Czech Technical University
Faculty of Civil Engineering

HABILITATION THESIS

Miloslav Vlasák

Analysis of time discretizations for parabolic problems with application to space discretizations

Abstract

This work summarizes some of the theoretical results of the author in last ten years, where the main area of the research was the numerical analysis for the stable higher order time discretization methods applied on parabolic problems. The main discretization scheme is the time discontinuous Galerkin method in combination with the conforming finite element method or the discontinuous Galerkin method in space. The thesis presents a priori error estimates for nonstationary singularly perturbed convection-diffusion problems, stability results for the problems with the domain evolving in time and a posteriori error estimates based on the equilibrated flux reconstructions. The technique presented for a posteriori analysis in time is applied to purely spatial problem and the quality of the recontruction is investigated with respect to the degree of polynomial approximation.

Keywords

discontinuous Galerkin method, convection-diffusion equation, error analysis, a posteriori analysis, p-robustness, time discontinuous Galerkin, arbitrary Lagrangian-Eulerian description.

Thanks

I would like to thank all of my coworkers, mainly to Monika Balazsová, Vít Dolejší, Miloslav Feistauer, Václav Kučera, Hans-Görg Roos and Filip Roskovec. I would like to thank for the general support to both institutions (MFF UK and FSv ČVUT), where I was employed, and especially to the corresponding departments (Department of Numerical Mathematics and Department of Mathematics).

I would like to thank also to my family and mainly to my wife Zuzana, who tried to help me and support me all the time.

Contents

Chapter 1

Introduction

There is number of areas for application of parabolic problems (mathematics, engineering, physics, biology, chemistry, economy, sociology, etc.). These problems are often discretized in space variables and the resulting large system of *stiff* ordinary differential equations (ODEs) needs to be solved by a suitable method. Backward differentiation formulae (BDF) were often considered as the method of the first choice for *stiff* problems, see e.g. [26], since they are robust and quite cheap. Nevertheless, BDF methods suffer from number of disadvantages. Namely, the order of convergence is limited by order 6, BDF are A-stable only to the order 2 and the robustness (area of stability) of the method decreases with the increasing order. Moreover, these methods are multi-step methods and suffers from usual disadvantages of multi-step methods in general, e.g. the necessity to define artificial starting values and stability issues connected with the step-size adaptation.

On the other hand, certain implicit Runge-Kutta methods and Galerkin time discretizations do not suffer these disadvantages. These methods are A-stable one-step methods of arbitrary order, for the overview about these methods see e.g. [29] and [30] and the citations therein. The main disadvantage of these methods that prevented the use of them in past years was their expensiveness, where the computational costs significantly increase with the order of the method. In comparison, BDF methods remain at the same cost independently of the order. Fortunately, the increase in computational power and advancements in numerical linear algebra in last two decades enabled practical applications of implicit RK or Galerkin methods. This makes implicit Runge-Kutta and Galerkin methods competitive with more traditional approaches like BDF.

This thesis presents some results achieved by the author and his coworkers in last 10 years about theoretical (numerical) analysis of Galerkin time discretizations for unsteady convection-diffusion problems. The main part of the thesis consists from 5 papers [6], [17], [34], [47] and [48] published in impact journals and presented here as Chapters 3–7. Each of these papers is presented in the same form as it is published. Therefore, all of these papers have their own individual style, page numbering, notation and references.

Chapter 3 and Chapter 4 study unsteady singularly perturbed convection-diffusion problems. The convection-diffusion problems appear in many practical applications, especially as a simplified model to Navier-Stokes equations. This problem represents a serious challenge to discretize, whenever the diffusion term is small in comparison to the other terms or data. Such a situation represents the transition state between parabolic and hyperbolic problems, where sharp boundary layers often appear. Usual finite element or finite difference discretizations fail in this situation, since they lead to the solution with highly oscillatory behavior around these layers that pollutes the solution not only in the vicinity of the layer, but at the all

computational domain. The overview of discretization techniques and their analysis for linear singularly perturbed problems can be found in [41]. The analysis of unsteady linear singularly perturbed problems can be found in e.g. [1] and [16]. The application to unsteady nonlinear problems can be found in [22]. For the analysis of Runge-Kutta methods applied to hyperbolic problems see [51].

Chapter 5 is devoted to the higher order analysis of unsteady convection-diffusion problems in time dependent domains, where the domain change is driven by a given smooth mapping. There are number of approaches dealing with time dependent domain problems, e.g. the fictitious domain method or the immersed boundary method. Another popular approach is Arbitrary Lagrangian-Eulerian (ALE) method based on one-to-one ALE mapping between the current evolving domain and the fixed reference domain. ALE method was analyzed mainly for the lower (first or second) order time discretization methods in combination with the classical conforming finite element method, see e.g. [23] and [25]. Analysis of higher order discretizations based on the discontinuous Galerkin method can be found in [8], [9] and [44].

Chapter 6 studies a posteriori error estimates for nonlinear parabolic problems. The aim of this chapter is to derive a posteriori error estimates that are cheap in comparison with the original discrete problem, fully computable, reliable and locally efficient. There are number of results devoted to a posteriori error estimates for parabolic problems. Most of these results assume lower (first or second) order time discretizations, see e.g. [27] or [40]. The aposteriori analysis of linear parabolic problems discretized by higher order methods in time based on the discontinuous Galerkin method can be found in [3], [20] and [42]. Nonlinear parabolic problems and higher order time discretizations are addressed in [36], where the upper bound consists from a dual norm and therefore it is not directly computable. For a general overview on a posteriori error concepts see e.g. [45].

Chapter 7 apply the reconstruction principle developed for the time discretization in [17] to the space discretization. Moreover, the efficiency of the derived a posteriori error estimate is studied with respect to the polynomial degree in one dimension. The topic of polynomial robustness (or polynomial dependence of the estimates) is important for the save application of a posteriori error bounds in *hp*-adaptive strategies with high polynomial degrees and it started to be very popular in the community of a posteriori error analysis in recent years. The first results for residual based estimates can be found in [37]. Very important results showing complete polynomial independence of equilibrated reconstructions are in [10]. The results from [10] are applied to large number of numerical methods in [21]. Paper [20] shows a complete polynomial independence of efficiency estimates for the discontinuous Galerkin time discretization for parabolic problems.

A general overview chapter precedes these main chapters. This overview contains a brief description of Chapters 3–7. Moreover, it contains a general description of several concepts for discretizations as well as the corresponding numerical analysis. The notation in this chapter is unified for convenience of the reader and is chosen as close as possible to the notation used in following chapters. The full explanation of the ideas and the full description of the concepts from the original papers can be rather long and technical in many situations. Therefore, the precision of the formulations is not always perfect in this overview, e.g. mean values, penalization parameters, reconstructions, etc., are defined only inside of the computational domain. The complete precise formulations can be found in the original papers or in Chapters 3–7.

Chapter 2

Overview

2.1 Notation

Here, we summarize a basic notation for the upcoming discretizations.

2.1.1 Space discretization notation

Let us assume a bounded polygonal domain $\Omega \subset \mathbb{R}^d$ with Lipschitz continuous boundary. We assume a partition of this domain into closed subsets K with mutually disjoint interiors and covering $\overline{\Omega}$, often called elements. For simplicity, we assume that elements K are simplices and that the partition is conforming, i.e. that the neigbouring elements share the entire edge or face depending on the dimension d. To simplify further notation, we call these boundary objects of co-dimension 1 edges regardless of the dimension d and denote them e.

We assume patches of elements ω_a denoting the patch consisting of the elements containing the common vertex a and ω_K denoting the patch consisting of the elements surrounding K and K itself.

We assume that the elements are shaped regular, i.e. the ratio of the diameters of the inscribed and circumscribed ball is bounded. We denote the local mesh-size $h_K = \mathrm{diam}(K)$ and the global mesh-size $h = \max_K h_K$. Finally, we assume that the mesh is locally quasi-uniform, i.e. the ratio $h_K/h_{K'}$ is bounded for neighnouring elements K and K'.

Moreover, we denote unit normals on edge e as n. The direction of the normals is arbitrary but fixed for the inner edges and outward for the boundary edges.

For piece-wise discontinuous function v, we need to define one-sided values on the edges

$$v_L(x) = \lim_{\epsilon \to 0+} v(x - \epsilon n), \qquad v_R(x) = \lim_{\epsilon \to 0+} v(x + \epsilon n) \tag{2.1}$$

depending on the orientation of n, jumps and mean values

$$[v] = v_L - v_R, \qquad \langle v \rangle = \frac{v_L + v_R}{2}. \tag{2.2}$$

We denote by $(.,.)_M$ and $\|.\|_M$ $L^2(M)$-scalar product and norm, respectively. Typically, we apply this notation with $M = K$ or $M = e$. The global $L^2(\Omega)$-scalar product and norm are denoted by $(.,.)$ and $\|.\|$, respectively. We denote the sum over all elements K or over all edges e of the mesh by $\sum_K$ or $\sum_e$, respectively.

2.1.2 Time discretization notation

Let us assume time interval $I = (0, T)$, where $T > 0$. We assume time partition of $\overline{I}$ by partition nodes $0 = t_0 < t_1 < \ldots < t_r = T$. Although the papers discussed often assume a general time partition, we assume here for simplicity that the partition is equidistant, i.e. $t_m = m\tau$, where τ is a global step-size. We denote local time subintervals $I_m = (t_{m-1}, t_m)$.

Combining the space and time discretization, we denote by $(.,.)_{M,m}$ and $\|.\|_{M,m}$ $L^2(M \times I_m)$-scalar product and norm, respectively. We denote the sum over all elements of the mesh and all the time subintervals by $\sum_{K,m}$.

For any function $f(t)$ defined in $\overline{I}$ we denote one sided nodal values $f(t_m\pm) = f^m_\pm$, where the subscript $\pm$ can be omitted for continuous functions, and we denote the corresponding jump in time as $\{v\}_m = v^m_+ - v^m_-$. The time derivative of function $f(t)$ is denoted as $f'(t)$.

