

Chapter 3

Solution of nonlinear systems

Introduction

This chapter concerns the numerical solution of nonlinear equations of the general form

$$\mathbf{f}(\mathbf{x}) = 0, \quad \mathbf{f}: \mathbf{R}^n \rightarrow \mathbf{R}^n. \quad (3.1)$$

A solution to this equation is called a *zero* of the function f . Except in particular cases (for example linear systems), there does not exist a numerical method for solving (3.1) in a finite number of operations, so iterative methods are required.

In contrast with the previous chapter, it may not be the case that (3.1) admits one and only one solution. For example, the equation $1 + x^2 = 0$ does not have a (real) solution, and the equation $\cos(x) = 0$ has infinitely many. Therefore, convergence results usually contain assumptions on the function f that guarantee the existence and uniqueness of a solution in \mathbf{R}^n or a subset of \mathbf{R}^n .

For an iterative method generating approximations $(\mathbf{x}_k)_{k \geq 0}$ of a root \mathbf{x}_* , we define the error as $\mathbf{e}_k = \mathbf{x}_k - \mathbf{x}_*$. If the sequence $(\mathbf{x}_k)_{k \geq 0}$ converges to \mathbf{x}_* in the limit as $k \rightarrow \infty$ and if

$$\lim_{k \rightarrow \infty} \frac{\|\mathbf{e}_{k+1}\|}{\|\mathbf{e}_k\|^q} = r, \quad (3.2)$$

then we say that $(\mathbf{x}_k)_{k \geq 0}$ converges with *order of convergence* q and *rate of convergence* r . In addition, we say that the convergence is linear $q = 1$, and quadratic if $q = 2$. The convergence is said to be superlinear if

$$\lim_{k \rightarrow \infty} \frac{\|\mathbf{e}_{k+1}\|}{\|\mathbf{e}_k\|} = 0. \quad (3.3)$$

In particular, the convergence is superlinear if the order of convergence is $q > 1$.

Remark 3.1. The definition (3.3) for the order and rate of convergence is not entirely satisfactory, as the limit may not exist. A more general definition for the order of convergence of

a sequence $(\mathbf{x}_k)_{k \geq 0}$ is the following:

$$q(\mathbf{x}_0) = \inf \left\{ p \in [1, \infty) : \limsup_{k \rightarrow \infty} \frac{\|\mathbf{e}_{k+1}\|}{\|\mathbf{e}_k\|^p} = \infty \right\},$$

or $q(\mathbf{x}_0) = \infty$ if the numerator and denominator of the fraction are zero for sufficiently large k . It is possible to define similarly the order of convergence of an iterative method for an initial guess in a neighborhood V of \mathbf{x}_* :

$$q = \inf \left\{ p \in [1, \infty) : \sup_{\mathbf{x}_0 \in V} \left(\limsup_{k \rightarrow \infty} \frac{\|\mathbf{e}_{k+1}\|}{\|\mathbf{e}_k\|^p} \right) = \infty \right\},$$

where the fraction should be interpreted as 0 if the numerator and denominator are zero. A more detailed discussion of this subject is beyond the scope of this course.

The rest of chapter is organized as follows:

- In [Section 3.1](#), by way of introduction to the subject of numerical methods for nonlinear equations, we present and analyze the bisection method.
- In [Section 3.2](#), we present a general method based on a fixed point iteration for solving (3.1). The convergence of this method is analyzed in [Section 3.3](#).
- In [Section 3.4](#), two concrete examples of fixed point methods are studied: the chord method and the Newton–Raphson method.

3.1 The bisection method

As an introduction to numerical methods for solving nonlinear equations, we present the bisection method. This method applies only in the case of a real-valued function $f: \mathbf{R} \rightarrow \mathbf{R}$, and relies on the knowledge of two points $a < b$ such that $f(a)$ and $f(b)$ have different signs. By the intermediate value theorem, there necessarily exists $x_* \in (a, b)$ such that $f(x_*) = 0$. The idea of the bisection method is to successively divide the interval in two equal parts, and to retain, based on the sign of f at the midpoint $x_{1/2}$, the one that necessarily contains a root. If $f(x_{1/2})f(a) \geq 0$, then $f(x_{1/2})f(b) \leq 0$ and so there necessarily exists a root of f in the interval $[x_{1/2}, b)$ by the intermediate value theorem. In contrast, if $f(x_{1/2})f(a) < 0$, then there necessarily is a root in the interval $(a, x_{1/2})$. The algorithm is presented in [Algorithm 7](#).

The following result establishes the convergence of the method.

Proposition 3.1. *Assume that $f: \mathbf{R} \rightarrow \mathbf{R}$ is a continuous function. Let $[a_j, b_j]$ denote the interval obtained after j iterations of the bisection method, and let x_j denote the midpoint $(a_j + b_j)/2$. Then there exists a root x_* of f such that*

$$|x_j - x_*| \leq (b_0 - a_0)2^{-(j+1)}. \quad (3.4)$$

Proof. By construction, $f(a_j)f(b_j) \leq 0$ and $f(b) \neq 0$. Therefore, by the intermediate value

Algorithm 7 Bisection method

Assume that $f(a)f(b) < 0$ with $a < b$.
 Pick $\varepsilon > 0$.
 $x \leftarrow a/2 + b/2$
while $|b - a| \geq \varepsilon$ **do**
 if $f(x)f(a) \geq 0$ **then**
 $a \leftarrow x$
 else
 $b \leftarrow x$
 end if
 $x \leftarrow a/2 + b/2$
end while

theorem, there exists a root of f in the interval $[a_j, b_j]$, implying that

$$|x_j - x_*| \leq \frac{b_j - a_j}{2}.$$

Since $b_j - a_j = 2^{-j}(b_0 - a_0)$, the statement follows. \square

Although the limit in (3.2) may not be well-defined (for example, x_1 may be a root of f), the error $x_j - x_*$ is bounded in absolute value by the sequence $(\tilde{e}_j)_{j \geq 0}$, where $\tilde{e}_j = (b_0 - a_0)2^{-(j+1)}$ by Proposition 3.1. Since the latter sequence exhibits linear convergence to 0, the convergence of the bisection method is said to be linear, by a slight abuse of terminology.

