Some analytical results for an algebraic flux correction scheme for a steady convection–diffusion equation in 1D

Gabriel R. Barrenechea[a], Volker John[b,c], Petr Knobloch[d,*]

[a] *Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland*

[b] *Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39, 10117 Berlin, Germany*

[c] *Free University of Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany*

[d] *Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 18675 Praha 8, Czech Republic*

Abstract

Algebraic flux correction schemes are nonlinear discretizations of convection dominated problems. In this work, a scheme from this class is studied for a steady-state convection–diffusion equation in one dimension. It is proved that this scheme satisfies the discrete maximum principle. Also, as it is a nonlinear scheme, the solvability of the linear subproblems arising in a Picard iteration is studied, where positive and negative results are proved. Furthermore, the non-existence of solutions for the nonlinear scheme is proved by means of counterexamples. Therefore, a modification of the method, which ensures the existence of a solution, is proposed. A weak version of the discrete maximum principle is proved for this modified method.

Keywords: finite element method, convection–diffusion equation, algebraic flux correction, discrete maximum principle, fixed point iteration, solvability of linear subproblems, solvability of nonlinear problem

1. Introduction

Scalar convection–diffusion equations model the convective and diffusive transport of a scalar quantity, like temperature or concentration. Solutions of convection-dominated convection–diffusion equations typically possess layers, which cannot be resolved unless the given mesh is sufficiently fine in layer regions. Standard discretizations, like central finite differences or the Galerkin finite element method, cannot cope with this situation and the computed solutions are globally polluted with spurious oscillations. It is well known that so-called stabilized discretizations have to be applied. There are many proposals of such discretizations, see the monograph [22] for an extensive review.

*Corresponding author.

Email addresses: gabriel.barrenechea@strath.ac.uk (Gabriel R. Barrenechea), john@wias-berlin.de (Volker John), knobloch@karlin.mff.cuni.cz (Petr Knobloch)

In the past few years, comprehensive numerical studies revealed, however, that none of the proposed stabilized discretizations satisfies the following three requirements: accuracy, efficiency, and numerical solution without spurious oscillations (discrete maximum principle). This statement holds true for the steady-state equation [2, 3, 8, 9, 13] as well as for the time-dependent equation [6, 11, 12]. Indeed, most of the methods fail to satisfy a discrete maximum principle. However, this property is particularly important in applications, where numerical results, e.g., with negative concentrations, will be considered to be worthless. Even if such quantities are not of primary interest, spurious oscillations have been shown to lead to blow-ups in the simulation of coupled problems [10]. Altogether, the validity a discrete maximum principle is, in our opinion, of utmost importance for simulations of applications.

There are few discretizations that satisfy a discrete maximum principle, like the upwind finite difference scheme [22], a finite volume scheme on Delaunay meshes [7], and algebraic flux correction schemes. The first two methods are generally rather inaccurate, while the algebraic flux correction schemes are usually nonlinear discretizations and their application might be time consuming. However, applications often lead to nonlinear models, and then a nonlinear discretization of a linear equation in such a model seems not to be a severe disadvantage. Altogether, from the point of view of applications, algebraic flux correction schemes are very attractive.

The basic philosophy of flux correction schemes was formulated already in the 1970s in [4, 23]. Later, the idea was applied in the finite element context, e.g., in [21] and [1]. In the last decade, the methods have been further developed and refined, in particular in [20, 14–19]. Until not long ago, two limiting techniques within algebraic flux correction schemes were pursued, so-called flux-corrected transport (FCT) schemes for the time-dependent equation and total variation diminishing (TVD) schemes for the steady-state equation. Finally, a scheme was presented in [18] that can handle both situations. For the time-dependent problem, a linear variant of a FCT scheme was proposed in [17].

Despite the attractiveness of algebraic flux correction schemes, there seems to be no rigorous numerical analysis for this class of methods. The main reason lies probably in their construction, which does not allow to apply the usual tools of the analysis of finite element discretizations. Unlike almost all other stabilized methods, which modify the bilinear form of the discrete problem in some way, algebraic flux correction schemes work on the algebraic level. They manipulate the matrix and the right-hand side of the algebraic system of equations. A few basic properties of these schemes can be deduced immediately from their construction, like mass conservation or the discrete maximum principle for transport equations [19].

In this work we study some properties of a nonlinear discrete problem that generalizes the algebraic flux correction method of TVD-type from [15] applied to the 1D steady-state convection–diffusion equation. We present both theoretical and computational results; the latter ones are obtained by solving the nonlinear discrete problem using a fixed point iteration. While the linear subproblems in the fixed point iteration are proved to be well-posed, the nonlinear problem is shown to be not solvable in general. However, we prove the solvability for a modified nonlinear discrete problem. To the authors' best knowledge, the results concerning the solvability of the linear subproblems and the nonlinear problem are the first results of this kind for algebraic flux correction schemes. In addition, the present work represents a basis for analyzing algebraic flux correction schemes applied to multi-dimensional

problems.

The paper is organized in the following way. First, the algebraic flux correction method will be introduced in Section 2. In Section 3, the 1D model problem will be formulated and its finite element discretization will be presented. The application of the algebraic flux correction method to this problem is the topic of Section 4. It will be shown there that the discrete operator of this scheme can be written as a nonlinear finite difference operator with an artificial diffusion vector whose components are bounded by a data-dependent constant $\tilde{\varepsilon}$. In Section 5, the discrete maximum principle for this operator will be proved for appropriately chosen values of $\tilde{\varepsilon}$. Different choices of $\tilde{\varepsilon}$, for which the discrete maximum principle is satisfied, will be studied numerically in Section 6. The unique solvability of the linear subproblems arising in the fixed point iteration is studied in Section 7 under more general conditions on the artificial diffusion vector than from the actual method [15]. Some positive but also a negative result are proved. Section 8 starts with a number of counterexamples concerning the solvability of the nonlinear discrete problem. Then, the existence of a solution of the nonlinear problem is proved for a modification of the method. A concrete realization of this modification is proposed in Section 9, where a weak form of the discrete maximum principle is proved and numerical results are presented. Finally, a summary and an outlook are given in Section 10.

2. An algebraic flux correction scheme

Consider a linear boundary value problem whose solution is (mainly) determined by convection and for which the maximum principle holds. Let us discretize this problem by the finite element method. Then, the discrete solution can be represented by a vector $\mathrm{U} \in \mathbb{R}^N$ of its coefficients with respect to a basis of the respective finite element space. Let us assume that the last $N - M$ components of U ($0 < M < N$) correspond to nodes where Dirichlet boundary conditions are prescribed whereas the first M components of U are computed using the finite element discretization of the underlying partial differential equation. Then $\mathrm{U} \equiv (u_1, \ldots, u_N)$ satisfies a system of linear equations of the form

$$\sum_{j=1}^{N} a_{ij}\, u_j = g_i\,, \qquad i = 1, \ldots, M\,, \tag{1}$$

$$u_i = u_i^b\,, \qquad i = M+1, \ldots, N\,. \tag{2}$$

We assume that

$$a_{ii} > 0\,, \qquad \sum_{j=1}^{N} a_{ij} = 0\,, \qquad i = 1, \ldots, M\,, \tag{3}$$

which is often the case when incompressible convection fields are considered.

Since the original problem satisfies the maximum principle, it is natural to require that this property is inherited by the discrete problem. Unfortunately, the discrete maximum principle does not hold for many finite element discretizations of convection dominated problems, in particular, for the Galerkin discretization and most stabilized methods, see, e.g., [22]. The aim of algebraic flux correction approaches is to cure this deficiency by manipulating the

algebraic system in such a way that the solution satisfies the discrete maximum principle and layers are not excessively smeared.

The starting point of the algebraic flux correction algorithm is the finite element matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$ corresponding to the above finite element discretization in the case where homogeneous natural boundary conditions are used instead of the Dirichlet ones. We introduce the symmetric artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ possessing the entries

$$d_{ij} = -\max\{a_{ij}, 0, a_{ji}\} \qquad \forall\, i \neq j\,, \qquad\qquad d_{ii} = -\sum_{j \neq i} d_{ij}\,.$$

Then, the matrix $\tilde{\mathbb{A}} := \mathbb{A} + \mathbb{D}$ has nonpositive off-diagonal entries and each of its row sums vanishes. A vector $\mathrm{U} \in \mathbb{R}^N$ being a solution of a linear system with the matrix $\tilde{\mathbb{A}}$ satisfies the discrete maximum principle in the sense that for any $i \in \{1, \ldots, M\}$ the following holds

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i \leq 0 \quad \Rightarrow \quad u_i \leq \max_{j \neq i,\, \tilde{a}_{ij} \neq 0} u_j\,.$$

This property immediately follows from the fact that, using (3), one gets

$$\tilde{a}_{ii}\,u_i \leq -\sum_{j \neq i} \tilde{a}_{ij}\,u_j = \tilde{a}_{ii}\,c - \sum_{j \neq i} \tilde{a}_{ij}\,(u_j - c) \leq \tilde{a}_{ii}\,c \qquad \forall\, c \geq \max_{j \neq i,\, \tilde{a}_{ij} \neq 0} u_j\,.$$

Going back to the solution of the system (1), this system is equivalent to

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + (\mathbb{D}\,\mathrm{U})_i\,, \qquad i = 1, \ldots, M\,. \tag{4}$$

Since the row sums of the matrix $\mathbb{D}$ vanish, it follows that

$$(\mathbb{D}\,\mathrm{U})_i = \sum_{j \neq i} f_{ij}\,, \qquad i = 1, \ldots, N\,,$$

where $f_{ij} = d_{ij}\,(u_j - u_i)$. Clearly, $f_{ij} = -f_{ji}$ for all $i, j = 1, \ldots, N$. Now the idea of the algebraic flux correction schemes is to limit those anti-diffusive fluxes f_{ij} that would otherwise cause spurious oscillations. To this end, system (1) (or, equivalently (4)) is replaced by

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + \sum_{j \neq i} \alpha_{ij}\,f_{ij}\,, \qquad i = 1, \ldots, M\,, \tag{5}$$

with solution-dependent correction factors $\alpha_{ij} \in [0, 1]$. For $\alpha_{ij} = 1$, the original system (1) is recovered. Hence, intuitively, the coefficients α_{ij} should be as close to 1 as possible to limit the modifications of the original problem.

The coefficients α_{ij} can be chosen in various ways but their definition is always based on the above fluxes f_{ij}, see [14–18] for examples. In this work we consider coefficients α_{ij} proposed in [15]. This definition relies on the values P_i^+, P_i^-, Q_i^+, Q_i^- computed for $i = 1, \ldots, N$ in the following way. First, one initializes all these quantities by zero. Then one goes through all pairs of indices $i, j \in \{1, \ldots, N\}$ and if $a_{ji} \leq a_{ij}$, one performs the updates

$$P_i^+ := P_i^+ + \max\{0, f_{ij}\}\,, \qquad\qquad P_i^- := P_i^- - \max\{0, f_{ji}\}\,, \tag{6}$$

$$Q_i^+ := Q_i^+ + \max\{0, f_{ji}\}\,, \qquad\qquad Q_i^- := Q_i^- - \max\{0, f_{ij}\}\,, \tag{7}$$

$$Q_j^+ := Q_j^+ + \max\{0, f_{ij}\}\,, \qquad\qquad Q_j^- := Q_j^- - \max\{0, f_{ji}\}\,. \tag{8}$$

After having computed the values P_i^+, P_i^-, Q_i^+, Q_i^-, $i = 1, \ldots, N$, one sets

$$R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \qquad i = 1, \ldots, N\,.$$

Finally, the coefficients α_{ij} are defined by

$$\alpha_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0\,, \\ R_i^- & \text{if } f_{ij} < 0\,, \end{cases} \qquad\qquad i, j = 1, \ldots, N\,.$$

3. Finite element discretization of a 1D convection–diffusion equation

To better understand the algebraic flux correction method described in the previous section, we shall apply it to a finite element discretization of a scalar one-dimensional convection–diffusion equation. In this section we formulate the 1D problem, introduce its discretization, and for completeness, we review its main characteristics.

We consider the boundary value problem

$$-\varepsilon\, u'' + b\, u' = g \quad \text{in } (0,1)\,, \qquad u(0) = u_L\,, \quad u(1) = u_R\,, \tag{9}$$

where, for simplicity, ε and b are assumed to be positive constants. Moreover, g is supposed to belong to $L^2(0,1)$ and u_L, u_R are any real numbers. If g is constant, then the solution of (9) is given by the formula

$$u(x) = u_L + \frac{g}{b}\, x + \gamma\, \frac{e^{-(1-x)\, b/\varepsilon} - e^{-b/\varepsilon}}{1 - e^{-b/\varepsilon}} \tag{10}$$

with $\gamma := u_R - u_L - g/b$. Thus, for $\gamma \neq 0$ and $\varepsilon \ll b$, the solution of (9) possesses a boundary layer at the right-hand boundary point.