2.2 One-step higher order time discretizations

Here, we present some classical one-step discretization techniques. For the overview see e.g. [29] and [30].

2.2.1 One-step discretizations

Let us consider ordinary differential equation (ODE)

$$\begin{aligned} y'(t) &= f(t, y(t)), \quad t \in (0, T), \\ y(0) &= \alpha. \end{aligned} \tag{2.3}$$

Let us denote the approximate solution $\{Y^m\}_{m=0}^r$ such that $y(t_m) \approx Y^m$. We can define three classes of one-step methods.

Runge–Kutta methods: Let $a_{i,j}$, b_i, c_i, $i, j = 1, \ldots, q+1$ be suitable coefficients. Then we call the sequence Y^m satisfying $Y^0 = \alpha$

$$\begin{aligned} g^m_i &= Y^{m-1} + \tau \sum_{j=1}^{q+1} a_{i,j} f(t_{m-1} + \tau c_j, g^m_j), \quad \forall i = 1, \ldots, q+1, \\ Y^m &= Y^{m-1} + \tau \sum_{i=1}^{q+1} b_i f(t_{m-1} + \tau c_i, g^m_i) \end{aligned} \tag{2.4}$$

the Runge-Kutta (RK) solution of (2.3). The auxiliary values g^m_i called *inner stages* represent the approximation of the exact solution in $t_{m-1} + \tau c_i$.

Collocation methods: Let c_i, $i = 1, \ldots, q+1$ be suitable coefficients. Let $Y^0 = \alpha$. In every step we construct polynomial p of degree at most $q+1$ such that

$$\begin{aligned} p(t_{m-1}) &= Y^{m-1}, \\ p'(t_{m-1} + \tau c_i) &= f(t_{m-1} + \tau c_i, p(t_{m-1} + \tau c_i)), \quad \forall i = 1, \ldots, q+1. \end{aligned} \tag{2.5}$$

Then we put $Y^m = p(t_m)$. We call the resulting sequence the collocation solution of (2.3). The points $t_{m-1} + \tau c_i$ are called collocation points. The method produces a piecewise polynomial function that satisfies the original equation (2.3) in these collocation points only.

Continuous and discontinuous Galerkin method: Let us define function spaces

$$X^\tau = \{v \in L^2(0, T) : v|_{I_m} \in P^q(I_m)\}, \tag{2.6}$$

$$Y^\tau = \{v \in C(0, T) : v|_{I_m} \in P^{q+1}(I_m), v(0) = \alpha\}, \tag{2.7}$$

where P^q and P^{q+1} are spaces of polynomials of degree q and $q+1$, respectively. It should be pointed out that both these spaces have the same dimension. We call $u \in Y^\tau$ the continuous Galerkin solution of (2.3) if

$$\int_{I_m} u'(t)v(t)\mathrm{d}t = \int_{I_m} f(t,u(t))v(t)\mathrm{d}t, \quad \forall v \in X^\tau. \tag{2.8}$$

We call $u \in X^\tau$ the discontinuous Galerkin solution of (2.3) if $u_-^0 = \alpha$ and

$$\int_{I_m} u'(t)v(t)\mathrm{d}t + \{u\}_{m-1}v_+^{m-1} = \int_{I_m} f(t,u(t))v(t)\mathrm{d}t, \quad \forall v \in X^\tau. \tag{2.9}$$

For comparison with previous methods we focus mainly on endpoints of intervals: $u_-^m = Y^m \approx y(t_m)$.

The integrals in the definition of continuous and discontinuous Galerkin method are often approximated by quadratures. Suitable quadratures are Gauss or right Radau quadratures on $q+1$ quadrature nodes, respectively, since they approximate all linear terms involved in the integrals exactly. We refer to these Galerkin methods approximated by Gauss or Radau quadrature as to quadrature variants.

2.2.2 Mutual connection between Runge-Kutta methods and Galerkin methods

It is very useful in the numerical analysis to understand the mutual connections among Runge-Kutta methods, collocation methods and Galerkin methods. This connection can be described by following lemmae.

Lemma 2.2.1 *Let the RK coefficients be chosen in the following way*

$$a_{i,j} = \int_0^{c_i} \ell_j(t)\mathrm{d}t, \quad i,j = 1,\dots,q+1, \tag{2.10}$$

$$b_i = \int_0^1 \ell_i(t)\mathrm{d}t, \quad i = 1,\dots,q+1, \tag{2.11}$$

where ℓ_i is the Lagrange interpolation basis function

$$\ell_i(t) = \prod_{j\neq i} \frac{t-c_j}{c_i-c_j}. \tag{2.12}$$

Then the values g_i^m, $i = 1,\dots,q+1$ and Y^m produced by such a RK method are equal to the values $p(t_{m-1}+\tau c_i)$, $i = 1,\dots,q+1$ and Y^m produced by the collocation method with the same coefficients c_i.

The proof can be found in [28] or [50].

Lemma 2.2.2 *Let $p \in P^{q+1}$ be the collocation polynomial on I_m associated to the collocation method with coefficients c_i chosen as Gauss quadrature nodes on $(0,1)$, $u \in P^{q+1}$ be the quadrature variant of continuous Galerkin solution on I_m. Then*

$$p(t) = u(t). \tag{2.13}$$

Lemma 2.2.3 *Let $p \in P^{q+1}$ be the collocation polynomial on I_m associated to the collocation method with coefficients c_i chosen as right Radau quadrature nodes, $u \in P^q$ be the quadrature variant of discontinuous Galerkin solution on I_m and $r_m \in P^{q+1}$ satisfy $r_m(t_{m-1}) = 1$, $r_m(t_m) = 0$ and $r_m \perp P^{q-1}$ on I_m. Then*

$$p(t) = u(t) - \{u\}_{m-1}r_m(t). \tag{2.14}$$

The proof for continuous Galerkin version can be found directly in [31]. The proof for discontinuous Galerkin version can be made similarly, see e.g. [49].

Summarizing these results, it is possible to realize that both Galerkin methods (up to corresponding quadrature and mild reconstruction (2.14) in case of discontinuous version) are special variants of the collocation methods and the collocation methods are special variants of the implicit Runge-Kutta methods. This can be exploited in the numerical analysis by application of the knowledge from one area to another area, especially by using the results about very well understood Runge-Kutta methods for the analysis of the Galerkin methods. The variants of Runge-Kutta methods corresponding to continuous and discontinuous Galerkin method are well known Kuntzmann-Butcher method (also known as Gauss-Legendre method) and Radau IIA method, respectively. For more details see [35] and [18], respectively.

2.3 Discontinuous Galerkin space discretization

Although most of the papers in this thesis are devoted to the time discretization techniques and their analysis, the space discretization is often made with the aid of the discontinuous Galerkin method. We shall briefly describe the discontinuous Galerkin method on simplified example of the Poisson equation

$$-\Delta u = f, \quad \text{in } \Omega. \tag{2.15}$$

We assume for simplicity the homogeneous Dirichlet boundary conditions. The other possibilities can be found in [15].

We apply the notation from Section 2.1.1. The difference between the classical finite element method and the discontinuous Galerkin method is in application of the discontinuous finite element space

$$X_h = \{v \in L^2(\Omega) : v|_K \in P^p(K)\}. \tag{2.16}$$

Since $X_h \not\subset H_0^1(\Omega)$, we could not apply the week formulation of problem (2.15) directly. In fact, we enhance the classical week formulation with additional terms. Among many variants of the discontinuous Galerkin method, one of the most popular approaches is the interior penalty method

$$\begin{aligned}(-\Delta u, v) \approx A_h(u,v) = \sum_K (\nabla u, \nabla v)_K - \sum_e (\langle \nabla u \rangle \cdot n, [v])_e \\ -\theta \sum_e (\langle \nabla v \rangle \cdot n, [u])_e + \sum_e (\alpha [u], [v])_e,\end{aligned} \tag{2.17}$$

where the choice of the parameter $\theta = 1, 0, -1$ corresponds to the symmetric (SIPG), incomplete (IIPG) and nonsymmetric (NIPG) variant. The parameter α is usually chosen as

$$\alpha = \frac{C_W}{h_e}, \tag{2.18}$$

where h_e is some intermediate value between h_K and $h_{K'}$ for neigbouring elements K and K' sharing the edge e. The constant $C_W > 0$ needs to be chosen large enough to guarantee the positivity of $A_h(.,.)$ on X_h. The detailed information about the suitable choice of the constant C_W can be found in [15].

The resulting discrete formulation of problem (2.15) is: find $u_h \in X_h$ such that

$$A_h(u_h, v_h) = (f, v_h), \quad \forall v_h \in X_h. \tag{2.19}$$

The corresponding error analysis can be found in [15].

2.4 Analysis of discontinuous Galerkin time discretization

In this section is described the most common approach to the derivation of a priori error estimates for the discontinuous Galerkin time discretization of parabolic problems. For simplicity, let us assume the heat equation

$$\begin{aligned} u' - \Delta u &= f, \quad \text{in } \Omega \times (0,T) \\ u(0) &= u^0, \quad \text{in } \Omega \end{aligned} \tag{2.20}$$

with homogeneous Dirichlet boundary condition.

