Analysis of a group finite element formulation

Gabriel R. Barrenechea[a], Petr Knobloch[b,*]

[a]*Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland*
[b]*Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 18675 Praha 8, Czech Republic*

Abstract

The group finite element formulation is a strategy aimed at speeding the assembly of finite element matrices for time-dependent problems. This process modifies the Galerkin matrix of the problem in a non-consistent way. This may cause a deterioration of both the stability and convergence of the method. In this paper we prove results for a group finite element formulation of a convection–diffusion–reaction equation showing that the stability of the original discrete problem remains unchanged under appropriate conditions on the data of the problem and on the discretization parameters. A violation of these conditions may lead to non-existence of solutions, as one of our main results shows. An analysis of the consistency error introduced by the group finite element formulation and its skew-symmetric variant is given.

Keywords: group finite element formulation, existence of solutions, stability, error estimates, convection–diffusion–reaction equation

1. Introduction

The numerical solution of convection-dominated transient problems is a topic that has received much attention over the last couple of decades. If the interest is to produce discretizations that preserve properties such as positivity, then the family of flux-corrected transport (FCT) schemes [3, 16, 13, 14, 12, 10, 8] has been actively used over the past years. These methods are related to the shock-capturing idea, and thus are nonlinear, but the main advantage is that they have provided some of the best results to date (see, e.g., [6, 7] for computational surveys).

When dealing with the numerical solution of the transient transport problem (or any time-dependent problem with time-varying coefficients) by means of the finite element method (FEM), a very costly part of the computations is the assembly of the finite element matrix at every time step. This is due to the possible time dependence of the convective field. Then, in order to make the implementation more efficient, the group finite element formulation can be applied. This technique was introduced in [5, 4] to simplify the implementation of nonlinear

*Corresponding author.
Email addresses: gabriel.barrenechea@strath.ac.uk (Gabriel R. Barrenechea), knobloch@karlin.mff.cuni.cz (Petr Knobloch)

(convective) terms and to increase the efficiency of computations. Its main idea is to represent products (i.e., groups) of variables by single finite element functions. In this way, assembling the matrix corresponding to the convective term reduces to the multiplication of the nodal values of the convective field by a collection of matrix entries that are computed only once at the beginning of the computation. This formulation can be interpreted also as evaluating the convective term by using nodal quadrature. Over the years, the group finite element formulation has been frequently applied in the context of explicit piecewise linear finite element discretizations of compressible flow problems. However, the group formulation has been also used intensively in implicit FCT discretizations of conservation laws and transport and convection–diffusion problems with incompressible convection fields (see, e.g., [10, 11, 2, 9, 6]), with very satisfactory numerical results. The main focus of this work is on implicit schemes for convection–diffusion–reaction equations with divergence-free convection fields.

There is, nevertheless, a lack of theoretical exploration on the limits of the group formulation. In particular, no results seem to be available on the impact that the lack of antisymmetry of the discrete convective term has in the formulation. One particular point that, in our opinion, deserves attention is the following. The FCT-like schemes can be reinterpreted as nonlinear stabilized finite element methods, where the stabilizing term is positive semidefinite and, in particular, may vanish for some meshes and discrete solutions. Consequently, the possible stability of the whole discretization relies on the stability of the group formulation of the underlying Galerkin scheme. Thus, the impact of the modification made by the group finite element method on the Galerkin scheme needs to be studied more in detail.

The purpose of this work is to fill the gap that was described in the last paragraph. To this end we consider the convection–diffusion–reaction equation as a model problem. Our main objective is to explore what is the impact of replacing the original convective term by its group formulation, both in terms of stability and lack of consistency. Concerning the stability of the method, the situation is as follows. For the steady-state case, the ellipticity of the approximate bilinear form can be proved by supposing that the convection is small enough or the mesh is sufficiently fine. For the time-dependent case, this requirement can be overcome by supposing, in turn, that the time step is small enough, which, in practice, reduces to imposing a CFL condition. On the other hand, if the assumptions that guarantee the stability are not fulfilled, the discrete problems based on the group formulation are not solvable in general, as we demonstrate by constructing a counterexample. We then move onto the analysis of the error introduced by the group finite element formulation. Our aim in this paper is not to perform a detailed error analysis of FCT schemes for time-dependent problems, and we will thus only present results estimating the consistency error induced by the group formulation. This will, in turn, give us an insight of what sort of convergence results can be expected for the considered schemes.

The plan of this work is as follows. In Section 2 we summarize the FCT methodology and we motivate then why we require the group formulation of the Galerkin part to be stable. Then, in Section 3 we present the problem of interest, namely the transient convection–diffusion–reaction equation, and the basic formulation of the group finite element strategy. The main result of that section is the aforementioned negative result in Theorem 3.1, where we show that, without further assumptions, the discrete problem may not have a solution. Next, in Section 4 we lie down conditions on the data, the mesh and the time step to make

sure that the bilinear form associated to the discrete problem is elliptic and hence that the discrete problem is solvable. Moreover, we present an alternative skew-symmetric group formulation that is stable without any additional assumptions on the data and discretization parameters. Finally, in Section 5 we estimate the consistency errors caused by the two group formulations.

2. A flux-corrected transport scheme

Consider a linear initial-boundary value problem and let us discretize it in space by the finite element method. Then, at a time instant $t \in [0,T]$, the approximate solution can be represented by a vector $\mathrm{U}(t) \in \mathbb{R}^N$ of its coefficients with respect to a basis of the respective finite element space. Let us assume that the last $N-M$ components of $\mathrm{U}(t)$ $(0 < M < N)$ correspond to nodes where Dirichlet boundary conditions are prescribed whereas the first M components of $\mathrm{U}(t)$ are computed using the semidiscretization of the underlying partial differential equation. Then $\mathrm{U}(t) \equiv (u_1(t), \ldots, u_N(t))$ satisfies a system of linear ordinary differential equations equipped with boundary and initial conditions of the form

$$\mathbb{M}\,\frac{\mathrm{dU}}{\mathrm{d}t}(t) + \mathbb{A}(t)\,\mathrm{U}(t) = \mathrm{F}(t)\,, \qquad t \in (0,T]\,, \tag{2.1}$$

$$u_i(t) = u_i^b(t)\,, \qquad i = M+1, \ldots, N\,, \quad t \in (0,T]\,, \tag{2.2}$$

$$\mathrm{U}(0) = \mathrm{U}_0\,, \tag{2.3}$$

where $\mathbb{M} = (m_{ij})_{j=1,\ldots,N}^{i=1,\ldots,M}$ is the mass matrix and $\mathbb{A}(t) = (a_{ij}(t))_{j=1,\ldots,N}^{i=1,\ldots,M}$ is the stiffness matrix. It is assumed that the entries of the mass matrix are nonnegative. Introducing discrete time instants $0 = t_0 < t_1 < \cdots < t_K = T$ and approximating the time derivative by a difference formula, one obtains a discrete scheme for the approximations $U^n \in \mathbb{R}^N$ of $U(t^n)$. For example, the Crank–Nicholson method leads to

$$\mathbb{M}\,\frac{\mathrm{U}^n - \mathrm{U}^{n-1}}{\Delta t_n} + \frac{1}{2}\left(\mathbb{A}^n\,\mathrm{U}^n + \mathbb{A}^{n-1}\,\mathrm{U}^{n-1}\right) = \frac{1}{2}\left(\mathrm{F}^n + \mathrm{F}^{n-1}\right), \quad n = 1, \ldots, K\,, \tag{2.4}$$

$$u_i^n = u_i^b(t_n)\,, \qquad i = M+1, \ldots, N\,, \quad n = 1, \ldots, K\,, \tag{2.5}$$

$$\mathrm{U}^0 = \mathrm{U}_0\,, \tag{2.6}$$

where $\Delta t_n = t_n - t_{n-1}$, $\mathbb{A}^n = \mathbb{A}(t_n)$, and $\mathrm{F}^n = \mathrm{F}(t_n)$.

