Chapter 5

Solution of nonlinear systems

5.1 The bisection method 131
5.2 Fixed point methods 132
5.3 Convergence of fixed point methods 133
5.4 Examples of fixed point methods 137
5.4.1 The chord method 137
5.4.2 The Newton–Raphson method 137
5.4.3 The secant method 143
5.5 A numerical experiment 145
5.6 Exercises 147
5.7 Discussion and bibliography 150

Introduction

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

$$\boldsymbol{f}(\boldsymbol{x}) = 0, \qquad \boldsymbol{f}\colon \mathbf{R}^n \to \mathbf{R}^n. \tag{5.1}$$

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 (5.1) in a finite number of operations, so iterative methods are required.

In contrast with the previous chapter, it may not be the case that (5.1) 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 $(\boldsymbol{x}_k)_{k \geqslant 0}$ of a root $\boldsymbol{x}_*$, we define the error as $\boldsymbol{e}_k = \boldsymbol{x}_k - \boldsymbol{x}_*$. If the sequence $(\boldsymbol{x}_k)_{k \geqslant 0}$ converges to $\boldsymbol{x}_*$ in the limit as $k \to \infty$ and if

$$\lim_{k\to\infty} \frac{\|\boldsymbol{e}_{k+1}\|}{\|\boldsymbol{e}_k\|^q} = r, \tag{5.2}$$

then we say that $(\boldsymbol{x}_k)_{k\geqslant 0}$ 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

$$\lim_{k\to\infty} \frac{\|\boldsymbol{e}_{k+1}\|}{\|\boldsymbol{e}_k\|} = 0. \tag{5.3}$$

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

Remark 5.1. The notion of order of convergence may be defined also when the limit in (5.2) does not exist. A more general definition for the order of convergence of a sequence $(\boldsymbol{x}_k)_{k\geqslant 0}$ converging to $\boldsymbol{x}_*$ is the following:

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

or $q(\boldsymbol{x}_0) = \infty$ 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 $\boldsymbol{x}_*$:

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

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 5.1, by way of introduction to the subject of numerical methods for nonlinear equations, we present and analyze the bisection method.
- In Section 5.2, we present a general method based on a fixed point iteration for solving (5.1). The convergence of this method is analyzed in Section 5.3.
- In Section 5.4, two concrete examples of fixed point methods are studied: the chord method and the Newton–Raphson method.

5.1 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 $f\colon \mathbf{R} \to \mathbf{R}$, 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 it 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 $f(x_{1/2})f(a) \geqslant 0$, then $f(x_{1/2})f(b) \leqslant 0$ 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.

Algorithm 7 Bisection method

```
Assume that f(a)f(b) < 0 with a < b.
Pick ε > 0.
x ← a/2 + b/2
while |b − a| ⩾ ε do
    if f(x)f(a) ⩾ 0 then
        a ← x
    else
        b ← x
    end if
    x ← a/2 + b/2
end while
```

The following result establishes the convergence of the method.

Proposition 5.1. *Assume that $f\colon \mathbf{R} \to \mathbf{R}$ is a continuous function and $f(a)f(b) < 0$. Let $[a_j, b_j]$ denote the interval obtained after j iterations of the bisection method, and let $x_j = (a_j + b_j)/2$ denote the midpoint of the interval, Then there exists a root x_* of f such that*

$$|x_j - x_*| \leqslant (b_0 - a_0)2^{-(j+1)}. \tag{5.4}$$

Proof. By construction, $f(a_j)f(b_j) \leqslant 0$ and $f(b) \neq 0$. Therefore, by the intermediate value theorem, there exists a root of f in the interval $[a_j, b_j)$, implying that

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

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

Although the limit in (5.2) 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 $(\tilde{e}_j)_{j \geqslant 0}$, where $\tilde{e}_j = (b_0 - a_0)2^{-(j+1)}$ by Proposition 5.1. 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.

5.2 Fixed point methods

Let $\boldsymbol{x}_*$ denote a zero of the function $\boldsymbol{f}$. The idea of iterative methods for (5.1) is to construct, starting from an initial guess $\boldsymbol{x}_0$, a sequence $(\boldsymbol{x}_k)_{k=0,1,\dots}$ approaching $\boldsymbol{x}_*$. A number of iterative methods for solving (5.1) are based on an iteration of the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{F}(\boldsymbol{x}_k), \tag{5.5}$$

for an appropriate continuous function F. Assume that $\boldsymbol{x}_k$ converges to some point $\boldsymbol{x}_* \in \mathbf{R}^n$ in the limit as $k \to \infty$. Then, taking the limit $k \to \infty$ in (5.5), we find that $\boldsymbol{x}_*$ satisfies

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

Such a point $\boldsymbol{x}_*$ is called a *fixed point* of the function $\boldsymbol{F}$. Several definitions of the function $\boldsymbol{F}$ can be employed in order to ensure that a fixed point of $\boldsymbol{F}$ coincides with a zero of $\boldsymbol{f}$. One may, for example, define $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x} - \alpha^{-1}\boldsymbol{f}(\boldsymbol{x})$, for some nonzero scalar coefficient α. Then $\boldsymbol{F}(\boldsymbol{x}_*) = \boldsymbol{x}_*$ if and only if $\boldsymbol{f}(\boldsymbol{x}_*) = 0$. Later in this chapter, in Section 5.4, we study two instances of numerical methods which can be recast in the form (5.5). Before this, we study the convergence of the iteration (5.5) for a general function $\boldsymbol{F}$.

5.3 Convergence of fixed point methods

Equation (5.5) may be viewed as a *discrete-time* dynamical system. In order to study the behavior of the system as $k \to \infty$, 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 $\boldsymbol{x} \in \mathbf{R}^n$:

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

Definition 5.1 (Stability of fixed points)**.** Let $(\boldsymbol{x}_k)_{k\geqslant 0}$ denote iterates obtained from (5.5) when starting from an initial vector $\boldsymbol{x}_0$. Then we say that a fixed point $\boldsymbol{x}_*$ is

- an *attractor* if there exists a neighborhood $\mathcal{V}$ of $\boldsymbol{x}_*$ such that

$$\forall \boldsymbol{x}_0 \in \mathcal{V}, \qquad \boldsymbol{x}_k \xrightarrow[k\to\infty]{} \boldsymbol{x}_*. \tag{5.6}$$

  The largest neighborhood for which this is true, i.e. the set of values of $\boldsymbol{x}_0$ such that (5.6) holds true, is called the basin of attraction of $\boldsymbol{x}_*$.

