Chapter 3
Finite Difference Methods

3.1 Notations

Remark 3.1. Idea. The basic idea of finite difference methods consists in approximating the derivatives of a differential equation with appropriate finite differences. Consider a decomposition of the interval $[0, 1]$, which is at the moment assumed to be equidistant:

$$\begin{aligned} x_i &= ih, \quad i = 0, \ldots, N, \quad h = 1/N, \\ \omega_h &= \{x_i \ : \ i = 0, \ldots, N\} \ - \text{grid, mesh.} \end{aligned}$$

The values x_i are called grid points or nodes and h is the mesh width. □

Definition 3.2. Grid function. A vector $\boldsymbol{u}_h = (u_0, \ldots, u_N)^T \in \mathbb{R}^{N+1}$, which assigns to each node a value, is called grid function. The restriction of a function $u \in C([0, 1])$ to a grid function is denoted by $R_h u$, i.e.,

$$R_h u := (u(x_0), u(x_1), \ldots, u(x_N))^T .$$

□

Example 3.3. Grid function. Consider a grid with the nodes $\{0, 0.25, 0.5, 0.75, 1\}$. Then, the grid function of $u(x) = x^2$ is

$$R_h u = \left(0, \frac{1}{16}, \frac{1}{4}, \frac{9}{16}, 1\right)^T .$$

Different functions might have for a given grid the same grid function. Consider, e.g., $u(x) = \sin(4\pi x)$ on the same grid as used above. The corresponding grid function is

$$R_h u = (0, 0, 0, 0, 0)^T .$$

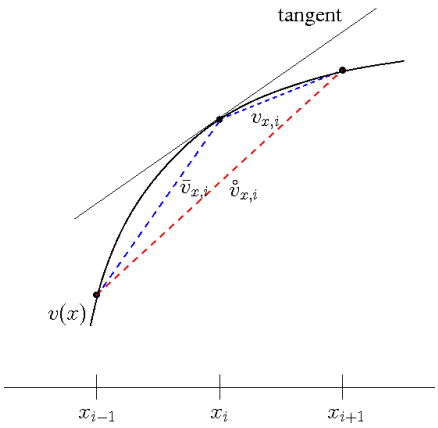


Fig. 3.1 Finite differences.

This grid function is obviously also the grid function of $u(x) = 0$. The considered grid is too coarse to represent the $u(x) = \sin(4\pi x)$ in a reasonable way. □

Definition 3.4. Finite difference operators. Let $v(x)$ be a sufficiently smooth function and denote $v_i = v(x_i)$, where x_i are the nodes of the grid. The following difference quotients (finite differences) are called

$$\begin{aligned}
D^+v(x_i) = v_{x,i} &= \frac{v_{i+1} - v_i}{h} && \text{– forward difference,}\\
D^-v(x_i) = v_{\bar{x},i} &= \frac{v_i - v_{i-1}}{h} && \text{– backward difference,}\\
D^0v(x_i) = v_{\mathring{x},i} &= \frac{v_{i+1} - v_{i-1}}{2h} && \text{– central difference,}\\
D^+D^-(v)(x_i) = v_{\bar{x}x,i} &= \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} && \text{– second order difference,}
\end{aligned}$$

see Figure 3.1. □

Remark 3.5. To the finite differences. The formula for $D^+D^-(v)(x_i)$ can be checked with a direct calculation. In addition, it is

$$D^0v(x_i) = \frac{1}{2}\left((D^+v(x_i) + D^-v(x_i)\right).$$

□

Definition 3.6. Consistency of a finite difference operator, discrete maximum norm. Let L be a differential operator. The finite difference operator $L_h : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ is said to be consistent with L of order k if

Fig. 3.2 Brook Taylor (1685 – 1731).

$$\max_{0\le i\le N} |(Lu)(x_i) - (L_h R_h u)_i| =: \|Lu - L_h R_h u\|_{\infty,d} = \mathcal{O}(h^k).$$

Here, $\|\cdot\|_{\infty,d}$ is the discrete maximum norm in the space of grid functions. □

Example 3.7. Orders of consistency for standard finite difference operators. The consistency is a measure of the approximation property of L_h. Applying a Taylor series expansion for $v(x)$ at the node x_i yields

$$\begin{aligned} D^+v(x_i) &= v'(x_i) + \mathcal{O}(h), \\ D^-v(x_i) &= v'(x_i) + \mathcal{O}(h), \\ D^0v(x_i) &= v'(x_i) + \mathcal{O}(h^2), \\ D^+D^-(v)(x_i) &= v''(x_i) + \mathcal{O}(h^2). \end{aligned}$$

The finite difference operators $D^+v(x_i), D^-v(x_i), D^0v(x_i)$ are consistent to $L = \frac{d}{dx}$ of first, first, and second order, respectively. The operator $D^+D^-(v)(x_i)$ is consistent of second order to $L = \frac{d^2}{dx^2}$. □

3.2 Classical Convergence Theory for Central Difference Schemes

Remark 3.8. Contents of this section. This section considers the two-point boundary value problem

$$Lu := -u'' + b(x)u' + c(x)u = f(x), \quad \text{for } x \in (0,1), \quad u(0) = u(1) = 0, \tag{3.1}$$

i.e., the model problem with $\varepsilon = 1$. The classical convergence theory will be presented. It will be assumed that the parameter functions b, c, f are sufficiently smooth and that $c(x) \geq 0$ for all $x \in [0,1]$. □

Definition 3.9. Central difference scheme. The central difference scheme for (3.1) has the form

$$\begin{aligned} (L_h u_h)_i := -D^+D^- u_i + b_i D^0 u_i + c_i u_i = f_i, \quad \text{for } i = 1, \ldots, N-1, \\ u_0 = u_N = 0, \end{aligned} \tag{3.2}$$

with $u_h = (u_0, u_1, \ldots, u_N)^T$. □

Remark 3.10. To the central difference scheme.

- The central difference scheme leads to a tridiagonal system of linear equations

$$r_i u_{i-1} + s_i u_i + t_i u_{i+1} = f_i, \quad i = 1, \ldots, N-1, \quad u_0 = u_N = 0,$$

with

$$r_i = -\frac{1}{h^2} - \frac{1}{2h} b_i, \quad s_i = c_i + \frac{2}{h^2}, \quad t_i = -\frac{1}{h^2} + \frac{1}{2h} b_i.$$

- The following questions have to be answered:
  - Which properties has the discrete problem (3.2)?
  - What can be said about the error $\|u - u_h\|_{\infty,d}$?

  For answering these questions, the concepts of consistency and stability will be used.

□

Definition 3.11. Consistency of a difference scheme and order of consistency. Consider a difference scheme of the form $L_h u_h := R_h(Lu) = R_h(f)$. The boundary conditions should be integrated into this scheme such that the first and last row of L_h are identical to the first and last row of the identity matrix and it is $R_h(Lu)_0 = u_0$, $R_h(Lu)_N = u_N$. The scheme is called consistent of order k in the discrete maximum norm, if

$$\|L_h R_h u - R_h(Lu)\|_{\infty,d} \leq c h^k$$

for all functions u from a given function space, where the positive constants c and k are independent of h.

Thus the consistency measures the difference of applying the first grid function operator and then the finite difference operator to applying first the differential operator and then the grid function operator. □

Lemma 3.12. Consistency order of the central difference scheme. *Assume that $u \in C^4([0,1])$, then the central difference scheme* (3.2) *has consistency order* 2.

Proof. The proof uses a Taylor series expansion, *exercise.* ■

Definition 3.13. Stability of a difference scheme. A difference scheme $L_h u_h = f_h$ is called stable in the discrete maximum norm, if there is a stability constant c_S, which is independent of h, with

$$\|u_h\|_{\infty,d} \leq c_S \|L_h u_h\|_{\infty,d} = c_S \|f_h\|_{\infty,d}$$

for all grid functions u_h. □

Definition 3.14. Convergence of a difference scheme and order of convergence. A difference scheme for (3.1) is convergent of order k in the discrete maximum norm, if there are positive constants c and k, which are independent of h, such that

$$\|u_h - R_h u\|_{\infty,d} \leq c h^k.$$

□

Theorem 3.15. Consistency + stability $\Longrightarrow$ convergence. *A consistent and stable difference scheme is convergent. The orders of consistency and convergence are the same.*

Proof. It is

$$\begin{aligned} \|u_h - R_h u\|_{\infty,d} &\overset{\text{stab.}}{\leq} c_S \|L_h (u_h - R_h u)\|_{\infty,d} \overset{\text{lin.}}{=} c_S \|L_h u_h - L_h R_h u\|_{\infty,d} \\ &= c_S \|f_h - L_h R_h u\|_{\infty,d} = c_S \|R_h f - L_h R_h u\|_{\infty,d} \\ &= c_S \|R_h L u - L_h R_h u\|_{\infty,d} \overset{\text{cons.}}{\leq} K h^k, \end{aligned}$$

where the constant K is the product of the constants from the stability and consistency condition. ■

Remark 3.16. To consistency and stability. One has to prove consistency and stability.