3.2 Fixed point methods

Let \mathbf{x}_* denote a zero of the function \mathbf{f} . The idea of iterative methods for (3.1) is to construct, starting from an initial guess \mathbf{x}_0 , a sequence $(\mathbf{x}_k)_{k=0,1,\dots}$ approaching \mathbf{x}_* . A number of iterative methods for solving (3.1) are based on an iteration of the form

$$\mathbf{x}_{k+1} = \mathbf{F}(\mathbf{x}_k), \tag{3.5}$$

for an appropriate continuous function \mathbf{F} . Assume that \mathbf{x}_k converges to some point $\mathbf{x}_* \in \mathbf{R}^n$ in the limit as $k \rightarrow \infty$. Then, taking the limit $k \rightarrow \infty$ in (3.5), we find that \mathbf{x}_* satisfies

$$\mathbf{F}(\mathbf{x}_*) = \mathbf{x}_*.$$

Such a point \mathbf{x}_* is called a *fixed point* of the function \mathbf{F} . Several definitions of the function \mathbf{F} can be employed in order to ensure that a fixed point of \mathbf{F} coincides with a zero of \mathbf{f} . One may, for example, define $\mathbf{F}(\mathbf{x}) = \mathbf{x} - \alpha^{-1}\mathbf{f}(\mathbf{x})$, for some nonzero scalar coefficient α . Then $\mathbf{F}(\mathbf{x}_*) = \mathbf{x}_*$ if and only if $\mathbf{f}(\mathbf{x}_*) = 0$. Later in this chapter, in Section 3.4, we study two instances of numerical methods which can be recast in the form (3.5). Before this, we study the convergence of the iteration (3.5) for a general function \mathbf{F} .

3.3 Convergence of fixed point methods

Equation (3.5) may be viewed as a *discrete-time* dynamical system. In order to study the behavior of the system as $k \rightarrow \infty$, it is important to understand the concept of stability of a fixed point. The concept of stability appears also in the field of ordinary differential equations, which are *continuous-time* dynamical systems. Before we define this concept, we introduce the following notation for the open ball of radius δ around $\mathbf{x} \in \mathbf{R}^n$:

$$B_\delta(\mathbf{x}) := \{\mathbf{y} \in \mathbf{R}^n : \|\mathbf{y} - \mathbf{x}\| < \delta\}.$$

Definition 3.1 (Stability of fixed points). Let $(\mathbf{x}_k)_{k \geq 0}$ denote iterates obtained from (3.5) when starting from an initial vector \mathbf{x}_0 . Then we say that a fixed point \mathbf{x}_* is

- an *attractor* if there exists a neighborhood \mathcal{V} of s such that

$$\forall \mathbf{x}_0 \in \mathcal{V}, \quad \mathbf{x}_k \xrightarrow[k \rightarrow \infty]{} \mathbf{x}_*. \quad (3.6)$$

The largest neighborhood for which this is true, i.e. the set of values of \mathbf{x}_0 such that (3.6) holds true, is called the basin of attraction of \mathbf{x}_* .

- stable (in the sense of Lyapunov) if for all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\forall \mathbf{x}_0 \in B_\delta(\mathbf{x}_*), \quad \|\mathbf{x}_k - \mathbf{x}_*\| < \varepsilon.$$

- asymptotically stable if it is stable and an attractor.
- exponentially stable if there exists $C > 0$, $\alpha \in (0, 1)$, and $\delta > 0$ such that

$$\forall \mathbf{x}_0 \in B_\delta(\mathbf{x}_*), \quad \forall k \in \mathbf{N}, \quad \|\mathbf{x}_k - \mathbf{x}_*\| \leq C\alpha^k \|\mathbf{x}_0 - \mathbf{x}_*\|.$$

- globally exponentially stable if there exists $C > 0$, $\alpha \in (0, 1)$ such that

$$\forall \mathbf{x}_0 \in \mathbf{R}^n, \quad \forall k \in \mathbf{N}, \quad \|\mathbf{x}_k - \mathbf{x}_*\| \leq C\alpha^k \|\mathbf{x}_0 - \mathbf{x}_*\|.$$

- Unstable if it is not stable.

Clearly, global exponential stability implies exponential stability, which itself implies asymptotic stability and stability. If \mathbf{x}_* is globally exponentially stable, then \mathbf{x}_* is the unique fixed point of \mathbf{F} ; showing this is the aim of [Exercise 3.3](#). If \mathbf{x}_* is an attractor, then the dynamical system (3.5) is said to be locally convergent to \mathbf{x}_* . The larger the basin of attraction of \mathbf{x}_* , the less careful we need to be when picking the initial guess \mathbf{x}_0 . Global exponential stability of a fixed point can sometimes be shown provided that \mathbf{F} satisfies a strong hypothesis.

Definition 3.2 (Lipschitz continuity). A function $\mathbf{F}: \mathbf{R}^n \rightarrow \mathbf{R}^n$ is said to be *Lipschitz*

continuous with constant L if

$$\forall (\mathbf{x}, \mathbf{y}) \in \mathbf{R}^n \times \mathbf{R}^n, \quad \|\mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\|.$$

A function $\mathbf{F}: \mathbf{R}^n \rightarrow \mathbf{R}^n$ that is Lipschitz continuous with a constant $L < 1$ is called a *contraction*. For such a function, it is possible to prove that (3.5) has a unique globally exponentially stable fixed point.

Theorem 3.2. *Assume that \mathbf{F} is a contraction. Then there exists a unique fixed point of (3.5), and it holds that*

$$\forall \mathbf{x}_0 \in \mathbf{R}^n, \quad \forall k \in \mathbf{N}, \quad \|\mathbf{x}_k - \mathbf{x}_*\| \leq L^k \|\mathbf{x}_0 - \mathbf{x}_*\|.$$

Proof. It holds that

$$\|\mathbf{x}_{k+2} - \mathbf{x}_{k+1}\| = \|\mathbf{F}(\mathbf{x}_{k+1}) - \mathbf{F}(\mathbf{x}_k)\| \leq L\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \dots \leq L^{k+1}\|\mathbf{x}_1 - \mathbf{x}_0\|.$$

Therefore, for any $n \geq m$, we have by the triangle inequality

$$\begin{aligned} \|\mathbf{x}_n - \mathbf{x}_m\| &\leq \|\mathbf{x}_n - \mathbf{x}_{n-1}\| + \dots + \|\mathbf{x}_{m+1} - \mathbf{x}_m\| \\ &\leq (L^{n-1} + \dots + L^m)\|\mathbf{x}_1 - \mathbf{x}_0\| \leq L^m(1 + L + \dots)\|\mathbf{x}_1 - \mathbf{x}_0\| = \frac{L^m}{1-L}\|\mathbf{x}_1 - \mathbf{x}_0\|. \end{aligned}$$