In this work we are mainly interested in solving convection-dominated problems. Then, if the semidiscrete equation (2.1) corresponds to a standard (conforming) finite element method, an additional stabilization has to be considered, see, e.g., [15]. One possibility is to apply a flux-corrected transport scheme, see, e.g., [12, 10, 8]. To formulate it, one first extends the matrices $\mathbb{A}^n$ to $(a_{ij}^n)_{i,j=1,\ldots,N}$. A common way is to use the stiffness matrices corresponding to the above-mentioned finite element discretization in the case where homogeneous natural boundary conditions are used instead of the Dirichlet ones. Then one introduces artificial diffusion matrices $\mathbb{D}^n = (d_{ij}^n)_{j=1,\ldots,N}^{i=1,\ldots,M}$ possessing the entries

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

In addition, one introduces the lumped mass matrix $\mathbb{M}_L = (m_{ij}^L)_{j=1,\dots,N}^{i=1,\dots,M}$ with the entries

$$m_{ij}^L = 0 \qquad \forall\, i \neq j\,, \qquad\qquad m_{ii}^L = \sum_{j=1}^N m_{ij}\,.$$

Denoting $\tilde{\mathbb{A}}^n := \mathbb{A}^n + \mathbb{D}^n$, (2.4) can be written in the form

$$\mathbb{M}_L\,\frac{\mathrm{U}^n - \mathrm{U}^{n-1}}{\Delta t_n} + \frac{1}{2}\left(\tilde{\mathbb{A}}^n\,\mathrm{U}^n + \tilde{\mathbb{A}}^{n-1}\,\mathrm{U}^{n-1}\right) = \frac{1}{2}\left(\mathrm{F}^n + \mathrm{F}^{n-1}\right) + \mathrm{R}^n(\mathrm{U}^n, \mathrm{U}^{n-1})$$

with

$$\mathrm{R}^n(\mathrm{U}^n, \mathrm{U}^{n-1}) = -(\mathbb{M} - \mathbb{M}_L)\,\frac{\mathrm{U}^n - \mathrm{U}^{n-1}}{\Delta t_n} + \frac{1}{2}\left(\mathbb{D}^n\,\mathrm{U}^n + \mathbb{D}^{n-1}\,\mathrm{U}^{n-1}\right)\,.$$

Note that the matrix $\tilde{\mathbb{A}}^n$ has non-positive off-diagonal entries. The matrix $\mathbb{D}^n$ has zero row sums and hence

$$(\mathbb{D}^n\,\mathrm{U})_i = \sum_{j=1}^N d_{ij}^n\,(u_j - u_i)\,, \qquad i = 1, \dots, M\,,$$

for any $\mathrm{U} = (u_1, \dots, u_N)$. Since also the matrix $\mathbb{M} - \mathbb{M}_L$ has zero row sums, one deduces that

$$(\mathrm{R}^n(\mathrm{U}^n, \mathrm{U}^{n-1}))_i = \sum_{j=1}^N r_{ij}^n\,, \qquad i = 1, \dots, M\,,$$

with fluxes

$$\begin{aligned} r_{ij}^n = &-\frac{1}{\Delta t_n}\,m_{ij}\,(u_j^n - u_i^n) + \frac{1}{\Delta t_n}\,m_{ij}\,(u_j^{n-1} - u_i^{n-1}) \\ &+ \frac{1}{2}\,d_{ij}^n\,(u_j^n - u_i^n) + \frac{1}{2}\,d_{ij}^{n-1}\,(u_j^{n-1} - u_i^{n-1})\,. \end{aligned}$$

Now the idea of the flux correction is to limit those fluxes r_{ij}^n that would otherwise cause spurious oscillations. To this end, $(\mathrm{R}^n(\mathrm{U}^n, \mathrm{U}^{n-1}))_i$ is replaced by

$$(\tilde{\mathrm{R}}^n(\mathrm{U}^n, \mathrm{U}^{n-1}))_i = \sum_{j=1}^N \alpha_{ij}^n\,r_{ij}^n$$

with solution-dependent correction factors $\alpha_{ij}^n \in [0, 1]$ satisfying

$$\alpha_{ij}^n = \alpha_{ji}^n\,, \qquad i, j = 1, \dots, M\,. \tag{2.7}$$

Then U^n satisfies

$$\frac{1}{\Delta t_n}\,\mathbb{M}\,\mathrm{U}^n + \frac{1}{2}\,\mathbb{A}^n\,\mathrm{U}^n + \mathbb{S}^n\,\mathrm{U}^n = \tilde{\mathrm{F}}^{n,n-1}\,, \tag{2.8}$$

where

$$(\mathbb{S}^n\,\mathrm{U})_i = -\frac{1}{\Delta t_n}\,\sum_{j=1}^N (1 - \alpha_{ij}^n)\,m_{ij}\,(u_j - u_i) + \frac{1}{2}\,\sum_{j=1}^N (1 - \alpha_{ij}^n)\,d_{ij}^n\,(u_j - u_i)\,, \quad i = 1, \dots, M\,,$$

and the right-hand side $\tilde{\mathrm{F}}^{n,n-1}$ may depend on U^n only through the factors α_{ij}^n.

To obtain a well-defined numerical scheme, it is necessary to guarantee that the principal $M \times M$ submatrix (i.e., with indices $i, j = 1, \ldots, M$) of the matrix $(1/\Delta t_n)\,\mathbb{M} + \frac{1}{2}\,\mathbb{A}^n + \mathbb{S}^n$ is non-singular for any values $\alpha_{ij}^n \in [0,1]$ satisfying (2.7). Since $\mathbb{S}^n$ may vanish, a necessary condition for this is the invertibility of the principal $M \times M$ submatrix of $(1/\Delta t_n)\,\mathbb{M} + \frac{1}{2}\,\mathbb{A}^n$. On the other hand, a sufficient condition is that the principal $M \times M$ submatrix of $(1/\Delta t_n)\,\mathbb{M} + \frac{1}{2}\,\mathbb{A}^n$ is positive definite since the principal $M \times M$ submatrix of $\mathbb{S}^n$ is positive semidefinite (see [1, Lemma 1]).

From the above considerations, it is clear that the positive definiteness of the principal $M \times M$ submatrix of $(1/\Delta t_n)\,\mathbb{M} + \frac{1}{2}\,\mathbb{A}^n$ is of fundamental importance for a FCT-type method to be well defined. The aim of this work is then to give sufficient conditions to ensure this positive definiteness in the case of finite element discretizations of transient convection–diffusion–reaction equations using a group formulation of the convective term.

