EPFL

MATH-251(a) Numerical analysis

Guillaume Olikier

December 18, 2025

Contents

Chapter 1

Linear systems: direct methods

Given $n \in \mathbb{N} \setminus \{0, 1\}$, $A \in \mathbb{R}^{n \times n}$, and $b \in \mathbb{R}^n$, find $x \in \mathbb{R}^n$ such that $Ax = b$. If A is invertible, then the unique solution is $A^{-1}b$. This chapter introduces Gaussian elimination, an algorithm that computes $A^{-1}b$ without computing A^{-1} explicitly. Nonsquare systems are briefly covered in Section 1.5. Notation:

$$A = [a_{i,j}]_{i,j=1}^n = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix}, \qquad b = [b_i]_{i=1}^n = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

1.1 Triangular systems

The matrix $A \in \mathbb{R}^{n \times n}$ is said to be *lower* (resp. *upper*) *triangular* if $a_{i,j} = 0$ for all $i, j \in \{1, \ldots, n\}$ such that $i < j$ (resp. $i > j$). The matrix A is said to be triangular if it is lower or upper triangular. Notation for A respectively lower and upper triangular:

$$\begin{bmatrix} a_{1,1} & & \\ \vdots & \ddots & \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix}, \qquad \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ & \ddots & \vdots \\ & & a_{n,n} \end{bmatrix},$$

where the empty entries are those that are structurally zero.

If A is triangular, then it is invertible if and only if its diagonal entries are nonzero, in which case $x \coloneqq A^{-1}b$ can be computed as follows:

- if A is lower triangular, $x_i \coloneqq (b_i - \sum_{j=1}^{i-1} a_{i,j}x_j)/a_{i,i}$ for $i = 1, \ldots, n$ (forward substitution);
- if A is upper triangular, $x_i \coloneqq (b_i - \sum_{j=i+1}^{n} a_{i,j}x_j)/a_{i,i}$ for $i = n, \ldots, 1$ (back substitution).

Forward and back substitutions each require $\sum_{i=1}^n (2i - 1) = n^2$ arithmetic operations.

1.2 Gaussian elimination and LU factorization

A unit triangular matrix is a triangular matrix whose diagonal entries equal 1. Gaussian elimination factorizes the matrix A as the product of a unit lower triangular matrix L and an upper triangular matrix U, thereby reducing the system $Ax = b$ to two triangular systems, $Ly = b$ and then $Ux = y$.

A basic step of Gaussian elimination works as follows:

- if the upper-left entry, called the *pivot*, is nonzero, then a multiple of the first row is substracted from each of the other rows to make zeros appear in the first column of these rows;
- the coefficient associated with a row is stored in the first entry of the row, at the place of the zero.

The complete process involves at most $n-1$ basic steps: for $i = 0, \ldots, n-2$, apply a basic step to the order-$(n-i)$ lower-right submatrix.

In practice, the solution y to the system $Ly = b$ and the LU factorization of A can be computed simultaneously, as in Algorithm 1.1 and Example 1.2.1.

Algorithm 1.1 Gaussian elimination and LU factorization

Input: (A, b), where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$.

Output: $(L, U, L^{-1}b)$, where $L \in \mathbb{R}^{n\times n}$ is unit lower triangular, $U \in \mathbb{R}^{n\times n}$ is upper triangular, and $LU = A$. The upper triangular part of A is overwritten by U, the strict lower triangular part of A is overwritten by the strict lower part of L, and b is overwritten by $L^{-1}b$.

```
for i = 1, ..., n − 1 do
    if a_{i,i} ≠ 0 then
        for j = i + 1, ..., n do
            a_{j,i} ← a_{j,i}/a_{i,i};
            for k = i + 1, ..., n do
                a_{j,k} ← a_{j,k} − a_{j,i}a_{i,k};
            end for
            b_j ← b_j − a_{j,i}b_i;
        end for
    else
        Stop;
    end if
end for
```

Example 1.2.1 ([TB97, Lecture 20])**.** Let

$$A \coloneqq \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}, \qquad b \coloneqq \begin{bmatrix} 2 \\ 3 \\ 5 \\ 0 \end{bmatrix}.$$

Step 1:

$$\begin{bmatrix} 2 & 1 & 1 & 0 & 2 \\ 2 & 1 & 1 & 1 & -1 \\ 4 & 3 & 5 & 5 & -3 \\ 3 & 4 & 6 & 8 & -6 \end{bmatrix}, \qquad \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & 3 & 5 & 5 \\ & 4 & 6 & 8 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & & & \\ -2 & 1 & & \\ -4 & & 1 & \\ -3 & & & 1 \end{bmatrix}}_{=:L_1} \underbrace{\begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}}_{=A}.$$

Step 2:

$$\begin{bmatrix} 2 & 1 & 1 & 0 & 2 \\ 2 & 1 & 1 & 1 & -1 \\ 4 & 3 & 2 & 2 & 0 \\ 3 & 4 & 2 & 4 & -2 \end{bmatrix}, \qquad \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & 2 & 4 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & & & \\ & 1 & & \\ & -3 & 1 & \\ & -4 & & 1 \end{bmatrix}}_{=:L_2} \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & 3 & 5 & 5 \\ & 4 & 6 & 8 \end{bmatrix}.$$

Step 3:

$$\begin{bmatrix} 2 & 1 & 1 & 0 & 2 \\ 2 & 1 & 1 & 1 & -1 \\ 4 & 3 & 2 & 2 & 0 \\ 3 & 4 & 1 & 2 & -2 \end{bmatrix}, \qquad \underbrace{\begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & & 2 \end{bmatrix}}_{=:U} = \underbrace{\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & -1 & 1 \end{bmatrix}}_{=:L_3} \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & 2 & 4 \end{bmatrix}.$$

Thus, $L_3L_2L_1A = U$. Observe that

$$L_1^{-1} = \begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & & 1 & \\ 3 & & & 1 \end{bmatrix}, \qquad L_2^{-1} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & 3 & 1 & \\ & 4 & & 1 \end{bmatrix}, \qquad L_3^{-1} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & 1 & 1 \end{bmatrix},$$

and

$$L := L_1^{-1}L_2^{-1}L_3^{-1} = \begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & 3 & 1 & \\ 3 & 4 & 1 & 1 \end{bmatrix}.$$

Conclusion: $A = LU$ and, by back substitution, $x = \begin{bmatrix} 1 & -1 & 1 & -1 \end{bmatrix}^\top$.

Number of arithmetic operations required by Algorithm 1.1:

- for updating b,
$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2 = 2\sum_{i=1}^{n-1}(n-i) = 2\sum_{i=1}^{n-1} i = n^2 - n;$$
- for updating A, i.e., computing L and U,
$$\begin{aligned} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(1 + \sum_{k=i+1}^{n} 2\right) &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 1 + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=i+1}^{n} 1 \\ &= \sum_{i=1}^{n-1}(n-i) + 2\sum_{i=1}^{n-1}(n-i)^2 \\ &= \sum_{i=1}^{n-1} i + 2\sum_{i=1}^{n-1} i^2 \\ &= \frac{n(n-1)}{2} + 2\frac{n(2n-1)(n-1)}{6} \\ &= \frac{2}{3}n^3 - \frac{1}{2}n^2 - \frac{1}{6}n. \end{aligned}$$

Computing x by back substitution requires n^2 arithmetic operations. In conclusion, solving the linear system with Gaussian elimination requires

$$\frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n$$

arithmetic operations.

With matrix notation, Algorithm 1.1 becomes Algorithm 1.2, which has the same input and output.

Algorithm 1.2 Gaussian elimination and LU factorization in matrix notation [GV13, Algorithm 3.2.1]

```
for i = 1, ..., n − 1 do
    if a_{i,i} ≠ 0 then
        ρ ← i + 1:n;
        a_{ρ,i} ← a_{ρ,i}/a_{i,i};
        a_{ρ,ρ} ← a_{ρ,ρ} − a_{ρ,i}a_{i,ρ};
        b_ρ ← b_ρ − a_{ρ,i}b_i;
    else
        Stop;
    end if
end for
```

As defined in this section, Gaussian elimination can fail even if A is invertible, as illustrated by the following.

Example 1.2.2. Let

$$A \coloneqq \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 2 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}.$$

Step 1:

$$\begin{bmatrix} 1 & & & \\ -2 & 1 & & \\ -4 & & 1 & \\ -3 & & & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 2 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 0 & 1 & 1 \\ & 3 & 5 & 5 \\ & 4 & 6 & 8 \end{bmatrix}.$$

Step 2 cannot be applied because the pivot, the $(2, 2)$ entry, is zero. However, A is invertible, as shown in Example 1.3.1.

1.3 Gaussian elimination with pivoting

If A is invertible and a zero pivot is encountered, then the corresponding row can be permuted with a lower row to obtain a nonzero pivot. (Indeed, the matrix at step $i \in \mathbb{N}$ is

$$\begin{bmatrix} \hat{U} & \hat{A}_1 \\ 0_{n-i \times i} & \hat{A}_2 \end{bmatrix},$$

where $\hat{U} \in \mathbb{R}^{i \times i}$ is upper triangular with nonzero diagonal entries, $\hat{A}_1 \in \mathbb{R}^{i \times n-i}$, and $\hat{A}_2 \in \mathbb{R}^{n-i \times n-i}$. Thus, $\det A = \det \hat{U} \det \hat{A}_2$ and $\det \hat{U} \neq 0$. Therefore, $\det A = 0$ if and only if $\det \hat{A}_2 = 0$.) For numerical stability, a pivot close to zero can be problematic as well. This motivates to choose the permutation that maximizes the absolute value of the pivot, as in Algorithm 1.3.

Algorithm 1.3 Gaussian elimination with pivoting [GV13, Algorithm 3.4.1]

Input: (A, b), where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$.

Output: $(L, U, p, L^{-1}b)$, where $L \in \mathbb{R}^{n \times n}$ is unit lower triangular, $U \in \mathbb{R}^{n \times n}$ is upper triangular, p is a permutation of $[1 \ \cdots \ n]^\top$, and $LU = A_{p,:}$. The upper triangular part of A is overwritten by U, the strict lower triangular part of A is overwritten by the strict lower part of L, and b is overwritten by $L^{-1}b$.

$p \leftarrow [1 \ \cdots \ n]^\top$;
for $i = 1, \ldots, n-1$ **do**
  Choose $j \in \operatorname{argmax}_{k \in \{i, \ldots, n\}} |a_{k,i}|$;
  if $a_{j,i} = 0$ **then**
    Stop: A is singular;
  else
    $a_{i,:} \leftrightarrow a_{j,:}$;
    $b_i \leftrightarrow b_j$;
    $p_i \leftrightarrow p_j$;
    $\rho \leftarrow i + 1{:}n$;
    $a_{\rho,i} \leftarrow a_{\rho,i} / a_{i,i}$;
    $a_{\rho,\rho} \leftarrow a_{\rho,\rho} - a_{\rho,i} a_{i,\rho}$;
    $b_\rho \leftarrow b_\rho - a_{\rho,i} b_i$;
  end if
end for

Number of comparisons of real numbers performed for pivoting: $\sum_{i=1}^{n-1}(n-i) = \sum_{i=1}^{n-1} i = \frac{1}{2}n(n-1)$.

Example 1.3.1. Let

$$A := \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 2 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}, \qquad b := \begin{bmatrix} 2 \\ 4 \\ 5 \\ 0 \end{bmatrix}.$$

Step 1:

- permutation of two rows:

$$\begin{bmatrix} 8 & 7 & 9 & 5 & 5 \\ 4 & 2 & 3 & 1 & 4 \\ 2 & 1 & 1 & 0 & 2 \\ 6 & 7 & 9 & 8 & 0 \end{bmatrix}, \qquad p = \begin{bmatrix} 3 \\ 2 \\ 1 \\ 4 \end{bmatrix}, \qquad \begin{bmatrix} 8 & 7 & 9 & 5 \\ 4 & 2 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{bmatrix} = \underbrace{\begin{bmatrix} & & 1 & \\ & 1 & & \\ 1 & & & \\ & & & 1 \end{bmatrix}}_{=:P_1} \underbrace{\begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 2 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}}_{=A};$$

- Gaussian elimination:

$$\begin{bmatrix} 8 & 7 & 9 & 5 & 5 \\ \frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} & -\frac{3}{2} & \frac{3}{2} \\ \frac{1}{4} & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} & -\frac{15}{4} \end{bmatrix}, \qquad \begin{bmatrix} 8 & 7 & 9 & 5 \\ & -\frac{3}{2} & -\frac{3}{2} & -\frac{3}{2} \\ & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & & & \\ -\frac{1}{2} & 1 & & \\ -\frac{1}{4} & & 1 & \\ -\frac{3}{4} & & & 1 \end{bmatrix}}_{=:L_1} \begin{bmatrix} 8 & 7 & 9 & 5 \\ 4 & 2 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{bmatrix}.$$

Step 2:

- permutation of two rows:

$$\begin{bmatrix} 8 & 7 & 9 & 5 & 5 \\ \frac{3}{4} & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} & -\frac{15}{4} \\ \frac{1}{4} & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} & \frac{3}{4} \\ \frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} & -\frac{3}{2} & \frac{3}{2} \end{bmatrix}, \quad p = \begin{bmatrix} 3 \\ 4 \\ 1 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ & -\frac{3}{2} & -\frac{3}{2} & -\frac{3}{2} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & & & \\ & & & 1 \\ & & 1 & \\ & 1 & & \end{bmatrix}}_{=:P_2} \begin{bmatrix} 8 & 7 & 9 & 5 \\ & -\frac{3}{2} & -\frac{3}{2} & -\frac{3}{2} \\ & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \end{bmatrix};$$

- Gaussian elimination:

$$\begin{bmatrix} 8 & 7 & 9 & 5 & 5 \\ \frac{3}{4} & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} & -\frac{15}{4} \\ \frac{1}{4} & -\frac{3}{7} & -\frac{2}{7} & \frac{4}{7} & -\frac{6}{7} \\ \frac{1}{2} & -\frac{6}{7} & \frac{3}{7} & \frac{15}{7} & -\frac{12}{7} \end{bmatrix}, \qquad \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & -\frac{2}{7} & \frac{4}{7} \\ & & \frac{3}{7} & \frac{15}{7} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & & & \\ & 1 & & \\ & \frac{3}{7} & 1 & \\ & \frac{6}{7} & & 1 \end{bmatrix}}_{=:L_2} \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ & -\frac{3}{2} & -\frac{3}{2} & -\frac{3}{2} \end{bmatrix}.$$

Step 3:

- permutation of two rows:

$$\begin{bmatrix} 8 & 7 & 9 & 5 & 5 \\ \frac{3}{4} & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} & -\frac{15}{4} \\ \frac{1}{2} & -\frac{6}{7} & \frac{3}{7} & \frac{15}{7} & -\frac{12}{7} \\ \frac{1}{4} & -\frac{3}{7} & -\frac{2}{7} & \frac{4}{7} & -\frac{6}{7} \end{bmatrix}, \quad p = \begin{bmatrix} 3 \\ 4 \\ 2 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & \frac{3}{7} & \frac{15}{7} \\ & & -\frac{2}{7} & \frac{4}{7} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & & 1 \\ & & 1 & \end{bmatrix}}_{=:P_3} \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & -\frac{2}{7} & \frac{4}{7} \\ & & \frac{3}{7} & \frac{15}{7} \end{bmatrix};$$

- Gaussian elimination:

$$\begin{bmatrix} 8 & 7 & 9 & 5 & 5 \\ \frac{3}{4} & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} & -\frac{15}{4} \\ \frac{1}{2} & -\frac{6}{7} & \frac{3}{7} & \frac{15}{7} & -\frac{12}{7} \\ \frac{1}{4} & -\frac{3}{7} & -\frac{2}{3} & 2 & -2 \end{bmatrix}, \quad \underbrace{\begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & \frac{3}{7} & \frac{15}{7} \\ & & & 2 \end{bmatrix}}_{=:U} = \underbrace{\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & \frac{2}{3} & 1 \end{bmatrix}}_{=:L_3} \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & \frac{3}{7} & \frac{15}{7} \\ & & -\frac{2}{7} & \frac{4}{7} \end{bmatrix}.$$

Thus, $L_3P_3L_2P_2L_1P_1A = U$. Observe that

$$L_3P_3L_2P_2L_1P_1 = L_3(\underbrace{P_3L_2P_3}_{=:\tilde{L}_2})(\underbrace{P_3P_2L_1P_2P_3}_{=:\tilde{L}_1})(\underbrace{P_3P_2P_1}_{=:P}),$$

$$\tilde{L}_1^{-1} = \begin{bmatrix} 1 & & & \\ \frac{3}{4} & 1 & & \\ \frac{1}{2} & & 1 & \\ \frac{1}{4} & & & 1 \end{bmatrix}, \quad \tilde{L}_2^{-1} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & -\frac{6}{7} & 1 & \\ & -\frac{3}{7} & & 1 \end{bmatrix}, \quad L_3^{-1} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & -\frac{2}{3} & 1 \end{bmatrix}, \quad P = \begin{bmatrix} & & 1 & \\ & & & 1 \\ & 1 & & \\ 1 & & & \end{bmatrix},$$

and

$$L := \tilde{L}_1^{-1}\tilde{L}_2^{-1}L_3^{-1} = \begin{bmatrix} 1 & & & \\ \frac{3}{4} & 1 & & \\ \frac{1}{2} & -\frac{6}{7} & 1 & \\ \frac{1}{4} & -\frac{3}{7} & -\frac{2}{3} & 1 \end{bmatrix}.$$

Conclusion: $PA = LU$ and, by back substitution, $x = \begin{bmatrix} 1 & -1 & 1 & -1 \end{bmatrix}^\top$.

Conclusion: in exact arithmetic, linear systems can be solved exactly within a finite number of arithmetic operations. Specifically, if A is invertible, then Gaussian elimination with pivoting computes $A^{-1}b$ within $\frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n$ arithmetic operations and $\frac{1}{2}n(n-1)$ comparisons of real numbers.

1.4 Effect of round-off errors

Computers represent real numbers in a floating-point format and use floating-point arithmetic (which is not exact arithmetic). With backward error analysis, it can be shown that, given floating-point representations of A and b, Gaussian elimination in floating-point arithmetic returns an exact solution to a perturbed system. Therefore, after reviewing background material about matrix norms and condition numbers (Section 1.4.1), we state a perturbation theorem (Section 1.4.2). Then, after reviewing the two main properties of floating-point arithmetic (Section 1.4.3), we analyze the effect of round-off errors by combining the perturbation theorem and a backward error analysis of Gaussian elimination (Section 1.4.4).

1.4.1 Matrix norm and condition number

See Section B.2 for backbround material on the concept of norm.

Proposition 1.4.1. *If $\|\cdot\|$ is a norm on $\mathbb{R}^n$, then the function*

$$\mathbb{R}^{n\times n} \to \mathbb{R} : M \mapsto \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Mx\|}{\|x\|}$$

is a norm on $\mathbb{R}^{n\times n}$ called the induced norm *and denoted also by $\|\cdot\|$.*

Examples are given in Table 1.1.

Norm on $\mathbb{R}^n$	Induced norm on $\mathbb{R}^{n\times n}$
$\|x\|_1 \coloneqq \sum_{i=1}^n \|x_i\|$	$\|M\|_1 = \max_{j\in\{1,\dots,n\}} \sum_{i=1}^n \|m_{i,j}\|$
$\|x\|_2 \coloneqq \sqrt{\sum_{i=1}^n x_i^2}$	$\|M\|_2 = \sigma_{\max}(M) = \sqrt{\lambda_{\max}(M^\top M)}$
$\|x\|_\infty \coloneqq \max_{i\in\{1,\dots,n\}} \|x_i\|$	$\|M\|_\infty = \max_{i\in\{1,\dots,n\}} \sum_{j=1}^n \|m_{i,j}\|$

Table 1.1: Examples of norms on $\mathbb{R}^n$ and induced norms on $\mathbb{R}^{n\times n}$.

Proposition 1.4.2. *Every induced norm on $\mathbb{R}^{n\times n}$ is submultiplicative: for all $M, N \in \mathbb{R}^{n\times n}$,*

$$\|MN\| \le \|M\|\|N\|.$$

For example, the function

$$\|\cdot\|_{\max} : \mathbb{R}^{n\times n} \to \mathbb{R} : M \mapsto \max_{i,j\in\{1,\dots,n\}} |m_{i,j}|$$

is a norm on $\mathbb{R}^{n\times n}$ that is not induced by a norm on $\mathbb{R}^n$ because it is not submultiplicative.

Definition 1.4.3. The *condition number* of an invertible matrix $M \in \mathbb{R}^{n\times n}$ is $\kappa(M) \coloneqq \|M\|\|M^{-1}\|$.

For every $p \in \{1, 2, \infty\}$, the condition number associated with the norm $\|\cdot\|_p$ from Table 1.1, denoted by κ_p, is at least one. Indeed, for all invertible $M \in \mathbb{R}^{n\times n}$,

$$1 = \|I_n\|_p = \|MM^{-1}\|_p \le \|M\|_p\|M^{-1}\|_p = \kappa_p(M).$$

1.4.2 Sensitivity of linear systems

Theorem 1.4.4 ([Dem97, (2.4)])**.** *Let $\|\cdot\|$ denote both a norm on $\mathbb{R}^n$ and the induced norm on $\mathbb{R}^{n\times n}$. Let $A \in \mathbb{R}^{n\times n}$ be invertible, $b \in \mathbb{R}^n\setminus\{0\}$, $\tilde{A} \in \mathbb{R}^{n\times n}$, and $\tilde{b} \in \mathbb{R}^n$. If $\frac{\|A-\tilde{A}\|}{\|A\|} < \frac{1}{\kappa(A)}$, then $\tilde{A}$ is invertible. If, moreover, $x, \tilde{x} \in \mathbb{R}^n$ satisfy $Ax = b$ and $\tilde{A}\tilde{x} = \tilde{b}$, then*

$$\frac{\|x-\tilde{x}\|}{\|x\|} \le \frac{\kappa(A)}{1-\kappa(A)\frac{\|A-\tilde{A}\|}{\|A\|}}\left(\frac{\|A-\tilde{A}\|}{\|A\|} + \frac{\|b-\tilde{b}\|}{\|b\|}\right).$$

1.4.3 Floating-point arithmetic

Computers represent real numbers in a floating-point format—typically the IEEE 754 double-precision binary floating-point format (binary64)—and use floating-point arithmetic. The floating-point number that is closest to $x \in \mathbb{R}$ is denoted by $\mathrm{fl}(x)$. The analysis of round-off errors is based on the existence of a positive real number $\varepsilon_{\text{machine}}$, called *machine epsilon*, such that:

1. for every $x \in \mathbb{R}$, there exists $\varepsilon \in [-\varepsilon_{\text{machine}}, \varepsilon_{\text{machine}}]$ such that $\mathrm{fl}(x) = x(1+\varepsilon)$, i.e., $\frac{\mathrm{fl}(x)-x}{x} = \varepsilon$;

2. for every arithmetic operation $*$ (addition, substraction, multiplication, or division) and every pair (x, y) of floating-point numbers, there exists $\varepsilon \in [-\varepsilon_{\text{machine}}, \varepsilon_{\text{machine}}]$ such that $x \circledast y = (x * y)(1+\varepsilon)$, where $\circledast$ is $*$ in floating-point arithmetic.

Property 1 is a property of the floating-point format, while Property 2 is a property of floating-point arithmetic. Property 1 actually holds only for $x \in [-N_{\max}, N_{\max}] \setminus (-N_{\min}, N_{\min})$ but this has no practical impact. For example, with binary64, $N_{\max} = (2-2^{-52})2^{1023} \approx 10^{308}$, $N_{\min} = 2^{-1022} \approx 10^{-308}$, and $\varepsilon_{\text{machine}}$ is about $2^{-52} \approx 2.22 \cdot 10^{-16}$ or even $2^{-53} \approx 1.11 \cdot 10^{-16}$.