It follows that the sequence $(\mathbf{x}_k)_{k \geq 0}$ is Cauchy in \mathbf{R}^n , implying by completeness that $\mathbf{x}_k \rightarrow \mathbf{x}_*$ in the limit as $k \rightarrow \infty$, for some limit $\mathbf{x}_* \in \mathbf{R}^n$. Being a contraction, the function \mathbf{F} is continuous, and so taking the limit $k \rightarrow \infty$ in (3.5) we obtain that \mathbf{x}_* is a fixed point of \mathbf{F} . Then

$$\|\mathbf{x}_k - \mathbf{x}_*\| = \|\mathbf{F}(\mathbf{x}_{k-1}) - \mathbf{F}(\mathbf{x}_*)\| \leq L\|\mathbf{x}_{k-1} - \mathbf{x}_*\| \leq \dots \leq L^k\|\mathbf{x}_0 - \mathbf{x}_*\|, \quad (3.7)$$

proving the statement. To show the uniqueness of the fixed point, assume \mathbf{y}_* is another fixed point. Then,

$$\|\mathbf{y}_* - \mathbf{x}_*\| = \|\mathbf{F}(\mathbf{y}_*) - \mathbf{F}(\mathbf{x}_*)\| \leq L\|\mathbf{y}_* - \mathbf{x}_*\|,$$

which implies that $\mathbf{y}_* = \mathbf{x}_*$ since $L < 1$. □

It is possible to prove a weaker, local result under a less restrictive assumptions on the function \mathbf{F} .

Theorem 3.3. *Assume that \mathbf{x}_* is a fixed point of (3.5) and that $\mathbf{F}: \mathbf{R}^n \rightarrow \mathbf{R}^n$ satisfies the local Lipschitz condition*

$$\forall \mathbf{x} \in B_\delta(\mathbf{x}_*), \quad \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_*)\| \leq L\|\mathbf{x} - \mathbf{x}_*\|, \quad (3.8)$$

with $0 \leq L < 1$ and $\delta > 0$. Then \mathbf{x}_ is the unique fixed point of \mathbf{F} in $B_\delta(\mathbf{x}_*)$ and, for all $\mathbf{x}_0 \in B_\delta(\mathbf{x}_*)$, it holds that*

- All the iterates $(\mathbf{x}_k)_{k \in \mathbf{N}}$ belong to $B_\delta(\mathbf{x}_*)$.
- The sequence $(\mathbf{x}_k)_{k \in \mathbf{N}}$ converges exponentially to \mathbf{x}_* .

Proof. See Exercise 3.4. □

It is possible to guarantee that condition (3.8) holds provided that we have sufficiently good control of the derivatives of the function \mathbf{F} . The function \mathbf{F} is differentiable at \mathbf{x} (in the sense of Fréchet) if there exists a linear operator $D\mathbf{F}_{\mathbf{x}}: \mathbf{R}^n \rightarrow \mathbf{R}^n$ such that

$$\lim_{\mathbf{h} \rightarrow 0} \frac{\|\mathbf{F}(\mathbf{x} + \mathbf{h}) - \mathbf{F}(\mathbf{x}) - D\mathbf{F}_{\mathbf{x}}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0. \quad (3.9)$$

If \mathbf{F} is differentiable, then all its first partial derivatives exist. The Jacobian matrix of \mathbf{F} at \mathbf{x} is defined as

$$\mathbf{J}_F(\mathbf{x}) = \begin{pmatrix} \partial_1 F_1(\mathbf{x}) & \dots & \partial_n F_1(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \partial_1 F_n(\mathbf{x}) & \dots & \partial_n F_n(\mathbf{x}) \end{pmatrix},$$

and it holds that $D\mathbf{F}_{\mathbf{x}}(\mathbf{h}) = \mathbf{J}_F(\mathbf{x})\mathbf{h}$.

Proposition 3.4. *Let \mathbf{x}_* be a fixed point of (3.5), and assume that there exists δ and a subordinate matrix norm such that \mathbf{F} is differentiable everywhere in $B_\delta(\mathbf{x}_*)$ and*

$$\forall \mathbf{x} \in B_\delta(\mathbf{x}_*), \quad \|\mathbf{J}_F(\mathbf{x})\| \leq L < 1.$$

Then condition (3.8) is satisfied in the associated vector norm, and so the fixed point \mathbf{x}_ is locally exponentially stable.*

Proof. Let $\mathbf{x} \in B_\delta(\mathbf{x}_*)$. By the fundamental theorem of calculus and the chain rule, we have

$$\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_*) = \int_0^1 \frac{d}{dt} \left(\mathbf{F}(\mathbf{x}_* + t(\mathbf{x} - \mathbf{x}_*)) \right) dt = \int_0^1 \mathbf{J}_F(\mathbf{x}_* + t(\mathbf{x} - \mathbf{x}_*)) (\mathbf{x} - \mathbf{x}_*) dt.$$

Therefore, it holds that

$$\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_*)\| \leq \int_0^1 \|\mathbf{J}_F(\mathbf{x}_* + t(\mathbf{x} - \mathbf{x}_*))\| dt \|\mathbf{x} - \mathbf{x}_*\| = \int_0^1 L dt \|\mathbf{x} - \mathbf{x}_*\| = L \|\mathbf{x} - \mathbf{x}_*\|,$$

which is the statement. □

In fact, it is possible to prove that a fixed point \mathbf{x}_* is exponentially locally stable under an even weaker condition, involving only the derivative of \mathbf{F} at \mathbf{x}_* .

Proposition 3.5. *Let \mathbf{x}_* be a fixed point of (3.5) and that f is differentiable at \mathbf{x}_* with*

$$\|\mathbf{J}_F(\mathbf{x}_*)\| = L < 1,$$

in a subordinate vector norm. Then there exists $\delta > 0$ such that condition (3.8) is satisfied in the associated vector norm, and so the fixed point \mathbf{x}_* is locally exponentially stable.