3. Transient convection–diffusion–reaction equation and its group finite element formulation

Let us consider the transient convection–diffusion–reaction equation

$$u_t - \varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u + c\,u = f \qquad \text{in } (0,T]\times\Omega\,, \tag{3.1}$$

$$u = u_b \qquad \text{on } [0,T]\times\partial\Omega\,, \tag{3.2}$$

$$u(0,\cdot) = u_0 \qquad \text{in } \Omega, \tag{3.3}$$

where $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is a bounded polygonal or polyhedral domain with a Lipschitz-continuous boundary $\partial\Omega$, $[0,T]$ is a time interval, $\varepsilon > 0$ is a constant diffusivity, $\boldsymbol{b} : [0,T] \to W^{1,\infty}(\Omega)^d$ is a divergence-free convection field, $c : [0,T] \to L^\infty(\Omega)$ is a nonnegative reaction coefficient, $f : [0,T] \to L^2(\Omega)$ is an outer source of the unknown quantity u, $u_b : [0,T] \to H^{1/2}(\partial\Omega)$ is the boundary condition, and $u_0 \in H_0^1(\Omega)$ is the initial condition.

To define a finite element discretization of (3.1)–(3.3) having the form (2.4)–(2.6), we introduce a triangulation $\mathscr{T}_h$ of Ω consisting of simplices possessing the usual compatibility properties and define the finite element spaces

$$W_h = \{v_h \in C(\overline{\Omega})\,;\ v_h|_T \in \mathbb{P}_1(T)\ \forall\, T \in \mathscr{T}_h\}\,, \qquad V_h = W_h \cap H_0^1(\Omega)$$

consisting of continuous piecewise linear functions. We denote by $x_1, \ldots, x_N$ the vertices of the triangulation $\mathscr{T}_h$ and assume that the first M vertices ($0 < M < N$) lie in Ω whereas $x_{M+1}, \ldots, x_N \in \partial\Omega$. We denote by $\varphi_1, \ldots, \varphi_N$ the standard basis functions of W_h assigned to these vertices that satisfy $\varphi_i(x_j) = \delta_{ij}$, $i, j = 1, \ldots, N$, where δ_{ij} is the Kronecker symbol. Then the functions $\varphi_1, \ldots, \varphi_M$ form a basis of V_h. For later use, we introduce the Lagrange interpolation operator $i_h : C(\overline{\Omega}) \to W_h$ by

$$i_h v = \sum_{j=1}^{N} v(x_j)\,\varphi_j\,, \qquad v \in C(\overline{\Omega})\,.$$

Using the standard Galerkin finite element discretization, the entries of the matrices $\mathbb{M}$ and $\mathbb{A}^n$ in (2.4) are given by

$$m_{ij} = (\varphi_j, \varphi_i)\,, \qquad a_{ij}^n = a^n(\varphi_j, \varphi_i)\,,$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$ and

$$a^n(u,v) = \varepsilon\,(\nabla u, \nabla v) + (\boldsymbol{b}^n \cdot \nabla u, v) + (c^n u, v)\,,$$

with the notation $\boldsymbol{b}^n = \boldsymbol{b}(t_n,\cdot)$, $c^n = c(t_n,\cdot)$. Obviously, the matrix $(m_{ij})_{i,j=1,\ldots,M}$ is positive definite. Moreover, since $\boldsymbol{b}^n$ is divergence-free and c^n is nonnegative, one has $a^n(v,v) \geq \varepsilon\,|v|^2_{1,\Omega}$ for any $v \in H^1_0(\Omega)$ so that the matrix $(a^n_{ij})_{i,j=1,\ldots,M}$ is positive definite as well. Next, since for the considered finite element space $m_{ij} \geq 0$, the principal $M \times M$ submatrix of $\mathbb{S}^n$ is positive semidefinite (the proof of this fact is essentially the same as in [1, Lemma 1]). Therefore, the principal $M \times M$ submatrix of the matrix $(1/\Delta t_n)\,\mathbb{M} + \frac{1}{2}\,\mathbb{A}^n + \mathbb{S}^n$ introduced in the preceding section is positive definite and hence non-singular. Thus, for any given correction factors $\alpha^n_{ij} \in [0,1]$ satisfying (2.7), the linearized FCT scheme (2.8) has a unique solution.

Now, as was mentioned before, assembling the stiffness matrix can be costly if $\boldsymbol{b}$ varies in time, since the computation of the entries $(\boldsymbol{b}^n \cdot \nabla\varphi_j, \varphi_i)$ is needed at each time step. The group finite element formulation [5, 4] appears as a cheaper alternative. It is based on writing

$$(\boldsymbol{b}^n \cdot \nabla u_h, v_h) = (\nabla \cdot (\boldsymbol{b}^n u_h), v_h)\,, \qquad u_h \in W_h,\ v_h \in V_h\,,$$

and replacing the product $\boldsymbol{b}^n u_h$ by one finite element function

$$i_h(\boldsymbol{b}^n u_h) = \sum_{j=1}^N \boldsymbol{b}^n_j\, u_j\, \varphi_j\,,$$

where we use the notation $\boldsymbol{b}^n_j = \boldsymbol{b}^n(x_j)$, $u_j = u_h(x_j)$. Note that then

$$(\nabla \cdot [i_h(\boldsymbol{b}^n u_h)], \varphi_i) = \sum_{j=1}^N (\boldsymbol{b}^n_j \cdot \nabla\varphi_j, \varphi_i)\, u_j = \sum_{k=1}^d \sum_{j=1}^N (\boldsymbol{b}^n_j)_k\, (\partial_k\varphi_j, \varphi_i)\, u_j\,. \tag{3.4}$$

Thus, it suffices to assemble the matrices $((\partial_k\varphi_j, \varphi_i))_{i,j=1,\ldots,N}$ for $k = 1,\ldots,d$ only once and the convection matrix at time level t_n is obtained very efficiently by multiplying this precomputed matrices by components of the nodal values of the convection field $\boldsymbol{b}^n$ instead of applying costly numerical quadrature.

Note that the group finite element method can be interpreted as an evaluation of the convective term by simple nodal quadrature. Indeed, denoting by $\mathcal{V}(T)$ the set of the $d+1$ vertices of any simplex $T \in \mathscr{T}_h$, one has

$$\begin{aligned}
(\boldsymbol{b}^n \cdot \nabla u_h, v_h) &= -(u_h, \boldsymbol{b}^n \cdot \nabla v_h) = -\sum_{T\in\mathscr{T}_h} (u_h, \boldsymbol{b}^n \cdot \nabla v_h)_T \\
&\approx -\sum_{T\in\mathscr{T}_h} \frac{|T|}{d+1} \sum_{x\in\mathcal{V}(T)} (u_h\, \boldsymbol{b}^n \cdot \nabla v_h|_T)(x) \\
&= -\sum_{T\in\mathscr{T}_h} \frac{|T|}{d+1} \sum_{x\in\mathcal{V}(T)} (i_h(\boldsymbol{b}^n u_h) \cdot \nabla v_h|_T)(x) \\
&= -\sum_{T\in\mathscr{T}_h} (i_h(\boldsymbol{b}^n u_h), \nabla v_h)_T = -(i_h(\boldsymbol{b}^n u_h), \nabla v_h) = (\nabla \cdot [i_h(\boldsymbol{b}^n u_h)], v_h)\,,
\end{aligned}$$

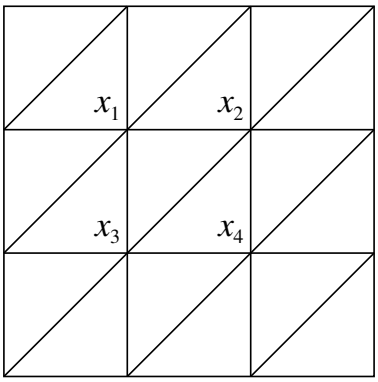

Figure 1: Triangulation used in the proof of Theorem 3.1.

where $(\cdot,\cdot)_T$ is the inner product in $L^2(T)$ or $L^2(T)^d$, $|T|$ is the d-dimensional measure of T, and we used the fact that the function $i_h(\boldsymbol{b}^n u_h)\cdot\nabla v_h|_T$ is linear on T so that it is integrated exactly using the nodal quadrature formula.

