Chapter 5

Computer-Assisted Proofs for Nonlinear Equations

The focus of this section is on developing a particular technique that we call a Newton-Kantorovich Theorem (NKT) for proving the existence of zeros and obtaining bounds on their location. The philosophy of the strategy is quite simple. Transform the problem of finding a zero to the problem of finding a fixed point. Newton's method, discussed in Section 5.2, provides a conceptual approach for constructing a function T with the property that if $T(\tilde{x}) = \tilde{x}$, then $f(\tilde{x}) = 0$. Of course, this is only of use if there is a general method by which one can prove that T has a fixed point. For this we make use of a classical result, the Contraction Mapping Theorem that is presented in Section 5.1. This is an extremely powerful theorem in that it guarantees the existence of a unique zero. As the name suggests the key requirement is that T be a contraction mapping.

Therefore, the challenge in implementing this strategy is identifying readily checkable conditions under which T is a contraction mapping. This is the role of the above mentioned NKT. Based on the fact that given today's technology people study dynamical systems numerically we assume that $\bar{x}$, an approximate zero to f, has been proposed. As detailed in Theorem 5.2.5, the required bounds are associated with $\bar{x}$, f and T, and if some inequalities hold then an explicit domain on which T is a contraction mapping is provided. As a consequence one obtains the existence of a unique zero and bounds on its location.

5.1 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 $T(\tilde{x}) = \tilde{x}$.

In general finding fixed points of a function is a nontrivial task. An exception is when iterations of the function lead to a fixed point. Given $x_0 \in X$, inductively define $x_{k+1} \stackrel{\text{def}}{=} T(x_k)$ for $k \in \mathbb{N}$. A fixed point is *globally attracting* if $\lim_{k\to\infty} x_k = \tilde{x}$ for all

$x_0 \in X$. The contraction mapping theorem, discussed below, provides a simple criterion that guarantees that a function has a globally attracting fixed point.

To effectively identify globally attracting fixed points it is essential that limits exist under reasonable bounds. This restricts the type of topological space of interest.

Definition 5.1.1. A metric space $(X, \mathbf{d})$ is said to be *complete* if every Cauchy sequence converges in X, i.e., if given any $\epsilon > 0$ there exists $N(\epsilon)$ such that $k_1, k_2 > N(\epsilon)$ implies $\mathbf{d}(x_{k_1}, x_{k_2}) < \epsilon$, then there is some $y \in X$ with $\lim_{k\to\infty} x_k = y$.

Definition 5.1.2. 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 5.1.3 (Contraction Mapping Theorem). *Let $(X, \mathbf{d})$ be a complete metric space. If $T\colon X \to X$ is a contraction with contraction constant κ, then there exists a unique fixed point $\tilde{x} \in X$ of T. Furthermore, $\tilde{x}$ is globally attracting, and for any $x \in X$,*

$$\mathbf{d}(T^k(x), \tilde{x}) \leq \frac{\kappa^k}{1-\kappa} \mathbf{d}(T(x), x). \tag{5.1}$$

Proof. Choose $x_0 \in X$ and recursively define $x_{k+1} \stackrel{\text{def}}{=} T(x_k)$. By the assumption that T is a contraction,

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

Thus, by induction

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

Applying the triangle inequality, we have for $k < m$

$$\begin{aligned} \mathbf{d}(x_k, x_m) &\leq \sum_{j=k}^{m-1} \mathbf{d}(x_{j+1}, x_j) \\ &\leq \sum_{j=k}^{m-1} \kappa^j \mathbf{d}(x_1, x_0) \\ &\leq \kappa^k \left(\sum_{k=0}^{\infty} \kappa^k \right) \mathbf{d}(x_1, x_0) \\ &\leq \kappa^k \frac{1}{1-\kappa} \mathbf{d}(x_1, x_0). \end{aligned} \tag{5.2}$$

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

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

By continuity of T,

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

This establishes the existence of a fixed point.

We prove by contradiction that the fixed point is unique. Assume $\tilde{x}$ is as above and $\tilde{y}$ is another fixed point of T, that is $T(\tilde{y}) = \tilde{y}$ and $\mathbf{d}(\tilde{y}, \tilde{x}) > 0$. Then

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

Dividing by $\mathbf{d}(\tilde{y}, \tilde{x})$ gives $\kappa \geq 1$, which is the desired contradiction.

Turning to (5.1), note that since $x_0 \in X$ was arbitrary, the estimates of (5.2) are independent of the base point x_0. Applying the triangle inequality we have

$$\begin{aligned}\mathbf{d}(T^k(x), \tilde{x}) &\leq \mathbf{d}(T^k(x), T^m(x)) + \mathbf{d}(T^m(x), \tilde{x}) \\ &\leq \frac{\kappa^k}{1-\kappa}\mathbf{d}(T(x), x) + \mathbf{d}(T^m(x), \tilde{x}).\end{aligned}$$

for any $x \in X$, and all $m > k$. Taking the limit as $m \to \infty$ yields the bound claimed in (5.1). Finally, (5.1) and the fact that $\kappa \in [0, 1)$ proves that $\tilde{x}$ is a globally attracting fixed point. □

Remark 5.1.4. Observe that given a contraction mapping $T\colon X \to X$, the rate at which points in X converge to the globally attracting fixed point $\tilde{x}$ is determined by κ. In particular, the smaller κ is, the faster iterates under T converge to $\tilde{x}$.

