.5.7. Observe that the definition of topologically equivalent in Definition 2.4.25 requires a homeomorphism $h\colon X \to Y$ as opposed to diffeomorphism. Since every diffeomorphism is a homeomorphism, our definition leads to larger equivalence classes. Using scalar differential equations provide an explanation for this choice. More explicitly, consider the family of linear scalar differential equation $\{\dot{x} = ax \mid a \in \mathbb{R}\}$.

(i) Determine the topological equivalence classes.

(ii) In the definition of topological equivalence replace the homeomorphism $h\colon X \to Y$ with the requirement that h be a diffeomorphism. Determine the equivalence classes.

The point of this exercise is to show that if we use a diffeomorphism, then the equivalence relation requires that the eigenvalues of the derivative at the equilibrium be the same.

Exercise 2.5.8 (Existence and uniqueness for linear differential equations). Consider $J \subset \mathbb{R}$ an open interval. Suppose that $g\colon J \to \mathbb{R}^n$ and $a_{i,j}\colon J \to \mathbb{R}$, $1 \leq i, j \leq n$, are continuous functions. Set

$$A(t) = \begin{pmatrix} a_{1,1}(t) & \dots & a_{1,n}(t) \\ \vdots & \ddots & \vdots \\ a_{n,1}(t) & \dots & a_{n,n}(t) \end{pmatrix}.$$

Prove that for any $t_0 \in J$ the initial value problem

$$\dot{x}(t) = A(t)x + g(t), \quad x(t_0) = x_0, \tag{2.33}$$

has a unique solution $x\colon J \to \mathbb{R}^n$.

<u>Hint</u>: Define an *integrating factor* $C(t)$ by writing down a solution of the equation

$$C'(t) = -C(t)A(t),$$

that is write down an explicit formula for such a $C(t)$. Show directly that

- $C(t)$ is well defined, continuous, and differentiable on J.
- $C(t_0) = Id$.
- $C(t)$ is invertible for every $t \in J$.

Now multiply both sides of Equation (2.33) by $C(t)$, apply the product rule on the left, integrate both sides from t_0 to t with $t \in J$, and solve for $x(t)$. This provides an explicit formula for the solution $x(t)$. Show that the formula gives a well defined, continuous, differentiable function for each $t \in J$.

Exercise 2.5.9 (Differentiability of the flow with respect to initial conditions). Let $\varphi\colon \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ be the flow generated by $\dot{x} = f(x)$, where $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is C^1 and bounded on $\mathbb{R}^n$. Proposition 2.2.9 guarantees that $\varphi(t,x)$ is continuous with respect to initial conditions. The following approach shows that φ is actually differentiable with respect to x.

Observe that differentiability of φ with respect to x is equivalent to the statement that for any $t\in\mathbb{R}$, $x_0\in\mathbb{R}^n$, there exists a unique $n\times n$ matrix $D\varphi(t,x_0)$ satisfying

$$\lim_{\|h\|\to 0} \frac{\|\varphi(t,x_0+h)-\varphi(t,x_0)-D\varphi(t,x_0)h\|}{\|h\|} = 0. \tag{2.34}$$

Let $x_0\in\mathbb{R}^n$ and $\gamma(t)\equiv\varphi(t,x_0)$ denote the orbit segment through x_0. Define the matrix $D\varphi(t,x_0)$ to be the solution of the *first variation equation*,

$$\frac{d}{dt}D\varphi(t,x_0) = Df(\gamma(t))D\varphi(t,x_0), \quad D\varphi(0,x_0) = Id.$$

Use the results of problem 2.5.8 to establish that the matrix solving this equation exists assuming only that $\phi(t,x_0)\equiv\gamma(t)$ exists and is continuous. Now show that $D\varphi(t,x_0)$ satisfies (2.34).

Exercise 2.5.10 (Flow Box Theorem). Consider the differential equation $\dot{x} = f(x)$ where $f\colon\mathbb{R}^n\to\mathbb{R}^n$ is a C^r vector field for $r\geq 1$. A point $\tilde{x}\in\mathbb{R}^n$ is called a *ordinary point* if $f(\tilde{x})\neq 0$. Prove that there exists a C^r change of coordinates $x = g(y)$ defined on a neighborhood of $\tilde{x}$ such that

$$\begin{cases} \dot{y}_1 = 1 \\ \dot{y}_2 = 0 \\ \vdots \\ \dot{y}_n = 0 \end{cases}$$

Exercise 2.5.11. Show that X with the $C^0(J)$ norm (as defined in the proof of Proposition 2.2.5) is a complete metric space, with the metric given by

$$d(x,y) = \|x-y\|_{C^0(J)} = \sup_{t\in J}\|x(t)-y(t)\|.$$

Exercise 2.5.12. Show that if x_0 is a heteroclinic point from x_- to x_+, then $\alpha(x_0) = x_-$ and $\omega(x_0) = x_+$.

Exercise 2.5.13. Give an example of a flow $\varphi\colon\mathbb{R}\times\mathbb{R}^2\to\mathbb{R}^2$ for which there exists a point $x\in\mathbb{R}^2$ such that $\omega(x)\neq\emptyset$, but $\omega(x)$ is not connected.

Exercise 2.5.14 (Lyapunov's stability theorem). Let $\tilde{x}$ be an equilibrium for $\dot{x} = f(x)$, $x \in \mathbb{R}^n$ where $f \in C^1(U, \mathbb{R}^n)$ for some open set $U \subset \mathbb{R}^n$. Let W be a neighborhood of $\tilde{x}$, $V \colon W \to \mathbb{R}$ be a continuous function that is differentiable on $W \setminus \{\tilde{x}\}$. Define

$$\dot{V}(x) := \frac{d}{dt} V(x(t)) = DV(x) \cdot \dot{x}(t) = DV(x) \cdot f(x),$$

and assume that V satisfies the following properties:

(i) $V(\tilde{x}) = 0$ and $V(x) > 0$ if $x \in W \setminus \{\tilde{x}\}$.

(ii) $\dot{V}(x) \leq 0$ for all $x \in W \setminus \{\tilde{x}\}$.

Then, $\tilde{x}$ is stable. Furthermore, if in addition $\dot{V}(x) < 0$ for all $x \in W \setminus \{\tilde{x}\}$, then $\tilde{x}$ is asymptotically stable.

Chapter 3

Equilibria and Radii Polynomials in Finite Dimension

Consider an open set $U \subset \mathbb{R}^n$ and a Lipschitz continuous function $f\colon U \to \mathbb{R}^n$. Theorem 2.2.12 guarantees the existence of a unique maximal solution to the initial value problem problem

$$\dot{x} = f(x), \quad x(0) = x_0$$

for any $x_0 \in U$. Therefore, we can now turn our attention to studying solutions with specific properties. The simplest, and hence the starting point for our investigations, is that of an *equilibrium point*, also called a *fixed point*, *steady state*, or *critical point*, i.e., a solution that is constant in time. One of our goal is to develop a constructive method to prove the existence of equilibria. Observe that $x_0 \in \mathbb{R}^n$ is an equilibrium point if and only if $f(x_0) = 0$. Thus, it is sufficient for us to provide a constructive approach for proving the existence of zeros of a function defined on a finite dimensional space X (in our case $\mathbb{R}^n$ or $\mathbb{C}^n$). More precisely, the goal of this section is twofold: prove the existence of a point $\tilde{x} \in X$ such that $f(\tilde{x}) = 0$ and provide bounds on the location of $\tilde{x}$. This is done using the radii polynomial approach which is a variant on Newton's method. With this in mind we recall Newton's method in Section 3.1 and then introduce the radii polynomial approach in finite dimension in Section 3.2.

Throughout this book, given two vectors $x, y \in \mathbb{R}^n$, we use the notation $x \ll y$ if $x_k \leq y_k$ for all $k = 1, \ldots, n$. Moreover, given a matrix $A = \{a_{ij}\}_{i,j}$ (real or complex valued) we use the notation $|A|$ to denote the matrix $|A| = \{|a_{ij}|\}_{i,j}$, where $|\cdot|$ denotes the absolute value.

3.1 Newton's Method

We begin with a trivial proposition that sets the stage for our strategy for finding zeros of a function.

Proposition 3.1.1. *Let X be $\mathbb{R}^n$ or $\mathbb{C}^n$, and let $U, V \subset X$ be open sets. Consider $f\colon U \to V$. Assume that $A\colon X \to X$ is an invertible linear map. Let $T\colon U \to X$ be defined by*

$$T(x) := x - Af(x). \tag{3.1}$$

If $T(\tilde{x}) = \tilde{x}$, then $f(\tilde{x}) = 0$.

Proposition 3.1.1 allows us to replace the problem of directly finding a zero of f with that of proving the existence of a fixed point of T. As is indicated at the beginning of Chapter 2 the contraction mapping theorem (Theorem 2.1.2) provides existence and uniqueness if T is a contraction. Furthermore, it gives bounds on the location of $\tilde{x}$ as a function of an initial guess (recall (2.1)). Thus the problems we are trying to address are reduced to finding an injective linear map A that makes T a contraction. This leads us to Newton's method. We begin with a simple example.

Example 3.1.2. Consider $f \in C^1(\mathbb{R}, \mathbb{R})$. Recall that in Newton's method, $T\colon \mathbb{R} \to \mathbb{R}$ applied to f is given by

$$T(x) := x - \frac{f(x)}{f'(x)}$$

and that T is used iteratively to find an approximate value of a root of f. More explicitly, given an initial guess $\widehat{x} \in \mathbb{R}$, set $x^0 = \widehat{x}$ and inductively define $x^{k+1} := T(x^k)$. Assume

$$\lim_{k\to\infty} x^k = \tilde{x}.$$

Observe that if T is continuous at $\tilde{x}$ (a sufficient condition for this is that $f'(\tilde{x}) \neq 0$), then $T(\tilde{x}) = \tilde{x}$ and hence $f(\tilde{x}) = 0$. Thus the problem of proving the existence of a zero of f is essentially reduced to finding and/or identifying whether an initial guess $\widehat{x} \in \mathbb{R}$ leads to convergence of Newton's method.

The existence of convergence is precisely the conclusion of the contraction mapping theorem. With this in mind assume $f(\tilde{x}) = 0$ and $f'(\tilde{x}) \neq 0$. Observe that for $|h|$ small

$$\begin{aligned}
|T(\tilde{x}+h) - T(\tilde{x})| &= \left|\tilde{x} + h - \frac{f(\tilde{x}+h)}{f'(\tilde{x}+h)} - \left(\tilde{x} - \frac{f(\tilde{x})}{f'(\tilde{x})}\right)\right| \\
&= \left|h - \frac{f(\tilde{x}+h)}{f'(\tilde{x}+h)}\right| \\
&\approx \left|h - \frac{f(\tilde{x}) + hf'(\tilde{x})}{f'(\tilde{x}+h)}\right| \\
&\approx |h| \left|1 - \frac{f'(\tilde{x})}{f'(\tilde{x}+h)}\right|.
\end{aligned}$$

Since $h = \tilde{x} + h - \tilde{x}$, the contraction constant for Newton near the fixed point $\tilde{x}$ is $\left|1 - \frac{f'(\tilde{x})}{f'(\tilde{x}+h)}\right| \approx 0$.

The intended take away message from Example 3.1.2 is that in a sufficiently small neighborhood of a nondegenerate zero of f the associated Newton operator is an extremely strong contraction. In particular, returning to the question concerning the choice of an injective linear map A, a naive interpretation of this example suggests setting $A = A(x) = (f'(x))^{-1}$. There are several reasons why this choice is not appropriate. The first is that in general $(f'(x))^{-1}$ cannot be represented using a finite binary expansion. Hence, it does not have an exact floating point representation, and therefore a faithful expression for the associated map T on a computer becomes a nontrivial task.

Returning to the general problem of finding equilibria, let $f\colon U \to \mathbb{R}^n$ be a C^1 function defined on an open set $U \subset \mathbb{R}^n$. In this case the Newton operator is given by

$$T(x) := x - (Df(x))^{-1} f(x). \tag{3.2}$$

An argument similar to that presented in Example 3.1.2 demonstrates that if $f(\tilde{x}) = 0$ and $Df(\tilde{x})$ is invertible, then in a small neighborhood of $\tilde{x}$, T is a contraction mapping with small contraction constant. Again, this suggests the choice of $A(x) = (Df(x))^{-1}$. However, the cost of computing the inverse of a $n \times n$ matrix is of order n^3 and thus for high dimensional problems repeatedly computing the inverse is prohibitively expensive.

One final caveat on the choice of A arises from the fact that for most of this book, i.e., Chapters **??** onward, we are interested in applying these techniques to maps that are defined on infinite dimensional Banach spaces for which an explicit representation of $(Df(x))^{-1}$ is not possible.

Of course, as presented in (4.24) A need not equal $(Df(x))^{-1}$. Thus the approach we adopt is as follows. We assume that we are given an initial guess $\bar{x}$ for a zero of f. In practice $\bar{x}$ is obtained using a standard numerical method. We then use some, typically problem dependent, form of approximation of $(Df(\bar{x}))^{-1}$ to choose A. This produces a function T. What remains is the challenge of proving that T is a contraction mapping.

3.2 Radii Polynomial Approach in Finite Dimension

Consider $f(x) = x^2 - 1$. The associated Newton operator $T\colon \mathbb{R} \setminus \{0\} \to \mathbb{R}$ is given by $T(x) = x - (2x)^{-1}(x^2 - 1)$. Since T has two distinct fixed points, $T(\pm 1) = \pm 1$, T cannot be a contraction mapping over its entire domain. However, the analysis of Example 3.1.2 suggests that there exists a neighborhood U^+ of 1 such that $T\colon U^+ \to \mathbb{R}$ defined by $T(x) := x - \frac{1}{2}(x^2 - 1)$ is a contraction mapping. The theorem below provides a mechanism for rigorously identifying a domain on which T is a contraction mapping.

Throughout this section we make use of the *sup norm* on $\mathbb{R}^n$, i.e., given $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ define

$$\|x\|_\infty := \max_{k=1,\ldots,n} \{|x_k|\}.$$

In this norm the closed ball of radius r centered at x is denoted by

$$\overline{B_r(x)} := \{y \in \mathbb{R}^n \mid \|x - y\|_\infty \leq r\}.$$

Theorem 3.2.1. *Let $U \subset \mathbb{R}^n$ be an open set and let $T = (T_1, \dots, T_n) \in C^1(U, \mathbb{R}^n)$, where $T_k \colon \mathbb{R}^n \to \mathbb{R}$. Let $\bar{x} \in U$. Assume that $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$ and $Z(r) = (Z_1(r), \dots, Z_n(r)) \in \mathbb{R}^n$ provide the following bounds:*

$$|T_k(\bar{x}) - \bar{x}_k| \leq Y_k \quad \text{and} \quad \sup_{b,c \in \overline{B_r(0)}} |DT_k(\bar{x} + b)c| \leq Z_k(r) \tag{3.3}$$

for all $k = 1, \dots, n$. If $\|Y + Z(r)\|_\infty < r$, then $T \colon \overline{B_r(\bar{x})} \to \overline{B_r(\bar{x})}$ is a contraction mapping with contraction constant

$$\kappa := \frac{\|Z(r)\|_\infty}{r} < 1.$$

In particular, there exists a unique $\tilde{x} \in \overline{B_r(\bar{x})}$ such that $T(\tilde{x}) = \tilde{x}$.

Proof. The mean value theorem applied to T_k implies that for any $x, y \in \overline{B_r(\bar{x})}$ there exists $z \in \{tx + (1-t)y \mid t \in [0,1]\} \subset \overline{B_r(\bar{x})}$ such that

$$T_k(x) - T_k(y) = DT_k(z)(x - y).$$

Thus,

$$|T_k(x) - T_k(y)| = \left| DT_k(z) \frac{r(x-y)}{\|x-y\|_\infty} \right| \frac{\|x-y\|_\infty}{r} \leq Z_k(r) \frac{\|x-y\|_\infty}{r}. \tag{3.4}$$

Setting $y = \bar{x}$ and noting that $\|x - y\|_\infty \leq r$, (3.4) yields

$$|T_k(x) - T_k(\bar{x})| \leq Z_k(r).$$

By the triangle inequality

$$|T_k(x) - \bar{x}_k| \leq |T_k(x) - T_k(\bar{x})| + |T_k(\bar{x}) - \bar{x}_k| \leq Z_k(r) + Y_k \leq \|Y + Z(r)\|_\infty < r.$$

That proves that $T(\overline{B_r(\bar{x})}) \subseteq \overline{B_r(\bar{x})}$.

From (3.4), it follows that

$$\|T(x) - T(y)\|_\infty \leq \|Z(r)\|_\infty \frac{\|x-y\|_\infty}{r}.$$

By assumption $\|Z(r)\|_\infty \leq \|Y + Z(r)\|_\infty < r$. Therefore T is a contraction on $\overline{B_r(\bar{x})}$ with a contraction constant $\kappa = \frac{\|Z(r)\|_\infty}{r} < 1$, and hence, by the contraction mapping theorem there exists a unique $\tilde{x} \in \overline{B_r(\bar{x})}$ such that $T(\tilde{x}) = \tilde{x}$. □

Observe that Theorem 3.2.1 does not prescribe a specific value of r. In fact, to emphasize the freedom to choose r we introduce the following concept.

Definition 3.2.2. Given $T \in C^1(U, \mathbb{R}^n)$, $U \subset \mathbb{R}^n$ open, and vectors $Y, Z(r) \in \mathbb{R}^n$ satisfying (3.3) the associated *radii polynomials* $p_k(r)$, $k = 1, \dots, n$ are given by

$$p_k(r) := Y_k + Z_k(r) - r.$$

Using the radii polynomials we restate Theorem 3.2.1 in the form in which we make primary use of it.

Corollary 3.2.3. *Let $U \subset \mathbb{R}^n$ be open, $f \in C^1(U, \mathbb{R}^n)$ and $A\colon \mathbb{R}^n \to \mathbb{R}^n$ be an invertible linear map. Define $T\colon U \to \mathbb{R}^n$ by*

$$T(x) := x - Af(x).$$

Let $\bar{x} \in U$, let $Y, Z(r) \in \mathbb{R}^n$ satisfy (3.3), and let $p_k(r)$, $k = 1, \ldots, n$ be the associated radii polynomials. If there exists $r > 0$ such that $p_k(r) < 0$, for all $k = 1, \ldots, n$, then there exists a unique $\tilde{x} \in \overline{B_r(\bar{x})}$ such that $f(\tilde{x}) = 0$.

Proof. Suppose that $r > 0$ is such that $p_k(r) < 0$ for all $k = 1, \ldots, n$. Hence,

$$\|Y + Z(r)\|_\infty = \max_{k=1,\ldots,n} \{(Y + Z(r))_k\} < r.$$

From Theorem 3.2.1 there exists a unique $\tilde{x} \in \overline{B_r(\bar{x})}$ such that $T(\tilde{x}) = \tilde{x}$ and therefore by Proposition 3.1.1 such that $f(\tilde{x}) = 0$. □

Remark 3.2.4. Observe that if there exists $r > 0$ such that $p_k(r) < 0$, for all $k = 1, \ldots, n$, then there exists a range of values $I = (r_-, r_+) \subset [0, \infty)$ over which the inequalities are satisfied. Since $\tilde{x}$ is the unique zero of f in $\overline{B_r(\bar{x})}$ for all $r \in I$, r_- provides tight bounds on the location of $\tilde{x}$, while r_+ provides information about the domain of isolation of $\tilde{x}$. The maximal such interval is called the *existence interval* for the radii polynomials. Observe that this allows us to rephrase Corollary 3.2.3 as follows; if the existence interval for the radii polynomials is nonempty, then one can present an explicit domain in which there exists a unique zero of f.

The reader will note that the form of T presented in Corollary 3.2.3 plays no role in the proof. However, in practice the choice of A is heavily influenced by $(Df(\bar{x}))^{-1}$. The following result is used to show that if $f(\tilde{x}) = 0$ and $Df(\tilde{x})$ is invertible, then the radii polynomial approach can be used to identify $\tilde{x}$ provided the matrix norm of A is not too large and that the initial guess $\bar{x}$ is a good enough approximation of $\tilde{x}$.

Given a matrix A, denote by $\|A\|_\infty = \max_{x \in \overline{B_1(0)}} \|Ax\|_\infty$ the matrix norm induced by the norm $\|\cdot\|_\infty$.

Theorem 3.2.5. *Consider a C^1 map $f\colon U \to \mathbb{R}^n$. Let $\bar{x} \in U$ and $A \in M_n(\mathbb{R})$. Assume there exist constants $\alpha, \beta, r_+ > 0$ for which the following inequalities are satisfied:*

$$\alpha + 2\beta < 1 \tag{3.5}$$

$$\|I - ADf(\bar{x})\|_\infty < \alpha \tag{3.6}$$

$$\|Df(\bar{x} + b) - Df(\bar{x})\|_\infty \|A\|_\infty < \beta, \quad \forall\, b \in \overline{B_{r_+}(0)} \tag{3.7}$$

$$\|A\|_\infty \|f(\bar{x})\|_\infty < \beta r_+. \tag{3.8}$$

Then, the existence interval for the radii polynomials associated with $T(x) := x - Af(x)$ contains the non-empty interval

$$\left(\beta^{-1}\|A\|_\infty\|f(\bar{x})\|_\infty, r_+\right).$$

Proof. Let $r_- := \beta^{-1}\|A\|_\infty\|f(\bar{x})\|_\infty$. By (3.8) $r_- < r_+$. Thus it is sufficient to show that (r_-, r_+) is contained in the existence interval for the radii polynomials associated with T. For the remainder of the proof we assume $r \in (r_-, r_+)$.