- Consistency proofs are based often on Taylor series expansions and they are performed in a standard way.
- Stability proofs are not performed only at functions but they are performed at matrices and functions, see Definition 3.13. They are generally not simple and they require the introduction of some new notations.

□

Definition 3.17. Natural order of vectors and matrices, inverse-monotone matrices. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Then one writes $\boldsymbol{x} \leq \boldsymbol{y}$ if and only if $x_i \leq y_i$ for all $i = 1, \ldots, n$. The notation $\boldsymbol{x} \geq \mathbf{1}$ means that $x_i \geq 1$ for all $i = 1, \ldots, n$. Analogously, the notation $A \geq 0$ means for a matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ that $a_{ij} \geq 0$ for all $i, j = 1, \ldots, n$.

A matrix A, for which the inverse A^{-1} exists with $A^{-1} \geq 0$ is called inverse-monotone matrix. □

Fig. 3.3 Hermann Minkowski (1864 – 1909).

Lemma 3.18. Discrete comparison principle. *Let $A \in \mathbb{R}^{n\times n}$ be inverse-monotone. If $A\boldsymbol{v} \leq A\boldsymbol{w}$ for $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$, then it follows that $\boldsymbol{v} \leq \boldsymbol{w}$.*

Proof. From the assumption, it follows that

$$A(\boldsymbol{v} - \boldsymbol{w}) := \boldsymbol{b} \leq \boldsymbol{0}.$$

Multiplication with A^{-1} gives

$$\boldsymbol{v} - \boldsymbol{w} = A^{-1}\boldsymbol{b} \leq \boldsymbol{0}.$$

The last inequality follows from the property that A is an inverse-monotone matrix. Non-negative matrix entries are multiplied with nonpositive vector entries of $\boldsymbol{b}$. The result is a vector with nonpositive components. ■

Remark 3.19. To Lemma 3.18. Lemma 3.18 is a discrete analog to the comparison principle from Corollary 2.34. Thus, in the case of inverse-monotone matrices, one can transfer one important property from the continuous to the discrete setting. □

Definition 3.20. M-matrix. A matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ is called an M-matrix, if:

1. $a_{ij} \leq 0$ for $i \neq j$,
2. A^{-1} exists with $A^{-1} \geq 0$.

Then name refers to Hermann Minkowski. □

Lemma 3.21. M-matrices have positive diagonal entries. *Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ be an M-matrix, then it is $a_{ii} > 0$, $i = 1, \ldots, n$.*

Proof. *Exercise.* ■

Remark 3.22. On M-matrices. M-matrices are an important subclass of inverse-monotone matrices, since the first property is an additional property, which

will become very useful in the analysis. However, in practice, the second condition of their definition is hard to check. But there are characterizations of M-matrices that are easier to check. □

Theorem 3.23. M-matrix criterion. *Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ with $a_{ij} \leq 0$ for $i \neq j$. Then A is an M-matrix if and only if there exists a vector $\boldsymbol{e} \in \mathbb{R}^n$, $\boldsymbol{e} > \mathbf{0}$, such that $A\boldsymbol{e} > \mathbf{0}$. In this case, one obtains for the row sum norm*

$$\left\|A^{-1}\right\|_\infty \leq \frac{\|\boldsymbol{e}\|_{\infty,d}}{\min_k \left(A\boldsymbol{e}\right)_k}. \tag{3.3}$$

The vector $\boldsymbol{e}$ is called majorizing element.

Proof. See the literature, e.g., Bohl (1981); Axelsson & Kolotilina (1990). ■

Remark 3.24. Concerning the M-matrix criterion.

- The following approach is often successful for the construction of a majorizing element.
  - ◦ Find a function $e(x) > 0$ such that $(Le)(x) > 0$ for $x \in (0,1)$. This function is a majorizing element of the differential operator L.
  - ◦ Restrict $e(x)$ to its corresponding grid function e_h.

  If the first step of this approach is possible and the discretization L_h of L is consistent, then this approach generally works, at least if the mesh width is sufficiently small. The matrix A is the matrix representation of L_h.
- The constant c_S in the definition of the stability can be estimated with (3.3)

$$\|u_h\|_{\infty,d} = \left\|A^{-1}f_h\right\|_{\infty,d} \leq \left\|A^{-1}\right\|_\infty \|f_h\|_{\infty,d} = \left\|A^{-1}\right\|_\infty \|L_h u_h\|_{\infty,d}\,.$$

  Hence, it holds for this constant

$$c_S \leq \frac{\|\boldsymbol{e}\|_{\infty,d}}{\min_k \left(A\boldsymbol{e}\right)_k}.$$

  Thus, from the M-matrix criterion, which is equivalent to the M-matrix property, it follows stability.
- If Dirichlet boundary conditions are prescribed, then the variables u_0 and u_N should be eliminated before Theorem 3.23 is applied.
- For the central finite difference scheme, the first requirement of the M-matrix criterion, $a_{ij} \leq 0$ for $i \neq j$, is satisfied if h is sufficiently small, see Remark 3.10 for the coefficients of the matrix. For sufficiently fine grids, $-1/h^2$ dominates $\pm b/h$.

□

Example 3.25. M-matrix criterion. Consider (3.1) with $b(x) \equiv 0$, i.e.,

$$Lu(x) = -u''(x) + c(x)u(x), \quad u(0) = u(1) = 0, \quad c(x) \geq 0 \text{ in } [0,1].$$

Choose $e(x) := \frac{1}{2}x(1-x)$, then it follows that

$$Le(x) = 1 + c(x)e(x) \geq 1.$$

Setting $e_h := R_h e$ gives for all nodes x_i

$$(L_h e_h)_i = -D^+ D^- e_{h,i} + c_i e_{h,i} = 1 + c_i e_{h,i} \geq 1,$$

because the second order finite difference discretizes the second derivative of a quadratic function in the interior nodes exactly, see Example 3.7. One gets

$$L_h e_h \geq (1, \ldots, 1)^T \quad \Longleftrightarrow \quad A\boldsymbol{e} \geq \mathbf{1}.$$

As bound for the stability constant, one obtains

$$c_S \leq \frac{\|\boldsymbol{e}\|_{\infty,d}}{\min_k (A\boldsymbol{e})_k} \leq \frac{e_h(1/2)}{1} = \frac{1/8}{1} = \frac{1}{8}.$$

This example shows that in the case $b(x) \equiv 0$ the M-matrix property holds without restrictions on the fineness of the grid. □

Lemma 3.26. Stability of the central finite difference scheme for sufficiently fine grids. *If the mesh width h is sufficiently small, then the central finite difference scheme* (3.2) *for the two-point boundary value problem* (3.1) *is stable in the discrete maximum norm. The stiffness matrix is an M-matrix.*

Proof. A majorizing element will be constructed. To this end, let $e(x)$ be the solution of the two-point boundary value problem

$$-e'' + b(x)e' = 1, \quad e(0) = e(1) = 0.$$

From the maximum principle, Lemma 2.31, it follows that $e(x) \geq 0$ for $x \in (0,1)$. In addition, $e(x)$ has no local minima in $(0,1)$ since in this case, one obtains from the equation that $-e''(x) = 1$ for the local minima, which is a contradiction. From the equation, it follows that there is no interval $(x_0, x_1) \subset (0,1)$ where $e(x) \equiv 0$. Altogether, it is $e(x) > 0$ for $x \in (0,1)$. Since $c(x) \equiv 0$, one gets with Corollary 2.39 that the given problem possesses a unique solution and that $e \in C([0,1])$. Hence, $e(x)$ is bounded. By construction, it is

$$Le(x) = -e''(x) + b(x)e'(x) + c(x)e(x) = 1 + c(x)e(x).$$

Let e_h be the grid function of $e(x)$. For interior nodes, one obtains with $c(x) \geq 0$

$$\begin{aligned}
(L_h e_h)_i &= (R_h Le)_i + (L_h e_h - R_h Le)_i \\
&= (R_h (1 + c(x)e(x)))_i + \left(-D^+ D^- e_h + b_i D^0 e_h + c_i e_h - 1 - c_i e_h\right)_i \\
&\geq 1 + \left(-D^+ D^- e_h + b_i D^0 e_h - 1\right)_i \\
&= \left(-D^+ D^- e_h + b_i D^0 e_h\right)_i .
\end{aligned}$$