Proof. Let $\varepsilon = \frac{1}{2}(1 - L) > 0$. By the definition of differentiability (3.9), there exists $\delta > 0$ such that

$$\forall \mathbf{h} \in B_\delta(\mathbf{0}) \setminus \{\mathbf{0}\}, \quad \frac{\|\mathbf{F}(\mathbf{x}_* + \mathbf{h}) - \mathbf{F}(\mathbf{x}_*) - \mathbf{J}_F(\mathbf{x}_*)\mathbf{h}\|}{\|\mathbf{h}\|} \leq \varepsilon.$$

By the triangle inequality, this implies

$$\begin{aligned} \forall \mathbf{h} \in B_\delta(\mathbf{0}), \quad \|\mathbf{F}(\mathbf{x}_* + \mathbf{h}) - \mathbf{F}(\mathbf{x}_*)\| &\leq \|\mathbf{F}(\mathbf{x}_* + \mathbf{h}) - \mathbf{F}(\mathbf{x}_*) - \mathbf{J}_F(\mathbf{x}_*)\mathbf{h}\| + \|\mathbf{J}_F(\mathbf{x}_*)\mathbf{h}\| \\ &\leq \varepsilon\|\mathbf{h}\| + \|\mathbf{J}_F(\mathbf{x}_*)\|\|\mathbf{h}\| = (\varepsilon + L)\|\mathbf{h}\| = (1 - \varepsilon)\|\mathbf{h}\|. \end{aligned}$$

This shows that \mathbf{F} satisfies the condition (3.8) in the neighborhood $B_\delta(\mathbf{x}_*)$. \square

The estimate in Theorem 3.2 suggests that when the fixed point iteration (3.5) converges, the convergence is linear. While this is usually the case, the convergence is superlinear if $\mathbf{J}_F(\mathbf{x}_*) = \mathbf{0}$.

Proposition 3.6. *Assume that \mathbf{x}_* is a fixed point of (3.5) and that $\mathbf{J}_F(\mathbf{x}_*) = \mathbf{0}$. Then the convergence to \mathbf{x}_* is superlinear, in the sense that if $\mathbf{x}_k \rightarrow \mathbf{x}_*$ as $k \rightarrow \infty$, then*

$$\lim_{k \rightarrow \infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_*\|}{\|\mathbf{x}_k - \mathbf{x}_*\|} = 0.$$

Proof. By Proposition 3.5, there exists $\delta > 0$ such that $(\mathbf{x}_k)_{k \geq 0}$ is a sequence converging to \mathbf{x}_* for all $\mathbf{x}_0 \in B_\delta(\mathbf{x}_*)$. It holds that

$$\frac{\|\mathbf{x}_{k+1} - \mathbf{x}_*\|}{\|\mathbf{x}_k - \mathbf{x}_*\|} = \frac{\|\mathbf{F}(\mathbf{x}_k) - \mathbf{F}(\mathbf{x}_*)\|}{\|\mathbf{x}_k - \mathbf{x}_*\|} = \frac{\|\mathbf{F}(\mathbf{x}_k) - \mathbf{F}(\mathbf{x}_*) - \mathbf{J}_F(\mathbf{x}_*)(\mathbf{x}_k - \mathbf{x}_*)\|}{\|\mathbf{x}_k - \mathbf{x}_*\|}.$$

Since $\mathbf{x}_k - \mathbf{x}_* \rightarrow \mathbf{0}$ as $k \rightarrow \infty$, the right-hand side converges to 0 by (3.9). \square

Similarly, if there exist $\delta > 0$, $C > 0$ and $q \in (1, \infty)$ such that

$$\forall \mathbf{x} \in B_\delta(\mathbf{x}_*), \quad \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_*)\| \leq C\|\mathbf{x} - \mathbf{x}_*\|^q,$$

then assuming that $(\mathbf{x}_k)_{k \geq 0}$ converges to \mathbf{x}_* , it holds for sufficiently large k that

$$\frac{\|\mathbf{x}_{k+1} - \mathbf{x}_*\|}{\|\mathbf{x}_k - \mathbf{x}_*\|^q} = \frac{\|\mathbf{F}(\mathbf{x}_k) - \mathbf{F}(\mathbf{x}_*)\|}{\|\mathbf{x}_k - \mathbf{x}_*\|^q} \leq C.$$

In this case, the order of convergence is at least q .

3.4 Examples of fixed point methods

As we mentioned in Section 3.2, there are several choices for the function \mathbf{F} that guarantee the equivalence $\mathbf{F}(\mathbf{x}) = \mathbf{x} \Leftrightarrow \mathbf{f}(\mathbf{x}) = \mathbf{0}$. In the case where f is a function from \mathbf{R} to \mathbf{R} , the simplest approach, sometimes called the *chord method*, is to define

$$F(x) = x - \alpha^{-1}f(x).$$

The fixed point iteration (3.4) in this case admits a simple geometric interpretation: at each step, the function f is approximated by the affine function $x \mapsto f(x_k) + \alpha(x - x_k)$, and the new iterate is defined as the zero of this affine function, i.e.

$$x_{k+1} = x_k - \alpha^{-1}f(x_k) = F(x_k). \quad (3.10)$$

By Proposition 3.5, a sufficient condition to ensure local convergence is that

$$|F'(x_*)| = |1 - \alpha^{-1}f'(x_*)| < 1. \quad (3.11)$$

In order for this condition to hold true, the slope α must be of the same sign as $f'(x_*)$ and the inequality $|\alpha| \geq |f'(x_*)|/2$ must be satisfied. If $f'(x_*) = 0$, then the sufficient condition (3.11) is never satisfied; in this case, the convergence must be studied on a case-by-case basis. By Proposition 3.6, the convergence of the chord method is superlinear if $\alpha = f'(x_*)$. In practice, the solution x_* is unknown, and so this choice is not realistic. Nevertheless, the above reasoning suggests that, by letting the slope α vary from iteration to iteration in such a manner that α_k approaches $f'(x_*)$ as $k \rightarrow \infty$, fast convergence can be obtained. This is precisely what the Newton–Raphson method aims to achieve; see Section 3.4.1

When \mathbf{f} is a function from \mathbf{R}^n to \mathbf{R}^n , the above approach generalizes to

$$x_{k+1} = \mathbf{F}(x_k), \quad \mathbf{F}(x) = x - \mathbf{A}^{-1}\mathbf{f}(x),$$

where \mathbf{A} is an invertible matrix. The geometric interpretation of the method in this case is the following: at each step, the function \mathbf{f} is approximated by the affine function $x \mapsto x_k + \mathbf{A}(x - x_k)$, and the next iterate is given by the unique zero of the latter function. Superlinear convergence is achieved when $\mathbf{A} = \mathbf{J}_f(x_*)$. Notice that each iteration requires to calculate $\mathbf{y} := \mathbf{A}^{-1}\mathbf{f}(x_k)$, which is generally achieved by solving the linear system $\mathbf{A}\mathbf{y} = \mathbf{f}(x_k)$.