We apply the notation from Section 2.1.1 and Section 2.1.2. We discretize this problem in space by the classical finite element method with the finite element space

$$X_h = \{v \in H_0^1(\Omega) : v|_K \in P^p(K)\}. \tag{2.21}$$

The resulting semidiscrete problem assumes the solution $u_h \in C^1(0,T,X_h)$ such that

$$\begin{aligned} (u_h', v) + (\nabla u_h, \nabla v) &= (f,v), \quad \forall v \in X_h \\ (u_h(0), v) &= (u^0, v), \quad \forall v \in X_h. \end{aligned} \tag{2.22}$$

The semidiscrete problem (2.22) represents the system of ODEs that can be solved by the discontinuous Galerkin method. Similarly as in Section 2.2.1, we define the fully discrete space

$$X_h^\tau = \{v \in L^2(0,T,X_h) : v|_{I_m} \in P^q(I_m, X_h)\}. \tag{2.23}$$

Then the fully discrete solution $U \in X_h^\tau$ satisfies

$$\begin{aligned} \int_{I_m} (U', v) + (\nabla U, \nabla v)\mathrm{d}t + (\{U\}_{m-1}, v_+^{m-1}) &= \int_{I_m} (f,v)\mathrm{d}t, \quad \forall v \in X_h^\tau, \\ (U_-^0, v) &= (u^0, v), \quad \forall v \in X_h. \end{aligned} \tag{2.24}$$

We may apply the technique of the error analysis described in [43]. Typically, we are interested in upper bounds of the error $e = U - u$ and most often in the nodes of the time partition t_m, i.e. $e_-^m = U_-^m - u(t_m)$. The error analysis most often consists from construction of suitable projection π on X_h^τ and dividing the error into projection part of the error $\eta = \pi u - u$, i.e. the error of the projection of the exact solution, and the rest of the error $\xi = U - \pi u \in X_h^\tau$. We gain the error equation by integrating relation (2.20) in weak form over I_m and subtracting this relation from (2.24). After dividing the error into ξ and η we gain for any $v \in X_h^\tau$

$$\begin{aligned} \int_{I_m} (\xi', v) + (\nabla \xi, \nabla v)\mathrm{d}t + (\{\xi\}_{m-1}, v_+^{m-1}) &= -\int_{I_m} (\nabla \eta, \nabla v)\mathrm{d}t \\ &\quad - \left(\int_{I_m} (\eta', v)\mathrm{d}t + (\{\eta\}_{m-1}, v_+^{m-1}) \right). \end{aligned} \tag{2.25}$$

The most usual projection $\pi : L^2(0,T,L^2(\Omega)) \to X_h^\tau$ is defined as

$$\begin{aligned} \int_{I_m} (\pi u - u, v t^j)\mathrm{d}t &= 0, \quad \forall v \in X_h,\ j \le q-1, \\ ((\pi u)_-^m, v) &= (u(t_m), v), \quad \forall v \in X_h. \end{aligned} \tag{2.26}$$

Advantage of this projection is that the terms on the second row of (2.25) vanish for any $v \in X_h^\tau$. Setting $v = 2\xi$ we gain

$$2\int_{I_m} (\xi', \xi)\mathrm{d}t + 2(\{\xi\}_{m-1}, \xi_+^{m-1}) = \|\xi_-^m\|^2 - \|\xi_-^{m-1}\|^2 + \|\{\xi\}_{m-1}\|^2, \tag{2.27}$$

cf [19]. Using (2.27) together with Cauchy inequality gives the error estimate for $\|\xi_-^m\|$ in terms of η

$$\|\xi_-^m\|^2 - \|\xi_-^{m-1}\|^2 + \|\{\xi\}_{m-1}\|^2 + \int_{I_m} \|\nabla\xi\|^2\mathrm{d}t \le \int_{I_m} \|\nabla\eta\|^2\mathrm{d}t. \tag{2.28}$$

The estimate

$$\int_{I_m} \|\nabla\eta\|^2 \le C\tau(h^{2p} + \tau^{2q+2}), \tag{2.29}$$

where the constant C depends on the corresponding derivatives of the exact solution u, are most often derived by the standard scaling argument using Bramble-Hilbert trick applied for Bochner spaces, see e.g. [46]. Since η_-^m is the error of L^2-orthogonal projection of u^m that satisfies $\|\eta_-^m\| \le Ch^{p+1}$, we gain from (2.28) and (2.29) the final desired estimate

$$\|e_-^m\| = \|U_-^m - u^m\| \le \|\xi_-^m\| + \|\eta_-^m\| \le C(h^p + \tau^{q+1}). \tag{2.30}$$

This estimate is usually considered optimal with respect to the polynomial degree in time, but suboptimal with respect to the polynomial degree in space, since h^{p+1} is usually expected for the finite element error in L^2-norm. The improvement to h^{p+1} can be found in [46]. Moreover, the basic theory of Runge-Kutta methods suggests that the nodal errors should converge with the rate τ^{2q+1} instead of τ^{q+1}. This faster convergence in a finite element setting is usually described as nodal *superconvergence*. Unfortunately, these faster rates appear only exceptionally for parabolic problems when certain compatibility conditions are met, cf. [3]. This *order reduction* phenomenon is analyzed in [11]. See also [24], where the investigation of convergence rate τ^{q+2} for $q \ge 1$ is presented.

2.5 A posteriori error estimates

Let us consider the Poisson problem

$$-\Delta u = f, \quad \text{in } \Omega, \tag{2.31}$$

where we assume for simplicity the homogeneous boundary condition. The resulting weak solution of problem (2.31) satisfies $u \in H_0^1(\Omega)$. Moreover, it is possible to find out that

$$\nabla u \in H(\mathrm{div}, \Omega) = \{w \in L^2(\Omega)^d : \nabla \cdot w \in L^2(\Omega)\} \tag{2.32}$$

whenever the right-hand side $f \in L^2(\Omega)$.

Denoting

$$X_h = \{v \in H_0^1(\Omega) : v|_K \in P^p(K)\} \tag{2.33}$$

the finite element space, we can define the finite element solution $u_h \in X_h$ satisfying

$$(\nabla u_h, \nabla v) = (f, v), \quad \forall v \in X_h. \tag{2.34}$$

In comparison with a priori analysis, where the convergence of the error with respect to the discretization data is studied and the error bound typically depends on the high derivatives of the unknown exact solution, a posteriori error analysis provides the error bounds depending on the discrete solution itself. There are many techniques for a posteriori error estimates, for overview see e.g. [45].

The goal of this section is to briefly describe the upper bound construction to the error $u_h - u$ by the so called *equilibrated flux reconstruction* technique. The resulting a posteriori error estimate can be viewed as a generalization of the hyper-circle theorem, cf. [39].

Theorem 2.5.1 (Hyper-circle) *Let $u \in H_0^1(\Omega)$ be the exact solution of problem (2.31), $\sigma \in H(\mathrm{div}, \Omega)$ satisfies $f + \nabla \cdot \sigma = 0$ and $v \in H_0^1(\Omega)$ be arbitrary. Then*

$$\|\nabla u - \nabla v\|^2 + \|\sigma - \nabla u\|^2 = \|\nabla v - \sigma\|^2. \tag{2.35}$$

When such a σ is available, then the estimate can be achieved be setting $v = u_h$ and omitting the term $\|\sigma - \nabla u\|^2$, i.e. $\|\nabla u - \nabla u_h\| \leq \|\nabla u_h - \sigma\|$.

Unfortunately, it is not easy to find a suitable $\sigma \in H(\mathrm{div}, \Omega)$ satisfying $f + \nabla \cdot \sigma = 0$ globally. Here, we describe the construction of $\sigma \approx \sigma_h \in H(\mathrm{div}, \Omega)$ that satisfies the relation $f + \nabla \cdot \sigma_h = 0$ in a weaker sense. Let us denote the local Raviart-Thomas space on element K as $\mathrm{RT}(K) = xP^p(K) + (P^p(K))^d$. This space is the usual finite element approximation space to $H(\mathrm{div}, \Omega)$ in the mixed finite element method. For the overview on the mixed finite element method and corresponding polynomial approximations see e.g. [7]. We construct the extension of Raviart-Thomas space to patches ω_a for given vertex a

$$W(\omega_a) = \{w \in H(\mathrm{div}, \omega_a) : w|_K \in \mathrm{RT}(K),\ w|_{\partial\omega_a} \cdot n = 0\}. \tag{2.36}$$

Denoting the space $P_*^p(\omega_a)$ as the space of piece-wise polynomial functions with zero mean value, we can formulate the local *patch-wise* mixed finite element problem: find $\sigma_a \in W(\omega_a)$ and $r_a \in P_*^p(\omega_a)$ such that

$$\begin{aligned} (\sigma_a, v)_{\omega_a} - (r_a, \nabla \cdot v)_{\omega_a} &= (\psi_a \nabla u_h, v)_{\omega_a}, \quad \forall v \in W(\omega_a), \\ (\nabla \cdot \sigma_a, \varphi)_{\omega_a} &= (\nabla \psi_a \cdot \nabla u_h - \psi_a f, \varphi)_{\omega_a}, \quad \forall \varphi \in P_*^p(\omega_a), \end{aligned} \tag{2.37}$$

where ψ_a is the hat function associated with the vertex a and serves as the discrete decomposition of the unity. The final reconstruction σ_h is the sum of all the local contributions σ_a, i.e. $\sigma_h = \sum_a \sigma_a$. Since each of the local contributions $\sigma_a \in H(\mathrm{div}, \Omega)$ if prolongated by zero outside of the patch ω_a then also the complete reconstruction satisfies $\sigma_h \in H(\mathrm{div}, \Omega)$. Moreover, it is possible to show that

$$(f + \nabla \cdot \sigma_h, 1)_K = 0. \tag{2.38}$$

The property (2.38) is usually called the flux equilibration property.