As is clear from the hypothesis of Theorem 5.1.3 the contraction mapping theorem is applicable in the setting of arbitrary complete metric spaces. For purposes of this book we always work in finite dimensional normed vector spaces. Recall that a norm $\|\cdot\|$ on X defines a metric $\mathbf{d}$ on X by

$$\mathbf{d}(x, y) \stackrel{\text{def}}{=} \|x - y\|.$$

The *open ball of radius r centered at x_0* is given by

$$B_r(x_0) \stackrel{\text{def}}{=} \{x \in X \mid \mathbf{d}(x_0, x) < r\} = \{x \in X \mid \|x_0 - x\| < r\},$$

and the *closed ball of radius r centered at x_0* is given by

$$\overline{B_r(x_0)} \stackrel{\text{def}}{=} \{x \in X \mid \mathbf{d}(x_0, x) \leq r\} = \{x \in X \mid \|x_0 - x\| \leq r\}.$$

Standard norms on $\mathbb{R}^n$ or $\mathbb{C}^n$ are the 1-*norm*

$$\|x\|_1 \stackrel{\text{def}}{=} \sum_{k=1}^{n} |x_k|,$$

the 2-*norm* or *Euclidean norm*

$$\|x\|_2 \stackrel{\text{def}}{=} \left(\sum_{k=1}^{n} |x_k|^2 \right)^{\frac{1}{2}},$$

and the ∞-*norm* or *sup norm*

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

A fundamental result is that these norms induce the same topology on $\mathbb{R}^n$ ($\mathbb{C}^n$) and thus we are free to choose the most convenient norm for our purposes. With this in mind when working with a finite dimensional vector space we denote the norm by $\|\cdot\|$ unless we make use of specific properties of a particular norm.

5.2 Newton's method and a Newton-Kantorovich theorem

As indicated earlier, the goal of this section is to develop techniques by which we can guarantee the existence of zeros of a function. The contraction mapping theorem provides a means of proving the existence of a unique fixed points in a purely topological setting. The following proposition – the proof is left to the reader – relates fixed points with zeros. For the remainder of this book we use the notation $\tilde{x}$ to denote a zero of a function, that is, $f(\tilde{x}) = 0$.

Proposition 5.2.1. *Let X be a finite dimensional vector space and let $U \subset X$. Consider $f\colon U \to X$ and assume that $A\colon X \to X$ is an injective linear map. Let $T\colon U \to X$ be defined by*

$$T(x) \stackrel{\text{def}}{=} x - Af(x). \tag{5.3}$$

Then, $T(\tilde{x}) = \tilde{x}$ if and only if $f(\tilde{x}) = 0$.

Via Proposition 5.2.1 to find a zero of f it is sufficient to find an injective linear map A that makes T a contraction. This leads us to Newton's method and the beginning of the analytic approach of this book.

We make use of the following notation. Given subsets of normed vector spaces $U \subset X$ and $V \subset Y$, the statement $f \in C^r(U,V)$ indicates that $f\colon U \to V$, f is r-times differentiable, and the r-th derivative of f is continuous.

Consider $f \in C^1(\mathbb{R},\mathbb{R})$. Recall that in Newton's method the fixed point map T is given by

$$T(x) \stackrel{\text{def}}{=} x - \frac{1}{f'(x)} f(x)$$

and (as is described in Section 5.1) T is used iteratively to find a fixed point $\tilde{x}$, and hence a zero of f. 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 $\bar{x} \in \mathbb{R}$ leads to convergence of Newton's method.

Newton's method and its variants are widely used to find zeros of functions. The contraction mapping theorem provides insight into why this is such a powerful formalism. For simplicity assume that $f \in C^2$ and, more importantly, that $f'(\tilde{x}) \neq 0$. In this case we can compute the derivative

$$T'(\tilde{x}) = 1 - \frac{(f'(\tilde{x}))^2 - f(\tilde{x})f''(\tilde{x})}{(f'(\tilde{x}))^2} = 0. \tag{5.4}$$

This implies that T' is small near the fixed point $\tilde{x}$ of T, hence, as we shall argue, T has a small contraction constant in a neighborhood of $\tilde{x}$.

Indeed, consider an open set $B_\delta(\tilde{x})$ where $\delta > 0$ is sufficiently small enough that $\sup_{\xi \in B_\delta(\tilde{x})} |T'(\xi)| < 1$. The mean value theorem implies that for any $x, y \in B_\delta(\tilde{x})$, there exists $z = z(x, y) \in B_\delta(\tilde{x})$ such that

$$T(x) - T(y) = T'(z)(x - y).$$

Let $\kappa \stackrel{\text{def}}{=} \sup_{\xi \in B_\delta(\tilde{x})} |T'(\xi)| < 1$. Then, given any $x, y \in B_\delta(\tilde{x})$

$$|T(x) - T(y)| = |T'(z)||x - y| \leq \kappa |x - y|,$$

and therefore, T has a contraction constant κ on $B_\delta(\tilde{x})$. Furthermore, κ goes to 0 as δ goes to 0.

