J Sci Comput
DOI 10.1007/s10915-016-0285-7

The Effect of the Consistent Mass Matrix on the Maximum-Principle for Scalar Conservation Equations

Jean-Luc Guermond[1] · **Bojan Popov**[1] · **Yong Yang**[2]

Received: 26 April 2016 / Revised: 17 August 2016 / Accepted: 6 September 2016

Abstract In this paper we study the effect of the use of the consistent mass matrix when solving scalar nonlinear conservation equations. It is shown that a continuous finite element method based on artificial viscosity in space and explicit time stepping using the consistent mass matrix cannot satisfy the maximum principle.

Keywords Consistent mass matrix · Maximum principle · Mass lumping · Nonlinear conservation equations · Finite element method

Mathematics Subject Classfication 65M60 · 35L65

1 Introduction

It is a well established fact that the unique entropy solution of any scalar conservation equation satisfies the maximum principle (see, e.g., Theorem 6.9.3 in Dafermos [5]). It is often desirable to reproduce this property at the discrete level to avoid nonphysical over- or under-shoots. Many finite volume methods using piecewise constant or piecewise linear approximation in space preserve the maximum principle (see e.g., Lax [15], Leer [16], Osher [19], and Nessyahu and Tadmor [18]). We refer to Mehmetoglu and Popov [17] for one of the very rare proofs of convergence for a class of second-order maximum-principle preserving finite

This material is based upon work supported in part by the National Science Foundation Grants DMS-1217262, by the Air Force Office of Scientific Research, USAF, under Grant/Contract Number FA9550-15-1-0257, and by the Army Research Office under Grant/Contract Number W911NF-15-1-0517. Draft version, September 7, 2016.

✉ Jean-Luc Guermond
guermond@math.tamu.edu

[1] Department of Mathematics, Texas A& M University, 3368 TAMU, College Station, TX 77843, USA

[2] Department of Mathematics, Penn State University, University Park, State College, PA 16802, USA

Published online: 17 September 2016

volume methods. There are also many maximum-principle preserving finite element methods based on the discontinuous Galerkin approximation; we refer for instance to Zhang and Shu [21] for a review. But, in this paper we focus our interest exclusively on continuous Galerkin methods. The type of methods we are interested in are for instance the so-called local extrema diminishing (LED) schemes developed in Jameson [11, §2.1], Kuzmin and Turek [14, Eq. (32)–(33)], Kuzmin et al. [13, p. 163], where the stabilization is based on an algebraic point of view, or other approaches where the stabilization is based on either a nonlinear or a linear artificial viscosity like in Burman [2], Burman and Ern [3], Badia and Hierro [1], Guermond and Nazarov [8], Guermond and Popov [10]. We note in passing that the method proposed in [10] is very general and extends naturally to any hyperbolic systems.

A common factor of all the continuous finite element methods referred to above is that they all assume that the mass matrix is lumped. This is an important deficiency, at least for piecewise linear approximation, since it is well-known that lumping the mass matrix induces dispersion errors that have adverse effects when solving transport-like equations with non-smooth initial data, see e.g., Christon et al. [4], Gresho et al. [7], Guermond and Pasquetti [9]. The objective of the present paper is to investigate whether it is possible to construct an explicit maximum-principle preserving method using the consistent mass matrix where the stabilization is induced by artificial viscosity in space. We are going to restrict the analysis to one space dimension and continuous piecewise linear finite elements. The conclusion, and main result of this paper, is that even in this very simplistic setting, it is impossible to construct an artificial viscosity that makes the method maximum-principle preserving, (whether the explicit artificial viscosity is linear or nonlinear). This will be demonstrated both for the Cauchy problem and for the periodic boundary value problem.

This paper is organized as follows. We introduce in Sect. 2 some definitions, the notation, the model problem and the finite element method that is used to solve the model problem. The Cauchy problem and the periodic boundary value problem are considered in Sects. 3 and 4, respectively. The main results of the paper are stated in Theorems 3.2 and 4.3.

2 Preliminaries

In this section, we first formulate the problem, then we introduce the finite element setting.

2.1 Model Problem

Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a non-constant Lipschitz function such that $f(0) = 0$. Consider the following nonlinear conservation equation in one space dimension: Given an open interval Ω in $\mathbb{R}$ and some initial data in $L^\infty(\Omega)$, find $u \in L^\infty(\Omega\times(0,\infty))$ such that

$$\begin{cases} \partial_t u(x,t) + \partial_x f(u(x,t)) = 0, & \text{a.e. } (x,t) \text{ in } \Omega\times(0,\infty), \\ u(x,0) = u^0(x), & \text{a.e. } x \text{ in } \Omega \end{cases} \tag{2.1}$$