We can derive the error bound using the reconstruction σ_h. Let us assume $v \in H_0^1(\Omega)$. Then

$$(f, v) - (\nabla u_h, \nabla v) = (f + \nabla \cdot \sigma_h, v) + (\sigma_h - \nabla u_h, \nabla v). \tag{2.39}$$

Estimating these terms individually and using the flux equilibration property (2.38) we get

$$(f, v) - (\nabla u_h, \nabla v) \leq \sum_K (C_P h_K \|f + \nabla \cdot \sigma_h\|_K + \|\sigma_h - \nabla u_h\|_K) \|\nabla v\|_K, \tag{2.40}$$

where C_P is the known constant form the Poincare inequality, cf. [38]. Since

$$\|\nabla u - \nabla u_h\| = \sup_{v \in H_0^1(\Omega)} \frac{(\nabla u - \nabla u_h, \nabla v)}{\|\nabla v\|} = \sup_{v \in H_0^1(\Omega)} \frac{(f, v) - (\nabla u_h, \nabla v)}{\|\nabla v\|}, \tag{2.41}$$

we can conclude that

$$\|\nabla u - \nabla u_h\|^2 \leq \sum_K (C_P h_K \|f + \nabla \cdot \sigma_h\|_K + \|\sigma_h - \nabla u_h\|_K)^2. \tag{2.42}$$

The estimate (2.42) is a guaranteed upper bound and the right-hand side contains only the terms that are fully computable from the discrete solution u_h. Since the construction of σ_h is based on the local problems only, cf. (2.37), the evaluation of this reconstruction σ_h as well as the evaluation of the estimator itself is essentially computationally cheaper than the original problem (2.34).

It is possible to provide the efficiency estimates, i.e. the opposite bounds, for the individual local estimators. These estimates are traditionally done under the assumption that the right-hand side f is a piece-wise polynomial, otherwise the additional oscillation term appears in the estimates. Denoting by $\lesssim$ the inequality up to some fixed constant that does not depend on the exact solution u nor the discrete solution u_h nor the mesh-size h, it is possible to derive following local efficiency estimates

$$\begin{aligned} h_K \|f + \nabla \cdot \sigma_h\|_K &\lesssim \|\nabla u - \nabla u_h\|_{\omega_K}, \\ \|\sigma_h - \nabla u_h\|_K &\lesssim \|\nabla u - \nabla u_h\|_{\omega_K}, \end{aligned} \tag{2.43}$$

see e.g. [21]. The proofs are quite technical and therefore they are skipped in this overview. These estimates (2.43) show that large local estimators correspond to large local contributions to the complete error. This property is important for the identifications of the source of the error in possible adaptive strategies. Unfortunately, the locality grows from elements K to patches ω_K.

2.6 Overview of Chapter 3: Linear unsteady singularly perturbed convection-diffusion problems

Chapter 3 is based on the paper *An optimal uniform a priori error estimate for an unsteady singularly perturbed problem* published in International Journal of Numerical Analysis and Modeling in 2014, [48].

The paper deals with the numerical analysis of unsteady singularly perturbed convection-diffusion problems on a square $\Omega = (0,1)^2$

$$u' - \varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{in } \Omega \times (0,T) \tag{2.44}$$

with homogeneous Dirichlet boundary condition and corresponding initial condition. The paper presents an optimal a priori error estimate for any general mesh-adapted space discretization and discontinuous Galerkin time discretization.

The paper [48] assumes a singularly perturbed case, where the diffusion coefficient ε is small in comparison to the rest of the data. The goal of the paper is to derive error estimates for mesh-adapted spatial methods in combination with the discontinuous Galerkin method in time that are independent of ε.

2.6.1 Discretization

A possible remedy comes from two different sources: high adaptation of the meshes around the layers (Shishkin meshes, Bakhvalov meshes, etc.) and stabilizations of the method (SUPG, local projection stabilization, etc.), see e.g. [41]. Both approaches are very often used together. The paper assumes the layer-adaptated

S-type meshes in combination with any consistent discretization method either stabilized or not.

The construction of the mesh in each direction is similar. Therefore we describe them in x direction only. Let us assume increasing and differentiable generating function ϕ satisfying $\phi(0) = 0$ and $\phi(1/2) = \ln(N)$, where $N + 1$ is number of discretization nodes in x direction including boundaries. Then the partition nodes x_i can be defined by

$$x_i = \frac{2i}{N}\left(1 - \frac{\sigma\varepsilon}{\beta_1}\phi\left(\frac{1}{2}\right)\right), \quad \forall i = 0, \ldots, N/2 \tag{2.45}$$

$$x_i = 1 - \frac{\sigma\varepsilon}{\beta_1}\phi\left(\frac{N-i}{N}\right), \quad \forall i = N/2, \ldots, N, \tag{2.46}$$

where $b = (\beta_1, \beta_2)$ and $\sigma \geq 5/2$. For instance, classical Shishkin mesh coresponds to the choice $\phi(s) = 2\ln(N)s$ and the choice $\phi(s) = -\ln(1 - 2s(1 - N^{-1}))$ coresponds to Bakhvalov-type mesh. Such an approach leads to the ε-uniform spatial error estimates even with respect to resulting norms of the exact solution. For the detail see e.g. [41].

The discretization in space is made with the aid of conforming bilinear space V_N, bilinear form $a_{st}(.,.)$ representing the discretization of the spatial terms from (2.44) and corresponding right-hand side f_{st}

$$(u', v) + a_{st}(u, v) = (f_{st}, v), \quad \forall v \in V_N. \tag{2.47}$$

The time discretization is made using the discontinuous Galerkin discretization described in Section 2.2.1, i.e. discrete solution $U \in V_N^\tau$ satisfies

$$\int_{I_m} (U', v) + a_{st}(U, v)\mathrm{d}t + (\{U\}_{m-1}, v_+^{m-1}) = \int_{I_m} (f_{st}, v)\mathrm{d}t \quad \forall v \in V_N^\tau, \tag{2.48}$$

where

$$V_N^\tau = \{v \in L^2(0, T, V_N) : v|_{I_m} \in P^q(I_m, V_N)\}. \tag{2.49}$$

Let us denote right Radau quadrature on $q + 1$ quadrature nodes

$$\int_{I_m} f(t)\mathrm{d}t \approx Q^m[f]. \tag{2.50}$$

Assuming for simplicity that f or f_{st}, respectively, is a polynomial in time of the same degree as the discrete solution we can replace all the integrals in (2.48) by the right Radau quadratures, since all the terms in (2.48) are linear, i.e.

$$Q^m[(U', v)] + Q^m[a_{st}(U, v)] + (\{U\}_{m-1}, v_+^{m-1}) = Q^m[(f_{st}, v)], \quad \forall v \in V_N^\tau. \tag{2.51}$$

2.6.2 Error analysis

We may apply a similar technique of the proof as in Section 2.4. We design a suitable projection π and divide the error into projection part $\eta = \pi u - u$ and $\xi = U - \pi u \in V_N^\tau$. Then the error equation is as follows

$$\begin{aligned} Q^m[(\xi', v)] + Q^m[a_{st}(\xi, v)] + (\{\xi\}_{m-1}, v_+^{m-1}) = &-Q^m[a_{st}(\eta, v)] \\ &- Q^m[(\eta', v)] - (\{\eta\}_{m-1}, v_+^{m-1}) \end{aligned} \tag{2.52}$$

There are two sources of difficulties in the error analysis comparing with the analysis presented in Section 2.4. The first difficulty is that it is not possible to

provide ellipticity and continuity estimates of $a_{st}(.,.)$ in any norm in such a way that the constants in these estimates would be independent of ε and N. The second difficulty is that L^2-orthogonal projection on V_N that is essentially involved in the definition of space-time projection π in Section 2.4 is very unsuitable for deriving accurate error estimates with respect to space variables for this specific problem, see e.g. [32], where suboptimal error analysis is presented due to this fact.

To overcome these difficulties, the projection π is designed differently to respect the Runge-Kutta nature of the discontinuous Galerkin method in time, cf. Section 2.2.2, and with the aid of classical Ritz projection in space, namely $\pi = P^\tau R_N$, where P^τ is the Lagrange interpolation operator on right Radau quadrature nodes and $R_N : H_0^1(\Omega) \to V_N$ is the Ritz projection satisfying

$$a_{st}(R_N u - u, v) = 0, \quad \forall v \in V_N. \tag{2.53}$$

Then it is possible to show that

$$\sup_{I_m} \|\pi u - u\| \leq C(\tau^{q+1} + g(N)), \tag{2.54}$$

where the constant C is completely independent of ε even with respect to the derivatives of the exact solution u and the term $g(N)$ depends on the choice of the mesh adaptation and the stabilization, e.g. $g(N) = N^{-2}\ln^2(N)$ for the Shishkin mesh or $g(N) = N^{-2}$ for the Bakhvalov mesh when the classical bilinear finite element method without any stabilization is used.

The advantage of this projection π described above is that the energy term $Q^m[a_{st}(\eta, v)]$ vanishes in (2.52) and it is only necessary to deal with the terms on the second row of the right hand side of (2.52). Following estimate is derived for these terms in the paper [48] or in Chapter 3

$$Q^m[(\eta', v)] + (\{\eta\}_{m-1}, v_+^{m-1}) \leq \tau C(\tau^{q+1} + g(N)) \sup_{I_m} \|v\|. \tag{2.55}$$

2.6.3 Estimates inside of intervals I_m

Since the estimate (2.55) contains the supremum over I_m, we need to handle this supremum term which represents a significant difficulty in comparison with the basic approach described in Section 2.4, where only the nodal values need to be handled. Following paper [2], it can be shown that

$$2Q^m[(\xi', \tilde{\xi})] + 2(\{\xi\}_{m-1}, \tilde{\xi}_+^{m-1}) = \|\xi_-^m\|^2 + \frac{1}{\tau} Q^m[\|\tilde{\xi}\|^2], \tag{2.56}$$

where

$$\tilde{\xi} = P^\tau \left(\frac{\tau \xi(t)}{t - t_{m-1}} \right) \in V_N^\tau. \tag{2.57}$$

Moreover, it is possible to show that

$$0 \leq Q^m[a_{st}(\xi, \xi)] \leq Q^m[a_{st}(\xi, \tilde{\xi})]. \tag{2.58}$$

Since the norms

$$\frac{1}{\tau} Q^m[\|\tilde{\xi}\|^2], \qquad \sup_{I_m} \|\xi\|^2, \qquad \sup_{I_m} \|\tilde{\xi}\|^2 \tag{2.59}$$

are equivalent, we can apply these relations to derive the error estimate. The details are shown in the paper [48] or in Chapter 3.