The intended take away message from this discussion is that if the derivative is non-zero, then in a sufficiently small neighborhood of a zero of f the associated Newton operator is an extremely strong contraction. In particular, returning to the question concerning the choice of an invertible linear map A, the straightforward but naive interpretation of this example suggests setting $A = A(x) = (f'(x))^{-1}$.

We wish to generalize this discussion to higher dimensions.

Definition 5.2.2. Let $f \in C^1(U, \mathbb{R}^n)$ where $U \subset \mathbb{R}^n$ is an open set. A point $\tilde{x} \in U$ is a *nondegenerate zero* of f if $f(\tilde{x}) = 0$ and $Df(\tilde{x})$ is invertible.

If $g \in C^2(U, \mathbb{R}^n)$ where $U \subset \mathbb{R}^n$ is an open set, then the associated Newton operator is given by

$$T(x) \stackrel{\text{def}}{=} x - (Df(x))^{-1} f(x). \tag{5.5}$$

An argument similar to that just presented demonstrates that if $\tilde{x}$ is a nondegenerate zero, 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, as one shall

see later, it is often enough to choose A as an approximation of $Df(\bar{x}))^{-1}$. Now, the n-dimensional analogue of (5.4) is

$$DT(\tilde{x}) = I - D(Df(x))^{-1}\Big|_{x=\tilde{x}} f(\tilde{x}) - (Df(\tilde{x}))^{-1}Df(\tilde{x}) = 0$$

from which we concluded that the linear map $DT(\hat{x})$ is small if $\|\hat{x} - \tilde{x}\|$ is small. The following norm allows us to make the meaning of 'small' precise. As the careful reader may have noted, we have implicitly made use of a metric on linear maps.

Definition 5.2.3. Let $(X, \|\cdot\|)$ be a normed linear space and let $A\colon X \to X$ be a linear map. The *operator norm* on A is given by

$$\|A\| \stackrel{\text{def}}{=} \sup_{x\in X\setminus\{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|.$$

Remark 5.2.4. It is important to note that the choice of norm $\|\cdot\|$ has an impact on the associated operator norm. Let $M_n(\mathbb{R})$ and $M_n(\mathbb{C})$ denote the set of $n \times n$ matrices over $\mathbb{R}$ and $\mathbb{C}$, respectively. A matrix $A = [a_{i,j}] \in M_n(\mathbb{C})$ (or equivalently $A \in M_n(\mathbb{R})$) then the associated matrix norms are

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{i,j}| \tag{5.6}$$

$$\|A\|_2 = \sqrt{r_\sigma(A^*A)} \tag{5.7}$$

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{i,j}| \tag{5.8}$$

where A^* is the conjugate transpose of A and $r_\sigma(A^*A)$ denotes the maximum of the magnitudes of the eigenvalues of A^*A. Observe that if the matrix A is known explicitly, then computing $\|A\|_1$ or $\|A\|_\infty$ is much easier than determining $\|A\|_2$.

We are now ready to introduce a variant of the well-known Newton-Kantorovich Theorem, which we use throughout this book to obtain our computer-assisted proofs.

Consider a finite dimensional Banach space X (in our context $X = \mathbb{R}^n$ or $X = \mathbb{C}^n$, for some $n \in \mathbb{N}$). Choose a norm $\|\cdot\|$ on X. Given a point $y \in X$ and a radius $r > 0$, denote by $B_r(y) = \{x \in X : \|y - x\| < r\}$ the open ball of radius r centered at y. Similarly, denote by $\overline{B_r(y)}$ the closed ball.

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 discussion of the previous section suggests that there exists a neighborhood U^+ of 1 such that $T\colon U^+ \to \mathbb{R}$ defined by $T(x) \stackrel{\text{def}}{=} x - \frac{1}{2}(x^2 - 1)$ is a contraction mapping. The theorem below provides a mechanism for rigorously identifying a domain on which T is a contraction mapping.

Theorem 5.2.5 (A Newton-Kantorovich theorem). *Let $(X, \|\cdot\|)$ be a finite dimensional Banach space (typically $\mathbb{R}^n$ or $\mathbb{C}^n$). Let $U \subset X$ be an open set. Let $f : U \to X$ be differentiable, fix a point $\bar{x} \in U$ and let $A : X \to X$ be a linear map. Fix $r_* > 0$. Suppose that the bounds $Y, Z = Z(r_*) > 0$ satisfy*

$$\|Af(\bar{x})\| \leq Y \quad \textit{and} \quad \sup_{z \in \overline{B_{r_*}(\bar{x})}} \|I - ADf(z)\| \leq Z. \tag{5.9}$$

If

$$Z < 1 \quad \textit{and} \quad \frac{Y}{1-Z} < r_*, \tag{5.10}$$

then for each $r \in \left(\frac{Y}{1-Z}, r_*\right]$, there is a unique $\tilde{x} \in B_r(\bar{x})$ such that $f(\tilde{x}) = 0$.

In practice, the point $\bar{x}$ is chosen to be an approximate zero of f, and A is chosen to be an approximate inverse of the Jacobian matrix $Df(\bar{x})$ (that is $\|I - ADf(\bar{x})\| \ll 1$).