Since e_h is the grid function that corresponds to $e(x)$, the expression in the last line approximates $-e''(x_i) + b(x_i)e'(x_i)(= 1)$ sufficiently well, if h sufficiently small, see the consistency estimate in Example 3.7. In particular, there is a $H > 0$ such that for all $h \in (0, H]$ it is

$$(L_h e_h)_i \geq \frac{1}{2}.$$

Since $a_{ij} \leq 0$ for sufficiently fine meshes, see Remark 3.24, one finds that L_h represents an M-matrix. Now, the proof of the theorem is finished since from the M-matrix property, it follows stability, see Remark 3.24. ■

Corollary 3.27. Second order convergence of the central difference scheme. *If $u \in C^4((0,1)) \cap C([0,1])$, then the central difference scheme* (3.2) *is convergent with second order.*

Proof. The statement follows with Theorem 3.15 by combining Lemma 3.12 and Lemma 3.26. ■

Example 3.28. Second order convergence of the central difference scheme for sufficiently fine meshes. Consider the two-point boundary value problem

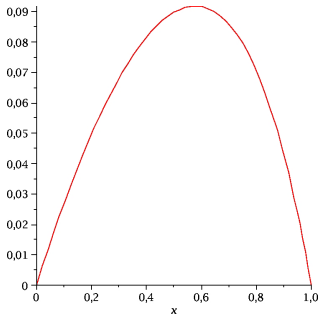
$$-u''(x) + 2u'(x) + 3u(x) = 1 \quad \text{in } (0,1), \quad u(0) = u(1) = 0.$$

The solution of this problem is

$$u(x) = \frac{1}{3}\left(1 + \frac{1 - e^{-1}}{e^{-1} - e^3}e^{3x} + \frac{e^3 - 1}{e^{-1} - e^3}e^{-x}\right).$$

One obtains the following errors for different mesh widths

no. of intervals N	$\|u - u_h\|_{\infty,d}$
4	4.2388e-4
8	9.8811e-5
16	2.4529e-5
32	6.1537e-6
64	1.5368e-6
128	3.8440e-7
256	9.6093e-8
512	2.4023e-8
1024	6.0058e-9



It can be observed that the error is reduced by the factor four if the mesh width is reduced by the factor two. This behavior is second order convergence. □

3.3 Upwind Schemes

Remark 3.29. Singularly perturbed two-point boundary value problem. From now on, finite difference schemes will be studied for the two-point boundary

value problem

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x), \quad \text{for } x \in (0,1), \tag{3.4}$$

with the boundary conditions

$$u(0) = u(1) = 0, \tag{3.5}$$

and the assumptions

$$\begin{aligned} \varepsilon &> 0, \\ b(x) &> 0 \quad \text{for all } x \in [0,1], \\ c(x) &\geq 0 \quad \text{for all } x \in [0,1], \end{aligned}$$

with sufficiently smooth functions $b(x), c(x)$, and $f(x)$. This problem is called singularly perturbed if $\varepsilon \ll \|b\|_{L^\infty((0,1))}$. The parameter ε is called singular perturbation parameter. With respect to the convection, it is only important that $b(x) \neq 0$ for all $x \in [0,1]$. If $b(x) < 0$ in $[0,1]$, then one obtains a problem of form (3.4) by applying the variable transform $x \mapsto 1 - x$.

If ε is small, then the solution of (3.4), (3.5) has in general a boundary layer at $x = 1$, see Example 2.8. This layer influences both the stability and the consistency of the numerical method. If the boundary values are chosen such that there is no boundary layer, then the consistency improves but the stability of the method might be still a problem. □

Example 3.30. Application of the central difference scheme to a simplified singularly perturbed problem. Consider the problem

$$-\varepsilon u'' + u' = 0 \text{ in } (0,1), \quad u(0) = 0,\ u(1) = 1.$$

The solution of this problem is

$$u(x) = \frac{e^{-(1-x)/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}.$$

Applying the transform $u(x) := x + v(x)$, one gets a problem with homogeneous boundary conditions. But one can apply the difference scheme directly to the problem with inhomogeneous boundary conditions. This approach is used in practice. The discrete problem has the form

$$-\varepsilon D^+ D^- u_i + D^0 u_i = 0, \quad u_0 = 0,\ u_N = 1$$

and it has the solution (*exercise*)

$$u_i = \frac{r^i - 1}{r^N - 1} \quad \text{with} \quad r = \frac{2\varepsilon + h}{2\varepsilon - h}.$$

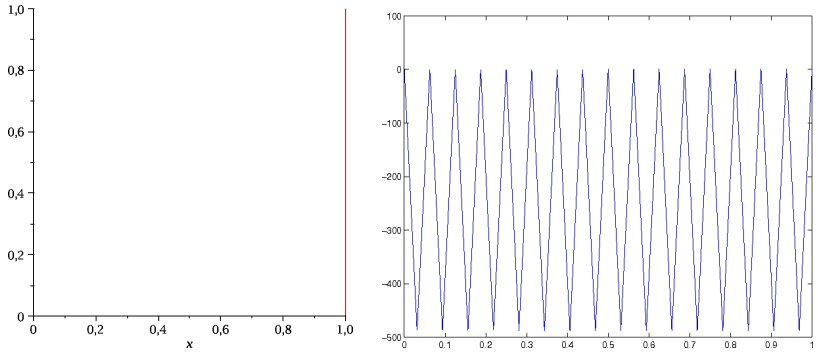


Fig. 3.4 Solution (left) and discrete solution with the central difference scheme (right) for $\varepsilon = 10^{-6}$ and $h = 1/32$.

In particular, it is $|r| > 1$, since the numerator is the sum of two positive numbers. The absolute value of this sum is always larger than the absolute value of the difference of these numbers.

If $h \gg 2\varepsilon$ then it is $r \approx -1$ and it follows that

$$u_i \approx \frac{(-1)^i - 1}{(-1)^N - 1}.$$

If N even, then one divides by a very small positive number, since $|r| > 1$. For i even, the numerator is also small and it is positive. Hence, the quotient is positive, too. For i odd, the numerator is negative and of order -2. In this case, the quotient is negative and its absolute value is large. The discrete solution is highly oscillating, see Figure 3.4.

If $h < 2\varepsilon$, then one obtains with the central difference scheme a useful approximation of the solution. However, in applications, it is often $\varepsilon \leq 10^{-6}$, i.e., one needs very fine grids in order to apply the central difference scheme. The use of such grids is possible in one dimension, but not in two or three dimensions. □

Remark 3.31. Application of the central difference scheme to the general singularly perturbed problem. Consider now the singularly perturbed problem (3.4), (3.5) and write the difference scheme in the form presented in Remark 3.10

$$r_i u_{i-1} + s_i u_i + t_i u_{i+1} = f_i, \quad i = 1, \ldots, N-1, \quad u_0 = u_N = 0,$$

with

$$r_i = -\frac{\varepsilon}{h^2} - \frac{1}{2h} b_i, \quad s_i = c_i + \frac{2\varepsilon}{h^2}, \quad t_i = -\frac{\varepsilon}{h^2} + \frac{1}{2h} b_i, \quad b_i > 0.$$

The first property for an M-matrix is satisfied if one assumes that

$$t_i \leq 0 \quad \Longleftrightarrow \quad h \leq h_0(\varepsilon) = \frac{2\varepsilon}{\|b\|_\infty}.$$

This assumption generalizes the observation from Example 3.30. Note that $h_0(\varepsilon) \to 0$ for $\varepsilon \to 0$. □

Remark 3.32. Motivations for upwind schemes. Another heuristic explanation for the failure of the central difference scheme in the case $\varepsilon \ll h$ is as follows. For small ε, the method applied to Example 3.30 has essentially the form

$$D^0 u_i = 0 \quad \Longleftrightarrow \quad \frac{u_{i+1} - u_{i-1}}{2h} = 0.$$

It follows for $i = N-1$ that $u_{N-2} \approx u_N = 1$, which is a very bad approximation of the exact value $u(x_{N-2}) \approx 0$.

This observations leads to the idea that for the approximation of $u'(x_{N-1})$, it is better not to use the value u_N. The simplest candidate which has this feature is the backward difference

$$u'(x_i) \approx \frac{u_i - u_{i-1}}{h}.$$

If one has the goal to modify the matrix entries which are obtained with the central difference scheme in such a way that one gets an M-matrix, i.e., $t_i \leq 0$, then one can also motivate the backward difference approximation of the first derivative, because this condition is satisfied at any rate if for the discretization of the convective term a contribution from the node x_{i+1} is not used. □

Definition 3.33. Simple upwind scheme. The simple upwind scheme for the singularly perturbed two-point boundary value problem (3.4), (3.5) has the form

$$\begin{aligned} -\varepsilon D^+ D^- u_i + b_i D^{\mathcal{N}} u_i + c_i u_i &= f_i, \quad \text{for } i = 1, \ldots, N-1, \\ u_0 = u_N &= 0, \end{aligned} \tag{3.6}$$

with

$$D^{\mathcal{N}} := \begin{cases} D^+ \text{ for } b < 0, \\ D^- \text{ for } b > 0. \end{cases}$$

□

Remark 3.34. Concerning the simple upwind scheme.