The interesting question arise: Why the choice of the test function $\tilde{\xi}$ defined in (2.57) gives such a nice result (2.56)? The answer can be found in the connection between Runge-Kutta methods and discontinuous Galerkin methods described in Section 2.2.2 and in the classical analysis for the error estimates of the inner stages of RK. For the overview see e.g. [30], where the results from the original paper [13] are presented.

2.7 Overview of Chapter 4: Semilinear unsteady singularly perturbed convection-diffusion problems

Chapter 4 is based on the paper *A priori diffusion-uniform error estimates for nonlinear singularly perturbed problems: BDF2, midpoint and time DG* published in Mathematical Modelling and Numerical Analysis in 2017, [34].

The paper deals with the numerical analysis of unsteady singularly perturbed semilinear convection-diffusion problems

$$u' - \varepsilon\Delta u + \nabla \cdot f(u) = g \quad \text{in } \Omega \times (0,T) \tag{2.60}$$

with homogeneous Dirichlet boundary condition and corresponding initial condition. The paper presents a priori error estimates for discontinuous Galerkin space discretization in combination with either the second order backward differentiation formula (BDF2) or the midpoint rule or the discontinuous Galerkin method in time.

Once again, we are mostly interested in the singularly perturbed situation, where the parameter ε is small. Since the problem (2.60) is nonlinear, it represents even more difficult challenge then the linear problem from Section 2.6 respectively from [48].

2.7.1 Discretization

We can apply the same notation as in Section 2.1.1 and Section 2.1.2. The space discretization is made with the aid of the discontinuous Galerkin method. The diffusion term $-\Delta u$ is discretized by SIPG formulation described in Section 2.3. The discretization of the convective term $\nabla \cdot f(u)$ is made similarly as in the finite volume method

$$(\nabla \cdot f(u), v) \approx b_h(u,v) = -\sum_K (f(u), \nabla v)_K + \sum_e (H(u_L, u_R, n), [v])_e, \tag{2.61}$$

where the flux $f(u) \cdot n$ is approximated on the edge e by the value $H(u_L, u_R, n)$ called *numerical flux*. We assume that the numerical flux can be arbitrary function satisfying following assumptions

- $H(u,v,n)$ is Lipschitz continuous, i.e.

$$|H(u,v,n) - H(\bar{u},\bar{v},n)| \le C(|u-\bar{u}| + |v-\bar{v}|), \tag{2.62}$$

- $H(u,v,n)$ is consistent, i.e.

$$H(u,u,n) = f(u) \cdot n, \tag{2.63}$$

- $H(u,v,n)$ is conservative, i.e.

$$H(u,v,n) = -H(v,u,-n), \tag{2.64}$$

- $H(u,v,n)$ is E-flux, i.e.

$$H(u,v,n) - f(q)\cdot n \geq 0, \quad \forall q \ \text{ between } \ u,v. \tag{2.65}$$

We shall point out that every monotone numerical flux is E-flux.

The semidiscrete formulation of problem (2.60) is

$$(u_h', v) + \varepsilon A_h(u_h, v) + b_h(u_h, v) = (g, v), \quad \forall v \in X_h. \tag{2.66}$$

This problem is discretized in time by either BDF2

$$\left(\frac{3}{2}U^m - 2U^{m-1} + \frac{1}{2}U^{m-2}, v\right) + \tau\varepsilon A_h(U^m, v) + \tau b_h(U^m, v) = \tau(g, v) \quad \forall v \in X_h, \tag{2.67}$$

where the starting value U^1 is obtained by the backward Euler method, or by the midpoint rule

$$(U^m - U^{m-1}, v_h) + \frac{\tau}{2}\varepsilon A_h(U^m + U^{m-1}, v_h) + \tau b_h\left(\frac{U^m + U^{m-1}}{2}\right) = \tau(g(t_{m-1} + \tau/2), v_h), \quad \forall v_h \in X_h \tag{2.68}$$

or by the quadrature version of the discontinuous Galerkin method

$$\int_{I_m} (U', v) + \varepsilon A_h(U, v)\mathrm{d}t + Q^m[b_h(U, v)] + (\{U\}_{m-1}, v_+^{m-1}) = Q^m[(g, v)], \quad \forall v_h \in X_h^\tau, \tag{2.69}$$

where

$$X_h^\tau = \{v \in L^2(0, T, X_h) : v|_{I_m} \in P^q(I_m, X_h)\} \tag{2.70}$$

and the right Radau quadrature $Q^m[.]$ is defined in Section 2.6.1.

2.7.2 Error analysis

The error analysis follows the idea from the paper [33] following the results from the paper [51]. The complete description of the idea is quite long and very technical. Here, we summarize the most important steps.

The resulting nonlinear form $b_h(.,.)$ with Lipschitz continuous, consistent and conservative numerical fluxes with the E-flux property satisfies following important estimate

$$b_h(v_h, v_h - \Pi u) - b_h(u, v_h - \Pi u) \leq C\left(1 + \frac{\|v_h - u\|^2_{L^\infty(\Omega)}}{h^2}\right)(h^{2p+1} + \|v_h - \Pi u\|^2), \tag{2.71}$$

where Π is L^2-orthogonal projection on X_h, u is any sufficiently regular function and $v_h \in X_h$, cf. [33]. Difficulties come from the term $\|v_h - u\|^2_{L^\infty(\Omega)}/h^2$, where v_h is typically chosen as the discrete solution U or some term directly derived from U . If it is possible to estimate a priori the error as $\|U - u\|_{L^\infty(\Omega)} = O(h)$, then standard application of the technique will give the desired error estimate that is usually much smaller then the considered bound $O(h)$, typically it is $\|U - u\|^2 \leq C(h^{2p+1} + \tau^{2q+2})$, where $q = 1$ for BDF2 and the midpoint rule and q is the degree of the polynomial approximation in time for the discontinuous Galerkin method. Unfortunately, it is

not easy to prove the error bound $O(h)$ a priori, since the error is the object of investigation and is unknown.

This problem is solved by the continuous mathematical induction, cf [33]. Let us assume that the discretization parameters h, p, τ and s are chosen in such a way that

$$\|U-u\|^2 \leq C(h^{2p+1}+\tau^{2q+2}) \quad \Longrightarrow \quad \|U-u\|_{L^\infty(\Omega)} \leq \frac{h}{2}. \tag{2.72}$$

If the error is represented by a continuous function and if the error is at some point $t=t_*$ sufficiently small, e.g. $\|U-u\|_{L^\infty(\Omega)}=h/2$, then it takes some time $\delta>0$ to grow the error over the bound h. Then it is possible to avoid the term $\|U-u\|^2_{L^\infty(\Omega)}/h^2$ on interval $[t_*,t_*+\delta]$ in the estimate (2.71) and it is possible to derive the desired error estimate $\|U-u\|^2 \leq C(h^{2p+1}+\tau^{2q+2})$ on $[t_*,t_*+\delta]$ by rather standard technique, where the constant is independent of ε. Moreover, it is possible to see that (2.72) implies $\|U-u\|_{L^\infty(\Omega)} \leq \frac{h}{2}$ at a new time $t=t_*+\delta$. Since the continuity of the error holds on the bounded interval $[0,T]$, i.e. on a compact set, there exists a minimal finite δ necessary for such a grow and we can deplete the set $[0,T]$ in a finite number of steps. It should be pointed out that the starting error in the initial condition is inherently small.

Alternatively, the concept of the continuous mathematical induction can be replaced by the argument that the error under the assumption (2.72) cannot hit the value $\|U-u\|_{L^\infty(\Omega)}=h$ and assuming the error evolves continuously and is started from the small initial condition error it is not possible to grow over h and therefore the square of the error behaves as $O(h^{2p+1}+\tau^{2q+2})$.

2.7.3 Discrete solution continuation

Since we assume that the exact solution u is continuous in time, the aim of this section is to describe how to reconstruct the discrete solution that is defined nodal-wise as U^m in the case of BDF2 and the midpoint rule and interval-wise (element-wise) as $U|_{I_m}$ in the case of the time discontinuous Galerkin, as a continuous function $U(t)$ that corresponds to the error in the nodal point, i.e. $U(t_m)=U^m$ or $U(t_m)=U^m_-$.

The idea of the construction of the nodal-wise defined solution as a continuous function can be found in [33], where the backward Euler method is discussed. Let us assume that the continuation is well defined on the interval $[0,t_{m-1}]$ and the goal is to define the continuation on the next time interval $(t_{m-1},t_m]$. Then the value of $U(t_{m-1}+s)$, where $s\in(0,\tau]$, is defined as the discrete solution for the given method by replacing the step-size τ by the new step-size s. Still, it remains to prove a number of technical results that imply that the resulting continuation $U(t_{m-1}+s)$ exists uniquely for arbitrary $s\in(t_{m-1},t_m]$ and that the resulting function $U(t)$ is really continuous. These results are described in detail in the paper [34] or Chapter 4. It should be pointed out that the BDF2 analysis also applies the stability theory for the multistep methods with non-equidistant time steps, see e.g. [30].

The time discontinuous Galerkin discretization is more complicated, since the solution at the final time of each interval depends on the corresponding inner stages, see Section 2.2.2, and the continuation should respect this fact. Let us assume that the continuation is constructed to the time level $y=t_{m-1}$. The approach from [34] defines the continuation on $(t_{m-1},t_m]$ as a set of functions on I_m

$$\{U_y\}_{y\in(t_{m-1},t_m]} \subset X^\tau_h. \tag{2.73}$$

Denoting $s\in(0,\tau]$ such that $y=t_{m-1}+s$ and denoting Radau quadrature rescaled from interval I_m to the new interval (t_{m-1},y) as $Q^m_s[.]$, then each function U_y of

the continuation is defined on I_m as

$$\int_{t_{m-1}}^{y} (U_y', v) + \varepsilon A_h(U_y, v)\mathrm{d}t + Q_s^m[b_h(U_y, v)] + (\{U\}_{m-1}, v_+^{m-1}) \\ = Q_s^m[(f, v)], \quad \forall v_h \in X_h^\tau. \tag{2.74}$$

The resulting *continuity* is described by the relations

$$\sup_{(t_{m-1}, \min(y,\bar{y}))} \|U_y - U_{\bar{y}}\| \to 0, \text{ as } |y - \bar{y}| \to 0, \tag{2.75}$$

$$\sup_{(t_{m-1}, y)} \|U_y - U_-^{m-1}\| \to 0, \text{ as } y \to t_{m-1}+ . \tag{2.76}$$

The proof of this continuity with respect to y is very technical and the details are presented in the paper [34] or Chapter 4.