Define

$$Y_\infty := \|A\|_\infty\|f(\bar{x})\|_\infty \quad \text{and} \quad Z_\infty(r) := 2\beta r.$$

Observe that

$$\|T(\bar{x}) - \bar{x}\|_\infty = \|Af(\bar{x})\|_\infty \leq \|A\|_\infty\|f(\bar{x})\|_\infty = \beta r_- < \beta r.$$

Furthermore,

$$\begin{aligned}
\sup_{b,c\in\overline{B_r(0)}} \|DT(\bar{x}+b)c\|_\infty &= \sup_{b,c\in\overline{B_r(0)}} \|[I - ADf(\bar{x}+b)]c\|_\infty \\
&= \sup_{b,c\in\overline{B_r(0)}} \|[I - ADf(x) + ADf(x) - ADf(\bar{x}+b)]c\|_\infty \\
&\leq \sup_{b,c\in\overline{B_r(0)}} \left(\|I - ADf(\bar{x})\|_\infty\|c\|_\infty + \|A(Df(\bar{x}+b) - Df(\bar{x}))\|_\infty\|c\|_\infty\right) \\
&\leq \sup_{b\in\overline{B_r(0)}} \left(\|I - ADf(\bar{x})\|_\infty + \|A\|_\infty\|Df(\bar{x}+b) - Df(\bar{x})\|_\infty\right) r \\
&\overset{①}{\leq} \left(\alpha + \|A\|_\infty \sup_{b\in\overline{B_r(0)}} \|Df(\bar{x}+b) - Df(\bar{x})\|_\infty\right) r \\
&\overset{②}{\leq} \left(\alpha + \|A\|_\infty \sup_{b\in\overline{B_{r_+}(0)}} \|Df(\bar{x}+b) - Df(\bar{x})\|_\infty\right) r \\
&\overset{③}{\leq} (\alpha+\beta)r
\end{aligned}$$

where ① follows from (3.6), ② follows from the assumption that $r < r_+$, and ③ follows from (3.7).

Hence,

$$\|T(\bar{x}) - \bar{x}\|_\infty + \sup_{b,c\in\overline{B_r(0)}} \|DT(\bar{x}+b)c\|_\infty < \beta r + (\alpha+\beta)r < r$$

where the last inequality follows from (3.5). Therefore, for all $r \in (r_-, r_+)$ the radii polynomials for T are negative. □

Remark 3.2.6. Theorem 3.2.5 provides considerable insight into the analytic issues associated with the radii polynomials. First, observe that by (3.5), $\alpha < 1$. Therefore, (3.6) implies that A and $Df(\bar{x})$ must be invertible matrices. Since A is invertible, if the hypotheses of Theorem 3.2.5 are satisfied, then by Corollary 3.2.3 there exists a unique solution $\tilde{x} \in \overline{B_{r_+}(\bar{x})}$ to $f(x) = 0$.

Observe that if $\bar{x} = \tilde{x}$, then $f(\bar{x}) = 0$ and hence the existence interval contains $(0, r_+)$. Furthermore, (3.8) is satisfied for all $\beta, r_+ > 0$. If in addition, $A = (Df(\bar{x}))^{-1}$, then $I - ADf(\bar{x}) = 0$ and hence (3.6) is satisfied for all $\alpha > 0$. The assumption that $f \in C^1$ implies that β can be chosen arbitrarily small by choosing r_+ sufficiently small. Therefore, given $\tilde{x} \in \mathbb{R}^n$ for which $f(\tilde{x}) = 0$ and $Df(\tilde{x})$ is invertible, it is, at least theoretically, always possible to use the radii polynomial approach to prove the existence of and provide bounds on the location of $\tilde{x}$; a sufficient condition is a good initial approximation $\bar{x} \approx \tilde{x}$ and a good approximation of $(Df(\tilde{x}))^{-1}$.

However, Theorem 3.2.5 also indicates that this is not a necessary condition. The initial approximation $\bar{x}$ may differ significantly from $\tilde{x}$, if $\|f(\bar{x})\|_\infty$ and/or $\|A\|_\infty$ is sufficiently small. In addition, as (3.6) indicates $A \approx (Df(\bar{x}))^{-1}$ is not a necessary condition. Thus, the radii polynomial provide considerable freedom in terms of their application in concrete problems.

To demonstrate how the radii polynomials are used in practice we consider several simple examples.

Example 3.2.7 (The radii polynomial approach for a one-dimensional example). Consider the simplest nonlinear function $f(x) = x^2 - 2$. Following Corollary 3.2.3 the first step is to choose an initial guess $\bar{x}$ for the zero of f. In a typical application $\bar{x}$ is taken as the output for a standard numerical procedure for finding zeros of a function. The next step is to fix a value for $A \in \mathbb{R}$. Example 3.1.2 suggests that A be chosen based on $Df(\bar{x})^{-1}$. We leave it general for now. To prove that

$$T(x) := x - A(x^2 - 2)$$

is a contraction mapping we need to determine bounds Y and $Z(r)$ that satisfy (3.3). Consider any Y such that

$$|T(\bar{x}) - \bar{x}| = |A(\bar{x}^2 - 2)| \leq Y.$$

Obtaining $Z(r)$ is slightly more complicated. It is convenient to write $b, c \in \overline{B_r(0)}$ as where $u, v \in \overline{B_1(0)}$. Using this notation

$$\begin{aligned} DT(\bar{x} + b)c &= (1 - 2A(\bar{x} + b))c \\ &= (1 - 2A(\bar{x} + ru))rv \\ &= [(1 - 2\bar{x}A)v]r + [-2Auv]r^2. \end{aligned}$$

Thus

$$\begin{aligned}\sup_{b,c\in\overline{B_r(0)}} |DT(\bar{x}+b)c| &= \sup_{u,v\in\overline{B_1(0)}} \left|[(1-2\bar{x}A)v]r + [-2Auv]r^2\right| \\ &\leq \sup_{u,v\in\overline{B_1(0)}} |(1-2\bar{x}A)v|\, r + \left|[2Auv]r^2\right| \\ &\leq |1-2\bar{x}A|r + 2Ar^2\end{aligned}$$

and hence we choose $Z(r) := |1-2\bar{x}A|r + 2Ar^2$. For the choice $Y = |A(\bar{x}^2-2)|$ and $Z(r)$ as above the associated radii polynomial is given by

$$\begin{aligned}p(r) &= Y + Z(r) - r \\ &= |A(\bar{x}^2-2)| + |1-2\bar{x}A|r + 2Ar^2 - r \\ &= 2Ar^2 + (|1-2\bar{x}A| - 1)\, r + |A(\bar{x}^2-2)|.\end{aligned}$$

Let us now make some explicit choices.

(a) The best possible approximations

Let $\bar{x} = \sqrt{2}$ (this is an exact solution) and $A = Df(\bar{x})^{-1} = \frac{1}{2\bar{x}}$ (the exact inverse). In this case, the radii polynomial is $p(r) = 2Ar^2 + (|1-2\bar{x}A| - 1)\, r + |A(\bar{x}^2-2)| = \frac{\sqrt{2}}{2}r^2 - r$. Then, the existence interval for the radii polynomial is given by $I = \left(0, \sqrt{2}\right)$.

(b) Not the best approximations

For the purpose of applications it is not practical to assume that $\bar{x}$ is the exact solution, nor is it reasonable to assume that $A = Df(\bar{x})^{-1}$. However, the proposed approach works well even with coarse approximations.

Choose $\bar{x} = 1.3$. Since $Df(\bar{x})^{-1} = (2\bar{x})^{-1} \approx 0.384615$, we set the approximate inverse to be $A = 0.38$. To prove that

$$T(x) := x - 0.38(x^2-2)$$

is a contraction mapping we need to determine bounds Y and $Z(r)$ that satisfy (3.3). To obtain Y observe that

$$|T(\bar{x}) - \bar{x}| = |-0.38(1.3^2-2)| = 0.1178.$$

Since we only need a bound we choose $Y := 0.12$. As above, we have $Z(r) = |1-2\bar{x}A|r + 2Ar^2 = 0.012r + 0.76r^2$.

For this choice of Y and $Z(r)$ the associated radii polynomial is given by

$$p(r) = Y + Z(r) - r = 0.12 + 0.012r + 0.76r^2 - r = 0.76r^2 - 0.988r + 0.12.$$

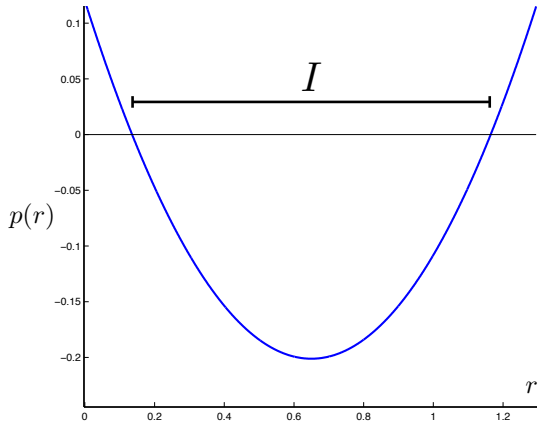

Figure 3.1: The radii polynomial $p(r) = 0.76r^2 - 0.988r + 0.12$ of Example 3.2.7 and $I = [0.136, 1.164]$, where the radii polynomial is strictly negative.

See Figure 3.1 for a geometrical interpretation of the radii polynomial $p(r)$. Using the quadratic formula we see that $p(r) < 0$ for all $r \in I = [0.136, 1.164]$. By Corollary 3.2.3 we can conclude that $\tilde{x} \in \overline{B_{0.136}(1.3)} = [1.164, 1.436]$. Given that the actual root contained in this interval is $\sqrt{2} \approx 1.414$ we see that the relative error for this bound is

$$\frac{1.414 - 1.164}{1.414} \approx 0.1768.$$

We can also conclude that there is a unique root in the interval $\tilde{x} \in \overline{B_{1.164}(1.3)} = [0.136, 2.464]$. Given that $-\sqrt{2} \approx -1.414$ is also a root, this is a reasonably accurate statement.

It is reasonable to ask about optimal results involving the radii polynomial approach. For example, over how large a domain can one hope to show the unique existence of a root? Let $\bar{x} = \tilde{x} = \sqrt{2}$ and let $A = Df(\bar{x})^{-1} = \frac{1}{2\sqrt{2}}$. As is discussed in Remark 3.2.6, this choice of $\bar{x}$ and A essentially allows us to assume that $\alpha = 0$ and hence by (3.5) that $\beta = \frac{1}{2}$. Thus the only constraint remaining for Theorem 3.2.5 is (3.7), which for this particular example gives rise to

$$\begin{aligned} \|Df(\bar{x}+b) - Df(\bar{x})\|_\infty \|A\|_\infty &= \left|2(\sqrt{2}+b) - 2\sqrt{2}\right| \frac{1}{2\sqrt{2}} \\ &= \frac{b}{\sqrt{2}} \\ &< \frac{1}{2} = \beta. \end{aligned}$$

Thus $r_+ = \frac{\sqrt{2}}{2}$.

Therefore, by Theorem 3.2.5 we can conclude that there is a unique root in the interval

$$(\sqrt{2}/2, 3\sqrt{2}/2) \approx (0.707, 2.121) \subset [0.136, 2.464]$$

where the last interval comes from the computations at the beginning of this example. This demonstrates that the bounds presented in the hypothesis of Theorem 3.2.5 are sufficient, but far from necessary.

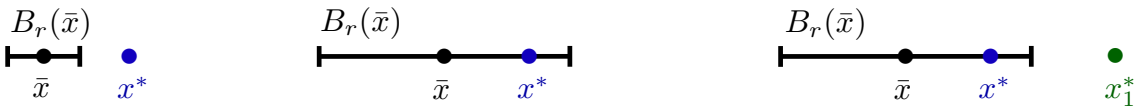

Figure 3.2: For different values of $r > 0$, we can have different scenarios. On the left, the radius r is too small and $p(r) \geq 0$. Hence the method based on the radii polynomials cannot conclude about the existence of a unique solution $x^* \in \overline{B_r(\bar{x})}$. In the context of Example 3.2.7, that would correspond for instance to $r = 0.1 \notin I = [0.136, 1.164]$. In the middle, the radius $r \in I$ is chosen so that $p(r) < 0$, which implies that the ball $\overline{B_r(\bar{x})}$ contains a unique solution. In the context of Example 3.2.7, that would correspond for instance to $r = 0.15 \in I$. On the right, the radius r is too large and $p(r) \geq 0$. Hence, it could happen that there exist more than one solution in the ball $\overline{B_r(\bar{x})}$, that is there exist $x^*, x_1^* \in \overline{B_r(\bar{x})}$ both solutions of $f = 0$. In the context of Example 3.2.7, that would correspond for instance to $r = 3$, where $x^* = \sqrt{2}$ and $x_1^* = -\sqrt{2}$.

Example 3.2.7 demonstrates how the radii polynomials can be employed. It also makes clear that the choice of A plays a central role in the values of Y and $Z(r)$, which in turn determines whether the radii polynomial inequalities can be satisfied. With this in mind we return to this example to gain some insight into how broad a range of As are possible.

Example 3.2.8. As a slight generalization of Example 3.2.7, let $f(x) = x^2 - \lambda$. Recall that the philosophy of the approach we are taking is that we begin with a guess $\bar{x}$ of the location of a zero of f and from that attempt to prove that a zero exists. Using Newton's method as a guide we set $A = (f'(\bar{x}))^{-1} = (2\bar{x})^{-1}$ and hence

$$T(x) = x - \frac{x^2 - \lambda}{2\bar{x}}.$$

Applying the same analysis as in Example 3.2.7, set

$$Y = \frac{|\bar{x}^2 - \lambda|}{2\bar{x}} = |T(\bar{x}) - \bar{x}|$$

and

$$Z(r) = \frac{r^2}{\bar{x}} = \sup_{u,v \in \overline{B_1(0)}} \frac{|-r^2 uv|}{\bar{x}} = \sup_{b,c \in \overline{B_r(0)}} |DT(\bar{x} + b)c|.$$

Thus, the radii polynomial inequality takes the form

$$p(r) = r^2 - \bar{x}r + \frac{|\bar{x}^2 - \lambda|}{2} < 0.$$

This in turn implies that the interval of convergence $I = (r_-, r_+)$ for $\lambda \geq 0$ is given by

$$r_\pm = \frac{\bar{x} \pm \sqrt{\bar{x}^2 - 2\,|\bar{x}^2 - \lambda|}}{2}.$$

This in turn tells us that if

$$\bar{x} \in \left(\sqrt{\frac{2\lambda}{3}}, \sqrt{2\lambda} \right),$$

then the radii polynomials will provide a positive answer to the question of existence of a zero of f.

Example 3.2.9 (The radii polynomial approach for a two-dimensional example). Consider the problem of looking for equilibria of

$$\begin{cases} \dot{x_1} = x_2 + 4x_1^2 - \lambda \\ \dot{x_2} = x_1 + x_2^2 - 1, \end{cases} \tag{3.9}$$

where $\lambda \in \mathbb{R}$ is a parameter. An equilibrium solution $x = (x_1, x_2)$ is a solution of $f(x) = 0$ where $f : \mathbb{R}^2 \to \mathbb{R}^2$ is given by the rand-hand side of (3.9). At some given parameter values $\lambda \in \mathbb{R}$, there are up to four real solutions.

Given an initial guess $\bar{x}$

$$Df(\bar{x}) = \begin{pmatrix} 8\bar{x}_1 & 1 \\ 1 & 2\bar{x}_2 \end{pmatrix}.$$

and the exact formula for the inverse is

$$Df(\bar{x})^{-1} = \frac{1}{16\bar{x}_1\bar{x}_2 - 1} \begin{pmatrix} 2\bar{x}_2 & -1 \\ -1 & 8\bar{x}_1 \end{pmatrix}.$$

Set $A := Df(\bar{x})^{-1}$, and let $T(x) := x - Af(x)$. To apply the radii polynomial approach we compute the bounds $Y, Z(r) \in \mathbb{R}^2$ satisfying (3.3).

To obtain Y, realize that

$$T(\bar{x}) - \bar{x} = -Af(\bar{x}) = -Df(\bar{x})^{-1} f(\bar{x}) = -\frac{1}{16\bar{x}_1\bar{x}_2 - 1} \begin{pmatrix} 2\bar{x}_2 & -1 \\ -1 & 8\bar{x}_1 \end{pmatrix} \begin{pmatrix} \bar{x}_2 + 4\bar{x}_1^2 - \lambda \\ \bar{x}_1 + \bar{x}_2^2 - 1 \end{pmatrix}.$$

Using this expression, we can choose Y_k such that $|[T(\bar{x}) - \bar{x}]_k| \leq Y_k$, for $k = 1, 2$.

The next step is to determine $Z_k(r)$ such that

$$\sup_{b,c \in \overline{B_r(0)}} |DT_k(\bar{x} + b)c| \leq Z_k(r), \quad k = 1, 2.$$

In order to simplify the computation of Z_k we rescale the variables b and c. For $r > 0$, let $\tilde{b} := b/r$ and $\tilde{c} := c/r$, in which case the desired bounds become

$$\sup_{\tilde{b},\tilde{c}\in\overline{B_1(0)}} |DT_k(\bar{x} + \tilde{b}r)\tilde{c}|r \leq Z_k(r), \quad k = 1, 2.$$

To improve the estimates we consider the following splitting

$$\begin{aligned} DT(\bar{x} + \tilde{b}r)\tilde{c}r &= \left(I - ADf(\bar{x} + \tilde{b}r)\right)\tilde{c}r \\ &= \left(I - ADf(\bar{x})\right)\tilde{c}r - A\left(Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\right)\tilde{c}r \\ &= -A\left(Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\right)\tilde{c}r \end{aligned}$$

where the last equality follows from the choice $A = Df(\bar{x})^{-1}$. It is important to realize that in general this is not the case, but does suggest why, in more complicated examples, it is useful to be able to choose $A \approx Df(\bar{x})^{-1}$.

To bound the second term in the splitting we note that

$$\begin{aligned} \left(Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\right)\tilde{c}r &= \left[\begin{pmatrix} 8\bar{x}_1 + 8\tilde{b}_1 r & 1 \\ 1 & 2\bar{x}_2 + 2\tilde{b}_2 r \end{pmatrix} - \begin{pmatrix} 8\bar{x}_1 & 1 \\ 1 & 2\bar{x}_2 \end{pmatrix}\right]\tilde{c}r \\ &= \begin{pmatrix} 8\tilde{b}_1\tilde{c}_1 \\ 2\tilde{b}_2\tilde{c}_2 \end{pmatrix} r^2 \end{aligned}$$

and hence

$$A\left(Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\right)\tilde{c}r = -\frac{1}{16\bar{x}_1\bar{x}_2 - 1}\begin{pmatrix} 2\bar{x}_2 & -1 \\ -1 & 8\bar{x}_1 \end{pmatrix}\begin{pmatrix} 8\tilde{b}_1\tilde{c}_1 \\ 2\tilde{b}_2\tilde{c}_2 \end{pmatrix} r^2.$$

By definition $\tilde{b}, \tilde{c} \in \overline{B_1(0)}$ implies that $|\tilde{b}_1|, |\tilde{c}_1|, |\tilde{b}_2|, |\tilde{c}_2| \leq 1$, thus

$$\left|\left(A\left(Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\right)\tilde{c}r\right)_1\right| \leq \left(\frac{16|\bar{x}_2| + 2}{|16\bar{x}_1\bar{x}_2 - 1|}\right) r^2$$

and

$$\left|\left(A\left(Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\right)\tilde{c}r\right)_2\right| \leq \left(\frac{16|\bar{x}_1| + 8}{|16\bar{x}_1\bar{x}_2 - 1|}\right) r^2.$$

Set

$$Z_1(r) := \left(\frac{16|\bar{x}_2| + 2}{|16\bar{x}_1\bar{x}_2 - 1|}\right) r^2 \quad \text{and} \quad Z_2(r) := \left(\frac{16|\bar{x}_1| + 8}{|16\bar{x}_1\bar{x}_2 - 1|}\right) r^2.$$

Finally, the two radii polynomials are defined by

$$p_1(r) := \left(\frac{16|\bar{x}_2| + 2}{|16\bar{x}_1\bar{x}_2 - 1|}\right) r^2 - r + Y_1 \quad \text{and} \quad p_2(r) := \left(\frac{16|\bar{x}_1| + 8}{|16\bar{x}_1\bar{x}_2 - 1|}\right) r^2 - r + Y_2. \quad (3.10)$$

Observe that the definition of the radii polynomials (3.10) is the same for any approximate solution $\bar{x} \in \mathbb{R}^2$. This means that we have derived explicit formulas of the radii polynomials that can be applied to any initial approximation $\bar{x}$. If the initial approximation is reasonable, then we expect to be able to prove the existence of a true solution $\tilde{x}$. As an example, set $\lambda = 3$, and using some numerical scheme, e.g., Newton's method, find $\bar{x} = (\bar{x}_1, \bar{x}_2) \in \mathbb{R}^2$ such that $\|f(\bar{x})\| < \texttt{tol}$ for some fixed tolerance `tol`. We chose $\texttt{tol} = 10^{-15}$, and computed four approximate solutions with Newton's method obtaining

$$\begin{aligned} \bar{x}^{(1)} &= \begin{pmatrix} -0.6545436118927946 \\ 1.286290640521338 \end{pmatrix}, \quad \bar{x}^{(2)} = \begin{pmatrix} 0.7986333753610425 \\ 0.4487389270377125 \end{pmatrix}, \\ \bar{x}^{(3)} &= \begin{pmatrix} 0.9086121587039679 \\ -0.3023042197787387 \end{pmatrix}, \quad \bar{x}^{(4)} = \begin{pmatrix} -1.052701922172216 \\ -1.432725347780312 \end{pmatrix}. \end{aligned} \tag{3.11}$$

For each $i = 1, 2, 3, 4$, we the following intervals $I^{(i)}$ are contained in the existence intervals for the associated radii polynomials:

$$\begin{aligned} I^{(1)} &= [1.608556563336234 \times 10^{-16},\ 0.6402146110280825] \\ I^{(2)} &= [6.468926557835299 \times 10^{-16},\ 0.2276100489022837] \\ I^{(3)} &= [1.218510871878330 \times 10^{-15},\ 0.2391290678185533] \\ I^{(4)} &= [4.855265252055830 \times 10^{-16},\ 0.9271769229945959]. \end{aligned} \tag{3.12}$$

Figure 3.3 shows the largest enclosures for each equilibrium of (3.9) for $\lambda = 3$. For each $i = 1, 2, 3, 4$, the radius around $\bar{x}^{(i)}$ is the largest value of $I^{(i)}$.