Let us divide the interval $[0,1]$ into $n+1$ subintervals $[x_i, x_{i+1}]$, $i = 0, \ldots, n$, with $x_i = i\, h$ and $h = 1/(n+1)$. We define the finite element space

$$W_h = \{v_h \in C([0,1])\,;\ v_h|_{[x_i, x_{i+1}]} \in P_1([x_i, x_{i+1}]),\ i = 0, \ldots, n\}$$

consisting of continuous piecewise linear functions and set

$$V_h = \{v_h \in W_h\,;\ v_h(0) = v_h(1) = 0\}\,.$$

Then the Galerkin finite element discretization of (9) reads: Find $u_h \in W_h$ such that $u_h(0) = u_L$, $u_h(1) = u_R$ and

$$\varepsilon\, (u_h', v_h') + (b\, u_h', v_h) = (g, v_h) \qquad \forall\ v_h \in V_h\,, \tag{11}$$

where $(\cdot, \cdot)$ denotes the inner product in $L^2(0,1)$.

Let us denote by $\varphi_1, \ldots, \varphi_n \in V_h$ the usual basis functions of V_h, i.e., $\varphi_i(x_j) = \delta_{ij}$ for $i, j = 1, \ldots, n$. We define

$$g_i = \frac{1}{h}\, (g, \varphi_i)\,, \qquad i = 1, \ldots, n\,.$$

Setting $u_i = u_h(x_i)$, $i = 0, \ldots, n+1$, then (11) is equivalent to the system

$$-\varepsilon \, \frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b \, \frac{u_{i+1} - u_{i-1}}{2\,h} = g_i \,, \quad i = 1, \ldots, n \,. \tag{12}$$

This system can be also obtained by discretizing (9) using the central finite difference method. Then, however, $g_i = g(x_i)$.

Let us introduce the Péclet number

$$Pe = \frac{b\,h}{2\,\varepsilon}$$

and let g be constant. If $Pe = 1$, then (12) reduces to

$$b \, \frac{u_i - u_{i-1}}{h} = g \,, \quad i = 1, \ldots, n \,,$$

and hence $u_i = u_L + (g/b)\, x_i$, $i = 0, \ldots, n$. Thus, in this case,

$$u_h(x) = u_L + \frac{g}{b}\, x \,, \quad x \in [0, 1-h] \,.$$

If $Pe \neq 1$, then

$$u_i = \frac{g}{b}\, x_i + A + B \left(\frac{1+Pe}{1-Pe} \right)^i \,, \quad i = 0, \ldots, n+1 \,, \tag{13}$$

where A and B are determined by the conditions $u_0 = u_L$ and $u_{n+1} = u_R$. We observe that, for $Pe < 1$, the discrete solution is the sum of two monotone grid functions but, for $Pe > 1$, the discrete solution u_i generally possesses spurious oscillations. This shows that the Galerkin discretization is not appropriate for solving (9) numerically if $Pe > 1$.

4. The algebraic flux correction scheme applied to the 1D problem

To suppress the spurious oscillations in the solutions of the Galerkin finite element discretization of (9) given by (11), we shall apply the algebraic flux correction scheme described in Section 2. We shall assume that $Pe > 1$, which is the case interesting in practice.

The Galerkin discretization of (9) introduced in the previous section corresponds to the system from Section 2 with $N = n + 2$ but with a different numbering of the nodes. The matrices $\mathbb{A}$ and $\mathbb{D}$ are three-diagonal $(n+2) \times (n+2)$ matrices with entries (cf. (12))

$$\begin{aligned}
&a_{0,0} = \frac{\varepsilon}{h^2} - \frac{b}{2\,h} \,, \quad a_{0,1} = -\frac{\varepsilon}{h^2} + \frac{b}{2\,h} \,, \\
&a_{i,i-1} = -\frac{\varepsilon}{h^2} - \frac{b}{2\,h} \,, \quad a_{i,i} = \frac{2\,\varepsilon}{h^2} \,, \quad a_{i,i+1} = -\frac{\varepsilon}{h^2} + \frac{b}{2\,h} \,, \qquad i = 1, \ldots, n \,, \\
&a_{n+1,n} = -\frac{\varepsilon}{h^2} - \frac{b}{2\,h} \,, \quad a_{n+1,n+1} = \frac{\varepsilon}{h^2} + \frac{b}{2\,h} \,, \\
&d_{i,i+1} = \frac{\varepsilon}{h^2} - \frac{b}{2\,h} \,, \qquad i = 0, \ldots, n \,.
\end{aligned} \tag{14}$$

The vector U in (5) is given by $\mathrm{U} = (u_0, u_1, \ldots, u_{n+1})^T$. Note that the assumption (3) is satisfied.

Now let us compute the values α_{ij} in (5). The values α_{ij} are needed only for $i = 1, \dots, n$ and $|i - j| = 1$, and they are not important if $f_{ij} = 0$. Since $f_{ij} \neq 0$ only if $|i - j| = 1$, and $a_{i+1,i} < a_{i,i+1}$ for $i = 0, \dots, n$, the updates (6)–(8) have to be computed only for $j = i + 1$, $i = 0, \dots, n$. This readily gives

$$\begin{aligned} P_i^+ &= \max\{0, f_{i,i+1}\}\,, & P_i^- &= -\max\{0, f_{i+1,i}\}\,, \\ Q_i^+ &= \max\{0, f_{i-1,i}\} + \max\{0, f_{i+1,i}\}\,, & Q_i^- &= -\max\{0, f_{i,i-1}\} - \max\{0, f_{i,i+1}\} \end{aligned}$$

for $i = 1, \dots, n$. Thus, for $i = 1, \dots, n$, one obtains

$$\alpha_{i,i-1} = \begin{cases} \min\left\{1, \dfrac{\max\{0, f_{i+1,i}\}}{\max\{0, f_{i,i+1}\}}\right\} & \text{if } f_{i,i-1} > 0\,, \\ \min\left\{1, \dfrac{\max\{0, f_{i,i+1}\}}{\max\{0, f_{i+1,i}\}}\right\} & \text{if } f_{i,i-1} < 0\,, \end{cases}$$

$$\alpha_{i,i+1} = \begin{cases} \min\left\{1, \dfrac{\max\{0, f_{i-1,i}\}}{f_{i,i+1}}\right\} & \text{if } f_{i,i+1} > 0\,, \\ \min\left\{1, \dfrac{\max\{0, f_{i,i-1}\}}{f_{i+1,i}}\right\} & \text{if } f_{i,i+1} < 0\,. \end{cases}$$

It is not completely clear, how to interpret the definition of $\alpha_{i,i-1}$ when the denominator vanishes. In this case we always set $\alpha_{i,i-1} = 1$. This leads to

$$\begin{aligned} &\alpha_{i,i-1} = \alpha_{i,i+1} = 0 \qquad \text{if} \quad f_{i,i-1}\, f_{i,i+1} > 0\,, \\ &\alpha_{i,i-1} = 1\,, \quad \alpha_{i,i+1} = \min\left\{1, \frac{f_{i-1,i}}{f_{i,i+1}}\right\} \qquad \text{if} \quad f_{i,i-1}\, f_{i,i+1} \leq 0\,. \end{aligned}$$

Setting

$$\beta_i = \begin{cases} 1 & \text{if } f_{i,i+1} \neq 0 \quad \text{and} \quad \dfrac{f_{i-1,i}}{f_{i,i+1}} < 1\,, \\ 0 & \text{otherwise}\,, \end{cases} \qquad i = 1, \dots, n\,,$$

system (5) is equivalent to

$$\begin{aligned} &u_0 = u_L\,, \\ &(\mathbb{A}\,\mathrm{U})_i + \beta_i\,(f_{i,i-1} + f_{i,i+1}) = g_i\,, \qquad i = 1, \dots, n\,, \\ &u_{n+1} = u_R\,. \end{aligned}$$

The definition of the coefficients β_i can be written also in the form

$$\beta_i = \begin{cases} 1 & \text{if } u_i \neq u_{i+1} \quad \text{and} \quad \dfrac{u_i - u_{i-1}}{u_{i+1} - u_i} < 1\,, \\ 0 & \text{otherwise}\,, \end{cases} \qquad i = 1, \dots, n\,. \tag{15}$$

Finally, applying that

$$f_{i,i-1} + f_{i,i+1} = \left(\frac{\varepsilon}{h^2} - \frac{b}{2\,h}\right)(u_{i-1} - 2\,u_i + u_{i+1})\,, \qquad i = 1, \dots, n\,,$$

and setting

$$\tilde{\varepsilon} = \frac{b\,h}{2} - \varepsilon = \varepsilon\,(Pe - 1)\,, \tag{16}$$

one arrives at the following final version of the algebraic flux correction scheme:

Find $u_0, \ldots, u_{n+1}$ such that:

$$u_0 = u_L\,, \qquad\qquad u_{n+1} = u_R\,, \tag{17}$$

and

$$-(\varepsilon + \beta_i\,\tilde{\varepsilon})\,\frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b\,\frac{u_{i+1} - u_{i-1}}{2\,h} = g_i\,, \qquad i = 1, \ldots, n\,. \tag{18}$$

Since also other definitions of β_i than (15) may be convenient (see the end of this section), we shall analyze the flux correction scheme (17), (18) for a class of functions β_i satisfying

$$\beta_i \in \{0, 1\}\,, \qquad \beta_i = 1 \quad \text{if} \quad (u_i - u_{i-1})(u_{i+1} - u_i) < 0\,, \qquad\qquad i = 1, \ldots, n\,. \tag{19}$$

Note that functions β_i defined by (15) satisfy (19).

Remark 1. *Some comments on this method are in order:*

1. *Condition* (19) *assures that artificial diffusion is added to the equation at the node x_i whenever the discrete solution has a local extremum at x_i.*
2. *If $\beta_i = 1$, then the corresponding equation in* (18) *reduces to*

$$b\,\frac{u_i - u_{i-1}}{h} = g_i\,. \tag{20}$$

   Thus, in this case the method transforms (locally) the original Galerkin method into an upwinded discretization of the hyperbolic equation $b\,u' = g$.
3. *There are alternative ways to define the matrix $\mathbb{D}$. For example, if it is defined with respect to the convection matrix only, i.e., setting $\varepsilon = 0$ in* (14)*, one obtains* (18) *with*

$$\tilde{\varepsilon} = \frac{b\,h}{2}\,. \tag{21}$$

   If $\beta_i = 1$, then the scheme (18) *becomes*

$$-\varepsilon\,\frac{u_{i-1} - 2\,u_i + u_{i+1}}{h^2} + b\,\frac{u_i - u_{i-1}}{h} = g_i\,, \tag{22}$$

   which is the usual upwind discretization of (9) *at the node x_i. This approach was used, e.g., in [18]. The definition of $\mathbb{D}$ using the whole matrix $\mathbb{A}$, as it was considered in this section, makes the implementation of the method simpler (and more economical) and was used, e.g., in [12, 2]. Furthermore, another possible alternative to define the matrix $\mathbb{D}$ is to use the sum of the convection matrix and the diffusion matrix multiplied by a constant from the interval $(0, 1)$. This approach leads to* (18) *with $\tilde{\varepsilon} \in (b\,h/2 - \varepsilon, b\,h/2)$, i.e., a method that can be viewed as intermediate with respect to the two upwinding strategies expressed by* (20) *and* (22)*.*

Let us now present two choices of β_i different from (15). For simplicity, we shall assume that $u_i = u(x_i)$, $i = 0, \ldots, n+1$. If u is increasing and strictly convex in $[0,1]$ or decreasing and strictly concave in $[0,1]$, then the definition (15) gives $\beta_i = 1$, $i = 1, \ldots, n$. Thus, the artificial diffusion may be added in regions where it is not needed at all, i.e., where no layer occurs. A partial remedy is to set $\beta_i = 1$ only at nodes where the increase or decrease of u sufficiently accelerates. For example, one can set

$$\beta_i = \begin{cases} 1 & \text{if } u_i \neq u_{i+1} \quad \text{and} \quad \dfrac{u_i - u_{i-1}}{u_{i+1} - u_i} < L\,, \\ 0 & \text{otherwise}\,, \end{cases} \qquad i = 1, \ldots, n\,, \tag{23}$$

with a constant $L \in (0,1)$, e.g., $L = 0.5$.

Unfortunately, the relation (23) does not prevent the method from adding artificial diffusion in regions where the solution is nearly constant with respect to its global behavior. For example, for $u(x) = 1 + x^5$ and any $n > 5$, the definition (23) with $L = 0.5$ leads to $\beta_1 = \cdots = \beta_5 = 1$ and $\beta_i = 0$ for $i > 5$, i.e., artificial diffusion is added on the interval $[0, x_5]$. However, $u(x) \in [1, 1.001]$ and $u'(x) \in [0, 0.02]$ for $x \in [0, 0.25]$, whereas $u(x) \in [1,2]$ and $u'(x) \in [0,5]$ for $x \in [0,1]$ so that u can be regarded as nearly constant in $[0, 0.25]$. Hence artificial diffusion is not needed at nodes near to 0. This suggests to replace (23) by

$$\beta_i = \begin{cases} 1 & \text{if } (u_i - u_{i-1})(u_{i+1} - u_i) < 0\,, \\ & \text{or } \dfrac{|u_{i+1} - u_i|}{h} > D \quad \text{and} \quad \dfrac{u_i - u_{i-1}}{u_{i+1} - u_i} < L\,, \\ 0 & \text{otherwise}\,, \end{cases} \qquad i = 1, \ldots, n\,, \tag{24}$$

with some suitable threshold D, e.g.,

$$D = \kappa\, \frac{\Delta u}{\Delta x}\,, \qquad \kappa = 0.5\,, \tag{25}$$

where Δx is a characteristic length scale and Δu a corresponding characteristic variation of u. For the above example of u, one gets $D = 0.5$ and $\beta_i = 0$, $i = 1, \ldots, n$, if $h \leq 0.1$. Note that if (23) leads to $\beta_i = 0$ so does (24) and if the values of β_i provided by (23) and (24) differ, then $|u_i - u_{i-1}|/h < D\,L$.