where the PDE is understood in the weak sense. When $\Omega = \mathbb{R}$ and $f(u) = \beta u$ with $\beta \in \mathbb{R}$, the method of characteristics gives $u(t,x) = u^0(x - \beta t)$. The same formula holds in the periodic case upon replacing u^0 by its periodic extension. For a general flux f it is known that this problem has a unique entropy solution; i.e., a weak solution that additionally satisfies the entropy inequalities $\partial_t E(u) + \partial_x F(u) \le 0$ for all convex entropies $E \in \text{Lip}(\mathbb{R};\mathbb{R})$ and associated entropy fluxes $F(u) = \int_0^u E'(v) f'(v)\,\mathrm{d}v$. Moreover, this solution satisfies the maximum principle, i.e., ess $\inf_{x\in\Omega} u^0(x) \le u(x,t) \le$ ess $\sup_{x\in\Omega} u^0(x)$ for every $t > 0$ and a.e. x in Ω. We refer to Kružkov [12] for more details.

In the rest of the paper we are going to assume that either $\Omega = \mathbb{R}$ or Ω is a bounded interval and periodic boundary conditions are imposed. We will refer to the first case as the Cauchy problem and to the second case as the periodic boundary value problem. For the Cauchy problem we assume that there exists $a < b$ and $u_a^0, u_b^0 \in \mathbb{R}$ such that $u^0(x) = u_a^0$ for all $x < a$ and $u^0(x) = u_b^0$ for all $x > b$.

2.2 Finite Element Approximation

Let $N \in \mathbb{N}$ and let $\{x_i\}_{i\in\{0:N\}}$ be a sequence of equidistributed points in Ω which we henceforth call Lagrange nodes. We denote $I_i := [x_i, x_{i+1}]$, $h := |I_i|$ and $\Omega_{\text{comp}} := \text{int}(\bigcup_{i=0}^{N-1} I_i)$. For the Cauchy problem we assume that the interval (a,b) defined above is such that $(a,b) \subset \Omega_{\text{comp}} \subsetneq \mathbb{R}$, and for the periodic boundary value problem we assume that $\Omega_{\text{comp}} = \Omega$. We define the mesh $\mathcal{T}_h := \{I_i\}_{i\in\{0:N-1\}}$. Let $\widehat{I} := [0,1]$ be the reference element, we denote by $T_{I_i} : \widehat{I} \longrightarrow I_i$ the affine geometric transformation such that $T_{I_i}(\widehat{x}) = x_i(1-\widehat{x}) + \widehat{x}x_{i+1}$.

We solve problem (2.1) by using the C^0 finite elements. For the Cauchy problem we define the discrete space

$$V_h := \{v_h \in \mathcal{C}^0(\overline{\Omega}_{\text{comp}}) \mid v_h \circ T_I \in \mathbb{P}_1,\ \forall I \in \mathcal{T}_h\}. \tag{2.2}$$

For the periodic boundary value problem we will use

$$V_h^{\text{per}} = \{v_h \in \mathcal{C}^0(\overline{\Omega}) \mid v_h \circ T_I \in \mathbb{P}_1,\ \forall I \in \mathcal{T}_h;\ v_h(x_0) = v_h(x_N)\}. \tag{2.3}$$

Note that the periodic boundary condition is built into V_h^{per}. Let $\{\varphi_0, \dots, \varphi_N\}$ be the Lagrange nodal basis functions associated with the Lagrange nodes of the mesh $\mathcal{T}_h$, i.e., $\varphi_i(x_j) = \delta_{ij}$. It follows that

$$\begin{aligned} V_h &= \text{span}\{\varphi_0, \dots, \varphi_N\}, \\ V_h^{\text{per}} &= \text{span}\{\varphi_0 + \varphi_N, \dots, \varphi_{N-1}\}. \end{aligned}$$

We also introduce a space for the artificial viscosity

$$D_h = \{v_h \in L^\infty(\Omega_{\text{comp}}) \mid v_h \circ T_I \in \mathbb{P}_0,\ \forall I \in \mathcal{T}_h\}.$$

3 Cauchy Problem

Let $\tau > 0$ be the time step. Let $u_h^0 \in V_h$ be a reasonable approximation of u^0. For instance u_h^0 can be chosen to be the L^2 projection or Lagrangian interpolation of u^0 (provided the quantities $\{u^0(x_i)\}_{i\in\{0:N\}}$ makes sense). The forward Euler finite element approximation to the Cauchy problem (2.1) is formulated as follows: For $n \ge 0$, find $u_h^{n+1} \in V_h$ such that

$$\int_\Omega \left(\frac{u_h^{n+1} - u_h^n}{\tau} + \partial_x \left(f_h(u_h^n)\right) \right) v_h \,\mathrm{d}x + \int_\Omega \nu^n(x) \partial_x u_h^n \partial_x v_h \,\mathrm{d}x = 0, \quad \forall v_h \in V_h, \tag{3.1}$$

where ν^n is any specified distribution of artificial viscosity. The discrete flux $f_h : V_h \longrightarrow V_h$ is defined by $f_h(u_h) = \sum_{i=0}^N f(U_i^n)\varphi_i$ where $u_h^n(x) := \sum_{i=0}^N U_i^n \varphi_i(x) \in V_h$. Note that this approximation is second-order accurate since it is exact if f is a linear function. The viscous term is estimated exactly as follows $\int_\Omega \nu^n(x)\partial_x u_h^n \partial_x \varphi_i \,\mathrm{d}x = \sum_j (U_i^n - U_j^n) d_{ij}^n$ where