In Example 3.2.9, we defined A as the exact inverse of $Df(\bar{x})$. For general finite dimensional problems, this is not always possible. In fact, as the dimension of the problem grows, this becomes difficult. For infinite dimensional problems, obtaining an exact inverse is almost impossible. However, as the following example demonstrates, having the exact inverse is not necessary.

Example 3.2.10. The Lorenz system is given by

$$\begin{cases} \dot{x}_1 = \sigma(x_2 - x_1) \\ \dot{x}_2 = \rho x_1 - x_2 - x_1 x_3 \\ \dot{x}_3 = -\beta x_3 + x_1 x_2 \end{cases} \tag{3.13}$$

For any $\beta > 0$ and $\rho > 1$, the set of equilibria of (3.13) is given by

$$\left\{(0,0,0), \left(\pm\sqrt{\beta(\rho-1)}, \pm\sqrt{\beta(\rho-1)}, \rho - 1\right)\right\}$$

which is obtained by solving $f(x) = 0$, where $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ is given by the right-hand side of (3.13). At the classical parameter values, $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$, $\left(\sqrt{72}, \sqrt{72}, 27\right)$

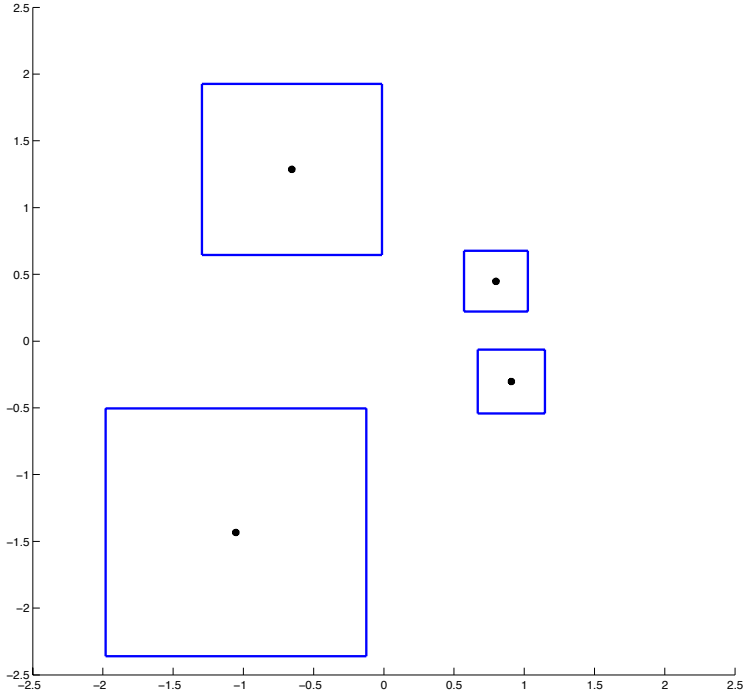

Figure 3.3: Largest existence and uniqueness enclosures for each equilibrium of (3.9) for $\lambda = 3$. For each $i = 1, 2, 3, 4$, the radius around $\bar{x}^{(i)}$ is the largest value of $I^{(i)}$. The quantities $\bar{x}^{(i)}$ and $I^{(i)}$ are found in (3.11) and (3.12) respectively. The smallest enclosure is too small to represent, which implies that the dots representing $\bar{x}^{(i)}$ also represent the true equilibria $\tilde{x}^{(i)}$.

is an equilibrium solution, which is approximated by

$$\bar{x} := \begin{pmatrix} 8.4853 \\ 8.4853 \\ 27 \end{pmatrix}. \tag{3.14}$$

Using that

$$Df(\bar{x}) = \begin{pmatrix} -\sigma & \sigma & 0 \\ \rho - \bar{x}_3 & -1 & -\bar{x}_1 \\ \bar{x}_2 & \bar{x}_1 & -\beta \end{pmatrix}$$

we compute

$$Df(\bar{x})^{-1} \approx A := \begin{pmatrix} -0.051851843722001 & -0.018518437220007 & 0.058925435753597 \\ 0.048148156277999 & -0.018518437220007 & 0.058925435753597 \\ -0.011785087150719 & -0.117850871507195 & 0 \end{pmatrix}.$$

Using interval arithmetic, we can compute bounds Y_k such that

$$|[T(\bar{x}) - \bar{x}]_k| = |[Af(\bar{x})]_k| \leq Y_k, \tag{3.15}$$

for each $k = 1, 2, 3$.

The next step is to compute bounds Z_k such that

$$\sup_{b,c \in \overline{B_r(0)}} |DT_k(\bar{x} + b)c| \leq Z_k(r), \quad k = 1, 2, 3.$$

We follow the procedure as in Example 3.2.9. For $r > 0$, let $\tilde{b} := b/r$ and $\tilde{c} := c/r$ and consider

$$\sup_{\tilde{b},\tilde{c} \in \overline{B_1(0)}} |DT_k(\bar{x} + \tilde{b}r)\tilde{c}|r \leq Z_k(r), \quad k = 1, 2, 3.$$

Again, we perform the splitting

$$\begin{aligned} DT(\bar{x} + \tilde{b}r)\tilde{c}r &= \Big(I - ADf(\bar{x} + \tilde{b}r)\Big)\,\tilde{c}r \\ &= (I - ADf(\bar{x}))\,\tilde{c}r - A\Big(Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\Big)\,\tilde{c}r. \end{aligned} \tag{3.16}$$

However, since $A \neq Df(\bar{x})$, the first term does not vanish and thus we need to obtain additional bounds. Using interval arithmetic, we can compute a bound $Z_k^{(1)}$ such that

$$|[(I - ADf(\bar{x}))\,\tilde{c}r]_k| \leq \left[|I - ADf(\bar{x})| \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\right]_k r \leq Z_k^{(1)} r. \tag{3.17}$$

Let us now focus our attention on the second term of (3.16):

$$\begin{aligned} &\Big(Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\Big)\,\tilde{c}r \\ &= \left[\begin{pmatrix} -\sigma & \sigma & 0 \\ \rho - (\bar{x}_3 + \tilde{b}_3 r) & -1 & -(\bar{x}_1 + \tilde{b}_1 r) \\ \bar{x}_2 + \tilde{b}_2 r & \bar{x}_1 + \tilde{b}_1 r & -\beta \end{pmatrix} - \begin{pmatrix} -\sigma & \sigma & 0 \\ \rho - \bar{x}_3 & -1 & -\bar{x}_1 \\ \bar{x}_2 & \bar{x}_1 & -\beta \end{pmatrix}\right] \begin{pmatrix} \tilde{c}_1 \\ \tilde{c}_2 \\ \tilde{c}_3 \end{pmatrix} r \\ &= \begin{pmatrix} 0 & 0 & 0 \\ -\tilde{b}_3 r & 0 & -\tilde{b}_1 r \\ \tilde{b}_2 r & \tilde{b}_1 r & 0 \end{pmatrix} \begin{pmatrix} \tilde{c}_1 \\ \tilde{c}_2 \\ \tilde{c}_3 \end{pmatrix} r = \begin{pmatrix} 0 \\ -\tilde{b}_3\tilde{c}_1 - \tilde{b}_1\tilde{c}_3 \\ \tilde{b}_2\tilde{c}_1 + \tilde{b}_1\tilde{c}_2 \end{pmatrix} r^2 \end{aligned}$$

and then

$$\left|A\Big(Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\Big)\,\tilde{c}r\right| \ll |A| \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix} r^2 =: \begin{pmatrix} Z_1^{(2)} \\ Z_2^{(2)} \\ Z_3^{(2)} \end{pmatrix} r^2, \tag{3.18}$$

where the notation $\ll$ denotes component-wise inequalities. Combining (3.15), (3.17) and (3.18), we can define the three radii polynomials as

$$p_k(r) = Z_k^{(2)} r^2 + \left(Z_k^{(1)} - 1\right) r + Y_k.$$

Applying Corollary 3.2.3 we can conclude that there is a unique equilibrium to (3.13) in the ball $\overline{B_r(\bar{x})}$ for any

$$r \in I := \left[1.8626 \times 10^{-5},\ 4.2427\right] \subset \bigcap_{k=1}^{3} \left\{r > 0 \mid p_k(r) < 0\right\}.$$

We want to emphasize the fact that careful choices of inverses and approximate solutions are not necessary. Keeping $\bar{x}$ as in (3.14) consider the crude approximation of $Df(\bar{x})^{-1}$

$$A = \begin{pmatrix} -0.05 & -0.01 & 0.05 \\ 0.04 & -0.01 & 0.05 \\ -0.01 & -0.11 & 0 \end{pmatrix}.$$

Using the same analysis we get that the interval

$$I = \left[2.0485 \times 10^{-5},\ 4.1517\right]$$

is a subset of the existence interval. In fact, with this latest choice of A, if instead of $\bar{x}$ given in (3.14), we let

$$\bar{x} := \begin{pmatrix} 8.48528137423857 \\ 8.48528137423857 \\ 27 \end{pmatrix},$$

then we would get that

$$I = \left[9.2097 \times 10^{-16},\ 4.1517\right].$$

Theorem 3.2.11. *If the vector field $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is polynomial of degree d, then the degree of the associated radii polynomials equals d.*

Proof. Recall that $p_k(r) = Y_k + Z_k(r) - r$, for $k = 1, \ldots, n$, where $Z_k(r)$ is a uniform upper bound for $|DT_k(\bar{x} + b)c|$, over all $b, c \in B_r(0)$. Let $\tilde{b}, \tilde{c} \in B_1(0)$ such that $b = \tilde{b}r$, $c = \tilde{c}r$. Hence,

$$\begin{aligned} |DT_k(\bar{x} + b)c| &= \left|\left[\left(I - ADf(\bar{x} + \tilde{b}r)\right)\tilde{c}\right]_k\right| r \\ &\leq \left|\left[\left(I - ADf(\bar{x})\right)\tilde{c}\right]_k\right| r + \left|\left[A\left(Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\right)\tilde{c}\right]_k\right| r \\ &\leq \left[\left|I - ADf(\bar{x})\right| \mathbb{1}_n\right]_k r + \left[|A|\left|Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\right| \mathbb{1}_n\right]_k r \end{aligned}$$

where $\mathbb{1}_n = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$. Since $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is polynomial of degree d, then $\left[|A|\left|Df(\bar{x} + \tilde{b}r) - Df(\bar{x})\right| \mathbb{1}_n\right]_k r$ is polynomial of degree d. Hence, the bound $Z_k(r)$ can be chosen as a polynomial of degree d. $\square$

For vector fields with polynomial nonlinearities, Theorem 3.2.11 shows that the degree of the radii polynomials equals the degree of the nonlinearity of the vector field under study. While this is very practical, the radii polynomial approach also works for non polynomial vector fields. This is done by applying the Mean Value Theorem appropriately, or by using a Taylor expansion about the approximate solution. Let us present an explicit example.

Example 3.2.12 (The radii polynomials for a non polynomial vector field). Consider the following model

$$\begin{cases} \dot{x_1} = 3x_1(1-x_1) - x_1x_2 - \lambda(1-e^{-5x_1}) \\ \dot{x_2} = -x_2 + 3x_1x_2 \end{cases}. \tag{3.19}$$

At some given parameter values $\lambda \in \mathbb{R}$, there are up to four real equilibrium solutions.

Let $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ the right hand-side of (3.19). Hence,

$$Df(\bar{x}) = \begin{pmatrix} 3 - 6\bar{x}_1 - \bar{x}_2 - 5\lambda e^{-5\bar{x}_1} & -\bar{x}_1 \\ 3\bar{x}_2 & -1 + 3\bar{x}_1 \end{pmatrix}.$$

Since we are working with a two dimensional example, it is easy to obtain an explicit expression for $Df(\bar{x})^{-1}$ and thus we can set $A := Df(\bar{x})^{-1}$, and let $T(x) := x - Af(x)$. To apply the radii polynomial approach we compute the bounds Y and Z satisfying (3.3). The bound Y can be easily computed with interval arithmetic. To compute the bound Z, consider any $b, c \in \overline{B_r(0)}$. For $r > 0$, let $\tilde{b} := b/r$ and $\tilde{c} := c/r$. As before, thanks to the (perfect) choice $A = Df(\bar{x})^{-1}$

$$DT(\bar{x}+b)c = -A\left(Df(\bar{x}+\tilde{b}r) - Df(\bar{x})\right)\tilde{c}r.$$

Now,

$$\begin{aligned} \left(Df(\bar{x}+\tilde{b}r) - Df(\bar{x})\right)\tilde{c}r &= \left[\begin{pmatrix} 3 - 6(\bar{x}_1+\tilde{b}_1r) - (\bar{x}_2+\tilde{b}_2r) - 5\lambda e^{-5(\bar{x}_1+\tilde{b}_1r)} & -(\bar{x}_1+\tilde{b}_1r) \\ 3(\bar{x}_2+\tilde{b}_2r) & -1+3(\bar{x}_1+\tilde{b}_1r) \end{pmatrix}\right. \\ &\quad \left. - \begin{pmatrix} 3 - 6\bar{x}_1 - \bar{x}_2 - 5\lambda e^{-5\bar{x}_1} & -\bar{x}_1 \\ 3\bar{x}_2 & -1+3\bar{x}_1 \end{pmatrix}\right]\tilde{c}r \\ &= \begin{pmatrix} -6\tilde{b}_1r - \tilde{b}_2r - 5\lambda e^{-5\bar{x}_1}(e^{-5\tilde{b}_1r}-1) & -\tilde{b}_1r \\ 3\tilde{b}_2r & 3\tilde{b}_1r \end{pmatrix}\tilde{c}r. \end{aligned}$$

Applying the mean value theorem to the function $e^{-5\tilde{b}_1rt}$ over the interval $t \in [0,1]$ we conclude that there exists $\xi \in [0,1]$ such that

$$e^{-5\tilde{b}_1r} - 1 = -5\tilde{b}_1re^{-5\tilde{b}_1r\xi}.$$

Now assume the *a-priori* existence of a constant r^* such that

$$r \leq r^*. \tag{3.20}$$

Hence, $|e^{-5\tilde{b}_1 r} - 1| = \left(5e^{5r^*}\right) r$. Using that $|\tilde{b}_1|, |\tilde{c}_1|, |\tilde{b}_2|, |\tilde{c}_2| \leq 1$, we get

$$\begin{aligned}\left|\left(Df(\bar{x}+\tilde{b}r) - Df(\bar{x})\right)\tilde{c}r\right| &= \left|\begin{pmatrix} -6\tilde{b}_1 r - \tilde{b}_2 r - 5\lambda e^{-5\bar{x}_1}(e^{-5\tilde{b}_1 r} - 1) & -\tilde{b}_1 r \\ 3\tilde{b}_2 r & 3\tilde{b}_1 r \end{pmatrix}\tilde{c}r\right| \\ &\ll \begin{pmatrix} 8 + 25|\lambda|e^{-5\bar{x}_1 + 5r^*} \\ 6 \end{pmatrix} r^2.\end{aligned}$$

Therefore,

$$\left|A\left(Df(\bar{x}+\tilde{b}r) - Df(\bar{x})\right)\tilde{c}r\right| \ll |A| \begin{pmatrix} 8 + 25|\lambda|e^{-5\bar{x}_1 + 5r^*} \\ 6 \end{pmatrix} r^2$$

Set

$$\begin{pmatrix} Z_1(r) \\ Z_2(r) \end{pmatrix} := |A| \begin{pmatrix} 8 + 25|\lambda|e^{-5(\bar{x}_1 - r^*)} \\ 6 \end{pmatrix} r^2$$

which results in the two radii polynomials

$$p_1(r) := Y_1 + Z_1(r) - r \quad \text{and} \quad p_2(r) := Y_2 + Z_2(r) - r.$$

To demonstrate the use of these radii polynomials we fix $\lambda = 1/2$ and consider $\bar{x} = (\bar{x}_1, \bar{x}_2) \in \mathbb{R}^2$ such that $\|f(\bar{x})\| < \texttt{tol}$ for some fixed tolerance `tol`. In this example, we choose $\texttt{tol} = 10^{-15}$, and compute four approximate solutions with Newton's method to obtain

$$\begin{aligned} \bar{x}^{(1)} &= \begin{pmatrix} 0.7940710596186642 \\ 0 \end{pmatrix}, & \bar{x}^{(2)} &= \begin{pmatrix} 0.3333333333333334 \\ 0.7833134042563428 \end{pmatrix}, \\ \bar{x}^{(3)} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, & \bar{x}^{(4)} &= \begin{pmatrix} -0.1095121750607109 \\ 0 \end{pmatrix}. \end{aligned} \tag{3.21}$$

For each $i = 1, 2, 3, 4$, we compute an interval $I^{(i)}$ on which each radii polynomial is negative:

$$\begin{aligned} I^{(1)} &= [1.533640748312133 \times 10^{-16},\ 0.1514305298303711] \\ I^{(2)} &= [7.392992376471959 \times 10^{-16},\ 0.02924694092691947] \\ I^{(3)} &= (0,\ 0.02268040322752661] \\ I^{(4)} &= [5.844698641248233 \times 10^{-16},\ 0.02048663024277441]. \end{aligned} \tag{3.22}$$

To obtain $I^{(1)}$, we set $r^* = 0.152$, to obtain $I^{(2)}$, we set $r^* = 0.0293$, to obtain $I^{(3)}$, we set $r^* = 0.023$ and to obtain $I^{(4)}$, we set $r^* = 0.0206$.

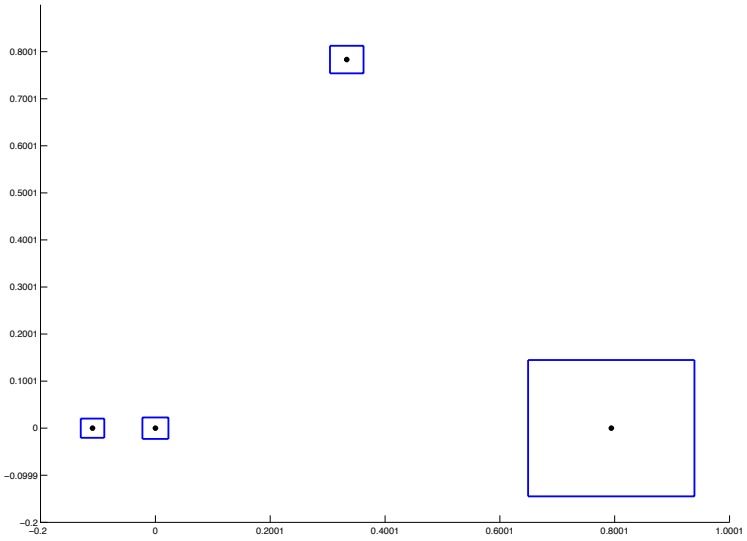

Figure 3.4: Largest existence and uniqueness enclosures for each equilibria of (3.19) for $\lambda = 1/2$. For each $i = 1, 2, 3, 4$, the radius around $\bar{x}^{(i)}$ is the largest value of $I^{(i)}$. The quantities $\bar{x}^{(i)}$ and $I^{(i)}$ are found in (3.21) and (3.22) respectively.

3.3 Exercises

1. Let $f\colon \mathbb{R} \to \mathbb{R}$ be differentiable with at least two real roots. Prove that there exist points $x^0 \in \mathbb{R}$ such that Newton's method applied to x^0 does not converge.

2. Consider the FitzHugh-Nagumo equation

$$f(x) = \begin{pmatrix} x_1(x_1 - a)(1 - x_1) - x_2 \\ \epsilon(x_1 - \gamma x_2) \end{pmatrix} \tag{3.23}$$

   at the parameter values $(a, \epsilon, \gamma) = (5, 1, 2)$. Consider the initial point $x^0 = (1, 0)^T \in \mathbb{R}^2$ and apply Newton's method to converge to a point $\bar{x} \in \mathbb{R}^2$. What is the point $\bar{x}$ you obtain? Decrease the parameter γ by .05 and apply Newton's method with initial point $x^0 = \bar{x}$, where $\bar{x}$ is the value obtained at the previous step. This initial guess is denoted as a 0th order predictor. Applying Newton's method, you should converge to a new point that we denote again by $\bar{x}$. Repeat this procedure 40 times. What do you observe? This procedure is the simplest form of a predictor-corrector method.