1.4.4 Gaussian elimination in floating-point arithmetic

Two sorts of errors are involved. The first sort comes from the representation of A and b in the chosen floating-point format: A and b become respectively $\mathrm{fl}(A)$ and $\mathrm{fl}(b)$, where fl is applied componentwise. As illustrated next, if A is ill-conditioned, i.e., $\kappa(A)$ is large, then the perturbation of fl can be significant. For example, if $\varepsilon_{\text{machine}}\kappa_\infty(A) = \frac{1}{2}$, then the upper bound given by the perturbation theorem (Theorem 1.4.4) is merely upper bounded by 2. Indeed, by property 1, $|\mathrm{fl}(A)-A| \le \varepsilon_{\text{machine}}|A|$ and $|\mathrm{fl}(b)-b| \le \varepsilon_{\text{machine}}|b|$, where the absolute value is applied componentwise. Thus, $\|\mathrm{fl}(A)-A\|_\infty \le \varepsilon_{\text{machine}}\|A\|_\infty$ and $\|\mathrm{fl}(b)-b\|_\infty \le \varepsilon_{\text{machine}}\|b\|_\infty$.

Example 1.4.5. Let $A \coloneqq \begin{bmatrix} 1 & 2^{-54} \\ 1 & 0 \end{bmatrix}$ and $b \coloneqq \begin{bmatrix} 1+2^{-54} \\ 1 \end{bmatrix}$. Then, $A^{-1}b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. With binary64, $\mathrm{fl}(A) = A$ and $\mathrm{fl}(b) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, thus $\mathrm{fl}(A)^{-1}\mathrm{fl}(b) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Since $A^{-1} = \begin{bmatrix} 0 & 1 \\ 2^{54} & -2^{54} \end{bmatrix}$, $\|A\|_\infty = 1+2^{-54}$, and $\|A^{-1}\|_\infty = 2^{55}$, it holds that $\kappa_\infty(A) = 2^{55}+2$ and the upper bound given by the perturbation theorem (Theorem 1.4.4) equals 2, twice the actual relative error.

In conclusion, if A is ill-conditioned, then algorithms cannot be expected to solve the system with high accuracy.

The second sort of errors comes from floating-point arithmetic. Backward error analysis shows that the round-off errors made in the course of Gaussian elimination can be projected back on the original matrix, as stated in the following.

Theorem 1.4.6 ([Hig02, Theorems 9.3 and 9.4])**.** *Given A and b in floating-point format, Gaussian elimination yields (L,U) such that*

$$|LU - A| \le \frac{n\varepsilon_{\text{machine}}}{1-n\varepsilon_{\text{machine}}}|L||U|$$

provided that $n\varepsilon_{\text{machine}} < 1$. Moreover, the computed solution $\tilde{x}$ satisfies $\tilde{A}\tilde{x} = b$ with

$$|A - \tilde{A}| \le \frac{3n\varepsilon_{\text{machine}}}{1-3n\varepsilon_{\text{machine}}}|L||U|$$

provided that $3n\varepsilon_{\text{machine}} < 1$. Thus,

$$\frac{\|A-\tilde{A}\|_\infty}{\|A\|_\infty} \le \frac{3n^3\varepsilon_{\text{machine}}}{1-3n\varepsilon_{\text{machine}}}\frac{\|U\|_{\max}}{\|A\|_{\max}}.$$

Proof of the last statement. It holds that

$$\|A-\tilde{A}\|_\infty \le \frac{3n\varepsilon_{\text{machine}}}{1-3n\varepsilon_{\text{machine}}}\||L||U|\|_\infty.$$

Moreover,

$$\||L||U|\|_\infty \le \||L|\|_\infty\||U|\|_\infty = \|L\|_\infty\|U\|_\infty \le n\|U\|_\infty \le n^2\|U\|_{\max}.$$

The result follows from the inequality $\|A\|_{\max} \le \|A\|_\infty$. □

The number $\|U\|_{\max}/\|A\|_{\max}$ is called the *growth factor*.

Theorem 1.4.7 ([Ste98, Theorem 4.12])**.** *Let $PA = LU$ be an LU factorization obtained by Gaussian elimination with pivoting in exact arithmetic. Then, $\|U\|_{\max} \le 2^{n-1}\|A\|_{\max}$. Moreover, equality is reached if*

$$A = \begin{bmatrix} 1 & \cdots & \cdots & 1 \\ -1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -1 & \cdots & -1 & 1 \end{bmatrix}.$$

In practice, the growth factor is often observed to grow as $n^{\frac{2}{3}}$. For tridiagonal matrices, it is upper bounded by 2 [Ste98, p. 240].

In conclusion, backward error analysis reduces the analysis of round-off errors to the analysis of the sensitivity of the system: these errors can be estimated by combining Theorems 1.4.6 and 1.4.7 with the perturbation theorem (Theorem 1.4.4). However, the estimation is generally very pessimistic. Thus, other methods have been proposed to estimate round-off errors in practice.

1.5 Least-squares problems

Let $m, n \in \mathbb{N} \setminus \{0, 1\}$ such that $m \geq n$. Let $A \in \mathbb{R}^{m \times n}$ such that $\operatorname{rank} A = n$. Let $b \in \mathbb{R}^m$. If $b \notin \operatorname{im} A$, there is no $x \in \mathbb{R}^n$ such that $Ax = b$. However, the optimization problem $\min_{x \in \mathbb{R}^n} \|Ax - b\|$ can still be considered. It is called a least-squares problem if the norm $\|\cdot\|$ is the 2-norm $\|\cdot\|_2$. Considering the 2-norm is convenient because the 2-norm is induced by the usual inner product on $\mathbb{R}^m$ (see Section B.5). Furthermore, the choice of the 2-norm can be defended by geometric and statistical reasons. In the rest of this section, only the 2-norm is considered.

Solution: $x \in \mathbb{R}^n$ minimizes the function $\mathbb{R}^n \to \mathbb{R} : z \mapsto \|Az - b\|_2$ if and only if $Ax = y$ and $y \in \operatorname{im} A$ minimizes the function $\operatorname{im} A \to \mathbb{R} : z \mapsto \|z - b\|_2$, i.e., y is the orthogonal projection of b onto $\operatorname{im} A$. Thus, it suffices to find $x \in \mathbb{R}^n$ such that Ax is the orthogonal projection of b onto $\operatorname{im} A$, i.e., $b - Ax \perp \operatorname{im} A$, i.e., $\langle b - Ax, Az \rangle = 0$ for all $z \in \mathbb{R}^n$, i.e., $z^\top A^\top (b - Ax) = 0$ for all $z \in \mathbb{R}^n$, i.e., $A^\top (b - Ax) = 0$. This yields the so-called normal equations:

$$A^\top A x = A^\top b.$$

Since $\operatorname{rank} A^\top A = \operatorname{rank} A = n$, this linear system has a unique solution.

Example 1.5.1. Let

$$A \coloneqq \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \qquad b \coloneqq \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}.$$

Then, $b \notin \operatorname{im} A$ and

$$A^\top A = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}, \qquad A^\top b = \begin{bmatrix} 4 \\ 8 \end{bmatrix}.$$

Thus, the least-squares solution is $\begin{bmatrix} 4/3 \\ 0 \end{bmatrix}$.

1.6 Notes and references

Section 1.1 is based on [GV13, §3.1], [QSS07, §3.2], [SB02, §4.1], [Ste98, Chap. 3, §1], and [Hig02, Chap. 8]. Section 1.2 is based on [GV13, §3.2], [TB97, Lecture 20], [Ste98, Chap. 3, §1], [QSS07, §3.3], and [Hig02, Chap. 9]. Section 1.3 is based on [GV13, §3.4], [TB97, Lecture 21], [Dem97, §2.3], [Ste98, Chap. 3, §1], [QSS07, §3.5], [SB02, §4.1], and [Hig02, Chap. 9]. Section 1.4 is based on [Dem97, Ste98, Hig02, GV13]. Section 1.5 is based on [NW06, Lecture 11].

Chapter 2

Nonlinear systems

Given $n \in \mathbb{N} \setminus \{0\}$ and a continuous function $f : \mathbb{R}^n \to \mathbb{R}^n$, find a zero of f, i.e., a point $x \in \operatorname{dom} f$ such that $f(x) = 0$.

This problem, ubiquitous in science and engineering, has been studied by some of the most brilliant minds and is still the subject of intense research nowadays. The reason is that, in general, there is no algorithm that computes a zero of f within finitely many arithmetic operations. Thus, iterative methods are necessary.

Example 2.0.1. Example of a function f that has no zero: $f : \mathbb{R} \to \mathbb{R} : x \mapsto x^2 + 1$. Examples of a function f whose zeros cannot be computed within finitely many arithmetic operations: most polynomial functions of degree at least five, e.g., $f : \mathbb{R} \to \mathbb{R} : x \mapsto x^5 - 4x - 2$, and many other functions, e.g., $f : \mathbb{R} \to \mathbb{R} : x \mapsto \cos x - x$, $f : \mathbb{R} \to \mathbb{R} : x \mapsto \mathrm{e}^x - x - 2$, and $f : \mathbb{R}^2 \to \mathbb{R}^2 : (x, y) \mapsto (\sin(x + y) - 3x, \cos(x - y) - 3y)$.

The general approach in numerical analysis to tackle such problems is establishing conditions that ensure the existence of a solution and designing an iterative method, i.e., a method that, given an initial guess, generates a sequence that hopefully converges to a solution. The two are related: the existence theorems that we are going to see in this chapter admit a constructive proof that shows that an iterative method converges to a solution. The speed of convergence is a criterion to compare iterative methods. Another criterion is the computational cost per iteration.

2.1 Bisection method and intermediate-value theorem

In this section, we focus on the case where $n = 1$. The following theorem is equivalent to the intermediate-value theorem.

Theorem 2.1.1. *Let $a, b \in \mathbb{R}$ such that $a < b$. If f is continuous on $[a, b]$ and $f(a)f(b) < 0$, then there exists $c \in (a, b)$ such that $f(c) = 0$.*

The proof simply establishes that the bisection method, defined in Algorithm 2.1 and illustrated in Figure 2.1, converges. Without loss of generality, assume that $f(a) < 0$ and $f(b) > 0$.

Algorithm 2.1 Bisection method

Require: (f, a, b), where $a, b \in \mathbb{R}$, $a < b$, $f : \mathbb{R} \to \mathbb{R}$ is continuous on $[a, b]$, $f(a) < 0$, and $f(b) > 0$.

$a_0 \leftarrow a$; $b_0 \leftarrow b$; $x_0 \leftarrow \frac{a+b}{2}$; $i \leftarrow 0$;
while $f(x_i) \neq 0$ **do**
    if $f(x_i) < 0$ **then**
        $a_{i+1} \leftarrow x_i$; $b_{i+1} \leftarrow b_i$;
    else
        $a_{i+1} \leftarrow a_i$; $b_{i+1} \leftarrow x_i$;
    end if
    $i \leftarrow i + 1$; $x_i \leftarrow \frac{a_i+b_i}{2}$;
end while

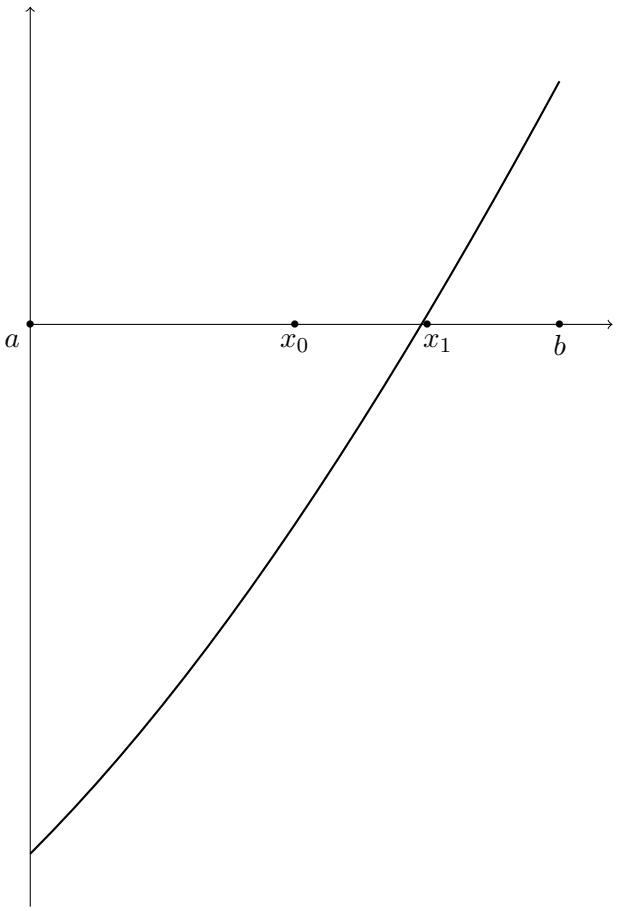

Figure 2.1: First two iterates of the bisection method for $f : [0, 1] \to \mathbb{R} : x \mapsto x - \cos x$.

The bisection method either finds a zero of $f|_{[a,b]}$ after a finite number of iterations or generates infinite sequences $(a_i)_{i\in\mathbb{N}}$, $(b_i)_{i\in\mathbb{N}}$, and $(x_i)_{i\in\mathbb{N}}$ such that, for all $i \in \mathbb{N}$, $a_i < x_i < b_i$, $b_i - a_i = 2^{-i}(b-a)$, $f(a_i) < 0$, and $f(b_i) > 0$. Since $(a_i)_{i\in\mathbb{N}}$ is monotonically nondecreasing and bounded from above, $(a_i)_{i\in\mathbb{N}}$ converges to $\sup_{i\in\mathbb{N}} a_i$. Similarly, since $(b_i)_{i\in\mathbb{N}}$ is monotonically nonincreasing and bounded from below, $(b_i)_{i\in\mathbb{N}}$ converges to $\inf_{i\in\mathbb{N}} b_i$. Moreover, $(a_i)_{i\in\mathbb{N}}$ and $(b_i)_{i\in\mathbb{N}}$ have the same limit. Letting i tend to infinity in $f(a_i) < 0$ and $f(b_i) > 0$ shows that the limit is a zero of f since f is continuous.

Unfortunately, no easy extension to the case where $n > 1$.

2.2 Fixed-point iteration and Banach fixed-point theorem

Reformulating Theorem 2.1.1 based on the following definition turns out to be particularly fruitful.

Definition 2.2.1. A *fixed point* of a function $g : \mathbb{R}^n \to \mathbb{R}^n$ is a point $x \in \operatorname{dom} g$ such that $g(x) = x$.

Theorem 2.2.2 (Brouwer fixed-point theorem)**.** *Let $a, b \in \mathbb{R}$ such that $a < b$. If $g : [a, b] \to \mathbb{R}$ is continuous and $g([a, b]) \subseteq [a, b]$, then there exists $c \in [a, b]$ such that $g(c) = c$.*

Proof. Apply Theorem 2.1.1 to $f : [a, b] \to \mathbb{R} : x \mapsto g(x) - x$. □

The proof of Theorem 2.2.2 shows that the problem of finding a zero of $f : \mathbb{R}^n \to \mathbb{R}^n$ can always be reformulated as the problem of finding a fixed point of a function $g : \mathbb{R}^n \to \mathbb{R}^n$. Furthermore, to find a fixed point of a function $g : \mathbb{R}^n \to \mathbb{R}^n$, it is natural to consider the iteration $x_{i+1} \coloneqq g(x_i)$ for all $i \in \mathbb{N}$. Indeed, if g is continuous, then the generated sequence can converge only to fixed points of g. However, the sequence may not converge, as illustrated in Figure 2.2. We are going to establish a sufficient condition for convergence based on Definition 2.2.3.

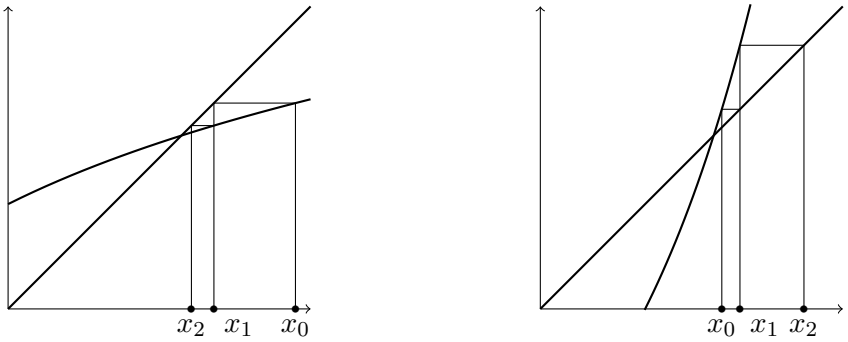

Figure 2.2: The fixed-point iteration converges for $g : (-2, \infty) \to \mathbb{R} : x \mapsto \ln(x+2)$ (left) but diverges for $g^{-1} : \mathbb{R} \to (-2, \infty) : x \mapsto \mathrm{e}^x - 2$ (right).

Definition 2.2.3. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be defined on a nonempty subset S of $\mathbb{R}^n$. The function g is said to be *Lipschitz continuous* on S if

$$\operatorname{Lip}_S g \coloneqq \sup_{\substack{x,y \in S \\ x \neq y}} \frac{\|g(x) - g(y)\|}{\|x - y\|} < \infty.$$

The function g is called a *contraction* on S if $\operatorname{Lip}_S g < 1$.

The name "contraction" stems from the fact that, for all $x, y \in S$, the distance between $g(x)$ and $g(y)$ is smaller than that between x and y. The following is a tool to identify Lipschitz continuous functions.

Proposition 2.2.4. *Let $\|\cdot\|$ denote both a norm on $\mathbb{R}^n$ and the induced norm on $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ (see Section B.2). Let U be an open convex subset of $\mathbb{R}^n$ and $g : U \to \mathbb{R}^n$ be differentiable on U. Then, $\operatorname{Lip}_U g = \sup_{x \in U} \|g'(x)\|$. Thus, g is Lipschitz continuous if and only if g' is bounded.*

Proof. Let x and y be two distinct points in U. By the mean-value theorem (Theorem B.3.2),

$$\|g(x) - g(y)\| \leq \|x - y\| \sup_{t \in (0,1)} \|g'(x + t(y - x))\|.$$

Thus, $\operatorname{Lip}_U g \leq \sup_{z \in U} \|g'(z)\|$.

Conversely, let $x \in U$ and $h \in \mathbb{R}^n \setminus \{0\}$. For all $t \in (0, \infty)$ sufficiently small, $x + th \in U$ and

$$\frac{\|g'(x)h\|}{\|h\|} = \frac{\|g'(x)th\|}{\|th\|} \leq \underbrace{\frac{\|g(x+th) - g(x) - g'(x)th\|}{\|th\|}}_{\to 0 \text{ as } t \to 0} + \underbrace{\frac{\|g(x+th) - g(x)\|}{\|th\|}}_{\leq \operatorname{Lip}_U g}.$$

Thus, $\|g'(x)\| \leq \operatorname{Lip}_U g$. □

Notation: $g^0 \coloneqq g$ and, for all $i \in \mathbb{N}$, $g^{i+1} \coloneqq g \circ g^i$.

Theorem 2.2.5 (Banach fixed-point theorem)**.** *Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be defined and continuous on a nonempty closed subset C of $\mathbb{R}^n$. If g is a contraction on C and $g(C) \subseteq C$, then g has a unique fixed point $x_* \in C$ and, for every $x \in C$, $\|g^i(x) - x_*\| \leq (\operatorname{Lip}_C g)^i \|x - x_*\|$ for all $i \in \mathbb{N}$, which implies that $(g^i(x))_{i \in \mathbb{N}}$ converges to x_*.*

Proof. Let $L \coloneqq \operatorname{Lip}_C g$. Existence of the fixed point. First, for all $x, y \in C$,

$$\|x - y\| \leq \frac{\|x - g(x)\| + \|y - g(y)\|}{1 - L} \tag{2.1}$$

since

$$\|x - y\| \leq \|x - g(x)\| + \|g(x) - g(y)\| + \|g(y) - y\| \leq \|x - g(x)\| + L\|x - y\| + \|g(y) - y\|.$$

Based on (2.1), let us prove that, for every $x \in C$, $(g^i(x))_{i\in\mathbb{N}}$ is a Cauchy sequence (see Section B.2). For all $i, j \in \mathbb{N}$,

$$\|g^i(x) - g^j(x)\| \leq \frac{\|g^i(x) - g^{i+1}(x)\| + \|g^j(x) - g^{j+1}(y)\|}{1 - L} \leq \frac{L^i + L^j}{1 - L}\|g(x) - x\|.$$

Since $\mathbb{R}^n$ is complete (Theorem B.2.5), $(g^i(x))_{i\in\mathbb{N}}$ converges. Moreover, letting i tend to infinity in $g^{i+1}(x) = g(g^i(x))$ shows that the limit is a fixed point of g since g is continuous.

Uniqueness of the fixed point. If $x, y \in C$ are fixed points of g, then $\|x - y\| \leq 0$ by (2.1), hence $\|x - y\| = 0$, and therefore $x = y$.

Inequality. Let $x_* \in C$ be a fixed point of g. For every $x \in C$ and $i \in \mathbb{N}$,

$$\|g^i(x) - x_*\| = \|g^i(x) - g^i(x_*)\| \leq L^i\|x - x_*\|. \qquad \square$$

The preceding theorem ensures linear convergence.

There are several ways of turning the problem of finding a zero into that of finding a fixed point, and some ways may be better than others, as illustrated next.

Example 2.2.6. Let $f : \mathbb{R} \to \mathbb{R} : x \mapsto \mathrm{e}^x - x - 2$. Then, $f(-2) > 0$, $f(-1) < 0$, $f(1) < 0$, and $f(2) > 0$. Thus, f has a zero on $(-2, -1)$ and a zero on $(1, 2)$.

- Define $g : \mathbb{R} \to \mathbb{R} : x \mapsto \mathrm{e}^x - 2$. Then, the zeros of f are exactly the fixed points of g. For all $x \in (-2, -1)$, $-2 < g(-2) < g(x) < g(-1) < -1$. For all $x \in \mathbb{R}$, $g'(x) = \mathrm{e}^x$. Thus, for all $x \in (-2, -1)$, $0 < \mathrm{e}^{-2} = g'(-2) < g'(x) < g'(-1) = \mathrm{e}^{-1} < 1$. Hence, g is a contraction on $[-2, -1]$. However, for all $x \in [0, \infty)$, $g'(x) \geq 1$. Therefore, g is not a contraction on any subinterval of $[0, \infty)$. In conclusion, with g, the fixed-point iteration can be used to find the fixed point in $(-2, -1)$ but not the fixed point in $(1, 2)$.
- Define $g : (-2, \infty) \to \mathbb{R} : x \mapsto \ln(x + 2)$. For all $x \in (1, 2)$, $1 < \ln 3 = g(1) < g(x) < g(2) = 2 \ln 2 < 2$. For all $x \in (-2, \infty)$, $g'(x) = \frac{1}{x+2}$. Thus, for all $x \in (1, 2)$, $0 < \frac{1}{4} = g'(2) < g'(x) < g'(1) = \frac{1}{3} < 1$. Hence, g is a contraction on $[1, 2]$. However, for all $x \in (-2, -1)$, $g'(x) \geq 1$. Therefore, g is not a contraction on $(-2, -1)$. In conclusion, the fixed-point iteration can be used to find the fixed point in $(1, 2)$ but not the fixed point in $(-2, -1)$.

The respective graphs of these functions are represented in Figure 2.3.

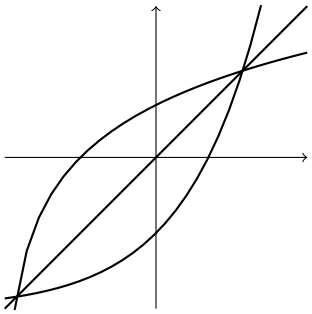

Figure 2.3: Functions from Example 2.2.6 on $(-2, 2)$.