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

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

- 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

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

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

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

- unstable if it is not stable.

Clearly, global exponential stability implies exponential stability, which itself implies asymptotic stability and stability. If $\boldsymbol{x}_*$ is globally exponentially stable, then $\boldsymbol{x}_*$ is the unique fixed point of $\boldsymbol{F}$; showing this is the aim of Exercise 5.3. If $\boldsymbol{x}_*$ is an attractor, then the dynamical system (5.5) is said to be locally convergent to $\boldsymbol{x}_*$. The larger the basin of attraction of $\boldsymbol{x}_*$, the less careful we need to be when picking the initial guess $\boldsymbol{x}_0$. Global exponential stability of a fixed point can sometimes be shown provided that $\boldsymbol{F}$ satisfies a strong hypothesis.

Definition 5.2 (Lipschitz continuity)**.** A function $\boldsymbol{F}\colon \mathbf{R}^n \to \mathbf{R}^n$ is said to be *Lipschitz* continuous with constant L if

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbf{R}^n \times \mathbf{R}^n, \qquad \big\|\boldsymbol{F}(\boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x})\big\| \leqslant L\|\boldsymbol{y} - \boldsymbol{x}\|.$$

A function $\boldsymbol{F}\colon \mathbf{R}^n \to \mathbf{R}^n$ that is Lipschitz continuous with a constant $L < 1$ is called a *contraction*. For such a function, it is possible to prove that (5.5) has a unique globally exponentially stable fixed point.

Theorem 5.2. *Assume that $\boldsymbol{F}$ is a contraction. Then there exists a unique fixed point of (5.5), and it holds that*

$$\forall \boldsymbol{x}_0 \in \mathbf{R}^n, \quad \forall k \in \mathbf{N}, \qquad \|\boldsymbol{x}_k - \boldsymbol{x}_*\| \leqslant L^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|. \tag{5.7}$$

Proof. Existence and uniqueness of the fixed point follows from the *Banach fixed point theorem*, see Theorem A.3, so here we show only global exponential convergence. Since $\boldsymbol{F}$ is a contraction, it holds that

$$\big\|\boldsymbol{x}_k - \boldsymbol{x}_*\big\| = \big\|\boldsymbol{F}(\boldsymbol{x}_{k-1}) - \boldsymbol{F}(\boldsymbol{x}_*)\big\| \leqslant L\|\boldsymbol{x}_{k-1} - \boldsymbol{x}_*\| \leqslant \ \ldots \ \leqslant L^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|, \tag{5.8}$$

which proves (5.7). □

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

Theorem 5.3. *Assume that $\boldsymbol{x}_*$ is a fixed point of (5.5) and that $\boldsymbol{F}\colon \mathbf{R}^n \to \mathbf{R}^n$ satisfies the local Lipschitz condition*

$$\forall \boldsymbol{x} \in B_\delta(\boldsymbol{x}_*), \qquad \big\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}_*)\big\| \leqslant L\|\boldsymbol{x} - \boldsymbol{x}_*\|, \tag{5.9}$$

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

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

Proof. See Exercise 5.4. □

It is possible to guarantee that condition (5.9) holds provided that we have sufficiently good control of the derivatives of the function $\boldsymbol{F}$. The function $\boldsymbol{F}$ is said to be differentiable at $\boldsymbol{x}$ (in the sense of Fréchet) if there exists a linear operator $\mathrm{D}\boldsymbol{F}_{\boldsymbol{x}}\colon \mathbf{R}^n \to \mathbf{R}^n$ such that

$$\lim_{\boldsymbol{h}\to 0} \frac{\|\boldsymbol{F}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{F}(\boldsymbol{x}) - \mathrm{D}\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h})\|}{\|\boldsymbol{h}\|} = 0. \tag{5.10}$$

If $\boldsymbol{F}$ is differentiable, then all its first partial derivatives $\partial_j F_i$ exist and, in addition, it holds that $\mathrm{D}\boldsymbol{F}_{\boldsymbol{x}}(\boldsymbol{h}) = \mathsf{J}_F(\boldsymbol{x})\boldsymbol{h}$ where $\mathsf{J}_F(\boldsymbol{x})$ is the Jacobian matrix of $\boldsymbol{F}$ at $\boldsymbol{x}$:

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

Proposition 5.4. *Let $\boldsymbol{x}_*$ be a fixed point of (5.5), and assume that there exists δ and a subordinate matrix norm such that $\boldsymbol{F}$ is differentiable everywhere in $B_\delta(\boldsymbol{x}_*)$ and*

$$\forall \boldsymbol{x} \in B_\delta(\boldsymbol{x}_*), \qquad \|\mathsf{J}_F(\boldsymbol{x})\| \leqslant L < 1.$$

Then condition (5.9) is satisfied in the associated vector norm, and so the fixed point $\boldsymbol{x}_*$ is locally exponentially stable.

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

$$\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}_*) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\Big(\boldsymbol{F}\big(\boldsymbol{x}_* + t(\boldsymbol{x}-\boldsymbol{x}_*)\big)\Big)\mathrm{d}t = \int_0^1 \mathsf{J}_F\big(\boldsymbol{x}_* + t(\boldsymbol{x}-\boldsymbol{x}_*)\big)(\boldsymbol{x}-\boldsymbol{x}_*)\mathrm{d}t.$$

Therefore, it holds that

$$\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}_*)\| \leqslant \int_0^1 \Big\|\mathsf{J}_F\big(\boldsymbol{x} + t(\boldsymbol{x}-\boldsymbol{x}_*)\big)\Big\|\,\mathrm{d}t\,\|\boldsymbol{x}-\boldsymbol{x}_*\| \leqslant \int_0^1 L\,\mathrm{d}t\,\|\boldsymbol{x}-\boldsymbol{x}_*\| = L\|\boldsymbol{x}-\boldsymbol{x}_*\|,$$

which is the statement. □

Remark 5.2. As a student observed during the lecture, in dimension $n = 1$, Proposition 5.4 can be proved by using the mean value theorem: since F is differentiable in $(x_* - \delta, x_* + \delta)$, there exists for all x in this interval a $\xi = \xi(x)$ also in this interval such that

$$F(x) - F(x_*) = F'(\xi)(x - x_*).$$

It then follows immediately that

$$\big|F(x) - F(x_*)\big| = |F'(\xi)(x-x_*)| \leqslant L\,|x - x_*|.$$

This proof does not carry over to the multi-dimensional setting, however.