3. Consider the problem $f(x) = 0$, where f is the right hand side of (3.23). Given an approximate solution $\bar{x}$ such that $f(x) \approx 0$, consider an approximate inverse A that is $A \approx Df(\bar{x})^{-1}$. Define $T(x) = x - Af(x)$ and compute the bounds $Y = (Y_1, Y_2)$ and

$Z(r) = (Z_1(r), Z_2(r))$ satisfying (3.3). Define the radii polynomials $p_k(r)$, $k = 1, 2$ using Definition 3.2.2. Use Corollary 3.2.3 to find $r > 0$ such that $\overline{B_r(\bar{x})}$ contains a unique solution of $f(x) = 0$. Write a MATLAB program that verifies the hypotheses of Corollary 3.2.3. Prove the existence of the fixed points close to the 40 numerical values obtain in Exercise 2.

Chapter 4

Linear Theory and Stability of Equilibria

4.1 Preliminaries

Definition 4.1.1. A function $\|\cdot\|: V \to [0, \infty)$ is a *norm* on a vector space V if it satisfies the following conditions:

1. $\|x\| = 0$ if and only if $x = 0 \in V$;

2. $\|\alpha x\| = |\alpha|\|x\|$ for all vectors $x \in V$ and scalars α; and

3. $\|x + y\| \leq \|x\| + \|y\|$ for all vectors $x, y \in V$.

It is easy to check that $d(u, v) = \|u - v\|$ defines a metric on V, which in turn defines a topology on V.

Definition 4.1.2. Let V and W be normed vector spaces with norms denoted by $\|\cdot\|_V$ and $\|\cdot\|_W$, respectively. Let $A: V \to W$ be a linear map. The *norm* of A is defined by

$$\|A\| := \sup_{x \neq 0} \frac{\|Ax\|_W}{\|x\|_V} = \sup_{\|x\|=1} \|Ax\|_W.$$

Observe that

$$\|Ax\|_W \leq \|A\|\|x\|_V.$$

Using different norms on vector spaces results in different norms on operators. It is left

to the reader the check that

$$\begin{array}{rl} \text{Vector Norm} & \text{Matrix Norm} \\ \|x\|_1 := \displaystyle\sum_{k=1}^{n} |x_k| & \|A\|_1 = \max\limits_{1\le i\le n} \displaystyle\sum_{j=1}^{n} |a_{i,j}| \\ \|x\|_2 := \sqrt{\displaystyle\sum_{k=1}^{n} x_k^2} & \|A\|_2 = \sqrt{r_\sigma(A^*A)} \\ \|x\|_\infty := \max\limits_{1\le k\le n} |x_k| & \|A\|_\infty = \max\limits_{1\le i\le n} \displaystyle\sum_{j=1}^{n} |a_{i,j}| \end{array} \tag{4.1}$$

where $r_\sigma(M)$ denotes the maximum of the magnitudes of the eigenvalues of the matrix M.

Definition 4.1.3. Let X be a linear space. Two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ are *equivalent* if there exists constants $K_0, K_1 > 0$ such that

$$K_0\|x\|_a \le \|x\|_b \le K_1\|x\|_a.$$

It is left to the reader to check that this defines an equivalence relation and furthermore that an alternative statement of the definition is that

$$K_0 \le \|x\|_b \le K_1.$$

for all $x \in X$ such that $\|x\|_a = 1$.

Proposition 4.1.4. *If X is a finite dimensional linear space, then all norms are equivalent.*

Proof. Assume the dimension of X is N. Let $\{e_n \mid n = 1, \dots, N\}$ be a basis for X. Then, for every $x \in X$, there are a unique set of coefficients $\{x_n \mid n = 1, \dots, N\}$ such that

$$x = \sum_{n=1}^{N} x_n e_n.$$

Observe that it is sufficient to prove all norms are equivalent to a single norm. It is left to the reader to check that

$$\|x\|_1 := \sum_{n=1}^{N} |x_n|$$

is a norm. Let $\|\cdot\|$ be an arbitrary norm on X we need to show that there are fixed positive constants K_0 and K_1 such that

$$K_0 \le \|x\| \le K_1 \tag{4.2}$$

for all $x \in X$ such that $\|x\|_1 = 1$. Observe that demonstrating (4.2) is equivalent to showing that $\|\cdot\|$ is a bounded function on the unit sphere S^{n-1} (under the $\|\cdot\|_1$ norm).

Since S^{n-1} is compact it is sufficient to show that $\|\cdot\|\colon V \to \mathbb{R}$ is a continuous function under the topology on V induced by $\|\cdot\|_1$. Fix $\epsilon > 0$ and define

$$\delta = \frac{\epsilon}{\max_n \|e_n\|}.$$

Choose $x, y \in X$ such that $\|x - y\|_1 < \delta$. Then

$$|\|x\| - \|y\|| \leq \|x - y\| \leq \sum_{n=1}^{N} |x_n - y_n| \|e_n\| \leq \|x - y\|_1 \max_n \|e_n\| \leq \epsilon.$$

Hence $\|\cdot\|\colon V \to \mathbb{R}$ is a continuous function. □

4.2 Homogeneous Linear Systems

In this section we study the homogeneous linear system

$$\dot{x} = A(t)x \tag{4.3}$$

where $A\colon \mathbb{R} \to \mathbb{R}^{n\times n}$ is a smooth matrix valued function. More general and more in-depth results can be found in [1, 4].

Proposition 4.2.1. *Given the homogeneous linear system* (4.3) *the solution to the initial value problem $x(t_0) = x_0$ is defined for all $t \in \mathbb{R}$.*

Proof. The proof is by contradiction. The existence of a solution $\varphi\colon J \to \mathbb{R}^n$ follows from Theorem 2.2.12. Assume that $J = (\alpha, \beta)$, is the maximal interval of existence and $\beta < \infty$. By Lemma 2.2.4

$$\varphi(t) = \varphi(t_0) + \int_{t_0}^{t} A(s)\varphi(s)\, ds. \tag{4.4}$$

By assumption A is continuous over $\mathbb{R}$, hence there exists M such that

$$\sup_{t\in[t_0,\beta]} \|A\| \leq M.$$

Applying this bound to (4.4) we obtain

$$|\varphi(t)| \leq |x_0| + \int_{t_0}^{t} M\|\varphi(s)\|\, ds,$$

which by Gronwall's inequality implies

$$|\varphi(t)| \leq |x_0| e^{M(t-t_0)}.$$

In particular, $\varphi(t)$ is bounded on $[t_0, \beta]$. This contradicts Theorem 2.2.18. Therefore $\beta = \infty$.

A similar argument shows that $\alpha = -\infty$. □

The principle of superposition, which follows directly from the linearity, is fundamental to our ability to characterize all solutions to (4.3).

Proposition 4.2.2. *If φ and ψ are solutions to* (4.3)*, then for all scalars α and β,*

$$\alpha\varphi + \beta\psi$$

is a solution.

The principle of superposition allows us to think of linear combinations of solutions, which in turn can be codified as follows. A matrix valued function $X\colon \mathbb{R} \to \mathbb{R}^{n\times n}$ is a *matrix solution* to (4.3) if each column is a solution to (4.3). Observe that if X is a matrix solution, then

$$\dot{X} = A(t)X.$$

Definition 4.2.3. A *fundamental matrix solution* to (4.3) is a matrix solution $X(t)$ of (4.3) such that $\det X(t) \neq 0$. A *principal matrix solution* $X(t)$ at initial time t_0 is a fundamental matrix solution satisfying $X(t_0) = I$, the identity matrix. This principal matrix solution is denoted by $X(t, t_0)$.

Proposition 4.2.4. *Let $X(t)$ be a matrix solution to* (4.3)*. Then, either* $\det X(t) \equiv 0$ *or* $\det X(t) \neq 0$ *for all $t \in \mathbb{R}$.*

Proof. If there exists τ such that $\det X(\tau) = 0$, then $X(\tau)$ is not invertible and hence there exists a nonzero vector v such that $X(\tau)v = 0$. Since $X(t)$ is a matrix solution to (4.3), $X(t)v$ is a solution to (4.3). In particular, it is a solution to the initial value problem $x(\tau) = 0$. However, 0 is an equilibrium for (4.3) and hence $X(t)v = 0$, for all $t \in \mathbb{R}$. Therefore, $\det X(t) = 0$ for all $t \in \mathbb{R}$. □

Corollary 4.2.5. *If $X(t)$ is a matrix solution of* (4.3) *satisfying $X(t_0) = X_0$ where X_0 is a non-singular matrix, then $X(t)$ is a fundamental matrix solution.*

Theorem 4.2.6. *If $X(t)$ is a matrix solution to* (4.3) *satisfying the initial condition $X(t_0) = X_0$, then*

$$\frac{d}{dt}(\det X) = (\operatorname{tr} A(t)) \det X$$

and hence

$$\det X(t) = \det X_0 \, e^{\int_{t_0}^{t} \operatorname{tr} A(s)\, ds}. \tag{4.5}$$

The last equation (4.5) is typically called *Liouville's equation.*

Proof. Let

$$X(t) = \begin{bmatrix} R_1(t) \\ R_2(t) \\ \vdots \\ R_n(t) \end{bmatrix}$$

where $R_j(t)$ denotes the j-th row. Since det is a multilinear function of its rows,

$$\frac{d}{dt}(\det X) = \sum_{i=1}^{n} \det \begin{bmatrix} R_1(t) \\ \vdots \\ \frac{d}{dt}R_i \\ \vdots \\ R_n(t) \end{bmatrix} \tag{4.6}$$

Writing out (4.3) in matrix form gives the expression

$$\frac{d}{dt}x_{i,j} = \sum_k a_{i,k}x_{k,j}$$

from which we can deduce that

$$\frac{dR_k}{dt} - \sum_{m \neq k} a_{i,m}R_m = a_{i,i}R_i.$$

Thus (4.6) becomes

$$\frac{d}{dt}(\det X) = \sum_{i=1}^{n} \det \begin{bmatrix} R_1(t) \\ \vdots \\ a_{i,i}R_i \\ \vdots \\ R_n(t) \end{bmatrix} = (\operatorname{tr} A) \det X$$

where the last equality follows from the multilinearity of det. □

4.3 Constant Coefficient Linear Systems

In this section we study properties of the linear system

$$\dot{x} = Ax \tag{4.7}$$

where A is a fixed $n \times n$ matrix.

We know that if $n = 1$, then $A \in \mathbb{R}$ and the solution to (4.7) takes the form e^{At}. Observe that if $x_0 = 1$, then e^{At} is the principal matrix solution. In analogy to this we define

$$e^{At} := X(t)$$

where $X(t) = X(t, 0)$ is the principal matrix solution to (4.7).

The following proposition provides essential properties of the function e^{At} and indicates this it is equivalent to the standard definition of the exponential map.

Proposition 4.3.1. *Let A and B be $n \times n$ matrices. The function e^{At} satisfies the following properties:*

(i) $e^{A(t+s)} = e^{At}e^{As}$;

(ii) $\left(e^{At}\right)^{-1} = e^{-At}$;

(iii) $\frac{d}{dt}e^{At} = Ae^{At} = e^{At}A$;

(iv) $e^{At} = \sum_{k=0}^{\infty} \frac{1}{k!}A^k t^k$;

(v) If $AB = BA$, then $e^{(A+B)t} = e^{At}e^{Bt}$;

(vi) If B is an invertible matrix, then

$$e^{B^{-1}ABt} = B^{-1}e^{At}B.$$

Proof. (i) Let $s \in \mathbb{R}$. Observe that $e^{A(t+s)}$ and $e^{At}e^{As}$ are fundamental matrix solutions to (4.7) that agree at $t = 0$. Therefore, the desired equality follows from uniqueness of solutions.

(ii) In (i) let $s = -t$.

(iii) The first equality follows from the definition of e^{At}. The second equality follows from the observation that both are matrix solutions to (4.7) that agree at $t = 0$.

(iv) Let $\|A\| = M$ and let $E^{(K)}(t) := \sum_{k=0}^{K} \frac{1}{k!}A^k t^k$. Then

$$\|E^{(K)}(t)x\| \leq \sum_{k=0}^{K} \frac{1}{k!}\|A\|^k t^k \|x\| \leq e^{Mt}\|x\|.$$

Furthermore,

$$\|E^{(K+1)}(t)x - E^{(K)}(t)x\| \leq \frac{1}{(K+1)!}M^{K+1}t^{K+1}\|x\|.$$

Therefore, for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$, the sequence $E^{(K)}(t)x$ converges uniformly as $K \to \infty$. Define

$$\varphi(t, x) := \lim_{K\to\infty} E^{(K)}(t)x.$$

It remains to be shown that $\varphi(t,x) = e^{At}x$.

Observe that

$$E^{(K+1)}(t)x = E^{(0)}(t)x + \int_0^t AE^{(K)}(s)x\,ds$$

Therefore,

$$\lim_{K\to\infty} E^{(K+1)}(t)x = x + \lim_{K\to\infty}\int_0^t AE^{(K)}(s)x\,ds$$

$$\varphi(t,x) = x + \int_0^t \lim_{K\to\infty} AE^{(K)}(s)x\,ds = x + \int_0^t A\varphi(s,x)\,ds,$$

which implies that $\varphi(t,x)$ is a solution to (4.7) with initial condition x. By uniqueness of solutions this implies the desired equality $e^{At}x = \varphi(t,x)$.

(v) Using (iv)

$$Be^{At} = B\sum_{k=0}^{\infty}\frac{1}{k!}A^kt^k = \sum_{k=0}^{\infty}\frac{1}{k!}A^kt^kB = e^{At}B.$$

By definition $e^{(A+B)t}$ is the principal matrix solution to $\dot{x} = (A+B)x$. Observe that

$$\frac{d}{dt}(e^{At}e^{Bt}) = Ae^{At}e^{Bt} + e^{At}Be^{Bt} = Ae^{At}e^{Bt} + Be^{At}e^{Bt} = (A+B)(e^{At}e^{Bt}).$$

Thus the desired equality follows from uniqueness of solutions.

(vi) Since B is invertible it acts as a linear change of coordinates. In the coordinate system $y = B^{-1}x$, (4.7) becomes

$$\dot{y} = B^{-1}ABy$$

with principal matrix solution $e^{B^{-1}ABt}$. Let $y_0 = B^{-1}x_0$ be an initial condition, then

$$x(t) = By(t) = Be^{B^{-1}ABt}y_0 = Be^{B^{-1}ABt}B^{-1}x_0.$$

By uniqueness of solutions

$$e^{At} = Be^{B^{-1}ABt}B^{-1}$$

or equivalently

$$B^{-1}e^{At}B = e^{B^{-1}ABt}.$$ □

Recall that the *real Jordan normal form* is a real valued block diagonal matrix where the diagonal blocks take one of the following four forms

$$[\lambda],\quad \begin{bmatrix}\alpha & -\beta\\ \beta & \alpha\end{bmatrix},\quad \begin{bmatrix}\lambda & 1 & & \\ & \ddots & & \\ & & \ddots & 1\\ & & & \lambda\end{bmatrix},\quad \begin{bmatrix}D & I & & \\ & \ddots & & \\ & & \ddots & I\\ & & & D\end{bmatrix}$$

where I is the 2×2 identity matrix and

$$D = \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix}.$$

Furthermore, up to ordering of the diagonal blocks any real valued matrix A has a unique real Jordan normal form J and there exists an invertible matrix P such that

$$J = PAP^{-1}.$$

Corollary 4.3.2. *Consider two n-dimensional constant coefficient linear systems*

$$\dot{x} = Ax \quad \textit{and} \quad \dot{x} = Bx.$$

If A and B have the same real Jordan normal form then the corresponding flows φ_A and φ_B are topologically equivalent.

Proof. Since A and B have the same real Jordan normal form, there exist P_1 and P_2 invertible matrices such that

$$P_1 A P_1^{-1} = J = P_2 B P_2^{-1}.$$

This implies that

$$P_1 e^{At} P_1^{-1} = e^{P_1 A P_1^{-1} t} = e^{P_2 B P_2^{-1} t} = P_2 e^{Bt} P_2^{-1},$$

and letting $S := P_2^{-1} P_1$ which is invertible, we get

$$e^{Bt} = S e^{At} S^{-1}.$$

Recall that two flows $\varphi\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ and $\psi\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ are topologically equivalent if there exists a homeomorphism $h\colon \mathbb{R}^n \to \mathbb{R}^n$ and a reparameterization function $\tau\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ such that

$$\psi(t, h(x)) = h(\varphi(\tau(t, x), x)),$$

for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$. In this case, letting $\varphi(t, x) = \varphi_A(t, x) := e^{At}x$, $\psi(t, x) = \varphi_B(t, x) := e^{Bt}x$, $h = S$ and $\tau(t, x) = t$ for all $x \in \mathbb{R}^n$, we get the result. In fact, since the reparameterization function is $\tau(t, x) = t$ for all $x \in \mathbb{R}^n$, we have even more: the flows φ_A and φ_B are topologically conjugate, as defined in Definition 4.4.6. □

As is made clear in Section 4.4 the converse of Corollary 4.3.2 is not true.

The following two propositions allow us to give an explicit formula for solutions to a homogeneous linear differential equation.

Proposition 4.3.3. *Let A be a matrix with real entries.*

(i) $\lambda \in \mathbb{R}$ *is an eigenvalue of* A *with associated eigenvector* v *if and only if* $t \mapsto e^{\lambda t}v$ *is a solution to* (4.7)*.*

(ii) If $\lambda = \alpha + i\beta$ *is an eigenvalue of* A *with associated eigenvector* $v = u + iw$*, then*

$$t \mapsto e^{\alpha t}(\cos(\beta t)u - \sin(\beta t)w) \quad \textit{and} \quad t \mapsto e^{\alpha t}(\sin(\beta t)u + \cos(\beta t)w)$$

are solutions to (4.7) *and if* $\beta \neq 0$*, then they are linearly independent.*

The proof of Proposition 4.3.3 is left as an exercise.

Proposition 4.3.4. *Let* $A = \lambda I + N$ *where* I *denotes the identity matrix and* N *is nilpotent, i.e.* $N^k = 0$*, for some* $k \geq 1$*. Then,*

$$e^{At} = e^{\lambda t}\left(\sum_{j=0}^{k-1} \frac{t^j}{j!}N^j\right).$$

Proof. Since $IN = NI$, by Proposition 4.3.1(v)

$$e^{At} = e^{(\lambda I+N)t} = e^{\lambda It}e^{Nt} = e^{\lambda t}\left(\sum_{j=0}^{k-1} \frac{t^j}{j!}N^j\right). \qquad \square$$

Given the homogeneous linear differential equation (4.7), let $J = PAP^{-1}$ denote a Jordan normal form for A. Then, Propositions 4.3.3 and 4.3.4 can be used to express e^{Jt} and by Proposition 4.3.1(vi)

$$e^{At} = P^{-1}e^{Jt}P. \tag{4.8}$$

4.4 Hyperbolic Linear Systems

Definition 4.4.1. A matrix is *hyperbolic* if all of its eigenvalues have nonzero real parts.

The focus of this section is on characterizing the dynamics of constant coefficient linear systems under the assumption that the matrix is hyperbolic. We begin by studying the case where all eigenvalues have negative real part.

Theorem 4.4.2. *Consider the differential equation* $\dot{x} = Ax$ *under the assumption that* A *is a real* $n \times n$ *matrix. The following statements are equivalent.*

(i) There is a norm $\|\cdot\|_a$ *on* $\mathbb{R}^n$ *and a real number* $\lambda > 0$ *such that for all* $x \in \mathbb{R}^n$ *and all* $t \geq 0$*,*

$$\left\|e^{At}x\right\|_a \leq e^{-\lambda t}\|x\|_a.$$

(ii) If $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^n$, then there is a constant $C > 0$ and a real number $\lambda > 0$ such that for all $x \in \mathbb{R}^n$ and all $t \geq 0$,

$$\left\|e^{At}x\right\| \leq Ce^{-\lambda t}\|x\|.$$

(iii) The real part of every eigenvalue of A is negative.

Proof. (i) $\Rightarrow$ (ii) By Proposition 4.1.4 all norms on $\mathbb{R}^n$ are equivalent, thus there exists positive constants K_0 and K_1 such that

$$K_0\|x\| \leq \|x\|_a \leq K_1\|x\|.$$

Therefore,

$$\|e^{At}x\| \leq \frac{1}{K_0}\|e^{At}x\|_a \leq \frac{e^{-\lambda t}}{K_0}\|x\|_a \leq \frac{K_1}{K_0}e^{-\lambda t}\|x\|.$$

(ii) $\Rightarrow$ (iii) The proof is by contradiction. Observe that by (ii) the omega limit set $\omega(x) = \{0\}$ for every $x \in \mathbb{R}^n$. Let $\mu = \alpha + i\beta$ be an eigenvalue of A and assume $\alpha \geq 0$. Let $v = u + iw \neq 0$ be a corresponding eigenvector. By Proposition 4.3.3(ii),

$$t \mapsto e^{\alpha t}(\cos(\beta t)u - \sin(\beta t)w)$$

is a solution. But $\omega(u) \neq 0$, a contradiction.

(iii) $\Rightarrow$ (i) Let $\left\{v^j \mid j = 1, \ldots, n\right\}$ be a set of linearly independent generalized eigenvectors for A. Let $\nu^j \colon \mathbb{R} \to \mathbb{R}^n$ denote the solution to $\dot{x} = Ax$ satisfying $\nu^j(0) = v^j$. Given an initial condition $x = \sum_{j=1}^n c_j v^j$,

$$e^{At}x = \sum_{j=1}^{n} c_j \nu^j(t).$$

Explicitly writing out (4.8) (that is $e^{At} = P^{-1}e^{Jt}P$) allows us to conclude that the components of e^{At} can be written as a sum of terms of the form $p(t)e^{\alpha t}\sin(\beta t)$ and $p(t)e^{\alpha t}\cos(\beta t)$, where $\alpha, \beta \in \mathbb{R}$ are the real and imaginary parts of an eigenvalue of A and the polynomial has degree bounded by the geometric dimension of the corresponding eigenspace. Choose $a > 0$ such that $\alpha < -a$ for all eigenvalues $\alpha + i\beta$ of A. Then for each x there exists $\tau(x)$ (continuous in x) such that

$$\left\|e^{At}x\right\| \leq e^{-at}\|x\|$$

for all $t \geq \tau(x)$. By compactness $\tau := \sup_{\|x\|=1} \tau(x) < \infty$. Thus, for $\|x\| = 1$, if $t > \tau$, then

$$\left\|e^{At}x\right\| \leq e^{-at}.$$

Finally, for arbitrary x

$$\left\|e^{At}x\right\| = \|x\| \left\|e^{At}\frac{x}{\|x\|}\right\| \leq \|x\|e^{-at} \tag{4.9}$$

for all $t \geq \tau$.