2.3 Newton iteration

Faster convergence can be guaranteed if f is differentiable. The idea is to replace f with its first-order Taylor polynomial at every iteration. Thus, the first iteration works as follows: given x_0 in the interior of $\operatorname{dom} f$, solve $f(x_0) + f'(x_0)(x - x_0) = 0$ instead of $f(x) = 0$. If $f'(x_0)$ is invertible, then the linear system has a unique solution $x_1 \coloneqq x_0 - f'(x_0)^{-1}f(x_0)$. This yields the following.

Algorithm 2.2 Newton iteration

Require: (f, x_0), where $f : \mathbb{R}^n \to \mathbb{R}^n$ is differentiable and x_0 is in the interior of $\operatorname{dom} f$.

$i \leftarrow 0$;
while $f(x_i) \neq 0$ **do**
  try to find $x \in \mathbb{R}^n$ such that $f'(x_i)x = -f(x_i)$, e.g., by Gaussian elimination;
  if $f'(x_i)$ is not invertible **then**
    stop;
  else
    $x_{i+1} \leftarrow x_i + x$; $i \leftarrow i + 1$;
  end if
end while

The Newton iteration is a fixed-point iteration for $g : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto x - f'(x)^{-1}f(x)$.

Theorem 2.3.1 ([OR00, §10.2.2])**.** *Let $\|\cdot\|$ denote both a norm on $\mathbb{R}^n$ and the induced norm on $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ (see Section B.2). Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be defined and differentiable on a nonempty open subset U of $\mathbb{R}^n$. Let $x_* \in U$ be a zero of f. If f' is continuous at x_* and $f'(x_*)$ is invertible, then there exists $\delta \in (0, \infty)$ such that $B(x_*, \delta) \subseteq U$, $f'(x)$ is invertible for every $x \in B(x_*, \delta)$, and $\lim_{x \to x_*} \|g(x) - x_*\|/\|x - x_*\| = 0$, which implies that, for every $x \in B(x_*, \delta)$ sufficiently close to x_*, the sequence $(g^i(x))_{i \in \mathbb{N}}$ converges to x_*. If, moreover, there exist $L \in [0, \infty)$ and $p \in (0, 1]$ such that, for all $x \in B(x_*, \delta)$, $\|f'(x) - f'(x_*)\| \leq L\|x - x_*\|^p$, then $\|g(x) - x_*\| \leq 4L\|f'(x_*)^{-1}\|\|x - x_*\|^{p+1}$ for all $x \in B(x_*, \delta)$.*

The preceding theorem ensures local superlinear convergence, and even local quadratic convergence if $p = 1$.

2.4 Secant method

In this section, we focus on the case where $n = 1$. To avoid computing the derivative of f, we can replace f with its secant instead of its tangent. Thus, given $x_0, x_1 \in \operatorname{dom} f$, $x_0 \neq x_1$, we solve $f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) = 0$ instead of $f(x) = 0$. This yields the following.

Algorithm 2.3 Secant method

Require: (f, x_0, x_1), where $f : \mathbb{R} \to \mathbb{R}$ and $x_0, x_1 \in \operatorname{dom} f$.

$i \leftarrow 0$;
while $0 \neq f(x_i) \neq f(x_{i+1}) \neq 0$ **do**
  $x_{i+2} \leftarrow \dfrac{x_i f(x_{i+1}) - x_{i+1} f(x_i)}{f(x_{i+1}) - f(x_i)}$; $i \leftarrow i + 1$;
end while

If $f(x) = 0 \neq f'(x)$, then it can be proven that the secant converges superlinearly, although not quadratically, around x. The Newton iteration and the secant method are illustrated in Figure 2.4.

2.5 Stopping criteria

Stop when $\|f(x_i)\|$ or $\|x_{i+1} - x_i\|$ is smaller than a tolerance $\varepsilon \in (0, \infty)$. The inequalities obtained in the convergence analysis sometimes enable to find $i \in \mathbb{N}$ such that the second stopping criterion is satisfied. For illustration, let us find $i \in \mathbb{N}$ such that $\|x_i - x_*\| \leq \varepsilon$ for the bisection method, the fixed-point iteration, and the Newton iteration.

- Bisection method: $\left\lceil \log_2 \left(\frac{b-a}{\varepsilon} \right) - 1 \right\rceil$.

  For all $i \in \mathbb{N}$, $|x_i - x_*| \leq \frac{1}{2}(b_i - a_i) = 2^{-i-1}(b - a)$ thus $|x_i - x_*| \leq \varepsilon$ if $2^{-i-1}(b - a) \leq \varepsilon$, i.e., $i \geq \log_2 \left(\frac{b-a}{\varepsilon} \right) - 1$.

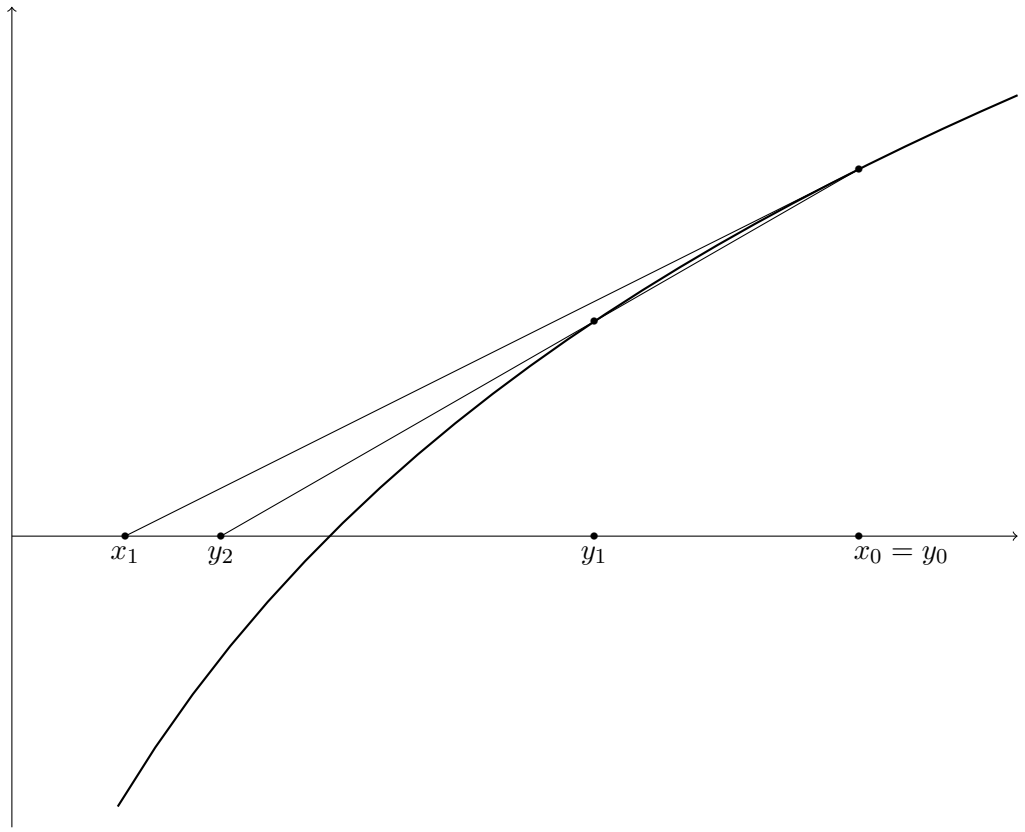

Figure 2.4: First iterates x_1 and y_2 respectively generated by the Newton iteration and the secant method.

- Fixed-point iteration: $\left\lceil \ln\left(\frac{\varepsilon(1-\operatorname{Lip}_C g)}{\|x_0-x_1\|}\right) / \ln \operatorname{Lip}_C g \right\rceil$.

  For all $i \in \mathbb{N}$, $\|x_i - x_*\| \leq (\operatorname{Lip}_C g)^i \|x_0 - x_*\|$ thus $\|x_i - x_*\| \leq \varepsilon$ if $(\operatorname{Lip}_C g)^i \|x_0 - x_*\| \leq \varepsilon$, i.e., $i \geq \ln(\varepsilon/\|x_0 - x_*\|)/\ln \operatorname{Lip}_C g$.

  Moreover, $\|x_0 - x_*\| \leq \|x_0 - x_1\| + \|x_1 - x_*\|$ and $\|x_1 - x_*\| = \|g(x_0) - g(x_*)\| \leq \operatorname{Lip}_C g \|x_0 - x_*\|$. Thus, $\|x_0 - x_*\| \leq \frac{1}{1-\operatorname{Lip}_C g}\|x_0 - x_1\|$. Hence, $\|x_i - x_*\| \leq \varepsilon$ if $\frac{(\operatorname{Lip}_C g)^i}{1-\operatorname{Lip}_C g}\|x_0 - x_1\| \leq \varepsilon$, i.e., $i \geq \ln\left(\frac{\varepsilon(1-\operatorname{Lip}_C g)}{\|x_0-x_1\|}\right) / \ln \operatorname{Lip}_C g$.

- Newton iteration in the case where $p = 1$: $\left\lceil \log_2\left(\frac{\ln(c(x_*)\varepsilon)}{\ln(c(x_*)\|x_0-x_*\|)}\right)\right\rceil$ with $c(x_*) \coloneqq 4L\|f'(x_*)^{-1}\|$.

  For all $x \in B(x_*, \delta)$, $\|g(x) - x_*\| \leq c(x_*)\|x - x_*\|^2$, thus $g(x) \in B(x_*, \delta)$ if $c(x_*)\delta\|x - x_*\| \leq \delta$, i.e., $\|x - x_*\| \leq 1/c(x_*)$. Let $\rho \coloneqq \min\{\delta, 1/c(x_*)\}$. Then, $g(B(x_*, \rho)) \subseteq B(x_*, \rho)$. Thus, for all $x_0 \in B(x_*, \rho)$ and $i \in \mathbb{N}$, $\|x_i - x_*\| \leq \frac{1}{c(x_*)}(c(x_*)\|x_0 - x_*\|)^{2^i}$, hence $\|x_i - x_*\| \leq \varepsilon$ if $\frac{1}{c(x_*)}(c(x_*)\|x_0 - x_*\|)^{2^i} \leq \varepsilon$, i.e., $i \geq \log_2\left(\frac{\ln(c(x_*)\varepsilon)}{\ln(c(x_*)\|x_0-x_*\|)}\right)$.

Example 2.5.1. The fixed-point iteration, the secant method, and the Newton iteration were applied to the problem of finding a fixed point of the cosine function, with $x_0 \coloneqq 0$ (and $x_1 \coloneqq 0.1$ for the secant method). The methods were stopped as soon as the distance between two consecutive iterates becomes smaller than or equal to 10^{-12}. Figure 2.5 represents the logarithm of the error as a function of the iteration counter. The point x_* is the last iterate generated by the methods. The difference between linear and superlinear convergence is clearly observable. The respective running times in seconds are $1.2 \cdot 10^{-4}$, $3.4 \cdot 10^{-5}$, and $1.7 \cdot 10^{-5}$.

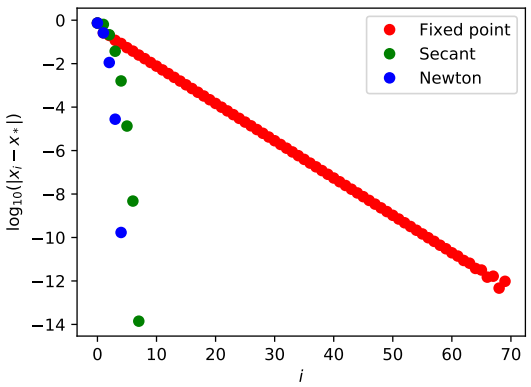

Figure 2.5: Empirical comparison of three numerical methods to compute a fixed point of the cosine function (see Example 2.5.1).

Chapter 3

Linear systems: iterative methods

Given $n \in \mathbb{N} \setminus \{0, 1\}$, $A \in \mathbb{R}^{n \times n}$ invertible, and $b \in \mathbb{R}^n$, find $x \in \mathbb{R}^n$ such that $Ax = b$. Gaussian elimination provides the exact solution in exact arithmetic but requires about $\frac{2}{3}n^3$ arithmetic operations, which is prohibitive if n is large. Iterative methods are an alternative to Gaussian elimination that is interesting if n is large.

3.1 Jacobi and Gauss–Seidel methods

Both methods rely on a decomposition, or splitting, $A = M - N$, where M is easy to invert and, for all $i, j \in \{1, \dots, n\}$, $m_{i,j}$ equals $a_{i,j}$ or 0. The system $Ax = b$ becomes $Mx = Nx + b$ or $x = M^{-1}Nx + M^{-1}b$, which is a fixed-point equation. Given $x_0 \in \mathbb{R}^n$, the fixed-point iteration is $Mx_{i+1} = Nx_i + b$ for all $i \in \mathbb{N}$. Thus, every iteration amounts to solving an easy linear system:

- M is the diagonal of A in the Jacobi method;
- M is the lower part of A in the Gauss–Seidel method.

The matrix M is called the *preconditioning matrix* or *preconditioner*. Here is why. The simplest way of rewriting $Ax = b$ as a fixed-point equation is arguably $x = (I_n - A)x + b$. The method that has just been described is that simple fixed-point iteration for the system $M^{-1}Ax = M^{-1}b$, which is called the *preconditioned system*.

Computational cost:

- $Nx_i + b$ requires (at most) $2n^2 + n$ arithmetic operations;
- solving the easy linear system requires n arithmetic operations for the Jacobi method and n^2 for the Gauss–Seidel method.

Convergence: Lipschitz constant of $g : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto M^{-1}Nx + M^{-1}b$. Clearly, $\mathrm{Lip}_{\mathbb{R}^n} g = \|M^{-1}N\|$, where $\|\cdot\|$ is the norm on $\mathbb{R}^{n \times n}$ induced by the norm on $\mathbb{R}^n$. Thus, convergence if $\|M^{-1}N\| < 1$. A more precise result can be stated based on the following.

Definition 3.1.1. For every $X \in \mathbb{R}^{n \times n}$, $\rho(X) \coloneqq \max_{i \in \{1,\dots,n\}} |\lambda_i(X)|$ is called the *spectral radius* of X.

Theorem 3.1.2 ([Saa03, Theorem 4.1])**.** *The iteration $x_{i+1} \coloneqq M^{-1}Nx_i + M^{-1}b$ converges for every $x_0 \in \mathbb{R}^n$ if and only if $\rho(M^{-1}N) < 1$.*

The matrix A is said to be *strictly diagonally dominant* if, for all $j \in \{1, \dots, n\}$,

$$|a_{j,j}| > \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{i,j}|.$$

If A is symmetric and, for all $x \in \mathbb{R}^n \setminus \{0\}$, $x^\top Ax > 0$, then A is said to be *positive-definite* (see Section A.2).

Theorem 3.1.3 ([Saa03, Theorem 4.9])**.** *If A is strictly diagonally dominant, then the associated Jacobi and Gauss–Seidel iterations converge for every $x_0 \in \mathbb{R}^n$.*

Theorem 3.1.4 ([Saa03, Theorem 4.10])**.** *If A is symmetric positive-definite, then the associated Gauss–Seidel iteration converges for every $x_0 \in \mathbb{R}^n$.*

Stopping criterion: stop when $\frac{\|b-Ax_i\|}{\|b\|}$ is smaller than a tolerance $\varepsilon \in (0,\infty)$. If this inequality is satisfied and $Ax = b$, then, by Theorem 1.4.4,

$$\frac{\|x - x_i\|}{\|x\|} \leq \kappa(A)\frac{\|b - Ax_i\|}{\|b\|} \leq \kappa(A)\varepsilon.$$

3.2 Gradient methods for symmetric positive-definite matrices

This section focuses on the case where A is symmetric positive-definite. Then, the function $\phi : \mathbb{R}^n \to \mathbb{R} : x \mapsto \frac{1}{2}x^\top Ax - b^\top x$ is strictly convex and $A^{-1}b$ is its unique global minimizer. Indeed, for all $x \in \mathbb{R}^n$, $\nabla\phi(x) = Ax - b$ and $\nabla^2\phi(x) = A$. The gradient can be computed as follows: for all $x, v \in \mathbb{R}^n$, $\langle\nabla\phi(x), v\rangle = v^\top\nabla\phi(x) = \phi'(x)v = \lim_{t\to 0}\frac{\phi(x+tv)-\phi(x)}{t}$ and, since

$$\begin{aligned}\phi(x + tv) - \phi(x) &= \frac{1}{2}(tv^\top Ax + tx^\top Av + t^2v^\top Av) - tb^\top v\\ &= \frac{t^2}{2}v^\top Av + tv^\top(Ax - b),\end{aligned}$$

$\lim_{t\to 0}\frac{\phi(x+tv)-\phi(x)}{t} = v^\top(Ax - b)$. Thus, for all $x, v \in \mathbb{R}^n$, $v^\top\nabla\phi(x) = v^\top(Ax - b)$, which implies $\nabla\phi(x) = Ax - b$, as announced.

Therefore, solving the linear system amounts to minimizing ϕ, which can be done by line-search methods. These methods rely on the following.

Definition 3.2.1. A *descent direction* for ϕ at $x \in \mathbb{R}^n$ is a vector $v \in \mathbb{R}^n$ such that $\phi'(x)v < 0$.

Given a descent direction v for ϕ at $x \in \mathbb{R}^n$, there exists $\alpha_* \in (0,\infty)$ such that, for all $\alpha \in (0, \alpha_*]$, $\phi(x + \alpha v) < \phi(x)$. However, since ϕ is quadratic, it is possible to compute $\operatorname{argmin}_{\alpha\in(0,\infty)}\phi(x + \alpha v)$. Indeed, as seen above, for all $\alpha \in (0,\infty)$, $\phi(x + \alpha v) = \phi(x) + \alpha v^\top\nabla\phi(x) + \frac{\alpha^2}{2}v^\top Av$, and the unique global minimizer is

$$-\frac{v^\top\nabla\phi(x)}{v^\top Av}. \tag{3.1}$$

Moving along a descent direction with this optimal step size is called exact line search. If the descent direction is chosen to be the steepest descent direction, given by the negative gradient, then the method is called the steepest descent method or the gradient method.

Algorithm 3.1 Steepest descent with exact line search for ϕ

Require: (A, x_0), where $A \in \mathbb{R}^{n\times n}$ is symmetric positive-definite and $x_0 \in \mathbb{R}^n$.

$i \leftarrow 0$; $r_0 \leftarrow b - Ax_0$;
while $r_i \neq 0$ **do**
  $\alpha_i \leftarrow \dfrac{r_i^\top r_i}{r_i^\top Ar_i}$;
  $x_{i+1} \leftarrow x_i + \alpha_i r_i$;
  $r_{i+1} \leftarrow r_i - \alpha_i Ar_i$;
  $i \leftarrow i + 1$;
end while

For all $i \in \mathbb{N}$, $b - Ax_{i+1} = b - A(x_i + \alpha_i r_i) = r_i - \alpha_i Ar_i$.

Define, for all $x \in \mathbb{R}^n$, $\|x\|_A := \sqrt{x^\top Ax}$. Then, $\|\cdot\|_A$ is a norm on $\mathbb{R}^n$ and, for all $x \in \mathbb{R}^n$, $\sqrt{\lambda_{\min}(A)}\|x\|_2 \leq \|x\|_A \leq \sqrt{\lambda_{\max}(A)}\|x\|_2$.

Theorem 3.2.2 ([NW06, Theorem 3.3]). *For every $x_0 \in \mathbb{R}^n$, steepest descent with exact line search for ϕ generates a sequence $(x_i)_{i\in\mathbb{N}}$ such that, for all $i \in \mathbb{N}$,*

$$\|x_i - x\|_A \leq \left(\frac{\kappa_2(A) - 1}{\kappa_2(A) + 1}\right)^i \|x_0 - x\|_A.$$

Thus, for all $i \in \mathbb{N}$,

$$\|x_i - x\|_2 \leq \sqrt{\kappa_2(A)} \left(\frac{\kappa_2(A) - 1}{\kappa_2(A) + 1}\right)^i \|x_0 - x\|_2.$$

Preconditioning enables to improve convergence. Recall that, for all symmetric positive-definite $P, Q \in \mathbb{R}^{n\times n}$, the eigenvalues of PQ are real and positive (because $PQ = P^{\frac{1}{2}}(P^{\frac{1}{2}}QP^{\frac{1}{2}})P^{-\frac{1}{2}}$; see Sections A.1–A.2), although PQ is not necessarily symmetric. If it is possible to find a symmetric positive-definite $M \in \mathbb{R}^{n\times n}$ such that M is easy to invert and $\lambda_{\max}(M^{-1}A)/\lambda_{\min}(M^{-1}A)$ is smaller than $\kappa_2(A)$, then the following can be considered.

Algorithm 3.2 Preconditioned steepest descent with preconditioner M

Require: (A, M, x_0), where $A, M \in \mathbb{R}^{n\times n}$ are symmetric positive-definite and $x_0 \in \mathbb{R}^n$.

$i \leftarrow 0$; $r_0 \leftarrow b - Ax_0$;
while $r_i \neq 0$ **do**
  find $z \in \mathbb{R}^n$ such that $Mz = r_i$;
  $\alpha_i \leftarrow \dfrac{z^\top r_i}{z^\top Az}$;
  $x_{i+1} \leftarrow x_i + \alpha_i z$;
  $r_{i+1} \leftarrow r_i - \alpha_i Az$;
  $i \leftarrow i + 1$;
end while

Theorem 3.2.3 ([NW06, Theorem 3.3]). *For every $x_0 \in \mathbb{R}^n$, preconditioned steepest descent with preconditioner M generates a sequence $(x_i)_{i\in\mathbb{N}}$ such that, for all $i \in \mathbb{N}$,*

$$\|x_i - x\|_A \leq \left(\frac{\lambda_{\max}(M^{-1}A) - \lambda_{\min}(M^{-1}A)}{\lambda_{\max}(M^{-1}A) + \lambda_{\min}(M^{-1}A)}\right)^i \|x_0 - x\|_A.$$

Proof. Preconditioned steepest descent with preconditioner M is steepest descent with exact line search for $\hat{\phi} : \mathbb{R}^n \to \mathbb{R} : x \mapsto \frac{1}{2}x^\top \hat{A}x - \hat{b}^\top x$, where $\hat{A} \coloneqq M^{-\frac{1}{2}}AM^{-\frac{1}{2}}$ and $\hat{b} \coloneqq M^{-\frac{1}{2}}b$. □

Choosing a descent direction different from the negative gradient can also improve convergence.

Algorithm 3.3 Conjugate gradient method for ϕ

Require: (A, x_0), where $A \in \mathbb{R}^{n\times n}$ is symmetric positive-definite and $x_0 \in \mathbb{R}^n$.

$i \leftarrow 0$; $r_0 \leftarrow b - Ax_0$; $p_0 \leftarrow r_0$;
while $r_i \neq 0$ **do**
  $\alpha_i \leftarrow \dfrac{r_i^\top r_i}{p_i^\top Ap_i}$;
  $x_{i+1} \leftarrow x_i + \alpha_i p_i$;
  $r_{i+1} \leftarrow r_i - \alpha_i Ap_i$;
  $\beta_{i+1} \leftarrow \dfrac{{r_{i+1}}^\top r_{i+1}}{r_i^\top r_i}$;
  $p_{i+1} \leftarrow r_{i+1} + \beta_{i+1} p_i$;
  $i \leftarrow i + 1$;
end while

For all $i \in \mathbb{N}$, $\nabla\phi(x_i)^\top p_i = -r_i^\top p_i = -\|r_i\|_2^2$. Thus, the step size satisfies (3.1), i.e., corresponds to an exact line search.