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

Proposition 5.5. *Let $\boldsymbol{x}_*$ be a fixed point of (5.5) and that F is differentiable at $\boldsymbol{x}_*$ with*

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

in a subordinate vector norm. Then the fixed point $\boldsymbol{x}_*$ is locally exponentially stable.

Proof. In this proof, the vector norm used is that associated with the matrix norm in the statement of the proposition. By the definition of differentiability (5.10), there exists for all $\varepsilon > 0$ a $\delta > 0$ such that

$$\forall \boldsymbol{x} \in B_\delta(\boldsymbol{x}_*)\backslash\{\boldsymbol{x}_*\}, \qquad \frac{\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}_*) - \mathsf{J}_F(\boldsymbol{x}_*)(\boldsymbol{x} - \boldsymbol{x}_*)\|}{\|\boldsymbol{x} - \boldsymbol{x}_*\|} \leqslant \varepsilon.$$

By the triangle inequality, this implies that for all $\boldsymbol{x} \in B_\delta(\boldsymbol{x}_*)$,

$$\begin{aligned}\|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}_*)\| &\leqslant \|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}_*) - \mathsf{J}_F(\boldsymbol{x}_*)(\boldsymbol{x} - \boldsymbol{x}_*)\| + \|\mathsf{J}_F(\boldsymbol{x}_*)(\boldsymbol{x} - \boldsymbol{x}_*)\| \\ &\leqslant \varepsilon\|\boldsymbol{x} - \boldsymbol{x}_*\| + \|\mathsf{J}_F(\boldsymbol{x}_*)\|\|(\boldsymbol{x} - \boldsymbol{x}_*)\| = (L + \varepsilon)\|\boldsymbol{x} - \boldsymbol{x}_*\|.\end{aligned}$$

We have thus shown that for all $\varepsilon > 0$, there exists $\delta > 0$ such that condition (5.9) is satisfied with constant $L + \varepsilon$. By taking ε sufficiently small, we can ensure that $L + \varepsilon < 1$, and so the fixed point $\boldsymbol{x}_*$ is locally exponentially stable by Theorem 5.3. □

The estimate in Theorem 5.2 suggests that when the fixed point iteration (5.5) converges, the convergence is linear. While this is usually the case, the convergence is superlinear if $\mathsf{J}_F(\boldsymbol{x}_*) = 0$.

Proposition 5.6. *Assume that $\boldsymbol{x}_*$ is a fixed point of (5.5) and that $\mathsf{J}_F(\boldsymbol{x}_*) = 0$. Then the convergence to $\boldsymbol{x}_*$ is superlinear, in the sense that if $\boldsymbol{x}_k \to \boldsymbol{x}_*$ as $k \to \infty$, then*

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

Proof. By Proposition 5.5, there exists $\delta > 0$ such that $(\boldsymbol{x}_k)_{k\geqslant 0}$ is a sequence converging to $\boldsymbol{x}_*$ for all $\boldsymbol{x}_0 \in B_\delta(\boldsymbol{x}_*)$. It holds that

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

Since $\boldsymbol{x}_k - \boldsymbol{x}_* \to \boldsymbol{0}$ as $k \to \infty$, the right-hand side converges to 0 by (5.10). □

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

$$\forall \boldsymbol{x} \in B_\delta(\boldsymbol{x}_*), \qquad \|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}_*)\| \leqslant C\|\boldsymbol{x} - \boldsymbol{x}_*\|^q, \tag{5.11}$$

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

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

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

5.4 Examples of fixed point methods

As we mentioned in Section 5.2, there are several choices for the function $\boldsymbol{F}$ that guarantee the equivalence $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x} \Leftrightarrow \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$.

5.4.1 The chord method

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 (5.4) 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.

$$x_{k+1} = x_k - \alpha^{-1} f(x_k) = F(x_k). \tag{5.12}$$

This is illustrated in Figure 5.1. By Proposition 5.5, a sufficient condition to ensure local convergence is that

$$|F'(x_*)| = |1 - \alpha^{-1} f'(x_*)| < 1. \tag{5.13}$$

In order for this condition to hold true, the slope α must be of the same sign as $f'(x_*)$ and the inequality $|\alpha| \geqslant |f'(x_*)|/2$ must be satisfied. If $f'(x_*) = 0$, then the sufficient condition (5.13) is never satisfied; in this case, the convergence must be studied on a case-by-case basis. By Proposition 5.6, 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 $k \to \infty$, fast convergence can be obtained. This is precisely what the Newton–Raphson method aims to achieve; see Section 5.4.2

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

$$x_{k+1} = \boldsymbol{F}(\boldsymbol{x}_k), \qquad \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x} - \mathsf{A}^{-1}\boldsymbol{f}(\boldsymbol{x}),$$

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

5.4.2 The Newton–Raphson method

Let us first consider the case of a function from $\mathbf{R}$ to $\mathbf{R}$. A necessary condition for the Newton–Raphson method to apply is that f is differentiable. At each step, the function f is