As another example, let us consider the function $u(x) = e^{-(1-x)\,b/\varepsilon}$, $x \in [0,1]$ (cf. (10)), which possesses a boundary layer at the point 1 for large values of b/ε. For any $i \in \{1, \ldots, n\}$, one obtains

$$\frac{u_i - u_{i-1}}{u_{i+1} - u_i} = e^{-2\,Pe}\,, \qquad \frac{u_{i+1} - u_i}{h} = u(x_i)\, \frac{e^{2\,Pe} - 1}{h}$$

so that (15) gives $\beta_1 = \cdots = \beta_n = 1$. The definition (23) gives either the same result or $\beta_1 = \cdots = \beta_n = 0$ if $L \leq e^{-2\,Pe}$. However, using (24) with $L = 0.5$ and $D = 0.5$, one always has $\beta_n = 1$ and possibly $\beta_i = 1$ at some further nodes near to 1, depending on ε, b and h. At the remaining nodes, $\beta_i = 0$. In particular, for $n \geq 4$, one obtains $\beta_i = 0$ for $i \leq (n+1)/2$. Thus, artificial diffusion is added only near the layer region, as desired.

5. Discrete maximum principle

From the last point in Remark 1, one can see that it makes sense to consider (18) with any

$$\tilde{\varepsilon} \in \left[\frac{b\,h}{2} - \varepsilon, \frac{b\,h}{2} \right]. \tag{26}$$

In this section we prove that then the method satisfies the discrete maximum principle and we formulate various consequences of this fact.

Theorem 1. *Consider any $\tilde{\varepsilon} \geq b\,h/2 - \varepsilon$. Then any solution of the nonlinear problem (17)–(19) satisfies the discrete maximum principle, i.e., for any $i \in \{1, \ldots, n\}$, one has*

$$g_i \leq 0 \qquad \Rightarrow \qquad u_i \leq \max\{u_{i-1}, u_{i+1}\}\,, \tag{27}$$

$$g_i \geq 0 \qquad \Rightarrow \qquad u_i \geq \min\{u_{i-1}, u_{i+1}\}\,. \tag{28}$$

Moreover, for any $k, l \in \{0, 1, \ldots, n+1\}$ with $k+1 < l$, one has

$$g_i \leq 0\,, \quad i = k+1, \ldots, l-1 \qquad \Rightarrow \qquad u_i \leq \max\{u_k, u_l\}\,, \quad i = k, \ldots, l\,, \tag{29}$$

$$g_i \geq 0\,, \quad i = k+1, \ldots, l-1 \qquad \Rightarrow \qquad u_i \geq \min\{u_k, u_l\}\,, \quad i = k, \ldots, l\,. \tag{30}$$

Proof. Let the values $u_0, u_1, \ldots, u_{n+1}$ satisfy (17)–(19). Consider any $i \in \{1, \ldots, n\}$ and let $g_i \leq 0$. If $u_i > \max\{u_{i-1}, u_{i+1}\}$, then $\beta_i = 1$ and hence

$$\begin{aligned} 0 \geq g_i\,h^2 &= -\left(\varepsilon + \tilde{\varepsilon} + \frac{b\,h}{2}\right) u_{i-1} + 2\,(\varepsilon + \tilde{\varepsilon})\,u_i - \left(\varepsilon + \tilde{\varepsilon} - \frac{b\,h}{2}\right) u_{i+1} \\ &> -\left(\varepsilon + \tilde{\varepsilon} + \frac{b\,h}{2}\right) u_i \quad + 2\,(\varepsilon + \tilde{\varepsilon})\,u_i - \left(\varepsilon + \tilde{\varepsilon} - \frac{b\,h}{2}\right) u_i = 0\,, \end{aligned}$$

which is a contradiction. Therefore, $u_i \leq \max\{u_{i-1}, u_{i+1}\}$.

Now consider any $k, l \in \{0, 1, \ldots, n+1\}$ with $k+1 < l$ and let $g_i \leq 0$ for $i = k+1, \ldots, l-1$. Let $j \in \{k, \ldots, l\}$ be such that $u_j \geq u_i$ for $i = k, \ldots, l$. If $j \in \{k, l\}$, then the right-hand side of the implication (29) holds. Thus, let $k < j < l$. If $u_j > u_{j+1}$ then $u_{j-1} = u_j$ in view of (27). If $u_j = u_{j+1}$, then if follows from (18) that

$$0 \geq g_j = \left(\frac{\varepsilon + \beta_j\,\tilde{\varepsilon}}{h^2} + \frac{b}{2\,h}\right)(u_j - u_{j-1}) \geq 0$$

and hence again $u_{j-1} = u_j$. Repeating the above argument, one deduces that $u_j = u_{j-1} = \cdots = u_k$ so that the right-hand side of (29) is satisfied.

The implications (28) and (30) follow analogously. □

Corollary 1. *Consider any $\tilde{\varepsilon} \geq b\,h/2 - \varepsilon$. Let $u_0, \ldots, u_{n+1}$ be a solution of the nonlinear problem (17)–(19) with $g_i \geq 0$, $i = 1, \ldots, n$. Let $j \in \{0, \ldots, n+1\}$ satisfy $u_j \geq u_i$, $i = 0, \ldots, n+1$. Then the solution increases monotonically until u_j and, after that, it decreases monotonically, i.e.,*

$$u_0 \leq u_1 \leq \cdots \leq u_j\,, \qquad u_j \geq u_{j+1} \geq \cdots \geq u_{n+1}\,. \tag{31}$$

If $g_i = 0$, $i = 1, \ldots, n$, then the solution is monotone, i.e.,

$$u_0 \leq u_1 \leq \cdots \leq u_{n+1} \qquad \textit{or} \qquad u_0 \geq u_1 \geq \cdots \geq u_{n+1}\,. \tag{32}$$

Proof. If $0 < i < j$, then $u_i \geq \min\{u_j, u_{i-1}\} = u_{i-1}$. If $j < i < n+1$, then $u_i \geq \min\{u_j, u_{i+1}\} = u_{i+1}$. Therefore, (31) holds. If $g_i = 0$, $i = 1, \ldots, n$, then $u_j = \max\{u_0, u_{n+1}\}$ according to (29) so that (32) follows from (31). □

Corollary 2. *Consider any $\tilde{\varepsilon} > b\,h/2 - \varepsilon$. Let $u_0, \ldots, u_{n+1}$ be a solution of the nonlinear problem (17)–(19) with $g_i \geq 0$, $i = 1, \ldots, n$. Let $j \in \{0, \ldots, n+1\}$ satisfy $u_j \geq u_i$, $i = 0, \ldots, n+1$. If $j < n$, $i \in \{j+1, \ldots, n\}$, and $g_i > 0$, then $u_i > u_{i+1}$ and $\beta_i = 1$. If $g_i = 0$ for some $i \in \{1, \ldots, n\}$, then either $u_{i-1} = u_i = u_{i+1}$ or*

$$\frac{u_i - u_{i-1}}{u_{i+1} - u_i} < 1\,.$$

Finally, if $u_L > u_R$, one obtains

$$g_i = 0\,, \quad i = 1, \ldots, n \qquad \Rightarrow \qquad u_0 > u_1 > \cdots > u_{n+1}\,, \quad \beta_1 = \beta_2 = \cdots = \beta_n = 1\,.$$

Proof. According to (18), one has

$$\left(\varepsilon + \beta_i\,\tilde{\varepsilon} + \frac{b\,h}{2}\right)(u_i - u_{i-1}) + \left(\varepsilon + \beta_i\,\tilde{\varepsilon} - \frac{b\,h}{2}\right)(u_i - u_{i+1}) = g_i\,h^2 \tag{33}$$

for $i = 1, \ldots, n$. If $i > j$, then $u_{i-1} \geq u_i \geq u_{i+1}$ due to (31) and hence the first term on the left-hand side of (33) is nonpositive. Therefore, (33) can be satisfied with $g_i > 0$ only if the second term on the left-hand side of (33) is positive, which implies that $u_i > u_{i+1}$ and $\beta_i = 1$. Furthermore, for any $i \in \{1, \ldots, n\}$ such that $g_i = 0$ and $u_i \neq u_{i+1}$, one deduces from (33) that

$$\frac{u_i - u_{i-1}}{u_{i+1} - u_i} = \frac{\varepsilon + \beta_i\,\tilde{\varepsilon} - \dfrac{b\,h}{2}}{\varepsilon + \beta_i\,\tilde{\varepsilon} + \dfrac{b\,h}{2}} < 1\,. \tag{34}$$

If $g_i = 0$ and $u_i = u_{i+1}$, then obviously also $u_i = u_{i-1}$.

Finally, let $g_i = 0$, $i = 1, \ldots, n$. If $u_k = u_{k+1}$ for some $k \in \{0, \ldots, n\}$, then according to (33) with $i = k$ and $i = k+1$, one obtains $u_k = u_{k-1}$ (if $k > 0$) and $u_{k+1} = u_{k+2}$ (if $k < n$). Thus, one deduces that $u_0 = u_1 = \cdots = u_{n+1}$. Therefore, if $u_L > u_R$, one gets $u_i \neq u_{i+1}$ for $i = 0, \ldots, n$ and hence (32) implies that $u_0 > u_1 > \cdots > u_{n+1}$. Consequently, for any $i \in \{1, \ldots, n\}$, the left-hand side of (34) is positive and therefore $\beta_i = 1$. □

Corollary 3. *Let $\tilde{\varepsilon} = b\,h/2 - \varepsilon$. Let $u_0, \ldots, u_{n+1}$ be a solution of the nonlinear problem (17)–(19) with $g_i \geq 0$, $i = 1, \ldots, n$. Let $j \in \{0, \ldots, n+1\}$ satisfy $u_j \geq u_i$, $i = 0, \ldots, n+1$. Then either $j \geq n$ or $g_{j+1} = \cdots = g_n = 0$ and $u_j = u_{j+1} = \cdots = u_n$.*

If $i \in \{1, \ldots, n\}$ and $g_i = 0$, then either $u_{i-1} = u_i = u_{i+1}$ or $u_i = u_{i-1}$ and $\beta_i = 1$. Consequently,

$$g_i = 0\,, \quad i = 1, \ldots, n \qquad \Rightarrow \qquad u_i = u_L\,, \quad i = 1, \ldots, n\,.$$

Proof. Let $j < n$ and $i \in \{j+1, \ldots, n\}$. Then the left-hand side of (33) is nonpositive due to (31) and hence (33) cannot hold with $g_i > 0$. Therefore, $g_{j+1} = \cdots = g_n = 0$. If $g_i = 0$ for some $i \neq j$, then it follows from (31) and (33) that $u_i = u_{i-1}$, which completes the proof of the first statement of the corollary. If $j \in \{1, \ldots, n\}$ and $g_j = 0$, then $u_j = u_{j-1}$ since otherwise $u_j > u_{j-1}$ and, in view of (33), $u_j > u_{j+1}$ and $\beta_j = 0$, which is in contradiction with (19). Thus, for any $i \in \{1, \ldots, n\}$ such that $g_i = 0$ one has $u_i = u_{i-1}$ and it follows from (33) that $u_i = u_{i+1}$ or $\beta_i = 1$. □

Remark 2. *Let $u_L > u_R$ and $g_i = 0$ for $i = 1, \ldots, n$. It follows from Corollaries 2 and 3 that if a solution of the nonlinear problem* (17)–(19) *exists, then it is determined uniquely. It is the solution of* (12) *with ε replaced by $\varepsilon + \tilde{\varepsilon}$. Thus, the nonlinear problem is solvable if this solution leads to $\beta_1 = \cdots = \beta_n = 1$ in case of $\tilde{\varepsilon} > b\,h/2 - \varepsilon$ and to $\beta_n = 1$ in case of $\tilde{\varepsilon} = b\,h/2 - \varepsilon$. If $\tilde{\varepsilon} = b\,h/2 - \varepsilon$, this means that $\beta_i = 1$ for $u_{i-1} = u_i \neq u_{i+1}$. This is the case for* (15) *and* (23) *but not necessarily for* (24). *If $\tilde{\varepsilon} > b\,h/2 - \varepsilon$, the solution is given by* (13) *with $g = 0$ and Pe replaced by*

$$Pe^* = \frac{b\,h}{2\,(\varepsilon + \tilde{\varepsilon})}\,.$$

Then, for any $i \in \{1, \ldots, n\}$,

$$\frac{u_i - u_{i-1}}{u_{i+1} - u_i} = \frac{1 - Pe^*}{1 + Pe^*} < \frac{1}{3} \quad \textit{for} \quad \tilde{\varepsilon} \in \left(\frac{b\,h}{2} - \varepsilon, \frac{b\,h}{2} \right].$$

Thus, the nonlinear problem is solvable if β_i is defined by (15) *or by* (23) *with $L \in [1/3, 1)$. On the other hand, if β_i satisfies* (19) *and $\beta_i = 0$ for $(u_i - u_{i-1})(u_{i+1} - u_i) \geq 0$, then the nonlinear problem is not solvable for any data. Unfortunately, also the favorable choice* (24) *does not lead to a solvable nonlinear problem in general. We shall return to this choice in Section 9, where it will be used for deriving a convenient definition of β_i.*

6. The solution of the nonlinear system and the choice of $\tilde{\varepsilon}$

In this section we report some numerical results obtained by solving the nonlinear problem (17), (18). We start by briefly describing the solution algorithm. The problem (17), (18) was solved by a fixed-point iteration: one chooses an initial guess $\underline{\boldsymbol{u}}^0$ for the solution $\underline{\boldsymbol{u}} := \{u_i\}_{i=0}^{n+1}$ and computes a sequence $\{\underline{\boldsymbol{u}}^k\}$ where each $\underline{\boldsymbol{u}}^k$ with $k = 1, 2, \ldots$ solves the linearized problem (17), (18) with β_i determined by means of the already known discrete solution $\underline{\boldsymbol{u}}^{k-1}$. In our case, the initial guess $\underline{\boldsymbol{u}}^0$ was computed as the solution of (17), (18) with $\beta_i = 1$, $i = 1, \ldots, n$. We shall prove in Section 7 that the linear problems defining this fixed-point algorithm are well-posed. The iteration was stopped if the coefficients β_i did not change.