Theorem 3.2.4 ([NW06, Theorem 5.3 and (5.36)])**.** *For every $x_0 \in \mathbb{R}^n$, the conjugate gradient method generates at most n iterates, which satisfy*

$$\|x_i - x\|_A \leq 2 \left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^i \|x_0 - x\|_A.$$

Chapter 4

Curve fitting

Given $n \in \mathbb{N} \setminus \{0\}$ and $(x_0, y_0), \dots, (x_n, y_n) \in \mathbb{R}^2$ such that $x_i < x_{i+1}$ for all $i \in \{0, \dots, n-1\}$, find a continuous function $f : \mathbb{R} \to \mathbb{R}$ such that, for all $i \in \{0, \dots, n\}$, $f(x_i) = y_i$ (interpolation) or $f(x_i) \approx y_i$ (approximation). Applications include resampling, noise reduction, and numerical integration.

4.1 Polynomial interpolation

Theorem 4.1.1. *There exists a unique polynomial function $p : \mathbb{R} \to \mathbb{R}$ of degree at most n such that $p(x_i) = y_i$ for all $i \in \{0, \dots, n\}$.*

Proof. Existence. For every $i \in \{0, \dots, n\}$, define the ith Lagrange polynomial function

$$L_i : \mathbb{R} \to \mathbb{R} : x \mapsto \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}, \tag{4.1}$$

which has degree n. For all $i, j \in \{0, \dots, n\}$, $L_i(x_j) = \delta_{i,j}$. Thus, $p \coloneqq \sum_{i=0}^{n} y_i L_i$ is a polynomial function of degree at most n such that $p(x_i) = y_i$ for all $i \in \{0, \dots, n\}$.

Uniqueness. Let $p : \mathbb{R} \to \mathbb{R}$ and $q : \mathbb{R} \to \mathbb{R}$ be two polynomial functions of degree at most n such that, for all $i \in \{0, \dots, n\}$, $p(x_i) = y_i = q(x_i)$. Then, $p - q$ is a polynomial function of degree at most n that has $n + 1$ zeros. Thus, $p - q$ is the zero function. □

Remark 4.1.2.

1. Theorem 4.1.1 is true on all fields, not only $\mathbb{R}$.
2. Theorem 4.1.1 implies that $\sum_{i=0}^{n} L_i = 1$. Indeed, if $y_i = 1$ for all $i \in \{0, \dots, n\}$, then $p = 1$.
3. If $p(x) = \sum_{i=0}^{n} a_i x^i$ and $a \coloneqq [a_0 \ \cdots \ a_n]^\top$, then $a = V^{-1} y$, where $V \coloneqq \left[x_i^j\right]_{i,j=0}^{n}$ is called the *Vandermonde matrix* and $y \coloneqq [y_0 \ \cdots \ y_n]^\top$. However, V tends to quickly get ill conditioned as n gets large. Under some conditions on the interpolation abscissae, the ∞-condition number of the Vandermonde matrix has a lower bound that grows exponentially with n [GI87]. The `polyfit` function must be used with care.
4. For practical computation, formulas other than $\sum_{i=0}^{n} y_i L_i$ are preferred.
   - Neville's algorithm [SB02, §2.1.2] is suitable to evaluate the interpolation polynomial at a single $x \in \mathbb{R}$, but less suited to determine the interpolation polynomial itself.
   - Newton's interpolation formula (see [SB02, §2.1.3] or [QSS07, §8.2]) gives the interpolation polynomial within fewer arithmetic operations than the formula $\sum_{i=0}^{n} y_i L_i$.
   - Barycentric interpolation formulas [QSS07, §8.3] are more stable than Newton's interpolation formula and have a similar computational cost.

The next theorem enables to estimate the interpolation error in the case where $(x_0, y_0), \dots, (x_n, y_n)$ belong to the graph of a sufficiently smooth function.

Theorem 4.1.3 ([SM03, Theorem 6.2])**.** *Let $a, b \in \mathbb{R}$ such that $a \leq x_0$ and $x_n \leq b$. Let $f \in \mathrm{C}^{n+1}([a,b],\mathbb{R})$. Let $p : \mathbb{R} \to \mathbb{R}$ be the polynomial function of degree at most n such that $p(x_i) = f(x_i)$ for all $i \in \{0,\dots,n\}$. Then, for every $x \in [a,b]$, there exists $c \in (a,b)$ such that*

$$f(x) = p(x) + \frac{f^{(n+1)}(c)}{(n+1)!}\prod_{i=0}^{n}(x - x_i).$$

The interpolation polynomial does not necessarily converge to f as n tends to infinity.

Example 4.1.4 (Runge's function)**.** Let $f : \mathbb{R} \to \mathbb{R} : x \mapsto \frac{1}{1+x^2}$, $a \coloneqq -5$, and $b \coloneqq 5$. Define equispaced interpolation abscissae in $[a,b]$: $x_i \coloneqq a + \frac{b-a}{n} i$ for all $i \in \{0,\dots,n\}$. Let p_n be the polynomial function of degree at most n such that $p_n(x_i) = f(x_i)$ for all $i \in \{0,\dots,n\}$. Then, $\lim_{n\to\infty} \|f - p_n\|_\infty = \infty$ (see Section B.2 for the definition of $\|\cdot\|_\infty$). The graph of p_n oscillates strongly near a and b. This is known as Runge's phenomenon.

To minimize $\|f-p\|_\infty$, it is tempting to choose the interpolation abscissae $x_0,\dots,x_n$ that minimize

$$\max_{x\in[a,b]} \prod_{i=0}^{n} |x - x_i|.$$

Exercise 4.1.5. Define equispaced interpolation abscissae in $[a,b]$: $x_i \coloneqq a + \frac{b-a}{n} i$ for all $i \in \{0,\dots,n\}$. Prove that, for all $x \in [a,b]$,

$$\prod_{i=0}^{n} |x - x_i| \leq \frac{n!}{4}\left(\frac{b-a}{n}\right)^{n+1}.$$

Theorem 4.1.6 ([SM03, Theorem 8.7])**.** *The abscissae $x_0,\dots,x_n$ that minimize*

$$\max_{x\in[a,b]} \prod_{i=0}^{n} |x - x_i|$$

are $x_i \coloneqq \frac{a+b}{2} - \frac{b-a}{2}\cos\left(\frac{(i+\frac{1}{2})\pi}{n+1}\right)$ for all $i \in \{0,\dots,n\}$, which are called the Chebyshev abscissae (of the first kind) *and represented in Figure 4.1. For those abscissae,*

$$\max_{x\in[a,b]} \prod_{i=0}^{n} |x - x_i| = \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

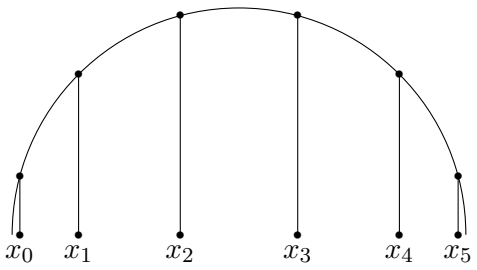

Figure 4.1: Chebyshev abscissae (of the first kind) for $n = 5$. Higher concentration of abscissae at the extremeties of the interpolation interval.

With those abscissae, for Runge's function (Example 4.1.4), $\lim_{n\to\infty} \|f-p_n\|_\infty = 0$ [Tre19, Chap. 13].

If the abscissae are imposed, other remedies must be found. Alternatives include spline interpolation and polynomial approximation. Before presenting these alternatives, we analyze the sensitivity of polynomial interpolation. Define the *Lebesgue constant*

$$\Lambda_n \coloneqq \max_{x\in[a,b]} \sum_{i=0}^{n} |L_i(x)|.$$

Let $\tilde{y}_0, \dots, \tilde{y}_n \in \mathbb{R}$ and $\tilde{y} \coloneqq [\tilde{y}_0 \ \cdots \ \tilde{y}_n]^\top$. Then, for all $x \in [a, b]$,

$$\left|\sum_{i=0}^{n} y_i L_i(x) - \sum_{i=0}^{n} \tilde{y}_i L_i(x)\right| = \left|\sum_{i=0}^{n} (y_i - \tilde{y}_i) L_i(x)\right| \leq \sum_{i=0}^{n} |y_i - \tilde{y}_i||L_i(x)| \leq \Lambda_n \|y - \tilde{y}\|_\infty.$$

Polynomial interpolation is well conditioned if Λ_n is small, and ill conditioned if Λ_n is large.

If $a = -1$ and $b = 1$, then:

- $\Lambda_n \geq \frac{2}{\pi} \ln(n+1) + 0.53$ [Bru78, (32)];
- for Chebyshev abscissae, $\Lambda_n \leq \frac{2}{\pi} \ln(n+1) + 1$ [Riv74];
- for equispaced abscissae, $\frac{2^{n-2}}{n^2} < \Lambda_n < \frac{2^{n+3}}{n}$ [TW91, Theorem 2].

4.2 Spline interpolation

A spline is a piecewise polynomial function with a certain degree of smoothness. Specifically, a *spline* of degree $k \in \mathbb{N} \setminus \{0\}$ with abscissae (or *knots*) $x_0, \dots, x_n$ is a function $s \in \mathrm{C}^{k-1}([x_0, x_n], \mathbb{R})$ such that, for every $i \in \{0, \dots, n-1\}$, $s|_{[x_i, x_{i+1}]}$ is a polynomial function of degre at most k [QSS07, Definition 8.1].

Exercise 4.2.1. Prove that the set $\mathcal{S}_k(x_0, \dots, x_n)$ of all splines of degree $k \in \mathbb{N} \setminus \{0\}$ with knots $x_0, \dots, x_n$ is a linear subspace of $\mathrm{C}^{k-1}([x_0, x_n], \mathbb{R})$. What is its dimension?

Sections 4.2.1 and 4.2.2 consider respectively $k = 1$ and $k = 3$.

4.2.1 Linear spline interpolation

The unique $s \in \mathcal{S}_1(x_0, \dots, x_n)$ such that $s(x_i) = y_i$ for all $i \in \{0, \dots, n\}$ satisfies, for all $i \in \{0, \dots, n-1\}$ and $x \in [x_i, x_{i+1}]$,

$$s(x) = \frac{x - x_{i+1}}{x_i - x_{i+1}} y_i + \frac{x - x_i}{x_{i+1} - x_i} y_{i+1} = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i).$$

Interpolation error Assume that, for all $i \in \{0, \dots, n\}$, $y_i = f(x_i)$ with $f \in \mathrm{C}^2([x_0, x_n], \mathbb{R})$. By Theorem 4.1.3, for every $i \in \{0, \dots, n-1\}$ and $x \in (x_i, x_{i+1})$, there exists $c_i \in (x_i, x_{i+1})$ such that

$$f(x) = s(x) + \frac{f''(c_i)}{2}(x - x_i)(x - x_{i+1}).$$

For every $i \in \{0, \dots, n-1\}$ and $x \in (x_i, x_{i+1})$, $0 < (x - x_i)(x_{i+1} - x) \leq (x_{i+1} - x_i)^2/4$. Thus,

$$\|f - s\|_\infty \leq \frac{\|f''\|_\infty}{8} \max_{i \in \{0, \dots, n-1\}} (x_{i+1} - x_i)^2.$$

Sensitivity Let $\tilde{y}_0, \dots, \tilde{y}_n \in \mathbb{R}$ and $\tilde{y} \coloneqq [\tilde{y}_0 \ \cdots \ \tilde{y}_n]^\top$. Let $s, \tilde{s} \in \mathcal{S}_1(x_0, \dots, x_n)$ such that, for all $i \in \{0, \dots, n\}$, $s(x_i) = y_i$ and $\tilde{s}(x_i) = \tilde{y}_i$. For every $i \in \{0, \dots, n-1\}$ and $x \in [x_i, x_{i+1}]$,

$$\begin{aligned} |s(x) - \tilde{s}(x)| &= \left|\frac{x - x_{i+1}}{x_i - x_{i+1}}(y_i - \tilde{y}_i) + \frac{x - x_i}{x_{i+1} - x_i}(y_{i+1} - \tilde{y}_{i+1})\right| \\ &\leq \left|\frac{x - x_{i+1}}{x_i - x_{i+1}}\right| |y_i - \tilde{y}_i| + \left|\frac{x - x_i}{x_{i+1} - x_i}\right| |y_{i+1} - \tilde{y}_{i+1}| \\ &\leq \|y - \tilde{y}\|_\infty, \end{aligned}$$

where the last inequality holds because

$$\left|\frac{x - x_{i+1}}{x_i - x_{i+1}}\right| + \left|\frac{x - x_i}{x_{i+1} - x_i}\right| = \frac{(x_{i+1} - x) + (x - x_i)}{x_{i+1} - x_i} = 1.$$

4.2.2 Cubic spline interpolation

Find $s \in \mathcal{S}_3(x_0, \dots, x_n)$ such that $s(x_i) = y_i$ for all $i \in \{0, \dots, n\}$. There are $4n$ degrees of freedom and $2n + 2(n-1) = 4n - 2$ conditions. Four popular pairs of additional conditions:

- $s''(x_0) = 0 = s''(x_n)$ (*natural* spline);
- prescribe $s'(x_0)$ and $s'(x_n)$;
- continuity of s''' at x_1 and x_{n-1} (*not-a-knot condition*);
- $s'(x_0) = s'(x_n)$ and $s''(x_0) = s''(x_n)$, especially relevant if $y_0 = y_n$ (*periodic* spline).

Computing s amounts to solving a tridiagonal linear system [QSS07, §8.7.1].

Theorem 4.2.2 ([QSS07, Property 8.3])**.** *Let $f \in \mathrm{C}^4([x_0, x_n], \mathbb{R})$ and $s \in \mathcal{S}_3(x_0, \dots, x_n)$ such that $s(x_i) = f(x_i)$ for all $i \in \{0, \dots, n\}$. Define $c_0 \coloneqq \frac{5}{384}$, $c_1 \coloneqq \frac{1}{24}$, and $c_2 \coloneqq \frac{3}{8}$. Then, for every $j \in \{0, 1, 2\}$,*

$$\|f^{(j)} - s^{(j)}\|_\infty \leq c_j \|f^{(4)}\|_\infty \max_{i \in \{1, \dots, n\}} (x_i - x_{i-1})^{4-j}.$$

4.3 Least-squares polynomial approximation

Find a polynomial function $p : \mathbb{R} \to \mathbb{R}$ of degree at most $m \in \mathbb{N}$ that minimizes

$$\sum_{i=0}^{n} (p(x_i) - y_i)^2.$$

There exist $a_0, \dots, a_m \in \mathbb{R}$ such that $p(x) = \sum_{j=0}^{m} a_j x^j$ for all $x \in \mathbb{R}$. Thus, the problem can be rewritten as

$$\min_{a_0, \dots, a_m \in \mathbb{R}} \sum_{i=0}^{n} \left(\sum_{j=0}^{m} a_j x_i^j - y_i \right)^2.$$

Define $V \coloneqq \left[x_i^j \right]_{i,j=0}^{n,m}$, $y \coloneqq [y_0 \ \cdots \ y_n]^\top$, and $a \coloneqq [a_0 \ \cdots \ a_m]^\top$. Then, the problem can be rewritten as

$$\min_{a \in \mathbb{R}^{m+1}} \|Va - y\|_2^2.$$

Unique solution if $m \leq n$ because, in that case, all order-$(m+1)$ submatrices of V are invertible by Theorem 4.1.1.

Chapter 5

Numerical differentiation

Given $f : \mathbb{R} \to \mathbb{R}$ and $x \in \operatorname{dom} f$, estimate $f'(x)$ or $f''(x)$ by relying only on arithmetic operations and evaluations of f. Application: numerical methods for differential equations.

5.1 Finite differences

Since

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$

a natural estimation of $f'(x)$ is

$$\frac{f(x+h) - f(x)}{h} \tag{5.1}$$

for some $h \in \mathbb{R} \setminus \{0\}$. This is called the *forward* (resp. *backward*) *finite-difference formula* if $h > 0$ (resp. $h < 0$). If f is twice continuously differentiable between x and $x + h$, then Taylor's theorem (Theorem B.3.4) ensures the existence of a real number a between x and $x + h$ such that

$$f(x+h) = f(x) + f'(x)h + \frac{f''(a)h^2}{2}.$$

Thus,

$$\left| \frac{f(x+h) - f(x)}{h} - f'(x) \right| = \frac{|f''(a)|}{2}|h|.$$

Formula (5.1) is said to be a first-order formula.

A second-order formula can be obtained as follows. Let $h \in (0, \infty)$. If $f \in \mathrm{C}^3([x-h, x+h], \mathbb{R})$, then Taylor's theorem (Theorem B.3.4) ensures the existence of $a \in (x, x+h)$ and $b \in (x-h, x)$ such that

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)h^2}{2} + \frac{f'''(a)h^3}{6},$$
$$f(x-h) = f(x) - f'(x)h + \frac{f''(x)h^2}{2} - \frac{f'''(b)h^3}{6}.$$

Thus,

$$\left| \frac{f(x+h) - f(x-h)}{2h} - f'(x) \right| = \frac{|f'''(a) + f'''(b)|}{12} h^2.$$

By the intermediate-value theorem, there exists $c \in [b, a]$ such that $f'''(c) = (f'''(a) + f'''(b))/2$. Therefore,

$$\left| \frac{f(x+h) - f(x-h)}{2h} - f'(x) \right| = \frac{|f'''(c)|}{6} h^2.$$

This shows that

$$\frac{f(x+h) - f(x-h)}{2h}, \tag{5.2}$$

called the *central finite-difference formula*, is a second-order formula.

The analysis of the central finite-difference formula suggests the following estimation of $f''(x)$:

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} \tag{5.3}$$

with $h \in (0, \infty)$. If $f \in \mathrm{C}^4([x-h, x+h], \mathbb{R})$, then Taylor's theorem (Theorem B.3.4) ensures the existence of $a \in (x, x+h)$ and $b \in (x-h, x)$ such that

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)h^2}{2} + \frac{f'''(x)h^3}{6} + \frac{f^{(4)}(a)h^4}{24},$$
$$f(x-h) = f(x) - f'(x)h + \frac{f''(x)h^2}{2} - \frac{f'''(x)h^3}{6} + \frac{f^{(4)}(b)h^4}{24}.$$

Thus,

$$\left|\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - f''(x)\right| = \frac{|f^{(4)}(a) + f^{(4)}(b)|}{24}h^2.$$

By the intermediate-value theorem, there exists $c \in [b, a]$ such that $f^{(4)}(c) = (f^{(4)}(a) + f^{(4)}(b))/2$. Therefore,

$$\left|\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - f''(x)\right| = \frac{|f^{(4)}(c)|}{12}h^2.$$

This is a second-order formula.

Exercise 5.1.1. Prove that (5.2) and (5.3) converge respectively to $f'(x)$ and $f''(x)$ as h goes to 0.

Everything in this section can be extended to functions from $\mathbb{R}$ to $\mathbb{R}^n$ (use Theorem B.3.3 instead of Theorem B.3.4).

5.2 Effect of round-off errors

In floating-point arithmetic, formulas such as (5.1)–(5.3) must be used with care. For example, in Python, the lines

```
import numpy
h = 0.1**numpy.arange(1, 20, 1)
Dh = numpy.log(1+h)*1./h
print(Dh)
```

generate Table 5.1.

Assume that x, h, and $x+h$ are floating-point numbers. In floating-point arithmetic, there exists $\varepsilon \in (0, \infty)$ such that $f(x)$ and $f(x+h)$ are respectively computed as $f(x)(1+\varepsilon_1)$ and $f(x+h)(1+\varepsilon_2)$ with $\varepsilon_1, \varepsilon_2 \in [-\varepsilon, \varepsilon]$. If, between x and $x+h$, f is twice continuously differentiable and $|f|$ and $|f''|$ are upper-bounded respectively by c_0 and c_2, then the error associated with (5.1) becomes

$$\begin{aligned}
\left|\frac{f(x+h)(1+\varepsilon_2) - f(x)(1+\varepsilon_1)}{h} - f'(x)\right| &= \left|\frac{f(x+h) - f(x)}{h} - f'(x) + \frac{f(x+h)\varepsilon_2 - f(x)\varepsilon_1}{h}\right| \\
&= \left|\frac{f''(a)}{2}h + \frac{f(x+h)\varepsilon_2 - f(x)\varepsilon_1}{h}\right| \\
&\leq \frac{|f''(a)|}{2}|h| + \frac{|f(x+h)| + |f(x)|}{|h|}\varepsilon \\
&\leq \frac{c_2}{2}|h| + \frac{2c_0}{|h|}\varepsilon,
\end{aligned}$$

which is minimum if

$$|h| = 2\sqrt{\frac{c_0\varepsilon}{c_2}}.$$

If $c_0 \approx c_2$ and $\varepsilon \approx 10^{-16}$, then the minimizer is close to 10^{-8}, in agreement with the empirical observation made in Table 5.1.

Exercise 5.2.1. Repeat the preceding analysis for (5.2) and (5.3).

i	$\frac{\ln(1+10^{-i})}{10^{-i}}$
1	0.9531018
2	0.99503309
3	0.99950033
4	0.99995
5	0.999995
6	0.9999995
7	0.99999995
8	0.99999999
9	1.00000008
10	1.00000008
11	1.00000008
12	1.0000889
13	0.99920072
14	0.99920072
15	1.11022302
16	0
17	0
18	0
19	0

Table 5.1: Forward finite-difference formula applied to ln at 1 in Python. The estimate that is closest to $\ln' 1 = 1$ is the eighth one.

Chapter 6

Numerical integration

Given $a, b \in \mathbb{R}$, $a < b$, and $f : [a, b] \to \mathbb{R}$ continuous, estimate $\int_a^b f$ by relying only on arithmetic operations and evaluations of f.

6.1 Newton–Cotes formulas

Given $n \in \mathbb{N} \setminus \{0\}$, estimate $\int_a^b f$ as $\int_a^b p_n$, where p_n is the polynomial function of degree at most n such that, for all $i \in \{0, \dots, n\}$, $p_n(x_i) = f(x_i)$ with $x_i \coloneqq a + \frac{b-a}{n} i$. With the Lagrange polynomial functions (4.1), $p_n = \sum_{i=0}^n f(x_i) L_i$.