When applying the group finite element method to the discretization of the convective term, the stiffness matrix $\mathbb{A}^n$ has entries $a_{ij}^n = a_h^n(\varphi_j,\varphi_i)$, where

$$a_h^n(u,v) = \varepsilon\,(\nabla u,\nabla v) + (\nabla\cdot[i_h(\boldsymbol{b}^n u)],v) + (c^n u,v)\,.$$

Unfortunately, a careful inspection shows that then the principal $M\times M$ submatrix of $(1/\Delta t_n)\,\mathbb{M} + \frac{1}{2}\,\mathbb{A}^n$ can be singular. In other words, it can happen that there is a nontrivial function $u_h\in V_h$ satisfying

$$\frac{2}{\Delta t_n}\,(u_h,\varphi_i) + a_h^n(u_h,\varphi_i) = 0\,,\qquad i=1,\dots,M\,. \tag{3.5}$$

Note that the first term on the left-hand side of (3.5) is analogous as the reaction term $(c^n u_h,\varphi_i)$ in $a_h^n(u_h,\varphi_i)$. Therefore, instead of showing (3.5), it is sufficient to prove that, for a suitable mesh and for any ε and c^n, there is a divergence-free convective field $\boldsymbol{b}^n$ such that there exists $u_h\in V_h\setminus\{0\}$ satisfying

$$a_h^n(u_h,\varphi_i) = 0\,,\qquad i=1,\dots,M\,. \tag{3.6}$$

This will be done in the following theorem.

Theorem 3.1. *There is a polygonal domain $\Omega\subset\mathbb{R}^2$ and a triangulation $\mathscr{T}_h$ of Ω such that, for any $\varepsilon>0$ and $c^n\in L^\infty(\Omega)$, one can find a divergence-free function $\boldsymbol{b}^n\in W^{1,\infty}(\Omega)^2$ such that* (3.6) *holds for a nontrivial function $u_h\in V_h$.*

Proof. Let $\Omega=(0,3)^2$ and let $\mathscr{T}_h$ be the uniform triangulation of Ω depicted in Fig. 1. Then $M=4$. We use the following numbering of the interior vertices of $\mathscr{T}_h$:

$$x_1=(1,2)\,,\quad x_2=(2,2)\,,\quad x_3=(1,1)\,,\quad x_4=(2,1)\,.$$

Let $u_1,u_2,u_3,u_4\in\mathbb{R}\setminus\{0\}$ be arbitrary and set

$$u_h = \sum_{j=1}^{4} u_j\,\varphi_j\,.$$

Our aim is to show that, given $\varepsilon > 0$ and $c^n \in L^\infty(\Omega)$, there is a divergence-free convection field $\boldsymbol{b}^n \in W^{1,\infty}(\Omega)^2$ such that (3.6) is satisfied. As a matter of fact, it suffices to find suitable values of $\boldsymbol{b}^n$ at the vertices $x_1, \ldots, x_4$. We shall consider them in the form $\boldsymbol{b}_j^n = v_j\, \boldsymbol{z}$, $j = 1, \ldots, 4$, with a fixed vector $\boldsymbol{z} \in \mathbb{R}^2$. To obtain the divergence-free function $\boldsymbol{b}^n$, we first introduce smooth divergence-free vector fields $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \boldsymbol{\psi}_3, \boldsymbol{\psi}_4$ such that $\boldsymbol{\psi}_i(x_j) = \delta_{ij}\, \boldsymbol{z}$ for $i, j = 1, \ldots, 4$. For example, one may set $\boldsymbol{\psi}_i = (\partial_2 \eta_i, -\partial_1 \eta_i)$, $i = 1, \ldots, 4$, where $\eta_i = (\boldsymbol{z} \times x)\, \xi_i$ and $\xi_i \in C_0^\infty(\Omega)$ vanishes outside the ball $\{x \in \mathbb{R}^2\,;\ |x - x_i| < \frac{1}{2}\}$ and equals 1 in the ball $\{x \in \mathbb{R}^2\,;\ |x - x_i| < \frac{1}{4}\}$. Hence, once the values $v_1, \ldots, v_4$ are computed, $\boldsymbol{b}^n = \sum_{j=1}^4 v_j\, \boldsymbol{\psi}_j$ is the desired divergence-free vector field.

Using (3.4), one gets

$$(\nabla \cdot [i_h(\boldsymbol{b}^n u_h)], \varphi_i) = \sum_{j=1}^4 (\boldsymbol{z} \cdot \nabla \varphi_j, \varphi_i)\, u_j\, v_j\,.$$

Denoting

$$q_{ij} = (\boldsymbol{z} \cdot \nabla \varphi_j, \varphi_i)\,, \quad g_i = -\varepsilon\, (\nabla u_h, \nabla \varphi_i) - (c^n u_h, \varphi_i)\,, \qquad i, j = 1, \ldots, 4\,,$$

one can write (3.6) equivalently in the form

$$\sum_{j=1}^4 q_{ij}\, u_j\, v_j = g_i\,, \qquad i = 1, \ldots, 4\,. \tag{3.7}$$

If $|z_1| \neq |z_2|$, then the matrix $\mathbb{Q} := (q_{ij})_{i,j=1}^4$ is non-singular. For example, setting $\boldsymbol{z} = (1, 0)$, it is easy to verify that

$$\mathbb{Q} = \frac{1}{6} \begin{pmatrix} 0 & 2 & 1 & 0 \\ -2 & 0 & -1 & 1 \\ -1 & 1 & 0 & 2 \\ 0 & -1 & -2 & 0 \end{pmatrix}$$

and $\det \mathbb{Q} = \frac{1}{144}$. This implies that there are uniquely determined values $v_1, \ldots, v_4$ satisfying (3.7), which finishes the proof. □

In the next section, we shall formulate conditions on h and Δt_n under which the bilinear form $(1/\Delta t_n)\, (\cdot, \cdot) + \frac{1}{2}\, a_h^n(\cdot, \cdot)$ is elliptic on V_h and hence the principal $M \times M$ submatrix of $(1/\Delta t_n)\, \mathbb{M} + \frac{1}{2}\, \mathbb{A}^n$ is positive definite.