At this point we have proved that given the norm $\|\cdot\|$ the flow eventually produces a contraction. We claim that

$$\|x\|_a := \int_0^\tau e^{as} \left\|e^{As}x\right\| \, ds$$

defines a norm that contracts for all positive time. It is left to the reader to check that $\|\cdot\|_a$ is a norm.

Given $t \geq 0$ choose a positive integer n and $0 \leq T < \tau$ such that $t = n\tau + T$. Observe that

$$\begin{aligned}
\left\|e^{At}x\right\|_a &= \int_0^\tau e^{as} \left\|e^{As}e^{At}x\right\| \, ds \\
&= \int_0^\tau e^{as} \left\|e^{A(n\tau+T+s)}x\right\| \, ds \\
&= \int_0^{\tau-T} e^{as} \left\|e^{An\tau}e^{A(T+s)}x\right\| \, ds + \int_{\tau-T}^\tau e^{as} \left\|e^{A(n+1)\tau}e^{A(T-\tau+s)}x\right\| \, ds \\
&\leq \int_0^{\tau-T} e^{a(s-n\tau)} \left\|e^{A(T+s)}x\right\| \, ds + \int_{\tau-T}^\tau e^{a(s-(n+1)\tau)} \left\|e^{A(T-\tau+s)}x\right\| \, ds \\
&= \int_T^\tau e^{a(u-T-n\tau)} \left\|e^{Au}x\right\| \, du + \int_0^T e^{a(u+\tau-T-(n+1)\tau)} \left\|e^{Au}x\right\| \, du \\
&= e^{-at} \int_0^\tau e^{au} \left\|e^{Au}x\right\| \, du \\
&= e^{-at} \left\|x\right\|_a ,
\end{aligned}$$

where the inequality is obtained using (4.9) and we use the substitutions $u = T + s$ in the first integral and $u = T - \tau + s$ in the second integral. Thus (i) holds for the norm $\|\cdot\|_a$. □

Corollary 4.4.3. *If every eigenvalue of A has negative real part, then 0 is an asymptotically stable equilibrium for $\dot{x} = Ax$.*

The fundamental conclusion of Theorem 4.4.2 is that if the every eigenvalue of A has negative real parts then every solution to $\dot{x} = Ax$ converges to the origin exponentially fast and furthermore there is a uniform lower bound on the exponential rate. To extend this theorem and its conclusions to hyperbolic linear equations it is useful to introduce the following subspaces.

Definition 4.4.4. Let A be an $n \times n$ matrix. The *stable, unstable*, and *center* eigenspaces of A are defined by

$$\mathbb{E}^s_A := \text{span}\,\{v \mid v \text{ a generalized eigenvector associated with eigenvalue } \alpha + i\beta,\ \alpha < 0\}$$
$$\mathbb{E}^u_A := \text{span}\,\{v \mid v \text{ a generalized eigenvector associated with eigenvalue } \alpha + i\beta,\ \alpha > 0\}$$
$$\mathbb{E}^c_A := \text{span}\,\{v \mid v \text{ a generalized eigenvector associated with eigenvalue } \alpha + i\beta,\ \alpha = 0\}$$

respectively.

By definition if A is a hyperbolic matrix, then $\mathbb{E}^c = \emptyset$. This implies that

$$\mathbb{R}^n = \mathbb{E}^u_A \oplus \mathbb{E}^s_A,$$

which is called a *hyperbolic splitting*.

Theorem 4.4.5. *Consider the differential equation $\dot{x} = Ax$ where A is an $n \times n$ matrix.*

(i) The subspace $\mathbb{E}^u_A$, $\mathbb{E}^s_A$ and $\mathbb{E}^c_A$ are invariant sets under the flow e^{At}.

(ii) $x \in \mathbb{E}^s_A$ if and only if there exists $a > 0$ and $C \geq 1$ such that

$$\|e^{At}x\| \leq Ce^{-a|t|} \quad \forall\, t \geq 0.$$

(iii) $x \in \mathbb{E}^u_A$ if and only if there exists $a > 0$ and $C \geq 1$ such that

$$\|e^{At}x\| \leq Ce^{-a|t|} \quad \forall\, t \leq 0.$$

(iv) $x \in \mathbb{E}^c_A$ if and only if for all $a > 0$

$$\lim_{t \to \pm\infty} e^{-a|t|}\|e^{At}x\| = 0.$$

Proof. (i) This follows directly from (4.8).

(ii) By (i) $\mathbb{E}^s_A$ is invariant. Restricting the differential equation to $\mathbb{E}^s_A$ implies that the conditions of Theorem 4.4.2(iii) are satisfied and hence the result follows from Theorem 4.4.2(ii).

(iii) This is essentially the same as (ii) after applying the transformation $t \mapsto -t$.

(iv) If $x \in \mathbb{E}^c_A$, then by (4.8) the growth rate in forward or backward time is subexponential.

So assume $x \notin \mathbb{E}^c_A$. Since $\mathbb{R}^n = \mathbb{E}^s_A \oplus \mathbb{E}^U_A \oplus \mathbb{E}^c_A$, we can write $x = x_s + x_v + x_c$ where $x_* \in E^*_A$, $* \in \{s, v, c\}$ and $x_s \neq 0$ or $x_u \neq 0$. By (i), (ii) and (iii), $\|x_s(t)\|$ and $\|x_u(t)\|$ grows at an exponential rate for negative or positive time, respectively, This contradicts $\lim_{t\to\pm\infty} e^{-a|t|}\|e^{At}x\| = 0$. □

Definition 4.4.6. Two flows $\varphi\colon \mathbb{R}\times X \to X$ and $\psi\colon \mathbb{R}\times Y \to Y$ are *topologically conjugate* if there exists a homeomorphism $h\colon X \to Y$ such that

$$h \circ \varphi(t,x) = \psi(t, h(x))$$

for all $t \in$ and $x \in X$.

Observe that if two flows are topologically conjugate, then they are topologically equivalent, but the converse need not be true.

Theorem 4.4.7. *Let $A, B \in M_n(\mathbb{R})$ be hyperbolic matrices.*

(i) If all eigenvalues of A and B have negative real parts, then the flows e^{At} and e^{Bt} are topologically conjugate.

(ii) If the number of eigenvalues with negative real parts are the same for A and B, then the flows e^{At} and e^{Bt} are topologically conjugate.

(iii) There exists $\epsilon > 0$ such that if $\|A-B\| < \epsilon$, then the flows e^{At} and e^{Bt} are topologically conjugate.

Proof. (i) Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be the norms for $\dot{x} = Ax$ and $\dot{y} = By$ guaranteed by Theorem 4.4.2(i), respectively, that is

$$\|e^{At}x\|_a \le e^{-\lambda_a t}\|x\|_a, \qquad \|e^{Bt}y\|_b \le e^{-\lambda_b t}\|y\|_b,$$

for some $\lambda_a, \lambda_b > 0$ and for all $x, y \in \mathbb{R}^n$ and $t \ge 0$.

Let $S_A := \{x \in \mathbb{R}^n \mid \|x\|_a = 1\}$ and $S_B := \{y \in \mathbb{R}^n \mid \|y\|_b = 1\}$. Define $h_0\colon S_A \to S_B$ by

$$h_0(x) = \frac{x}{\|x\|_b}.$$

Because $\|\cdot\|_a$ is strictly decreasing along the orbits $e^{At}x$ we can define $\tau\colon \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$ to be the solution to

$$e^{A\tau(x)}x \in S_A.$$

Note that

$$\begin{cases} \tau(x) > 0, & \text{if } \|x\|_a > 1 \\ \tau(x) = 0, & \text{if } \|x\|_a = 1 \\ \tau(x) < 0, & \text{if } \|x\|_a < 1. \end{cases}$$

Define $h\colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$h(x) := \begin{cases} e^{-B\tau(x)}h_0\left(e^{A\tau(x)}x\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

By construction h is continuous except perhaps at 0. To show continuity at this point, consider a sequence $x_k \to 0$ and observe that this implies that $\tau(x_k) \to -\infty$. Since by definition $\|h_0(e^{A\tau(x_k)}x_k)\|_b = 1$, $\lim_{k\to\infty} h(x_k) = 0 = h(0)$. Thus h is continuous.

That h defines a conjugation follows from the following calculations. If $x = 0$, then

$$h(e^{At}0) = 0 = e^{Bt}0 = e^{Bt}h(0).$$

If $x \neq 0$, then

$$\begin{aligned} h(e^{At}x) &= e^{-B\tau(e^{At}x)}h_0\left(e^{A\tau(e^{At}x)}e^{At}x\right) \\ &= e^{-B(\tau(x)-t)}h_0\left(e^{A(\tau(x)-t)}e^{At}x\right) \\ &= e^{Bt}e^{-B\tau(x)}h_0\left(e^{A\tau(x)}e^{-At}e^{At}x\right) \\ &= e^{Bt}e^{-B\tau(x)}h_0\left(e^{A\tau(x)}x\right) \\ &= e^{Bt}h(x). \end{aligned}$$

It is left to the reader to prove that h is a homeomorphism.

(ii) Since A and B are hyperbolic we write them as $A = A_s \oplus A_u$ and $B = B_s \oplus B_u$ by restricting them to their stable and unstable eigenspaces $\mathbb{E}^s$ and $\mathbb{E}^u$, respectively. Assume $\dim \mathbb{E}^s = k$. Let $h_s \colon \mathbb{R}^k \to \mathbb{R}^k$ and $h_u \colon \mathbb{R}^{n-k} \to \mathbb{R}^{n-k}$ be defined as in (i) using A_s and B_s, and A_u and B_u, respectively. The reader can check that $h := h_s \oplus h_u$ is a conjugacy.

(iii) This follows from Exercise 4.7.5 where it is shown that hyperbolic matrices are generic. □

4.5 Linear Approximations of Nonlinear Systems

The goal of this section is to prove that if $\tilde{x}$ is an equilibrium to $\dot{x} = f(x)$, and the real parts of all the eigenvalues of $Df(\tilde{x})$ are negative, then $\tilde{x}$ is asymptotically stable. We also discuss the case where $Df(\tilde{x})$ is a hyperbolic matrix.

Proposition 4.5.1. *Let $x(t)$ be the unique solution of*

$$\dot{x} = Ax + g(t), \qquad x(t_0) = x_0. \tag{4.10}$$

Then

$$x(t) = e^{At}\left(e^{-At_0}x_0 + \int_{t_0}^{t} e^{-As}g(s)\,ds\right). \tag{4.11}$$

Equation (4.11) is called the *variation of constants* formula.

Proof. Set

$$y(t) := e^{-At}x(t)$$

and observe that

$$\dot{y} = -Ae^{-At}x + e^{-At}\dot{x} = -Ae^{-At}x + e^{-At}(Ax + g(t)) = e^{-At}g(t). \tag{4.12}$$

An initial condition $x(t_0) = x_0$ to (4.10) is equivalent to an initial condition $y(t_0) = e^{-At_0}x_0$ to (4.12). We can solve (4.12) by direct integration, i.e.

$$y(t) = e^{-At_0}x_0 + \int_{t_0}^{t} e^{-As}g(s)\,ds.$$

Rewriting this in the original coordinates results in equation (4.11). □

Theorem 4.5.2. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ a C^1 function and let $\tilde{x}$ be an equilibrium solution to*

$$\dot{x} = f(x), \quad x \in \mathbb{R}^n. \tag{4.13}$$

If all the eigenvalues of $Df(\tilde{x})$ have negative real part strictly less than $-a$, for some $a > 0$, then for any norm $\|\cdot\|$ on $\mathbb{R}^n$ there exists a neighborhood U of $\tilde{x}$ and a constant $C \geq 1$ such that for any initial condition $x_0 \in U$ the solution $\varphi(\cdot, x_0) \colon [0, \infty) \to U$ satisfies

$$\|\varphi(t, x_0) - \tilde{x}\| \leq Ce^{-at}\|x_0 - \tilde{x}\|, \quad \forall t \geq 0.$$

Proof. Let $y(t) = x(t) - \tilde{x}$. We are interested in the dynamics in a neighborhood of $\tilde{x}$. Observe that using a Taylor approximation (4.13) can be written as

$$\dot{y} = f(\tilde{x} + y) = f(\tilde{x}) + Df(\tilde{x})y + g(y) = Df(\tilde{x})y + g(y) \tag{4.14}$$

where $g(0) = 0$ and $Dg(0) = 0$.

Let $y(t)$ be a solution to (4.14) with initial condition $y(0) = y_0$ and view $g(y(t))$ as a function of time. Using the variation of constants formula (4.11)

$$\begin{aligned} y(t) &= e^{At}\left(y_0 + \int_0^t e^{-As}g(y(s))\,ds\right) \\ &= e^{At}y_0 + \int_0^t e^{A(t-s)}g(y(s))\,ds. \end{aligned}$$

Since the negative real parts of the eigenvalues of A are strictly less than $-a$, there exists $\lambda = a + \delta$ for some $\delta > 0$ and a constant $C \geq 1$ which satisfy Theorem 4.4.2(ii), that is

$$\left\|e^{At}x\right\| \leq Ce^{-\lambda t}\|x\|, \tag{4.15}$$

for all $x \in \mathbb{R}^n$ and all $t \geq 0$.

Choose m such that $mC < \delta$. Because $g(0) = 0 \in \mathbb{R}^n$ and $Dg(0) = 0 \in M_n(\mathbb{R})$ there exists $\epsilon > 0$ such that

$$\|g(y)\| \leq m\|y\| \quad \text{for all } \|y\| < \epsilon.$$

By (4.15),

$$\|y(t)\| \leq Ce^{-\lambda t}\|y_0\| + \int_0^t Ce^{-\lambda(t-s)}m\|y(s)\|\,ds$$

as long as $\|y(s)\| < \epsilon$ for $s \in [0, t]$. Thus,

$$e^{\lambda t}\|y(t)\| \leq C\|y_0\| + \int_0^t Ce^{\lambda s}m\|y(s)\|\,ds.$$

By Gronwall's inequality (with $\alpha(t) = e^{\lambda t}\|y(t)\|$ and $\beta(t) = Cm$),

$$e^{\lambda t}\|y(t)\| \leq C\|y_0\|e^{Cmt},$$

or equivalently

$$\|y(t)\| \leq C\|y_0\|e^{-(\lambda - Cm)t} < C\|y_0\|e^{-at}. \tag{4.16}$$

Observe that if $\|y_0\| \leq \epsilon/C$, then (4.16) guarantees that $\|y(t)\| < \epsilon$ for all $t > 0$. □

To understand the dynamics near arbitrary equilibria it is useful to introduce the following notation. Let $\varphi\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ denote the flow associated with (4.13). The *stable* and *unstable* sets for $\tilde{x}$ are defined to be

$$W^s(\tilde{x}) := \left\{x \in \mathbb{R}^n \mid \lim_{t\to\infty} \varphi(t, x) = \tilde{x}\right\} \tag{4.17}$$

$$W^u(\tilde{x}) := \left\{x \in \mathbb{R}^n \mid \lim_{t\to-\infty} \varphi(t, x) = \tilde{x}\right\}, \tag{4.18}$$

respectively. Given a neighborhood V of $\tilde{x}$ the associated *local* stable and unstable manifolds are given by

$$W^s_{\text{loc}}(\tilde{x}) := W^s(\tilde{x}, V) := \{x \in W^s(\tilde{x}) \mid \varphi([0, \infty), x) \subset V\} \tag{4.19}$$

$$W^u_{\text{loc}}(\tilde{x}) := W^u(\tilde{x}, V) := \{x \in W^u(\tilde{x}) \mid \varphi((-\infty, 0], x) \subset V\}, \tag{4.20}$$

respectively.

The following theorems are fundamental. We will provide a constructive proof of Theorem 4.5.3 in the analytic setting in Chapter **??**. The proof of Hartman-Grobman can be found in [1, 6]. Consider

$$\dot{x} = Ax + g(x) \tag{4.21}$$

where $A \in M_n(\mathbb{R})$ and $g \in C^r(\mathbb{R}^n, \mathbb{R}^n)$ satisfies $g(0) = 0$.

Theorem 4.5.3. **[(Un)Stable Manifold Theorem]** *Consider* (4.21) *and assume A is hyperbolic of the form*

$$A = \begin{bmatrix} A^s & 0 \\ 0 & A^u \end{bmatrix}$$

where $A^s \in M_d(\mathbb{R})$ and $A^s \in M_{n-d}(\mathbb{R})$, and the real parts of the all the eigenvalues of A^s and A^u are negative and positive, respectively. Then, there there exists a neighborhood $U = U_s \times U_u \subset \mathbb{R}^d \times \mathbb{R}^{n-d}$ of 0 and C^r functions $P^s \colon U_s \to \mathbb{R}^n$ and $P^u \colon U_u \to \mathbb{R}^n$ tangent to $\mathbb{R}^d \times 0$ and $0 \times \mathbb{R}^{n-d}$ at the point 0, respectively, such that

$$W^s_{loc}(0) = P^s(U_s) \quad \text{and} \quad W^u_{loc}(0) = P^u(U_u).$$

Let $C^r_b(\mathbb{R}^n, \mathbb{R}^n)$ denote the set of bounded functions in $C^r(\mathbb{R}^n, \mathbb{R}^n)$. Set

$$B^r_\mu(\mathbb{R}^n, \mathbb{R}^n) := \{ f \in C^r_b(\mathbb{R}^n, \mathbb{R}^n) \mid \|f\|_{C^r} < \mu \}.$$

Theorem 4.5.4. **[Hartman-Grobman Theorem]** *Consider* (4.21) *and assume A is hyperbolic. Let φ be the flow generated by* (4.21) *and let ψ be the flow generated by $\dot{x} = Ax$. Then, there exists $\mu > 0$ such that for any $g \in B^1_\mu(\mathbb{R}^n, \mathbb{R}^n)$ there exists a homeomorphism*

$$h = h(g) = I + \rho(g)$$

where $\rho \colon B^1_\mu \to C^0_b$ is continuous, unique, and $\rho(0) = 0$, and for all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$

$$h(\varphi(t, x)) = \psi(t, h(x)).$$

4.6 Rigorous Computation of Eigenvalues and Eigenvectors

As became clear in the previous section, determining the linear stability of an equilibrium solution requires knowledge of the eigenvalues and determining a linear approximation of the stable and unstable manifolds requires knowledge of the eigenvectors of a matrix. In practice, for all but the simplest systems, computation of eigenvalues and eigenvectors must be done numerically and there are a variety of techniques for doing so [**?**, Chapter 9]. With this in mind we consider the problem of rigorously determining solutions (λ, v) of the equation

$$Av = \lambda v, \tag{4.22}$$

for a given matrix $A \in \mathbb{C}^{n \times n}$ under the assumption that an approximate eigenpair $(\bar{\lambda}, \bar{v})$ has been determined numerically. In particular, we show how the radii polynomial approach can be used to obtain these quantities rigorously. For sake of simplicity we do not present how to verify generalized eigenvectors.

Recall from Section 3.2 that radii polynomials provide a domain of existence of a unique zero of a function. In this case the most obvious function is

$$f(\lambda, v) := Av - \lambda v$$

where $f\colon \mathbb{C}^{n+1} \to \mathbb{C}^{n+1}$. However, given a solution $(\tilde{\lambda}, \tilde{v})$ of (4.22) and any $\theta \in \mathbb{C} \setminus \{0\}$, $(\tilde{\lambda}, \theta\tilde{v})$ is also a solution. This implies that the solution $(\tilde{\lambda}, \tilde{v})$ is not *isolated*, i.e. that there is no neighborhood of $(\tilde{\lambda}, \tilde{v})$ in $\mathbb{C}^{n+1}$ on which $(\tilde{\lambda}, \tilde{v})$ is the unique solution to $f(\tilde{\lambda}, \tilde{v}) = 0$. To address this issue, we introduce the notion of a *phase condition* which will ensure that solutions are isolated. Observe that uniqueness fails along a two dimensional parameter space $\mathbb{C} \setminus \{0\}$, thus one expects to obtain uniqueness by reducing by two the dimension of the space on which the function f acts. With this in mind we choose a phase condition that involves fixing one of the components of v to be a given constant.

To be more precise, suppose that an approximate eigenpair of A has been computed, that is $(\bar{\lambda}, \bar{v})$ such that $A\bar{v} \approx \bar{\lambda}\bar{v}$. Choose k such that

$$|\bar{v}_k| = \max\{|\bar{v}_j| \mid j = 1, \ldots, n\}$$

and define $f : \mathbb{C}^n \to \mathbb{C}^n$ by

$$f(x) := A \begin{bmatrix} v_1 \\ \vdots \\ \bar{v}_k \\ \vdots \\ v_n \end{bmatrix} - \lambda \begin{bmatrix} v_1 \\ \vdots \\ \bar{v}_k \\ \vdots \\ v_n \end{bmatrix} \tag{4.23}$$

where $x = (\lambda, v_1, v_2, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n)$. By definition, a solution x of $f(x) = 0$ corresponds to an eigenpair (λ, v) of A with the eigenvalue λ given by the first component of x and the eigenvector $v = (v_1, \ldots, v_{k-1}, \bar{v}_k, v_{k+1}, \ldots, v_n)$.