Since this section focuses on the choice of $\tilde{\varepsilon}$, we only shall present results obtained for β_i defined by (15). To suppress the influence of the rounding errors on the validity of the conditions in (15) for setting $\beta_i = 1$, we replaced (15) by

$$\beta_i = \begin{cases} 1 & \text{if} \quad u_i + \tau < u_{i+1} \quad \text{and} \quad 2\,u_i + \tau < u_{i-1} + u_{i+1} \\ & \text{or} \quad u_i - \tau > u_{i+1} \quad \text{and} \quad 2\,u_i - \tau > u_{i-1} + u_{i+1}\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{35}$$

with a suitable positive constant τ. In the computations presented in this section, we used $\tau = 10^{-12}$. For $\tau = 0$, the relations (15) and (35) are equivalent.

As we pointed out in the previous section, any $\tilde{\varepsilon}$ satisfying (26) can be used in (18). Then, a natural question is which choice of $\tilde{\varepsilon}$ is most convenient. It is well known that if all the coefficients β_i in (18) are set to 1, then

$$\tilde{\varepsilon} = \frac{b\,h}{2}\left(\coth Pe - \frac{1}{Pe}\right) \tag{36}$$

is optimal in the sense that, for constant g, the discrete solution is nodally exact, i.e., $u_i = u(x_i)$ for $i = 1, \ldots, n$, see [5]. On the other hand, in general, the parameter $\tilde{\varepsilon}$ cannot be chosen in such a way that the discrete solution is nodally exact if the coefficients β_i are defined by (15). However, it is well known that the performance of most stabilized methods is primarily affected by the amount of artificial diffusion introduced near the numerical layers, and quite insensitive to the changes on it far away from them. Thus, since we expect that $\beta_i = 1$ in numerical boundary layers, it may be of advantage to use $\tilde{\varepsilon}$ given by (36) also when the coefficients β_i are defined by (15). Then what is required is that the exact solution solves the scheme (18) for the nodes x_i where $\beta_i = 1$. Note that the parameter $\tilde{\varepsilon}$ defined in (36) is larger than $\tilde{\varepsilon}$ from (16) and smaller than $\tilde{\varepsilon}$ from (21).

In what follows, we shall compare solutions of the problem given by (17), (18), (35) for $\tilde{\varepsilon}$ defined by (16), (21), and (36). We shall consider

$$b = g = 1\,, \qquad u_L = u_R = 0\,, \tag{37}$$

and various choices of ε and n.

First, we notice that if $\tilde{\varepsilon}$ is defined by (16), it is easy to verify that, for the data (37) and any ε and n,

$$u_i = i\,h\,, \quad i = 0, \ldots, n\,, \qquad u_{n+1} = 0$$

is a solution of (17), (18) with β_i given by (35) or any β_i satisfying (19) (it is the only solution of the respective nonlinear problem). In this case, $\beta_n = 1$ and if β_i is defined by (35) or (15), one has $\beta_i = 0$ for $i = 1, \ldots, n-1$. Since the discrete solution is independent of ε, one cannot expect a good approximation of the exact solution for the whole range of the values of ε. Indeed, according to (10), the error of the discrete solution satisfies

$$u_i - u(x_i) = \frac{e^{-(1-x_i)/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}\,, \qquad i = 0, \ldots, n\,, \tag{38}$$

so that the largest error appears at the node x_n and, for $\varepsilon \le 0.1$, one has

$$u_n - u(x_n) = \frac{e^{-h/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}} > 0.135 \qquad \text{for} \quad Pe \to 1\,.$$

To see the impact of the nonlinear artificial diffusion in (18) on the discrete solutions, we computed the errors

$$\left(\frac{1}{n}\sum_{i=1}^{n}\left(u(x_i) - u_i\right)^2\right)^{1/2} \tag{39}$$

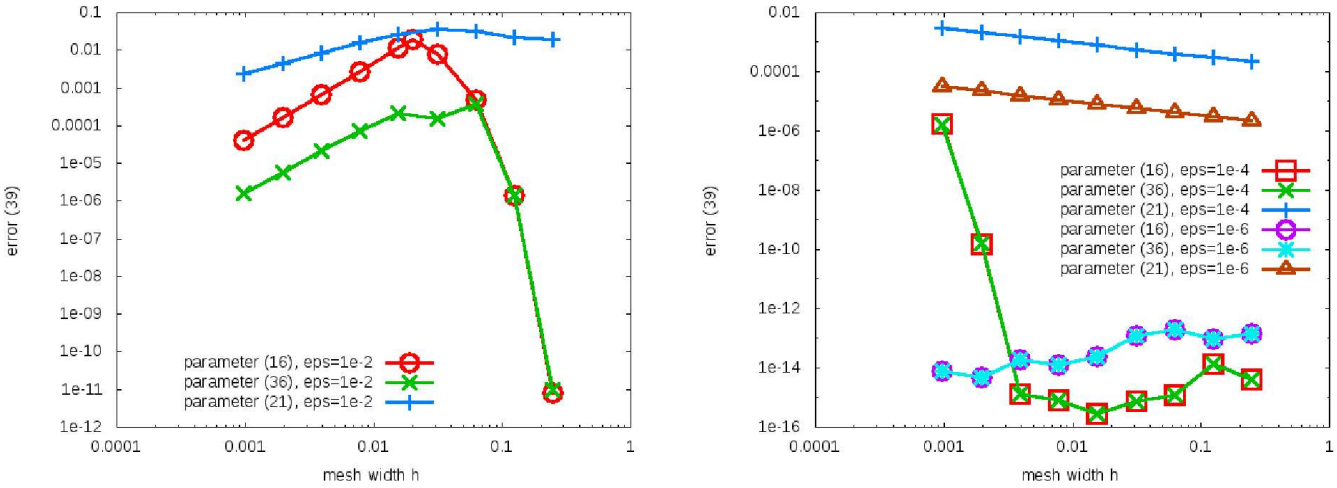

Figure 1: Dependence of the errors of the solutions of (17), (18), (35) on h for $\tilde{\varepsilon}$ defined by (16), (36), and (21) and for $\varepsilon = 10^{-2}$ (left) and $\varepsilon \in \{10^{-4}, 10^{-6}\}$ (right).

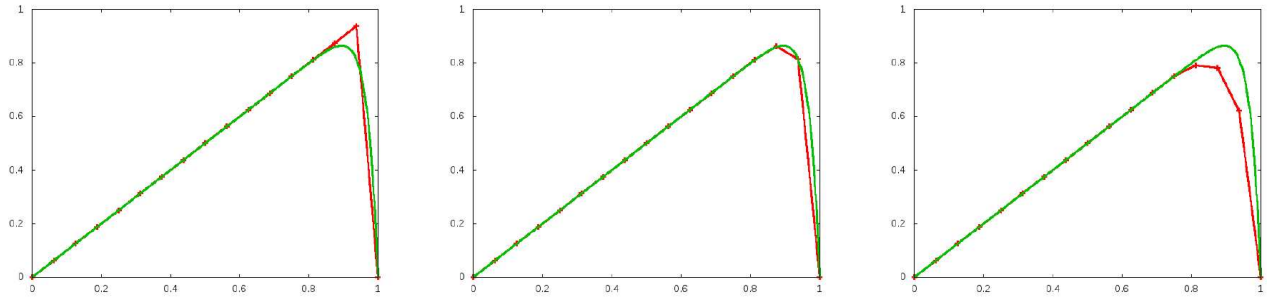

Figure 2: Comparisons of the exact solution (green) and solutions of (17), (18), (35) for $\varepsilon = 0.03$, $n = 15$, and $\tilde{\varepsilon}$ defined by (16), (36), and (21).

for different values of h, the three definitions of $\tilde{\varepsilon}$ (cf. (16), (36), and (21)), and for $\varepsilon \in \{10^{-2}, 10^{-4}, 10^{-6}\}$. If $\tilde{\varepsilon}$ is defined by (16), we set $\tilde{\varepsilon} = 0$ for $Pe \leq 1$ (this situation occurs only for $\varepsilon = 10^{-2}$). The results are depicted in Fig. 1, where we observe that the best results are obtained for $\tilde{\varepsilon}$ defined by (36). For large Péclet numbers, comparable errors are also obtained for $\tilde{\varepsilon}$ defined by (16). The choice (21) always adds too much artificial diffusion and leads to the worst results. To further stress this, Fig. 2 depicts the discrete solutions corresponding to $Pe = 25/24$ and clearly demonstrates the differences between the three choices of $\tilde{\varepsilon}$.

One final comment is in place for the case where $\tilde{\varepsilon}$ is given by (16). In this case, according to (38), the error (39) is bounded by $e^{-h/\varepsilon}$. This shows that, for $\varepsilon = 10^{-6}$ (and partly also for $\varepsilon = 10^{-4}$), the errors depicted in Fig. 1 are results of rounding errors and are much larger than actual values of the errors.

7. Solvability of the linear subproblems

At the beginning of the previous section, the solution of the nonlinear problem (17), (18) using a fixed-point iteration was described. In this section, we shall discuss under which conditions the corresponding linear subproblems are uniquely solvable.

We shall consider the following more general problem: given positive numbers $d_1, \ldots, d_n$, find $u_1, \ldots, u_n$ such that

$$-d_i\,(u_{i-1} - 2\,u_i + u_{i+1}) + u_{i+1} - u_{i-1} = \tilde{g}_i\,, \quad i = 1, \ldots, n\,, \tag{40}$$

where $u_0 = u_L$ and $u_{n+1} = u_R$. This problem corresponds to (18) for $d_i = 2\,(\varepsilon + \beta_i\,\tilde{\varepsilon})/(b\,h)$ and $\tilde{g}_i = 2\,h\,g_i/b$.

The following theorem proves the unique solvability of problem (40) in the case that the coefficients d_i are allowed to take the values 1 and d with $d > 0$. As a consequence, the unique solvability of the linearized problem (18) with $\tilde{\varepsilon}$ given by (16) follows.

Theorem 2. *Let $d_1, \ldots, d_n \in \{1, d\}$ with an arbitrary $d > 0$. Then problem (40) has a unique solution.*

Proof. It suffices to show that the homogeneous problem corresponding to (40) has only the trivial solution, i.e., that if

$$-d_i\,(u_{i-1} - 2\,u_i + u_{i+1}) + u_{i+1} - u_{i-1} = 0\,, \quad i = 1, \ldots, n\,, \tag{41}$$

with $u_0 = u_{n+1} = 0$, then

$$u_1 = u_2 = \cdots = u_n = 0\,. \tag{42}$$

Let $1 \le K \le L \le n$ and $d_K = d_{K+1} = \cdots = d_L = d$. Multiplying the i-th equation in (41) by u_i and summing up over $i = K, \ldots, L$, one obtains

$$d\,u_K^2 + d\sum_{i=K}^{L-1}(u_i - u_{i+1})^2 + d\,u_L^2 - (1+d)\,u_{K-1}\,u_K + (1-d)\,u_L\,u_{L+1} = 0\,. \tag{43}$$

Thus, if $d_1 = d_2 = \cdots = d_n = d$, one may set $K = 1$ and $L = n$, and (43) readily implies (42). Of course, this result also follows from the equivalence between (11) and (12) and the fact that (11) is uniquely solvable.