$d_{ij}^n := -\int_\Omega \nu^n(x)\partial_x\varphi_j\partial_x\varphi_i\,\mathrm{d}x$, and we have used that $\sum_j d_{ij}^n = 0$ due to the partition of unity property. Note also that $d_{ij}^n = d_{ji}^n$. We set $d_{i+\frac{1}{2}}^n := d_{ii+1}^n = d_{i+1i}^n$ and $d_{i-\frac{1}{2}}^n := d_{ii-1}^n = d_{i-1i}^n$, and we denote by $d^n \in D_h$ the piece-wise constant function such that $d_{|I_i}^n = d_{i+\frac{1}{2}}^n$. Then upon setting $\lambda := \frac{\tau}{h}$, the discrete formulation of the finite element method (3.1) can be written as

$$\begin{aligned}\frac{U_{i+1}^{n+1}+4U_i^{n+1}+U_{i-1}^{n+1}}{6} = &\frac{U_{i+1}^n+4U_i^n+U_{i-1}^n}{6}\\ &+\frac{\lambda}{2}(f(U_{i-1}^n)-f(U_{i+1}^n))\\ &+\lambda d_{i-\frac{1}{2}}^n(U_{i-1}^n-U_i^n)+\lambda d_{i+\frac{1}{2}}^n(U_{i+1}^n-U_i^n), \end{aligned}\tag{3.2}$$

for $i \in \{1{:}N-1\}$. For $i = 0, N$, we can choose $U_0^{n+1} = U_0^n = u_a^0$, $U_N^{n+1} = U_N^n = u_b^0$. Since f is not constant and $f(0) = 0$ by assumption, there is $\gamma \in \mathbb{R}$ such that $f(\gamma) \neq 0$. Let us now define $L := \lfloor \frac{N}{2} \rfloor$, where $\lfloor\cdot\rfloor$ is the floor function, and let us choose the special initial data u_h^0 satisfying that

$$U_i^0 = \begin{cases}\gamma, & \text{if}\, i \in \{0{:}L\}\\ 0, & \text{if}\, i \in \{L+1{:}N\}.\end{cases}\tag{3.3}$$

The following result shows that the maximum principle is violated in (3.2).

Lemma 3.1 *Let u_h^0 as defined in (3.3). Then*

$$|\gamma| = \max_i\{U_i^0\} - \min_i\{U_i^0\} < \max_i\{U_i^1\} - \min_i\{U_i^1\},\tag{3.4}$$

for every $d^0 \in D_h$ and every $\lambda > 0$, i.e., the solution u_h^1 of (3.1) at $t^1 := \tau$ violates the maximum principle.

Proof (1) For simplicity, we shift the indices of the Lagrange nodes and shape functions to make the range of indices be $\{-L, \cdots, N-L\}$. Then the initial data (3.3) is rewritten

$$U_i^0 = \begin{cases}\gamma, & \text{if}\, i \in \{-L{:}0\}\\ 0, & \text{if}\, i \in \{1{:}N-L\}.\end{cases}$$

From (3.2), we obtain that u_h^1 satisfies the following equation

$$\frac{U_{i+1}^1+4U_i^1+U_{i-1}^1}{6} = \begin{cases}\gamma & \text{for } i<0\\ \frac{5\gamma}{6}+\lambda(\frac{f(\gamma)}{2}-\gamma d_{\frac{1}{2}}^0) & \text{for } i=0\\ \frac{\gamma}{6}+\lambda(\frac{f(\gamma)}{2}+\gamma d_{\frac{1}{2}}^0) & \text{for } i=1\\ 0 & \text{for } i>1.\end{cases}\tag{3.5}$$