- In the upwind scheme, the finite difference approximation of the convective term is computed with values from the upwind direction. For convection-dominated problems, the transport of information occurs in the direction of convection. Hence, the upwind direction is the direction from which information is coming.

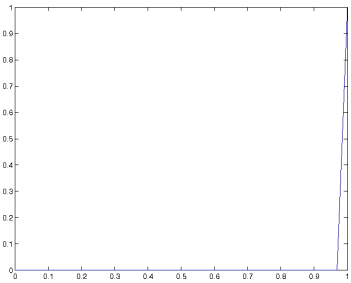


Fig. 3.5 Numerical solution of Example 3.30 with the simple upwind scheme for $\varepsilon = 10^{-6}$ and $h = 1/32$.

- Using the simple upwind scheme, one obtains a much better numerical solution for Example 3.30, see Figure 3.5.
- In the simple upwind scheme, the second order approximation D^0 is replaced by the first order approximation D^+ or D^-. This reduced order can be observed in the accuracy of the numerical results.
- Let L_h be the matrix of the simple upwind scheme after having eliminated the boundary values u_0 and u_N. In the form of Remark 3.10, this matrix has the form

$$r_i = -\frac{\varepsilon}{h^2} - \frac{1}{h}\max\{0, b_i\}, \quad s_i = c_i + \frac{2\varepsilon}{h^2} + \frac{1}{h}\left|b_i\right|, \quad t_i = -\frac{\varepsilon}{h^2} + \frac{1}{h}\min\{0, b_i\}.$$

  One can see that all non-diagonal entries are negative, independently of the size of ε and h.

□

Theorem 3.35. Stability of the simple upwind scheme. *Under the assumptions from Remark 3.29, the matrix L_h of the simple upwind scheme* (3.6) *is an M-matrix. The simple upwind scheme is uniformly stable with respect to the parameter ε, i.e., it is*

$$\|u_h\|_{\infty,d} \leq c_S \|L_h u_h\|_{\infty,d},$$

where the stability constant $c_S > 0$ is independent of ε and h.

Proof. Consider only the case $b(x) \geq \beta > 0$. The goal consists in constructing an appropriate majorizing element and to apply the M-matrix criterion. To this end, choose $e(x) = x$, which gives

$$Le(x) = -\varepsilon e''(x) + b(x)e'(x) + c(x)e(x) = b(x) + xc(x) \geq \beta.$$

For the simple upwind scheme and the corresponding grid function e_h, one obtains

$$\begin{aligned}(L_h e_h)_i &= r_i x_{i-1} + s_i x_i + t_i x_{i+1} \\ &= \left(-\frac{\varepsilon}{h^2} - \frac{1}{h} b_i\right)(x_i - h) + \left(c_i + \frac{2\varepsilon}{h^2} + \frac{1}{h} b_i\right) x_i - \frac{\varepsilon}{h^2}(x_i + h) \\ &= \left(-\frac{\varepsilon}{h^2} - \frac{1}{h} b_i + c_i + \frac{2\varepsilon}{h^2} + \frac{1}{h} b_i - \frac{\varepsilon}{h^2}\right) x_i + \left(\frac{\varepsilon}{h^2} + \frac{1}{h} b_i - \frac{\varepsilon}{h^2}\right) h \\ &= c_i x_i + b_i \geq \beta.\end{aligned}$$

It follows from Theorem 3.23, the M-matrix criterion, that L_h is an M-matrix. Using the estimate of the stability constant from Remark 3.24, one gets

$$c_S \leq \frac{\|e_h\|_{\infty,d}}{\min_k (L_h e_h)_k} = \frac{1}{\beta}.$$

■

Lemma 3.36. Estimates of the norm of derivatives of the solution. *Let $b(x) \geq \beta > 0$ and $b(x), c(x), f(x)$ be sufficiently smooth. Then, it is for the solution $u(x)$ of (3.4), (3.5)*

$$\left|u^{(i)}(x)\right| \leq C\left[1 + \varepsilon^{-i} \exp\left(-\beta \frac{1-x}{\varepsilon}\right)\right], \quad i = 1, 2, \ldots, q,$$

for $x \in [0,1]$. The maximal order q depends on the smoothness of the data.

Proof. The proof was performed in Kellogg & Tsan (1978), see also (Roos *et al.*, 2008, p. 21). ■

Theorem 3.37. Consistency of the simple upwind scheme. *Under the assumptions from Remark 3.29 with $b(x) \geq \beta > 0$, there is a positive constant β^*, which depends only on β, such that the error committed by the simple upwind scheme (3.6) in the inner nodes $\{x_i : i = 1, \ldots, N-1\}$ can be bounded as follows*

$$|u(x_i) - u_i| \leq \begin{cases} Ch\left[1 + \varepsilon^{-1} \exp\left(-\beta^* \dfrac{1-x_i}{\varepsilon}\right)\right] & \text{if } h < \varepsilon, \\ Ch + C \exp\left(-\beta^* \dfrac{1-x_{i+1}}{\varepsilon}\right) & \text{if } h \geq \varepsilon. \end{cases} \tag{3.7}$$

Proof. The proof was performed in Kellogg & Tsan (1978). Here, only the interesting case $h \geq \varepsilon$ will be considered and also for this case, not the complete proof will be presented. The complete proof can be found in (Roos *et al.*, 2008, pp. 49).

In the case $h \geq \varepsilon$, one decomposes the solution of (3.4), (3.5) into

$$u(x) = -u_0(1) \exp\left(-\frac{b(1)(1-x)}{\varepsilon}\right) + z(x) =: v(x) + z(x),$$

where $u_0(x)$ is the reduced solution. Those properties of $u(x)$ are transferred to the part $z(x)$ which $v(x)$ does not possess. One finds, analogously to Lemma 3.36, that

$$\left|z^{(k)}(x)\right| \leq C\left[1 + \varepsilon^{1-k} \exp\left(-\frac{b(1)(1-x)}{\varepsilon}\right)\right], \quad k = 1, 2, 3. \tag{3.8}$$

It is

$$L_h u_h = f_h = R_h(f) = R_h(Lu) = R_h\left(L(v+z)\right) = R_h(Lv) + R_h(Lz).$$

In this way, a decomposition $u_h = v_h + z_h$ is obtained, where the grid functions are defined as the solutions of the following discrete problems

$$L_h v_h = R_h(Lv) \quad \text{and} \quad L_h z_h = R_h(Lz). \tag{3.9}$$

The functions v_h and z_h coincide with $v(x)$ and $z(x)$, respectively, in x_0 and x_N. Applying the triangle inequality gives

$$|u(x_i) - u_i| = |v(x_i) + z(x_i) - (v_i + z_i)| \leq |v(x_i) - v_i| + |z(x_i) - z_i|\,.$$

The estimate of the second term starts with the stability estimate from Theorem 3.35, where it is used that z_h is the solution of a discrete problem with simple upwind, and the definition of $L_h z_i$ from (3.9)

$$|z(x_i) - z_i| \leq C\,|L_h z(x_i) - L_h z_i| = C\,|L_h z(x_i) - R_h(Lz(x_i))|\,.$$

The expression on the right-hand side is a consistency error. Applying the Taylor series expansion of $z(x)$ in the node x_i, one obtains for the first step of the consistency error estimate, *exercise*,

$$|\tau_i| := |L_h z(x_i) - R_h(Lz(x_i))| \leq C \int_{x_{i-1}}^{x_{i+1}} \left(\varepsilon \left|z^{(3)}(t)\right| + \left|z^{(2)}(t)\right|\right)\, dt.$$