It remains to investigate the case when the values of d_i are not all equal. Let $K \in \{1, \ldots, n\}$ be the smallest index such that $d_K = d$ and let $L \in \{K, \ldots, n\}$ be the largest index such that $d_K = d_{K+1} = \cdots = d_L = d$. Then, for any $i \in \{1, \ldots, K-1\}$, one has $d_i = 1$ and hence $u_i = u_{i-1}$. Consequently, $u_i = 0$ for $i = 0, \ldots, K-1$. Furthermore, if $L < n$, then $d_{L+1} = 1$ and hence $u_{L+1} = u_L$, which implies that $d\,u_L^2 + (1-d)\,u_L\,u_{L+1} \ge 0$. This inequality is satisfied also if $L = n$ since then $u_{L+1} = 0$. Thus, one deduces from (43) that

$$d\,u_K^2 + d\sum_{i=K}^{L-1}(u_i - u_{i+1})^2 \le 0\,,$$

which gives $0 = u_K = u_{K+1} = \cdots = u_L$. Repeating the above arguments until $L = n$, one obtains (42). □

The following theorem proves the unique solvability of (40) for a more general choice of $d_1, \ldots, d_n$.

Theorem 3. *Let $d_1, \ldots, d_n \in (0, 1]$. Then problem (40) has a unique solution. Furthermore, if $d_1, \ldots, d_n \in [\delta, 1 + \delta]$ with $\delta \in (0, 1]$, then problem (40) has a unique solution as well. However, for any $\delta > 0$, there are $d_1, \ldots, d_n \in (0, 1 + \delta]$ such that problem (40) is not uniquely solvable.*

Proof. We introduce the $n \times n$ matrices

$$\mathbb{B} = \mathrm{diag}(d_1, d_2, \ldots, d_n)\,, \qquad \mathbb{C} = \mathrm{tridiag}(-1, 2, -1)\,, \qquad \mathbb{E} = \mathrm{tridiag}(-1, 0, 1)\,.$$

Then the matrix corresponding to (40) is $\mathbb{B}\,\mathbb{C} + \mathbb{E}$. This matrix will be transformed by operations which preserve full rank such that it becomes possible to see that its determinant does not vanish.

Let $\mathbb{G} = (g_{ij})_{i,j=1}^n$ be a symmetric matrix given by

$$g_{ij} = (n - i + 1)\, j\,, \qquad j = 1, \ldots, i\,, \quad i = 1, \ldots, n\,.$$

Then $\mathbb{C}\,\mathbb{G} = (n + 1)\,\mathbb{I}$, where $\mathbb{I}$ is the identity matrix. Setting $\mathbb{Q} = (\mathbb{B}\,\mathbb{C} + \mathbb{E})\,\mathbb{G}$, one obtains a matrix with the entries

$$\begin{aligned} q_{ij} &= -2\, j + 2\,(n + 1) && \text{for} \quad i = 1, \ldots, j - 1\,, \\ q_{jj} &= -2\, j + (n + 1)\,(1 + d_j)\,, \\ q_{ij} &= -2\, j && \text{for} \quad i = j + 1, \ldots, n\,, \end{aligned}$$

where $j = 1, \ldots, n$. Now, let us define the matrix $\mathbb{Z} = (z_{ij})_{i,j=1}^n$ by

$$z_{ij} = \frac{1}{n + 1}\,(q_{ij} - q_{i+1,j})\,, \quad i = 1, \ldots, n - 1\,, \qquad z_{nj} = \frac{1}{n + 1}\left(2\, q_{nj} + \sum_{i=1}^{n-1} q_{ij}\right),$$

where $j = 1, \ldots, n$. Then $\det(\mathbb{B}\,\mathbb{C} + \mathbb{E}) \neq 0$ if and only if $\det \mathbb{Z} \neq 0$ and one has

$$\begin{aligned} &z_{ii} = 1 + d_i\,, \quad z_{i,i+1} = 1 - d_{i+1}\,, \quad z_{ij} = 0 \ \text{ for } j \notin \{i, i + 1\}\,, \qquad i = 1, \ldots, n - 1\,, \\ &z_{nj} = -1 + d_j\,, \quad j = 1, \ldots, n - 1\,, \qquad z_{nn} = 2\, d_n\,. \end{aligned}$$

Let $\mathbb{Z}^{ij}$ be the $(n - 1) \times (n - 1)$ matrix obtained from $\mathbb{Z}$ by removing the i-th row and j-th column. Then

$$\det \mathbb{Z}^{nj} = \prod_{k=1}^{j-1} (1 + d_k) \prod_{l=j+1}^{n} (1 - d_l)\,. \tag{44}$$

Let n be odd and denote

$$\tilde{z}_{nj} = \sum_{\substack{i=1 \\ i \text{ is odd}}}^{n} z_{ij}\,, \qquad j = 1, \ldots, n\,.$$

Then, for $j = 1, \ldots, n$, one has

$$\tilde{z}_{nj} = 2\, d_j \qquad \text{if } j \text{ is odd}\,, \qquad\qquad \tilde{z}_{nj} = 0 \qquad \text{if } j \text{ is even}\,.$$

Thus,

$$\det \mathbb{Z} = 2 \sum_{\substack{j=1 \\ j \text{ is odd}}}^{n} d_j \, \det \mathbb{Z}^{nj} \,. \tag{45}$$

If $d_1, \ldots, d_n \in (0,1]$, then $\det \mathbb{Z}^{nj} \geq 0$, $j = 1, \ldots, n-1$, and $\det \mathbb{Z}^{nn} > 0$ so that $\det \mathbb{Z} > 0$. If n is even, then

$$\det \mathbb{Z} = (1 + d_1) \, \det \mathbb{Z}^{11} + (1 - d_1) \, \det \mathbb{Z}^{n1} \,. \tag{46}$$

Since $\mathbb{Z}^{11}$ has the same structure as $\mathbb{Z}$ and has an odd number of rows and columns, one has $\det \mathbb{Z}^{11} > 0$ for $d_1, \ldots, d_n \in (0,1]$. Moreover, $\det \mathbb{Z}^{n1} \geq 0$ in view of (44) and hence again $\det \mathbb{Z} > 0$, which proves the first part of the theorem.

Now let $d_1, \ldots, d_n \in [\delta, 1+\delta]$ with $\delta \in (0,1]$. We denote

$$A_s = \prod_{k=1}^{s} (1 + d_k) \,, \qquad B_s = \prod_{l=s}^{n} (1 - d_l) \,, \qquad\qquad s = 1, \ldots, n \,,$$

and we set $A_0 = 1$. If $B_s < 0$, then, for some $k \in \{1, \ldots, n\}$, we have $|1 - d_k| \leq \delta$. Therefore, since $|1 - d_l| \leq 1$ for any $l \in \{1, \ldots, n\}$, one gets

$$B_s \geq -\delta \,, \qquad s = 1, \ldots, n \,. \tag{47}$$

First, let n be odd and let us prove that, for any odd $m \in \{1, \ldots, n\}$, the matrices $\mathbb{Z}^{nj}$ satisfy

$$\sum_{\substack{j=m \\ j \text{ is odd}}}^{n} d_j \, \det \mathbb{Z}^{nj} \geq d_n \, A_{m-1} \,. \tag{48}$$

In view of (44), this inequality holds for $m = n$. Let us assume that (48) holds for a given odd $m \in \{3, \ldots, n\}$. Then, again in view of (44),

$$\begin{aligned} \sum_{\substack{j=m-2 \\ j \text{ is odd}}}^{n} d_j \, \det \mathbb{Z}^{nj} &\geq d_n \, A_{m-1} + d_{m-2} \, A_{m-3} \, B_{m-1} \\ &> d_n \, A_{m-3} + d_{m-2} \, A_{m-3} \left[d_n \, (1 + d_{m-1}) + B_{m-1} \right] > d_n \, A_{m-3} \end{aligned}$$

since $d_n \, (1 + d_{m-1}) > \delta$ and $B_{m-1} \geq -\delta$, see (47). Thus, (48) holds for any odd $m \in \{1, \ldots, n\}$ and hence, setting $m = 1$ and using (45), one gets $\det \mathbb{Z} \geq 2 \, d_n$. If n is even, then $\det \mathbb{Z}^{11} \geq 2 \, d_n$ and hence, according to (46), $\det \mathbb{Z} = (1 + d_1) \, \det \mathbb{Z}^{11} + B_1 > 2 \, d_n + B_1 \geq d_n$.

Finally, let us consider any $\delta > 0$ and set

$$d_1 = d_2 = \cdots = d_{n-2} = 1 \,, \qquad d_{n-1} = 1 + \delta \,, \qquad d_n = \frac{\delta}{3 \, \delta + 4} \,.$$

Then $d_1, \ldots, d_n \in (0, 1+\delta]$ and

$$\det \mathbb{Z} = 2^{n-2} \, \det \begin{pmatrix} 1 + d_{n-1} & 1 - d_n \\ -1 + d_{n-1} & 2 \, d_n \end{pmatrix} = 0 \,.$$

Consequently, the matrix corresponding to (40) is singular and hence problem (40) is not uniquely solvable. □

The following corollary states the unique solvability of the linearized problem (17), (18) for any $\tilde{\varepsilon}$ satisfying (26).

Corollary 4. *Consider any $\tilde{\varepsilon} \in [0, b\,h/2]$ and any $\beta_1, \ldots, \beta_n \in [0,1]$. Then the linear problem* (17), (18) *has a unique solution.*

Proof. Since (18) is equivalent to (40) with $d_1, \ldots, d_n \in [1/Pe, 1 + 1/Pe]$, the statement follows immediately from Theorem 3. □

8. Solvability of the nonlinear problem

The computations reported in Section 6 were the ones for which convergence of the fixed-point iteration was achieved. However, some other computations we performed did not converge at all. In some cases, a convergence was obtained after changing the value of τ in (35) (although we realized that the iterative process was still very sensitive to rounding errors). For some other cases though, we were not able to find any way to achieve a convergence and hence no solution at all was found. The ultimate conclusion of these numerical experiments was that the nonlinear problem (17)–(19) is not solvable in general. In this section we first describe examples of data for which the nonlinear problem has no solution, thus proving the above claim. This lack of solvability is due to the discontinuous character of the coefficients β_i. As a matter of fact, at the end of the present section we shall prove that the problem (17), (18) is solvable if one considers coefficients β_i depending on the discrete solution in a continuous way.

Let us start with the following remark. If the nonlinear problem (17), (18) with some functions β_i satisfying (19) has a solution, then there are numbers $\bar{\beta}_1, \ldots, \bar{\beta}_n \in \{0,1\}$ such that, after having computed the solution $\underline{\boldsymbol{u}} = \{u_i\}_{i=0}^{n+1}$ of (17), (18) with $\beta_i = \bar{\beta}_i$, $i = 1, \ldots, n$, one has $\beta_i(\underline{\boldsymbol{u}}) = \bar{\beta}_i$, $i = 1, \ldots, n$. Since there are only 2^n admissible choices of $\bar{\beta}_1, \ldots, \bar{\beta}_n$, one can easily check (at least for small n) whether the nonlinear problem is solvable or not. In what follows, we shall consider the three choices of $\tilde{\varepsilon}$ tested in Section 6 and, for each of them, we shall present an example of data such that the nonlinear problem (17), (18) is not solvable for any functions β_i satisfying (19) and

$$\beta_i = 0 \qquad \text{if} \qquad u_i \neq u_{i+1} \quad \text{and} \quad \frac{u_i - u_{i-1}}{u_{i+1} - u_i} > 1\,. \tag{49}$$

These requirements are met by all the three choices (15), (23), and (24). In all the cases, we shall use

$$n = 4\,, \qquad u_L = u_R = 0\,. \tag{50}$$

First, let us study the problem (17), (18), (15) with $\tilde{\varepsilon}$ defined by (16). We consider the data

$$\varepsilon = 0.03\,, \qquad b = 1\,, \qquad g_1 = 6\,, \quad g_2 = -6\,, \quad g_3 = 3\,, \quad g_4 = -2\,. \tag{51}$$

As explained above, for each of the 16 possible choices of $\bar{\beta}_1, \ldots, \bar{\beta}_4$, we compute the solution $\underline{\boldsymbol{u}} = \{u_i\}_{i=0}^{5}$ of (17), (18) with $\beta_i = \bar{\beta}_i$, $i = 1, \ldots, 4$. These solutions together with the values of $\beta_1(\underline{\boldsymbol{u}}), \ldots, \beta_4(\underline{\boldsymbol{u}})$ computed according to (15) are shown in Figs. 3 and 4. Since $(\beta_1(\underline{\boldsymbol{u}}), \ldots, \beta_4(\underline{\boldsymbol{u}}))$ always differs from $(\bar{\beta}_1, \ldots, \bar{\beta}_4)$, one concludes that the nonlinear problem (17), (18), (15) does not have any solution. Note that, for all choices of $\bar{\beta}_1, \ldots, \bar{\beta}_4$ except $\bar{\beta}_1 = \cdots = \bar{\beta}_4 = 1$, there always exists $j \in \{1,2,3,4\}$ such that $\bar{\beta}_j = 0$ and the solution $\underline{\boldsymbol{u}}$ has an extremum at the node x_j so that $\beta_j(\underline{\boldsymbol{u}}) = 1$ as soon as (19) holds. If $\bar{\beta}_1 = \cdots = \bar{\beta}_4 = 1$,