We have a non-homogeneous linear recurrence relations with constant coefficients $U_{i+1}^1 + 4U_i^1 + U_{i-1}^1 = b_i$. The characteristic equation is $r^2 + 4r + 1 = 0$, and the two roots are $r_+ = -2+\sqrt{3}$ and $r_- = -2-\sqrt{3}$. We propose the ansatz $U_i^1 = \alpha r_+^i$ for $i \geq 1$ and $U_i^1 = \beta r_-^i + \gamma$ for $i \leq 0$ where $\alpha, \beta \in \mathbb{R}$ are yet to be determined. It is clear that these two ansätze satisfy (3.5) for all $i \notin \{0, 1\}$. We are going to compute α and β by requesting that the ansätze also satisfy (3.5) for $i = 0$ and $i = 1$:

$$\begin{cases} \frac{U_1^1+4U_0^1+U_{-1}^1}{6} = \frac{5\gamma}{6} + \lambda\left(\frac{f(\gamma)}{2} - \gamma d^0_{\frac{1}{2}}\right) \\ \frac{U_2^1+4U_1^1+U_0^1}{6} = \frac{\gamma}{6} + \lambda\left(\frac{f(\gamma)}{2} + \gamma d^0_{\frac{1}{2}}\right). \end{cases} \tag{3.6}$$

Inserting the ansätze in the above two equations, we have

$$\begin{cases} \alpha r_+ + 4(\beta+\gamma) + (\beta r_-^{-1} + \gamma) = 5\gamma + 6\lambda(\frac{f(\gamma)}{2} - \gamma d^0_{\frac{1}{2}}) \\ \alpha r_+^2 + 4\alpha r_+ + (\beta+\gamma) = \gamma + 6\lambda(\frac{f(\gamma)}{2} + \gamma d^0_{\frac{1}{2}}), \end{cases}$$

i.e.,

$$\begin{cases} \alpha(-2+\sqrt{3}) + \beta(2+\sqrt{3}) = 6\lambda(\frac{f(\gamma)}{2} - \gamma d^0_{\frac{1}{2}}) \\ -\alpha + \beta = 6\lambda(\frac{f(\gamma)}{2} + \gamma d^0_{\frac{1}{2}}). \end{cases}$$

Solving these equations, we obtain the solution u_h^1 as follows

$$U_i^1 = \gamma \begin{cases} (-\sqrt{3}+3)\lambda[\frac{f(\gamma)}{2\gamma} - \sqrt{3}d^0_{\frac{1}{2}}]r_-^i + 1 & \text{for } i \le 0 \\ -(\sqrt{3}+3)\lambda[\frac{f(\gamma)}{2\gamma} + \sqrt{3}d^0_{\frac{1}{2}}]r_+^i & \text{for } i \ge 1. \end{cases} \tag{3.7}$$

Note that it is not possible that both $\frac{f(\gamma)}{2\gamma} + \sqrt{3}d^0_{\frac{1}{2}}$ and $\frac{f(\gamma)}{2\gamma} - \sqrt{3}d^0_{\frac{1}{2}}$ be zero since $f(\gamma) \neq 0$. Note also both $r_- < 0$ and $r_+ < 0$; hence depending on the parity of i the factor $r_\pm^i$ is either positive or negative.
(2) Assume first that either $\frac{f(\gamma)}{2\gamma} + \sqrt{3}d^0_{\frac{1}{2}} = 0$ or $\frac{f(\gamma)}{2\gamma} - \sqrt{3}d^0_{\frac{1}{2}} = 0$ (the "or" is exclusive, i.e., either $\frac{f(\gamma)}{2\gamma} + \sqrt{3}d^0_{\frac{1}{2}} \neq 0$ or $\frac{f(\gamma)}{2\gamma} - \sqrt{3}d^0_{\frac{1}{2}} \neq 0$ since $f(\gamma) \neq 0$). Assume now that $\gamma > 0$. Then either there exists $i_0 \le 0$ such that $U_{i_0}^1 > \gamma$ and $U_i^1 = 0$ for all $i \ge 1$ or there exists $i_0 \ge 1$ such that $U_{i_0}^1 < 0$ and $U_i^1 = \gamma$ for all $i \le 0$. In both cases $\max_i\{U_i^1\} - \min_i\{U_i^1\} > \gamma$. The same argument holds if $\gamma < 0$ and in this case we have $\max_i\{U_i^1\} - \min_i\{U_i^1\} > |\gamma|$.
(3) Assume that $\frac{f(\gamma)}{2\gamma} + \sqrt{3}d^0_{\frac{1}{2}} \neq 0$ and $\frac{f(\gamma)}{2\gamma} - \sqrt{3}d^0_{\frac{1}{2}} \neq 0$. Assume now that $\gamma > 0$. Then there is $i_0 \le 0$ such that $U_{i_0}^1 > \gamma$ and there $j_0 \ge 1$ such that $U_{j_0}^1 < 0$, i.e., $\max_i\{U_i^1\} - \min_i\{U_i^1\} > \gamma$. The same argument holds if $\gamma < 0$ and in this case we have $\max_i\{U_i^1\} - \min_i\{U_i^1\} > |\gamma|$. This concludes the proof. □

As an immediate consequence of the above lemma we derive the main result of this section.