The right-hand side comes from the remainder in the expansion. Using the estimates (3.8) for the derivatives of $z(x)$, $1 + \varepsilon = C$, and $b(x) \geq \beta > 0$ yields

$$\begin{aligned}
|\tau_i| &\leq C \int_{x_{i-1}}^{x_{i+1}} \left(\varepsilon + \varepsilon^{-1} \exp\left(-b(1)\frac{1-t}{\varepsilon}\right) + 1 + \varepsilon^{-1} \exp\left(-b(1)\frac{1-t}{\varepsilon}\right)\right)\, dt \\
&\leq C \int_{x_{i-1}}^{x_{i+1}} (\varepsilon + 1)\; dt + C\varepsilon^{-1} \int_{x_{i-1}}^{x_{i+1}} \left(\exp\left(-\beta\frac{1-t}{\varepsilon}\right) + \exp\left(-\beta\frac{1-t}{\varepsilon}\right)\right)\, dt \\
&\leq Ch + C\varepsilon^{-1} \int_{x_{i-1}}^{x_{i+1}} \exp\left(-\beta\frac{1-t}{\varepsilon}\right)\, dt \\
&= Ch + C\varepsilon^{-1} \left(\frac{\varepsilon}{\beta} \exp\left(-\beta\frac{1-t}{\varepsilon}\right)\Big|_{x_i-h}^{x_i+h}\right) \\
&= Ch + C\left[\exp\left(-\beta\frac{1-x_i-h}{\varepsilon}\right) - \exp\left(-\beta\frac{1-x_i+h}{\varepsilon}\right)\right] \\
&= Ch + C\exp\left(-\beta\frac{1-x_i}{\varepsilon}\right)\left[\exp\left(\frac{\beta h}{\varepsilon}\right) - \exp\left(-\frac{\beta h}{\varepsilon}\right)\right] \\
&= Ch + C\sinh\left(\frac{\beta h}{\varepsilon}\right) \exp\left(-\beta\frac{1-x_i}{\varepsilon}\right).
\end{aligned}$$

It is

$$\sinh(t) = \frac{e^t - e^{-t}}{2} \leq \frac{e^t}{2} = Ce^t.$$

Hence, it follows that

$$|\tau_i| \leq Ch + C\exp\left(-\beta\frac{1-x_i}{\varepsilon} + \frac{\beta h}{\varepsilon}\right) = Ch + C\exp\left(-\beta\frac{1-x_{i+1}}{\varepsilon}\right).$$

An estimate for the error of the other part $v(x)$ can be derived with the discrete comparison principle, Lemma 3.18, and using that $v(x)$ is known

$$|v(x_i) - v_i| \leq C \exp\left(-\beta \frac{1 - x_{i+1}}{\varepsilon}\right).$$

Combining both bounds gives the final estimate. ■

Corollary 3.38. Convergence of the simple upwind scheme away from layers. *Under the assumptions of Theorem 3.35 and Theorem 3.37, the simple upwind scheme converges in an interval* $[0, 1-\delta]$*, where* $\delta > 0$ *is fixed, of first order. The constant in the convergence estimate is independent of* ε*.*

Proof. This statement follows with Theorem 3.15, using Theorems 3.35 and 3.37. ■

Remark 3.39. Behavior in the layer based on estimate (3.7). Let $\varepsilon < h$ and let ε be fixed, then one obtains in the node x_{N-2} the estimate

$$\begin{aligned} |u(x_{N-2}) - u_{N-2}| &\leq Ch + C \exp\left(-\beta^* \frac{1 - x_{N-1}}{\varepsilon}\right) = Ch + C \exp\left(-\beta^* \frac{h}{\varepsilon}\right) \\ &\leq Ch + Ch = \mathcal{O}(h), \end{aligned}$$

since $\exp(-x) < x$ if x is sufficiently large. However, for x_{N-1} one gets

$$|u(x_{N-1}) - u_{N-1}| \leq Ch + C \exp\left(-\beta^* \frac{1 - x_N}{\varepsilon}\right) = Ch + C = \mathcal{O}(1),$$

since $x_N = 1$. If follows that the bound of the error in x_{N-1} does not converge toward zero. The notion convergence has to be understood in this context as $h \geq \varepsilon, h \to \varepsilon$. □

Example 3.40. Behavior in the layer. The observation in the previous remark is not a problem of the estimate. Consider

$$-\varepsilon u''(x) - u'(x) = 0, \quad u(0) = 0,\ u(1) = 1.$$

The solution of this problem has a layer at $x = 0$. One obtains with the simple upwind scheme, *exercise*,

$$u_i = \frac{1 - r^i}{1 - r^N}, \quad \text{with} \quad r = \frac{\varepsilon}{\varepsilon + h}.$$

For $h = \varepsilon$, one gets

$$u_1 = \frac{1 - r}{1 - r^N} = \frac{1 - 1/2}{1 - (1/2)^N} = \frac{1/2}{1 - (1/2)^N} \approx \frac{1}{2}.$$

However, for the solution, it is for $x_1 = h = \varepsilon$

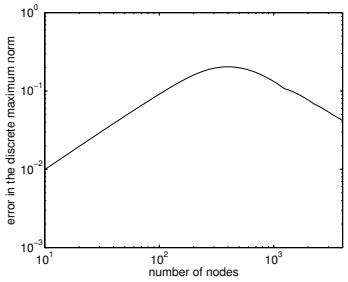


Fig. 3.6 Error of the simple upwind scheme for Example 2.8, $\varepsilon = 1e-3$, and different number of nodes.

$$u(x_1) = \frac{1 - e^{-1}}{1 - e^{-1/\varepsilon}} \approx 0.63$$

for small ε. Hence, the error is $\mathcal{O}(1)$ for small ε and there is no convergence for $h \geq \varepsilon$. Consequently, one cannot expect to improve substantially the estimate from Theorem 3.37. □

Remark 3.41. Typical behavior in the layer in numerical simulations. Consider constant ε and variable h. If h is sufficiently large, then all grid points are outside the layer. If h decreases, then the error increases since the last node moves into the layer, see Figure 3.6. If h becomes sufficiently small, then the error decreases. In this case the first estimate of Theorem 3.37 becomes effective. □

Remark 3.42. Interpretation of the simple upwind scheme as artificial diffusion. The difficulties in the numerical solution of singularly perturbed problems originate from the different magnitudes of diffusion and convection, and caused by this issue of the appearance of sharp (thin) layers. It is clear that the numerical solution becomes the simpler, the larger the diffusion, compared with the convection, becomes and the layers become wider.

Consider $b > 0$, then it is

$$\begin{aligned} b_i D^{\mathcal{N}} u_i = b_i D^- u_i = b_i \frac{u_i - u_{i-1}}{h} &= b_i \frac{u_{i+1} - u_{i-1}}{2h} + b_i \frac{-u_{i+1} + 2u_i - u_{i-1}}{2h} \\ &= b_i D^0 u_i - \frac{b_i h}{2} D^+ D^- u_i. \end{aligned}$$

Hence, the simple upwind scheme (3.6) can be written in the form

$$\begin{aligned} -\left(\varepsilon + \frac{b_i h}{2}\right) D^+ D^- u_i + b_i D^0 u_i + c_i u_i = f_i, \quad \text{for } i = 1, \ldots, N-1, \\ u_0 = u_N = 0. \end{aligned} \tag{3.10}$$

One can see that the diffusion coefficient is artificially increased and it has the magnitude $\mathcal{O}(h)$. Consequently, the simple upwind scheme is nothing else than a central difference scheme applied to a problem with sufficiently large, $\mathcal{O}(h)$, diffusion. It has been observed already in Example 3.30 that the central difference scheme gives useful results if the diffusion is of order $\mathcal{O}(h)$.

One can define methods with artificial diffusion also directly. □

Definition 3.43. Scheme with artificial diffusion, fitted upwind scheme. A finite difference scheme with artificial diffusion is defined by

$$\begin{aligned} -\varepsilon\sigma\left(q(x_i)\right) D^+D^-u_i + b_i D^0 u_i + c_i u_i &= f_i, \quad \text{for } i = 1, \ldots, N-1, \\ u_0 = u_N &= 0, \\ q(x) &:= \frac{b(x)h}{2\varepsilon}, \end{aligned} \tag{3.11}$$

and $\sigma(q)$ is an appropriate function. It is also called fitted upwind scheme. □

Remark 3.44. Fitted upwind schemes.

- The simple upwind scheme (3.6) is obtained for $\sigma(q) = 1 + q$, see (3.10).
- The introduction of artificial diffusion changes the original problem significantly. Consider, e.g.,

$$-\varepsilon u'' + u' = 1 \quad \text{in } (0,1), \quad u(0) = u(1) = 0,$$

with the solution

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}, \tag{3.12}$$

see Example 2.8. The second term is responsible for the satisfaction of the boundary condition at $x = 1$. It is basically different from zero only in the interval $[1-\varepsilon, 1]$, see Figure 3.7. Introducing artificial diffusion leads to a perturbed solution and the term that is responsible for the satisfaction of the boundary condition is in the interval $[1-\varepsilon\sigma(q(x_{N-1})), 1]$ considerably different from zero. That means, the layer is by far less steep. This effect is called smearing of the layer.
For the simple upwind scheme, it is

$$\varepsilon\sigma(q(x_{N-1})) = \varepsilon + \varepsilon\frac{b_{N-1}h}{2\varepsilon} = \varepsilon + \frac{b_{N-1}h}{2}.$$

This expression is in realistic situations, i.e., for $\varepsilon \ll b_{N-1}$ and $\varepsilon \ll h$, larger by orders of magnitude than ε.