2.8 Overview of Chapter 5: Nonlinear unsteady convection-diffusion problems in time-dependent domains

Chapter 5 is based on the paper *Stability of the ALE space-time discontinuous Galerkin method for nonlinear convection-diffusion problems in time-dependent domains* published in Mathematical Modelling and Numerical Analysis in 2018, [6].

The paper deals with the numerical analysis of unsteady nonlinear convection-diffusion problems

$$u' - \nabla \cdot (\beta(u)\nabla u) + \nabla \cdot f(u) = g \tag{2.77}$$

with Dirichlet boundary conditions and corresponding initial condition. The nonlinearity in the diffusion term described by the function $\beta(u)$ is considered bounded and Lipschitz continuous, i.e.

$$\beta : \mathbb{R} \to [\beta_0, \beta_1], \quad 0 < \beta_0 \le \beta_1 < \infty, \tag{2.78}$$

$$|\beta(u) - \beta(v)| \le C|u - v|. \tag{2.79}$$

The paper [6] does not assume the singularly perturbed case, where $\beta_0 \to 0$, but the dependence of the derived results on β_0 is tracked for further investigations.

In comparison with previous sections, the problem (2.77) is not considered in a fixed space-time cylinder $\Omega \times (0, T)$, but in an *evolving* space-time cylinder $\Omega_t \times (0, T)$, where the space domain Ω_t depends smoothly on time t. The goal of the paper [6] is to present a stability bound for discontinuous Galerkin space-time discretization.

2.8.1 Arbitrary Lagrangian-Eulerian description

The evolution of the domain Ω_t is described by a one-to-one mapping $\mathcal{A}_t : \Omega_{\text{ref}} \to \Omega_t$ which maps the point $X \in \Omega_{\text{ref}}$ onto the point $x \in \Omega_t$, i.e. $x = \mathcal{A}_t(X) \in \Omega_t$. Collecting these mappings for $t \in [0, T]$ we get the so called ALE mapping $\mathcal{A}$. Although such a mapping is assumed individually for each time interval I_m in the paper [6], we consider here only a single ALE mapping over all time interval $(0, T)$. We also assume that the evolution of the domain as well as the ALE mapping $\mathcal{A}$ is independent of the solution u of problem (2.77). We assume that the evolution of the domain is smooth and that the ALE mapping $\mathcal{A}$ and its inverse $\mathcal{A}^{-1}$satisfies

$$\mathcal{A} \in W^{1,\infty}(0, T, W^{1,\infty}(\Omega_{\text{ref}})), \\ \mathcal{A}^{-1} \in W^{1,\infty}(0, T, W^{1,\infty}(\Omega_t)). \tag{2.80}$$

The important concept in the ALE description is the ALE derivative. The ALE derivative D_t of function $f(x,t)$ is defined as the time derivative of the reference function $\tilde{f}(X,t) = f(\mathcal{A}_t(X),t)$, where $x = \mathcal{A}_t(X)$. By the chain rule we gain

$$D_t f(x,t) = \frac{\partial}{\partial t}\tilde{f}(X,t) = \frac{\mathrm{d}}{\mathrm{d}t} f(\mathcal{A}_t(X)),t) = \nabla f(x,t)\cdot\frac{\partial}{\partial t}\mathcal{A}_t(X) + f'(x,t). \quad (2.81)$$

Denoting the mesh velocity $z(x,t) = \tilde{z}(X,t)$, where $\tilde{z}(X,t) = \frac{\partial}{\partial t}\mathcal{A}_t(X)$, we can rewrite (2.81) as

$$D_t f = z\cdot\nabla f + f'. \quad (2.82)$$

The interpretation of the ALE derivative is the derivative along the ALE curve, where the ALE curve is defined as the evolution of the single point $X \in \Omega_{\mathrm{ref}}$.

Using the ALE derivative, we can reformulate the original problem (2.77) into equivalent problem

$$D_t u - \nabla\cdot(\beta(u)\nabla u) + \nabla\cdot f(u) - z\cdot\nabla u = g. \quad (2.83)$$

2.8.2 Discretization

The aim of this section is to discretize the problem (2.83) by the space-time discontinuous Galerkin method. We apply the notation from Section 2.3 and Section 2.4 on fixed space-time cylinder. Let us assume the discontinuous finite element space

$$\tilde{X}_h = \{\tilde{v} \in L^2(\Omega_{\mathrm{ref}}) : \tilde{v}|_{\tilde{K}} \in P^p(\tilde{K})\} \quad (2.84)$$

on Ω_{ref}. We can define fully discrete space

$$\tilde{X}_h^\tau = \{\tilde{v} \in L^2(0,T,\tilde{X}_h) : v|_{I_m} \in P^q(I_m,\tilde{X}_h)\} \quad (2.85)$$

on the fixed (reference) space-time cylinder. Finally, the fully discrete space X_h^τ on the evolving space-time cylinder is defined as

$$X_h^\tau = \{v : v\circ\mathcal{A} \in \tilde{X}_h^\tau\}. \quad (2.86)$$

Applying the space and time discontinuous Galerkin technique described in Section 2.3 and Section 2.2.1, we arrive to the discrete formulation of problem (2.77)

$$\int_{I_m} (D_t U, v)_t + A_h(U,v,t) + b_h(U,v,t) - (z\cdot\nabla U, v)_t \mathrm{d}t \quad (2.87)$$
$$+(\{U\}_{m-1}, v_+^{m-1})_{t_{m-1}} = \int_{I_m} \ell(v,t)\mathrm{d}t, \quad \forall v \in X_h^\tau,$$

where $(.,.)_t$ denotes the L^2-scalar product on Ω_t. The detailed description of the forms $A_h(.,.,t)$, $b_h(.,.,t)$ and $\ell(.,t)$ can be found in the paper [6] or in Chapter 5.

2.8.3 Stability analysis

The goal of this section is to derive the stability estimate, i.e. the estimate that bounds the discrete solution $U \in X_h^\tau$ in $L^\infty(L^2)$-norm by the data of the problem in suitable norms, i.e. by the initial and boundary conditions and by the right-hand side g. Setting $v = U$ in (2.87) gives after some manipulations

$$\|U_-^m\|_{t_m}^2 - \|U_-^{m-1}\|_{t_{m-1}}^2 + \|\{U\}_{m-1}\|_{t_{m-1}}^2 + \int_{I_m} A_h(U,U,t)\mathrm{d}t \quad (2.88)$$
$$\leq R_t + C\int_{I_m} \|U\|_t^2 \mathrm{d}t,$$

where $\|.\|_t$ denotes the L^2-norm on Ω_t and the term R_t consists of the norms of the boundary condition and the right-hand side. The main difficulty lies in the estimate of the $L^2(L^2)$-norm of the discrete solution U on the right-hand side of (2.88). Since the discrete solution is from the finite dimensional space, it is possible to show that the norms

$$\int_{I_m} \|U\|_t^2 \mathrm{d}t \quad \text{and} \quad \tau \sup_{I_m} \|U\|_t^2 \tag{2.89}$$

are equivalent. For piece-wise constant or piece-wise linear time approximations, i.e. $q = 0, 1$, it is possible to deal with the supremum term directly, since the supremum over I_m is gained only at the endpoints of the interval I_m, see [5]. The polynomial approximations of higher degree need to be treated more carefully.

2.8.4 Discrete characteristic function

Denoting $y \in [t_{m-1}, t_m]$ such that

$$\|U(y)\|_y^2 = \sup_{I_m} \|U\|_t^2, \tag{2.90}$$

the ideal choice of the test function in (2.87) is $v = U\chi_{(0,y)}$, where $\chi_{(0,y)}$ is the characteristic function of interval $(0, y)$. The applications of this test function in (2.87) leads after some manipulations to

$$\|U(y)\|_y^2 - \|U_-^{m-1}\|_{t_{m-1}}^2 + \|\{U\}_{m-1}\|_{t_{m-1}}^2 + \int_{t_{m-1}}^{y} A_h(U, U, t)\mathrm{d}t \tag{2.91}$$
$$\leq R_t + C \int_{t_{m-1}}^{y} \|U\|_t^2 \mathrm{d}t.$$

Then the proof of the stability can be finished by Gronwall lemma.

Unfortunately, this choice of the test function is not possible, since $U\chi_{(0,y)} \notin X_h^\tau$, and it is necessary to construct a discrete approximation of $U\chi_{(0,y)}$ in the space X_h^τ. In the paper [6], the approximation $U\chi_{(0,y)} \approx U_y \in X_h^\tau$ is made with the aid of the *discrete characteristic function* described in [12] for fixed domains. Denoting the corresponding function $\tilde{U} \in \tilde{X}_h^\tau$ to the original function $U \in X_h^\tau$, we can define the discrete characteristic function $\tilde{U}_y \in \tilde{X}_h^\tau$ on the fixed space-time cylinder by

$$\int_{I_m} (\tilde{U}_y, v)_{\mathrm{ref}} \mathrm{d}t = \int_{t_{m-1}}^{y} (\tilde{U}, v)_{\mathrm{ref}}, \quad \forall v \in P^{q-1}(I_m, \tilde{X}_h), \tag{2.92}$$
$$(\tilde{U}_y)_+^{m-1} = (\tilde{U})_+^{m-1}.$$

Then the final discrete characteristic function U_y is defined as the transformation of $\tilde{U}_y$ back to the evolving domain, i.e. $U_y(x, t) = \tilde{U}_y(\mathcal{A}_t^{-1}(x), t) \in X_h^\tau$.