3.4.1 The Newton–Raphson method

Let us first consider the case of a function from \mathbf{R} to \mathbf{R} . The Newton–Raphson method applies when f is continuously differentiable. At each step, the function f is approximated by the affine function $x \mapsto f(x_k) + f'(x_k)(x - x_k)$ and the unique zero of this function is returned. In other words, one iteration of the Newton–Raphson method reads

$$x_{k+1} = x_k - f'(x_k)^{-1}f(x_k). \quad (3.12)$$

For this iteration to be well-defined, it is necessary that $f'(x_k) \neq 0$. The Newton–Raphson method may be viewed as a variation on (3.10) where the slope α is adapted as the iteration progresses. If the method converges, then $f'(x_k) \rightarrow f'(x_*)$ in the limit as $k \rightarrow \infty$, which indicates that superlinear convergence could occur. Equation (3.12) may be recast as a fixed point iteration of the form (3.4) with

$$F(x) = x - \frac{f(x)}{f'(x)}.$$

Clearly, if x_* is a simple root of f , that is if $f(x_*) = 0$ and $f'(x_*) \neq 0$, then x_* is a fixed point of F . If the function f is twice continuously differentiable, then the convergence of the Newton–Raphson method is superlinear by [Proposition 3.6](#), because then

$$F'(x_*) = \frac{f(x_*)f''(x_*)}{f'(x_*)^2} = 0.$$

The Newton–Raphson method generalizes to nonlinear equations in \mathbf{R}^n of the form [\(3.1\)](#). In this case $\mathbf{F}(\mathbf{x}) = \mathbf{x} - \mathbf{J}_f(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$, and so an iteration of the method reads

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{J}_f(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k). \quad (3.13)$$

In the rest of this section, we show that the iteration [\(3.13\)](#) is well-defined in a small neighborhood of a root of \mathbf{f} under appropriate assumptions, and we demonstrate the *second order* convergence of the method. We begin by proving the following preparatory lemma, which we will then employ in the particular case where the matrix-valued function \mathbf{A} is equal to \mathbf{J}_f .

Lemma 3.7. *Let $\mathbf{A}: \mathbf{R}^n \rightarrow \mathbf{R}^{n \times n}$ denote a matrix-valued function on \mathbf{R}^n that is both continuous and nonsingular at \mathbf{x}_* , and let \mathbf{f} be a function that is differentiable at \mathbf{x}_* where $\mathbf{f}(\mathbf{x}_*) = 0$. Then the function*

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - \mathbf{A}(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$$

is well-defined in a neighborhood $B_\delta(\mathbf{x}_)$ of \mathbf{x}_* . In addition, \mathbf{G} is differentiable at \mathbf{x}_* with*

$$\mathbf{J}_G(\mathbf{x}_*) = \mathbf{I} - \mathbf{A}(\mathbf{x}_*)^{-1}\mathbf{J}_f(\mathbf{x}_*). \quad (3.14)$$

Proof. It holds that

$$\mathbf{A}(\mathbf{x}) = \left(\mathbf{A}(\mathbf{x}_*) - (\mathbf{A}(\mathbf{x}_*) - \mathbf{A}(\mathbf{x})) \right) = \mathbf{A}(\mathbf{x}_*) \left(\mathbf{I} - \mathbf{A}(\mathbf{x}_*)^{-1}(\mathbf{A}(\mathbf{x}_*) - \mathbf{A}(\mathbf{x})) \right). \quad (3.15)$$

Let $\beta = \|\mathbf{A}(\mathbf{x}_*)^{-1}\|$ and $\varepsilon = (2\beta)^{-1}$. By continuity of the matrix-valued function \mathbf{A} , there exists $\delta > 0$ such that

$$\forall \mathbf{x} \in B_\delta(\mathbf{x}_*), \quad \|\mathbf{A}(\mathbf{x}) - \mathbf{A}(\mathbf{x}_*)\| \leq \varepsilon.$$

For $\mathbf{x} \in B_\delta(\mathbf{x}_*)$ we have $\|\mathbf{A}(\mathbf{x}_*)^{-1}(\mathbf{A}(\mathbf{x}_*) - \mathbf{A}(\mathbf{x}))\| \leq \|\mathbf{A}(\mathbf{x}_*)^{-1}\| \|\mathbf{A}(\mathbf{x}_*) - \mathbf{A}(\mathbf{x})\| \leq \beta\varepsilon = \frac{1}{2}$, and so [Lemma 2.3](#) implies that the second factor on the right-hand side of [\(3.15\)](#) is invertible with a norm bounded from above by 2. Therefore, we deduce that $\mathbf{A}(\mathbf{x})$ is invertible with

$$\forall \mathbf{x} \in B_\delta(\mathbf{x}_*), \quad \|\mathbf{A}(\mathbf{x})^{-1}\| \leq 2\|\mathbf{A}(\mathbf{x}_*)^{-1}\| = 2\beta.$$

In order to prove [\(3.14\)](#), we need to show that

$$\lim_{\|\mathbf{h}\| \rightarrow 0} \frac{\|\mathbf{G}(\mathbf{x}_* + \mathbf{h}) - \mathbf{G}(\mathbf{x}_*) - (\mathbf{I} - \mathbf{A}(\mathbf{x}_*)^{-1}\mathbf{J}_f(\mathbf{x}_*))\mathbf{h}\|}{\|\mathbf{h}\|} = 0$$

By definition of \mathbf{G} , the argument of the norm in the numerator is equal to

$$\begin{aligned} & \mathbf{A}(\mathbf{x}_*)^{-1}\mathbf{f}(\mathbf{x}_*) - \mathbf{A}(\mathbf{x}_* + \mathbf{h})^{-1}\mathbf{f}(\mathbf{x}_* + \mathbf{h}) + \mathbf{A}(\mathbf{x}_*)^{-1}\mathbf{J}_f(\mathbf{x}_*)\mathbf{h} \\ &= \underbrace{(\mathbf{A}^{-1}(\mathbf{x}_*) - \mathbf{A}(\mathbf{x}_* + \mathbf{h})^{-1})\mathbf{J}_f(\mathbf{x}_*)\mathbf{h}}_{=: \mathbf{v}_1} - \underbrace{\mathbf{A}(\mathbf{x}_* + \mathbf{h})^{-1}(\mathbf{f}(\mathbf{x}_* + \mathbf{h}) - \mathbf{f}(\mathbf{x}_*) - \mathbf{J}_f(\mathbf{x}_*)\mathbf{h})}_{=: \mathbf{v}_2}. \end{aligned}$$