Remark 5.2.6. Observe that if (5.10) holds, then the *existence interval*

$$\mathsf{EI}(p) \stackrel{\text{def}}{=} \left(\frac{Y}{1-Z}, r_*\right] \neq \emptyset$$

and $\tilde{x}$ is the unique zero of f in $\overline{B_r(\bar{x})}$ for all $r \in \mathsf{EI}(p)$, r_0 provides tight bounds on the location of $\tilde{x}$, while r_* provides information about the domain of isolation of $\tilde{x}$. In particular, if the existence interval is nonempty, then one can present an explicit domain $U \subset \mathbb{R}^n$ in which there exists a unique zero of f.

Proof. Define the Newton-like operator $T : U \to X$ by

$$T(x) = x - Af(x),$$

and note that $DT(x) = I - ADf(x)$. Since (5.10) holds, pick any $r \in \left(\frac{Y}{1-Z}, r_*\right]$. Note that $\frac{Y}{1-Z} < r$ implies that

$$Zr + Y < r. \tag{5.11}$$

The idea of the proof is to show that $T : \overline{B_r(\bar{x})} \to B_r(\bar{x})$ is a contraction. From (5.11), we have that $Z \leq Z + \frac{Y}{r} < 1$. For $x, y \in \overline{B_r(\bar{x})}$ we use the Mean Value Inequality to get that

$$\begin{aligned}
\|T(x) - T(y)\| &\leq \sup_{z \in \overline{B_r(\bar{x})}} \|DT(z)\| \|x - y\| \\
&= \sup_{z \in \overline{B_r(\bar{x})}} \|I - ADf(z)\| \|x - y\| \\
&\leq \sup_{z \in \overline{B_{r_*}(\bar{x})}} \|I - ADf(z)\| \|x - y\| \\
&\leq Z\|x - y\|.
\end{aligned}$$

Since $Z < 1$, T is a contraction on $\overline{B_r(\bar{x})}$. To see that T maps the closed ball into itself (in fact in the open ball) choose $x \in \overline{B_r(\bar{x})}$, and observe that

$$\begin{aligned}\|T(x) - \bar{x}\| &\leq \|T(x) - T(\bar{x})\| + \|T(\bar{x}) - \bar{x}\| \\ &\leq Z\|x - \bar{x}\| + \|Af(\bar{x})\| \\ &\leq Zr + Y < r,\end{aligned}$$

which shows that $T(x) \in B_r(\bar{x})$ for all $x \in \overline{B_r(\bar{x})}$. It follows from the contraction mapping theorem that there exists a unique $\tilde{x} \in \overline{B_r(\bar{x})}$ such that $T(\tilde{x}) = \tilde{x} \in B_r(\bar{x})$. Since $Z < 1$, we get

$$\|I - ADf(\bar{x})\| \leq \sup_{z \in \overline{B_{r_*}(\bar{x})}} \|I - ADf(z)\| \leq Z < 1,$$

and hence $ADf(\bar{x})$ is invertible. From this we get that A is invertible. By invertibility of A and by definition of T, the fixed points of T are in one-to-one correspondence with the zeros of f. We conclude that there is a unique $\tilde{x} \in B_r(\bar{x})$ such that $f(\tilde{x}) = 0$. □

5.3 Examples and Numerics

In this section, we apply the radii polynomial approach, as introduced in Theorem 5.2.5 to prove the existence of zeros of different maps. Throughout this section, we fix the norm to be the sup norm $\|\cdot\| = \|\cdot\|_\infty$.

Example 5.3.1. Consider the use of Theorem 5.2.5 to the problem of verifying zeros of $f : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$f(x) = \begin{pmatrix} 4x_1^2 + x_2 - \alpha \\ x_1 + x_2^2 - 1 \end{pmatrix} \tag{5.12}$$

where $x = (x_1, x_2)$ and $\alpha \in \mathbb{R}$ is a parameter. This corresponds to finding the intersecting points of the parabolas $4x_1^2 + x_2 = \alpha$ and $x_1 + x_2^2 = 1$ (see Figure 5.1 for the case $\alpha = 3$).

To construct a radii polynomial we need to determine bounds Y and Z satisfying the inequalities in (5.9). Because (5.12) is relatively simple, we can explicitly compute

$$Df(x) = \begin{pmatrix} 8x_1 & 1 \\ 1 & 2x_2 \end{pmatrix} \quad \text{and} \quad Df(x)^{-1} = \frac{1}{16x_1x_2 - 1}\begin{pmatrix} 2x_2 & -1 \\ -1 & 8x_1 \end{pmatrix},$$

assuming of course that $16x_1x_2 - 1 \neq 0$. The optimal choice suggests setting $A \overset{\text{def}}{=} Df(\bar{x})^{-1}$. With this choice of A we can set $Y \overset{\text{def}}{=} \|Df(\bar{x})^{-1}f(\bar{x})\|$. To determine Z, we fix $r_* = 0.01$ and compute Z (with interval arithmetic) such that

$$\sup_{z \in \overline{B_{0.01}(\bar{x})}} \|I - ADf(z)\| \leq Z.$$

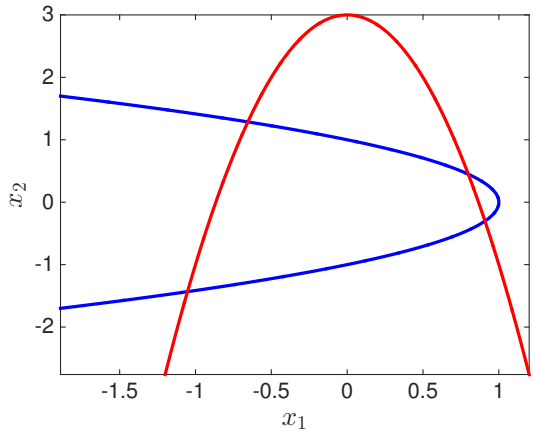

Figure 5.1: Intersection of the parabolas $4x_1^2 + x_2 = 3$ (in red) and $x_1 + x_2^2 = 1$ (in blue).