□

Remark 3.45. Key point of stabilized methods. Appropriate discretizations for convection-dominated problems are called stabilized methods. The introduction of artificial diffusion is the key issue of stabilized methods. The difficulty

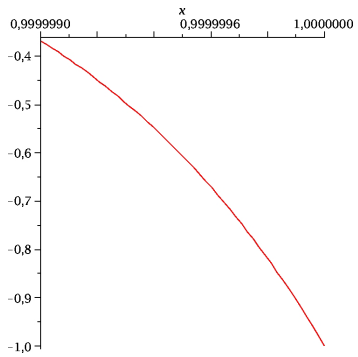


Fig. 3.7 Second term of the solution (3.12) for $\varepsilon = 10^{-6}$.

in the construction of stabilized methods is that one needs to apply the right amount of artificial diffusion, at the right locations, and in the correct directions (in multiple dimensions).

A question is whether one can construct stable methods which lead to less smearing of the layers than the simple upwind scheme. □

Theorem 3.46. Stability of schemes with artificial diffusion. *Let $b(x) > \beta > 0$, $c(x) \geq 0$, and $\sigma(q) > q$. Then, the matrix of the scheme with artificial diffusion* (3.11) *is an M-matrix and the method is stable in the discrete maximum norm. The stability constant does not depend on ε.*

Proof. The proof is very similar to the proof of Theorem 3.35, *exercise.* ■

Theorem 3.47. Consistency of schemes with artificial diffusion. *Let the assumptions of Theorem 3.46 be satisfied, let $u \in C^4([0,1])$, and let*

$$|\sigma(q) - 1| \leq \min\{q, C_M q^2\}, \quad q \geq 0,$$

with a constant $C_M > 0$. Then, for fixed ε, the consistency error of the scheme with artificial diffusion (3.11) *is of second order.*

Proof. The consistency error in the node x_i is

$$\begin{aligned} |\tau_i| &= \big|\big[-\varepsilon\sigma(q_i)\, D^+D^- u(x_i) + b_i D^0 u(x_i) + c_i u(x_i)\big] \\ &\quad -\big[-\varepsilon u''(x_i) + b_i u'(x_i) + c_i u(x_i)\big]\big| \\ &= \big|\varepsilon\sigma(q_i)\left(u''(x_i) - D^+D^- u_i\right) + \varepsilon\left(1-\sigma(q_i)\right) u''(x_i) + b_i\left(D^0 u(x_i) - u'(x_i)\right)\big|. \end{aligned}$$

From the consistency error estimate in Example 3.7, it follows that

$$|\tau_i| \leq C\left(\varepsilon\sigma(q_i)h^2 \left\|u^{(4)}\right\|_\infty + \varepsilon\left|1-\sigma(q_i)\right| \left\|u''\right\|_\infty + h^2 \left\|u^{(3)}\right\|_\infty\right).$$

Using the assumptions of the theorem and the definition of $q(x)$ gives

$$\begin{aligned} \sigma(q_i) &\leq |\sigma(q_i) - 1| + 1 \leq \min\left\{q_i, C_M q_i^2\right\} + 1 \leq q_i + 1 \leq C\frac{h}{\varepsilon} + 1, \\ |1-\sigma(q_i)| &\leq C_M q_i^2 \leq C\frac{h^2}{\varepsilon^2}. \end{aligned}$$

Inserting this estimate yields

$$\begin{aligned} |\tau_i| &\leq C\left(\left(\frac{h}{\varepsilon}+1\right)\varepsilon h^2 \left\|u^{(4)}\right\|_\infty + \varepsilon\frac{h^2}{\varepsilon^2}\left\|u''\right\|_\infty + h^2\left\|u^{(3)}\right\|_\infty\right) \\ &\leq C(\varepsilon)h^2. \end{aligned} \tag{3.13}$$

■

Remark 3.48. The consistency of the scheme with artificial diffusion.

- Examples of functions $\sigma(q)$ that satisfy the assumptions of Theorem 3.47 are (*exercise*)
$$\sigma(q) = \sqrt{1+q^2}, \quad \sigma(q) = 1 + \frac{q^2}{1+q}.$$
The second variant is called Samarskii upwind scheme.
- The consistency is of second order only for constant ε. The factor $C(\varepsilon)$ diverges to infinity for $\varepsilon \to 0$, see the middle term in (3.13). In addition, also derivatives of the solution might depend in a bad way on ε and they

Fig. 3.8 Alexander Andreewitsch Samarskii (1919 – 2008).

might explode for $\varepsilon \to 0$. It follows that these methods become worse and worse for $\varepsilon \to 0$. One can show that the consistency independent of ε away from the layer is only of first order. The typical behavior in the layer is analogously as for the simple upwind scheme, see Remark 3.41. □

Remark 3.49. Summary.

- The central difference scheme is not suited for singularly perturbed problems.
- The simple upwind scheme is stable, but too inaccurate (of first order consistent). Layers are heavily smeared.
- Upwind methods can be interpreted as methods with artificial diffusion.
- Fitted upwind methods might be of second order consistent, but only for fixed ε. This property is not uniform in ε.

The upwind schemes that have been introduced so far are not satisfactory since they are too inaccurate for small ε and the convergence in the layer depends on ε. □

3.4 Uniformly Convergent Methods

Remark 3.50. Motivation. The goal consists in the development of discretizations that converge uniformly in the whole interval $[0, 1]$, in particular in the layer. Two ways for achieving this goal will be presented:

- a scheme that is obtained with an appropriate choice of artificial diffusion $\sigma(q)$ in (3.11),
- schemes, that are defined by an appropriate choice of the grid.

In practice, one often has very small diffusion. Therefore it is important to construct numerical methods that provide accurate results in this case. The construction of such methods is not trivial. This might become clear already from the fact that the limit $\varepsilon \to 0$ is in a certain sense discontinuous, since the order of the differential equation changes. With this change of order, also other important features change, e.g., the needed boundary conditions to define a well-posed problem. Also the properties of the solutions of differential equations with different order are different, e.g., the smoothness properties. A uniformly convergent method has to deal with this limit without deterioration of the quality of the computed solution.

This section follows in part Großmann & Roos (2005). □

Definition 3.51. Uniform convergence. A scheme for the solution of (3.4), (3.5) is called uniformly convergent of order p with respect to the singularly perturbation parameter ε in the discrete maximum norm if an estimate of the form

$$\|u - u_h\|_{\infty,d} \leq Ch^p, \quad p > 0,$$

holds with a constant C that does not depend on ε. □

3.4.1 Sophisticated Artificial Diffusion

Remark 3.52. Idea. The choice of a sophisticated artificial diffusion $\sigma(q)$ can be motivated with a study of the solution of (3.4), (3.5) for $\varepsilon \to 0$. □

Lemma 3.53. Convergence to the reduced solution. *Let $u(x,\varepsilon)$ be the solution of (3.4), (3.5) with $b(x) \geq \beta > 0$, $c(x) \geq 0$, and let $u_0(x)$ be the solution of the reduced problem. Then, it holds for all $x \in [0, x_0)$ with $x_0 < 1$ that*

$$\lim_{\varepsilon \to 0} u(x,\varepsilon) = u_0(x).$$

Proof. The proof is based on the comparison principle, Corollary 2.34. Set

$$v_1(x) := \gamma \exp(\beta x), \quad \gamma > 0,$$

then it follows that

$$(Lv_1)(x) = \gamma(-\varepsilon\beta^2 + b(x)\beta + c(x)) \exp(\beta x) \geq \gamma\beta^2(1-\varepsilon)\exp(\beta x) \geq 1 \tag{3.14}$$

for sufficiently large γ. Now, let

$$v_2(x) := \exp\left(-\beta \frac{1-x}{\varepsilon}\right),$$

then, it holds

$$(Lv_2)(x) = \left(-\varepsilon\frac{\beta^2}{\varepsilon^2} + b(x)\frac{\beta}{\varepsilon} + c(x)\right)\exp\left(-\beta\frac{1-x}{\varepsilon}\right)$$
$$\geq \frac{\beta}{\varepsilon}\left(-\beta + b(x)\right)\exp\left(-\beta\frac{1-x}{\varepsilon}\right) \geq 0. \tag{3.15}$$