Theorem 3.2 *If the consistent mass matrix is used in* (3.1)*, then for every nonzero flux f there exists $u_h^0 \in V_h$ such that the solution u_h of* (3.1) *violates the maximum principle at the first time step for every $\lambda > 0$ and every artificial viscosity distribution $d^0 \in D_h$.*

Remark 3.1 (Viscosity distribution ν^0) Note that since $d^0 \in D_h$ is arbitrary, the distribution of viscosity ν^0 in (3.1) can be arbitrary as well. In particular ν^0 could be any nonlinear function of u_h^0. In conclusion there is not hope to recover the maximum principle by using an explicit nonlinear viscosity when the consistent mass matrix is used. This result is somewhat in agreement with Theorem 2.1 in Thomée and Wahlbin [20] where the authors study the semi-discrete finite element approximation of a general linear parabolic equation.

4 Periodic Boundary Value Problem

We now investigate the periodic boundary value problem. Let $\tau > 0$ be the time step and let $u_h^0 \in V_h^{\mathsf{per}}$ be a reasonable approximation of u^0. For all $n \geq 0$, the forward Euler finite element approximation to the periodic problem (2.1) consists of finding $u_h^{n+1} \in V_h^{\mathsf{per}}$, i.e., $u_h^{n+1}(x) = \sum_{i=0}^{N-1} U_i^{n+1}\varphi_i(x) + U_0^{n+1}\varphi_N(x)$, such that

$$\int_\Omega \left(\frac{u_h^{n+1} - u_h^n}{\tau} + \partial_x \left(f_h(u_h^n) \right) \right) v_h \,\mathrm{d}x + \int_\Omega \nu^n(x) \partial_x u_h^n \partial_x v_h \,\mathrm{d}x = 0, \quad \forall v_h \in V_h^{\mathsf{per}}, \tag{4.1}$$

where ν^n is the distribution of the artificial viscosity. Upon setting $U^n := [U_0^n, \cdots, U_{N-1}^n]^\mathsf{T}$, the algebraic form of (4.1) is as follows:

$$M(U^{n+1} - U^n) = \lambda F^n \tag{4.2}$$

where the consistent mass matrix M is a $N \times N$ circulant matrix:

$$M = \frac{1}{6}\mathrm{circ}[4, 1, 0, \cdots, 0, 1], \tag{4.3}$$

$F^n := [F_0^n, \cdots, F_{N-1}^n]^\mathsf{T}$, and $F_i^n = -\lambda[\frac{f(U_{i+1}^n) - f(U_{i-1}^n)}{2} + d_{i+\frac{1}{2}}^n (U_i^n - U_{i+1}^n) + d_{i-\frac{1}{2}}^n (U_i^n - U_{i-1}^n)]$ with the convention that $U_{-1}^n = U_{N-1}^n$, $U_N^n = U_0^n$, $d_{-\frac{1}{2}}^n = U_{N-\frac{1}{2}}^n$. As a Gram matrix, M is invertible. Since M is a circulant matrix, M^{-1} is also a circulant matrix. In fact, since the mesh is uniform, the inverse M^{-1} can be written out explicitly as stated in the next Lemma.

Lemma 4.1 *Assume* $N > 3$. *Then* $M^{-1} = \mathrm{circ}[a_0, a_1, \cdots, a_{N-1}]$, *where* $a_j := \sqrt{3}\frac{r_+^j + r_+^{N-j}}{1 - r_+^N}$, $j \in \{0{:}N-1\}$, *where* $r_+ := -2 + \sqrt{3}$.

Proof Assume $M^{-1} = \mathrm{circ}[a_0, a_1, \cdots, a_{N-1}]$. The condition $MM^{-1} = I$ gives

$$\begin{cases} 4a_0 + a_1 + a_{N-1} = 6, \\ a_{i-1} + 4a_i + a_{i+1} = 0, \quad i = 1, \ldots, N-1 \end{cases} \tag{4.4}$$

with the convention that $a_N = a_0$. The solution to the $(N-1)$ linear recurrence relations with constant coefficients $a_{i-1} + 4a_i + a_{i+1} = 0$ for $i = 1, \ldots, N-1$ is $a_i = Ar_+^i + Br_-^i$ where we recall that $r_\pm := -2 \pm \sqrt{3}$ are the roots of the characteristic equation $r^2 + 4r + 1 = 0$. Then we use the other two equations $4a_0 + a_1 + a_{N-1} = 6$ and $a_N = a_0$ to find the coefficients A and B:

$$\begin{cases} 4A + 4B + Ar_+ + Br_- + Ar_+^{N-1} + Br_-^{N-1} = 6, \\ Ar_+^N + Br_-^N = A + B. \end{cases} \tag{4.5}$$

That is, the pair (A, B) solves the following linear system

$$\begin{bmatrix} r_+^{N-1} + 4 + r_+ & r_-^{N-1} + 4 + r_- \\ r_+^N - 1 & r_-^N - 1 \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \end{bmatrix}.$$