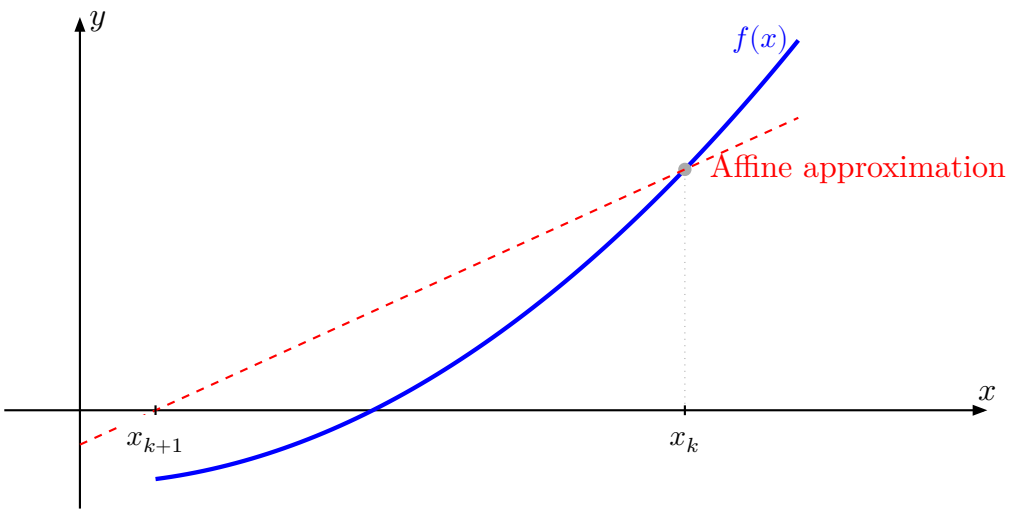

Figure 5.1: Graphical illustration of an iteration of the chord method.

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

$$x_{k+1} = x_k - f'(x_k)^{-1} f(x_k). \tag{5.14}$$

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 (5.12) where the slope α is adapted as the simulation progresses. If the method converges and f' is continuous, then $f'(x_k) \to f'(x_*)$ in the limit as $k \to \infty$, which is an indication that superlinear convergence could occur in view of our discussion in the previous section. Equation (5.14) may be recast as a fixed point iteration of the form (5.4) with

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

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 the function F. If the function f is twice continuously differentiable, then the convergence of the Newton–Raphson method is superlinear by Proposition 5.6, because then

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

The geometric interpretation of the Newton–Raphson method in dimension 1 is the following: at each step, the function $\boldsymbol{f}$ is approximated by the affine function $x \mapsto x_k + f'(x_k)(x - x_k)$, which is *the tangent line to* f *at* x_k, and the next iterate is given by the unique zero of the latter function. This is illustrated in Figure 5.2.

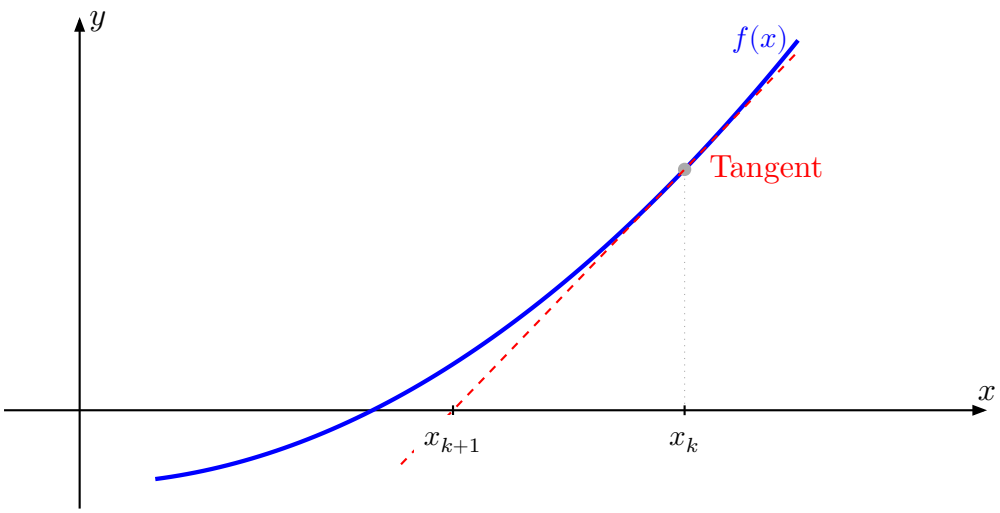

Figure 5.2: Graphical illustration of a Newton–Raphson iteration. The code used to create this figure is based on the answer https://tex.stackexchange.com/a/551205/125558 on LaTeX stack exchange.

The Newton–Raphson method may be generalized to nonlinear equations in $\mathbf{R}^n$ of the form (5.1). In this case $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x} - \mathsf{J}_f(\boldsymbol{x})^{-1}\boldsymbol{f}(\boldsymbol{x})$, and so an iteration of the method reads

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \mathsf{J}_f(\boldsymbol{x}_k)^{-1}\boldsymbol{f}(\boldsymbol{x}_k). \tag{5.15}$$

In the rest of this section, we show that the iteration (5.15) is well-defined in a small neighborhood of a root of $\boldsymbol{f}$ under appropriate assumptions, and we demonstrate the *second order* convergence of the method, first in dimension 1 under simplifying assumption involving the second derivative of f, and then in the multi-dimensional setting under more general assumptions.

Convergence in the one-dimensional setting

We assume in this section that $(x_k)_{k\geqslant 0}$ is generated from the Newton–Raphson method (5.14) and prove the following result.

Theorem 5.7 (Quadratic convergence of Newton–Raphson)**.** *Assume that $f \in C^2(\mathbf{R})$ and that the following assumptions are satisfied:*

- *The first derivative of f is uniformly bounded away from zero:*

$$\inf_{x\in\,\mathbf{R}} |f'(x)| = m > 0.$$

- *The second derivative of f is uniformly bounded from above in absolute value:*

$$\sup_{x\in\,\mathbf{R}} |f''(x)| = M < \infty.$$

Then $f(x)$ has a unique root x_* and it holds for all initial $x_0 \in \mathbf{R}$ that

$$\forall k \in \mathbf{N}, \qquad |x_{k+1} - x_*| \leqslant \frac{M}{2m}|x_k - x_*|^2. \tag{5.16}$$

Proof. By assumption, the function f is continuous and either strictly increasing everywhere or strictly decreasing everywhere. Therefore there exists a unique root $x_* \in \mathbf{R}$ of f. In order to prove (5.16), we note that

$$x_{k+1} - x_* = x_k - \frac{f(x_k)}{f'(x_k)} - x_* = \frac{1}{f'(x_k)}\Big(f'(x_k)(x_k - x_*) - f(x_k)\Big). \tag{5.17}$$

By Taylor's theorem, there is $\xi \in \mathbf{R}$ such that

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

Since x_* is a root of f, the left-hand side of this equation is zero. Combining this equation with (5.17), we deduce that

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

Taking absolute values and using the assumptions gives

$$|x_{k+1} - x_*| \leqslant \frac{M}{2m}(x_k - x_*)^2,$$

which concludes the proof. □

Remark 5.3. As a corollary of Theorem 5.7, we obtain that the Newton–Raphson method is convergent if

$$|x_k - x_*| \leqslant \frac{2m}{M}.$$

Convergence in the multi-dimensional setting

As a first step towards a proof of quadratic convergence for the Newton–Raphson method in the multi-dimensional setting, we begin by proving the following preparatory lemma, which we will then employ in the particular case where the matrix-valued function A is equal to J_f.