Consider now

$$v(x) := M_1\varepsilon v_1(x) + M_2 v_2(x),$$

then, one obtains, for appropriately chosen constants M_1 and M_2, using the estimates (3.14) for $(Lv_1)(x)$ and (3.15) for $(Lv_2)(x)$, and the definition of $u_0(x)$

$$\begin{aligned}
(Lv)(x) &= M_1\varepsilon(Lv_1)(x) + M_2(Lv_2)(x) \geq M_1\varepsilon(Lv_1)(x) \geq M_1\varepsilon \geq \varepsilon\left|u_0''(x)\right| \\
&= \left|L(u-u_0)(x)\right|, \\
v(0) &= M_1\varepsilon v_1(0) + M_2 v_2(0) = M_1\varepsilon\gamma + M_2\exp(-\beta/\varepsilon) \geq 0 \\
&= \left|(u-u_0)(0)\right|, \\
v(1) &= M_1\varepsilon v_1(1) + M_2 v_2(1) = M_1\varepsilon\gamma\exp(\beta) + M_2 \geq M_2 \\
&\geq \left|(u-u_0)(1)\right| = \left|u_0(1)\right|.
\end{aligned}$$

The constants have to be sufficiently large and they only depend on $u_0(x)$. There is no coefficient ε in the reduced problem such that $u_0(x)$ cannot depend on ε and hence the constants do not depend on ε. With the comparison principle, it follows that

$$\left|(u-u_0)(x)\right| \leq v(x) = M_1\varepsilon\gamma\exp(\beta x) + M_2\exp\left(-\beta\frac{1-x}{\varepsilon}\right).$$

Hence, one gets for $x < 1$

$$\lim_{\varepsilon\to 0}\left|(u-u_0)(x)\right| = 0.$$

■

Lemma 3.54. Estimate of the reduced solution plus a correction term. *Under the assumptions of Lemma 3.53, there is a constant C, which is independent of x and ε, such that for the solution of* (3.4), (3.5), *it holds*

$$\left|u(x,\varepsilon) - \left[u_0(x) - u_0(1)\exp\left(-b(1)\frac{1-x}{\varepsilon}\right)\right]\right| \leq C\varepsilon, \quad x \in [0,1].$$

Proof. The proof is similar to the proof of Lemma 3.53. ■

Remark 3.55. Necessary condition for a sophisticated artificial diffusion $\sigma(q)$. Let $\rho^* := h/\varepsilon$ be fixed, i.e., for $h \to 0$ it holds also $\varepsilon \to 0$. The goal consists in finding for this case a condition for an appropriate function $\sigma(q)$ in the fitted upwind scheme.

Let a node i be fixed. Since $\varepsilon \to 0$ for $h \to 0$, it follows from Lemma 3.54 that

$$\begin{aligned}\lim_{h\to 0} u(1-ih) &= \lim_{h\to 0}\left(u_0(1-ih) - u_0(1)\exp\left(-b(1)\frac{1-(1-ih)}{\varepsilon}\right)\right)\\ &= u_0(1) - u_0(1)\lim_{h\to 0}\exp\left(-b(1)\frac{ih}{\varepsilon}\right)\\ &= u_0(1) - u_0(1)\exp\left(-ib(1)\rho^*\right)\\ &= u_0(1)\left(1-\exp(-2iq(1))\right), \end{aligned} \tag{3.16}$$

where the definition of $q(x)$ has been used. The fitted upwind scheme has the form

$$-\varepsilon\sigma\left(q(b_i)\right)\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + b_i\frac{u_{i+1}-u_{i-1}}{2h} = f_i - c_i u_i,$$

or after multiplication with h^2/ε

$$\begin{aligned}-\sigma\left(q(b_i)\right)\left(u_{i+1}-2u_i+u_{i-1}\right) + q(b_i)\left(u_{i+1}-u_{i-1}\right) &= \frac{h^2}{\varepsilon}\left(f_i - c_i u_i\right)\\ &= h\rho^*\left(f_i - c_i u_i\right).\end{aligned}$$

For the right boundary, i.e., $i = N-1$, it is in particular

$$\begin{aligned}&\lim_{h\to 0}\left(-\sigma\left(q_{N-1}\right)\left(u_N - 2u_{N-1} + u_{N-2}\right) + q_{N-1}\left(u_N - u_{N-2}\right)\right)\\ &\qquad = \lim_{h\to 0} h\rho^*\left(f_{N-1} - c_{N-1}u_{N-1}\right) = 0,\end{aligned} \tag{3.17}$$

since $f(x), c(x)$ are bounded, ρ^* is a constant, and $u_{N-1} \to u(1)$. The goal of the derivation is that the finite difference solution should be as close as possible to the solution of the continuous problem. Inserting the representation (3.16) for the solution of the continuous problem in the finite element scheme (3.17) gives, where one has to take in (3.16) the indices $i \in \{0,1,2\}$ and one assumes without loss of generality $u_0(1) \neq 0$,

$$\begin{aligned}0 = &-\sigma\left(q(1)\right)\Big(-\exp(0) + 2\exp(-2q(1)) - \exp(-4q(1))\Big)\\ &+q(1)\Big(-\exp(0) + \exp(-4q(1))\Big).\end{aligned} \tag{3.18}$$

It is, using the binomial theorem,

$$\begin{aligned}\frac{-1+e^{-4x}}{-1+2e^{-2x}-e^{-4x}} = \frac{(e^{-2x}-1)(e^{-2x}+1)}{-(e^{-2x}-1)^2} &= \frac{1+e^{-2x}}{1-e^{-2x}} = \frac{e^x+e^{-x}}{e^x-e^{-x}}\\ &= \coth(x).\end{aligned}$$

Hence, one gets from (3.18) as necessary condition

$$\sigma\left(q(1)\right) = q(1)\frac{1-\exp(-4q(1))}{1-2\exp(-2q(1))+\exp(-4q(1))} = q(1)\coth\left(q(1)\right).$$

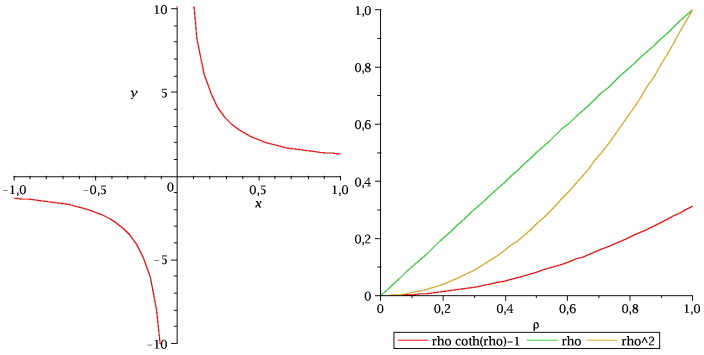


Fig. 3.9 $\coth(x)$ and comparison with the conditions of Theorem 3.47.

A choice that satisfies this limit is

$$\sigma(q) = q \coth(q).$$

This function satisfies also the conditions for the consistency of the scheme with artificial diffusion from Theorem 3.47, see Figure 3.9. □

Definition 3.56. Iljin scheme, Iljin–Allen–Southwell scheme. The finite difference scheme

$$-\frac{h}{2} b_i \coth\left(\frac{h}{2\varepsilon} b_i\right) D^+ D^- u_i + b_i D^0 u_i + c_i u_i = f_i, \quad \text{for } i = 1, \ldots, N-1,$$
$$u_0 = u_N = 0,$$

is called Iljin scheme or Iljin–Allen–Southwell (Il'in (1969); Allen & Southwell (1955)) scheme. In some applications it is called also Scharfetter–Gummel scheme (Scharfetter & Gummel (1969)). □

Theorem 3.57. Uniform convergence of the Iljin–Allen–Southwell scheme. *The Iljin–Allen–Southwell scheme converges in* $[0,1]$ *uniformly of first order in the discrete maximum norm, i.e., it holds*

$$\max_{i=1,\ldots,N-1} |u(x_i) - u_i| \leq Ch$$

with a constant C *that is independent of* ε *and* h.

Proof. The proof is rather long and involved, e.g., see Roos *et al.* (2008). ■

Example 3.58. Iljin–Allen–Southwell scheme. Consider

$$-\varepsilon u'' + u' = 1 \quad \text{in } (0,1), \quad u(0) = u(1) = 0,$$

with the solution

Table 3.1 Example 3.58, errors in the discrete maximum norm.