Lemma 6.1.1. *If $b = -a$, then $L_{n-i}(x) = L_i(-x)$ for all $i \in \{0, \dots, n\}$ and $x \in \mathbb{R}$. If, moreover, f is odd (resp. even), then p_n is odd (resp. even).*

Proof. Let $b = -a$ and $x \in \mathbb{R}$. For all $i \in \{0, \dots, n\}$, $x_{n-i} = -x_i$ and

$$L_{n-i}(x) = \prod_{\substack{j=0 \\ j \neq n-i}}^{n} \frac{x - x_j}{x_{n-i} - x_j} = \prod_{\substack{j=0 \\ j \neq n-i}}^{n} \frac{-x - x_{n-j}}{x_i - x_{n-j}} = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{-x - x_j}{x_i - x_j} = L_i(-x).$$

If n is odd, then

$$\sum_{i=0}^{n} f(x_i) L_i(x) = \sum_{i=0}^{\frac{n-1}{2}} \left(f(x_i) L_i(x) + f(x_{n-i}) L_{n-i}(x) \right) = \sum_{i=0}^{\frac{n-1}{2}} \left(f(x_i) L_i(x) + f(-x_i) L_i(-x) \right).$$

If n is even, then

$$\begin{aligned} \sum_{i=0}^{n} f(x_i) L_i(x) &= f(x_{\frac{n}{2}}) L_{\frac{n}{2}}(x) + \sum_{i=0}^{\frac{n}{2}-1} \left(f(x_i) L_i(x) + f(x_{n-i}) L_{n-i}(x) \right) \\ &= f(0) L_{\frac{n}{2}}(x) + \sum_{i=0}^{\frac{n}{2}-1} \left(f(x_i) L_i(x) + f(-x_i) L_i(-x) \right). \end{aligned}$$

In both cases, p_n is odd (resp. even) if f is odd (resp. even). □

The bijection $\sigma : [-1, 1] \to [a, b] : t \mapsto \frac{b-a}{2} t + \frac{b+a}{2}$, whose inverse is $\sigma^{-1} : [a, b] \to [-1, 1] : x \mapsto \frac{2}{b-a} x - \frac{b+a}{b-a}$, enables to reduce the interval of integration to $[-1, 1]$: $\int_a^b f = \frac{b-a}{2} \int_{-1}^1 f \circ \sigma$. This makes the symmetry properties given in Lemma 6.1.1 applicable. For all $i \in \{0, \dots, n\}$, define

$$t_i \coloneqq \sigma^{-1}(x_i) = -1 + \frac{2}{n} i, \qquad w_i \coloneqq \frac{1}{2} \int_{-1}^1 L_i \circ \sigma = \frac{1}{2} \int_{-1}^1 \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{t - t_j}{t_i - t_j} \mathrm{d}t.$$

Proposition 6.1.2. *Given $n \in \mathbb{N} \setminus \{0\}$, let $p_n : \mathbb{R} \to \mathbb{R}$ be the polynomial function of degree at most n such that, for all $i \in \{0, \dots, n\}$, $p_n(x_i) = f(x_i)$ with $x_i \coloneqq a + \frac{b-a}{n} i$. Then:*

1. $\int_a^b p_n = (b-a)\sum_{i=0}^n w_i f(x_i)$ *(*Newton–Cotes formula *of order n);*
2. $\sum_{i=0}^n w_i = 1$*;*
3. *for all $i \in \{0,\dots,n\}$, $w_{n-i} = w_i$;*
4. *if f is a polynomial function of degree at most n, then $\int_a^b p_n = \int_a^b f$;*
5. *if n is even and f is a polynomial function of degree at most $n+1$, then $\int_a^b p_n = \int_a^b f$.*

Proof. 1. Since $p_n = \sum_{i=0}^n f(x_i)L_i$, $\int_a^b p_n = \sum_{i=0}^n f(x_i)\int_a^b L_i$. Moreover, for all $i \in \{0,\dots,n\}$, $\int_a^b L_i = \frac{b-a}{2}\int_{-1}^1 L_i \circ \sigma = (b-a)w_i$.

2. Apply the preceding point to $f : x \mapsto 1$, observing that, in that case, $p_n = f$.

3. By Lemma 6.1.1, the function $L_i \circ \sigma - L_{n-i} \circ \sigma$ is odd, hence its integral on $[-1,1]$ is zero.

4. If f is a polynomial function of degree at most n, then $p_n = f$.

5. Let n be even and f be a polynomial function of degree at most $n+1$. By the first point, the equality to be proven, $\int_a^b p_n = \int_a^b f$, is equivalent to $\frac{1}{2}\int_{-1}^1 f \circ \sigma = \sum_{i=0}^n w_i f(x_i)$. Since $f \circ \sigma$ is a polynomial function of degree at most $n+1$, it can be decomposed as $f \circ \sigma = q_n + q_{n+1}$ with q_n a polynomial function of degree at most n and q_{n+1} a multiple of the odd function $t \mapsto t^{n+1}$. By the preceding point, $\frac{1}{2}\int_{-1}^1 q_n = \sum_{i=0}^n w_i q_n(t_i)$. By Lemma 6.1.1, $\sum_{i=0}^n w_i q_{n+1}(t_i) = 0 = \int_{-1}^1 q_{n+1}$. □

The vector $[w_0 \ \cdots \ w_n]$ can be computed as the solution to a linear system. Properties 2 and 3 of Proposition 6.1.2 give respectively 1 and $\lceil \frac{n}{2} \rceil$ equations. Thus, only $n - \lceil \frac{n}{2} \rceil$ additional equations are needed. The vector $[w_0 \ \cdots \ w_n]$ is given in Table 6.1 for $n \in \{1,2\}$. For $n = 1$, no additional equation is needed: $w_0 + w_1 = 1$ and $w_1 = w_0$, hence the result in the table. For $n = 2$, one additional equation is needed: by properties 1 and 4 of Proposition 6.1.2, $\frac{1}{3} = \frac{1}{2}\int_{-1}^1 t^2 \mathrm{d}t = w_0 + w_2$, hence the result in the table.

$n = 1$ (trapezium rule)	$w_0 = \frac{1}{2} = w_1$
$n = 2$ (Simpson's rule)	$w_0 = \frac{1}{6} = w_2$, $w_1 = \frac{2}{3}$

Table 6.1: Newton–Cotes weights for $n \in \{1,2\}$.

The formula for $n = 1$ is called the trapezium rule because it estimates $\int_a^b f$ as the area of the trapezium defined by the points $(a,0)$, $(a,f(a))$, $(b,f(b))$, and $(b,0)$, as illustrated in Figure 6.1.

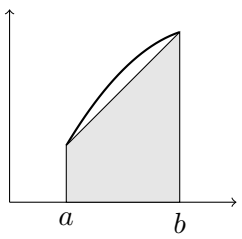

Figure 6.1: Illustration of the trapezium rule.

Error. By Theorem 4.1.3, if $f \in \mathrm{C}^{n+1}([a,b],\mathbb{R})$,

$$\left|\int_a^b f - \int_a^b p_n\right| = \left|\int_a^b (f - p_n)\right| \leq \int_a^b |f - p_n| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!}\int_a^b \prod_{i=0}^n |x - x_i| \mathrm{d}x.$$

Trapezium rule ($n = 1$):

$$\int_a^b |x-a||x-b|\mathrm{d}x = \int_a^b (x-a)(b-x)\mathrm{d}x = \frac{(b-a)^3}{6}.$$

Simpson's rule ($n = 2$):

$$\int_a^b |x-a|\left|x-\frac{a+b}{2}\right||x-b|\mathrm{d}x = \frac{(b-a)^4}{32}.$$

Theorem 6.1.3 ([SM03, Theorem 7.2])**.** *If $f \in \mathrm{C}^4([a,b],\mathbb{R})$, then there exists $c \in (a,b)$ such that*

$$\int_a^b f = (b-a)\left(\frac{f(a)}{6} + \frac{2f(\frac{a+b}{2})}{3} + \frac{f(b)}{6}\right) - \frac{(b-a)^5}{2880}f^{(4)}(c).$$

6.2 Composite formulas

In general, Newton–Cotes formulas do not work well for large n; remember the behavior of p_n for Runge's function (Example 4.1.4). To increase accuracy, composite formulas are a preferable alternative. Given $n \in \mathbb{N} \setminus \{0,1\}$, define $h \coloneqq \frac{b-a}{n}$ and $x_i \coloneqq a + hi$ for all $i \in \{0,\dots,n\}$. Based on the identity $\int_a^b f = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f$, each composite formulas applies a Newton–Cotes formula to $\int_{x_i}^{x_{i+1}} f$ for all $i \in \{0,\dots,n-1\}$.

- Composite trapezium rule:

$$T(n) \coloneqq h\left(\frac{f(a)+f(b)}{2} + \sum_{i=1}^{n-1} f(x_i)\right).$$

- Composite Simpson rule:

$$S(n) \coloneqq \frac{h}{3}\left(\frac{f(a)+f(b)}{2} + \sum_{i=1}^{n-1} f(x_i) + 2\sum_{i=0}^{n-1} f\left(\frac{x_i+x_{i+1}}{2}\right)\right).$$

Error.

- Composite trapezium rule with $f \in \mathrm{C}^2([a,b],\mathbb{R})$:

$$\left|\int_a^b f - T(n)\right| \le \frac{b-a}{12}\|f''\|_\infty h^2.$$

- Composite Simpson rule with $f \in \mathrm{C}^4([a,b],\mathbb{R})$:

$$\left|\int_a^b f - S(n)\right| \le \frac{b-a}{2880}\|f^{(4)}\|_\infty h^4.$$

These rules are said to be of order 2 and 4, respectively. For both rules, the error can be made arbitrarily small by choosing n sufficiently large.

6.3 Extrapolation methods

Theorem 6.3.1 (Euler–Maclaurin expansion [SM03, Theorem 7.4])**.** *Let $k \in \mathbb{N} \setminus \{0,1\}$. If $f \in \mathrm{C}^{2k}([a,b],\mathbb{R})$, then there exist $c_1,\dots,c_{k-1} \in \mathbb{R}$ such that*

$$\int_a^b f = T(n) + \sum_{i=1}^{k-1} c_i h^{2i} + \mathrm{O}(h^{2k}).$$

Proposition 6.3.2 (Romberg integration method)**.** *Define* $T_0(n) \coloneqq T(n)$ *and, for all* $k \in \mathbb{N}$*,*

$$T_{k+1}(n) \coloneqq \frac{4^{k+1}T_k(2n) - T_k(n)}{4^{k+1} - 1}.$$

For every $k \in \mathbb{N} \setminus \{0, 1\}$*, if* $f \in \mathrm{C}^{2k}([a, b], \mathbb{R})$*, then*

$$\int_a^b f = T_{k-1}(n) + \mathrm{O}(h^{2k}).$$

Proof. Let $k \in \mathbb{N}\setminus\{0, 1\}$ and $f \in \mathrm{C}^{2k}([a, b], \mathbb{R})$. Let us prove by induction that, for all $j \in \{0, \dots, k-1\}$,

$$\int_a^b f = T_j(n) + \frac{(-1)^j}{\prod_{l=1}^{j}(4^l - 1)} \sum_{i=j+1}^{k-1} c_i \left(\prod_{l=1}^{j} \left(1 - 4^{l-i}\right) \right) h^{2i} + \mathrm{O}(h^{2k}). \tag{6.1}$$

The equality to be proven is (6.1) with $j = k - 1$. By Theorem 6.3.1, (6.1) is true if $j = 0$. Assuming that (6.1) is true for some $j \in \{0, \dots, k - 2\}$, we now prove that it is true for $j + 1$: since

$$\int_a^b f = T_j(2n) + \frac{(-1)^j}{\prod_{l=1}^{j}(4^l - 1)} \sum_{i=j+1}^{k-1} c_i \left(\prod_{l=1}^{j} \left(1 - 4^{l-i}\right) \right) 4^{-i} h^{2i} + \mathrm{O}(h^{2k}), \tag{6.2}$$

a straightforward computation shows that (6.1) $- 4^{j+1}$(6.2) is equivalent to

$$\left(1 - 4^{j+1}\right) \int_a^b f = T_j(n) - 4^{j+1}T_j(2n) + \frac{(-1)^j}{\prod_{l=1}^{j}(4^l - 1)} \sum_{i=j+2}^{k-1} c_i \left(\prod_{l=1}^{j} \left(1 - 4^{l-i}\right) \right) \left(1 - 4^{j+1-i}\right) h^{2i} + \mathrm{O}(h^{2k})$$

or

$$\int_a^b f = T_{j+1}(n) + \frac{(-1)^{j+1}}{\prod_{l=1}^{j+1}(4^l - 1)} \sum_{i=j+2}^{k-1} c_i \left(\prod_{l=1}^{j+1} \left(1 - 4^{l-i}\right) \right) h^{2i} + \mathrm{O}(h^{2k}). \qquad \square$$

Half the evaluations of f required to compute $T(2n)$ are already known from the computation of $T(n)$.

Chapter 7

Differential equations

A differential equation is an equation that relates a function to its derivatives. Differential equations appeared to model the temporal or spatial evolution of physical systems. In applications, they always come with additional conditions that express constraints on the state of the system, at initial time (initial condition) for a temporal evolution and on the boundary of the domain (boundary condition) for a spatial evolution, and that bring important mathematical properties such as uniqueness of solutions. In general, a differential equation without an additional condition cannot have a unique solution since a constant of integration appears in the calculation of a solution. In physics, it is generally impossible to predict the evolution of a system without knowing its initial state.

With or without additional conditions, it is generally impossible to write explicit formulas for the solutions to a differential equation. This led to the development of numerical methods, describing quantitatively the solutions, and a theory, describing their qualitative properties such as existence, uniqueness, and dependence on initial or boundary conditions.

In every modeling exercise, the model must be validated before being used to study the phenomenon that it is supposed to represent. For a differential equation with an additional condition, the validation generally involves a both qualitative and quantitative study of its solutions. Possessing properties such as uniqueness can contribute to the validation.

There are several ways of classifying differential equations. One of the most important is based on the number of real variables on which the unknown function depends. An ordinary differential equation (ODE) is a differential equation whose unknown is a function of a real variable. A partial differential equation (PDE) is a differential equation whose unknown is a function of several real variables. An ODE with an initial condition is called an initial-value problem. A differential equation with a boundary condition is called a boundary-value problem.

7.1 Initial-value problems

Contents: theoretical foundations in Section 7.1.1, four basic one-step methods in Section 7.1.2, error analysis in Section 7.1.3, absolute stability in Section 7.1.4, and further topics in Section 7.1.5. In this section, $\|\cdot\|$ denotes both a norm on $\mathbb{R}^n$ and the induced norm on $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ (see Section A.3 and Proposition B.2.6). The norm of every $(t, x) \in \mathbb{R} \times \mathbb{R}^n$ is defined as $\max\{|t|, \|x\|\}$.

7.1.1 Theoretical foundations

This section provides basic definitions and results from the theory of ODEs. Let $n \in \mathbb{N} \setminus \{0\}$. A function from $\mathbb{R} \times \mathbb{R}^n$ to $\mathbb{R}^n$ is called a *vector field*. Every vector field defines an ODE, the solution to which is defined as an integral curve of the vector field.

Definition 7.1.1 (integral curve)**.** A function $u : \mathbb{R} \to \mathbb{R}^n$ is called an *integral curve* of $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ if $\operatorname{dom} u$ is an interval, u is differentiable, and, for all $t \in \operatorname{dom} u$, $(t, u(t)) \in \operatorname{dom} f$ and $u'(t) = f(t, u(t))$. The interval $\operatorname{dom} u$ is called the *interval of existence* of u.

Definition 7.1.2 (initial-value problem)**.** Given $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ and $(t_0, u_0) \in \operatorname{dom} f$, the triplet (f, t_0, u_0) is called an *initial-value problem (IVP)*. A solution to the IVP (f, t_0, u_0) is an integral curve u of f such that $t_0 \in \operatorname{dom} u$ and $u(t_0) = u_0$.

An integral curve is always defined on an interval for at least two reasons. First, from a modeling point of view, obeying a differential equation on a set that is not connected makes little physical sense. Second, from the mathematical point of view, it is a necessary condition to make the uniqueness of solutions to the IVP possible. It is sometimes desirable to find the largest interval possible, leading to the following.

Definition 7.1.3 (maximal integral curve)**.** Let $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. An integral curve v of f is said to be an *extension* of an integral curve u of f if $\operatorname{dom} u \subsetneq \operatorname{dom} v$ and, for all $t \in \operatorname{dom} u$, $u(t) = v(t)$. An integral curve of f is said to be *maximal* if it has no extension.

The maximal length of an interval of existence is an unknown of the IVP that depends on the constraint stating that the graph of an integral curve is in the domain of the vector field.

Two basic results are established in the next section. Then, the questions of existence, maximality, uniqueness, and asymptotic behavior of solutions are addressed.

Integral equation and regularity

In general, a vector field admits an integral curve in the sense of Definition 7.1.1 only if it is continuous. In that case, it is a direct consequence of the fundamental theorem of calculus (Theorem B.4.1) that the IVP can be rewritten as an integral equation, which is crucial from both the theoretical and the numerical point of view. The integral of a vector-valued function of a real variable is defined componentwise (see Section B.4).

Proposition 7.1.4. *Let (f, t_0, u_0) be an IVP and $u : \mathbb{R} \to \mathbb{R}^n$ be a continuous function whose domain is an interval that contains t_0. If f is continuous, then u is a solution to the IVP if and only if its graph is contained in* $\operatorname{dom} f$ *and it satisfies the integral equation*

$$u(t) = u_0 + \int_{t_0}^{t} f(s, u(s)) \, \mathrm{d}s$$

for all $t \in \operatorname{dom} u$; moreover, in that case, u' is continuous.

Proof. Since f and u are continuous, the function $\mathbb{R} \to \mathbb{R}^n : t \mapsto f(t, u(t))$ is continuous. Let u be a solution to the IVP. Then, u' is continuous. Thus, by the fundamental theorem of calculus, for all $t \in \operatorname{dom} u$,

$$u(t) = u(t_0) + \int_{t_0}^{t} u'(s) \, \mathrm{d}s = u_0 + \int_{t_0}^{t} f(s, u(s)) \, \mathrm{d}s.$$

Conversely, assume that the graph of u is in $\operatorname{dom} f$ and the integral equation is satisfied. Evaluating this equation at $t = t_0$ yields $u(t_0) = u_0$. Furthermore, by the fundamental theorem of calculus, u is differentiable and, for all $t \in \operatorname{dom} u$, $u'(t) = f(t, u(t))$. Moreover, u' is continuous. □

Proposition 7.1.5. *Let $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ and $k \in \mathbb{N}$. If* $\operatorname{dom} f$ *is open and f is k times differentiable, then every integral curve of f is $k + 1$ times differentiable.*

Proof. Let $\operatorname{dom} f$ be open, f be k times differentiable, and u be an integral curve of f. Let us prove by induction that u is $k + 1$ times differentiable. By definition of integral curve, u is differentiable and, for all $t \in \operatorname{dom} u$, $(t, u(t)) \in \operatorname{dom} f$ and $u'(t) = f(t, u(t))$. Assume that, for some $i \in \{0, \ldots, k\}$, u is i times differentiable. Since f is i times differentiable, by the chain rule (Theorem B.3.1), u' is i times differentiable, i.e., u is $i + 1$ times differentiable. □

Local existence

The next theorem is a quantitative and slightly more general version of the following: given an IVP (f, t_0, u_0), if (t_0, u_0) is in the interior of $\operatorname{dom} f$ and f is continuous, then there exists a solution defined on a compact interval whose interior contains t_0.

Theorem 7.1.6 (local existence)**.** *Let (f, t_0, u_0) be an IVP, $\tau_0, \tau_1 \in [0, \infty)$, and $r \in (0, \infty)$ such that $\max\{\tau_0, \tau_1\} > 0$, f is defined and continuous on $[t_0 - \tau_0, t_0 + \tau_1] \times B[u_0, r]$, and $\|f(t, x)\| \leq r/\max\{\tau_0, \tau_1\}$ for all $(t, x) \in [t_0 - \tau_0, t_0 + \tau_1] \times B[u_0, r]$. Then, there exists an integral curve $u : [t_0 - \tau_0, t_0 + \tau_1] \to B[u_0, r]$ of f such that $u(t_0) = u_0$.*

If $u(t)$ is a position at time t, then the required upper bound on the norm of f can be interpreted as a speed limit, which ensures that, if $u : [t_0 - \tau_0, t_0 + \tau_1] \to \mathbb{R}^n$ is a solution to the IVP (f, t_0, u_0), then $u([t_0 - \tau_0, t_0 + \tau_1]) \subseteq B[u_0, r]$ since, by Proposition 7.1.4, for all $t \in [t_0 - \tau_0, t_0 + \tau_1]$,

$$\|u(t) - u_0\| = \left\| \int_{t_0}^{t} f(s, u(s)) \, \mathrm{d}s \right\| \leq \left| \int_{t_0}^{t} \|f(s, u(s))\| \, \mathrm{d}s \right| \leq r|t - t_0| / \max\{\tau_0, \tau_1\} \leq r.$$

As explained next, the required upper bound always holds, possibly with smaller τ_0 and τ_1. By the extreme-value theorem (Theorem B.2.3), $\bar{\tau} \coloneqq r/\max_{(t,x) \in [t_0 - \tau_0, t_0 + \tau_1] \times B[u_0, r]} \|f(t, x)\| > 0$. Define, for all $i \in \{0, 1\}$, $\hat{\tau}_i \coloneqq \min\{\tau_i, \bar{\tau}\}$. Then, for all $(t, x) \in [t_0 - \hat{\tau}_0, t_0 + \hat{\tau}_1] \times B[u_0, r]$, $\|f(t, x)\| \leq r/\bar{\tau} \leq r/\max\{\hat{\tau}_0, \hat{\tau}_1\}$.

Maximality

Theorem 7.1.6 ensures local existence: it gives no information on the maximal length of an interval of existence.

Theorem 7.1.7. *Let (f, t_0, u_0) be an IVP. If $\operatorname{dom} f$ is open and f is continuous, then the IVP admits a maximal solution, whose interval of existence is open.*

Proof. This proof follows closely [LR14, proof of Theorem 4.8]. Let S be the set of all solutions of open interval of existence. The local existence theorem (Theorem 7.1.6) implies that S is nonempty and contains all maximal solutions. Define a partial order $\preceq$ on S by $u \preceq v$ if $\operatorname{dom} u \subseteq \operatorname{dom} v$ and, for all $t \in \operatorname{dom} u$, $u(t) = v(t)$. Every maximal element of S is a maximal solution. Thus, it suffices to prove that S has a maximal element. By Zorn's lemma, it suffices to prove that every totally ordered subset of S has a maximal element. Let T be a totally ordered subset of S. Define $I \coloneqq \bigcup_{v \in T} \operatorname{dom} v$. Since T is totally ordered, for every $t \in I$, all elements of T that are defined at t have the same value at t, denoted by $u(t)$. The function $u : I \to \mathbb{R}^n : t \mapsto u(t)$ is a maximal element of T. □

Theorem 7.1.8 (characterization of maximal integral curves)**.** *Let $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ be continuous and $\operatorname{dom} f$ be open. An integral curve u of f is maximal if and only if, for every $t_* \in \partial \operatorname{dom} u$ and every compact $K \subseteq \mathbb{R}^n$ such that $\{t_*\} \times K \subseteq \operatorname{dom} f$, there exists $\delta \in (0, \infty)$ such that, for all $t \in \operatorname{dom} u$ such that $|t - t_*| \leq \delta$, $u(t) \notin K$.*

Proof. If u is not maximal, then there exists an integral curve v of f such that $\operatorname{dom} u \subsetneq \operatorname{dom} v$ and $u(t) = v(t)$ for all $t \in \operatorname{dom} u$. Let $t_* \in (\partial \operatorname{dom} u) \cap \operatorname{dom} v$. As $\operatorname{dom} f$ is open, there exists $\varepsilon \in (0, \infty)$ such that $\{t_*\} \times B[v(t_*), \varepsilon] \subseteq \operatorname{dom} f$. As v is continuous at t_*, there exists $\eta \in (0, \infty)$ such that $v(t) \in B[v(t_*), \varepsilon] =: K$ if $t \in \operatorname{dom} v$ and $|t - t_*| \leq \eta$. Since $t_* \in \partial \operatorname{dom} u$, for every $\delta \in (0, \infty)$, there exists $t \in \operatorname{dom} u$ such that $|t - t_*| \leq \min\{\eta, \delta\}$ and thus $u(t) = v(t) \in K$.

Conversely, if there exist $t_* \in \partial \operatorname{dom} u$ and a compact $K \subseteq \mathbb{R}^n$ such that $\{t_*\} \times K \subseteq \operatorname{dom} f$ and, for every $\delta \in (0, \infty)$, there exists $t \in \operatorname{dom} u$ such that $|t - t_*| \leq \delta$ and $u(t) \in K$, then there exists a sequence $(t_i)_{i \in \mathbb{N}}$ in $\operatorname{dom} u$ converging to t_* such that $u(t_i) \in K$ for all $i \in \mathbb{N}$. By compactness of K, a subsequence of $(u(t_i))_{i \in \mathbb{N}}$ converges to $u_* \in K$. Theorem 7.1.6 applied with the initial condition (t_*, u_*) then enables to extend u. □

In the case where $\operatorname{dom} f = \mathbb{R} \times \mathbb{R}^n$, Theorem 7.1.8 implies that, if u is a maximal integral curve and $\sup \operatorname{dom} u < \infty$, then $\lim_{t \to \sup \operatorname{dom} u} \|u(t)\| = \infty$. It suffices to take $K \coloneqq B[0, r]$ for arbitrary $r \in (0, \infty)$. See an application in Exercise 7.1.16.