Using Vieta's formulas, i.e., $r_+ + r_- = -4$ and $r_+ r_- = 1$, the system can be rewritten as follows:

$$\begin{bmatrix} r_- & r_+ \\ 1 & 1 \end{bmatrix} \begin{bmatrix} (r_+^N - 1)A \\ (r_-^N - 1)B \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \end{bmatrix}.$$

Solving the above system, we get that $A = \frac{\sqrt{3}}{1-r_+^N}$ and $B = \frac{-\sqrt{3}}{1-r_-^N}$ which proves that $a_j := \sqrt{3}(\frac{r_+^j}{1-r_+^N} - \frac{r_-^j}{1-r_-^N})$. Eliminating r_-, we simplify the expression as follows

$$a_j = \sqrt{3}(\frac{r_+^j}{1-r_+^N} - \frac{r_-^j}{r_+^N r_-^N - r_-^N}) = \sqrt{3}(\frac{r_+^j}{1-r_+^N} + \frac{r_-^{j-N}}{1-r_+^N}) = \sqrt{3}\frac{r_+^j + r_+^{N-j}}{1-r_+^N}. \quad (4.6)$$

This completes the proof. □

Remark 4.1 Note that M is a banded Toeplitz matrix with modified corner terms. Hence, M^{-1} can also be obtained by using the technique described in Dow [6, §4] together with Theorem 1, page E199, therein.

Lemma 4.2 *The coefficients* $\{a_j\}_{j\in\{0:N\}}$ *satisfy the following properties:* $a_j = a_{N-j}$; $\text{sgn}(a_j) = (-1)^j$ *for* $2j \leq N$; $|a_j|$ *is decreasing for* $2j \leq N+1$.

Proof The proof follows from the expression of a_j given in Lemma 4.1. □

We are now in position to state the main result of this section.

Theorem 4.3 *Assume that* $N \geq 10$. *If the consistent mass matrix is used in* (4.1)*, then for every flux* f*, there exists* $u_h^0 \in V_h$ *such that the solution* u_h *of* (4.1) *violates the maximum principle at the first time step for every* $\lambda > 0$ *and every distribution* $d^0 \in D_h$.

Proof Let $\gamma \neq 0$ be such that $f(\gamma) \neq 0$. Let $m = \lfloor \frac{N}{2} \rfloor$ and consider the following initial data

$$U_i^0 = \begin{cases} 0 & 1 \leq i \leq m-1 \\ \gamma & \text{otherwise} \end{cases}$$

For this data, the right-hand side vector in (4.2) is as follows:

$$F_0^0 = \frac{f(\gamma)}{2} - \gamma d_{\frac{1}{2}}^0, \; F_1^0 = \frac{f(\gamma)}{2} + \gamma d_{\frac{1}{2}}^0, \; F_{m-1}^0 = -\frac{f(\gamma)}{2} + \gamma d_{m-\frac{1}{2}}^0, \; F_m^0 = -\frac{f(\gamma)}{2} - \gamma d_{m-\frac{1}{2}}^0,$$

and $F_i^0 = 0$ if $i \notin \{0, 1, m-1, m\}$. To simplify the notation we henceforth set $g = \frac{f(\gamma)}{2\gamma}$, $d_0 = d_{\frac{1}{2}}^0$, $d_m = d_{m-\frac{1}{2}}^0$. After left-multiplying the linear system (4.2) by M^{-1}, we obtain the following equations for lines $N-1, 0, 1, 2$:

$$\begin{aligned}
\tfrac{1}{\lambda\gamma}(U_{N-1}^1 - \gamma) &= a_1(g-d_0) + a_2(g+d_0) + a_m(-g+d_m) + a_{m+1}(-g-d_m)\\
\tfrac{1}{\lambda\gamma}(U_0^1 - \gamma) &= a_0(g-d_0) + a_1(g+d_0) + a_{m-1}(-g+d_m) + a_m(-g-d_m)\\
\tfrac{1}{\lambda\gamma}(U_1^1 - 0) &= a_1(g-d_0) + a_0(g+d_0) + a_{m-2}(-g+d_m) + a_{m-1}(-g-d_m)\\
\tfrac{1}{\lambda\gamma}(U_2^1 - 0) &= a_2(g-d_0) + a_1(g+d_0) + a_{m-3}(-g+d_m) + a_{m-2}(-g-d_m),
\end{aligned}$$

where we have used that $a_{N-k} = a_k$ and $m \geq 5$ (since $N \geq 10$). Similarly we obtain the following equations for lines $m-2, m-1, m, m+1$:

$$\begin{aligned}
\tfrac{1}{\lambda\gamma}(U_{m-2}^1 - 0) &= a_{m-2}(g-d_0) + a_{m-3}(g+d_0) + a_1(-g+d_m) + a_2(-g-d_m)\\
\tfrac{1}{\lambda\gamma}(U_{m-1}^1 - 0) &= a_{m-1}(g-d_0) + a_{m-2}(g+d_0) + a_0(-g+d_m) + a_1(-g-d_m)\\
\tfrac{1}{\lambda\gamma}(U_m^1 - \gamma) &= a_m(g-d_0) + a_{m-1}(g+d_0) + a_1(-g+d_m) + a_0(-g-d_m)\\
\tfrac{1}{\lambda\gamma}(U_{m+1}^1 - \gamma) &= a_{m+1}(g-d_0) + a_m(g+d_0) + a_2(-g+d_m) + a_1(-g-d_m).
\end{aligned}$$