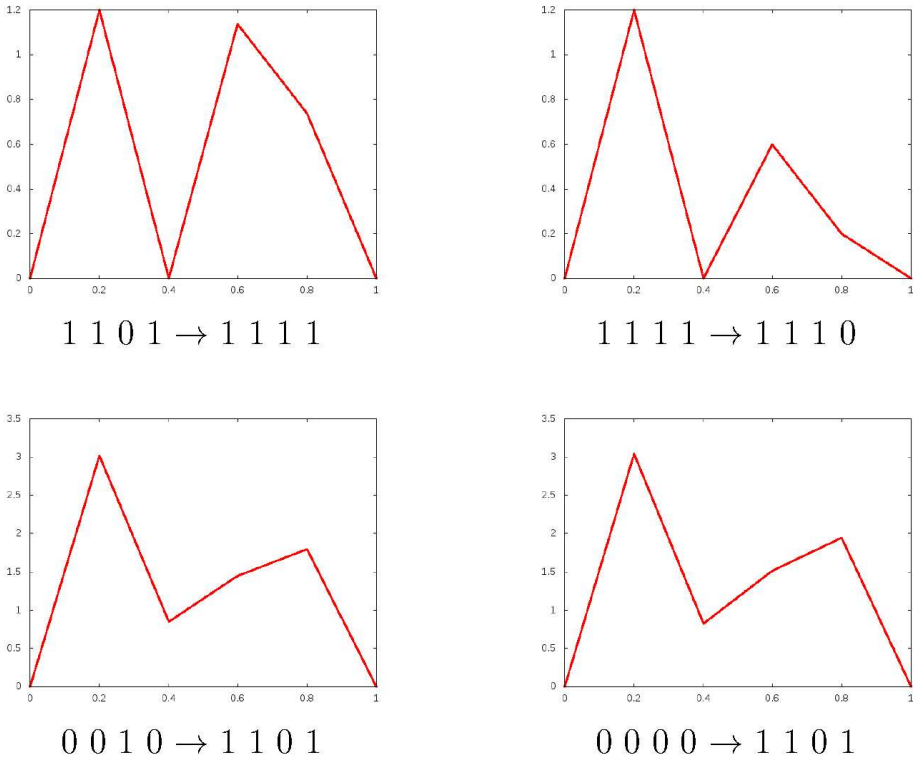

Figure 3: Solutions $\underline{\boldsymbol{u}}$ of (17), (18) with $\beta_i = \bar{\beta}_i$, $i = 1, \ldots, 4$ for the data (50), (51) and $\tilde{\varepsilon}$ defined by (16). The numbers left to '$\rightarrow$' represent $\bar{\beta}_1, \ldots, \bar{\beta}_4$, the numbers right to '$\rightarrow$' represent $\beta_1(\underline{\boldsymbol{u}}), \ldots, \beta_4(\underline{\boldsymbol{u}})$ corresponding to the respective solution according to (15).

one observes that $\beta_4(\underline{\boldsymbol{u}}) = 0$ as soon as (49) holds. This shows that the problem (17), (18) is not solvable for any functions β_i satisfying (19) and (49).

Similar non-existence studies were performed for the case in which $\tilde{\varepsilon}$ is defined by (36) and (21). For both cases we were able to find various right-hand sides for which the discrete problem does not have a solution. For example, if $\tilde{\varepsilon}$ is defined by (36), then the nonlinear problem with any β_i satisfying (19) and (49) is not solvable for the following data:

$$\varepsilon = 0.09\,, \qquad b = 1\,, \qquad g_1 = 6\,, \quad g_2 = g_3 = g_4 = 1\,. \tag{52}$$

Finally, if $\tilde{\varepsilon}$ is defined by (21), then the nonlinear problem with any β_i satisfying (19) and (49) is not solvable, e.g., for

$$\varepsilon = 0.064\,, \qquad b = 1\,, \qquad g_1 = g_2 = g_3 = g_4 = 1\,. \tag{53}$$

We have verified that the nonexistence of a solution to the nonlinear problem (17), (18) in the cases presented in this section is not caused by rounding errors.

Now, as we already stated, we present a result ensuring the solvability of the nonlinear problem (17), (18) under the hypothesis of continuity of the coefficients β_i.

Theorem 4. *Let $\beta_i : \mathbb{R}^{n+2} \rightarrow [0,1]$, $i = 1, \ldots, n$, be continuous functions and let $\tilde{\varepsilon} \in [0, b\,h/2]$. Then there exists a solution of the nonlinear problem* (17), (18).

Proof. We set $\boldsymbol{\beta}(\underline{\boldsymbol{u}}) := \{\beta_i(\underline{\boldsymbol{u}})\}_{i=1}^{n}$ with $\underline{\boldsymbol{u}} = \{u_i\}_{i=0}^{n+1}$. We also denote $\mathbb{M}(\boldsymbol{\beta}) \in \mathbb{R}^{n\times n}$ the matrix corresponding to system (18) for a particular choice of the coefficients $\boldsymbol{\beta} \in \mathbb{R}^n$. Then

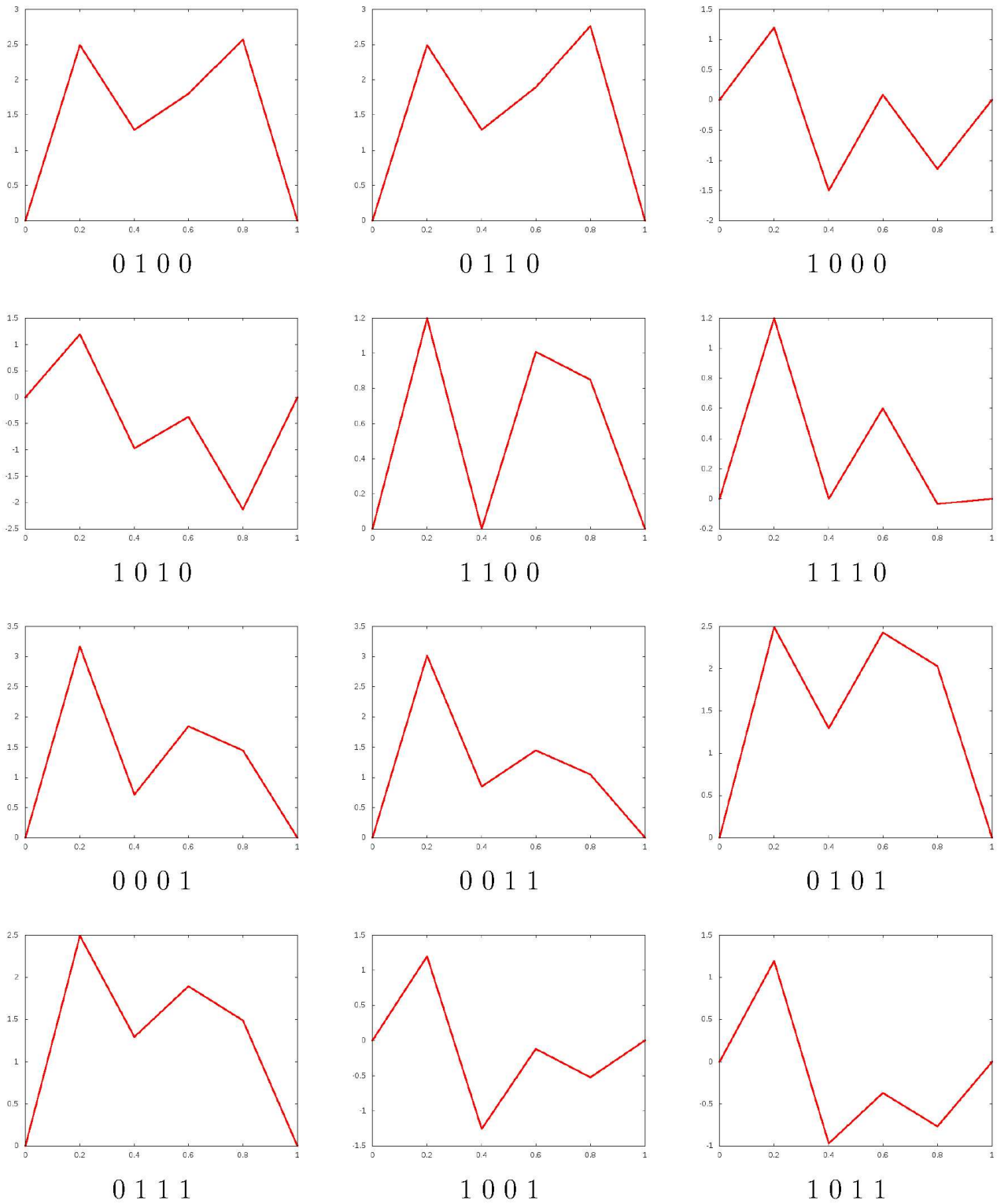

Figure 4: Solutions $\underline{\boldsymbol{u}}$ of (17), (18) with $\beta_i = \bar{\beta}_i$, $i = 1, \ldots, 4$ for the data (50), (51) and $\tilde{\varepsilon}$ defined by (16). The numbers below the graphs represent $\bar{\beta}_1, \ldots, \bar{\beta}_4$. For all solutions, the formula (15) gives $\beta_1(\underline{\boldsymbol{u}}) = \beta_2(\underline{\boldsymbol{u}}) = \beta_3(\underline{\boldsymbol{u}}) = \beta_4(\underline{\boldsymbol{u}}) = 1$.

the nonlinear problem (17), (18) can be written as: Find $\boldsymbol{u} \equiv \{u_i\}_{i=1}^n$ such that

$$\mathbb{M}(\boldsymbol{\beta}(\underline{\boldsymbol{u}}))\, \boldsymbol{u} = \tilde{\boldsymbol{g}}(\underline{\boldsymbol{u}})\,, \tag{54}$$

where $\underline{\boldsymbol{u}} = \{u_i\}_{i=0}^{n+1}$ with $u_0 = u_L$, $u_{n+1} = u_R$, and $\tilde{\boldsymbol{g}}(\underline{\boldsymbol{u}}) = \{\tilde{g}_i(\underline{\boldsymbol{u}})\}_{i=1}^n$ with $\tilde{g}_i(\underline{\boldsymbol{u}}) = g_i$ for

$i = 2, \ldots, n-1$, and

$$\tilde{g}_1(\underline{\boldsymbol{u}}) = g_1 + (\varepsilon + \beta_1(\underline{\boldsymbol{u}})\,\tilde{\varepsilon})\,\frac{u_L}{h^2} + b\,\frac{u_L}{2\,h}\,, \qquad \tilde{g}_n(\underline{\boldsymbol{u}}) = g_n + (\varepsilon + \beta_n(\underline{\boldsymbol{u}})\,\tilde{\varepsilon})\,\frac{u_R}{h^2} - b\,\frac{u_R}{2\,h}\,.$$

Since $|\beta_i(\underline{\boldsymbol{u}})| \le 1$ for $i = 1, \ldots, n$, one has

$$\|\tilde{\boldsymbol{g}}(\underline{\boldsymbol{u}})\| \le \|\boldsymbol{g}\| + \frac{\varepsilon + b\,h}{h^2}\,(|u_L| + |u_R|) \qquad \forall\ \underline{\boldsymbol{u}} \in \mathbb{R}^{n+2}\,, \tag{55}$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^n$ and $\boldsymbol{g} = \{g_i\}_{i=1}^n$.

Corollary 4 guarantees that the matrix $\mathbb{M}(\boldsymbol{\beta})$ is invertible for all $\boldsymbol{\beta}$ belonging to the hypercube $[0,1]^n$. Then, since the determinant of a matrix is a continuous function of its entries, there exists $\sigma_0 > 0$ such that

$$|\det \mathbb{M}(\boldsymbol{\beta})| \ge \sigma_0 \qquad \forall\ \boldsymbol{\beta} \in [0,1]^n\,.$$

Hence, the function $\boldsymbol{\beta} \mapsto [\mathbb{M}(\boldsymbol{\beta})]^{-1}$ is continuous on $[0,1]^n$, and there exists $C > 0$ such that

$$\|[\mathbb{M}(\boldsymbol{\beta})]^{-1}\| \le C \qquad \forall\ \boldsymbol{\beta} \in [0,1]^n\,, \tag{56}$$

where we use the matrix norm induced by the Euclidean norm on $\mathbb{R}^n$. Consequently, there exists a constant $C_0 > 0$ such that

$$\forall\ \boldsymbol{\beta} \in [0,1]^n,\ \boldsymbol{v} \in \mathbb{R}^n,\ \underline{\boldsymbol{u}} \in \mathbb{R}^{n+2}: \qquad \mathbb{M}(\boldsymbol{\beta})\,\boldsymbol{v} = \tilde{\boldsymbol{g}}(\underline{\boldsymbol{u}}) \quad \Rightarrow \quad \|\boldsymbol{v}\| \le C_0\,. \tag{57}$$

In view of (55) and (56), the constant C_0 depends on the data of (9) and, possibly, on h, but it does not depend on $\underline{\boldsymbol{u}}$.