Uniqueness

In general, a continuous vector field has several distinct integral curves passing through a given point.

Exercise 7.1.9. Prove that the IVP

$$\begin{cases} u'(t) = \sqrt{u(t)} \text{ for all } t \in \mathbb{R} \\ u(0) = 0 \end{cases}$$

has an uncountable number of solutions.

Uniqueness can be guaranteed based on the following.

Definition 7.1.10. A function $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is said to be *Lipschitz continuous* in the second argument, uniformly with respect to the first argument, on a nonempty subset X of $\operatorname{dom} f$ if

$$\operatorname{Lip}^2_X f \coloneqq \sup \left\{ \frac{\|f(t,x) - f(t,y)\|}{\|x - y\|} \mid t \in \mathbb{R},\, x, y \in \mathbb{R}^n,\, x \neq y,\, (t,x), (t,y) \in X \right\} < \infty.$$

The function f is said to be *locally Lipschitz continuous* in the second argument, uniformly with respect to the first argument, if, for every $(t,x) \in \operatorname{dom} f$, there exist $\tau, r \in (0,\infty)$ such that f is Lipschitz continuous in the second argument, uniformly with respect to the first argument, on $((t-\tau, t+\tau) \times B(x,r)) \cap \operatorname{dom} f$.

The next proposition is analogous to Proposition 2.2.4.

Proposition 7.1.11. *Let $I \subseteq \mathbb{R}$ be an interval, U be a nonempty open convex subset of $\mathbb{R}^n$, and $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ be defined on $I \times U$. Assume that, for every $t \in I$, $f(t,\cdot) : U \to \mathbb{R}^n : x \mapsto f(t,x)$ is differentiable. Then, $\operatorname{Lip}^2_{I\times U} f = \sup_{(t,x)\in I\times U} \|\partial_2 f(t,x)\|$. Thus, on $I \times U$, f is Lipschitz continuous in the second argument, uniformly with respect to the first argument, if and only if $\partial_2 f$ is bounded.*

Proof. Let $t \in I$. Let x and y be two distinct points in U. By the mean-value theorem (Theorem B.3.2),

$$\|f(t,x) - f(t,y)\| \leq \|x - y\| \sup_{s\in(0,1)} \|\partial_2 f(t, x + s(y-x))\|.$$

Thus, $\operatorname{Lip}^2_{I\times U} f \leq \sup_{(s,z)\in I\times U} \|\partial_2 f(s,z)\|$.

Conversely, let $t \in I$, $x \in U$, and $h \in \mathbb{R}^n \setminus \{0\}$. For all $s \in (0,\infty)$ sufficiently small, $x + sh \in U$ and

$$\frac{\|\partial_2 f(t,x)h\|}{\|h\|} = \frac{\|\partial_2 f(t,x)sh\|}{\|sh\|} \leq \underbrace{\frac{\|f(t,x+sh) - f(t,x) - \partial_2 f(t,x)sh\|}{\|sh\|}}_{\to 0 \text{ as } s\to 0} + \underbrace{\frac{\|f(t,x+sh) - f(t,x)\|}{\|sh\|}}_{\leq \operatorname{Lip}^2_{I\times U} f}.$$

Thus, $\|\partial_2 f(t,x)\| \leq \operatorname{Lip}^2_{I\times U} f$. □

Corollary 7.1.12. *A function $f : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz continuous in the second argument, uniformly with respect to the first argument, if $\operatorname{dom} f$ is open and:*

1. *for every $t \in \mathbb{R}$ such that there exists $x \in \mathbb{R}^n$ such that $(t,x) \in \operatorname{dom} f$, $f(t,\cdot) : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto f(t,x)$ is differentiable (note that $\operatorname{dom} f(t,\cdot) = \{x \in \mathbb{R}^n \mid (t,x) \in \operatorname{dom} f\}$ is open);*
2. *$\partial_2 f : \mathbb{R} \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) : (t,x) \mapsto \partial_2 f(t,x)$ is locally bounded (note that $\operatorname{dom} \partial_2 f = \operatorname{dom} f$).*

Theorem 7.1.13 (local uniqueness)**.** *Let (f, t_0, u_0) be an IVP, $\tau_0, \tau_1 \in [0,\infty)$, and $r \in (0,\infty)$ such that $\max\{\tau_0, \tau_1\} > 0$ and f is defined, continuous, and Lipschitz continuous in the second argument, uniformly with respect to the first argument, on $[t_0-\tau_0, t_0+\tau_1]\times B[u_0,r]$. Let $L \coloneqq \operatorname{Lip}^2_{[t_0-\tau_0,t_0+\tau_1]\times B[u_0,r]} f$, $\bar{\tau} \in (0, \min\{r/\max_{(t,x)\in[t_0-\tau_0,t_0+\tau_1]\times B[u_0,r]} \|f(t,x)\|, 1/L\})$, and $\hat{\tau}_i \coloneqq \min\{\tau_i, \bar{\tau}\}$ for all $i \in \{0,1\}$. If $u : [t_0 - \hat{\tau}_0, t_0 + \hat{\tau}_1] \to B[u_0,r]$ and $v : [t_0 - \hat{\tau}_0, t_0 + \hat{\tau}_1] \to B[u_0,r]$ are two integral curves of f such that $u(t_0) = u_0 = v(t_0)$, then $u = v$.*

Proof. Let $X \coloneqq \mathrm{C}^0([t_0 - \hat{\tau}_0, t_0 + \hat{\tau}_1], B[u_0, r])$. Define $\Phi : X \to X$ by

$$\Phi u(t) \coloneqq u_0 + \int_{t_0}^{t} f(s, u(s))\,\mathrm{d}s.$$

Then, $\operatorname{dom}\Phi = X$. By Proposition 7.1.4, $u \in X$ is a solution to the IVP if and only if u is a fixed point of Φ. Furthermore, Φ is a contraction for the norm $\|\cdot\|_\infty$: for all $u, v \in X$,

$$\begin{aligned}
\|\Phi u - \Phi v\|_\infty &= \max_{t\in[t_0-\hat{\tau}_0,t_0+\hat{\tau}_1]} \left\| \int_{t_0}^{t} (f(s,u(s)) - f(s,v(s)))\,\mathrm{d}s \right\| \\
&\le \max_{t\in[t_0-\hat{\tau}_0,t_0+\hat{\tau}_1]} \left| \int_{t_0}^{t} \|f(s,u(s)) - f(s,v(s))\|\,\mathrm{d}s \right| \\
&\le L \max_{t\in[t_0-\hat{\tau}_0,t_0+\hat{\tau}_1]} \left| \int_{t_0}^{t} \|u(s) - v(s)\|\,\mathrm{d}s \right| \\
&\le L \max\{\hat{\tau}_0, \hat{\tau}_1\} \|u - v\|_\infty \\
&\le L\bar{\tau}\|u - v\|_\infty.
\end{aligned}$$

Thus, as seen in the proof of the Banach fixed-point theorem (Theorem 2.2.5), Φ has at most one fixed point. □

The following global version of Theorem 7.1.13 states that two integral curves that meet each other at a point are necessarily equal on the intersection of their respective intervals of existence.

Theorem 7.1.14 (global uniqueness)**.** *Let $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ be continuous and locally Lipschitz continuous in the second argument, uniformly with respect to the first argument. Let u and v be two integral curves of f. If $\operatorname{dom} f$ is open and there exists $t_0 \in \operatorname{dom} u \cap \operatorname{dom} v$ such that $u(t_0) = v(t_0)$, then $u(t) = v(t)$ for all $t \in \operatorname{dom} u \cap \operatorname{dom} v$.*

Proof. Assume, for the sake of contradiction, that there exists $\tau \in \operatorname{dom} u \cap \operatorname{dom} v$ such that $u(\tau) \neq v(\tau)$ and, without loss of generality, that $\tau > t_0$. Define $t_* \coloneqq \inf\{t \in [t_0, \tau] \mid u(t) \neq v(t)\}$. Let us prove that $u(t_*) = v(t_*)$. This is clear if $t_* = t_0$. If $t_* \in (t_0, \tau]$, since $u(t) = v(t)$ for all $t \in [t_0, t_*)$ and u and v are continuous,

$$u(t_*) = \lim_{\substack{t\to t_* \\ t<t_*}} u(t) = \lim_{\substack{t\to t_* \\ t<t_*}} v(t) = v(t_*).$$

In particular, $t_* \in [t_0, \tau)$.

Let us now apply the local uniqueness theorem (Theorem 7.1.13) to obtain a contradiction with respect to the definition of t_*. The restrictions of u and v to $[t_*, \tau]$ are two integral curves of f taking the same value at t_* and, by the local uniqueness theorem, therefore coincide on $[t_*, \tau_*]$ for some $\tau_* \in (t_*, \tau]$, which is a contradiction with respect to the definition of t_*. □

A function from $\mathbb{R}^n$ to $\mathbb{R}^n$ is called an *autonomous vector field*.

Example 7.1.15. Let us compute the maximal solution to the IVP

$$\begin{cases} u' = u^2 \\ u(0) = u_0 \end{cases}$$

for every $u_0 \in \mathbb{R}$. The associated autonomous vector field is $f : \mathbb{R} \to \mathbb{R} : x \mapsto x^2$. Since f is continuously differentiable, it is locally Lipschitz continuous and Theorem 7.1.14 therefore applies. If $u_0 = 0$, the identically zero function is the unique maximal solution. Assume that $u_0 > 0$. Then, the solution u has no zero; indeed, if it has a zero, then it should be equal to the identically zero function by uniqueness. Thus, u is positive and

$$\frac{u'}{u^2} = 1.$$

Integrating both sides from 0 to t yields

$$\int_0^t \frac{u'}{u^2} = \int_0^t 1,$$

i.e.,

$$\int_0^t \left(-\frac{1}{u}\right)' = t$$

or, equivalently,

$$\frac{1}{u_0} - \frac{1}{u(t)} = t.$$

We conclude that the maximal solution is the function

$$u : \left(-\infty, \frac{1}{u_0}\right) \to \mathbb{R} : t \mapsto \frac{1}{\frac{1}{u_0} - t}.$$

The same reasoning shows that, if $u_0 < 0$, then the maximal solution is

$$u : \left(\frac{1}{u_0}, \infty\right) \to \mathbb{R} : t \mapsto \frac{1}{\frac{1}{u_0} - t}.$$

If $u_0 \neq 0$, the maximal solution is not defined on $\mathbb{R}$ although the autonomous vector field is defined on $\mathbb{R}$.

Exercise 7.1.16. Let $u_0, u_1 \in \mathbb{R}$. Prove that every solution to the IVP

$$\begin{cases} u'' = -u^3 \\ u(0) = u_0 \\ u'(0) = u_1 \end{cases}$$

satisfies $u^4/4 + u'^2/2 = u_0^4/4 + u_1^2/2$. Deduce that the maximal solution to the IVP is defined on $\mathbb{R}$.

Introduction to Lyapunov stability

A zero of an autonomous vector field is called an *equilibrium point* of the autonomous vector field. If u is an equilibrium point of an autonomous vector field $f : \mathbb{R}^n \to \mathbb{R}^n$, then the constant function $\mathbb{R} \to \mathbb{R}^n : t \mapsto u$ is a maximal integral curve of f.

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz continuous and $\operatorname{dom} f$ be open. Then, by Theorems 7.1.7 and 7.1.14, for every $u_0 \in \operatorname{dom} f$, the IVP $(f, 0, u_0)$ has a unique maximal solution.

Definition 7.1.17. An equilibrium point $\bar{u}$ of f is said to be:

- *stable* if, for every $\varepsilon \in (0, \infty)$ such that $B[\bar{u}, \varepsilon] \subseteq \operatorname{dom} f$, there exists $\delta \in (0, \infty)$ such that, for every $u_0 \in B[\bar{u}, \delta] \subseteq \operatorname{dom} f$, the unique maximal integral curve u of f such that $u(0) = u_0$ is defined on $[0, \infty)$ and $u([0, \infty)) \subseteq B[\bar{u}, \varepsilon]$;
- *attractive* if there exists $\delta \in (0, \infty)$ such that, for every $u_0 \in B[\bar{u}, \delta] \subseteq \operatorname{dom} f$, the unique maximal integral curve u of f such that $u(0) = u_0$ is defined on $[0, \infty)$ and $\lim_{t\to\infty} u(t) = \bar{u}$;
- *asymptotically stable* if it is both stable and attractive;
- *unstable* if it is not stable.

Stability and attractivity are two independent concepts: an equilibrium point can be stable while not attractive (an example is $(0, 0)$ for the simple pendulum $f : \mathbb{R}^2 \to \mathbb{R}^2 : (x_1, x_2) \mapsto (x_2, -\sin x_1)$) or attractive while not stable (see [LR14, Exercise 5.10] for an example).

If f is linear, then there exists $A \in \mathbb{R}^{n\times n}$ such that $f(x) = Ax$ for all $x \in \mathbb{R}^n$. In that case, 0 is an equilibrium point, which is:

- asymptotically stable if each eigenvalue of A has a negative real part;
- unstable if an eigenvalue of A has a positive real part.

This is easy to see if A is diagonalizable, i.e., there exists an invertible $P \in \mathbb{C}^{n\times n}$ such that

$$P^{-1}AP = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

where $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ are the eigenvalues of A (see Section A.1). With $\hat{u} \coloneqq P^{-1}u$, the ODE $u' = Au$ becomes $\hat{u}' = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)\hat{u}$ or, equivalently, $\hat{u}_i' = \lambda_i \hat{u}_i$ for all $i \in \{1, \ldots, n\}$, which yields $\hat{u}_i(t) = \exp(\lambda_i t)\hat{u}_i(0)$ for all $t \in \mathbb{R}$. Thus, for all $t \in \mathbb{R}$, $\hat{u}(t) = \operatorname{diag}(\exp(\lambda_1 t), \ldots, \exp(\lambda_n t))\hat{u}(0)$, i.e.,

$$u(t) = P\operatorname{diag}(\exp(\lambda_1 t), \ldots, \exp(\lambda_n t))P^{-1}u(0).$$

If A is not diagonalizable, a similar argument can be made based on the *Jordan canonical form* of A.

Exercise 7.1.18. Compute the solution to the IVP

$$\begin{cases} u'(t) = Au(t) \text{ for all } t \in \mathbb{R} \\ u(0) = u_0 \end{cases} \quad \text{where} \quad A \coloneqq \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad u_0 \in \mathbb{R}^2.$$

The stability of nonlinear autonomous vector fields can be studied by linearization.

Theorem 7.1.19 (Lyapunov's first method)**.** *Let $\bar{u}$ be an equilibrium point of f. Assume that f is differentiable at $\bar{u}$. Then $\bar{u}$ is:*

- *asymptotically stable if each eigenvalue of $f'(\bar{u})$ has a negative real part;*
- *unstable if an eigenvalue of $f'(\bar{u})$ has a positive real part.*

Exercise 7.1.20. Consider the autonomous vector field associated with a damped pendulum, namely $f : \mathbb{R}^2 \to \mathbb{R}^2 : (x_1, x_2) \mapsto (x_2, -\sin x_1 - x_2)$. What are the equilibrium points of f? Study their stability.

7.1.2 Basic one-step methods

In the rest of Section 7.1, $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, $(t_0, u_0) \in \operatorname{dom} f$, and $\tau, r \in (0, \infty)$ are such that f is defined, continous, and Lipschitz continuous in the second argument, uniformly with respect to the first argument, on $[t_0, t_0 + \tau] \times B[u_0, r]$. Moreover, for all $(t, x) \in [t_0, t_0 + \tau] \times B[u_0, r]$, $\|f(t, x)\| \leq r/\tau$. Thus, by Theorems 7.1.6 and 7.1.13–7.1.14, the IVP

$$\begin{cases} u'(t) = f(t, u(t)) \text{ for all } t \in [t_0, t_0 + \tau] \\ u(t_0) = u_0 \end{cases}$$

has a unique solution $u : [t_0, t_0 + \tau] \to B[u_0, r]$. The goal is to find an approximation to u. The quality of the approximation is evaluated based on an error analysis in Section 7.1.3 and a stability analysis in Section 7.1.4.

Given $m \in \mathbb{N} \setminus \{0\}$, $h \coloneqq \tau/m$, and $t_i \coloneqq t_0 + hi$ for all $i \in \{1, \ldots, m\}$, the four methods presented in this section each generate a sequence $(u_1, \ldots, u_m)$ in $B[u_0, r]$ such that, for all $i \in \{1, \ldots, m\}$, $u_i \approx u(t_i)$. Let $L \coloneqq \operatorname{Lip}^2_{[t_0, t_0+\tau]\times B[u_0, r]} f$.

The *forward Euler method* iterates

$$u_{i+1} \coloneqq u_i + hf(t_i, u_i) \tag{7.1}$$

for all $i \in \{0, \ldots, m-1\}$. This iteration comes from the forward finite-difference formula (5.1):

$$\frac{u_{i+1} - u_i}{h} \approx \frac{u(t_{i+1}) - u(t_i)}{h} \approx u'(t_i) = f(t_i, u(t_i)) \approx f(t_i, u_i).$$

The iteration (7.1) is said to be *explicit* because the unknown u_{i+1} appears only in the left-hand side, hence can be computed through an evaluation of f and arithmetic operations.

The *backward Euler method* proceeds similarly with the backward finite-difference formula (5.1):

$$\frac{u_{i+1} - u_i}{h} \approx \frac{u(t_{i+1}) - u(t_i)}{h} \approx u'(t_{i+1}) = f(t_{i+1}, u(t_{i+1})) \approx f(t_{i+1}, u_{i+1}).$$

This yields the formula

$$u_{i+1} = u_i + hf(t_{i+1}, u_{i+1}), \tag{7.2}$$

which is said to be *implicit* because the unknown u_{i+1} appears in both sides, hence, in general, cannot be computed exactly within a finite number of evaluations of f and arithmetic operations. Nevertheless, (7.2) generates a well-defined sequence if $hL < 1$, as indicated in Proposition 7.1.21.

Both methods have been derived from formulas of numerical differentiation. Unsurprisingly, they can also be derived from formulas of numerical integration since

$$u_{i+1} \approx u(t_{i+1}) = u(t_i) + \int_{t_i}^{t_{i+1}} f(t, u(t))\,\mathrm{d}t \approx u_i + \int_{t_i}^{t_{i+1}} f(t, u(t))\,\mathrm{d}t;$$

the formulas of numerical integration were defined for $n = 1$ but make sense for all $n \in \mathbb{N} \setminus \{0\}$. The forward and backward Euler methods are obtained by approximating the integral with $hf(t_i, u_i)$ and $hf(t_{i+1}, u_{i+1})$, respectively. Approximating the integral with the trapezium rule from Section 6.1 yields the Crank–Nicolson method,

$$u_{i+1} = u_i + \frac{h}{2}\left(f(t_i, u_i) + f(t_{i+1}, u_{i+1})\right), \tag{7.3}$$

which is implicit. Approximating u_{i+1} in the right-hand side of (7.3) with the forward Euler method yields Heun's method,

$$u_{i+1} \coloneqq u_i + \frac{h}{2}\left(f(t_i, u_i) + f(t_{i+1}, u_i + hf(t_i, u_i))\right), \tag{7.4}$$

which is explicit. The next proposition, illustrated in Figure 7.1, ensures that the four methods presented in this section are well defined.

Proposition 7.1.21. *The iterations* (7.1)–(7.4) *each generate a well-defined sequence* $(u_1, \dots, u_m)$ *such that, for all* $i \in \{0, \dots, m\}$, $u_i \in B[u_0, \frac{i}{m}r]$, *provided that* $hL < 1$ *for* (7.2) *and* $hL < 2$ *for* (7.3).

Proof. For (7.1) and (7.4), the result follows from the inequality $\|u_{i+1} - u_i\| \leq hr/\tau = r/m$, which holds for all $i \in \{0, \dots, m-1\}$. We now focus on (7.2), proceeding by induction; the argument is similar for (7.3). The statement is true for $i = 0$. Assume that, for some $i \in \{0, \dots, m-1\}$, u_i is defined and $u_i \in B[u_0, \frac{i}{m}r]$. Define $g : B[u_0, \frac{i+1}{m}r] \to \mathbb{R}^n : x \mapsto u_i + hf(t_{i+1}, x)$ and observe that $\operatorname{Lip}_{B[u_0, \frac{i+1}{m}r]} g \leq hL$. Moreover, $g(B[u_0, \frac{i+1}{m}r]) \subseteq B[u_0, \frac{i+1}{m}r]$ since, for all $x \in B[u_0, \frac{i+1}{m}r]$, $\|g(x) - u_0\| \leq \|g(x) - u_i\| + \|u_i - u_0\| \leq \frac{r}{m} + \frac{ir}{m} = \frac{i+1}{m}r$. By the Banach fixed-point theorem (Theorem 2.2.5), g has a unique fixed point $u_{i+1} \in B[u_0, \frac{i+1}{m}r]$. □

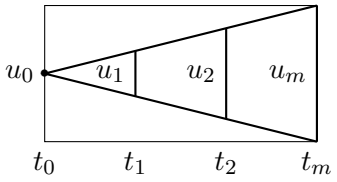

Figure 7.1: Illustration of Proposition 7.1.21 for $n = 1$ and $m = 3$. The rectangle is $[t_0, t_0+\tau] \times B[u_0, r]$. As seen in Section 7.1.1, the graph of u is contained in the thick cone of vertex (t_0, u_0). For every $i \in \{1, \dots, m\}$, the vertical thick segment above t_i is $B[u_0, \frac{i}{m}r]$, which contains u_i.

7.1.3 Error analysis of one-step methods

A method that generates a sequence $(u_1, \dots, u_m)$ in $B[u_0, r]$ such that, for all $i \in \{1, \dots, m\}$, $u_i \approx u(t_i)$ is said to be *convergent of order* $p \in \mathbb{N} \in \{0\}$ if there exists $c \in (0, \infty)$ such that

$$\|u(t_0 + \tau) - u_m\| \leq ch^p$$

if h is sufficiently small. This section develops a framework to justify the last column of Table 7.1.

Method	$\Phi_h(t,x,y)$	order of convergence
forward Euler	$f(t,x)$	1 if $f \in \mathrm{C}^1([t_0,t_0+\tau]\times B[u_0,r],\mathbb{R}^n)$
backward Euler	$f(t+h,y)$	1 if $f \in \mathrm{C}^1([t_0,t_0+\tau]\times B[u_0,r],\mathbb{R}^n)$
Crank–Nicolson	$(f(t,x)+f(t+h,y))/2$	2 if $f \in \mathrm{C}^2([t_0,t_0+\tau]\times B[u_0,r],\mathbb{R}^n)$
Heun	$(f(t,x)+f(t+h,x+hf(t,x)))/2$	2 if $f \in \mathrm{C}^2([t_0,t_0+\tau]\times B[u_0,r],\mathbb{R}^n)$

Table 7.1: One-step methods from Section 7.1.2.

A *one-step method* is defined by an iteration that can be written as

$$u_{i+1} = u_i + h\Phi_h(t_i, u_i, u_{i+1}) \tag{7.5}$$

for some continuous function $\Phi_h : \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}^n$. All methods defined in Section 7.1.2 are one-step methods, as indicated in Table 7.1. The method is said to be explicit if Φ_h does not depend on its third argument, in which case Φ_h is considered as a function from $\mathbb{R}\times\mathbb{R}^n$ to $\mathbb{R}^n$, and implicit otherwise.