Lemma 5.8. *Let* $\mathsf{A}\colon \mathbf{R}^n \to \mathbf{R}^{n\times n}$ *denote a matrix-valued function on* $\mathbf{R}^n$ *that is both continuous and nonsingular at* $\boldsymbol{x}_*$*, and let* $\boldsymbol{f}$ *be a function that is differentiable at* $\boldsymbol{x}_*$ *where* $\boldsymbol{f}(\boldsymbol{x}_*) = 0$*. Then the function*

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

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

$$\mathsf{J}_G(\boldsymbol{x}_*) = \mathsf{I} - \mathsf{A}(\boldsymbol{x}_*)^{-1}\mathsf{J}_f(\boldsymbol{x}_*). \tag{5.18}$$

Proof. It holds that

$$\mathsf{A}(\boldsymbol{x}) = \Big(\mathsf{A}(\boldsymbol{x}_*) - \big(\mathsf{A}(\boldsymbol{x}_*) - \mathsf{A}(\boldsymbol{x})\big)\Big) = \mathsf{A}(\boldsymbol{x}_*)\Big(\mathsf{I} - \mathsf{A}(\boldsymbol{x}_*)^{-1}\big(\mathsf{A}(\boldsymbol{x}_*) - \mathsf{A}(\boldsymbol{x})\big)\Big). \tag{5.19}$$

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

$$\forall \boldsymbol{x} \in B_\delta(\boldsymbol{x}_*), \qquad \|\mathsf{A}(\boldsymbol{x}) - \mathsf{A}(\boldsymbol{x}_*)\| \leqslant \varepsilon.$$

For $\boldsymbol{x} \in B_\delta(\boldsymbol{x}_*)$ we have $\|\mathsf{A}(\boldsymbol{x}_*)^{-1}\big(\mathsf{A}(\boldsymbol{x}_*) - \mathsf{A}(\boldsymbol{x})\big)\| \leqslant \|\mathsf{A}(\boldsymbol{x}_*)^{-1}\|\|\mathsf{A}(\boldsymbol{x}_*) - \mathsf{A}(\boldsymbol{x})\| \leqslant \beta\varepsilon = \frac{1}{2}$, and so Lemma 4.2 implies that the second factor on the right-hand side of (5.19) is invertible with a norm bounded from above by 2. Therefore, we deduce that $\mathsf{A}(\boldsymbol{x})$ is invertible with

$$\forall \boldsymbol{x} \in B_\delta(\boldsymbol{x}_*), \qquad \|\mathsf{A}(\boldsymbol{x})^{-1}\| \leqslant 2\|\mathsf{A}(\boldsymbol{x}_*)^{-1}\| = 2\beta, \tag{5.20}$$

which shows that $\boldsymbol{G}$ is well-defined in $B_\delta(\boldsymbol{x}_*)$. In order to prove (5.18), we need to show that

$$\lim_{\|\boldsymbol{h}\|\to 0} \frac{\Big\|\boldsymbol{G}(\boldsymbol{x}_* + \boldsymbol{h}) - \boldsymbol{G}(\boldsymbol{x}_*) - \big(\mathsf{I} - \mathsf{A}(\boldsymbol{x}_*)^{-1}\mathsf{J}_f(\boldsymbol{x}_*)\big)\boldsymbol{h}\Big\|}{\|\boldsymbol{h}\|} = 0$$

By definition of $\boldsymbol{G}$, and using the fact that $\boldsymbol{f}(\boldsymbol{x}_*) = \boldsymbol{0}$, we obtain that the argument of the norm in the numerator is equal to

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

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

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

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

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

Theorem 5.9 (Convergence of Newton–Raphson)**.** *Let $\boldsymbol{f}\colon \mathbf{R}^n \to \mathbf{R}^n$ denote a function that is differentiable in a neighborhood $B_\delta(\boldsymbol{x}_*)$ of a point $\boldsymbol{x}_*$ where $\boldsymbol{f}(\boldsymbol{x}_*) = 0$. Assume that the Jacobian matrix $\mathsf{J}_f(\boldsymbol{x})$ is nonsingular and continuous at $\boldsymbol{x}_*$. Then $\boldsymbol{x}_*$ is an attractor of the Newton–Raphson iteration (5.15) and the convergence is superlinear.*

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

$$\forall \boldsymbol{x} \in B_\delta(\boldsymbol{x}_*), \qquad \|\mathsf{J}_f(\boldsymbol{x}) - \mathsf{J}_f(\boldsymbol{x}_*)\| \leqslant \alpha \|\boldsymbol{x} - \boldsymbol{x}_*\|$$

is satisfied, there exists $d \in (0, \delta)$ *and* $C > 0$ *such that*

$$\forall \boldsymbol{x}_k \in B_d(\boldsymbol{x}_*), \qquad \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_*\| \leqslant C \|\boldsymbol{x}_k - \boldsymbol{x}_*\|^2.$$

In other words, the convergence is at least quadratic in $B_d(\boldsymbol{x}_*)$.