Let now $T : \mathbb{R}^n \to \mathbb{R}^n$ be the mapping defined by

$$T\boldsymbol{u} := [\mathbb{M}(\boldsymbol{\beta}(\underline{\boldsymbol{u}}))]^{-1}\tilde{\boldsymbol{g}}(\underline{\boldsymbol{u}}) \qquad \forall\ \boldsymbol{u} \equiv \{u_i\}_{i=1}^n \in \mathbb{R}^n\,,$$

where $\underline{\boldsymbol{u}} = \{u_i\}_{i=0}^{n+1}$ with $u_0 = u_L$ and $u_{n+1} = u_R$. Then T is continuous and, according to (57), it maps the closed ball $B(0, C_0) := \{\boldsymbol{v} \in \mathbb{R}^n\,;\ \|\boldsymbol{v}\| \le C_0\}$ into itself. Applying Brouwer's fixed point theorem, there exists $\boldsymbol{u} \in B(0, C_0)$ such that $T\boldsymbol{u} = \boldsymbol{u}$, i.e., $\boldsymbol{u}$ satisfies (54). □

9. An example of continuous β_i and properties of the resulting solvable nonlinear discrete problem

In this section we propose a definition of continuous coefficients β_i that, according to Theorem 4, leads to a solvable nonlinear discrete problem, prove a corresponding (weaker) variant of the discrete maximum principle, and present a few numerical results.

For $i = 1, \ldots, n$, let us denote the derivatives of the discrete solution to the left and to the right of a point x_i by

$$u'_{i-} = \frac{u_i - u_{i-1}}{h}\,, \qquad u'_{i+} = \frac{u_{i+1} - u_i}{h}\,,$$

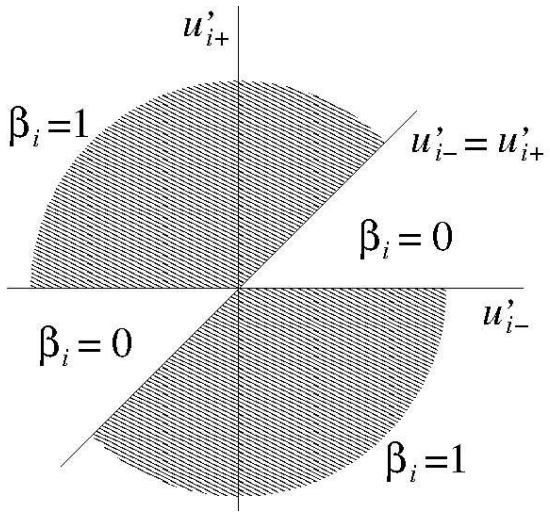

Figure 5: Values of β_i from (15) in dependence on u'_{i-} and u'_{i+}.

respectively. If β_i are defined by (15), then

$$\beta_i = \begin{cases} 1 & \text{if } u'_{i+} > \max\{0, u'_{i-}\} \quad \text{or} \quad u'_{i+} < \min\{0, u'_{i-}\}\,, \\ 0 & \text{if } \min\{0, u'_{i-}\} \le u'_{i+} \le \max\{0, u'_{i-}\}\,, \end{cases}$$

for $i = 1, \ldots, n$, see Fig. 5. Note that β_i is discontinuous along the lines $u'_{i-} = u'_{i+}$ and $u'_{i+} = 0$. Similarly, β_i is discontinuous if it is defined by (23) or (24).

Our aim is to introduce continuous coefficients β_i to guarantee the solvability of the nonlinear problem (17), (18). Based on the relation (24) and the discussion at the end of Section 4 and in Remark 2, we propose to set (cf. Fig. 6)

$$\beta_i = \begin{cases} 1 & \text{if } (u'_{i+} \ge \Delta + \max\{0, 2\,u'_{i-}\} \quad \text{or} \quad u'_{i+} \le -\Delta + \min\{0, 2\,u'_{i-}\})\,, \\ & \text{and } (u'_{i-}, u'_{i+}) \notin (-\Delta, D/2) \times (0, D + \Delta)\,, \\ & \text{and } (u'_{i-}, u'_{i+}) \notin (-D/2, \Delta) \times (-D - \Delta, 0)\,, \\ 0 & \text{if } \min\{0, 2\,u'_{i-}\} \le u'_{i+} \le \max\{0, 2\,u'_{i-}\}\,, \\ & \text{or } (u'_{i-}, u'_{i+}) \in [0, D/2] \times [0, D]\,, \\ & \text{or } (u'_{i-}, u'_{i+}) \in [-D/2, 0] \times [-D, 0]\,, \end{cases} \tag{58}$$

with positive parameters $\Delta \le D$. Furthermore, we require that β_i is continuous and that it is linear in each of the eight dark shadow subregions in Fig. 6. These requirements define the function β_i uniquely. The parameters D and Δ should be proportional to a characteristic derivative $\Delta u/\Delta x$, see (25).

Unfortunately, with the new definition of the coefficients β_i, we cannot guarantee the validity of the discrete maximum principle formulated in Theorem 1. Nevertheless, the following result shows that a possible violation of the discrete maximum principle is not significant if the parameter D or the mesh width h are small. The constant δ in the following theorem is related to the above definition of β_i by $\delta = D + \Delta$.

Theorem 5. *Consider any $\tilde{\varepsilon}$ satisfying* (26). *Let $u_0, \ldots, u_{n+1}$ be a solution of the nonlinear problem* (17), (18) *with any functions $\beta_1, \ldots, \beta_n \in [0, 1]$ satisfying*

$$\beta_i = 1 \qquad \text{if} \qquad u_i < \min\{u_{i-1}, u_{i+1} - \delta\,h\} \quad \text{or} \quad u_i > \max\{u_{i-1}, u_{i+1} + \delta\,h\}$$

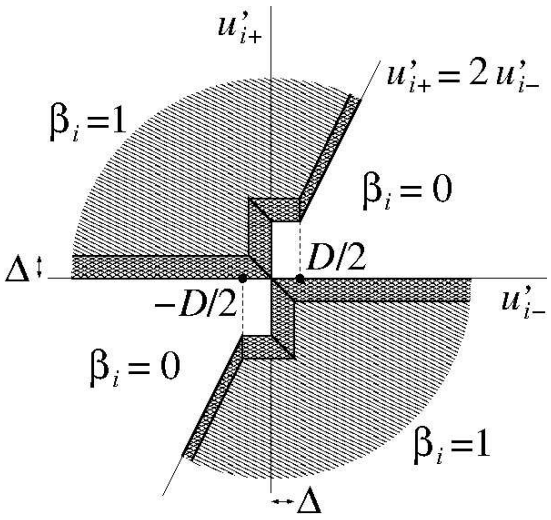

Figure 6: Definition of continuous β_i according to (58).

for some $\delta > 0$ *and* $i = 1, \ldots, n$. *Then*

$$\begin{aligned} g_i \le 0 \quad &\Rightarrow \quad u_i \le \max\{u_{i-1}, u_{i+1}\} \quad \textit{or} \quad u_i \le \min\{u_{i-1}, u_{i+1}\} + \delta\, h\,, \\ g_i \ge 0 \quad &\Rightarrow \quad u_i \ge \min\{u_{i-1}, u_{i+1}\} \quad \textit{or} \quad u_i \ge \max\{u_{i-1}, u_{i+1}\} - \delta\, h\,, \end{aligned}$$

for $i = 1, \ldots, n$. *Moreover, for any* $k, l \in \{0, 1, \ldots, n+1\}$ *with* $k + 1 < l$, *one has*

$$\begin{aligned} g_i \le 0\,, \quad i = k+1, \ldots, l-1 \quad &\Rightarrow \quad u_i < \max\{u_k, u_l\} + \delta\, h\,, \quad i = k, \ldots, l\,, \\ g_i \ge 0\,, \quad i = k+1, \ldots, l-1 \quad &\Rightarrow \quad u_i > \min\{u_k, u_l\} - \delta\, h\,, \quad i = k, \ldots, l\,. \end{aligned}$$

Proof. Consider any $i \in \{1, \ldots, n\}$ and let $g_i \le 0$. If $u_i - u_{i+1} \notin [0, \delta\, h]$, then $u_i \le \max\{u_{i-1}, u_{i+1}\}$ since the proof of Theorem 1 can be repeated without any changes. For $u_i - u_{i+1} \in [0, \delta\, h]$ it will be shown that $u_i \le \min\{u_{i-1}, u_{i+1}\} + \delta\, h$. To this end, assume that $u_i > \min\{u_{i-1}, u_{i+1}\} + \delta\, h$. Then

$$u_{i+1} + \delta\, h \ge u_i \ge u_{i+1}\,, \qquad u_i > u_{i-1} + \delta\, h\,.$$

Therefore, using (33) and noting that $(u_i - u_{i+1})$ is estimated either from below or from above depending on the sign of the term in front of it, one derives

$$\begin{aligned} 0 \ge g_i\, h^2 &= \left(\varepsilon + \beta_i\, \tilde{\varepsilon} + \frac{b\, h}{2}\right)(u_i - u_{i-1}) + \left(\varepsilon + \beta_i\, \tilde{\varepsilon} - \frac{b\, h}{2}\right)(u_i - u_{i+1}) \\ &> \left(\varepsilon + \beta_i\, \tilde{\varepsilon} + \frac{b\, h}{2}\right)\delta\, h + \min\left\{0, \varepsilon + \beta_i\, \tilde{\varepsilon} - \frac{b\, h}{2}\right\}\delta\, h > 0\,, \end{aligned}$$

which is a contradiction. Therefore, $u_i \le \min\{u_{i-1}, u_{i+1}\} + \delta\, h$.

Now consider any $k, l \in \{0, 1, \ldots, n+1\}$ with $k+1 < l$ and let $g_i \le 0$ for $i = k+1, \ldots, l-1$. First, we shall prove that, for any $i \in \{k+1, \ldots, l-1\}$, the following implication holds:

$$u_{i-1} \le u_i \quad \text{and} \quad u_i > u_{i+1} \qquad \Rightarrow \qquad u_k > u_{i+1}\,. \tag{59}$$

Thus, consider any $i \in \{k+1, \ldots, l-1\}$ such that the left-hand side of (59) is satisfied. Let $m \in \{k, \ldots, i-1\}$ be such that $u_s \le u_{s+1}$ for $s = m, \ldots, i-1$. We assume that m cannot be further decreased, i.e., either $m = k$ or $u_{m-1} > u_m$. According to (33), one has

$$\begin{aligned} 0 &\ge \left(\varepsilon + \beta_i\,\tilde{\varepsilon} + \frac{b\,h}{2}\right)(u_i - u_{i-1}) + \left(\varepsilon + \beta_i\,\tilde{\varepsilon} - \frac{b\,h}{2}\right)(u_i - u_{i+1}) \\ &> \left(\varepsilon + \frac{b\,h}{2}\right)(u_i - u_{i-1}) - \frac{b\,h}{2}\,(u_i - u_{i+1})\,. \end{aligned} \tag{60}$$

If $m < i-1$, then for $s = m+1, \ldots, i-1$, one derives in view of (33)

$$\begin{aligned} 0 &\ge \left(\varepsilon + \beta_s\,\tilde{\varepsilon} + \frac{b\,h}{2}\right)(u_s - u_{s-1}) + \left(-\varepsilon - \beta_s\,\tilde{\varepsilon} + \frac{b\,h}{2}\right)(u_{s+1} - u_s) \\ &\ge \left(\varepsilon + \frac{b\,h}{2}\right)(u_s - u_{s-1}) - \varepsilon\,(u_{s+1} - u_s)\,. \end{aligned} \tag{61}$$

Summing up the inequalities (60) and (61), one obtains

$$\begin{aligned} 0 &> \left(\varepsilon + \frac{b\,h}{2}\right)\sum_{s=m+1}^{i}(u_s - u_{s-1}) - \varepsilon\sum_{s=m+1}^{i-1}(u_{s+1} - u_s) - \frac{b\,h}{2}\,(u_i - u_{i+1}) \\ &= \left(\varepsilon + \frac{b\,h}{2}\right)(u_i - u_m) - \varepsilon\,(u_i - u_{m+1}) - \frac{b\,h}{2}\,(u_i - u_{i+1}) \ge \frac{b\,h}{2}\,(u_{i+1} - u_m)\,. \end{aligned}$$

Therefore, $u_m > u_{i+1}$, which is true also if $m = i-1$ according to (60). If $m = k$ or $u_s > u_{s+1}$ for $s = k, \ldots, m-1$, then the right-hand side of (59) holds. Otherwise $m \ge k+2$ and there is $i' \in \{k+1, \ldots, m-1\}$ for which left-hand side of (59) is satisfied and $u_{i'+1} > u_{i+1}$ holds. Hence the inequality $u_k > u_{i+1}$ follows by induction. For proving the statement of the theorem, let $j \in \{k, \ldots, l\}$ be such that $u_j = \max\{u_k, u_{k+1}, \ldots, u_l\}$ and let $u_j > \max\{u_k, u_l\}$. Then $u_j > u_{j+1}$ since otherwise $u_j = u_{j-1}$ in view of (33) and hence $u_j = u_k$ by induction. Thus, one has $u_k > u_{j+1}$ according to (59). Finally, applying the first part of the theorem, one obtains $u_j \le \min\{u_{j-1}, u_{j+1}\} + \delta\,h < u_k + \delta\,h \le \max\{u_k, u_l\} + \delta\,h$.