The *local discretization error* associated with Φ_h is defined as

$$\Delta_h : [t_0, t_0+\tau-h] \to \mathbb{R}^n : t \mapsto \frac{u(t+h)-u(t)}{h} - \Phi_h(t, u(t), u(t+h)).$$

Thus, $h\Delta_h$ measures the error made by the method on one step. The associated method is said to be *consistent* if $\lim_{\varepsilon\to 0^+}\Delta_\varepsilon = 0$. The local discretization error is said to be of order $p \in \mathbb{N}\setminus\{0\}$ if there exists $c \in (0,\infty)$ such that $\|\Delta_h(t)\| \le ch^p$ for all $t \in [t_0, t_0+\tau-h]$.

Proposition 7.1.22 (convergence of explicit one-step methods)**.** *Let $\Phi_h : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ be continuous and Lipschitz continuous in the second argument, uniformly with respect to the first argument, on $([t_0,t_0+h]\times B[u_0,r])\cap \operatorname{dom}\Phi_h$. A method is convergent of order $p \in \mathbb{N}\setminus\{0\}$ if its local discretization error is of order p.*

Lemma 7.1.23. *Let $a,b \in (0,\infty)$ and $(v_i)_{i\in\mathbb{N}}$ be a sequence in $[0,\infty)$ such that $v_0 = 0$ and, for all $i \in \mathbb{N}$,*

$$v_{i+1} \le (1+a)v_i + b.$$

Then, for all $i \in \mathbb{N}$,

$$v_i \le \frac{b}{a}(\exp(ai)-1).$$

Proof. The inequality is true for $i = 0$. If it is true for $i \in \mathbb{N}$, then it is also true for $i+1$:

$$v_{i+1} \le (1+a)v_i + b \le (1+a)\frac{b}{a}(\exp(ai)-1) + b = \frac{b}{a}\left((1+a)\exp(ai)-1\right) < \frac{b}{a}\left(\exp(a(i+1))-1\right),$$

where the last inequality holds because $1+a < \exp(a)$. □

Proof of Proposition 7.1.22. By assumption, there exists $c \in (0,\infty)$ such that, for all $t \in [t_0, t_0+\tau-h]$, $\|\Delta_h(t)\| \le ch^p$. Let $i \in \{0,\dots,m-1\}$. Then,

$$\begin{aligned} u(t_{i+1}) - u_{i+1} &= (\Delta_h(t_i)h + u(t_i) + h\Phi_h(t_i, u(t_i))) - (u_i + h\Phi_h(t_i, u_i)) \\ &= (u(t_i)-u_i) + h\left(\Phi_h(t_i,u(t_i)) - \Phi_h(t_i,u_i)\right) + \Delta_h(t_i)h. \end{aligned}$$

Let $\tilde{L} \in (0,\infty)$ be the Lipschitz constant. Then,

$$\begin{aligned} \|u(t_{i+1}) - u_{i+1}\| &\le \|u(t_i)-u_i\| + h\,\|\Phi_h(t_i,u(t_i)) - \Phi_h(t_i,u_i)\| + \|\Delta_h(t_i)\|h \\ &\le (1+h\tilde{L})\|u(t_i)-u_i\| + ch^{p+1}. \end{aligned}$$

Therefore, by Lemma 7.1.23,

$$\|u(t_i)-u_i\| \le \frac{c}{\tilde{L}}\left(\exp(h\tilde{L}i)-1\right)h^p.$$

In particular,

$$\|u(t_0+\tau)-u_m\| \le \frac{c}{\tilde{L}}\left(\exp(\tilde{L}\tau)-1\right)h^p.$$ □

The following ensures that the forward Euler and Heun methods satisfy the hypotheses of Proposition 7.1.22.

Proposition 7.1.24. *For the forward Euler method,*

$$\operatorname{Lip}^2_{[t_0,t_0+\tau]\times B[u_0,r]} \Phi_h = L.$$

For Heun's method,

$$\operatorname{Lip}^2_{[t_0,t_0+\tau-h]\times B[u_0,(1-\frac{1}{m})r]} \Phi_h \le L\left(1+\frac{hL}{2}\right).$$

The following is an analogous result for implicit one-step methods.

Proposition 7.1.25 (convergence of implicit one-step methods)**.** *Let $\Phi_h : \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}^n$ be continuous and assume that there exists $\tilde{L}\in(0,\infty)$ such that, for all $t\in[t_0,t_0+\tau]$ and $w,x,y,z\in B[u_0,r]$ such that $(t,x,y),(t,w,z)\in\operatorname{dom}\Phi_h$,*

$$\|\Phi_h(t,x,y)-\Phi_h(t,w,z)\| \le \tilde{L}(\|x-w\|+\|y-z\|).$$

If $\varepsilon\in(0,1)$, $h\tilde{L}\le 1-\varepsilon$, and the local discretization error is of order $p\in\mathbb{N}\setminus\{0\}$, then the associated method is convergent of order p.

Proof. By assumption, there exists $c\in(0,\infty)$ such that, for all $t\in[t_0,t_0+\tau-h]$, $\|\Delta_h(t)\|\le ch^p$. Let $i\in\{0,\dots,m-1\}$. Then,

$$\begin{aligned} u(t_{i+1})-u_{i+1} &= (\Delta_h(t_i)h+u(t_i)+h\Phi_h(t_i,u(t_i),u(t_{i+1})))-(u_i+h\Phi_h(t_i,u_i,u_{i+1})) \\ &= (u(t_i)-u_i)+h\left(\Phi_h(t_i,u(t_i),u(t_{i+1}))-\Phi_h(t_i,u_i,u_{i+1})\right)+\Delta_h(t_i)h. \end{aligned}$$

Thus,

$$\begin{aligned} \|u(t_{i+1})-u_{i+1}\| &\le \|u(t_i)-u_i\|+h\,\|\Phi_h(t_i,u(t_i),u(t_{i+1}))-\Phi_h(t_i,u_i,u_{i+1})\|+\|\Delta_h(t_i)\|h \\ &\le \frac{(1+h\tilde{L})\|u(t_i)-u_i\|+ch^{p+1}}{1-h\tilde{L}}. \end{aligned}$$

Therefore, by Lemma 7.1.23,

$$\|u(t_i)-u_i\| \le \frac{c}{2\tilde{L}}\left(\exp\left(\frac{2h\tilde{L}i}{1-h\tilde{L}}\right)-1\right)h^p.$$

In particular,

$$\|u(t_0+\tau)-u_m\| \le \frac{c}{2\tilde{L}}\left(\exp\left(\frac{2\tau\tilde{L}}{1-h\tilde{L}}\right)-1\right)h^p \le \frac{c}{2\tilde{L}}\left(\exp\left(\frac{2\tau\tilde{L}}{\varepsilon}\right)-1\right)h^p. \qquad \square$$

Proposition 7.1.26. *For the backward Euler method, $[t_0,t_0+\tau-h]\times\mathbb{R}^n\times B[u_0,r]\subseteq\operatorname{dom}\Phi_h$ and, for all $t\in[t_0,t_0+\tau-h]$, $w,x\in\mathbb{R}^n$, and $y,z\in B[u_0,r]$,*

$$\|\Phi_h(t,x,y)-\Phi_h(t,w,z)\| \le L\|y-z\|.$$

For the Crank–Nicolson method, $[t_0,t_0+\tau-h]\times B[u_0,r]\times B[u_0,r]\subseteq\operatorname{dom}\Phi_h$ and

$$\tilde{L}\le\frac{L}{2}.$$

To justify the last column of Table 7.1, it remains to prove that the local discretization error is of order 1 for the forward and backward Euler methods and of order 2 for the Crank–Nicolson and Heun methods.

Proposition 7.1.27. *If $f\in\mathrm{C}^1([t_0,t_0+\tau]\times B[u_0,r],\mathbb{R}^n)$, then the respective local discretization errors of the forward and backward Euler methods are of order 1.*

Proof. By Proposition 7.1.5, $u \in \mathrm{C}^2([t_0, t_0+\tau], B[u_0, r])$. Thus, by Taylor's theorem (Theorem B.3.3), for all $t \in [t_0, t_0+\tau-h]$,

$$\|\Delta_h(t)\| \le \frac{1}{2} \max_{s\in[t,t+h]} \|u''(s)\|h \le \frac{1}{2} \max_{s\in[t_0,t_0+\tau]} \|u''(s)\|h. \qquad \square$$

Proposition 7.1.28. *If $f \in \mathrm{C}^2([t_0, t_0+\tau]\times B[u_0, r], \mathbb{R}^n)$, then the respective local discretization errors of the Crank–Nicolson and Heun methods are of order 2.*

Proof. By Proposition 7.1.5, $u \in \mathrm{C}^3([t_0, t_0+\tau], B[u_0, r])$. Consider the Crank–Nicolson method. As can be deduced from Section 6.1, for all $g \in \mathrm{C}^2([a, b], \mathbb{R}^n)$,

$$\left\|(b-a)\frac{g(a)+g(b)}{2} - \int_a^b g\right\|_\infty \le \frac{1}{12}(b-a)^3 \max_{x\in[a,b]} \|g''(x)\|_\infty.$$

Thus, since, for all $t \in [t_0, t_0+\tau-h]$,

$$-\Delta_h(t) = \frac{u'(t)+u'(t+h)}{2} - \frac{1}{h}\int_t^{t+h} u',$$

it holds that

$$\|\Delta_h(t)\|_\infty \le \frac{1}{12} \max_{s\in[t,t+h]} \|u'''(s)\|_\infty h^2 \le \frac{1}{12} \max_{s\in[t_0,t_0+\tau]} \|u'''(s)\|_\infty h^2.$$

Consider Heun's method. For all $t \in [t_0, t_0+\tau-h]$, by Taylor's theorem (Theorem B.3.3),

$$\left\|u(t+h) - \left(u(t)+u'(t)h+u''(t)\frac{h^2}{2}\right)\right\| \le \frac{1}{6} \max_{s\in[t,t+h]} \|u'''(s)\|h^3,$$

hence

$$\left\|\frac{u(t+h)-u(t)}{h} - \left(u'(t)+u''(t)\frac{h}{2}\right)\right\| \le \frac{1}{6} \max_{s\in[t,t+h]} \|u'''(s)\|h^2.$$

For all $t \in [t_0, t_0+\tau]$,

$$\begin{aligned} u'(t) &= f(t, u(t)), \\ u''(t) &= \partial_1 f(t, u(t)) + \partial_2 f(t, u(t))u'(t) = \partial_1 f(t, u(t)) + \partial_2 f(t, u(t))f(t, u(t)). \end{aligned}$$

Thus, for all $t \in [t_0, t_0+\tau-h]$,

$$\left\|\frac{u(t+h)-u(t)}{h} - \left(f(t, u(t)) + (\partial_1 f(t, u(t)) + \partial_2 f(t, u(t))f(t, u(t)))\,\frac{h}{2}\right)\right\| \le \frac{1}{6} \max_{s\in[t,t+h]} \|u'''(s)\|h^2.$$

For all $i \in \{0, 2\}$, let $c_i \coloneqq \max_{(s,y)\in[t_0,t_0+\tau]\times B[u_0,r]} \|f^{(i)}(s, y)\|$; for every $(s, y) \in [t_0, t_0+\tau] \times B[u_0, r]$, $f''(s, y) \in \mathcal{L}_2(\mathbb{R}\times\mathbb{R}^n, \mathbb{R}^n)$ and the norm is that defined in Proposition B.2.7. By Taylor's theorem (Theorem B.3.3), for all $(t, x) \in [t_0, t_0+\tau-h] \times B[u_0, (1-\frac{1}{m})r]$,

$$\|f(t+h, x+hf(t, x)) - (f(t, x) + \partial_1 f(t, x)h + \partial_2 f(t, x)f(t, x)h)\| \le \frac{1}{2}\max\{1, c_0^2\}c_2h^2,$$

hence

$$\left\|\Phi_h(t, x) - \left(f(t, x) + (\partial_1 f(t, x) + \partial_2 f(t, x)f(t, x))\,\frac{h}{2}\right)\right\| \le \frac{1}{4}\max\{1, c_0^2\}c_2h^2.$$

Therefore, for all $t \in [t_0, t_0+\tau-h]$,

$$\begin{aligned} \|\Delta_h(t)\| \le{}& \left\|\frac{u(t+h)-u(t)}{h} - \left(f(t, u(t)) + (\partial_1 f(t, u(t)) + \partial_2 f(t, u(t))f(t, u(t)))\,\frac{h}{2}\right)\right\| \\ &+ \left\|\left(f(t, u(t)) + (\partial_1 f(t, u(t)) + \partial_2 f(t, u(t))f(t, u(t)))\,\frac{h}{2}\right) - \Phi_h(t, u(t))\right\| \\ \le{}& \left(\frac{1}{6} \max_{s\in[t,t+h]} \|u'''(s)\| + \frac{1}{4}\max\{1, c_0^2\}c_2\right)h^2 \\ \le{}& \left(\frac{1}{6} \max_{s\in[t_0,t_0+\tau]} \|u'''(s)\| + \frac{1}{4}\max\{1, c_0^2\}c_2\right)h^2. \end{aligned} \qquad \square$$

7.1.4 Absolute stability of one-step methods

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz continuous and $\operatorname{dom} f$ be open. Let $\bar{u}$ be an equilibrium point of f and assume that f is differentiable at $\bar{u}$ and the real part of each eigenvalue of $f'(\bar{u})$ is negative. Then, by Theorem 7.1.19, for every u_0 sufficiently close to $\bar{u}$, the solution u to the IVP $(f, 0, u_0)$ converges to $\bar{u}$ at infinity. Consider a method that, given $h \in (0, \infty)$, generates a sequence $(u_i)_{i\in\mathbb{N}\setminus\{0\}}$ such that, for all $i \in \mathbb{N} \setminus \{0\}$, $u_i \approx u(hi)$. Does the generated sequence converge to $\bar{u}$?

This section answers the question for one-step methods in the case where f is linear and diagonalizable. Then, as seen in Section 7.1.1, the IVP is equivalent to n decoupled equations of the form $\hat{u}_i' = \lambda_i \hat{u}_i$, where $\lambda_i \in \mathbb{C}$ and $\Re(\lambda_i) < 0$, for all $i \in \{1, \dots, n\}$. Thus, it suffices to consider the case where $n = 1$ and $f : (t, x) \mapsto \lambda x$ for some $\lambda \in \mathbb{C}$ such that $\Re(\lambda) < 0$. Then, the unique solution to the IVP is $u : \mathbb{R} \to \mathbb{C} : t \mapsto \exp(\lambda t)u_0$ and $\lim_{t\to\infty} |u(t)| = 0$. Furthermore, the iteration (7.5) simplifies to

$$u_{i+1} = \phi(h\lambda)u_i$$

for some continuous function $\phi : \mathbb{C} \to \mathbb{C}$, which yields

$$u_i = \phi(h\lambda)^i u_0.$$

Thus, $(u_i)_{i\in\mathbb{N}}$ converges to 0 if and only if $|\phi(h\lambda)| < 1$. The set

$$\{z \in \mathbb{C} \mid |\phi(z)| < 1\}$$

is called the *region of absolute stability* of the method. The method is said to be *absolutely stable* for a step size $h \in (0, \infty)$ if $h\lambda$ is in the region of absolute stability. The method is said to be *unconditionally absolutely stable* (or *A-stable*) if its region of absolute stability contains $\{z \in \mathbb{C} \mid \Re(z) < 0\}$ (in which case $(u_i)_{i\in\mathbb{N}}$ converges to 0 for all $h \in (0, \infty)$), and *conditionally absolutely stable* otherwise ($(u_i)_{i\in\mathbb{N}}$ converges to 0 for some $h \in (0, \infty)$).

The region of absolute stability of each one-step method from Section 7.1.2 is given in Table 7.2 and represented in Figure 7.2. The two implicit methods are unconditionally absolutely stable. In contrast, the two explicit methods are only conditionally absolutely stable. For example, if $\lambda \in \mathbb{R}$, then both methods are absolutely stable if and only if $h\lambda \in (-2, 0)$, i.e., $h < -2/\lambda$. Actually, it is a general fact that explicit methods are, at best, conditionally absolutely stable; see Section 7.1.5. Thus, explicit methods are not suitable for *stiff* problems, where $\Re(\lambda) \ll 0$.

Method	$\phi(z)$	region of absolute stability
forward Euler	$1 + z$	$\{z \in \mathbb{C} \mid \lvert z + 1\rvert < 1\}$
backward Euler	$\dfrac{1}{1 - z}$	$\{z \in \mathbb{C} \mid \lvert z - 1\rvert > 1\}$
Crank–Nicolson	$\dfrac{2 + z}{2 - z}$	$\{z \in \mathbb{C} \mid \Re(z) < 0\}$
Heun	$1 + z + z^2/2$	$\{z \in \mathbb{C} \mid \lvert(z + 1)^2 + 1\rvert < 2\}$

Table 7.2: Region of absolute stability of each one-step method from Section 7.1.2.

7.1.5 Further topics

This section gives an overview of more advanced methods that are out of the scope of this course.

Runge–Kutta methods

Given $s \in \mathbb{N} \setminus \{0\}$, $A \in \mathbb{R}^{s\times s}$, and $b, c \in \mathbb{R}^s$, an s-stage Runge–Kutta method iterates

$$\begin{cases} k_j = f(t_i + c_j h, u_i + h\sum_{l=1}^{s} a_{j,l}k_l) \text{ for all } j \in \{1, \dots, s\}, \\ u_{i+1} = u_i + h\sum_{j=1}^{s} b_j k_j. \end{cases} \tag{7.6}$$

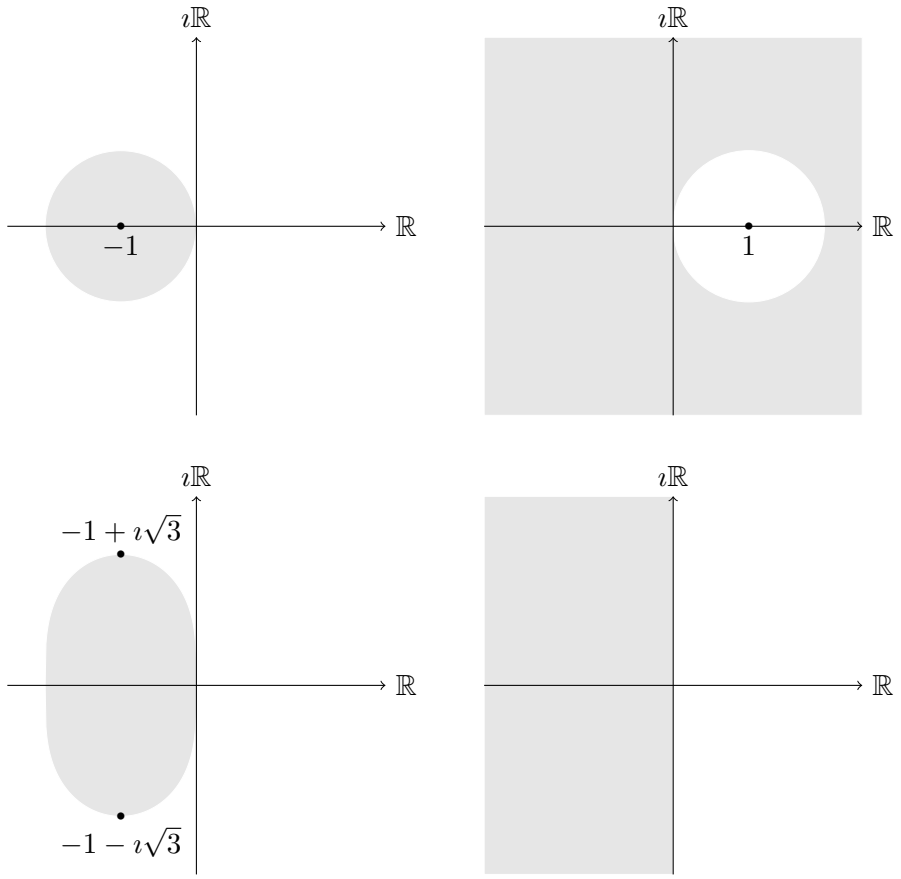

Figure 7.2: Region of absolute stability of each one-step method from Section 7.1.2: explicit methods on the left, implicit methods on the right, first-order methods on the top, second-order methods on the bottom.

The method is explicit if A is strictly lower triangular (i.e., $a_{i,j} = 0$ if $i \leq j$) and implicit otherwise. The method is said to be semi-implicit if A is lower triangular (i.e., $a_{i,j} = 0$ if $i < j$), in which case every iteration requires solving s decoupled equations in $\mathbb{R}$. The well-definedness of implicit Runge–Kutta methods is discussed in [HNW93, §II.7] and [HW96, §IV.14]. Runge–Kutta methods are one-step methods, and the four methods from Section 7.1.2 are Runge–Kutta methods.

The method (7.6) is consistent if and only if $\sum_{i=1}^{s} b_i = 1$. Other constraints on the triplet (A, b, c) are given in [SW22, Theorem 19.4] and [HNW93, §II.2].

The maximum order of an s-stage explicit Runge–Kutta method is s if $1 \leq s \leq 4$, $s-1$ if $5 \leq s \leq 7$, $s-2$ if $8 \leq s \leq 9$, and at most $s-2$ if $s \geq 10$. The maximum order of an s-stage implicit Runge–Kutta method is $2s$. See [Gau12, §5.6.5], [QSS07, p. 520], [MM02, p. 56], [HNW93, §§II.5 and II.7], and references therein. For example, the classic Runge–Kutta method, defined by

$$A = \begin{bmatrix} 1/2 & & & \\ & 1/2 & & \\ & & 1 & \end{bmatrix}, \qquad b = \begin{bmatrix} 1/6 \\ 1/3 \\ 1/3 \\ 1/6 \end{bmatrix}, \qquad c = \begin{bmatrix} 0 \\ 1/2 \\ 1/2 \\ 1 \end{bmatrix},$$

is of order 4.

The region of absolute stability of every explicit Runge–Kutta method is bounded; see [QSS07, §11.8.4], [SM03, p. 351], [HW96, §IV.2], and references therein.

Runge–Kutta methods are covered in [HNW93, Chap. II], [HW96, Chap. IV], [MM02, Chap. III], [SB02, §7.2.1], [SM03, §12.5], [QSS07, §11.8], [Gau12, §5.6.5], and [SW22, Chap. 19].

Linear multistep methods

Given $s \in \mathbb{N} \setminus \{0\}$ and $a_0, \dots, a_s, b_0, \dots, b_s \in \mathbb{R}$ with $a_s \neq 0 \neq a_0 b_0$, a linear s-step method iterates

$$\sum_{j=0}^{s} a_j u_{i+j} = h \sum_{j=0}^{s} b_j f(t_{i+j}, u_{n+j}). \tag{7.7}$$

The method is explicit if $b_s = 0$ and implicit otherwise. Except Heun's method, the one-step methods from Section 7.1.2 are linear multistep methods.

The method (7.7) requires s initial iterates $u_0, \dots, u_{s-1}$. Thus, the iterates $u_1, \dots, u_{s-1}$ must be computed with another method. Informally, a linear s-step method is said to be *zero-stable*, or *D-stable*, if, for sufficiently small h, a small perturbation of the initial iterates $u_0, \dots, u_{s-1}$ induces a small perturbation of the next iterates.

According to the Dahlquist first barrier theorem [Dah56, Theorem 4a], the order of a zero-stable linear s-step method is at most $s + 1$ if s is odd and at most $s + 2$ if s is even.

According to the Dahlquist second barrier theorem [Dah63, Theorems 2.1–2.2], in the class of linear multistep methods:

1. no explicit method is unconditionally absolutely stable;
2. the order of an unconditionally absolutely stable method is at most 2;
3. the second-order unconditionally absolutely stable method with the smallest error constant is the Crank–Nicolson method (constant 1/12).

Thus, unconditional absolute stability, also called A-stability, is a restrictive property.