Set $p(r) = (Z-1)r + Y$, where Y and Z depend on the approximate solution $\bar{x}$, the matrix A and the map f.

Our expectation is that if we are given any reasonable approximation of a zero of (5.12), then there will exist $r_0 > 0$ such that $p(r_0) < 0$.

To test our expectation, we set $\alpha = 3$ and applied a numerical scheme based on Newton's method to find $\bar{x} \in \mathbb{R}^2$ such that $\|f(\bar{x})\| < \texttt{tol}$ where we fixed the tolerance $\texttt{tol} = 5\times 10^{-16}$. This resulted in four approximate solutions

$$\begin{aligned} \bar{x}^{(1)} &= \begin{pmatrix} -0.6545436118927946 \\ 1.286290640521338 \end{pmatrix}, \quad \bar{x}^{(2)} = \begin{pmatrix} 0.7986333753610425 \\ 0.4487389270377125 \end{pmatrix}, \\ \bar{x}^{(3)} &= \begin{pmatrix} 0.9086121587039679 \\ -0.3023042197787387 \end{pmatrix}, \quad \bar{x}^{(4)} = \begin{pmatrix} -1.052701922172216 \\ -1.432725347780312 \end{pmatrix}. \end{aligned} \tag{5.13}$$

For $i = 1, 2, 3, 4$ one can check that the following intervals

$$\begin{aligned} I^{(1)} &= [1.7 \times 10^{-16},\ 0.01] \subset \mathsf{EI}(p^{(1)}) \\ I^{(2)} &= [2.6 \times 10^{-16},\ 0.01] \subset \mathsf{EI}(p^{(2)}) \\ I^{(3)} &= [3.2 \times 10^{-16},\ 0.01] \subset \mathsf{EI}(p^{(3)}) \\ I^{(4)} &= [3.7 \times 10^{-16},\ 0.01] \subset \mathsf{EI}(p^{(4)}) \end{aligned} \tag{5.14}$$

are subset of the existence interval. Moveover, for each of the steady state, we maximized the r_* such that the condition $Z < 1$ would still hold. Figure 5.2 shows the largest balls (of radius r_*) for which the proof succeeded.

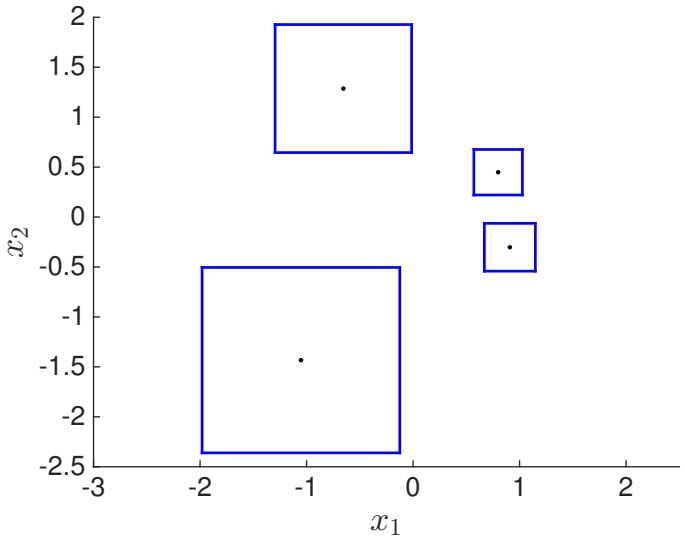

Figure 5.2: Largest existence and uniqueness enclosures for each equilibrium of (5.12) for $\alpha = 3$. For each $i = 1, 2, 3, 4$, the radius around $\bar{x}^{(i)}$ is the largest value of $I^{(i)}$. The quantities $\bar{x}^{(i)}$ and $I^{(i)}$ are found in (5.13) and (5.14) respectively. The smallest enclosure is too small to represent graphically, which implies that the dots representing $\bar{x}^{(i)}$ also represent the true equilibria $\tilde{x}^{(i)}$.