Proof. Using Lemma 5.8, we obtain that the Newton–Raphson update

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

is well-defined in a neighborhood $B_\delta(\boldsymbol{x}_*)$ of $\boldsymbol{x}_*$ for sufficiently small δ. In addition, the second statement in Lemma 5.8 gives that $\mathsf{J}_F(\boldsymbol{x}_*)^{-1} = \mathsf{I} - \mathsf{J}_F(\boldsymbol{x}_*)^{-1} \mathsf{J}_F(\boldsymbol{x}_*) = 0$, which establishes the superlinear convergence by Proposition 5.6.

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

$$\boldsymbol{f}(\boldsymbol{x}_k) = \int_0^t \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{f}\big(\boldsymbol{x}_* + t(\boldsymbol{x}_k - \boldsymbol{x}_x)\big) \,\mathrm{d}t = \int_0^t \mathsf{J}_f\big(\boldsymbol{x}_* + t(\boldsymbol{x}_k - \boldsymbol{x}_x)\big)(\boldsymbol{x}_k - \boldsymbol{x}_*) \,\mathrm{d}t,$$

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

$$\begin{aligned}
\|\boldsymbol{f}(\boldsymbol{x}_k) - \mathsf{J}_f(\boldsymbol{x}_*)(\boldsymbol{x}_k - \boldsymbol{x}_*)\| &= \left\| \int_0^1 \Big(\mathsf{J}_f\big(\boldsymbol{x}_* + t(\boldsymbol{x}_k - \boldsymbol{x}_*)\big) - \mathsf{J}_f(\boldsymbol{x}_*) \Big) (\boldsymbol{x}_k - \boldsymbol{x}_*) \,\mathrm{d}t \right\| \\
&\leqslant \int_0^1 \Big\| \mathsf{J}_f\big(\boldsymbol{x}_* + t(\boldsymbol{x}_k - \boldsymbol{x}_*)\big) - \mathsf{J}_f(\boldsymbol{x}_*) \Big\| \|\boldsymbol{x}_k - \boldsymbol{x}_*\| \,\mathrm{d}t \\
&\leqslant \int_0^1 \alpha t \|\boldsymbol{x}_k - \boldsymbol{x}_*\|^2 \,\mathrm{d}t \leqslant \frac{\alpha}{2} \|\boldsymbol{x}_k - \boldsymbol{x}_*\|^2. \qquad (5.21)
\end{aligned}$$

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

$$\forall \boldsymbol{x} \in B_d(\boldsymbol{x}_*), \qquad \|\mathsf{J}_f(\boldsymbol{x})^{-1}\| \leqslant 2 \|\mathsf{J}_f(\boldsymbol{x}_*)^{-1}\|.$$

There exists such a d by (5.20). Using the inequality (5.21), we have that for all $\boldsymbol{x}_k \in B_d(\boldsymbol{x}_*)$,

$$\begin{aligned}
\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_*\| &= \|\boldsymbol{F}(\boldsymbol{x}_k) - \boldsymbol{x}_*\| = \|\boldsymbol{x}_k - \boldsymbol{x}_* - \mathsf{J}_f(\boldsymbol{x}_k)^{-1} \boldsymbol{f}(\boldsymbol{x}_k)\| \\
&= \Big\| \mathsf{J}_f(\boldsymbol{x}_k)^{-1} \big(\boldsymbol{f}(\boldsymbol{x}_k) - \mathsf{J}_f(\boldsymbol{x}_k)(\boldsymbol{x}_k - \boldsymbol{x}_*) \big) \Big\| \leqslant \|\mathsf{J}_f(\boldsymbol{x}_k)^{-1}\| \|\boldsymbol{f}(\boldsymbol{x}_k) - \mathsf{J}_f(\boldsymbol{x}_k)(\boldsymbol{x}_k - \boldsymbol{x}_*)\| \\
&\leqslant \|\mathsf{J}_f(\boldsymbol{x}_k)^{-1}\| \Big(\|\boldsymbol{f}(\boldsymbol{x}_k) - \mathsf{J}_f(\boldsymbol{x}_*)(\boldsymbol{x}_k - \boldsymbol{x}_*)\| + \|\mathsf{J}_f(\boldsymbol{x}_*) - \mathsf{J}_f(\boldsymbol{x}_k)\| \|\boldsymbol{x}_k - \boldsymbol{x}_*\| \Big) \\
&\leqslant \frac{3\alpha}{2} \|\mathsf{J}_f(\boldsymbol{x}_k)^{-1}\| \|\boldsymbol{x}_k - \boldsymbol{x}_*\|^2 \leqslant 3\alpha \|\mathsf{J}_f(\boldsymbol{x}_*)^{-1}\| \|\boldsymbol{x}_k - \boldsymbol{x}_*\|^2,
\end{aligned}$$

which concludes the proof. □

5.4.3 The secant method

The Newton–Raphson method exhibits very fast convergence, but it requires the knowledge of the derivatives of the function $\boldsymbol{f}$. 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 $\boldsymbol{f}$. This method applies only when $\boldsymbol{f}$ 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 5.2, 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 in the following: given x_k and x_{k+1}, the function f is approximated by the unique linear function that passes through $\big(x_k, f(x_k)\big)$ and $\big(x_{k+1}, f(x_{k+1})\big)$, 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}:

$$x_{k+2} = \frac{f(x_{k+1})x_k - f(x_k)x_{k+1}}{f(x_{k+1}) - f(x_k)}, \tag{5.22}$$

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 5.10 (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: $\lim_{k\to\infty} x_k = x_*$. 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

$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|^\varphi} = y_\infty. \tag{5.23}$$

Proof. Equation (5.22) 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

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

Substituting this expression in (5.4.3) 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

$$\frac{e_{k+2}}{e_{k+1}} = \frac{\mu e_k}{\lambda + \mu(e_{k+1} + e_k)}. \tag{5.24}$$

By assumption, the right-hand side converges to zero, and so the left-hand side must also converge to zero; the convergence is superlinear.