Linear multistep methods are covered in [HNW93, Chap. III], [HW96, Chap. V], [MM02, Chap. VII], [SB02, §§7.2.6–7.2.13], [SM03, §12.6], [QSS07, §§11.5–11.6], [Gau12, Chap. 6], and [SW22, Chap. 20].

Step-size adaptation

Practical methods do not use a constant step size: they adapt the step size at every iteration. Step-size adaptation is covered in [SB02, §§7.2.5 and 7.2.13], in [HNW93, §§II.4 and II.9], [HW96, §§IV.2 and IV.8], [QSS07, §11.8.2], and [Gau12, §5.8] for Runge–Kutta methods, in [HNW93, §III.5] for linear multistep methods, and in [MM02, §III.6] for explicit methods.

7.1.6 Notes and references

The material from Section 7.1.1 can be found in textbooks such as [CL55, Chap. 1 and Chap. 13 §1], [Har02, Chap. II–III], [Kha02, Chap. 3–4], [Tes12, Chap. 2 and §6.5], and [LR14, Chap. 4–5]. The local existence theorem (Theorem 7.1.6) is due to Peano, who proved it in 1890. Modern proofs rely on the Arzelà–Ascoli theorem to show that the forward Euler method converges to a solution; see, e.g., [LR14, §4.1].

The rest of Section 7.1 is based on [SB02, §§7.0–7.2], [SM03, Chap. 12], [QSS07, Chap. 11], [Gau12, Chap. 5], and [SW22, Part IV] and the specialized books [HNW93, HW96, MM02]. Except [HNW93, HW96, SW22], these references focus on $n = 1$.

7.2 Boundary-value problems

Appendix A

Elements of linear algebra and matrix theory

A.1 Eigenvalues and eigenspaces

Let $n \in \mathbb{N} \setminus \{0,1\}$ and $A \in \mathbb{C}^{n\times n}$. The *characteristic polynomial* of A is $\det(XI_n - A)$, where X is the indeterminate. It holds that $\det(XI_n - A) = X^n + \sum_{i=0}^{n-1} \alpha_i X^i$ with $\alpha_{n-1} = -\operatorname{tr} A$ and $\alpha_0 = (-1)^n \det A$. Moreover, for every invertible $P \in \mathbb{C}^{n\times n}$, A and $P^{-1}AP$ have the same characteristic polynomial. By the fundamental theorem of algebra, the characteristic polynomial of A has n complex roots, which are called the *eigenvalues* of A. If $\lambda \in \mathbb{C}$ is an eigenvalue of A, then $\ker(\lambda I_n - A)$ is called the *eigenspace* associated with λ. If the sum of the eigenspaces of A equals $\mathbb{C}^n$, then A is said to be *diagonalizable* and there exists an invertible $P \in \mathbb{C}^{n\times n}$ such that

$$P^{-1}AP = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

where $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ are the eigenvalues of A.

A.2 Symmetric positive-semidefinite matrices

A symmetric $A \in \mathbb{R}^{n\times n}$ is said to be *positive-semidefinite* if, for all $x \in \mathbb{R}^n \setminus \{0\}$, $x^\top Ax \geq 0$; A is said to be *positive-definite* if the strict inequality holds. If $A \in \mathbb{R}^{n\times n}$ is symmetric, then its eigenvalues are real. A symmetric $A \in \mathbb{R}^{n\times n}$ is positive-semidefinite if and only if all its eigenvalues are nonnegative, and positive-definite if and only if all its eigenvalues are positive. If $A \in \mathbb{R}^{n\times n}$ is symmetric positive-semidefinite and $k \in \mathbb{N} \setminus \{0,1\}$, then there exists a unique symmetric positive-semidefinite $B \in \mathbb{R}^{n\times n}$ such that $B^k = A$; B is called the kth root of A and denoted by $A^{\frac{1}{k}}$ [HJ12, Theorem 7.2.6].

A.3 Linear and multilinear maps

This section is based on [Die69, §V.7]. Let X and Y be finite-dimensional real vector spaces. The real vector space of all linear maps from X to Y is denoted by $\mathcal{L}(X,Y)$. Let $p \in \mathbb{N} \setminus \{0\}$ and $X^p := \bigtimes_{i=1}^{p} X$ be the pth Cartesian power of X. A map $L : X^p \to Y$ is said to be p-linear if it is linear in each of its p variables. The real vector space of all p-linear maps from X^p to Y is denoted by $\mathcal{L}_p(X,Y)$. A map $L \in \mathcal{L}_p(X,Y)$ is said to be symmetric if, for all $x_1, \ldots, x_p \in X$ and every permutation $s : \{1,\ldots,p\} \to \{1,\ldots,p\}$, $L(x_{s(1)}, \ldots, x_{s(p)}) = L(x_1, \ldots, x_p)$. A map $L \in \mathcal{L}_2(X,Y)$ can be identified with the map $\tilde{L} \in \mathcal{L}(X, \mathcal{L}(X,Y))$ defined by $(\tilde{L}x_1)x_2 = L(x_1,x_2)$ for all $x_1, x_2 \in X$.

Appendix B

Elements of analysis

B.1 Function, graph, domain, and image

A *function* is a triplet $f \coloneqq (A, B, G)$, where A and B are nonempty sets and G is a nonempty subset of $A \times B$ such that, for every $x \in A$, there exists at most one $y \in B$ such that $(x, y) \in G$; if such a y exists, it is called the *image* of x under f and denoted by $f(x)$. The sets A, B, and G are respectively called the *set of departure*, the *set of destination*, and the *graph* of f. Moreover, f is said to be a function from A to B, written $f : A \to B$. The *domain* of f, denoted by $\operatorname{dom} f$, is the nonempty set of all $x \in A$ such that there exists $y \in B$ such that $(x, y) \in G$; f is said to be defined on a nonempty set X if $X \subseteq \operatorname{dom} f$. The *image* of a nonempty set $X \subseteq \operatorname{dom} f$ under f is $f(X) \coloneqq \{f(x) \mid x \in X\}$. The *image* of f is $f(\operatorname{dom} f)$.

B.2 Normed spaces

Definition B.2.1. A *norm* on a real vector space X is a real-valued function $\|\cdot\|$ defined on X such that:

1. $\|x\| \neq 0$ for all $x \in X \setminus \{0\}$;
2. $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{R}$ and $x \in X$;
3. $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$.

A real normed vector space is a pair $(X, \|\cdot\|)$, where X is a real vector space and $\|\cdot\|$ is a norm on X.

In this text, real normed vector spaces are simply called normed spaces. Moreover, if $(X, \|\cdot\|)$ is a normed space and there is no risk of confusion, X is sometimes called a normed space.

Let X be a normed space. The *open* and *closed balls* of center $x \in X$ and radius $r \in (0, \infty)$ are respectively $B_X(x, r) \coloneqq \{y \in X \mid \|x - y\| < r\}$ and $B_X[x, r] \coloneqq \{y \in X \mid \|x - y\| \leq r\}$; when there is no risk of confusion, $B_X(x, r)$ and $B_X[x, r]$ are simply denoted by $B(x, r)$ and $B[x, r]$, respectively.

A sequence $(x_i)_{i \in \mathbb{N}}$ in X is called a *Cauchy sequence* if, for every $\varepsilon \in (0, \infty)$, there exists $k \in \mathbb{N}$ such that, for all integers $i, j \geq k$, $\|x_i - x_j\| \leq \varepsilon$. A sequence $(x_i)_{i \in \mathbb{N}}$ in X is said to be *convergent* if there exists $x \in X$ such that, for every $\varepsilon \in (0, \infty)$, there exists $k \in \mathbb{N}$ such that, for all integers $i \geq k$, $\|x_i - x\| \leq \varepsilon$. Every convergent sequence is a Cauchy sequence. The normed space X is said to be *complete* if every Cauchy sequence in X converges in X. A subset of X is said to be *compact* if every sequence in the subset contains a convergent subsequence.

The *interior* of a subset U of X is the set of all $x \in U$ such that there exists $\delta \in (0, \infty)$ such that $B(x, \delta) \subseteq U$. A subset of X is said to be *open* if it is equal to its interior. A subset U of X is said to be *closed* if $X \setminus U$ is open. If a sequence contained in a closed subset of X converges, then the limit is in the subset. A subset U of X is said to be *bounded* if there exists $r \in (0, \infty)$ such that $U \subseteq B[0, r]$.

Let Y be a normed space. A function $f : X \to Y$ is said to be *continuous* at $x \in \operatorname{dom} f$ if, for every $\varepsilon \in (0, \infty)$, there exists $\delta \in (0, \infty)$ such that $f(B_X[x, \delta] \cap \operatorname{dom} f) \subseteq B_Y[f(x), \varepsilon]$. A function is said to be continuous on a nonempty subset of its domain if it is continuous at every point of the subset.

A function is said to be continuous if it is continuous on its domain. The reverse triangle inequality, which states that, for all $x, y \in X$,

$$|\|x\| - \|y\|| \leq \|x - y\|,$$

implies that the norm $\|\cdot\|$ is continuous. A point $y \in Y$ is called a *limit* of a function $f : X \to Y$ at $x \in X$ if, for every $\varepsilon \in (0, \infty)$, there exists $\delta \in (0, \infty)$ such that the set $(B_X[x, \delta] \cap \operatorname{dom} f) \setminus \{x\}$ is nonempty and its image under f is contained in $B_Y[y, \varepsilon]$; if such a y exists, it is unique and denoted by $\lim_x f$ or $\lim_{z \to x} f(z)$.

A norm $|||\cdot|||$ on X is said to be *equivalent* to the norm $\|\cdot\|$ if there exist $a, b \in (0, \infty)$ such that, for all $x \in X$,

$$a\|x\| \leq |||x||| \leq b\|x\|.$$

The equivalence of norms is an equivalence relation. Two equivalent norms have the same properties, e.g., the same Cauchy and convergent sequences, the same open, closed, bounded, and compact sets, and the same continuous functions.

Let $m, n \in \mathbb{N} \setminus \{0\}$. Examples of norms on $\mathbb{R}^n$ are given in Table 1.1.

Theorem B.2.2 (Bolzano–Weierstrass)**.** *A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Theorem B.2.3 (extreme-value theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ and C be a nonempty subset of $\operatorname{dom} f$. If f is continuous on C and C is compact, then there exist $x, y \in C$ such that $f(C) \subseteq [f(x), f(y)]$.*

Theorem B.2.4. *On a finite-dimensional real vector space, all norms are equivalent.*

Theorem B.2.5. *Every finite-dimensional normed space is complete.*

Proposition B.2.6. *Let $\|\cdot\|$ and $|||\cdot|||$ be norms on $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. Let X be $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ or $\mathbb{R}^{m \times n}$. Then, the function*

$$X \to \mathbb{R} : L \mapsto \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Lx\|}{|||x|||}$$

is a norm on X.

In the preceding proposition, if $m = n$ and $\|\cdot\| = |||\cdot|||$, then the norm on X is called the norm induced by $\|\cdot\|$ and is also denoted by $\|\cdot\|$.

The preceding proposition can be extended to multilinear maps.

Proposition B.2.7. *Let X and Y be two finite-dimensional normed spaces. For every $p \in \mathbb{N} \setminus \{0\}$, the function*

$$\mathcal{L}_p(X, Y) \to \mathbb{R} : L \mapsto \sup_{x_1, \ldots, x_p \in X \setminus \{0\}} \frac{\|L(x_1, \ldots, x_p)\|_Y}{\|x_1\|_X \ldots \|x_p\|_X}$$

is a norm on $\mathcal{L}_p(X, Y)$.

For every nonempty subset U of $\mathbb{R}^n$, $\mathrm{C}^0(U, \mathbb{R})$ denotes the set of all continuous $f : \mathbb{R}^n \to \mathbb{R}$ such that $\operatorname{dom} f = U$. If U is compact, then the function

$$\|\cdot\|_\infty : f \mapsto \max_{x \in U} |f(x)|$$

is a norm on $\mathrm{C}^0(U, \mathbb{R})$.

B.3 Derivative

In this section, based on [Die69, Chap. VIII], X and Y are finite-dimensional normed spaces. Let $f : X \to Y$ such that the interior U of $\operatorname{dom} f$ is nonempty. The function f is said to be *differentiable* at $x \in U$ if there exists $L \in \mathcal{L}(X, Y)$ such that

$$\lim_{h \to 0} \frac{\|f(x + h) - f(x) - Lh\|_Y}{\|h\|_X} = 0.$$

If such an L exists, it is unique, denoted by $f'(x)$, and called the *derivative* of f at x. If f is differentiable at every point of U, then f is said to be differentiable and the function $f' : U \to \mathcal{L}(X,Y) : x \mapsto f'(x)$ is called the derivative of f. If f is differentiable and f' is continuous, then f is said to be *continuously differentiable*.

If f is differentiable and f' is differentiable, then f is said to be *twice differentiable* and the derivative of f' is called the *second derivative* of f and denoted by f''. For every $x \in U$, $f''(x)$ is an element of the real vector space $\mathcal{L}(X, \mathcal{L}(X,Y))$, which can be identified with $\mathcal{L}_2(X,Y)$, as seen in Section A.3. Moreover, $f''(x)$ is symmetric: $f''(x)(v,u) = f''(x)(u,v)$ for all $u, v \in X$.

Define $f^{(0)} := f$, $f^{(1)} := f'$, and $f^{(2)} := f''$. By induction, for every $p \in \mathbb{N} \setminus \{0\}$, if $f^{(p-1)}$ is differentiable, then its derivative is called the pth derivative of f, denoted by $f^{(p)}$, and f is said to be p times differentiable. For every $x \in U$, $f^{(p)}(x) \in \mathcal{L}_p(X,Y)$ is symmetric.

For every nonempty subset V of X and every $p \in \mathbb{N} \setminus \{0\}$, $\mathrm{C}^p(V,Y)$ denotes the set of all $g : X \to Y$ such that the interior of $\operatorname{dom} g$ contains V, g is p times differentiable on V, and $g^{(p)}$ is continuous on V.

Theorem B.3.1 (chain rule)**.** *Assume that f is continuous on U. Let V be an open subset of Y that contains $f(x)$. Let Z be a finite-dimensional normed space. Let $g : V \to Z$ be continuous on V. If f is differentiable at $x \in U$ and g is differentiable at $f(x)$, then $g \circ f$ is differentiable at x and*

$$(g \circ f)'(x) = g'(f(x)) \circ f'(x).$$

Theorem B.3.2 (mean-value theorem)**.** *Let x and y be two distinct points in U such that U contains the segment $\{(1-t)x + ty \mid t \in [0,1]\}$. If f is continuous on U and differentiable at every point of $\{(1-t)x + ty \mid t \in (0,1)\}$, then*

$$\|f(x) - f(y)\|_Y \leq \|x - y\|_X \sup_{t \in (0,1)} \|f'((1-t)x + ty)\|,$$

where $\|\cdot\|$ is the norm defined in Proposition B.2.6 (or Proposition B.2.7 with $p = 1$).

Theorem B.3.3 (Taylor's theorem)**.** *Let x and y be two distinct points in U such that U contains the segment $\{(1-t)x + ty \mid t \in [0,1]\}$. Let $p \in \mathbb{N} \setminus \{0\}$. If f is p times differentiable on U, then*

$$\left\| f(y) - \sum_{i=0}^{p-1} \frac{1}{i!} f^{(i)}(x)(y-x)^i \right\|_Y \leq \frac{\|x - y\|_X^p}{p!} \sup_{t \in (0,1)} \|f^{(p)}((1-t)x + ty)\|,$$

where $\|\cdot\|$ is the norm defined in Proposition B.2.7.

If $X = \mathbb{R}$, then f is differentiable at $x \in U$ if and only if the limit

$$\lim_{h \to 0} \frac{f(t+h) - f(x)}{h}$$

exists, in which case, $f'(x) \in \mathcal{L}(\mathbb{R}, Y)$ can be identified with the limit, which is in Y. This limit can be computed if, for every $\delta \in (0, \infty)$, there exists $y \in \operatorname{dom} f \setminus \{x\}$ such that $|x - y| \leq \delta$; x does not need to be in the interior of $\operatorname{dom} f$.

A subset I of $\mathbb{R}$ is called an *interval* if it contains at least two elements and, for every $a, b \in I$ such that $a < b$, $[a,b] \subseteq I$. Thus, every interval can be written as (a,b), $[a,b)$, $(a,b]$, $[a,b]$, $(-\infty, a)$, $(-\infty, a]$, (a, ∞), or $[a, \infty)$ with $a, b \in \mathbb{R}$ and $a < b$.

If $X = \mathbb{R}$ and V is an interval, then the definition $\mathrm{C}^p(V,Y)$ can be broadened as follows: $\mathrm{C}^p(V,Y)$ denotes the set of all $g : X \to Y$ such that $V \subseteq \operatorname{dom} g$, g is p times differentiable on V, and $g^{(p)}$ is continuous on V. All results from this section extend to that broader setting.

Theorem B.3.4 (Taylor's theorem for real-valued functions of a real variable)**.** *Let $p \in \mathbb{N} \setminus \{0\}$ and $a, b \in \mathbb{R}$ such that $a < b$. If $f \in \mathrm{C}^{p-1}([a,b], \mathbb{R})$ and $f^{(p-1)}$ is differentiable on (a,b), then there exist $c \in (a,b)$ such that*

$$f(b) = \sum_{i=0}^{p-1} \frac{f^{(i)}(a)}{i!}(b-a)^i + \frac{f^{(p)}(c)}{p!}(b-a)^p.$$

B.4 Integral

Let $n \in \mathbb{N} \setminus \{0\}$ and $a, b \in \mathbb{R}$ such that $a < b$. For every $f \in \mathrm{C}^0([a,b], \mathbb{R}^n)$, $\int_a^b f$ is defined componentwise.

Theorem B.4.1 (fundamental theorem of calculus)**.** *If $f \in \mathrm{C}^1([a,b], \mathbb{R}^n)$, then*

$$\int_a^b f' = f(b) - f(a).$$

B.5 Inner product

Definition B.5.1. An *inner product* on a real vector space X is a real-valued function $\langle \cdot, \cdot \rangle$ defined on $X \times X$ such that:

1. $\langle x, x \rangle > 0$ for all $x \in X \setminus \{0\}$;
2. $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$ for all $\alpha, \beta \in \mathbb{R}$ and $x, y, z \in X$;
3. $\langle y, x \rangle = \langle x, y \rangle$ for all $x, y \in X$.

If $\langle \cdot, \cdot \rangle$ is an inner product on a real vector space X, then the function $X \to \mathbb{R} : x \mapsto \sqrt{\langle x, x \rangle}$ is a norm on X called the norm induced by $\langle \cdot, \cdot \rangle$.

Let $n \in \mathbb{N} \setminus \{0\}$. The function

$$\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} : (x, y) \mapsto y^\top x = \sum_{i=1}^n x_i y_i$$

is an inner product on $\mathbb{R}^n$ called the usual inner product on $\mathbb{R}^n$.

Bibliography

[Bru78] L. Brutman. On the Lebesgue function for polynomial interpolation. *SIAM J. Numer. Anal.*, 15(4):694–704., 1978. doi:10.1137/0715046.

[CL55] E. A. Coddington and N. Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, NY, 1955.

[Dah56] G. G. Dahlquist. Convergence and stability in the numerical integration of ordinary differential equations. *Math. Scand.*, 4:33–53, 1956. doi:10.7146/math.scand.a-10454.

[Dah63] G. G. Dahlquist. A special stability problem for linear multistep methods. *BIT*, 3:27–43, 1963. doi:10.1007/BF01963532.

[Dem97] J. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997. doi:10.1137/1.9781611971446.

[Die69] J. Dieudonné. *Foundations of Modern Analysis*, volume 10-I of *Pure Appl. Math.* Academic Press, New York, NY, 1969. Enlarged and Corrected Printing.

[Gau12] W. Gautschi. *Numerical Analysis*. Birkhäuser, Boston, MA, 2nd edition, 2012. doi:10.1007/978-0-8176-8259-0.

[GI87] W. Gautschi and G. Inglese. Lower bounds for the condition number of Vandermonde matrices. *Numer. Math.*, 52:241–250, May 1987. doi:10.1007/BF01398878.

[GV13] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 4th edition, 2013. doi:10.56021/9781421407944.

[Har02] P. Hartman. *Ordinary Differential Equations*, volume 38 of *Classics Appl. Math.* SIAM, Philadelphia, PA, 2002. doi:10.1137/1.9780898719222.

[Hig02] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, PA, 2nd edition, 2002. doi:10.1137/1.9780898718027.

[HJ12] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2nd edition, 2012. doi:10.1017/CBO9781139020411.

[HNW93] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Ser. Comput. Math.* Springer, Berlin, Heidelberg, 2nd edition, 1993. Corrected 3rd printing 2008. doi:10.1007/978-3-540-78862-1.

[HW96] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, volume 14 of *Springer Ser. Comput. Math.* Springer, Berlin, Heidelberg, 2nd edition, 1996. Corrected 2nd printing 2002. doi:10.1007/978-3-642-05221-7.

[Kha02] H. K. Khalil. *Nonlinear systems*. Prentice-Hall, Upper Saddle River, NJ, third edition, 2002.

[LR14] H. Logemann and E. P. Ryan. *Ordinary Differential Equations: Analysis, Qualitative Theory and Control*. Springer Undergrad. Math. Ser. Springer, London, 2014. doi:10.1007/978-1-4471-6398-5.

[MM02] R. Mattheij and J. Molenaar. *Ordinary Differential Equations in Theory and Practice*, volume 43 of *Classics Appl. Math.* SIAM, Philadelphia, PA, 2002. doi:10.1137/1.9780898719178.

[NW06] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Ser. Oper. Res. Financ. Eng. Springer, New York, NY, 2nd edition, 2006. doi:10.1007/978-0-387-40065-5.

[OR00] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*, volume 30 of *Classics Appl. Math.* SIAM, Philadelphia, PA, 2000. doi:10.1137/1.9780898719468.

[QSS07] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*, volume 37 of *Texts Appl. Math.* Springer, Berlin, Heidelberg, 2 edition, 2007. doi:10.1007/b98885.

[Riv74] T. J. Rivlin. The Lebesgue constants for polynomial interpolation. In H. G. Garnir, K. R. Unni, and J. H. Williamson, editors, *Functional Analysis and its Applications*, volume 399 of *Lecture Notes in Math.*, pages 422–437, Berlin, Heidelberg, 1974. Springer. doi:10.1007/BFb0063594.

[Saa03] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, 2nd edition, 2003. doi:10.1137/1.9780898718003.

[SB02] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*, volume 12 of *Texts Appl. Math.* Springer, New York, NY, 3rd edition, 2002. doi:10.1007/978-0-387-21738-3.

[SM03] E. Süli and D. F. Mayers. *An Introduction to Numerical Analysis*. Cambridge University Press, Cambridge, 2003. doi:10.1017/CBO9780511801181.

[Ste98] G. W. Stewart. *Matrix Algorithms: Basic Decompositions*. SIAM, Philadelphia, PA, 1998. doi:10.1137/1.9781611971408.

[SW22] A. J. Salgado and S. M. Wise. *Classical Numerical Analysis: A Comprehensive Course*. Cambridge University Press, Cambridge, 2022. doi:10.1017/9781108942607.

[TB97] L. N. Trefethen and D. Bau, III. *Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997. doi:10.1137/1.9781611977165.

[Tes12] Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Grad. Stud. Math.* American Mathematical Society, Providence, RI, 2012. URL: https://bookstore.ams.org/gsm-140.

[Tre19] L. N. Trefethen. *Approximation Theory and Approximation Practice, Extended Edition*. SIAM, Philadelphia, PA, 2019. doi:10.1137/1.9781611975949.

[TW91] L. N. Trefethen and J. A. C. Weideman. Two results on polynomial interpolation in equally spaced points. *J. Approx. Theory*, 65(3):247–260, 1991. doi:10.1016/0021-9045(91)90090-W.