4. Results on ellipticity of the bilinear form

The aim of this section is to investigate the ellipticity of the bilinear form

$$a_h(u, v) = \varepsilon\, (\nabla u, \nabla v) + (\nabla \cdot [i_h(\boldsymbol{b}\, u)], v) + (c\, u, v)$$

on the finite element space $V_h = W_h \cap H_0^1(\Omega)$ where

$$W_h = \{v_h \in C(\overline{\Omega})\,;\ v_h|_T \in \mathbb{P}_k(T)\ \forall\, T \in \mathscr{T}_h\}$$

now consists of continuous piecewise polynomial functions of degree less than or equal to $k \geq 1$. Again, $i_h : C(\overline{\Omega}) \to W_h$ is the Lagrange interpolation operator. We still assume that $\varepsilon > 0$ is constant, $\boldsymbol{b} \in W^{1,\infty}(\Omega)^d$ is divergence-free, and $c \in L^\infty(\Omega)$ is nonnegative. For $k = 1$, $\boldsymbol{b} = \boldsymbol{b}^n$ and $c = c^n + 2/\Delta t_n$, one obtains a bilinear form generating the matrix $(2/\Delta t_n)\,\mathbb{M} + \mathbb{A}^n$ discussed in the previous section. In this section we consider the more general case to cover also other applications of the group FEM.

The ellipticity of the bilinear form a_h will be studied with respect to the norm

$$\|v\|_G = \left(\varepsilon\,|v|_{1,\Omega}^2 + \|c^{1/2}\,v\|_{0,\Omega}^2\right)^{1/2},$$

which is a natural norm for the Galerkin discretization of the steady-state case of (3.1). In view of the Friedrichs inequality, $\|\cdot\|_G$ is a norm on $H_0^1(\Omega)$ also if $c \equiv 0$. We shall specify conditions under which

$$a_h(v_h, v_h) \geq \frac{1}{2}\,\|v_h\|_G^2 \qquad \forall\ v_h \in V_h\,. \tag{4.1}$$

We assume that all triangulations $\mathscr{T}_h$ are shape-regular, i.e.,

$$\frac{h_T}{\varrho_T} \leq \sigma \qquad \forall\ T \in \mathscr{T}_h\,, \tag{4.2}$$

where h_T is the diameter of T, ϱ_T is the diameter of the largest ball contained in T, and σ is a constant independent of h. Then, for any $T \in \mathscr{T}_h$, the interpolation operator i_h satisfies

$$|v - i_h v|_{m,T} \leq C\,h_T^{s-m}\,|v|_{s,T} \qquad \forall\ v \in H^s(T)\,,\ s = 2, \ldots, k+1\,,\ m = 0, 1\,, \tag{4.3}$$

where C is a constant depending only on σ and k. We shall also need the inverse inequality

$$|v_h|_{m,T} \leq C_{inv}\,h_T^{s-m}\,|v_h|_{s,T} \qquad \forall\ v_h \in V_h\,,\ T \in \mathscr{T}_h\,,\ 0 \leq s < m \leq k\,, \tag{4.4}$$

where C_{inv} again depends only on σ and k.

In what follows, we shall derive various estimates of the term $(\nabla\cdot[i_h(\boldsymbol{b}\,v_h)], v_h)$ enabling to formulate conditions that allow us to prove the ellipticity (4.1). For $v_h \in V_h$, the integration by parts gives

$$(\nabla \cdot [i_h(\boldsymbol{b}\,v_h)], v_h) = -(i_h(\boldsymbol{b}\,v_h), \nabla v_h) \tag{4.5}$$

and hence

$$|(\nabla \cdot [i_h(\boldsymbol{b}\,v_h)], v_h)| \leq \|i_h(\boldsymbol{b}\,v_h)\|_{0,\Omega}\,|v_h|_{1,\Omega}\,.$$

For any $T \in \mathscr{T}_h$, one gets

$$\|i_h(\boldsymbol{b}\,v_h)\|_{0,T} \leq C\,|T|^{1/2}\,\|\boldsymbol{b}\|_{0,\infty,T}\,\|v_h\|_{0,\infty,T}\,,$$

with C depending only on k. From the equivalence of norms on finite-dimensional spaces (applied on the reference element), it follows that

$$|T|^{1/2}\,\|v\|_{0,\infty,T} \leq C\,\|v\|_{0,T} \qquad \forall\ v \in \mathbb{P}_k(T)\,,$$

and hence

$$\|i_h(\boldsymbol{b}\,v_h)\|_{0,T} \leq C_1\|\boldsymbol{b}\|_{0,\infty,T}\,\|v_h\|_{0,T}\,, \tag{4.6}$$

where the constant C_1 again depends only on k. Consequently,

$$|(\nabla \cdot [i_h(\boldsymbol{b}\, v_h)], v_h)| \le C_1 \|\boldsymbol{b}\|_{0,\infty,\Omega}\, \|v_h\|_{0,\Omega}\, |v_h|_{1,\Omega}\,. \tag{4.7}$$

If $h\,|\boldsymbol{b}|_{1,\infty,\Omega} < \|\boldsymbol{b}\|_{0,\infty,\Omega}$, this estimate can be improved by employing that $(\nabla\cdot(\boldsymbol{b}\, v_h), v_h) = (\boldsymbol{b}\cdot\nabla v_h, v_h) = 0$. This property implies that

$$\begin{aligned}(\nabla \cdot [i_h(\boldsymbol{b}\, v_h)], v_h) &= (\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h), \nabla v_h) = \sum_{T\in\mathscr{T}_h} (\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h), \nabla v_h)_T \\ &\le \sum_{T\in\mathscr{T}_h} \|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,T}\, |v_h|_{1,T}\,. \end{aligned}\tag{4.8}$$

Consider any $T \in \mathscr{T}_h$ and set

$$\boldsymbol{b}_T = \frac{1}{|T|}\int_T \boldsymbol{b}\,\mathrm{d}x\,.$$

Then

$$\|\boldsymbol{b} - \boldsymbol{b}_T\|_{0,\infty,T} \le C\, h_T\, |\boldsymbol{b}|_{1,\infty,T}\,,$$

where C depends only on σ from (4.2). Since $\boldsymbol{b}_T\, v_h = i_h(\boldsymbol{b}_T\, v_h)$, it follows using (4.6) that

$$\begin{aligned}\|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,T} &= \|(\boldsymbol{b} - \boldsymbol{b}_T)\, v_h - i_h((\boldsymbol{b} - \boldsymbol{b}_T)\, v_h)\|_{0,T} \\ &\le (1 + C_1)\, \|\boldsymbol{b} - \boldsymbol{b}_T\|_{0,\infty,T}\, \|v_h\|_{0,T}\end{aligned}$$

and hence one obtains

$$\|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,T} \le C_2\, h_T\, |\boldsymbol{b}|_{1,\infty,T}\, \|v_h\|_{0,T}\,, \tag{4.9}$$

with a constant C_2 depending only on σ and k. Then

$$|(\nabla \cdot [i_h(\boldsymbol{b}\, v_h)], v_h)| \le C_2\, h\, |\boldsymbol{b}|_{1,\infty,\Omega}\, \|v_h\|_{0,\Omega}\, |v_h|_{1,\Omega}\,, \tag{4.10}$$

or, using (4.4),

$$|(\nabla \cdot [i_h(\boldsymbol{b}\, v_h)], v_h)| \le C_2\, C_{inv}\, |\boldsymbol{b}|_{1,\infty,\Omega}\, \|v_h\|_{0,\Omega}^2\,. \tag{4.11}$$

If $\boldsymbol{b} \in W^{k+1,\infty}(\Omega)^d$, one can apply the interpolation error estimate (4.3) and the inverse inequality (4.4) to obtain

$$\begin{aligned}\|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,T} &\le C\, h_T^{k+1}\, |\boldsymbol{b}\, v_h|_{k+1,T} \le \tilde{C}\, h_T^{k+1}\, \|\boldsymbol{b}\|_{k+1,\infty,T}\, \|v_h\|_{k,T} \\ &\le \bar{C}\, h_T^2\, \|\boldsymbol{b}\|_{k+1,\infty,T}\, \|v_h\|_{1,T}\,.\end{aligned}\tag{4.12}$$

Due to the Friedrichs inequality, this implies that

$$|(\nabla \cdot [i_h(\boldsymbol{b}\, v_h)], v_h)| \le C\, h^2\, \|\boldsymbol{b}\|_{k+1,\infty,\Omega}\, |v_h|_{1,\Omega}^2\,, \tag{4.13}$$

but it does not lead to any improvement of (4.10) and (4.11) if we want to keep the norms of v_h used in these estimates.