To conclude the proof, we first reason formally in order to guess the order convergence, and then give a rigorous proof that our guess is correct. If e_k is small, then it holds approximately by (5.24) that

$$\frac{e_{k+2}}{e_{k+1}} \approx \mu e_k. \tag{5.25}$$

Assume that there exists $q > 0$ such that the equation $e_{k+1} = Ce_k^q$ is valid for all k. Then it holds that $e_{k+2} = Ce_{k+1}^q = C(Ce_k^q)^q$ and (5.25) enables to determine q:

$$\frac{C(Ce_k^q)^q}{Ce_k^q} = \frac{\mu}{\lambda} e_k \qquad \Rightarrow \qquad C^q e_k^{q^2-q} = \frac{\mu}{\lambda} e_k \qquad \Rightarrow \qquad q^2 - q - 1 = 0. \qquad \Rightarrow \qquad q = \varphi.$$

Now comes the rigorous justification. Take absolute values in (5.24) 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

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

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

$$y_k = \exp\big(\log(y_k)\big) \xrightarrow[k\to\infty]{} \exp\left(\frac{c_\infty}{\varphi}\right) = \left|\frac{\mu}{\lambda}\right|^{\frac{1}{\varphi}}$$

and so we deduce (5.23). □

5.5 A numerical experiment

To conclude this chapter, we present the results of a numerical experiment. Specifically, we consider four different fixed point methods for calculating the square root of 2, i.e. for solving the nonlinear equation

$$f(x) := x^2 - 2 = 0. \tag{5.26}$$

The unique positive solution to this equation is $x_* = \sqrt{2}$. The methods we consider are the following:

- The chord method with large $\alpha = 10$.
- The chord method with the optimal parameter α, which is such that $F'(x_*) = 0$. The optimum value for α for solving (5.26) is given by $\alpha_* = 2\sqrt{2}$.
- The Newton–Raphson method, where each iteration is given by

  $$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^2 - 2}{2x_k} = \frac{1}{2}\left(x_k + \frac{2}{x_k}\right) =: F(x_k),$$

  with

  $$F(x) = \frac{1}{2}\left(x + \frac{2}{x}\right).$$

  Notice that $F'(x_*) = 0$ and that $F \in C^2\big((0,\infty)\big)$. Therefore, by Taylor's theorem it holds for all $x \in (x_* - 1, x_* + 1)$ that

  $$|F(x) - F(x_*)| = \left|F''\big(\xi(x)\big)\right| \leqslant L(x - x_*)^2, \qquad L := \sup_{|x - x_*| \leqslant 1} |F''(x)|.$$

  We deduce that the convergence is at least quadratic by (5.11).

Remark 5.4. Note that the ancient Babylonian method coincides with the Newton–Raphson method applied to (5.26).

The following code implements these methods. Note that we use the arbitrary precision `BigFloat` format with a precision we manually set to 2000 bits, which enables using a very small ε in the stopping criterion.

```
function count_digits(x, y)
    xdigits = split(string(x), "")
    ydigits = split(string(y), "")
    len = min(length(xdigits), length(ydigits))
    for i in 1:len
        xdigits[i] != ydigits[i] && return i-2
    end
end

function my_sqrt(a)
    exact = sqrt(a)
    f(x) = x*x - a
    fp(x) = 2x

    # Uncomment desired line
    F(x) = x - f(x)/10      # Chord method
    # F(x) = x - f(x)/(2√a)   # Chord method with optimal α
    # F(x) = 1/2 * (x + a/x)  # Babylonian / Newton Raphson

    r, ε = 1, 1e-200
    while abs(f(r)) > ε
        r = F(r)
        digits = ceil(Int, -log10(abs(r - exact)))
        println("Number of correct digits: $digits")
    end
end

# Sets the precision of BigFloats to 1000 bits
setprecision(2000)
my_sqrt(BigFloat(2))
```

For each of the methods, the number of correct digits of the approximation as the iterations progress is illustrated in Table 5.1. Observe that for all the methods except the first one, the number of correct digits is approximately doubled at each iteration, which is consistent with quadratic convergence.

Method	Chord $\alpha = 10$	Chord $\alpha = 2\sqrt{2}$	Newton–Raphson
# Iterations	**1357**	**8**	**9**
# Correct digits $i = 1$	1	1	1
# Correct digits $i = 2$	1	3	3
# Correct digits $i = 3$	1	6	6
# Correct digits $i = 4$	1	12	12
# Correct digits $i = 5$	1	26	24
# Correct digits $i = 6$	1	53	48
# Correct digits $i = 7$	1	107	97
# Correct digits $i = 8$	1	214	196
# Correct digits $i = 9$	1	n/a	392

Table 5.1: Comparison of different fixed point methods for calculating $\sqrt{2}$. Here i denotes the iteration index.

5.6 Exercises

Exercise 5.1. *Implement the bisection method for finding the solution(s) to the equation*

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

Exercise 5.2. *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 5.3. *Show that if* $\boldsymbol{x}_*$ *is a globally exponentially stable fixed point of* F*, then* F *does not have any other fixed point:* $\boldsymbol{x}_*$ *is the unique fixed point.*

Exercise 5.4. *Prove Theorem 5.3.*

Exercise 5.5. *Let* $\boldsymbol{x}_*$ *be a fixed point of (5.5). Show that if*

$$\rho\big(\mathsf{J}_F(\boldsymbol{x}_*)\big) < 1,$$

then $\boldsymbol{x}_*$ *is locally exponentially stable. It is sufficient by Proposition 5.5 to find a subordinate matrix norm such that* $\|\mathsf{J}_F(\boldsymbol{x}_*)\| < 1$*. In other words, this exercise amounts to showing that for any matrix* $\mathsf{A} \in \mathbf{R}^{n\times n}$ *with* $\rho(\mathsf{A}) < 1$*, there exists a matrix norm such that* $\|\mathsf{A}\| < 1$*.*
Hint: *One may employ a matrix norm of the form* $\|\mathsf{A}\|_{\mathsf{T}} := \|\mathsf{T}^{-1}\mathsf{A}\mathsf{T}\|_2$*, which is a subordinate norm by Exercise 4.10. The Jordan normal form is useful for constructing the matrix* T*, and equation (4.24) is also useful.*

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

$$\mathsf{E}_\varepsilon = \begin{pmatrix} \varepsilon & & & \\ & \varepsilon^2 & & \\ & & \ddots & \\ & & & \varepsilon^n \end{pmatrix}$$