Noting that $\mathbf{A}^{-1}(\mathbf{x}_*) - \mathbf{A}(\mathbf{x}_* + \mathbf{h})^{-1} = \mathbf{A}(\mathbf{x}_*)^{-1}(\mathbf{A}(\mathbf{x}_* + \mathbf{h}) - \mathbf{A}(\mathbf{x}_*))\mathbf{A}(\mathbf{x}_* + \mathbf{h})^{-1}$, we bound the norm of the first term on the right-hand as follows:

$$\forall \mathbf{h} \in B_\delta(\mathbf{0}), \quad \|\mathbf{v}_1\| \leq 2\beta^2 \|\mathbf{A}(\mathbf{x}_* + \mathbf{h}) - \mathbf{A}(\mathbf{x}_*)\| \|\mathbf{J}_f(\mathbf{x}_*)\| \|\mathbf{h}\|.$$

Clearly $\|\mathbf{v}_1\|/\|\mathbf{h}\| \rightarrow 0$ is the limit as $\mathbf{h} \rightarrow \mathbf{0}$ by continuity of the matrix function \mathbf{A} . It also holds that $\|\mathbf{v}_2\|/\|\mathbf{h}\| \rightarrow 0$ by differentiability of \mathbf{f} at \mathbf{x}_* , which concludes the proof. \square

Using this lemma, we can show the following result on the convergence of the multi-dimensional Newton–Raphson method.

Theorem 3.8 (Convergence of Newton–Raphson). *Let $\mathbf{f}: \mathbf{R}^n \rightarrow \mathbf{R}^n$ denote a function that is differentiable in a neighborhood $B_\delta(\mathbf{x}_*)$ of a point \mathbf{x}_* where $\mathbf{f}(\mathbf{x}_*) = \mathbf{0}$. Assume that $\mathbf{J}_f(\mathbf{x})$ is nonsingular and continuous at \mathbf{x}_* . Then \mathbf{x}_* is an attractor of the Newton–Raphson iteration (3.13) and the convergence is superlinear.*

In addition, there is $\alpha > 0$ and such that the Lipschitz condition

$$\forall \mathbf{x} \in B_\delta(\mathbf{x}_*), \quad \|\mathbf{J}_f(\mathbf{x}) - \mathbf{J}_f(\mathbf{x}_*)\| \leq \alpha \|\mathbf{x} - \mathbf{x}_*\|$$

is satisfied, then the convergence is at least quadratic: there exists $d \in (0, \delta)$ such that

$$\forall \mathbf{x}_k \in B_d(\mathbf{x}_*), \quad \|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq C \|\mathbf{x}_k - \mathbf{x}_*\|^2.$$

Proof. Using Lemma 3.7, we obtain that the Newton–Raphson update

$$\mathbf{F}(\mathbf{x}) = \mathbf{x} - \mathbf{J}_f(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x}),$$

is well-defined in a neighborhood $B_\delta(\mathbf{x}_*)$ of \mathbf{x}_* for sufficiently small δ . In addition, the second statement in Lemma 3.7 gives that $\mathbf{J}_F(\mathbf{x}_*)^{-1} = \mathbf{I} - \mathbf{J}_F(\mathbf{x}_*)\mathbf{J}_F(\mathbf{x}_*) = \mathbf{0}$, which establishes the superlinear convergence by Proposition 3.6.

In order to show that the convergence is quadratic, we begin by noticing that, since

$$\mathbf{f}(\mathbf{x}_k) = \int_0^1 \frac{d}{dt} \mathbf{f}(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) dt = \int_0^1 \mathbf{J}_f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) (\mathbf{x}_k - \mathbf{x}_*) dt,$$

it holds for all $\mathbf{x}_k \in B_\delta(\mathbf{x}_*)$ that

$$\begin{aligned} \|\mathbf{f}(\mathbf{x}_k) - \mathbf{J}_f(\mathbf{x}_*)(\mathbf{x}_k - \mathbf{x}_*)\| &= \left\| \int_0^1 \left(\mathbf{J}_f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) - \mathbf{J}_f(\mathbf{x}_*) \right) (\mathbf{x}_k - \mathbf{x}_*) dt \right\| \\ &\leq \int_0^1 \|\mathbf{J}_f(\mathbf{x}_* + t(\mathbf{x}_k - \mathbf{x}_*)) - \mathbf{J}_f(\mathbf{x}_*)\| \|\mathbf{x}_k - \mathbf{x}_*\| dt \\ &\leq \int_0^1 \alpha t \|\mathbf{x}_k - \mathbf{x}_*\|^2 dt \leq \frac{\alpha}{2} \|\mathbf{x}_k - \mathbf{x}_*\|^2. \end{aligned} \quad (3.16)$$

Let $d \in (0, \delta)$ be sufficiently small to ensure that

$$\forall \mathbf{x} \in B_d(\mathbf{x}_*), \quad \|\mathbf{J}_f(\mathbf{x})^{-1}\| \leq 2\|\mathbf{J}_f(\mathbf{x}_*)^{-1}\|.$$

Using the inequality (3.16), we have that for all $\mathbf{x}_k \in B_d(\mathbf{x}_*)$,

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}_*\| &= \|\mathbf{F}(\mathbf{x}_k) - \mathbf{x}_*\| = \|\mathbf{x}_k - \mathbf{x}_* - \mathbf{J}_f(\mathbf{x}_k)^{-1} \mathbf{f}(\mathbf{x}_k)\| \\ &= \|\mathbf{J}_f(\mathbf{x}_k)^{-1} (\mathbf{f}(\mathbf{x}_k) - \mathbf{J}_f(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}_*))\| \leq \|\mathbf{J}_f(\mathbf{x}_k)^{-1}\| \|\mathbf{f}(\mathbf{x}_k) - \mathbf{J}_f(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}_*)\| \\ &\leq \|\mathbf{J}_f(\mathbf{x}_k)^{-1}\| \left(\|\mathbf{f}(\mathbf{x}_k) - \mathbf{J}_f(\mathbf{x}_*)(\mathbf{x}_k - \mathbf{x}_*)\| + \|\mathbf{J}_f(\mathbf{x}_*) - \mathbf{J}_f(\mathbf{x}_k)\| \|\mathbf{x}_k - \mathbf{x}_*\| \right) \\ &\leq \frac{3\alpha}{2} \|\mathbf{J}_f(\mathbf{x}_k)^{-1}\| \|\mathbf{x}_k - \mathbf{x}_*\|^2 \leq 3\alpha \|\mathbf{J}_f(\mathbf{x}_*)\| \|\mathbf{x}_k - \mathbf{x}_*\|^2, \end{aligned}$$