The main properties of this discrete characteristic function $U_y \in X_h^\tau$ is that it behaves similarly as the true characteristic function $U\chi_{(0,y)}$ when applied on the term that corresponds to the discrete time derivative, i.e.

$$2\int_{I_m} (D_t U, U_y)_t \mathrm{d}t + 2(\{U\}_{m-1}, (U_y)_+^{m-1})_{t_{m-1}} \geq \|U(y)\|_y^2 - \|U_-^{m-1}\|_{t_{m-1}}^2 \tag{2.93}$$
$$-C \int_{I_m} \|U\|_t^2 \mathrm{d}t.$$

The application of U_y on all of the other terms in (2.87) is treated with the aid of the following continuity property of $U \to U_y$ proved in the paper [6] or in Chapter

5

$$\int_{I_m} \|U_y\|_t^2 \mathrm{d}t \le C \int_{I_m} \|U\|_t^2 \mathrm{d}t, \tag{2.94}$$
$$\int_{I_m} A_h(U_y, U_y, t)\mathrm{d}t \le C \int_{I_m} A_h(U, U, t)\mathrm{d}t.$$

2.9 Overview of Chapter 6: A posteriori error estimates for nonlinear parabolic problems

Chapter 6 is based on the paper *A posteriori error estimates for higher order space-time Galerkin discretizations of nonlinear parabolic problems* published in SIAM Journal on Numerical Analysis in 2021, [17].

The paper deals with the numerical analysis of unsteady singularly perturbed nonlinear convection-diffusion problems

$$u' - \nabla \cdot \sigma(u, \nabla u) + c(u) = 0 \quad \text{in } \Omega \times (0, T) \tag{2.95}$$

with homogeneous Dirichlet boundary condition and corresponding initial condition u^0. The nonlinearity is supposed to be monotone and continuous.

The paper [17] assumes either conforming or nonconforming Galerkin discretizations in space or time resulting in four different types of discretizations. The goal of the paper [17] is to present a unified a posteriori error analysis based on the equilibrated flux reconstructions for all these Galerkin discretizations.

To simplify forthcoming explanations, we only consider the heat equation instead of (2.95), i.e. $\sigma(u, \nabla u) = \nabla u$ and $c(u) = -f$, and the discontiunous Galerkin time discretization in combination with the classical finite element method.

2.9.1 Continuous problem and its discretization

Let us denote spaces

$$\begin{aligned} X &= L^2(0, T, H_0^1(\Omega)), \\ Y &= \{v \in X : v' \in L^2(0, T, L^2(\Omega))\} \subset C([0, T], L^2(\Omega)), \\ Y^0 &= \{v \in Y : v(0) = u^0\}. \end{aligned} \tag{2.96}$$

Then the weak solution satisfies $u \in Y^0$ and

$$\int_0^T (u', v) + (\nabla u, \nabla v)\mathrm{d}t = \int_0^T (f, v)\mathrm{d}t, \quad \forall v \in X. \tag{2.97}$$

We define X_h^τ in the same way as in Section 2.4. It shall be pointed out that X_h^τ is a very natural approximation space to the space X, but not to the spaces Y or Y^0, since $X_h^\tau \not\subset Y$. The fully discrete solution $U \in X_h^\tau$ satisfies

$$\int_{I_m} (U', v) + (\nabla U, \nabla v)\mathrm{d}t + (\{U\}_{m-1}, v_+^{m-1}) = \int_{I_m} (f, v)\mathrm{d}t, \quad \forall v \in X_h^\tau, \tag{2.98}$$

where $U_-^0 = u^0$.

2.9.2 Discrete solution reconstruction

Similarly in Section 2.5, we need to reconstruct the discrete solution. The spatial reconstruction $\sigma_h^\tau \in L^2(0, T, H(\mathrm{div}, \Omega))$ can be obtained in a similar way as described in Section 2.5 and the precise description can be found in the paper [17] or in Chapter 6.

It remains to reconstruct the discrete solution in time. The exact solution u belongs to Y^0. It is possible to see that any function $v \in X_h^\tau$ belongs to the space Y^0 if and only if v is continuous in time and satisfies the initial condition, i.e. $v(0) = u^0$. Then the reconstruction $R_h^\tau \in Y^0$ can be obtained directly from the discrete solution U

$$R_h^\tau(x,t) = U(x,t) - \{U\}_{m-1}(x) r_m(t), \quad t \in I_m,\ x \in \Omega, \tag{2.99}$$

where r_m are Radau polynomials defined in Lemma 2.2.3, i.e. $r_m \in P^{q+1}(I_m)$, $r_m(t_m) = 0$, $r_m(t_{m-1}) = 1$ and $r_m \perp P^{q-1}(I_m)$. In fact, the reconstruction (2.99) is identical to the reconstruction (2.14).

The resulting reconstruction R_h^τ satisfies $R_h^\tau \in Y^0$ and together with the spatial reconstruction σ_h^τ also satisfies the space-time version of the equilibration property (2.38), i.e.

$$(f - (R_h^\tau)' + \nabla \cdot \sigma_h^\tau, 1)_{K,m} = 0 \tag{2.100}$$

The details of the proof are presented in the paper [17].

2.9.3 Error measure

Inspired by the work [14], we design the error measure $\mathrm{Res}(U)$ as the dual norm of residual. Since the method (2.98) is nonconforming and the formulations for the exact and the discrete solutions differ, we design a common formulation for both these solutions.

Since the space $X_h^\tau \not\subset Y^0$, we design a new space

$$Y^\tau = \{v \in X : v'|_{I_m} \in L^2(I_m, L^2(\Omega))\}. \tag{2.101}$$

The space Y^τ can be considered as the broken Sobolev space with respect to time. Then this space satisfies $Y^0 \subset Y \subset Y^\tau \subset X$ and also $X_h^\tau \subset Y^\tau$. We can exploit these properties to define the extended formulation that covers the formulation for the exact solution (2.95) as well as the formulation for the discrete solution (2.98): find $u \in Y^\tau$ such that

$$\int_{I_m} (u', v) + (\nabla u, \nabla v)\mathrm{d}t + (\{u\}_{m-1}, v_+^{m-1}) = \int_{I_m} (f, v)\mathrm{d}t, \quad \forall v \in Y^\tau. \tag{2.102}$$

It shall be pointed out that the formulation (2.102) has a unique solution in Y^τ and this solution is the exact solution u of the original problem (2.97).

Then the error measure is defined as a dual norm of residual with respect to the extended formulation (2.102)

$$\mathrm{Res}(U) = \sup_{v \in Y^\tau} \frac{1}{\|v\|_{Y^\tau}} \sum_{K,m} (f - U', v)_{K,m} - (\nabla U, \nabla v)_{K,m} - (\{U\}_{m-1}, v_+^{m-1})_K, \tag{2.103}$$

where the norm $\|.\|_{Y^\tau}$ is designed locally and similarly as in [14]

$$\|v\|_{Y^\tau}^2 = \sum_{K,m} \|v\|_{Y^\tau,K,m}^2, \quad \text{where} \quad \|v\|_{Y^\tau,K,m}^2 = \frac{1}{d_{K,m}^2} h_K^2 \|\nabla v\|_{K,m}^2 + \tau^2 \|v'\|_{K,m}^2. \tag{2.104}$$

The norm $\|.\|_{Y^\tau}$ in [17] contains a user dependent local parameter $d_{K,m}$. To simplify the forthcoming exposition, we assume here $d_{K,m} = 1$.

2.9.4 Error estimate

The upper bound can be derived similarly as in Section 2.5. Let us assume $v \in Y^\tau$. Then

$$\begin{aligned}&\sum_{K,m}(f-U',v)_{K,m}-(\nabla U,\nabla v)_{K,m}-(\{U\}_{m-1},v_+^{m-1})_K \\ &=\sum_{K,m}(f-(R_h^\tau)'+\nabla\cdot\sigma_h^\tau,v)_{K,m}+\sum_{K,m}(\sigma_h^\tau-\nabla U,\nabla v)_{K,m} \\ &\quad+\sum_{K,m}((R_h^\tau)'-U',v)_{K,m}-(\{U\}_{m-1},v_+^{m-1})_K.\end{aligned} \tag{2.105}$$

Estimation of these terms individually with the aid of (2.100) leads to

$$\begin{aligned}&(f-(R_h^\tau)'+\nabla\cdot\sigma_h^\tau,v)_{K,m}\le C_P\|f-(R_h^\tau)'+\nabla\cdot\sigma_h^\tau\|_{K,m}\|v\|_{Y^\tau,K,m}, \\ &(\sigma_h^\tau-\nabla U,\nabla v)_{K,m}\le\|\sigma_h^\tau-\nabla U\|_{K,m}\|\nabla v\|_{K,m}, \\ &((R_h^\tau)'-U',v)_{K,m}-(\{U\}_{m-1},v_+^{m-1})_K\le\|R_h^\tau-U\|_{K,m}\|v'\|_{K,m},\end{aligned} \tag{2.106}$$

where C_P is again the constant from Poincare inequality, cf. [38]. Application of these estimates together and denoting the individual local estimators from (2.106) as

$$\begin{aligned}\eta_{R,K,m}&=C_P\|f-(R_h^\tau)'+\nabla\cdot\sigma_h^\tau\|_{K,m}, \\ \eta_{S,K,m}&=\frac{1}{h_K}\|\sigma_h^\tau-\nabla U\|_{K,m}, \\ \eta_{T,K,m}&=\frac{1}{\tau}\|R_h^\tau-U\|_{K,m}\end{aligned} \tag{2.107}$$

gives a posteriori error estimate

$$\mathrm{Res}(U)^2\le\eta^2=\sum_{K,m}\left(\eta_{R,K,m}+(\eta_{S,K,m}^2+\eta_{T,K,m}^2)^{1/2}\right)^2. \tag{2.108}$$

2.9.5 Efficiency estimates

Similarly as in Section 2.5, we can derive local efficiency estimates for the individual error estimators $\eta_{R,K,m}$, $\eta_{S,K,m}$ and $\eta_{T,K,m}$. We assume traditionally that f is a piece-wise polynomial function. Again, we denote by $\lesssim$ the inequality up to constant independent of the exact solution u, the discrete solution U, mesh-size h and step-size τ.