The implications for $g_i \ge 0$ follow analogously. □

Theorem 5 shows that if the discrete maximum principle is violated then the discrete solution is locally near to a constant function provided that δ or h are sufficiently small. Globally, the violation of the discrete maximum principle is smaller or equal to $\delta\,h$.

Remark 3. *Using a similar construction as above, one could modify the definition* (24) *in such a way that the resulting function β_i is continuous and equals 1 whenever the discrete solution attains an extremum at the node x_i. Then the statements of Theorem 5 hold with $\delta = 0$. However, the resulting method then adds artificial diffusion of magnitude $\tilde{\varepsilon}$ in regions where the discrete solution is constant, which is not desirable. Moreover, due to rounding errors, an approximation of a constant solution u typically possesses a lot of negligible extrema that also should not lead to adding a significant amount of artificial diffusion. The continuous function β_i defined at the beginning of this section satisfies this requirement.*

Now let us report a few numerical results for β_i defined by (58). We used $\Delta = D = 0.5$ so that $\delta = 1$. For decreasing δ, we encountered increasing difficulties with the solution of the nonlinear problem, whereas the resulting approximate solution was not affected significantly. We again applied the fixed-point iteration described at the beginning of Section 6 that was terminated if absolute values of all components of the residual vector were smaller than $5 \cdot 10^{-14}$.

First, we repeated the computations of Section 6 and realized that all results are very similar for the continuous β_i, at least for $Pe \geq 1$ (for $Pe < 1$, a difference stems from using $L = 0.5$ instead of $L = 1$, cf. the end of Section 4). Then, we considered the counterexamples from Section 8 for which the discrete problems with discontinuous β_i were not solvable. Now, solutions could be computed and we obtained the following values of $\beta_1, \ldots, \beta_4$:

$$\begin{aligned}
&\text{data (51):} \quad \beta_1 = 1\,, \quad \beta_2 = 1\,, \quad \beta_3 = 1\,, \quad \beta_4 = 0.041172246777\,,\\
&\text{data (52):} \quad \beta_1 = 0\,, \quad \beta_2 = 0\,, \quad \beta_3 = 0.016194286589\,, \quad \beta_4 = 1\,,\\
&\text{data (53):} \quad \beta_1 = 0\,, \quad \beta_2 = 0\,, \quad \beta_3 = 0.018436266748\,, \quad \beta_4 = 1\,.
\end{aligned}$$

Finally, we investigated numerically a possible violation of the discrete maximum principle by the method (17), (18) if β_i are defined by (58). We used $\tilde{\varepsilon}$ from (36) and considered the problem (9) with the data

$$b = 1\,, \qquad g = 0\,, \qquad u_L = 1\,, \qquad u_R = 0\,, \tag{62}$$

and various values of $\varepsilon > 0$. According to (10), the exact solution of this problem is a decreasing function with values in the interval $[0, 1]$. For small ε, the solution is nearly constant except for a small neighborhood of the right boundary point. Therefore, this problem is suitable for testing the validity of the discrete maximum principle by comparing the maximum value of the approximate solution

$$u_h^{\max} = \max_{i=0,\ldots,n+1} u_i$$

with the value 1. We used several values of ε and, for each of them, we computed approximate solutions for all values of $h \equiv 1/(n+1) \leq 0.25$ leading to $Pe \geq 1$. It turns out that it is reasonable to consider separately moderate Péclet numbers and large Péclet numbers. More precisely, we separately considered $Pe \in [1, 20)$ and $Pe \in [20, \infty)$. We denote by MAX the maximum of $u_h^{\max} - 1$ over all h for which the Péclet number belongs to the respective interval, by $RMAX$ the maximum of $(u_h^{\max} - 1)/h$ and by Pe_{RMAX} the value of Pe for which the maximum $RMAX$ is attained. The results are summarized in Table 1. We observe that the results are in agreement with Theorem 5 and that the largest violations of the discrete maximum principle appear for small Péclet numbers, i.e., when the mesh width approaches the thickness of the boundary layer. The numerical results also suggest that the violation of the discrete maximum principle is bounded by $0.2 \min\{h, \varepsilon \ln(1/\varepsilon)\}$ and is often significantly smaller so that it is negligible in the most cases. The results presented for $Pe \in [20, \infty)$ are influenced by rounding errors and hence differ from values that would be obtained in exact arithmetic.

Table 1: Violation of the discrete maximum principle for the data (62) and continuous β_i given by (58).

	$Pe \in [1, 20)$			$Pe \in [20, \infty)$		
ε	MAX	$RMAX$	Pe_{RMAX}	MAX	$RMAX$	Pe_{RMAX}
10^{-1}	6.62−3	2.65−2	1.25	no $Pe \geq 20$		
10^{-2}	3.55−3	9.27−2	1.85	no $Pe \geq 20$		
10^{-3}	7.14−4	1.28−1	2.79	4.88−15	4.88−14	25.0
10^{-4}	1.06−4	1.40−1	3.77	5.60−14	9.23−13	21.6
10^{-5}	1.41−5	1.47−1	4.80	4.81−13	5.59−10	21.6
10^{-6}	1.77−6	1.51−1	5.84	6.06−12	6.92−8	22.9

10. Conclusions and outlook

An algebraic flux correction scheme of TVD-type, generalizing the one proposed in [15], was studied in this work for 1D steady-state convection–diffusion equations. The discrete operator was reformulated as a nonlinear finite difference operator with a parameter vector. Possible choices of this parameter vector were studied numerically. A fixed point iteration was used for solving the nonlinear problem. The main results of this work are about properties of the nonlinear problem and the linear subproblems (discrete maximum principle, solvability). The unique solvability of the linear subproblems was studied under rather general conditions on the parameter vector of the scheme. Counterexamples concerning the existence of a solution of the nonlinear problem were provided. A modification of the scheme was proposed for which the existence of a solution and a weak variant of the discrete maximum principle were proved. Numerical experiments suggested that a good choice of the maximum artificial diffusion is $\tilde{\varepsilon} = b\,h\,(\coth Pe - 1/Pe)/2$. Then the modified nonlinear scheme is solvable and, in all numerical experiments, the approximate solutions were not smeared and the violation of the discrete maximum principle was negligible.

Future work will study alternative algebraic flux correction schemes proposed, e.g., in [18]. As first step, it has to be ensured that a solution of these nonlinear schemes exists. If this point is positively clarified, it makes sense to investigate the (order of) convergence to a solution. Of course, numerical analysis for multi-dimensional problems is of utmost interest. From our experience so far, we think that such an analysis should initially consider model problems, simple domains, and structured grids.

Acknowledgments

The authors are gratefully indebted to Professor Dmitri Kuzmin for many fruitful discussions on algebraic flux correction schemes. The work of G.R. Barrenechea has been partially funded by The Leverhulme Trust, through the Research Project Grant RPG-2012-483. The work of P. Knobloch has been partially supported through the grant No.13-00522S of the Czech Science Foundation.

References

[1] Paul Arminjon and Alain Dervieux. Construction of TVD-like artificial viscosities on two-dimensional arbitrary FEM grids. J. Comput. Phys., 106(1):176–198, 1993.

[2] Matthias Augustin, Alfonso Caiazzo, André Fiebach, Jürgen Fuhrmann, Volker John, Alexander Linke, and Rudolf Umla. An assessment of discretizations for convection-dominated convection–diffusion equations. Comput. Methods Appl. Mech. Engrg., 200(47-48):3395–3409, 2011.

[3] Markus Bause and Kristina Schwegler. Analysis of stabilized higher-order finite element approximation of nonstationary and nonlinear convection–diffusion–reaction equations. Comput. Methods Appl. Mech. Engrg., 209-212:184–196, 2012.

[4] Jay P. Boris and David L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. J. Comput. Phys., 11(1):38–69, 1973.

[5] Ian Christie, David F. Griffiths, Andrew R. Mitchell, and Olgierd C. Zienkiewicz. Finite element methods for second order differential equations with significant first derivatives. Int. J. Numer. Methods Eng., 10(6):1389–1396, 1976.

[6] Ramon Codina. Comparison of some finite element methods for solving the diffusion–convection–reaction equation. Comput. Methods Appl. Mech. Engrg., 156(1-4):185–210, 1998.

[7] Jürgen Fuhrmann and Hartmut Langmach. Stability and existence of solutions of time-implicit finite volume schemes for viscous nonlinear conservation laws. Appl. Numer. Math., 37(1-2):201–230, 2001.

[8] Volker John and Petr Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations. I. A review. Comput. Methods Appl. Mech. Engrg., 196(17-20):2197–2215, 2007.

[9] Volker John and Petr Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations. II. Analysis for P_1 and Q_1 finite elements. Comput. Methods Appl. Mech. Engrg., 197(21-24):1997–2014, 2008.

[10] Volker John, Teodora Mitkova, Michael Roland, Kai Sundmacher, Lutz Tobiska, and Andreas Voigt. Simulations of population balance systems with one internal coordinate using finite element methods. Chemical Engineering Science, 64(4):733–741, 2009.

[11] Volker John and Julia Novo. On (essentially) non-oscillatory discretizations of evolutionary convection–diffusion equations. J. Comput. Phys., 231(4):1570–1586, 2012.

[12] Volker John and Ellen Schmeyer. Finite element methods for time-dependent convection–diffusion–reaction equations with small diffusion. Comput. Methods Appl. Mech. Engrg., 198(3-4):475–494, 2008.

[13] Volker John and Liesel Schumacher. A study of isogeometric analysis for scalar convection–diffusion equations. Appl. Math. Lett., 27(1):43–48, 2014.

[14] Dmitri Kuzmin. On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection. J. Comput. Phys., 219(2):513–531, 2006.

[15] Dmitri Kuzmin. Algebraic flux correction for finite element discretizations of coupled systems. In Manolis Papadrakakis, Eugenio Oñate, and Bernard Schrefler, editors, Proceedings of the Int. Conf. on Computational Methods for Coupled Problems in Science and Engineering, pages 1–5. CIMNE, Barcelona, 2007.

[16] Dmitri Kuzmin. On the design of algebraic flux correction schemes for quadratic finite elements. J. Comput. Appl. Math., 218(1):79–87, 2008.

[17] Dmitri Kuzmin. Explicit and implicit FEM-FCT algorithms with flux linearization. J. Comput. Phys., 228(7):2517–2534, 2009.

[18] Dmitri Kuzmin. Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. J. Comput. Appl. Math., 236(9):2317–2337, 2012.

[19] Dmitri Kuzmin and Matthias Möller. Algebraic flux correction I. Scalar conservation laws. In Dmitri Kuzmin, Rainald Löhner, and Stefan Turek, editors, Flux-Corrected Transport. Principles, Algorithms, and Applications, pages 155–206. Springer-Verlag, Berlin, 2005.

[20] Dmitri Kuzmin and Stefan Turek. High-resolution FEM-TVD schemes based on a fully multidimensional flux limiter. J. Comput. Phys., 198(1):131–158, 2004.

[21] Rainald Löhner, Ken Morgan, Jaime Peraire, and Mehdi Vahdati. Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier-Stokes equations. Int. J. Numer. Methods Fluids, 7(10):1093–1109, 1987.

[22] Hans-Görg Roos, Martin Stynes, and Lutz Tobiska. Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion–Reaction and Flow Problems. 2nd ed. Springer-Verlag, Berlin, 2008.

[23] Steven T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. J. Comput. Phys., 31(3):335–362, 1979.

Response to the reports on the manuscript

Gabriel R. Barrenechea, Volker John, Petr Knobloch: Some analytical results for an algebraic flux correction scheme for a steady convection–diffusion equation in 1D

In their reports the referres raised four remarks which are repeated below together with a description of the changes made in the paper.

Report 1

1. *The only unclear point is the derivation of the form (15) for the numerical scheme, performed in page 7. The deduction of the expressions for coefficients $\alpha_{i,i-1}$ and $\alpha_{i,i+1}$ is based upon the definition of P_i^+, P_i^-, etc (end of page 4) which is not understandable. The rest of the deductions until eq. (15) is then not justified.*
   The last paragraph of Section 2 (where P_i^+, P_i^-, etc are defined) was written in a more understandable way. Moreover, the computation of P_i^+, P_i^-, Q_i^+, Q_i^- at the beginning of page 7 was also explained more clearly.

Report 2

1. *p4, -l1: It seems, that the values for Q_j^+ and Q_j^- are not used. Moreover, is there some initial value for these iterations missing?*
   This was clarified by the changes in the last paragraph of Section 2.

2. *p14, Fig. 1: Please provide labels for the axes.*
   The labels were added and the caption of Fig. 1 was changed.

3. *p26, Chap. 10: Is there a general recommendation for the choice of the parameters in view of solvability of the non-linear problem and the maximum-principle?*
   Two sentences concerning the choice of the parameter $\tilde{\varepsilon}$ were added at the end of the first paragraph of Section 10.