Continuing to follow the radii polynomial approach we define the operator $T : \mathbb{C}^n \to \mathbb{C}^n$ by

$$T(x) = x - Rf(x), \tag{4.24}$$

where R is a numerical inverse of $Df(\bar{x})$. We assume that R is invertible. For the purpose of constructing the necessary bounds for the radii polynomials we make use of the norm $||x||_\infty = \max_{i=1,\ldots,n}\{|x_i|\}$.

To simplify the expression of $Df(\bar{x})$ we introduce the following notation. Given a matrix $M \in \mathbb{C}^{n\times m}$, we let $(M)_{\widehat{k}}$ denote the $n \times (m-1)$ matrix obtained by deleting the k-th column of M.

At $\bar{x} = (\bar{\lambda}, \bar{v}_1, \bar{v}_2, \ldots, \bar{v}_{k-1}, \bar{v}_{k+1}, \ldots, \bar{v}_n)$

$$Df(\bar{x}) = \left[\begin{array}{c|c} -\bar{v}_1 & \\ \vdots & \\ -\bar{v}_k & (A - \bar{\lambda}I_n)_{\widehat{k}} \\ \vdots & \\ -\bar{v}_n & \end{array}\right]. \tag{4.25}$$

To apply the radii polynomial result, i.e. Corollary 3.2.3, we need to obtain the bounding vectors Y and $Z(r)$. Since $T(\bar{x}) - \bar{x} = -Rf(\bar{x})$, let

$$Y := |Rf(\bar{x})| \in \mathbb{R}^n_+. \tag{4.26}$$

To obtain a bound $Z(r)$ satisfying (3.3) we note that

$$\begin{aligned} DT(\bar{x}+b)c &= (I - RDf(\bar{x}+b))\, c \\ &= (I - RDf(\bar{x}))c + R[(Df(\bar{x}) - Df(\bar{x}+b))c] \end{aligned}$$

which implies that

$$|DT(\bar{x}+b)c| \ll |(I - RDf(\bar{x}))c| + |R[(Df(\bar{x}) - Df(\bar{x}+b))c]|. \tag{4.27}$$

Observe that

$$Df(\bar{x}+b) = \left[\begin{array}{c|c} \begin{matrix} -\bar{v}_1 - b_2 \\ \vdots \\ -\bar{v}_{k-1} - b_k \\ -\bar{v}_k \\ -\bar{v}_{k+1} - b_{k+1} \\ \vdots \\ -\bar{v}_n - b_n \end{matrix} & (A - (\bar{\lambda} + b_1)I_n)_{\widehat{k}} \end{array}\right]$$

Thus,

$$Df(\bar{x}) - Df(\bar{x}+b) = \left[\begin{array}{c|c} \begin{matrix} b_2 \\ \vdots \\ b_k \\ 0 \\ b_{k+1} \\ \vdots \\ b_n \end{matrix} & b_1(I_n)_{\widehat{k}} \end{array}\right].$$

Define

$$Z(r) = rZ_0 + r^2 Z_1, \tag{4.28}$$

where

$$Z_0 := |I_n - R \cdot Df(\bar{x})|\mathbb{1}_n, \quad Z_1 := 2|R|(\mathbb{1}_n - e_k), \tag{4.29}$$

with e_k the k-th element of the canonical basis of $\mathbb{R}^n$.

Returning to (4.27), it is left to the reader to check that

i) $\sup_{c \in B(r)} |(I - RDf(\bar{x}))c| \ll rZ_0$

ii) $\sup_{b,c\in B(r)} |R[(Df(\bar{x}) - Df(\bar{x}+b))c]| \ll r^2 Z_1.$

This combined with (4.26) guarantees that (3.3) is satisfied. Therefore, if the existence radius of the radii polynomials

$$p_k(r) := Y_k + Z_k(r) - r$$

is non-empty then we have proven the existence of an eigenpair $(\tilde{\lambda}, \tilde{v})$ for A.

Example 4.6.1. Recall that the Lorenz system is given by

$$\begin{cases} \dot{x}_1 = \sigma(x_2 - x_1) \\ \dot{x}_2 = \rho x_1 - x_2 - x_1 x_3 \\ \dot{x}_3 = -\beta x_3 + x_1 x_2. \end{cases}$$

For any $\beta > 0$ and $\rho > 1$, the set of equilibria of (3.13) is given by

$$\left\{(0,0,0), \left(\pm\sqrt{\beta(\rho-1)}, \pm\sqrt{\beta(\rho-1)}, \rho - 1\right)\right\}.$$

The stability of each hyperbolic equilibrium solution $x = (x_1, x_2, x_3)$ is determined by the eigenvalues of $Df(x)$. For instance, at the *positive eye* $x = \left(\sqrt{\beta(\rho-1)}, \sqrt{\beta(\rho-1)}, \rho - 1\right)$, the jacobin matrix is given by

$$Df(x) = \begin{pmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & -\sqrt{\beta(\rho-1)} \\ \sqrt{\beta(\rho-1)} & \sqrt{\beta(\rho-1)} & -\beta \end{pmatrix}.$$

For the classical parameter values $\sigma = 10$, $\beta = \frac{8}{3}$ and $\rho = 28$, Numerical calculations lead to the following candidates for eigenvalues

$$\bar{\lambda}_1 = -13.854577914596042, \quad \bar{\lambda}_{2,3} = 0.093955623964686 \pm 10.194505220927850i$$

and their associated eigenvectors

$$\bar{v}^1 = \begin{pmatrix} 0.855665024602210 \\ -0.329822750612395 \\ -0.398816146677856 \end{pmatrix}, \; \bar{v}^{2,3} = \begin{pmatrix} -0.266119316830765 \pm 0.295010166256352i \\ 0.032128610535763 \pm 0.569077429163104i \\ -0.719213558699417 \end{pmatrix}.$$

Consider the approximation $(\bar{\lambda}_1, \bar{v}^1)$. The component of $\bar{v}^1$ with largest magnitude is $\bar{v}^1_1$. Defining the radii polynomials as

$$p_k(r) = (Z_1)_k r^2 + ((Z_0)_k - 1)r + Y_k, \quad k = 1, 2, 3$$

where Y is defined by (4.26), and Z_0 and Z_1 are given by (4.29), we obtained an interval of existence $I = \left[9.3438 \times 10^{-15}\ ,\ 0.64912\right]$. For $\lambda_{2,3}$ we obtain the interval of existence $\left[5.6641 \times 10^{-15}\ ,\ 0.45322\right]$. This altogether yields a proof that the true eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of $Df(x)$ satisfy

$$\lambda_1 < 0 < Re(\lambda_2) = Re(\lambda_3)$$

and therefore that the positive eye $x = \left(\sqrt{\beta(\rho-1)}, \sqrt{\beta(\rho-1)}, \rho-1\right)$ is an unstable equilibrium solution.

4.7 Exercises

Exercise 4.7.1.

Prove that if $M(v_1(t), v_2(t), \ldots, v_n(t)$ is a multilinear function, then

$$\frac{dM}{dt}(t) = \sum_{k=1}^{n} M\left(v_1(t), \ldots, v_{k-1}(t), \frac{dv_k}{dt}(t), v_{k+1}, \ldots, v_n(t)\right)$$

Exercise 4.7.2. Prove Proposition 4.3.3.

Exercise 4.7.3. Consider $\dot{x} = Ax$ and assume the real parts of all eigenvalues of A are negative.

(i) Prove that $\{x \mid \|x\| \leq 1\}$ is an attracting neighborhood for 0.

(ii) Prove that there exists a trapping region containing 0.

Exercise 4.7.4. Complete the proof of Theorem 4.4.7(i) and show that the conjugacy $h\colon \mathbb{R}^n \to \mathbb{R}^n$ is a homeomorphism.

Exercise 4.7.5. Let X be a topological space. A set $G \subset X$ is *generic* if G is open and dense in X. Prove that hyperbolic matrices are open and dense in the space of linear maps on $\mathbb{R}^n$.

Hint: Given $A \in M_n(\mathbb{R})$, show that $\|A\| := \max_{1 \leq i,j \leq n} |a_{i,j}|$ defines a metric on $M_n(\mathbb{R})$.

Chapter 5

Continuation of Equilibria

5.1 Parameterized Families of Equilibria

Radii polynomials are introduce in Chapter 3 as a means of rigorously verifying the existence of an equilibrium to an ODE of the form $\dot{x} = f(x)$. However, in applications parameters play a fundamental role and thus it is important to be able to find equilibria for

$$\dot{x} = f(x, \lambda),$$

where $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ takes the form $(x, \lambda) \mapsto f(x, \lambda)$. Assume that $f(x_0, \lambda_0) = 0$. An obvious question is under what conditions does there exists a neighborhood $V \subset \mathbb{R}^m$ of λ_0 and a continuous function $\alpha\colon V \to \mathbb{R}^n$ such that $\alpha(\lambda_0) = x_0$ and $f(\alpha(\lambda), \lambda) = 0$. The implicit function theorem provides an answer.

Theorem 5.1.1. [Implicit Function Theorem] *Let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$ be open sets. Assume that $f\colon U \times V \to \mathbb{R}^n$ is C^1 and that $f(x_0, y_0) = 0$ for $(x_0, y_0) \in U \times V$. If $D_x f(x_0, y_0)$ is invertible, then there exists neighborhoods $U_0 \subset U$ and $V_0 \subset V$ of x_0 and λ_0, respectively, and a C^1 function $\alpha\colon V_0 \to U_0$ such that $f(\alpha(y), y) = 0$. Furthermore, if $(x, y) \in U_0 \times V_0$ and $f(x, y) = 0$, then $x = \alpha(y)$.*

We take a slightly different, more restrictive approach in this Chapter and assume that $\lambda \in \mathbb{R}$. Consider the following initial value problem

$$\frac{d\alpha}{d\lambda} = F(\alpha, \lambda) := -\left[D_x f(\alpha, \lambda)\right]^{-1} \frac{\partial f}{\partial \lambda}(\alpha, \lambda), \quad \alpha(\lambda_0) = x_0, \tag{5.1}$$

and assume that $f(x_0, \lambda_0) = 0$. By Theorem 2.2.12, if $D_x f(\alpha, \lambda)$ is invertible and if F is smooth, then there exists a smooth solution $\alpha\colon J \to \mathbb{R}^n$ where $J \subset \mathbb{R}$ is a neighborhood of

λ_0. To understand the significance of this observe that (5.1) can be rewritten as

$$\begin{aligned} D_x f(\alpha,\lambda)\frac{d\alpha}{d\lambda} &= -\frac{\partial f}{\partial \lambda}(\alpha,\lambda) \\ D_x f(\alpha,\lambda)\frac{d\alpha}{d\lambda} + \frac{\partial f}{\partial \lambda}(\alpha,\lambda) &= 0 \\ \frac{d}{d\lambda} f(\alpha,\lambda) &= 0 \\ f(\alpha,\lambda) &= C \end{aligned}$$

for some constant C. However, since we are assuming that $f(\alpha(\lambda_0),\lambda_0) = 0$, $C = 0$ and hence

$$f(\alpha(\lambda),\lambda) = 0. \tag{5.2}$$

Notice that we have essentially recovered a simplified form of the inverse function theorem.

While we cannot expect to obtain an exact closed form solution to (5.2) we can use the fact that α is the solution to a differential equation to compute an approximate solution. The most straightforward approach is to make use of the Euler approximation and set

$$x_1 := x_0 + hF(x_0,\lambda_0).$$

For sufficiently small step size $|h| > 0$ it is reasonable to assume that $f(x_1,\lambda_1) \approx 0$ where $\lambda_1 = \lambda_0 + h$. In fact, it is also reasonable to assume that the Newton operator

$$T(x) := x - D_x f(x_1,\lambda_1)^{-1} f(x,\lambda_1)$$

is a contraction mapping and thus that $0 \leq \|f(T(x_1),\lambda_1)\|_\infty < \|f(x_1,\lambda_1)\|_\infty$, i.e. $T(x_1)$ is a better approximation of a root of f at λ_1 than $\widehat{x}_1$. This technique of approximating equilibria is a standard numerical approach and is often referred to as a *predictor-corrector method*. The Euler step is used to predict a new root and Newton's method is used to improve upon the prediction.

Example 5.1.2. To make the above discussion more concrete consider

$$\dot{x} = f(x,\lambda) = x^2 - \lambda.$$

Example 3.2.7 allows us to conclude that $f(\tilde{x}(2),2) = 0$ for a unique value of $\tilde{x}(2) \in \overline{B_{1.3}(0.136)}$. Let us choose $(x_0,\lambda_0) = (x(\lambda_0),\lambda_0) = (1.3,2)$ as an initial condition for

$$\frac{dx}{d\lambda} = F(x,\lambda) = -\left[D_x f(x,\lambda)\right]^{-1}\frac{\partial f}{\partial \lambda}(x,\lambda) = \frac{1}{2x}.$$

Applying the Euler approximation with $h = 1$ we obtain

$$x_1 = x_0 + hF(x_0,\lambda_0) = x_0 + \frac{1}{2x_0} = 1.3 + \frac{1}{2(1.3)} \approx 1.68.$$

For the associated Newton operator at $\lambda_1 = \lambda_0 + h = 2 + 1 = 3$, we choose

$$T(x) = x - 0.3f(x, 3)$$

since $D_x f(x_1, \lambda_1)^{-1} = (2 \cdot 1.68)^{-1} \approx 0.3$. Set

$$x(3) = 1.73 \approx T(x_1).$$

We would like to know if we have found a reasonable representative of an equilibrium at $\lambda = 3$. To this we apply the method of radii polynomials. We use the mapping

$$T(x) := x - Af(x, 3)$$

where $A = 0.3$ since $D_x f(x(3), 3)^{-1} = (2 \cdot 1.73)^{-1} \approx 0.3$. The same analysis as in Example 3.2.7 leads to the radii polynomial

$$\begin{aligned} p(r) &= \left|0.3(1.73^2 - 3)\right| + (|1 - 2(0.3)(1.73)| - 1)r + 2(0.3)r^2 \\ &= 0.6r^2 - 0.962r + 0.00213 \end{aligned}$$

which is negative over the interval $I = [0.003, 1.6]$ and hence there exists $\tilde{x}(3) \in \overline{B_{1.73}(0.003)}$ such that $f(\tilde{x}(3), 3) = 0$.

Given that $\sqrt{3} \approx 1.73205080756888$ the computation in Example 5.1.2 is a fairly good result. The problem is that we have only established the existence and bounds on the roots $\tilde{x}(2)$ and $\tilde{x}(3)$, whereas we would like to have existence and bounds for $\tilde{x}(\lambda)$ for $\lambda \in [2, 3]$.

5.2 Computing Branches of Equilibria

Consider the differential equation

$$\dot{x} = f(x, \lambda),$$

where $f\colon \mathbb{R}^n \times \Lambda \to \mathbb{R}^n$ and $\Lambda \subset \mathbb{R}$. Assume that $f \in C^1(\mathbb{R}^n \times \Lambda, \mathbb{R}^n)$. Assume we have approximate solutions to $f(x, \lambda) = 0$ given by $(\bar{x}_0, \lambda_0)$ and $(\bar{x}_1, \lambda_1)$. Let

$$\lambda_s := (1 - s)\lambda_0 + s\lambda_1 \quad \text{and} \quad \bar{x}_s := (1 - s)\bar{x}_0 + s\bar{x}_1. \tag{5.3}$$

We want to use radii polynomials to conclude about the existence of a smooth curve of equilibria $\tilde{x}_s$ lying in a region centered about the segment $\{\bar{x}_s : s \in [0, 1]\}$.

Theorem 5.2.1. *Consider the differential equation*

$$\dot{x} = f(x, \lambda),$$

where $f \in C^k(\mathbb{R}^n \times \Lambda, \mathbb{R}^n)$, $k \geq 2$, and $\Lambda \subset \mathbb{R}$. Let $\bar{x}_s$ and λ_s as in (5.3). Set

$$B_r := \bigcup_{s \in [0,1]} \overline{B_r(\bar{x}_s)} \times \{\lambda_s\}.$$

Define $T : \mathbb{R}^n \times [0,1] \to \mathbb{R}^n$ *to be a Newton-like operator*

$$T(x,s) := x - Af(x,\lambda_s)$$

where $A \in M_n(\mathbb{R})$. *For* $k = 1,\dots,n$, *define*

$$p_k(r) := Y_k + Z_k(r) - r$$

where Y_k *and* $Z_k(r)$ *satisfy*

$$|[T(\bar{x}_s,s) - \bar{x}_s]_k| \le Y_k \quad \textit{and} \quad \sup_{b,c\in\overline{B_r(0)}} |D_x T_k(\bar{x}_s + b, s)c| \le Z_k(r), \tag{5.4}$$

for all $s \in [0,1]$.

If there exists $r_0 > 0$, *such that* $p_k(r_0) < 0$, *for all* $k = 1,\dots,n$, *and* $D_x f(x,\lambda)$ *is invertible for all* $(x,\lambda) \in B_{r_0}$, *then there exists a* C^{k-1} *function* $\alpha\colon [\lambda_0,\lambda_1] \to B_{r_0}$ *such that*

$$f(\alpha(\lambda),\lambda) = 0$$

and if $(x,\lambda) \in B_{r_0}$ *and* $f(x,\lambda) = 0$, *then* $x = \alpha(\lambda)$.

Proof. Because $p_k(r_0) < 0$, for all $k = 1,\dots,n$, by continuity, there exists $0 < r_- < r_0 < r_+$ such that if $r \in (r_-, r_+)$, then $p_k(r) < 0$, for all $k = 1,\dots,n$. Furthermore, by Corollary 3.2.3 for each $s \in [0,1]$, there exists a unique $\tilde{x}_s \in \overline{B_{r_-}(\bar{x}_s)}$ such that $f(\tilde{x}_s,\lambda_s) = 0$ and for all $x \in B_{r_+}(\bar{x}_s) \setminus B_{r_-}(\bar{x}_s)$, $f(x,\lambda_s) \neq 0$. It remains to be shown that $\tilde{x}_s$ lies on a smooth curve.

Let $\alpha\colon (\lambda_-,\lambda_+) \to \mathbb{R}^n$ be the maximal solution to the initial value problem

$$\frac{d\alpha}{d\lambda} = -\left[D_x f(\alpha,\lambda)\right]^{-1} \frac{\partial f}{\partial \lambda}(\alpha,\lambda), \quad \alpha(\lambda_{1/2}) = \tilde{x}_{1/2}.$$

Observe that by (5.2), $f(\alpha(\lambda),\lambda) = 0$. Our goal is to show that $\lambda_- < \lambda_0 < \lambda_1 < \lambda_+$ and that $\alpha([\lambda_0,\lambda_1]) \subset B_r$.

Assume $\lambda_+ < \lambda_1$. By assumption $D_x f(x,\lambda)$ is invertible on B_{r_0}, and hence, by continuity, $D_x f(x,\lambda)$ is invertible on a neighborhood of B_{r_0}. By Theorem 2.2.18, this implies that there exists $\lambda_s \in (\lambda_{1/2},\lambda_+)$ such that $\alpha(\lambda_s) \in B_{r_+}(\bar{x}_s) \setminus B_{r_-}(\bar{x}_s)$. However, by (5.2) $f(\alpha(\lambda_s),\lambda_s) = 0$, a contradiction. Thus, $\lambda_1 < \lambda_+$. Observe, that the same argument also implies that $\alpha([\lambda_{1/2},\lambda_1]) \subset B_{r_0}$. Yet another application of this argument allows us to conclude that $\alpha([\lambda_0,\lambda_1]) \subset B_{r_0}$. □

Example 5.2.2. Consider the Lorenz system, fix the values of σ and β, and leave ρ as a parameter. Denote the right-hand side of the system by

$$f(x,\rho) := \begin{pmatrix} \sigma(x_2 - x_1) \\ \rho x_1 - x_2 - x_1 x_3 \\ -\beta x_3 + x_1 x_2 \end{pmatrix}.$$

Assume that at $\rho = \rho_i$, we have an approximate solution $\bar{x}_i$ for $i = 0, 1$. As in (5.3), define $\rho_s = (1-s)\rho_0 + s\rho_1$, and $\bar{x}_s = (1-s)\bar{x}_0 + s\bar{x}_1$.