# intervals	central diff.	simple upwind	IAS scheme
2	124.5	0.00199	0
4	31.004	0.00398	0
8	7.715	0.00793	0
16	2.0235	0.01574	0
32	0.91132	0.03100	2.2204e-16
64	0.77305	0.06015	1.5543e-15
128	0.59276	0.11307	8.3598e-15
256	0.34287	0.18371	1.2388e-14
512	0.12997	0.19679	1.0976e-14
1024	0.03277	0.12933	5.8457e-14
2048	0.00750	0.07486	1.5675e-13
4096	0.00183	0.04076	2.8882e-13

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}.$$

For $\varepsilon = 10^{-3}$, one obtains the errors in the discrete maximum norm given in Table 3.1. It can be seen that the Iljin–Allen–Southwell scheme gives always the most accurate results. If the nodes are sufficiently away from the layer, then the computed solution is even exact in the nodes. □

3.4.2 Layer-Adapted Grids

Remark 3.59. Motivation. As already mentioned, the solution of a singularly perturbed problem consists of two parts:

- the solution of the reduced problem, which is generally smooth and easily to approximate numerically,
- the correction term, which enforces the boundary condition at the outflow boundary. This term is responsible for the appearance of the layer, i.e., for the dramatic change of the solution in a very small interval.

Consider as typical example the two-point boundary value problem from Example 3.58. In the interval $[0, 1-\varepsilon]$, the solution has practically the form $u(x) = x$. This solution can be easily approximated on a coarse grid. The interesting part of the solution is in the interval $[1-\varepsilon, 1]$. If one chooses an equidistant grid with the mesh width h, then it is generally $h \gg \varepsilon$. Hence, the interval $[1-\varepsilon, 1]$ is contained in $[x_{N-1}, x_N] = [1-h, 1]$. Consequently, one cannot expect with this choice to resolve the solution in $[1-\varepsilon, 1]$ in a good way.

The idea of layer-adapted grids is to choose in the layer region a considerably finer grid than outside the layer region. This approach offers the possibility to approximate the solution in the layer well. □

Remark 3.60. Shishkin[1] *mesh.* Consider, for simplicity of notation, a problem where the layer is situated at $x = 0$. In addition, let $b = -\beta$, $\beta \in \mathbb{R}^+$, a constant. The grid points are distributed in the form

$$x_i = \phi(i/N),$$

where one has to choose the function $\phi(\xi)$ in such a way that one obtains in a neighborhood of $x = 0$ a sufficiently fine mesh. The number N of intervals is given.

A mesh of Shishkin-type is defined by

$$\phi(\xi) = \begin{cases} \dfrac{\sigma\varepsilon}{\beta}\hat{\phi}(\xi) & \text{for } \xi \in [0, 1/2], \\ 1 - 2\left(1 - \dfrac{\sigma\varepsilon}{\beta}\ln(N)\right)(1-\xi) & \text{for } \xi \in [1/2, 1], \end{cases}$$

with $\hat{\phi}(1/2) = \ln(N)$ and the parameter $\sigma > 0$. The Shishkin mesh (1988) is obtained with

$$\hat{\phi}(\xi) = 2\ln(N)\xi.$$

With this choice, one has for the nodes $x_0, \ldots, x_{N/2}$, $i \geq 1$,

$$x_i - x_{i-1} = \phi\left(\frac{i}{N}\right) - \phi\left(\frac{i-1}{N}\right) = \frac{\sigma\varepsilon}{\beta} 2\ln(N)\left(\frac{i}{N} - \frac{i-1}{N}\right) = 2\frac{\sigma\varepsilon}{\beta}\frac{\ln(N)}{N},$$

independent of i. For the nodes $x_{N/2+1}, \ldots, x_N$, it is for $i \geq N/2 + 1$,

$$\begin{aligned} x_i - x_{i-1} &= \phi\left(\frac{i}{N}\right) - \phi\left(\frac{i-1}{N}\right) \\ &= 1 - 2\left(1 - \frac{\sigma\varepsilon}{\beta}\ln(N)\right)\left(1 - \frac{i}{N}\right) - 1 + 2\left(1 - \frac{\sigma\varepsilon}{\beta}\ln(N)\right)\left(1 - \frac{i-1}{N}\right) \\ &= \frac{2}{N} - 2\frac{\sigma\varepsilon}{\beta}\frac{\ln(N)}{N}, \end{aligned}$$

independent of i. Hence, a piecewise equidistant mesh is defined. The transition point from the very fine to the coarse grid is located at

$$\tau = x_{N/2} = \frac{\sigma\varepsilon}{\beta}\ln(N).$$

The use of the Shishkin mesh is supported by results from the numerical analysis. □

Theorem 3.61. Convergence of the simple upwind scheme on a Shishkin mesh. *Consider the simple upwind scheme on a Shishkin mesh with the transition point*

[1] Grigory I. Shishkin

$$\tau = \min\left\{\frac{1}{2}, \frac{\varepsilon}{\beta}\ln(N)\right\},$$

i.e., with $\sigma = 1$. Then, it holds the following error estimate

$$\max_{i=1,\ldots,N-1} |u(x_i) - u_i| \leq CN^{-1}\ln(N),$$

where the constant C is independent of ε and N.

Proof. The proof is based on the decomposition of the solution in the part coming from the reduced problem (smooth part) and the correction part. It is rather long and involved, see Roos *et al.* (2008). ■

Remark 3.62. Layer-adapted meshes.

- The convergence is slightly suboptimal because of the factor $\ln(N)$. However, one can see in numerical examples that the given error estimate is sharp, i.e., this factor cannot be omitted.
- The idea to use layer-adapted mesh goes back to Bahvalov[2] (Bahvalov (1969)). In Bahvalov meshes, there is a smooth transition from the fine to the coarse mesh. However, the numerical analysis for schemes on Bahvalov meshes is in general more complicated than on Shishkin meshes.
- The a priori (before the numerical simulation) construction of appropriate layer-adapted meshes needs more or less already the knowledge of the solution. This aspect is in applications not given, in particular for problems in two or three dimensions. Then, one needs an a posteriori (during the numerical simulation) construction of adapted grids. There are ways to perform this approach.
- An essential finding of the analysis of numerical methods on a priori layer-adapted grids is that one can use on an appropriate grid a simple numerical method and one obtains reasonable error estimates.
- Using a Shishkin mesh, one has to define the finite differences in the node $x_{N/2}$, where the distances to the neighbor nodes are of different lengths. Let x_i be a node and let the intervals $[x_{i-1}, x_i]$ and $[x_i, x_{i+1}]$ have the length h_i and h_{i+1}, respectively. There are no changes for the backward and forward finite difference compared with Definition 3.4, since for them one needs only one of the neighbor intervals. Define

  $$\tilde{h}_i := \frac{h_i + h_{i+1}}{2},$$

  then the central difference is the weighted average

  $$D^0 v(x_i) = \frac{1}{2\tilde{h}_i}\left(h_i D^+ v(x_i) + h_{i+1} D^- v(x_i)\right).$$

  The second derivative is approximated by

[2] Nikolai Sergejewitsch Bahvalov (1934 – 2005)

Table 3.2 Errors in the discrete maximum norm for simulations on a Shishkin mesh.

# intervals	$\|u - u_h\|_{\infty,d}$
4	0.25584
8	0.16455
16	0.10833
32	0.069125
64	0.043656
128	0.026335
256	0.015402
512	0.0087902
1024	0.0049257
2048	0.0027225
4096	0.0014891

$$v''(x_i) \approx \frac{1}{\tilde{h}_i}\left(D^+v(x_i) - D^-v(x_i)\right) = \frac{1}{\tilde{h}_i}\left(\frac{v_{i+1}-v_i}{h_{i+1}} - \frac{v_i - v_{i-1}}{h_i}\right).$$

- The matrices, which are obtained when using layer-adapted meshes, have generally a very bad condition number.
- Layer adapted meshes can be applied in the finite element context in the same fashion as for finite difference methods.
- A comprehensive monograph about numerical methods on layer-adapted meshes is Linß (2010).

□

Example 3.63. Shishkin mesh. Consider again

$$-\varepsilon u'' + u' = 1 \quad \text{in } (0,1), \quad u(0) = u(1) = 0,$$

with the solution

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)},$$

like in Example 3.58. In this example, the layer is at $x = 1$. Thus, the transition point is chosen at

$$\tau = x_{N/2} = 1 - \frac{\sigma\varepsilon}{\beta}\ln(N) = 1 - \sigma\varepsilon\ln(N).$$

The errors in the discrete maximum norm for $\varepsilon = 10^{-6}$ and $\sigma = 2$ are given in Table 3.2.

The results depend strongly on the choice of σ, *exercise.* □

Remark 3.64. Summary. There are two ways to construct uniformly convergent methods:

- by using an appropriately modified scheme on a simple grid,

- by using a simple scheme on an appropriately chosen grid.

□