Note that the estimate (4.9) cannot be improved by applying the interpolation error estimate (4.3) with $s \in \{2, \dots, k\}$. Indeed, one obtains using (4.4)

$$\|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,T} \le C\, h_T^s\, |\boldsymbol{b}\, v_h|_{s,T} \le \tilde{C}\, h_T^s\, \|\boldsymbol{b}\|_{s,\infty,T}\, \|v_h\|_{s,T} \le \bar{C}\, h_T\, \|\boldsymbol{b}\|_{s,\infty,T}\, \|v_h\|_{1,T}\,.$$

The estimates (4.7), (4.10), and (4.13) together with the Friedrichs inequality imply that there is a constant C_0 depending only on σ, k, and Ω such that

$$|(\nabla \cdot [i_h(\boldsymbol{b}\, v_h)], v_h)| \le C_0 \, \min\{\|\boldsymbol{b}\|_{0,\infty,\Omega}, \, h\, |\boldsymbol{b}|_{1,\infty,\Omega}, \, h^2\, \|\boldsymbol{b}\|_{k+1,\infty,\Omega}\}\, |v_h|^2_{1,\Omega}\,.$$

Thus, if

$$2\, C_0 \, \min\{\|\boldsymbol{b}\|_{0,\infty,\Omega}, \, h\, |\boldsymbol{b}|_{1,\infty,\Omega}, \, h^2\, \|\boldsymbol{b}\|_{k+1,\infty,\Omega}\} \le \varepsilon\,, \tag{4.14}$$

one obtains (4.1), i.e., the bilinear form a_h is elliptic. Note that $\boldsymbol{b}$ may possess boundary layers (typically if the flow field $\boldsymbol{b}$ satisfies a no-slip boundary condition) and then the minimum in (4.14) may be equal to $\|\boldsymbol{b}\|_{0,\infty,\Omega}$. Since ε is usually much smaller than $|\boldsymbol{b}|$ in applications, the condition (4.14) will often not be satisfied.

Another possibility how to prove (4.1) is to employ the contribution of the reaction term to the norm $\|\cdot\|_G$, assuming that

$$c_0 := \operatorname*{ess\,inf}_{\Omega} \, c > 0\,.$$

For this the local inverse inequality (4.4) is fundamental. In view of (4.11), the ellipticity inequality (4.1) holds if

$$2\, C_2\, C_{inv}\, |\boldsymbol{b}|_{1,\infty,\Omega} \le c_0\,. \tag{4.15}$$

If $\boldsymbol{b}$ possesses boundary layers, a less strict condition may be obtained by applying (4.4) and (4.6) as follows:

$$|(\nabla \cdot [i_h(\boldsymbol{b}\, v_h)], v_h)| \le \sqrt{d} \sum_{T\in\mathscr{T}_h} |i_h(\boldsymbol{b}\, v_h)|_{1,T}\, \|v_h\|_{0,T} \le \sqrt{d}\, C_1\, C_{inv} \sum_{T\in\mathscr{T}_h} h_T^{-1}\, \|\boldsymbol{b}\|_{0,\infty,T}\, \|v_h\|^2_{0,T}\,.$$

Thus, (4.1) also holds if

$$2\, \sqrt{d}\, C_1\, C_{inv}\, h_T^{-1}\, \|\boldsymbol{b}\|_{0,\infty,T} \le c_0 \tag{4.16}$$

for any $T \in \mathscr{T}_h$. However, it may be useful to formulate a condition for (4.1) involving both (4.16) and the local version of (4.15). To this end, one may use (4.8), apply the inverse inequality (4.4) and estimate $\|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,T}$ by taking the minimum of (4.9) and the estimate

$$\|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,T} \le (1 + C_1)\, \|\boldsymbol{b}\|_{0,\infty,T}\, \|v_h\|_{0,T}\,, \tag{4.17}$$

which follows from (4.6). Then one deduces that (4.1) is satisfied if

$$2\, C_{inv}\, \min\{(1 + C_1)\, h_T^{-1}\, \|\boldsymbol{b}\|_{0,\infty,T}, \, C_2\, |\boldsymbol{b}|_{1,\infty,T}\} \le c_0 \qquad \forall\, T \in \mathscr{T}_h\,. \tag{4.18}$$

Finally, let us mention that (4.1) holds also if

$$\min\{C_1 \|\boldsymbol{b}\|_{0,\infty,\Omega}, \, C_2\, h\, |\boldsymbol{b}|_{1,\infty,\Omega}\} \le \sqrt{\varepsilon\, c_0}\,. \tag{4.19}$$

Indeed, it then follows from (4.7) and (4.10) that

$$|(\nabla \cdot [i_h(\boldsymbol{b}\, v_h)], v_h)| \le \varepsilon^{1/2}\, |v_h|_{1,\Omega}\, c_0^{1/2}\, \|v_h\|_{0,\Omega} \le \frac{1}{2}\, \|v_h\|^2_G\,.$$

A local version of (4.19) reads

$$\min\{(1+C_1)\,\|\boldsymbol{b}\|_{0,\infty,T},\, C_2\,h_T\,|\boldsymbol{b}|_{1,\infty,T}\} \le \sqrt{\varepsilon\,c_0} \qquad \forall\, T \in \mathscr{T}_h \tag{4.20}$$

and follows from (4.8), (4.9), and (4.17).

If we now return to the transient problems of the previous sections, for which $\boldsymbol{b} = \boldsymbol{b}^n$ and $c = c^n + 2/\Delta t_n$, one has $c_0 \ge (2/\Delta t_n)$ and hence the conditions involving c_0 can be satisfied by choosing the time step appropriately. In particular, it follows from (4.16) that the FCT scheme is well defined if the time step Δt_n satisfies a CFL-like condition $\|\boldsymbol{b}^n\|_{0,\infty,T}\,\Delta t_n \le C\,h_T$ for every $T \in \mathscr{T}_h$.

Gathering all the above ellipticity conditions, we can state the following main result on the ellipticity of the bilinear form a_h.