which concludes the proof. \square

3.4.2 The secant method

The Newton–Raphson method exhibits very fast convergence, but it requires the knowledge of the derivatives of the function \mathbf{f} . To conclude this chapter, we describe a root-finding algorithm, known as the secant method, that enjoys superlinear convergence but does not require the derivatives of \mathbf{f} . This method applies only when \mathbf{f} is a function from \mathbf{R} to \mathbf{R} , and so we drop the vector notation in the rest of this section.

Unlike the other methods presented so far in Section 3.2, the secant method *can not* be recast as a fixed point iteration of the form $x_{k+1} = F(x_k)$. Instead, it is of the more general form $x_{k+2} = F(x_k, x_{k+1})$. The geometric intuition behind the method is the following: given x_k and x_{k+1} , the function f is approximated by the unique linear function that passes through $(x_k, f(x_k))$ and $(x_{k+1}, f(x_{k+1}))$, and the iterate x_{k+2} is defined as the root of this linear function. In other words, f is approximated as follows:

$$\tilde{f}(x) = f(x_k) + \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k} (x - x_k).$$

Solving $\tilde{f}(x) = 0$ gives the following expression for x_{k+2} :

$$x_{k+2} = \frac{f(x_{k+1})x_k - f(x_k)x_{k+1}}{f(x_{k+1}) - f(x_k)}, \quad (3.17)$$

Showing the convergence of the secant method rigorously under general assumptions is tedious, so in this course we restrict our attention to the case where f is a quadratic function. Extending

the proof of convergence to a more general smooth function can be achieved by using a quadratic Taylor approximation of f around the root x_* , which is accurate in a close neighborhood of x_* .

Theorem 3.9 (Convergence of the secant method). *Assume that f is a convex quadratic polynomial with a simple root at x_* and that the secant method converges: $\lim_{k \rightarrow \infty} x_k = x_*$. Then the order of convergence is given by the golden ratio*

$$\varphi = \frac{1 + \sqrt{5}}{2}.$$

More precisely, there exists a positive real number y_∞ such that

$$\lim_{k \rightarrow \infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|^\varphi} = y_\infty. \quad (3.18)$$

Proof. Equation (3.17) implies that

$$x_{k+2} - x_* = \frac{f(x_{k+1})(x_k - x_*) - f(x_k)(x_{k+1} - x_*)}{f(x_{k+1}) - f(x_k)}.$$

By assumption, the function f may be expressed as

$$f(x) = \lambda(x - x_*) + \mu(x - x_*)^2, \quad \lambda \neq 0.$$

Substituting this expression in (3.4.2) and letting $e_k = x_k - x_*$, we obtain

$$e_{k+2} = \frac{\mu e_k e_{k+1} (e_{k+1} - e_k)}{\lambda(e_{k+1} - e_k) + \mu(e_{k+1}^2 - e_k^2)} = \frac{\mu e_k e_{k+1}}{\lambda + \mu(e_{k+1} + e_k)}.$$

Rearranging this equation, we have

$$\frac{e_{k+2}}{e_{k+1}} = \frac{\mu e_k}{\lambda + \mu(e_{k+1} + e_k)}. \quad (3.19)$$

By assumption, the right-hand side converges to zero, and so the left-hand side must also converge to zero; the convergence is superlinear. In order to find the order of convergence, we take absolute values in the previous equation to obtain, after rearranging,

$$\frac{|e_{k+2}|}{|e_{k+1}|^\varphi} = \left(\frac{|e_{k+1}|}{|e_k|^{\frac{1}{\varphi-1}}} \right)^{1-\varphi} \frac{\mu}{|\lambda + \mu(e_{k+1} + e_k)|} = \left(\frac{|e_{k+1}|}{|e_k|^\varphi} \right)^{1-\varphi} \frac{|\mu|}{|\lambda + \mu(e_{k+1} + e_k)|},$$

where we used that $\varphi = \frac{1}{\varphi-1}$, since φ is a root of the equation $\varphi^2 - \varphi - 1 = 0$. Thus, introducing the ratio $y_k = |e_{k+1}|/|e_k|^\varphi$, we have

$$y_{k+1} = y_k^{1-\varphi} \frac{|\mu|}{|\lambda + \mu(e_{k+1} + e_k)|}.$$

Taking logarithms in this equation, we deduce

$$\log(y_{k+1}) = (1 - \varphi) \log(y_k) + c_k, \quad c_k := \log \left(\frac{|\mu|}{|\lambda + \mu(e_{k+1} + e_k)|} \right).$$

This is a recurrence equation for $\log(y_k)$, whose explicit solution can be obtained from the variation-of-constants formula:

$$\log(y_k) = (1 - \varphi)^{k-1} \log(y_1) + \sum_{i=1}^{k-1} (1 - \varphi)^{k-1-i} c_i.$$

Since $(c_k)_{k \geq 0}$ converges to the constant $c_\infty = |\mu/\lambda|$ by the assumption that $e_k \rightarrow 0$, the sequence $(\log(y_k))_{k \geq 0}$ converges to c_∞/φ (prove this!). Therefore, by continuity of the exponential function, it holds that

$$y_k = \exp(\log(y_k)) \xrightarrow{k \rightarrow \infty} \exp \left(\frac{c_\infty}{\varphi} \right),$$

and so we deduce (3.18). □

3.5 Exercises

□ **Exercise 3.1.** Implement the bisection method for finding the solution(s) to the equation

$$x = \cos(x).$$

⚙️ **Exercise 3.2.** Find a discrete-time dynamical system over \mathbf{R} of the form

$$x_{k+1} = F(x_k)$$

for which 0 is an attractor but is not stable.

Hint: Use a function F that is discontinuous.

⚙️ **Exercise 3.3.** Show that if \mathbf{x}_* is a globally exponentially stable fixed point of F , then F does not have any other fixed point: \mathbf{x}_* is the unique fixed point.

⚙️ **Exercise 3.4.** Prove Theorem 3.3.