By Eq. (4.24), the matrix $\mathsf{J}_\varepsilon := \mathsf{E}_\varepsilon^{-1}\mathsf{J}\mathsf{E}_\varepsilon$ coincides with J, except that the first superdiagonal is multiplied by ε. Let D denote the diagonal part of J_ε. We have that

$$\|\mathsf{J}_\varepsilon - \mathsf{D}\|_2 = \sqrt{\lambda_{\max}(\mathsf{E}_\varepsilon^T \mathsf{E}_\varepsilon)}.$$

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

$$\|\mathsf{J}_\varepsilon\| \leqslant \|\mathsf{D}\| + \|\mathsf{J}_\varepsilon - \mathsf{D}\|_2 \leqslant \rho(\mathsf{A}) + \varepsilon. \tag{5.27}$$

Let $\|\mathsf{A}\|_\varepsilon := \|\mathsf{E}_\varepsilon^{-1}\mathsf{P}^{-1}\mathsf{A}\mathsf{P}\mathsf{E}_\varepsilon\|$. By Exercise 4.10 with $\mathsf{T} = \mathsf{P}\mathsf{E}_\varepsilon$, this is indeed a subordinate matrix norm. By (5.27) and the assumption that $\rho(\mathsf{A}) < 1$, it is clear that $\|\mathsf{A}\|_\varepsilon < 1$ provided that ε is sufficiently small. △

Remark 5.5. A corollary of Exercise 4.10 is that the spectral radius of a matrix A is *the infimum* of $\|\mathsf{A}\|$ over all subordinate matrix norms.

Exercise 5.6. *Calculate* $x = \sqrt[3]{3 + \sqrt[3]{3 + \sqrt[3]{3 + \sqrt{\dots}}}}$ *using the bisection method.*

Exercise 5.7. *Solve the equation* $f(x) = \mathrm{e}^x - 2 = 0$ *using a fixed point iteration of the form*

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

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

Exercise 5.9. *Find the point* (x, y) *on the parabola* $y = x^2$ *that is closest to the point* $(3, 1)$*.*

Exercise 5.10. *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 5.11. *Find solutions* (ψ, λ)*, with* $\lambda > 0$*, to the following eigenvalue problem:*

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

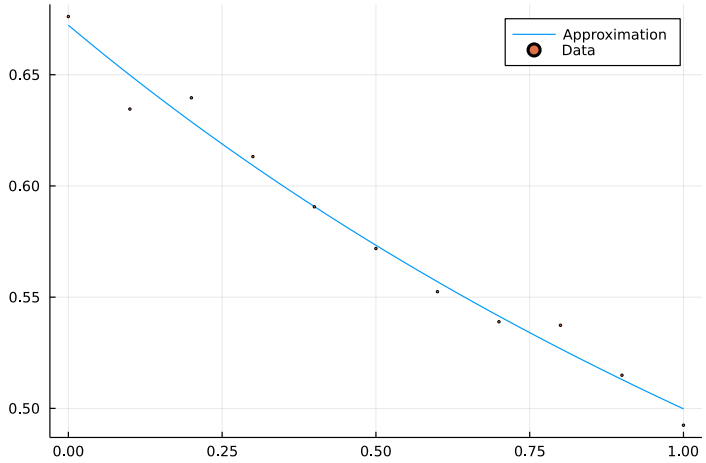

Figure 5.3: Solution to Exercise 5.12.

Exercise 5.12. *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

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

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 5.3.

Exercise 5.13 (Nonlinear least-squares)**.** *Suppose that we are given n data points (x_i, y_i) of an unknown function $y = f(x)$. We wish to approximate f by a straight line*

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

by minimizing the sum of squared Euclidean distances *between the data points and the straight line $\tilde{f}$. Since the distance between a point (x_i, y_i) and the straight line is given by*

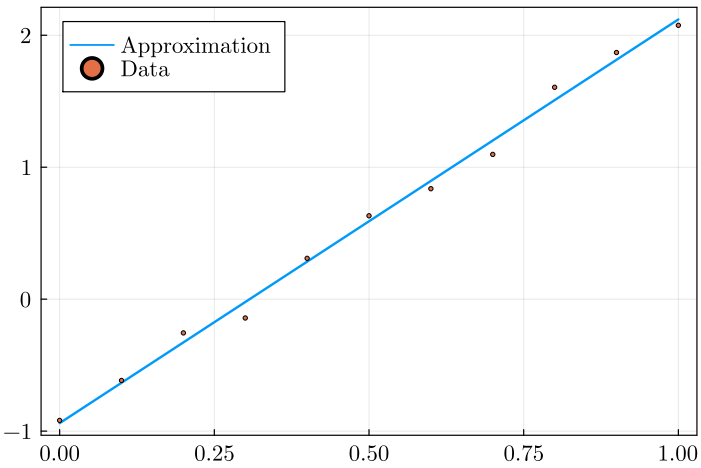

Figure 5.4: Solution to Exercise 5.13.

$$|y_i - ax_i - b\frac{|}{\sqrt{1+a^2}},$$

the objective function to minimize is given by

$$J(a,b) := \sum_{i=1}^{n} \frac{(y_i - ax_i - b)^2}{1+a^2}.$$

This is a smooth function of a and b, and so a necessary condition for a pair $(a_*, b_*) \in \mathbf{R}^2$ to be a minimizer is that

$$\nabla J(a_*, b_*) = \mathbf{0},$$

which is a nonlinear equation for the unknowns a_* and b_*. Solve this equation by using the Newton–Raphson method initialized at $(1,1)$, and then plot the data points together with the function $\tilde{f}$ over the interval $[0,1]$. Your plot should look like Figure 5.4. The data is given hereafter:

```
x = [0.0; 0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9; 1.0]
y = [-0.9187980789440975; -0.6159791344678258; -0.25568734869121856;
     -0.14269370171581808; 0.3094396057228459; 0.6318327173549161;
     0.8370437988106428; 1.0970402798788812; 1.6057799131867696;
     1.869090784869698; 2.075369730726694]
```

5.7 Discussion and bibliography

The content of this chapter is largely based on the lecture notes [15]. Several of the exercises are taken or inspired from [7]. The proof of convergence of the secant method is inspired from the general proof presented in the short paper [17]. For a detailed treatment of iterative methods for nonlinear equations, see the book [9].