Following (5.4), we first find a bound for $|T(\bar{x}_s, s) - \bar{x}_s| = |Af(x_s, \rho_s)|$. To simplify notation define $g : [0,1] \to \mathbb{R}^3$ by $g(s) = f(\bar{x}_s, \rho_s)$. Since g is quadratic in s,

$$\begin{aligned} g(s) &= g(0) + g'(0)s + \frac{1}{2}g''(0)s^2 \\ &= f(\bar{x}_0, \rho_0) + [D_x f(\bar{x}_0, \rho_0)(\bar{x}_1 - \bar{x}_0) + D_\rho f(\bar{x}_0, \rho_0)(\rho_1 - \rho_0)]\, s \\ &\quad + \frac{1}{2}\left[D_x^2 f(\bar{x}_0, \rho_0)(\bar{x}_1 - \bar{x}_0, \bar{x}_1 - \bar{x}_0)\right. \\ &\qquad \left. + D_{\rho,x} f(\bar{x}_0, \rho_0)(\bar{x}_1 - \bar{x}_0)(\rho_1 - \rho_0) + D_\rho^2 f(\bar{x}_0, \rho_0)(\rho_1 - \rho_0)^2\right] s^2. \end{aligned}$$

Since f is linear in ρ, $D_\rho^2 f(\bar{x}_0, \rho_0)(\rho_1 - \rho_0)^2 = 0$. To further simplify the notation set

$$\begin{aligned} y^{(0)} &:= f(\bar{x}_0, \rho_0) \\ y^{(1)} &:= D_x f(\bar{x}_0, \rho_0)(\bar{x}_1 - \bar{x}_0) + D_\rho f(\bar{x}_0, \rho_0)(\rho_1 - \rho_0) \\ y^{(2)} &:= \frac{1}{2}\left[D_x^2 f(\bar{x}_0, \rho_0)(\bar{x}_1 - \bar{x}_0, \bar{x}_1 - \bar{x}_0) + D_{\rho,x} f(\bar{x}_0, \rho_0)(\bar{x}_1 - \bar{x}_0)(\rho_1 - \rho_0)\right]. \end{aligned}$$

Then,

$$|T(\bar{x}_s, s) - \bar{x}_s| = |Af(x_s, \rho_s)| \ll \left|Ay^{(0)}\right| + \left|Ay^{(1)}\right| + \left|Ay^{(2)}\right|.$$

Choose $Y = (Y_1, Y_2, Y_3)$ such that

$$\left|Ay^{(0)}\right| + \left|Ay^{(1)}\right| + \left|Ay^{(2)}\right| \ll Y. \tag{5.5}$$

In practice this is usually done using an interval arithmetic computation. This provides the first bound of (5.4).

Turning to the second bound of (5.4), observe that

$$\begin{aligned} D_x T(\bar{x}_s + b, s)c &= [I - AD_x f(\bar{x}_s + b, \rho_s)]c \\ &= [I - AD_x f(\bar{x}_0 + sv + b, \rho_0 + s(\rho_1 - \rho_0))]c \end{aligned}$$

where $v := \bar{x}_1 - \bar{x}_0$. Making use of the fact that f is quadratic in x and linear in ρ, the Taylor series expansion of $D_x f(x, \rho)$ at $(\bar{x}_0, \rho_0)$ takes the form

$$D_x f(\bar{x}_0 + sv + b, \rho_0 + s(\rho_1 - \rho_0)) = D_x f(\bar{x}_0, \rho_0) + D_{xx}^2 f(\bar{x}_0, \rho_0)(sv + b) + D_{x\rho}^2 f(\bar{x}_0, \rho_0)(s(\rho_1 - \rho_0)).$$

Note that

$$D_x f(\bar{x}_0, \rho_0) = \begin{bmatrix} -\sigma & \sigma & 0 \\ \rho_0 - \bar{x}_{0,3} & -1 & -\bar{x}_{0,1} \\ \bar{x}_{0,2} & \bar{x}_{0,1} & -\beta \end{bmatrix}$$

$$D_{xx} f(\bar{x}_0, \rho_0)(sv + b) = \begin{bmatrix} 0 & 0 & 0 \\ -sv_3 - b_3 & 0 & -sv_1 - b_1 \\ sv_2 + b_2 & sv_1 + b_1 & 0 \end{bmatrix}$$

$$D^2_{x\rho} f(\bar{x}_0, \rho_0)(s(\rho_1 - \rho_0)) = \begin{bmatrix} 0 & 0 & 0 \\ s(\rho_1 - \rho_0) & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

To explicitly express the powers of r, let $\tilde{b} = b/r$ and $\tilde{c} = b/r$. Then using the Taylor series expansion we obtain

$$\begin{aligned} D_x T(\bar{x}_s + b, s)c &= [I - AD_x f(\bar{x}_s + \tilde{b}r, \rho_s)]\tilde{c}r \\ &= [I - AD_x f(\bar{x}_0, \rho_0)]\tilde{c}r - A \begin{bmatrix} 0 \\ (-sv_3 - \tilde{b}_3 r)\tilde{c}_1 + (-sv_1 - \tilde{b}_1 r)\tilde{c}_3 \\ (sv_2 + \tilde{b}_2 r)\tilde{c}_1 + (sv_1 + \tilde{b}_1 r)\tilde{c}_2 \end{bmatrix} r \\ &\quad - A \begin{bmatrix} 0 \\ s(\rho_1 - \rho_0)\tilde{c}_1 \\ 0 \end{bmatrix} r \\ &= [I - AD_x f(\bar{x}_0, \rho_0)]\tilde{c}r - A \begin{bmatrix} 0 \\ (\rho_1 - \rho_0)\tilde{c}_1 - v_3\tilde{c}_1 - v_1\tilde{c}_3 \\ v_2\tilde{c}_1 + v_1\tilde{c}_2 \end{bmatrix} sr \\ &\quad - A \begin{bmatrix} 0 \\ -\tilde{b}_3\tilde{c}_1 - \tilde{b}_1\tilde{c}_3 \\ \tilde{b}_2\tilde{c}_1 + \tilde{b}_1\tilde{c}_2 \end{bmatrix} r^2. \end{aligned}$$

Thus we define

$$Z^{(1)} := |I - AD_x f(\bar{x}_0, \rho_0)|\, \mathbb{1}_3 + |A| \begin{pmatrix} 0 \\ |\rho_1 - \rho_0| + |v_3| + |v_1| \\ |v_2| + |v_1| \end{pmatrix} \tag{5.6}$$

and

$$Z^{(2)} := 2|A| \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \tag{5.7}$$

and leave it to the reader to check that for all $s \in [0, 1]$,

$$\sup_{b,c \in \overline{B_r(0)}} |D_x T(\bar{x}_s + b, s)c| \ll Z(r) := Z^{(1)} r + Z^{(2)} r^2.$$

To do an explicit computation fix $\sigma = 10$ and $\beta = 8/3$ and at $\rho_0 = 28$ and 29 consider the numerical approximations

$$\bar{x}_0 := \begin{pmatrix} 8.485281374238570 \\ 8.485281374238570 \\ 27 \end{pmatrix} \quad \text{and} \quad \bar{x}_1 := \begin{pmatrix} 8.640987597877146 \\ 8.640987597877146 \\ 28 \end{pmatrix}.$$

Using the numerically computed inverse

$$D_x f(\bar{x}_0, \rho_0)^{-1} \approx A := \begin{pmatrix} -0.051851851851852 & -0.018518518518519 & 0.058925565098879 \\ 0.048148148148148 & -0.018518518518519 & 0.058925565098879 \\ -0.011785113019776 & -0.117851130197758 & 0 \end{pmatrix}.$$

and the bounds (5.5), (5.6) and (5.7), we obtain the radii polynomials $p_k(r) = Z_k^{(2)} r^2 + (Z_k^{(1)} - 1)r + Y_k$, for $k = 1, 2, 3$.

The existence interval contains $I = [0.0886, 3.0762]$. Choosing $r_0 = 0.0886$, we get that for each $s \in [0, 1]$, there is a true equilibrium solution $\tilde{x}(s) \in \overline{B_r(\bar{x}_s)}$, that is $f(\tilde{x}(s), 28+s) = 0$, for all $s \in [0, 1]$.

For the Lorenz equations we have a closed form expression for the this branch of equilibria

$$\tilde{x}(s) = \left(\sqrt{\beta(\rho_s - 1)}, \sqrt{\beta(\rho_s - 1)}, \rho_s - 1\right), \quad s \in [0, 1]$$

and thus we can check that

$$\sup_{s \in [0,1]} \|\tilde{x}(s) - \bar{x}_s\|_\infty \approx 3.539 \times 10^{-4}.$$

This is 2 orders of magnitude smaller than r_0 the optimal radius obtained from the existence interval for the radii polynomials.

5.3 Saddle-Node Bifurcation

Consider a differential equation $f \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ which induces a continuous family of flows

$$\begin{aligned} \varphi \colon \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m &\to \mathbb{R}^n \times \mathbb{R}^m \\ (t, x, \lambda) &\mapsto (\varphi(t, x, \lambda), \lambda) = (\varphi_\lambda(t, x), \lambda). \end{aligned}$$

A parameter value $\bar{\lambda} \in \mathbb{R}^m$ is a *bifurcation point* if given any $\epsilon > 0$ there exists $\lambda \in \mathbb{R}^m$ such that $\|\lambda - \bar{\lambda}\| < \epsilon$ and φ_λ is not topologically equivalent to $\varphi_{\bar{\lambda}}$.

There is no hope of classifying all bifurcation points. However there are some bifurcations that can be detected by changes in the eigenvalue structure of the linearized equation at equilibria. The equilibria for the differential equation $\dot{x} = f(x, \lambda)$, $f \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ are

solutions to $f(x,\lambda)=0$. By the implicit function theorem, if $f(\tilde{x},\tilde{\lambda})=0$ and $D_xf(\tilde{x},\tilde{\lambda})$ is invertible then there exists a smooth family of equilibria $\tilde{x}(\lambda)$ defined in a neighborhood of $\tilde{\lambda}$. This raises the question of what happens when $D_xf(\tilde{x},\tilde{\lambda})$ is not invertible. The lack of invertibility is equivalent to the existence of a zero eigenvalue. The simplest possible setting is a unique zero eigenvalue and a one dimensional parameter space, $m=1$. This leads to the following definitions.

Definition 5.3.1. Consider $f\colon \mathbb{R}^n\times\mathbb{R}\to\mathbb{R}^n$. A *saddle-node* for f is a point $(\tilde{x},\tilde{\lambda})\in\mathbb{R}^n\times\mathbb{R}$ such that

(i) $f(\tilde{x},\tilde{\lambda})=0$, and

(ii) 0 is an eigenvalue of $D_xf(\tilde{x},\tilde{\lambda})$ with algebraic multiplicity one and all other eigenvalues have non-zero real parts.

Definition 5.3.2. Given $f\colon \mathbb{R}^n\times\mathbb{R}\to\mathbb{R}^n$ a *saddle-node bifurcation* occurs at the saddle-node $(\tilde{x},\tilde{\lambda})\in\mathbb{R}^n\times\mathbb{R}$ if the following conditions are met.

SNB1 There exists a smooth curve $g\colon (-\delta,\delta)\to\mathbb{R}^n\times\mathbb{R}$ denoted by $s\mapsto(g_1(s),g_2(s))$ such that $g(0)=(\tilde{x},\tilde{\lambda})$ and $f(g_1(s),g_2(s))=0$.

SNB2 The curve defined by g has a quadratic tangency with $\mathbb{R}^n\times\{\tilde{\lambda}\}$ at $(\tilde{x},\tilde{\lambda})$; that is,

$$g_2(0)=\tilde{\lambda},\quad g_2'(0)=0,\quad \text{and}\quad g_2''(0)\neq 0.$$

SNB3 If $s\neq 0$ then $D_xf(g_1(s),g_2(s))$ is hyperbolic and if $\sigma(s)$ is the eigenvalue of $D_xf(g_1(s),g_2(s))$ that satisfies $\sigma(0)=0$, then $\sigma'(0)\neq 0$.

Before stating a general theorem we consider the simplest setting where the phase space is one dimensional.

Proposition 5.3.3. *Consider a C^2 function $f\colon\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ satisfying*

$$f(0,0)=0,\quad D_xf(0,0)=0,\quad D_x^2f(0,0)\neq 0,\quad \textit{and}\quad D_\lambda f(0,0)\neq 0.$$

Then, there is a saddle-node bifurcation at $(0,0)$.

Proof. Since $D_\lambda f(0,0)\neq 0$ the implicit function theorem implies the existence of a curve $g_2\colon(-\delta,\delta)\to\mathbb{R}$ such that $f(s,g_2(s))=0$. Observe that by defining $g_1(s)=s$, **SNB1** is satisfied.

Differentiating $f(s,g_2(s))=0$ with respect to s gives

$$D_xf(s,g_2(s))+D_\lambda f(s,g_2(s))g_2'(s)=0 \tag{5.8}$$

and in particular

$$D_xf(0,0)+D_\lambda f(0,0)g_2'(0)=0.$$

By assumption $D_x f(0,0) = 0$ and $D_\lambda f(0,0) \neq 0$, hence $g_2'(0) = 0$. Differentiating (5.8) gives

$$D_x^2 f(s, g_2(s)) + 2D_\lambda D_x f(s, g_2(s)) g_2'(s) + D_\lambda^2 f(s, g_2(s))(g_2'(s))^2 + D_\lambda f(s, g_2(s)) g_2''(s) = 0.$$

At the saddle-node we have

$$D_x^2 f(0,0) + D_\lambda f(0,0) g_2''(0) = 0$$

and hence

$$g_2''(0) = -\frac{D_x^2 f(0,0)}{D_\lambda f(0,0)} \neq 0.$$

This implies **SNB2**.

To check that **SNB3** is satisfied observe that

$$\frac{d}{ds} D_x f(s, g_2(s)) = D_x^2 f(s, g_2(s)) + D_\lambda D_x f(s, g_2(s)) g_2'(s)$$

and hence at $s = 0$,

$$\frac{d}{ds} D_x f(0,0) = D_x^2 f(0,0) \neq 0$$

which implies that for $|\delta|$ sufficiently small $D_x f(s, g_2(s)) \neq 0$. □

The following result, [1, Theorem 8.12], provides sufficient conditions in the setting of an n-dimensional phase space for the occurrence of a saddle-node bifurcation.

Theorem 5.3.4. *Assume $f\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is C^1, $(\tilde{x}, \tilde{\lambda}) \in \mathbb{R}^n \times \mathbb{R}$ is a saddle-node, and the kernel of $D_x f(\tilde{x}, \tilde{\lambda})$ is spanned by the non-zero vector $\tilde{v} \in \mathbb{R}^n$. If*

$$D_\lambda f(\tilde{x}, \tilde{\lambda}) \neq 0 \quad \textit{and} \quad D_x^2 f(\tilde{x}, \tilde{\lambda})(\tilde{v}, \tilde{v}) \neq 0$$

and both are not in the range of $D_x f(\tilde{x}, \tilde{\lambda})$, then there is a saddle-node bifurcation at $(\tilde{x}, \tilde{\lambda})$. Moreover, among all C^∞ one parameter families that have a saddle-node, those that undergo a saddle-node bifurcation form an open and dense subset.

The following lemma provides an elegant means of checking that if $u_1, u_2 \neq 0$, then they are not in the range of $D_x f(\tilde{x}, \tilde{\lambda})$.

Lemma 5.3.5. *Let $D : \mathbb{R}^n \to \mathbb{R}^n$ a linear transformation with* $\dim(\ker(D)) = 1$. *Then* $\ker D^T = \langle w \rangle$, *for some $w \in \mathbb{R}^n \setminus \{0\}$. Also, u is in the range of D if and only if $u \cdot w = 0$.*

Proof. By the rank-nullity theorem, $\operatorname{rank}(D) := \dim(\text{image}(D)) = n - \dim(\ker(D)) = n-1$. Since $\operatorname{rank}(D) = \operatorname{rank}(D^T)$ the rank-nullity theorem implies that $\dim(\ker(D^T)) = 1$. Hence, there is a non-zero vector $w \in \mathbb{R}^n$ such that $\ker D^T = \langle w \rangle$.

Now, if $u \in \mathbb{R}^n$ is in the range of D, there exists $y \in \mathbb{R}^n$ such that $u = Dy$, and then $u \cdot w = w^T u = w^T(Dy) = (w^T D)y = (D^T w)^T y = 0$. This implies that

$$\text{image}(D) \subset (\ker(D^T))^\perp = (\langle w \rangle)^\perp := \{u \in \mathbb{R}^n \ \mid \ u \cdot w = 0\}. \tag{5.9}$$

Conversely, assume $u \cdot w = 0$, that is $u \in (\langle w \rangle)^\perp = (\ker(D^T))^\perp$. Since $(\ker(D^T))^\perp$ is an $(n-1)$-dimensional subspace and $\dim(\text{image}(D)) = n-1$, we use (5.9) to get $\text{image}(D) = (\ker(D^T))^\perp$. Hence, $u \in (\ker(D^T))^\perp = \text{image}(D)$, that is u is in the range of D. □

Example 5.3.6. Consider the FitzHugh-Nagumo equation

$$\dot{x} = f(x, \gamma) = \begin{pmatrix} x_1(x_1 - a)(1 - x_1) - x_2 \\ \epsilon(x_1 - \gamma x_2) \end{pmatrix}, \tag{5.10}$$

with parameter values $(a, \epsilon) = (5, 1)$ and where γ is left as a free parameter. We use Theorem 5.3.4 to show the existence of a saddle-node bifurcation.

The first step is to identify a saddle-node, which we do numerically by using a predictor-corrector approach to compute two branches of hyperbolic equilibria, as portrayed in Figure 5.1. The computations suggest this occurs at $(\tilde{x}, \tilde{\gamma})$, with $\tilde{x} = (3, 12)^T$ and $\tilde{\gamma} = \frac{1}{4}$. Direct computation gives

$$f(\tilde{x}, \tilde{\gamma}) = \begin{pmatrix} 3(3-5)(1-3) - 12 \\ 1\left(3 - \frac{1}{4} \cdot 12\right) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad D := D_x f(\tilde{x}, \tilde{\gamma}) = \begin{pmatrix} 4 & -1 \\ 1 & -\frac{1}{4} \end{pmatrix},$$

and the eigenvalues of D are given by 0 and $\frac{15}{4}$. Therefore, $(\tilde{x}, \tilde{\gamma})$ is a saddle-node.

$\tilde{v} = (1, 4)^T$ is an eigenvector associated with eigenvalue 0 and hence $\ker D = \langle \tilde{v} \rangle$.

Since $\dim \ker D^T = \dim \ker D = 1$, we can solve for the eigenpair $(0, w)$ for D^T to obtain $w = (1, -4)^T \neq 0$. In particular, $\ker D^T = \langle w \rangle$.

Again, direct computation shows that

$$\begin{aligned} u_1 &:= D_\gamma f(\tilde{x}, \tilde{\gamma}) = \begin{pmatrix} 0 \\ -\epsilon \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -12 \end{pmatrix} \\ u_2 &:= D_x^2 f(\tilde{x}, \tilde{\gamma})(\tilde{v}, \tilde{v}) = \begin{pmatrix} 2\tilde{v}_1^2(a - 3\tilde{x}_1 + 1) \\ 0 \end{pmatrix} = \begin{pmatrix} 2(5 - 3 \cdot 3 + 1) \\ 0 \end{pmatrix} = \begin{pmatrix} -6 \\ 0 \end{pmatrix} \end{aligned}$$

and thus,

$$u_1 \cdot w = (0, -12)(1, -4)^T = 48 \neq 0 \quad \text{and} \quad u_2 \cdot w = (-6, 0)(1, -4)^T = -6 \neq 0.$$

Lemma 5.3.5 implies that both u_1, u_2 are not in the range of $D_x f(\tilde{x}, \tilde{\lambda})$.

Therefore, by Theorem 5.3.4 that there is a saddle-node bifurcation at $(\tilde{x}, \tilde{\gamma})$.

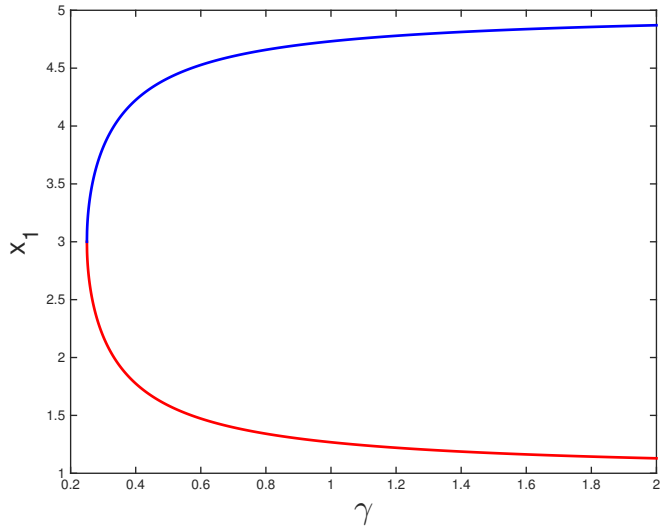

Figure 5.1: The below (red) branch was obtained by applying Newton to the initial point $(1,0)^T$ at $\gamma = 2$. After Newton converged to the point $\bar{x} = (1.129171306613029\ ,\ 0.564585653306515)^T$, we decreased the parameter γ up to $\frac{1}{4}$, where the branch seemed to disappear. The top (blue) branch was obtained by applying Newton to the initial point $(4,0)^T$ at $\gamma = 2$. After Newton converged to the point $\bar{x} = (4.870828693386970\ ,\ 2.435414346693485)^T$, we decrease the parameter γ up to $\frac{1}{4}$, where the branch seems to disappear.

Our success in directly applying Theorem 5.3.4 to demonstrate the existence of a saddle-node bifurcation for the FitzHugh-Nagumo equation made use of the fact that we had explicit values for the saddle-node $(\tilde{x}, \tilde{\gamma})$ and that the system is two dimensional and so we could perform the linear algebra computations by hand. Keeping with the spirit of this book we are interested in recasting Theorem 5.3.4 in such a way that we can rigorously verify the existence of a point at which a saddle-node bifurcation occurs for more general problems.