⚙️ **Exercise 3.5.** Let \mathbf{x}_* be a fixed point of (3.5). Show that if

$$\rho(\mathbf{J}_F(\mathbf{x}_*)) < 1,$$

then \mathbf{x}_* is locally exponentially stable. It is sufficient by Proposition 3.5 to find a subordinate matrix norm such that $\|\mathbf{J}_F(\mathbf{x}_*)\| < 1$. In other words, this exercise amounts to showing that for any matrix $\mathbf{A} \in \mathbf{R}^{n \times n}$ with $\rho(\mathbf{A}) < 1$, there exists a matrix norm such that $\|\mathbf{A}\| < 1$.

Hint: One may employ a matrix norm of the form $\|\mathbf{A}\|_{\mathbf{T}} := \|\mathbf{T}^{-1}\mathbf{A}\mathbf{T}\|_2$, which is a subordinate norm by Exercise 2.10. The Jordan normal form is useful for constructing the matrix \mathbf{T} , and equation (2.19) is also useful.

Solution. Let $J = P^{-1}AP$ denote the Jordan normal form of A , and let

$$E_\varepsilon = \begin{pmatrix} \varepsilon & & & \\ & \varepsilon^2 & & \\ & & \ddots & \\ & & & \varepsilon^n \end{pmatrix}$$

By Eq. (2.19), the matrix $J_\varepsilon := E_\varepsilon^{-1}JE_\varepsilon$ coincides with J , except that the first superdiagonal is multiplied by ε . Let D denote the diagonal part of J_ε . We have that

$$\|J_\varepsilon - D\|_2 = \sqrt{\lambda_{\max}(E_\varepsilon^T E_\varepsilon)}.$$

The matrix $E_\varepsilon^T E_\varepsilon$ is diagonal with entries equal to either 0 or ε^2 , and so $\|J_\varepsilon - D\|_2 < \varepsilon$. By the triangle inequality, we have

$$\|J_\varepsilon\| \leq \|D\| + \|J_\varepsilon - D\|_2 \leq \rho(A) + \varepsilon. \quad (3.20)$$

Let $\|A\|_\varepsilon := \|E_\varepsilon^{-1}P^{-1}APE_\varepsilon\|$. By (2.10) with $T = PE_\varepsilon$, this is indeed a subordinate matrix norm. By (3.20) and the assumption that $\rho(A) < 1$, it is clear that $\|A\|_\varepsilon < 1$ provided that ε is sufficiently small.

⚙ **Exercise 3.6.** Calculate $x = \sqrt[3]{3 + \sqrt[3]{3 + \sqrt[3]{3 + \sqrt{\dots}}}}$ using the bisection method.

⚙ **Exercise 3.7.** Solve the equation $f(x) = e^x - 2 = 0$ using a fixed point iteration of the form

$$x_{k+1} = F(x_k), \quad F(x) = x - \alpha^{-1}f(x).$$

Using your knowledge of the exact solution $x_* = \log 2$, write a sufficient condition on α to guarantee that x_* is locally exponentially stable. Verify your findings numerically and plot, using a logarithmic scale for the y axis, the error in absolute value as a function of k .

⚙ **Exercise 3.8.** Implement the Newton–Raphson method for solving $f(x) = e^x - 2 = 0$, and plot the error in absolute value as a function of the iteration index k .

⚙ **Exercise 3.9.** Find the point (x, y) on the parabola $y = x^2$ that is closest to the point $(3, 1)$.

⚙ **Exercise 3.10.** Consider the linear system

$$\begin{cases} y = (x - 1)^2 \\ x^2 + y^2 = 4 \end{cases}$$

By drawing these two constraints in the xy plane, find an approximation of the solution(s). Then calculate the solution(s) using a fixed-point method.

⚙ **Exercise 3.11.** Find solutions (ψ, λ) , with $\lambda > 0$, to the following eigenvalue problem:

$$\psi'' = -\lambda^2\psi, \quad \psi(0) = 0, \quad \psi'(1) = \psi(1).$$

⚙️ **Exercise 3.12.** Suppose that we have n data points (x_i, y_i) of an unknown function $y = f(x)$. We wish to approximate f by a function of the form

$$\tilde{f}(x) = \frac{a}{b+x}$$

by minimizing the sum of squares

$$\sum_{i=1}^n |\tilde{f}(x_i) - y_i|^2.$$

Write a system of nonlinear equations that the minimizer (a, b) must satisfy, and solve this system using the Newton–Raphson method starting from $(1, 1)$. The data is given below:

```
x = [0.0; 0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9; 1.0]
y = [0.6761488864859304; 0.6345697680852508; 0.6396283580587062; 0.6132010027973919;
     0.5906142598705267; 0.5718728461471725; 0.5524549902830562; 0.538938885654085;
     0.5373495476994958; 0.514904589752926; 0.49243437874655027]
```

Plot the data points together with the function \tilde{f} over the interval $[0, 1]$. Your plot should look like [Figure 3.1](#).

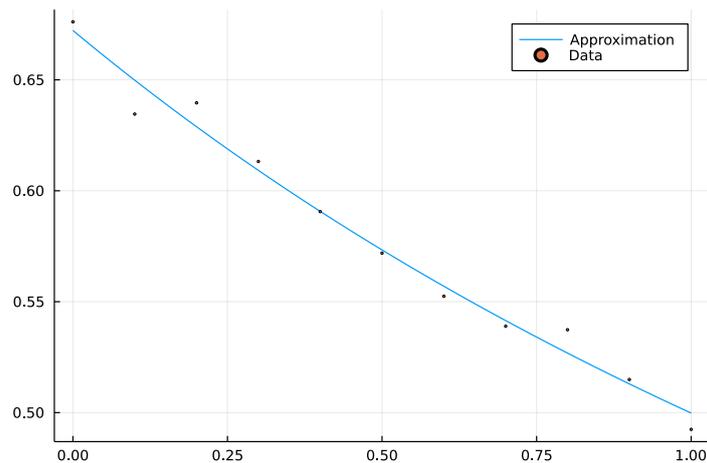


Figure 3.1: Solution to [Exercise 3.12](#).

3.6 Discussion and bibliography

The content of this chapter is largely based on the lecture notes [13]. Several of the exercises are taken or inspired from [6]. The proof of convergence of the secant method is inspired from the general proof presented in the short paper [14]. For a detailed treatment of iterative methods for nonlinear equations, see the book [8].