5.3.1 Periodic points in discrete dynamical systems

In the next couple of examples, we prove existence of periodic points in some discrete dynamical systems. Given a map $g : U \to \mathbb{R}^n$ (with $U \subset \mathbb{R}^n$ open), looking for periodic point of period k of the map g boils down to find $\tilde{x} \in (\mathbb{R}^n)^k = \mathbb{R}^{kn}$ such that $f(\tilde{x}) = 0$, where $f : \mathbb{R}^{kn} \to \mathbb{R}^{kn}$ is given by

$$f(x) \stackrel{\text{def}}{=} \begin{pmatrix} g(x_1) - x_2 \\ g(x_2) - x_3 \\ \vdots \\ g(x_{k-1}) - x_k \\ g(x_k) - x_1 \end{pmatrix}, \qquad x \stackrel{\text{def}}{=} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} \in (\mathbb{R}^n)^k. \tag{5.15}$$

The Jacobian matrix of f is given by

$$Df(x) = \begin{pmatrix} Dg(x_1) & -I & 0 & \cdots & 0 \\ 0 & Dg(x_2) & -I & \ddots & 0 \\ 0 & 0 & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & Dg(x_{k-1}) & -I \\ -I & 0 & \cdots & 0 & Dg(x_k) \end{pmatrix} \tag{5.16}$$

Using Newton's method, we can find a numerical approximation $\bar{x} \in \mathbb{R}^{kn}$ such that $f(\bar{x}) \approx 0$, fix $r_* > 0$ and compute (with interval arithmetics) the bounds Y and Z satisfying (5.9), and try to apply Theorem 5.2.5 to conclude about existence of periodic points. This is what is done in the next two examples.

Example 5.3.2 (Periodic points in the logistic map). Consider the logistic map $g\colon \mathbb{R} \to \mathbb{R}$ given by $g(x) = \alpha x(1-x)$, where we fix the parameter to be $\alpha = 4$. Fixing $k = 3$, we applied Newton's method and found two numerical approximations of $f = 0$ in (5.15) for periodic orbits of period 3, namely

$$\bar{x}^{(1)} = \begin{pmatrix} 0.188255099070633 \\ 0.611260466978157 \\ 0.950484433951210 \end{pmatrix} \quad \text{and} \quad \bar{x}^{(2)} = \begin{pmatrix} 0.413175911166535 \\ 0.969846310392954 \\ 0.116977778440511 \end{pmatrix}. \tag{5.17}$$

Fixing $r_* = .001$ and applying Theorem 5.2.5, for $\bar{x}^{(1)}$, we proved the existence of a true zero $\tilde{x}^{(1)}$ such that $\|\tilde{x}^{(1)} - \bar{x}^{(1)}\|_\infty \leq 3.2169 \times 10^{-16}$ while for $\bar{x}^{(2)}$, we proved the existence of a true zero $\tilde{x}^{(2)}$ such that $\|\tilde{x}^{(2)} - \bar{x}^{(2)}\|_\infty \leq 6.4359 \times 10^{-17}$. Similarly, we fixed $k = 5$ and found

$$\bar{x}^{(1)} = \begin{pmatrix} 0.663533981658711 \\ 0.893026547371394 \\ 0.382120532245286 \\ 0.944417724327462 \\ 0.209971545214401 \end{pmatrix}, \quad \bar{x}^{(2)} = \begin{pmatrix} 0.979746486807249 \\ 0.079373233584409 \\ 0.292292493499057 \\ 0.827430366972642 \\ 0.571157419136643 \end{pmatrix}, \quad \bar{x}^{(3)} = \begin{pmatrix} 0.997434661695948 \\ 0.010235029373753 \\ 0.040521094189885 \\ 0.155516540462158 \\ 0.525324584419360 \end{pmatrix}$$

$$\bar{x}^{(4)} = \begin{pmatrix} 0.138132980947465 \\ 0.476209042088128 \\ 0.997735961286542 \\ 0.009035651368647 \\ 0.035816033491964 \end{pmatrix}, \quad \bar{x}^{(5)} = \begin{pmatrix} 0.625326266129361 \\ 0.937173308072291 \\ 0.235517994836519 \\ 0.720197075778817 \\ 0.806052991273831 \end{pmatrix}, \quad \bar{x}^{(6)} = \begin{pmatrix} 0.879379061346395 \\ 0.424286111247712 \\ 0.977069628200024 \\ 0.089618279396362 \\ 0.326347373577590 \end{pmatrix}$$

Close to each of the above numerical approximation $\bar{x}^{(i)}$ $(i = 1, \ldots, 6)$, we proved the existence of $\tilde{x}^{(i)} \in \mathbb{R}^5$ such that $f(\tilde{x}^{(i)}) = 0$ and that $\|\tilde{x}^{(i)} - \bar{x}^{(i)}\|_\infty \leq r_i$ with

$$r_i = \begin{cases} 1.3445 \times 10^{-16}, & i = 1, \\ 7.8784 \times 10^{-17}, & i = 2, \\ 3.739 \times 10^{-15}, & i = 3, \\ 6.817 \times 10^{-16}, & i = 4, \\ 4.787 \times 10^{-16}, & i = 5, \\ 5.2741 \times 10^{-16}, & i = 6. \end{cases}$$

Example 5.3.3 (Periodic points in the Hénon map). Consider the Hénon map $g : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$g(x) = \begin{pmatrix} 1 - \alpha x_1^2 + x_2 \\ \beta x_1 \end{pmatrix},$$

where $\alpha = 1.4$ and $\beta = 0.3$ are the parameters. We fixed $k \in \{1, 2, 3, 4, 5, 6, 7\}$ and applied Newton's method (NM) to 1000 random initial conditions in the square $[-1.5, 1.5] \times [-0.4, 0.4]$. For $k = 1$ NM found two fixed points (period one orbit), for $k = 2$ NM found

one orbit, for $k = 3$ no orbit was found by NM, for $k = 4$ NM found one orbit, for $k = 5$ no orbit was found by NM, for $k = 6$ NM found two orbits and for $k = 7$ NM found four orbits. For all ten of these orbits, we applied Theorem 5.2.5 to prove existence of true orbits nearby. For instance, the numerical approximation of the period-2 orbit is

$$\begin{pmatrix} 0.975800051175056 \\ -0.142740015352517 \end{pmatrix}, \qquad \begin{pmatrix} -0.475800051175056 \\ 0.292740015352517 \end{pmatrix} \tag{5.18}$$

for which we proved that the true period-2 orbit lies within 2.3×10^{-16} away from the numerical solution. See Figure 5.3 for the visualization of the orbits proven.