Theorem 4.1. *The bilinear form a_h is elliptic if one of the conditions* (4.14), (4.15), (4.16), (4.18), (4.19), *or* (4.20) *is satisfied. If a_h corresponds to the time-discretized problem* (3.1), *then its ellipticity is guaranteed at the time level t^n under the CFL condition*

$$\|\boldsymbol{b}^n\|_{0,\infty,T}\,\Delta t_n \le C\,h_T \qquad \forall\, T \in \mathscr{T}_h\,, \tag{4.21}$$

or, more generally, under the condition

$$\min\{\|\boldsymbol{b}^n\|_{0,\infty,T},\, h_T\,|\boldsymbol{b}^n|_{1,\infty,T}\}\,\Delta t_n \le C\,h_T \qquad \forall\, T \in \mathscr{T}_h\,.$$

If $k = 1$, a possible remedy to avoid the use of one of the conditions listed in the above theorem is to consider a skew-symmetric discretization of the convective term. This is based on the fact that

$$(\boldsymbol{b}\cdot\nabla u, v) = \frac{1}{2}\,\{(\boldsymbol{b}\cdot\nabla u, v) - (u, \boldsymbol{b}\cdot\nabla v)\} \qquad \forall\, u \in H^1(\Omega),\ v \in H^1_0(\Omega)\,.$$

Applying the idea of the group FEM to this equivalent expression leads to the bilinear form

$$\tilde{a}_h(u,v) = \varepsilon\,(\nabla u, \nabla v) + \frac{1}{2}\,\{(\nabla\cdot[i_h(\boldsymbol{b}\,u)], v) - (u, \nabla\cdot[i_h(\boldsymbol{b}\,v)])\} + (c\,u, v)$$

that satisfies

$$\tilde{a}_h(v_h, v_h) = \|v_h\|_G^2 \qquad \forall\, v_h \in V_h\,.$$

Thus, the bilinear form $\tilde{a}_h$ is elliptic without any assumptions on the data and discretization parameters while keeping all the advantages of the group finite element formulation. However, for $k > 1$, this skew-symmetric discretization of the convective term is not appropriate since the corresponding consistency error is of first order (uniformly in ε) as we shall see in the next section.

Remark 4.1. *A further drawback to the alternative bilinear form $\tilde{a}_h$ is that the skew-symmetric rewriting of the convective term is not valid if a non-Dirichlet boundary condition is prescribed on $\Gamma \subset \partial\Omega$. Then the test functions v vanish only on $\partial\Omega \setminus \Gamma$ so that one has*

$$(\boldsymbol{b}\cdot\nabla u, v) = \frac{1}{2}\,\{(\boldsymbol{b}\cdot\nabla u, v) - (u, \boldsymbol{b}\cdot\nabla v)\} + \frac{1}{2}\int_\Gamma (\boldsymbol{b}\cdot\boldsymbol{n})\,u\,v\,\mathrm{d}s\,,$$

where $\boldsymbol{n}$ is the outward unit normal vector to $\partial\Omega$.

Remark 4.2. *The derivation of most of the conditions in this section relied on the identity* (4.5) *that is not valid if a non-Dirichlet boundary condition is prescribed on a part of* $\partial\Omega$ *(cf. the previous remark). This leads to more restrictive conditions for the ellipticity of* a_h*. For example, instead of* (4.14)*, one obtains* $\min\{\|\boldsymbol{b}\|_{1,\infty,\Omega},\, h\,\|\boldsymbol{b}\|_{k+1,\infty,\Omega}\} \le C\,\varepsilon$*. Nevertheless, the condition* (4.16) *and hence also the CFL condition* (4.21) *remain valid.*

5. Estimates of the consistency errors

A thorough error analysis of discretizations based on the group FEM is outside the scope of this work. Therefore, we confine ourselves to estimates of the consistency errors caused by replacing the standard Galerkin bilinear form

$$a(u,v) = \varepsilon\,(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u, v) + (c\,u, v)$$

by a_h or $\tilde{a}_h$. When deriving error estimates using the first Strang lemma, the difference $a(w_h, v_h) - a_h(w_h, v_h)$ or $a(w_h, v_h) - \tilde{a}_h(w_h, v_h)$ is estimated for w_h equal to an interpolant of the approximated solution u, which is assumed to be sufficiently regular. In what follows, we simply set $w_h := i_h u$.

Theorem 5.1. *Let* $\boldsymbol{b} \in W^{l+1,\infty}(\Omega)^d$ *for some* $l \in \{1,\dots,k\}$*, where* k *is the polynomial degree used for defining the finite element space* W_h*. Then, for any* $u \in H^{l+1}(\Omega)$ *and* $v_h \in V_h$*, one has*

$$|a(i_h u, v_h) - a_h(i_h u, v_h)| \le C\,h^l\,\max\{c_0, \varepsilon\,h^{-2}\}^{-1/2}\,\|\boldsymbol{b}\|_{l+1,\infty,\Omega}\,\|u\|_{l+1,\Omega}\,\|v_h\|_G\,,$$

where C *depends only on* $\operatorname{diam}(\Omega)$*,* σ*, and* k*.*

Proof. For $v_h \in V_h$ and $w_h \in W_h$, one obtains

$$\begin{aligned} a(w_h, v_h) - a_h(w_h, v_h) &= (\boldsymbol{b}\cdot\nabla w_h, v_h) - (\nabla\cdot[i_h(\boldsymbol{b}\,w_h)], v_h) \\ &= (\nabla\cdot[\boldsymbol{b}\,w_h - i_h(\boldsymbol{b}\,w_h)], v_h) \le \sqrt{d}\,|\boldsymbol{b}\,w_h - i_h(\boldsymbol{b}\,w_h)|_{1,\Omega}\,\|v_h\|_{0,\Omega}\,. \end{aligned}$$

On the other hand, integrating by parts before applying the Hölder inequality gives

$$a(w_h, v_h) - a_h(w_h, v_h) = -(\boldsymbol{b}\,w_h - i_h(\boldsymbol{b}\,w_h), \nabla v_h) \le \|\boldsymbol{b}\,w_h - i_h(\boldsymbol{b}\,w_h)\|_{0,\Omega}\,|v_h|_{1,\Omega}\,.$$

Setting $w_h := i_h u$ with $u \in H^{l+1}(\Omega)$, one has $i_h(\boldsymbol{b}\,w_h) = i_h(\boldsymbol{b}\,u)$ and hence, for $m = 0, 1$, one deduces from (4.3) that

$$\begin{aligned} |\boldsymbol{b}\,w_h - i_h(\boldsymbol{b}\,w_h)|_{m,\Omega} &\le |\boldsymbol{b}\,(u - i_h u)|_{m,\Omega} + |\boldsymbol{b}\,u - i_h(\boldsymbol{b}\,u)|_{m,\Omega} \\ &\le \|\boldsymbol{b}\|_{m,\infty,\Omega}\,\|u - i_h u\|_{m,\Omega} + C\,h^{l+1-m}\,|\boldsymbol{b}\,u|_{l+1,\Omega} \le \tilde{C}\,h^{l+1-m}\,\|\boldsymbol{b}\|_{l+1,\infty,\Omega}\,\|u\|_{l+1,\Omega}\,, \end{aligned}$$

where $\tilde{C}$ depends also on $\operatorname{diam}(\Omega)$ due to the estimate of $\|u - i_h u\|_{1,\Omega}$. Now the theorem follows by combining the above estimates. □