Let us assume that the artificial viscosity d is such that the maximum principle holds. Then $\frac{U_k^1-\gamma}{\gamma} \leq 0$ and $\frac{U_k^1-0}{\gamma} \geq 0$ for any $k \in \{0{:}N-1\}$. As a result, the above equations give the following two sets of inequalities:

$$d_0(a_2-a_1)+d_m(a_m-a_{m+1})+g(a_1+a_2-a_m-a_{m+1}) \leq 0 \tag{4.7a}$$
$$d_0(a_1-a_0)+d_m(a_{m-1}-a_m)+g(a_0+a_1-a_{m-1}-a_m) \leq 0 \tag{4.7b}$$
$$d_0(a_0-a_1)+d_m(a_{m-2}-a_{m-1})+g(a_1+a_0-a_{m-2}-a_{m-1}) \geq 0 \tag{4.7c}$$
$$d_0(a_1-a_2)+d_m(a_{m-3}-a_{m-2})+g(a_2+a_1-a_{m-3}-a_{m-2}) \geq 0. \tag{4.7d}$$

and

$$d_0(a_{m-3}-a_{m-2})+d_m(a_1-a_2)+g(-a_1-a_2+a_{m-3}+a_{m-2}) \geq 0 \tag{4.8a}$$
$$d_0(a_{m-2}-a_{m-1})+d_m(a_0-a_1)+g(-a_0-a_1+a_{m-2}+a_{m-1}) \geq 0 \tag{4.8b}$$
$$d_0(a_{m-1}-a_m)+d_m(a_1-a_0)+g(-a_1-a_0+a_{m-1}+a_m) \leq 0 \tag{4.8c}$$
$$d_0(a_m-a_{m+1})+d_m(a_2-a_1)+g(-a_2-a_1+a_m+a_{m+1}) \leq 0. \tag{4.8d}$$

After adding (4.7a) and (4.8d) and adding (4.7b) and (4.8c), we obtain

$$(d_0+d_m)(a_2-a_1+a_m-a_{m+1}) \leq 0, \quad \text{and} \quad (d_0+d_m)(a_1-a_0+a_{m-1}-a_m) \leq 0.$$

A direct computation shows that $|a_m-a_{m+1}| < a_2-a_1$ (since $m \geq 5$); hence $d_0+d_m \leq 0$. The same argument implies that $|a_m-a_{m-1}| < a_0-a_1$; hence $d_0+d_m \geq 0$. In conclusion $d_0+d_m=0$. After substituting d_m by $-d_0$ in (4.7b), (4.8c), (4.7d) and (4.8a) we obtain

$$\begin{cases} d_0\alpha_2+g\beta_2 \leq 0 \\ d_0\alpha_2+g\beta_2 \geq 0, \end{cases} \qquad \begin{cases} d_0\alpha_4+g\beta_4 \geq 0 \\ d_0\alpha_4+g\beta_4 \leq 0, \end{cases}$$

where $\alpha_2 := a_1-a_0-a_{m-1}+a_m$, $\beta_2 := a_0+a_1-a_{m-1}-a_m$, $\alpha_4 := a_1-a_2-a_{m-3}+a_{m-2}$, $\beta_4 := a_2+a_1-a_{m-3}-a_{m-2}$. This immediately implies that

$$d_0\alpha_2+g\beta_2=0, \quad \text{and} \quad d_0\alpha_4+g\beta_4=0.$$

Let us now show that the determinant $\alpha_2\beta_4-\alpha_4\beta_2$ is not equal to zero. Essentially we want to show that the $\left(\frac{1}{2}(\frac{\beta_2}{\alpha_2}+1)\right)^{-1}$ is different from $\left(\frac{1}{2}(\frac{\beta_4}{\alpha_4}+1)\right)^{-1}$. A direct computation shows that

$$\frac{1}{\frac{1}{2}(\frac{\beta_2}{\alpha_2}+1)} = 1-\frac{a_0-a_m}{a_1-a_{m-1}}, \qquad \frac{1}{\frac{1}{2}(\frac{\beta_4}{\alpha_4}+1)} = 1-\frac{a_2-a_{m-2}}{a_1-a_{m-3}}.$$