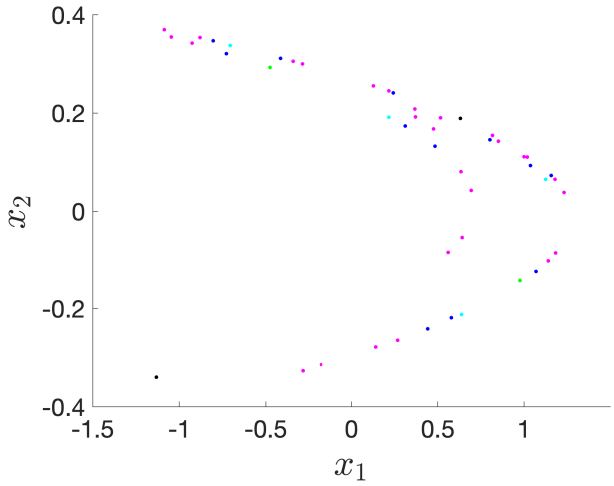

Figure 5.3: Rigorous computations of two fixed points (black), one orbit of period 2 (green), one orbit of period 4 (cyan), two orbits of period six (blue) and four orbits of period seven (magenta), in the Hénon map at parameter values $\alpha = 1.4$ and $\beta = 0.3$.

5.4 Exercises

Exercise 5.4.1. Assume $f \in C^2(\mathbb{R}^n, \mathbb{R}^n)$ and that $\tilde{x}$ is a nondegenerate zero of f. Consider the Newton operator $T(x) = x - Df(x)^{-1} f(x)$. Prove that there exists $r > 0$ such that if $x^0 \in B_r(\tilde{x})$ and $x^{n+1} = T^n(x^0)$, then the sequence x^n converges to $\tilde{x}$ at a *quadratic rate*, that is, if

$$\epsilon_n = \|f(x_n)\|,$$

then

$$\epsilon_{n+1} \leq M\epsilon_n^2,$$

for some $M > 0$.

Exercise 5.4.2 (Mean values). Prove the following version of the mean value theorem and its corollary.

Theorem 5.4.3 (The mean value theorem). *Suppose that $U \subset \mathbb{R}^n$ is open and that $f\colon U \to \mathbb{R}^m$ is C^1. Assume that $x, y \in U$ are such that the line segment from x to y is contained in U, that is that*

$$(1-t)x + ty \in U,$$

for all $t \in [0,1]$. Then

$$f(y) - f(x) = \left(\int_0^1 Df((1-t)x + ty)\, dt \right) (y - x).$$

Corollary 5.4.4 (The mean value inequality). *Suppose that $f\colon V \to \mathbb{R}^m$ is C^1 with $V \subset \mathbb{R}^n$ open and convex. Assume further that that there exists an $M > 0$ so that*

$$\sup_{y \in V} \|Df(y)\|_{B(\mathbb{R}^n, \mathbb{R}^m)} \leq M.$$

Then

$$\|f(x) - f(y)\|_{\mathbb{R}^m} \leq M \|x - y\|_{\mathbb{R}^n}.$$

Exercise 5.4.5. Use Newton-Kantorovich's Theorem 5.2.5 to give rigorous bounds on the value of $\sqrt{7}$.

Exercise 5.4.6. Use Newton-Kantorovich's Theorem 5.2.5 to give rigorous bounds on the zeros of $f(x) = xe^x - 1$.

Exercise 5.4.7. Use Newton's method in the complex variable setting to obtain numerical approximations of the three zeros of $z^3 + z + 1 = 0$. Use Newton-Kantorovich's Theorem 5.2.5 to provide rigorous bounds on the values of the true zeros.

Exercise 5.4.8 (Dynamics of a plant-herbivore model). Plant-herbivore interactions exhibit natural oscillations in the populations of both plants and the herbivore. In 1935, Nicholson-Bailey examined a discrete-time plant-herbivore model of the form

$$g(x) = g(x_1, x_2) = \begin{pmatrix} x_1 e^{\alpha(1-x_1) - \beta x_2} \\ x_1 e^{\alpha(1-x_1)}(1 - e^{-\beta x_2}) \end{pmatrix}, \tag{5.19}$$

where x_1 is the population of plant and x_2 is the population of herbivore. The parameter α is the growth rate of plants and the parameter β measures average area of leaves consumed by a herbivore. We assume the herbivore does not attack the plant before the plant grows.