Algorithm 5.3.7. *Let $f\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ be smooth. Assume there exists $(\bar{x}, \bar{\lambda}) \in \mathbb{R}^n \times \mathbb{R}$ and $\mu \in \sigma(D_x f(\bar{x}, \bar{\lambda}))$ such that $f(\bar{x}, \bar{\lambda}) \approx 0$ and $\mu \approx 0$. Perform the following steps.*

(i) Using the techniques of Section 4.6 compute eigenpairs $\left\{(\tilde{\mu}_i, \tilde{v}^i) \mid i = 1, \ldots, N\right\}$ such that $\tilde{\mu}_i \in \overline{B_{r_i}(\bar{\mu}_i)}$ where

$$0 \in \overline{B_{r_1}(\bar{\mu}_1)} \quad \text{and} \quad \overline{B_{r_i}(\bar{\mu}_i)} \cap i\mathbb{R} = \emptyset, \quad i = 2, \ldots, n.$$

(ii) Verify that $\tilde{\mu}_1 = 0$.

(iii) Define $u^1 := D_\lambda f(\tilde{x}, \tilde{\lambda})$ *and* $u^2 := D_x^2 f(\tilde{x}, \tilde{\lambda})(\tilde{v}^1, \tilde{v}^1)$ *and verify that*

$$w^T u^i \neq 0, \quad i = 1, 2, \tag{5.11}$$

where $w \in \mathbb{R}^n \setminus \{0\}$ *such that* $w^T D_x f(\tilde{x}, \tilde{\lambda}) = 0$.

If steps (i)-(iii) are carried out successfully, then a saddle-node bifurcation occurs at the saddle-node $(\tilde{x}, \tilde{\lambda}) \in \overline{B_{\tilde{r}}\left(\bar{x}, \bar{\lambda}\right)}$ *where* $\tilde{r}$ *is obtain explicitly in the verification of step (ii).*

There are two parts to the justification of this algorithm. The first part is to show that if steps (i)-(iii) are carried out, then the hypotheses of Theorem 5.3.4 are satisfied and hence a saddle-node bifurcation occurs. We leave this to the reader.

The second part is to demonstrate that there are rigorous numerical procedures capable of carrying out steps (i)-(iii). As indicated in the statement of Algorithm 5.3.7 step (i) is to be carried out using the techniques of Section 4.6. We assume that this computation is successful.

Step (ii) requires us to rigorously verify that $\tilde{\mu}_1 = 0$, which in turn requires us to modify the problem being address in Section 4.6. By Step (i) there is at most one zero eigenvalue, thus it is sufficient to prove the existence of an eigenvector v that lies in the kernel of $D_x f(\tilde{x}, \tilde{\lambda})$. Observe that if it exists, then $v \in \mathbb{R}^n$, and thus to obtain isolation it is sufficient to require that $\|v\| = 1$. This leads to the following problem: prove the existence of $X = (x, \lambda, v) \in \mathbb{R}^{2n+1}$ satisfying

$$F(X) := \begin{pmatrix} f(x, \lambda) \\ \|v\|^2 - 1 \\ D_x f(x, \lambda) v \end{pmatrix} = 0. \tag{5.12}$$

This problem can be addressed using the radii polynomial approach described in Section 3.2. This solution to (5.12) identifies the existence of $(\tilde{x}, \tilde{\lambda}) \in \overline{B_{\tilde{r}}\left(\bar{x}, \bar{\lambda}\right)}$ for some $\tilde{r} > 0$. We now assume that step (ii) has been successfully completed.

Let $D := D_x f(\tilde{x}, \tilde{\lambda})$. Observe that step (iii) requires the existence of a vector $w \in \mathbb{R}^n \setminus \{0\}$ such that $w^T D = 0$. Let us now introduce how to use the radii polynomial approach to compute rigorously a non-zero vector w such that $\ker D^T = \langle w \rangle$. At this point, it is known rigorously that $\dim \ker(D) = 1$, which implies that $\operatorname{rank}(D) = \operatorname{rank}(D^T) = n - 1$. Therefore, when looking for a non-zero w satisfying $D^T w = 0$, we can get rid of one row of D^T without changing the solution space. Now the question is which equation to get rid of? As usual, we use a numerical approximation to answer that question. Assume that $\bar{v} \neq 0$ satisfies $D\bar{v} \approx 0$. Let k the component of $\bar{v}$ with the largest magnitude, that is

$$|\bar{v}_k| = \max_{i=1,\dots,n} \{|\bar{v}_i|\} \neq 0.$$

Denote by $C_1, \dots, C_n$ the columns of D and $R_1, \dots, R_n$ the corresponding rows of D^T that is $R_i = C_i^T$ for $i = 1, \dots, n$. Then since $D\bar{v} \approx 0$,

$$C_k \approx \frac{1}{\bar{v}_k} \sum_{\substack{i=1 \\ i \neq k}}^{n} \bar{v}_i C_i \Longrightarrow R_k = C_k^T \approx \frac{1}{\bar{v}_k} \sum_{\substack{i=1 \\ i \neq k}}^{n} \bar{v}_i C_i^T = \frac{1}{\bar{v}_k} \sum_{\substack{i=1 \\ i \neq k}}^{n} \bar{v}_i R_i.$$

Since the k-th row R_k of D^T is a linear combination of the other rows, we get rid of it, or equivalently we get rid of the k-th column C_k of D. Denote $M := \left(D_{\hat{k}}\right)^T$, with $D_{\hat{k}}$ the $n \times (n-1)$ matrix defined by D without its k-th column C_k. A non-zero unit vector w such that $\ker D^T = \langle w \rangle$ is an isolated solution of

$$g(w) := \begin{pmatrix} \|w\|^2 - 1 \\ Mw \end{pmatrix} = 0, \tag{5.13}$$

which we solve using the radii polynomial approach as introduced in Section 3.2.

The final step is to verify (5.11) which, in practice, is done using interval arithmetic.

Example 5.3.8. The Hodgkin-Huxley model for the action potential of a space-clamped squid axon is defined by the four dimensional vector field

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{pmatrix} = \begin{pmatrix} -20 - 120 x_2^3 x_3 \left(x_1 - 25.1 \ln(\frac{23}{1350}(550 - \lambda)) \right) - 36 x_4^4 \left(x_1 - 25.1 \ln(\frac{\lambda}{400}) \right) - 0.3(x_1 + 24.3) \\ \frac{9}{25}(1 - x_2) \dfrac{x_1 - \Delta(\lambda) + 35}{1 - \exp\left(-\frac{x_1 - \Delta(\lambda) + 35}{10} \right)} - \frac{72}{5} x_2 \exp\left(-\dfrac{x_1 - \Delta(\lambda) + 60}{18} \right) \\ \frac{63}{250}(1 - x_3) \exp\left(-\dfrac{x_1 - \Delta(\lambda) + 60}{20} \right) - \frac{18}{5} x_3 \dfrac{1}{\exp\left(-\frac{x_1 - \Delta(\lambda) + 30}{10} \right) + 1} \\ \frac{9}{250}(1 - x_4) \dfrac{x_1 - \Delta(\lambda) + 50}{1 - \exp\left(-\frac{x_1 - \Delta(\lambda) + 50}{10} \right)} - \frac{9}{20} x_4 \exp\left(-\dfrac{x_1 - \Delta(\lambda) + 60}{80} \right) \end{pmatrix} \tag{5.14}$$

where

$$\Delta(\lambda) := 9.32 \ln(\frac{11}{10} - \frac{\lambda}{500}).$$

The variable x_1 is the membrane potential, x_2 is the activation of a sodium current, x_3 is the activation of a potassium current, x_4 is the inactivation of the sodium current and the parameter λ is the external potassium concentration.

Using a standard pseudo-arclength continuation technique (see Exercise **??**), we obtain the bifurcation diagram of Figure 5.2. From this numerical simulation, we can conjecture the existence of two saddle-node bifurcations at $\lambda \approx 426.42$ and at $\lambda \approx 53.61$.

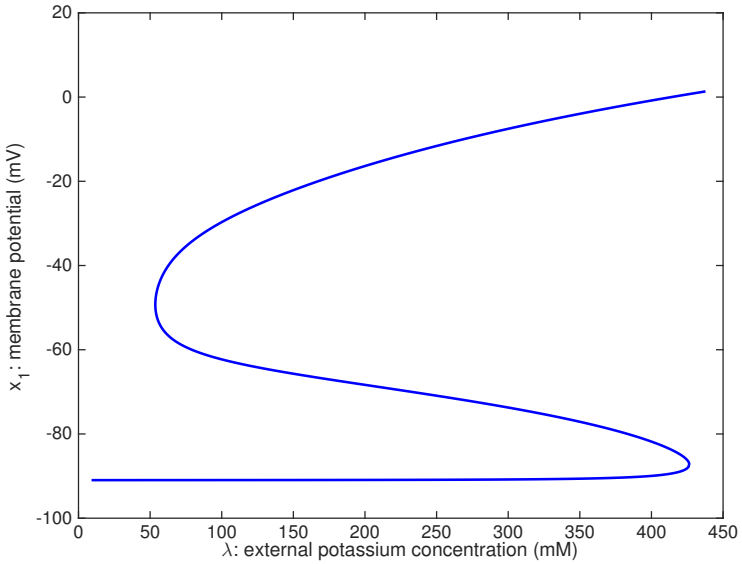

Figure 5.2: A branch of equilibria which undergoes two saddle-node bifurcations.

Denote by $f(x,\lambda)$ the right-hand side of (5.14). Let $X = (x,\lambda,v) \in \mathbb{R}^9$, and define $F : \mathbb{R}^9 \to \mathbb{R}^9$ as in (5.12). Applying Newton's method to problem (5.12), we compute two approximate solutions

$$\bar{X}_1 := \begin{pmatrix} -87.2515605439908 \\ 0.0089034888748 \\ 0.9231857631959 \\ 0.1363687823857 \\ 426.4159725555050 \\ 0.9998982043007 \\ 0.0011653907919 \\ -0.0105518817291 \\ 0.0095331365494 \end{pmatrix} \quad \text{and} \quad \bar{X}_2 := \begin{pmatrix} -49.270516282150226 \\ 0.170894916154782 \\ 0.241946685250340 \\ 0.487852905937695 \\ 53.607108413683697 \\ 0.999427022035232 \\ 0.016569867126847 \\ -0.025152384230317 \\ 0.015441006985167 \end{pmatrix} \tag{5.15}$$

corresponding to the two possible saddle-node points.

Theorem 5.3.9. *There is a saddle-node bifurcation at a point*

$$(\tilde{x}_1, \tilde{\lambda}_1) \in \overline{B_{r_1}\left((\bar{x}_1, \bar{\lambda}_1)\right)}, \qquad r_1 := 1.715 \times 10^{-12},$$

where $(\bar{x}_1, \bar{\lambda}_1) \in \mathbb{R}^5$ *is given by the first five components of* $\bar{X}_1$ *in* (5.15).

Theorem 5.3.10. *There is a saddle-node bifurcation at a point*

$$(\tilde{x}_2, \tilde{\lambda}_2) \in \overline{B_{r_2}\left((\bar{x}_2, \bar{\lambda}_2)\right)}, \qquad r_2 := 2.034 \times 10^{-12},$$

where $(\bar{x}_2, \bar{\lambda}_2) \in \mathbb{R}^5$ *is given by the first five components of* $\bar{X}_2$ *in* (5.15).

The proofs of Theorem 5.3.9 and Theorem 5.3.10 are both obtained by applying Algorithm 5.3.7.

Proof. (a) Let $\bar{X}$ one of the two points $\bar{X}_1$ or $\bar{X}_2$ given in (5.15). We already have that $F(\bar{X}) \approx 0$, with F given by (5.12). Note that

$$D_X F(\bar{X}) = \begin{pmatrix} D_x f(\bar{x}, \bar{\lambda}) & D_\lambda f(\bar{x}, \bar{\lambda}) & 0 \\ 0 & 0 & 2\bar{v}^T \\ D_x \left(D_x f(\bar{x}, \bar{\lambda}) \bar{v} \right) & D_x D_\lambda f(\bar{x}, \bar{\lambda}) v & D_x f(\bar{x}, \bar{\lambda}) \end{pmatrix}. \tag{5.16}$$

Using INTLAB in MATLAB, compute the exact inverse $A = D_X F(\bar{X})^{-1}$. Note that in practice, the so obtained A will have interval entries. Define $T : \mathbb{R}^9 \to \mathbb{R}^9$ by

$$T(X) = X - AF(X).$$

Using interval arithmetic, compute the upper bound Y such that $\left| AF(\bar{X}) \right| \ll Y$.

Now, thanks to the (perfect) choice of A

$$DT(\bar{X} + b)c = -A \left(DF(\bar{X} + b) - DF(\bar{X}) \right) c.$$

Defining $h : [0, 1] \to \mathbb{R}^9$ by $h(s) = D_X F(\bar{X} + sb)c$, $h(1) - h(0) = \left(D_X F(\bar{X} + b) - D_X F(\bar{X}) \right) c$. For each $k \in \{1, \ldots, 9\}$, there exists $s_k \in [0, 1]$ such that

$$\left(D_X F_k(\bar{X} + b) - D_X F_k(\bar{X}) \right) c = h_k(1) - h_k(0) = h_k'(s_k) = D_X^2 F_k(\bar{X} + s_k b)(b, c).$$

Now, let $\tilde{b}, \tilde{c} \in \overline{B_1(0)}$ such that $b, c \in \overline{B_r(0)}$ are given by $b = \tilde{b}r$ and $c = \tilde{c}r$. In this case,

$$\left(D_X F_k(\bar{X} + b) - D_X F_k(\bar{X}) \right) c = D_X^2 F_k(\bar{X} + s_k b)(\tilde{b}, \tilde{c}) r^2.$$

Set $r^* = 10^{-4}$ an a-priori upper bound for the left point of the existence interval of the radii polynomials. We will have to show a-posteriori that $r \leq r^*$. Denote by $\mathbf{b}^* = [-r^*, r^*]^9$ a vector in $\mathbb{R}^9$ whose entries are given by the interval $[-r^*, r^*]$. Denote by $\mathbf{X}^* = \bar{X} + \mathbf{b}^*$ a vector in $\mathbb{R}^9$ with its k-th entry given by the interval $[\bar{X}_k - r^*, \bar{X}_k + r^*]$. Denote by $\boldsymbol{\delta} = [-1, 1]^9$ a vector in $\mathbb{R}^9$ whose entries are given by the interval $[-1, 1]$. Then, for each $b, c \in \overline{B_r(0)}$, it is left to the reader to verify that

$$\left| A \left(DF(\bar{X} + b) - DF(\bar{X}) \right) c \right| \in \left| AD_X^2 F(\mathbf{X}^*)(\boldsymbol{\delta}, \boldsymbol{\delta}) \right|.$$

Using interval arithmetic, compute $Z_1 \in \mathbb{R}^9$ such that

$$\left| AD_X^2 F(\mathbf{X}^*)(\boldsymbol{\delta}, \boldsymbol{\delta}) \right| \ll Z^{(2)}.$$

Using the previous bounds, define the radii polynomials $p_k(r) = Z_k^{(2)} r^2 - r + Y_k$. For each of the point $\bar{X}_1$ and $\bar{X}_2$ given in (5.15), we computed the radii polynomials and obtained the existence intervals

$$I_1 = \left[1.715 \times 10^{-12}\ ,\ 8.052 \times 10^{-6}\right] \quad \text{and} \quad I_2 = \left[2.034 \times 10^{-12}\ ,\ 1.974 \times 10^{-5}\right],$$

respectively. Since $8.052 \times 10^{-6}, 1.974 \times 10^{-5} < r^* = 10^{-4}$, then the existence intervals are valid. Let $r_1 := 1.715 \times 10^{-12}$ and $r_2 := 2.034 \times 10^{-12}$. Recall (5.15), then by Corollary 3.2.3, there exists a unique $\tilde{X}_1 = (\tilde{x}_1, \tilde{\lambda}_1, \tilde{v}_1) \in \overline{B_{r_1}\left(\bar{X}_1\right)}$ such that $F(\tilde{X}_1) = 0$ and there exists a unique $\tilde{X}_2 = (\tilde{x}_2, \tilde{\lambda}_2, \tilde{v}_2) \in \overline{B_{r_2}\left(\bar{X}_2\right)}$ such that $F(\tilde{X}_2) = 0$. Hence, for $j = 1, 2$, $f(\tilde{x}_j, \tilde{\lambda}_j) = 0$ and the kernel of $D_x f(\tilde{x}_j, \tilde{\lambda}_j)$ must be one dimensional, as otherwise we would not have that $\tilde{X}_j$ isolated solution in $\mathbb{R}^9$.

(b) Choose $j \in \{1, 2\}$ and let $I := I_j$ the existence interval associated to $\bar{X} := \bar{X}_j$. Let r the smallest radius of the existence interval I. Define $\mathbf{B} = \overline{B_r\left((\bar{x}, \bar{\lambda})\right)} \subset \mathbb{R}^5$, that is

$$\mathbf{B} = \prod_{k=1}^{4} \left[\bar{x}_k - r, \bar{x}_k + r\right] \times \left[\bar{\lambda} - r, \bar{\lambda} + r\right].$$

Let $D := D_x f(\tilde{x}, \tilde{\lambda})$ and $\mathbf{D} := D_x f(\mathbf{B})$ a 4×4 interval matrix computed with interval arithmetic. Note that $D \subset \mathbf{D}$. Using the radii polynomial approach as introduced in Section 4.6, we now show that $\sigma(\mathbf{D}) \subset \bigcup_{j=1}^{n} B_j$, for some small balls $B_j \in \mathbb{C}$. The only modification from the theory of Section 4.6 is that we now have a matrix whose entries are intervals. Hence, the bounds Y in (4.26) and the bonds Z_0, Z_1 in (4.29) have to bound all possible error coming from $\mathbf{D}$. Interval arithmetic can be used to do this. Using the above procedure and INTLAB, we proved that the eigenvalues of $D_x f(\tilde{x}_1, \tilde{\lambda}_1)$ are enclosed in $\cup_{j=1}^{4} B_j$, where

$$\begin{aligned}
B_1 &= \left\{z \in \mathbb{C} \ :\ |z + 32.02633660454969| \leq 3.394274197807681 \times 10^{-11}\right\} \\
B_2 &= \left\{z \in \mathbb{C} \ :\ |z - 9.305978062941147 \times 10^{-16}| \leq 1.069542699059249 \times 10^{-11}\right\} \\
B_3 &= \left\{z \in \mathbb{C} \ :\ |z + 0.9317141708275124| \leq 9.975910428070430 \times 10^{-12}\right\} \\
B_4 &= \left\{z \in \mathbb{C} \ :\ |z + 0.5558951614569074| \leq 1.774408327300926 \times 10^{-12}\right\}.
\end{aligned}$$

Therefore, we obtain that $0 \in B_k$ for the unique $k = 2 \in \{1, 2, 3, 4\}$, and that $B_j \cap i\mathbb{R} = \emptyset$, for all $j \in \{1, 3, 4\}$. This shows that $(\tilde{x}_1, \tilde{\lambda}_1)$ is a saddle-node. We repeated the same procedure to show that $(\tilde{x}_2, \tilde{\lambda}_2)$ is also a saddle-node.

(c) Let

$$\bar{w}_1 := \begin{pmatrix} 0.019179161523012 \\ 0.001671623761223 \\ 0.000309560202397 \\ 0.999814617621566 \end{pmatrix} \quad \text{and} \quad \bar{w}_2 := \begin{pmatrix} -0.009242057198078 \\ -0.253654297666144 \\ -0.937890513625427 \\ 0.236506799280007 \end{pmatrix}, \tag{5.17}$$

numerical approximations satisfying $D_x f(\bar{x}_1, \bar{\lambda_1})^T \bar{w}_1 \approx 0$ and $D_x f(\bar{x}_2, \bar{\lambda_2})^T \bar{w}_2 \approx 0$.

Choose $j \in \{1, 2\}$, let $D = D_x f(\tilde{x}_j, \tilde{\lambda_j})$ and $\bar{w} = \bar{w}_j$. Based on D, construct the 3×4 matrix M as above in order to define the problem $g(w) = 0$ as in(5.13). Using the radii polynomial approach of Section 3.2 applied on (5.13), we showed the existence of (i) $\tilde{w}_1 \in \overline{B_{7.246\times10^{-13}}(\bar{w}_1)}$ such that $D_x f(\tilde{x}_1, \tilde{\lambda_1})\tilde{w}_1 = 0$ and (ii) $\tilde{w}_2 \in \overline{B_{1.595\times10^{-11}}(\bar{w}_2)}$ such that $D_x f(\tilde{x}_2, \tilde{\lambda_2})\tilde{w}_2 = 0$.

(d) For $j = 1, 2$, define

$$u_1^{(j)} = D_\gamma f(\tilde{x}_j, \tilde{\lambda}_j) \quad \text{and} \quad u_2^{(j)} = D_x^2 f(\tilde{x}_j, \tilde{\lambda}_j)(\tilde{v}_{,j}\, \tilde{v}_j).$$

With interval arithmetic, we showed that for both $j = 1, 2$, $u_1^{(j)} \cdot \tilde{w}_j \neq 0$ and $u_2^{(j)} \cdot \tilde{w}_j \neq 0$. This concludes the proofs of Theorem 5.3.9 and Theorem 5.3.10. □

Bibliography

[1] Carmen Chicone. *Ordinary differential equations with applications*, volume 34 of *Texts in Applied Mathematics*. Springer, New York, second edition, 2006.

[2] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[3] Gerald B. Folland. *Introduction to partial differential equations*. Princeton University Press, Princeton, NJ, second edition, 1995.

[4] Jack K. Hale. *Ordinary differential equations*. Robert E. Krieger Publishing Co. Inc., Huntington, N.Y., second edition, 1980.

[5] James M. Ortega. The Newton-Kantorovich theorem. *Amer. Math. Monthly*, 75:658–660, 1968.

[6] Clark Robinson. *Dynamical systems*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, second edition, 1999. Stability, symbolic dynamics, and chaos.