It can be shown that $\frac{a_0-a_m}{a_1-a_{m-1}} - \frac{a_2-a_{m-2}}{a_1-a_{m-3}} > 3$; thereby proving that $\alpha_2\beta_4-\alpha_4\beta_2 \neq 0$. Hence $d_0=0$ and $g=0$, which is a contradiction since γ has been chosen so that $g := \frac{f(\gamma)}{\gamma} \neq 0$. The proof is complete. □

5 Conclusions

We have shown in this paper that it is impossible to satisfy the maximum principle for an explicit finite element method using the consistent mass matrix and artificial viscosity for stabilization in space. This statement is proved for any nonlinear scalar conservation equation in one space dimension. The approximation in space is done with piecewise linear finite elements on uniform meshes, the approximation in time is done with forward Euler, and the stabilization is done with piecewise constant artificial viscosity. For both the Cauchy problem and the periodic boundary value problem, we have proved that if the consistent mass matrix is used, there exists u_h^0 such the maximum principle is violated at the first time step for every time step τ and every distribution of artificial viscosity. Note that this result holds for any type of explicit artificial viscosity – whether the artificial viscosity is a nonlinear function of u_h^0 or not.

References

1. Badia, S., Hierro, A.: On monotonicity-preserving stabilized finite element approximations of transport problems. SIAM J. Sci. Comput. **36**(6), A2673–A2697 (2014)
2. Burman, E.: On nonlinear artificial viscosity, discrete maximum principle and hyperbolic conservation laws. BIT **47**(4), 715–733 (2007)
3. Burman, E., Ern, A.: Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence. Math. Comput. **74**(252), 1637–1652 (2005). (electronic)
4. Christon, M.A., Martinez, M.J., Voth, T.E.: Generalized Fourier analyses of the advection-diffusion equation-part I: one-dimensional domains. Int. J. Numer. Methods Fluids **45**(8), 839–887 (2004)
5. Dafermos, C.: Hyperbolic Conservation Laws in Continuum Physics. Grundlehren der mathematischen Wissenschaften. Springer, Berlin (2009)
6. Dow, M.: Explicit inverses of toeplitz and associated matrices. ANZIAM J. **44**(E), E185–E215 (2003)
7. Gresho, P., Sani, R., Engelman, M.: Incompressible Flow and the Finite Element Method: Advection-Diffusion and Isothermal Laminar Flow. Incompressible Flow & the Finite Element Method. Wiley, New York (1998)
8. Guermond, J.-L., Nazarov, M.: A maximum-principle preserving C^0 finite element method for scalar conservation equations. Comput. Methods Appl. Mech. Eng. **272**, 198–213 (2014)
9. Guermond, J.-L., Pasquetti, R.: A correction technique for the dispersive effects of mass lumping for transport problems. Comput. Methods Appl. Mech. Eng. **253**, 186–198 (2013)
10. Guermond, J.-L., Popov, B.: Invariant domains and first-order continuous finite element approximation for hyperbolic systems. SIAM J. Numer. Anal. **54**(4), 2466–2489 (2016) arXiv:1509.07461
11. Jameson A.: Positive schemes and shock modelling for compressible flows. Int. J. Numer. Methods Fluids **20**(8-9), 743–776 (1995). Finite elements in fluids—new trends and applications (Barcelona, 1993)
12. Kružkov, S.N.: First order quasilinear equations with several independent variables. Mat. Sb. **81**(123), 228–255 (1970)
13. Kuzmin, D., Löhner, R., Turek, S.: Flux–Corrected Transport. Scientific Computation. Springer, ISBN: 3-540-23730-5 (2005)
14. Kuzmin, D., Turek, S.: Flux correction tools for finite elements. J. Comput. Phys. **175**(2), 525–558 (2002)
15. Lax, P.D.: Weak solutions of nonlinear hyperbolic equations and their numerical computation. Commun. Pure Appl. Math. **7**, 159–193 (1954)
16. Leer, V.: Towards the ultimate conservative difference scheme. II. Monotonicity and conservation combined in a second-order scheme. J. Comput. Phys. **14**, 361–370 (1974)
17. Mehmetoglu, O., Popov, B.: Maximum principle and convergence of central schemes based on slope limiters. Math. Comput. **81**(277), 219–231 (2012)
18. Nessyahu, H., Tadmor, E.: Non-oscillatory central differencing for hyperbolic conservation laws. J. Comput. Phys. **87**, 408–463 (1990)
19. Osher, S.: Riemann solvers, the entropy condition, and difference approximations. SIAM J. Numer. Anal. **21**(2), 217–235 (1984)
20. Thomée, V., Wahlbin, L .B.: On the existence of maximum principles in parabolic finite element equations. Math. Comput. **77**(261), 11–19 (2008). (electronic)
21. Zhang, X., Shu, C.-W.: Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **467**(2134), 2752–2776 (2011)