a) Fix the parameters $(\alpha, \beta) = (2.975, 1.5)$ and iterate the map 500 times starting from a random initial condition. Do you observe interesting organized dynamics for larger iterations? If not, repeat the process with different random initial conditions until you observe an asymptotic (organized) pattern. Use your observation to compute a periodic orbit of period 6. Prove its existence using NKT and Interval Arithmetic.

b) Starting from the periodic orbit of part a), perform a *continuation* in the parameter α starting at $\alpha = 2.975$ and finishing at $\alpha = 2.965$. Proceed as follows. Denote by $\bar{x}^{(2.975)}$ the periodic orbit of period $k = 6$ you computed at $(\alpha, \beta) = (2.975, 1.5)$. Then decrease the parameter α by 0.001 to $\alpha = 2.974$, and starting with initial condition $\bar{x}^{(2.975)}$ apply Newton's method to obtain a new periodic orbit $\bar{x}^{(2.974)}$ of period 6. Prove the existence of a true orbit close to $\bar{x}^{(2.974)}$. Repeat the process by decreasing $\alpha = 2.974$ by 0.001, and so one. Always try to prove the existence of a period 6 orbit as you decrease α. What do you observe? Try to explain.

c) Fix the parameters $(\alpha, \beta) = (2.82, 1.5)$ and redo the same as in part a). In this case, compute another periodic orbit. What is its period? Prove its existence with a computer-assisted proof.

Exercise 5.4.9 (Relative equilibria in the CRFBP). We consider the particular case of the restricted four body problem where three point masses move in circular periodic orbits around their center of mass and a fourth massless particle interacts with this system without affecting their motion. This particular case is known as the *Circular Equilateral Restricted Four-Body Problem* (CR4BP). The equations of motion of the massless particle relative to a reference (synodic) frame can be written as

$$\ddot{x} - 2\dot{y} = \Omega_x, \quad \ddot{y} + 2\dot{x} = \Omega_y, \quad \ddot{z} = \Omega_z,$$

where $\dot{x} = dx/dt$, $\ddot{x} = d^2x/dt^2$, etc, and where

$$\Omega(x, y, z, m_1, m_2, m_3) \stackrel{\text{def}}{=} \frac{1}{2}(x^2 + y^2) + \sum_{i=1}^{3} \frac{m_i}{r_i}, \tag{5.20}$$

is called the *effective potential*, and $r_i \stackrel{\text{def}}{=} \sqrt{(x - x_i)^2 + (y - y_i)^2 + z^2}$, for $i = 1, 2, 3$. The general expressions of the coordinates of the three *primaries* in terms of the masses of the three point masses are given by

$$(x_1, y_1, z_1) = \left(\frac{-|K|\sqrt{m_2^2 + m_2 m_3 + m_3^2}}{K}, 0, 0 \right)$$
$$(x_2, y_2, z_2) = \left(\frac{|K|[(m_2 - m_3)m_3 + m_1(2m_2 + m_3)]}{2K\sqrt{m_2^2 + m_2 m_3 + m_3^2}}, \frac{-\sqrt{3}m_3}{2m_2^{3/2}} \sqrt{\frac{m_2^3}{m_2^2 + m_2 m_3 + m_3^2}}, 0 \right)$$
$$(x_3, y_3, z_3) = \left(\frac{|K|}{2\sqrt{m_2^2 + m_2 m_3 + m_3^2}}, \frac{\sqrt{3}}{2\sqrt{m_2}} \sqrt{\frac{m_2^3}{m_2^2 + m_2 m_3 + m_3^2}}, 0 \right)$$

where $K \stackrel{\text{def}}{=} m_2(m_3 - m_2) + m_1(m_2 + 2m_3)$ and the three masses satisfy the relation $m_1 + m_2 + m_3 = 1$. The equilibria of the system are given by the critical points of the

effective potential $\Omega(x, y, z, m_1, m_2, m_3)$ given by (5.20), that is they satisfy the equations $\Omega_x = 0$, $\Omega_y = 0$ and $\Omega_z = 0$. A straightforward computation shows that

$$\Omega_x = x - \sum_{i=1}^{3} \frac{m_i}{r_i^3}(x - x_i), \quad \Omega_y = y - \sum_{i=1}^{3} \frac{m_i}{r_i^3}(y - y_i), \quad \Omega_z = -z \sum_{i=1}^{3} \frac{m_i}{r_i^3},$$

and therefore as a consequence all the equilibrium points are coplanar, since $\Omega_z = 0 \Rightarrow z = 0$. Hence, the relative equilibria $(x, y, 0)$ satisfy the equation

$$f(x, y) \stackrel{\text{def}}{=} \begin{pmatrix} x - \displaystyle\sum_{i=1}^{3} \frac{m_i(x - x_i)}{\left(\sqrt{(x - x_i)^2 + (y - y_i)^2}\right)^3} \\ y - \displaystyle\sum_{i=1}^{3} \frac{m_i(y - y_i)}{\left(\sqrt{(x - x_i)^2 + (y - y_i)^2}\right)^3} \end{pmatrix} = 0 \tag{5.21}$$

Using the Newton-Kantorovich Theorem 5.2.5, prove that (5.21) has at least ten equilibrium points for the case $m_1 = m_2 = m_3 = 1/3$ and eight equilibrium points for the non equal masses case $m_1 = 0.9987451087$, $m_2 = 0.0010170039$ and $m_3 = 0.0002378873$.