Theorem 5.2. *Let $\boldsymbol{b} \in W^{2,\infty}(\Omega)^d$. Then, for any $u \in H^2(\Omega)$ and $v_h \in V_h$, one has*

$$|a(i_h u, v_h) - \tilde{a}_h(i_h u, v_h)| \leq C\, h\, \max\{c_0, \varepsilon\, h^{-2}\}^{-1/2}\, \|\boldsymbol{b}\|_{2,\infty,\Omega}\, \|u\|_{2,\Omega}\, \|v_h\|_G\,,$$

where C depends only on $\mathrm{diam}(\Omega)$*, σ and the polynomial degree k used for defining the finite element space W_h.*

Proof. For any $v_h \in V_h$ and $w_h := i_h u$ with $u \in H^2(\Omega)$, one obtains

$$a(w_h, v_h) - \tilde{a}_h(w_h, v_h) = \frac{1}{2}\,(\nabla \cdot [\boldsymbol{b}\, w_h - i_h(\boldsymbol{b}\, w_h)], v_h) + \frac{1}{2}\,(\nabla w_h, \boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h))\,.$$

The first term on the right-hand side of this identity was estimated in the proof of the previous theorem. For the second one, it follows that

$$(\nabla w_h, \boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)) \leq C\, \|u\|_{2,\Omega}\, \|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,\Omega}\,, \tag{5.1}$$

where we used the estimate $|w_h|_{1,\Omega} \leq |u|_{1,\Omega} + |u - i_h u|_{1,\Omega} \leq (1 + Ch)\|u\|_{2,\Omega}$ so that C in (5.1) depends on $\mathrm{diam}(\Omega)$. Using (4.9), one obtains

$$\|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,\Omega} \leq C_2\, h\, |\boldsymbol{b}|_{1,\infty,\Omega}\, \|v_h\|_{0,\Omega}\,. \tag{5.2}$$

To obtain a second order estimate (with the H^1 norm of v_h on the right-hand side) without the need of a higher regularity of $\boldsymbol{b}$ than assumed, we first introduce a piecewise linear interpolant $\boldsymbol{b}_h$ of $\boldsymbol{b}$. Then

$$|\boldsymbol{b} - \boldsymbol{b}_h|_{m,\infty,\Omega} \leq C\, h^{2-m}\, |\boldsymbol{b}|_{2,\infty,\Omega}\,, \qquad m = 0, 1\,, \tag{5.3}$$

where C depends only on σ. The inequalities (4.12) and (5.3) imply

$$\|\boldsymbol{b}_h v_h - i_h(\boldsymbol{b}_h v_h)\|_{0,\Omega} \leq C\, h^2\, \|\boldsymbol{b}_h\|_{1,\infty,\Omega}\, \|v_h\|_{1,\Omega} \leq \tilde{C}\, h^2\, \|\boldsymbol{b}\|_{2,\infty,\Omega}\, \|v_h\|_{1,\Omega}\,.$$

Furthermore, applying (4.6) and (5.3), one derives

$$\|(\boldsymbol{b} - \boldsymbol{b}_h)\, v_h - i_h((\boldsymbol{b} - \boldsymbol{b}_h)\, v_h)\|_{0,\Omega} \leq (1 + C_1)\, \|\boldsymbol{b} - \boldsymbol{b}_h\|_{0,\infty,\Omega}\, \|v_h\|_{0,\Omega} \leq C\, h^2\, |\boldsymbol{b}|_{2,\infty,\Omega}\, \|v_h\|_{0,\Omega}\,.$$

Summing the last two estimates and applying the Friedrichs inequality gives

$$\|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,\Omega} \leq C\, h^2\, \|\boldsymbol{b}\|_{2,\infty,\Omega}\, \|v_h\|_{1,\Omega} \leq \tilde{C}\, h^2\, \|\boldsymbol{b}\|_{2,\infty,\Omega}\, |v_h|_{1,\Omega}\,. \tag{5.4}$$

Combining (5.2) and (5.4), one obtains

$$\|\boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h)\|_{0,\Omega} \leq C\, h\, \max\{c_0, \varepsilon\, h^{-2}\}^{-1/2}\, \|\boldsymbol{b}\|_{2,\infty,\Omega}\, \|v_h\|_G\,,$$

which completes the proof. □

The above theorems show that, if $c_0 > 0$, the consistency error of the bilinear form a_h is of the order $O(h^l)$ uniformly for $\varepsilon \to 0$ so that the group formulation does not decrease the order of the method. On the other hand, the skew-symmetric group formulation represented by the bilinear form $\tilde{a}_h$ leads to a first order estimate only due to the term $(\nabla i_h u, \boldsymbol{b}\, v_h - i_h(\boldsymbol{b}\, v_h))$. This is the case also if $\boldsymbol{b}$ possesses a higher regularity than assumed in Theorem 5.2. If $c_0 = 0$, the consistency errors are at least of the order $O(\varepsilon^{-1/2}\, h^{l+1})$, resp. $O(\varepsilon^{-1/2}\, h^2)$.

Acknowledgments

The work of P. Knobloch has been partially supported through the grant No. 16-03230S of the Czech Science Foundation.

References

[1] G.R. Barrenechea, V. John, P. Knobloch, Analysis of algebraic flux correction schemes, SIAM J. Numer. Anal. 54 (2016) 2427–2451.

[2] O. Boiarkine, D. Kuzmin, S. Čanić, G. Guidoboni, A. Mikelić, A positivity-preserving ALE finite element scheme for convection–diffusion equations in moving domains, J. Comput. Phys. 230 (2011) 2896–2914.

[3] J.P. Boris, D.L. Book, Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works, J. Comput. Phys. 11 (1973) 38–69.

[4] C.A.J. Fletcher, A comparison of finite element and finite difference solutions of the one- and two-dimensional Burgers' equations, J. Comput. Phys. 51 (1983) 159–188.

[5] C.A.J. Fletcher, The group finite element formulation, Comput. Methods Appl. Mech. Engrg. 37 (1983) 225–244.

[6] V. John, J. Novo, On (essentially) non-oscillatory discretizations of evolutionary convection–diffusion equations, J. Comput. Phys. 231 (2012) 1570–1586.

[7] V. John, E. Schmeyer, Finite element methods for time-dependent convection–diffusion–reaction equations with small diffusion, Comput. Methods Appl. Mech. Engrg. 198 (2008) 475–494.

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

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

[10] D. Kuzmin, M. Möller, Algebraic flux correction I. Scalar conservation laws, in: D. Kuzmin, R. Löhner, S. Turek (Eds.), Flux-Corrected Transport. Principles, Algorithms, and Applications, Springer-Verlag, Berlin, 2005, pp. 155–206.

[11] D. Kuzmin, M. Möller, J.N. Shadid, M. Shashkov, Failsafe flux limiting and constrained data projections for equations of gas dynamics, J. Comput. Phys. 229 (2010) 8766–8779.

[12] D. Kuzmin, S. Turek, Flux correction tools for finite elements, J. Comput. Phys. 175 (2002) 525–558.

[13] R. Löhner, K. Morgan, J. Peraire, M. Vahdati, Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier–Stokes equations, Int. J. Numer. Methods Fluids 7 (1987) 1093–1109.

[14] R. Löhner, K. Morgan, M. Vahdati, J.P. Boris, D.L. Book, FEM-FCT: Combining unstructured grids with high resolution, Commun. Appl. Numer. Methods 4 (1988) 717–729.

[15] H.G. Roos, M. Stynes, L. Tobiska, Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion–Reaction and Flow Problems. 2nd ed., Springer-Verlag, Berlin, 2008.

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