To be able to provide the efficiency estimates locally, we need to define a local version of the error norm $\mathrm{Res}(U)$. Since the error norm is dual norm of residual of the extended formulation, i.e. certain supremum term over all functions $v \in Y^\tau$, see (2.103), we define local versions of the error norm as

$$\begin{aligned}\mathrm{Res}_{M,m}(U)=\sup_{v\in Y_{M,m}^\tau}\frac{1}{\|v\|_{Y_{M,m}^\tau}}&\sum_{K,m}(f-U',v)_{K,m} \\ &-(\nabla U,\nabla v)_{K,m}-(\{U\}_{m-1},v_+^{m-1})_K,\end{aligned} \tag{2.109}$$

where $Y_{M,m}^\tau \subset Y^\tau$ is a space consisting from functions supported by $\overline{M\times I_m}$, where M is some collection of elements K.

The efficiency estimates for $\eta_{R,K,m}$ and $\eta_{S,K,m}$ can be derived by generalizing the stationary technique, see [21] and [45]. The proof of the efficiency estimate for

$\eta_{T,K,m}$ is made more directly with the aid of a suitable test function, for the details see [17]. The resulting efficiency estimates are following

$$\begin{aligned}\eta_{R,K,m} &= C_P\|f-(R_h^\tau)'+\nabla\cdot\sigma_h^\tau\|_{K,m} \lesssim \mathrm{Res}_{\omega_K,m}(U), \\ \eta_{S,K,m} &= \frac{1}{h_K}\|\sigma_h^\tau-\nabla U\|_{K,m} \lesssim \mathrm{Res}_{\omega_K,m}(U), \\ \eta_{T,K,m} &= \frac{1}{\tau}\|R_h^\tau-U\|_{K,m} \lesssim \mathrm{Res}_{K,m}(U).\end{aligned} \tag{2.110}$$

Since

$$\sum_{K,m}\mathrm{Res}_{K,m}(U) \le \sum_{K,m}\mathrm{Res}_{\omega_K,m}(U) \lesssim \mathrm{Res}(U), \tag{2.111}$$

we can derive from (2.110) the global efficiency estimate

$$\eta = \sum_{K,m}\left(\eta_{R,K,m}+(\eta_{S,K,m}^2+\eta_{T,K,m}^2)^{1/2}\right) \lesssim \mathrm{Res}(U). \tag{2.112}$$

2.10 Overview of Chapter 7: Polynomial robustness of efficiency estimates

Chapter 7 is based on the paper *On polynomial robustness of flux reconstructions* published in Appl. Math. in 2020, [47].

The paper deals with the numerical analysis of convection-diffusion-reaction problems

$$-\Delta u + b\cdot\nabla u + cu = f \quad \text{in } \Omega \tag{2.113}$$

with homogeneous Dirichlet boundary condition. The problem is discretized by the standard finite element method. A posteriori error estimate, where the flux reconstructions are designed element-wise, is derived. The main result of the paper show that the efficiency constant of the flux reconstruction in 1D ($d=1$) depends on the discretization polynomial degree as $p^{1/2}$ at most. The main contribution behind this paper lies in the application of the reconstruction developed for the time discretization in [17] for the space discretization as well.

To simplify forthcoming explanations, we only consider the Poisson equation instead of (2.113), i.e. $b=0$ and $c=0$.

2.10.1 Discretization and upper bound

We descretize the problem (2.113) by the standard finite element method. We can apply the notation from Section 2.1.1. The finite element space is defined as

$$X_h = \{v\in H_0^1(\Omega) : v|_K \in P^p(K)\} \tag{2.114}$$

and we can formulate the discrete problem: find $u_h\in X_h$ such that

$$(\nabla u_h,\nabla v) = (f,v), \quad \forall v_h\in X_h. \tag{2.115}$$

In contrary to Section 2.5, we compose the reconstruction $\sigma_h\in H(\mathrm{div},\Omega)$ from the local element-wise information, i.e. we define $\sigma_h|_K\in \mathrm{RT}(K)$ such that

$$\begin{aligned}\sigma_h|_e\cdot n &= \langle\nabla u_h\rangle|_e\cdot n, \\ (\sigma_h,w)_K &= (\nabla u_h,w)_K, \quad \forall w\in P^{p-1}(K)^d.\end{aligned} \tag{2.116}$$

The resulting global function σ_h is in $H(\mathrm{div}, \Omega)$, since the normal component of σ_h is continuous across the edges. Moreover, σ_h is equilibrated in generalized sense

$$(f + \nabla \cdot \sigma_h, v) = 0, \quad \forall v \in X_h. \tag{2.117}$$

The advantage of the reconstruction defined by (2.116) in comparison with the reconstruction defined in Section 2.5 is its simplicity that enables to evaluate the reconstruction directly without solving an artificial mixed finite element problem on patches ω_a. It shall be pointed out that the relations from (2.116) correspond to the classical (natural) degrees of freedom for $\mathrm{RT}(K)$, see e.g. [7].

Instead of Poincare inequality applied in Section 2.5, we need a more accurate estimate

$$\inf_{v_h \in X_h} \|v - v_h\|_K \leq C_{Fl} \frac{h_K}{p} \|\nabla v\|_K \tag{2.118}$$

that holds for any function $v \in H_0^1(K)$, see e.g. [4]. The constant C_{Fl} is unknown in general, but it can be determined in some special cases. E.g., it is possible to take

$$C_{Fl} = \frac{p}{\sqrt{(2p+3)(2p-1)}} \tag{2.119}$$

in 1D ($d = 1$), see [47] or Chapter 7.

Denoting local estimators

$$\begin{aligned} \eta_{R,K} &= C_{Fl} \frac{h_K}{p} \|f + \nabla \cdot \sigma_h\|_K, \\ \eta_{F,K} &= \|\sigma_h - \nabla u_h\|_K \end{aligned} \tag{2.120}$$

and applying (2.117) together with (2.118) imply a posteriori error estimate

$$\|\nabla u - \nabla u_h\|^2 = \mathrm{Res}(u_h)^2 \leq \eta^2 = \sum_K (\eta_{R,K} + \eta_{F,K})^2. \tag{2.121}$$

The idea of the proof is similar as in Section 2.5.

2.10.2 Efficiency

We derive local efficiency estimates for the individual error estimators $\eta_{R,K}$ and $\eta_{F,K}$ in 1D ($d = 1$). We traditionally assume that $f \in X_h$, similarly as in Section 2.5. Again, we denote by $\lesssim$ the inequality up to constant independent of the exact solution u, the discrete solution u_h and mesh-size h. Since we are also interested in the polynomial dependence of this constant, we assume that this constant is also independent of the discretization polynomial degree p and denote this polynomial dependence separately.

Similarly as in [17], we define local errors

$$\mathrm{Res}_M(u_h) = \sup_{v \in H_0^1(\Omega),\, \mathrm{supp}(v) \subset M} \frac{(f, v) - (\nabla u_h, \nabla v)}{\|\nabla v\|}, \tag{2.122}$$

where M is some collection of elements K. Similarly as in [17], it is possible to show that

$$\sum_K \mathrm{Res}_K(u_h) \leq \sum_K \mathrm{Res}_{\omega_K}(u_h) \lesssim \mathrm{Res}(u_h). \tag{2.123}$$

Moreover, it is possible to see that the reconstruction defined by (2.116) in 1D can be equivalently rewritten on element $K = [a, b]$

$$\sigma_h|_K = \nabla u_h + (\langle \nabla u_h \rangle(a) - \nabla u_h(a))r_a + (\langle \nabla u_h \rangle(b) - \nabla u_h(b))r_b, \tag{2.124}$$

where the values of $\nabla u_h(a)$ and $\nabla u_h(b)$ are taken from inside of K and $r_a, r_b \in P^{p+1}$ are Radau orthogonal polynomials on K oriented either to the left endpoint a or the right endpoint b. Comparing with the reconstruction (2.14), we can see that the reconstruction of σ_h is defined according to the similar principle, but assumes the jump term on both sides of the interval K.

The efficiency of $\eta_{F,K}$ can be proved by similar argument as in the proof of efficiency of $\eta_{T,K,m}$ in [17]. The advantage of the directness of the proof enables to track the dependence on the polynomial degree

$$\eta_{F,K} = \|\sigma_h - \nabla u_h\|_K \lesssim p^{1/2} \operatorname{Res}_{\omega_K}(u_h). \tag{2.125}$$

The proof is rather technical and therefore it is skipped here. The details can be found in [47] or in Chapter 7. This estimate can be applied for the proof of the efficiency of $\eta_{R,K}$, where it is possible to show that the polynomial dependence is the same as in $\eta_{F,K}$

$$\eta_{R,K} = C_{Fl}\frac{h_K}{p}\|f + \nabla \cdot \sigma_h\|_K \lesssim p^{1/2} \operatorname{Res}_{\omega_K}(u_h). \tag{2.126}$$

Again, the proof is quite technical and the details can be found in [47] or in Chapter 7.

The estimate of $\eta_{R,K}$ is quite interesting, since usually the authors in the literature are focused in the efficiency of $\eta_{F,K}$ only. The problem with traditional concept of the term $\eta_{R,K}$ is that only standard Poincare inequality (or a similar inequality like Friedrichs inequality etc.) is applied. This enables to determine the constant C_{Fl} as Poincare constant C_P that is known in standard situations, e.g. on convex domains. On the other hand, classical Poincare inequality only contains the term $C_P h_K$ instead of $C_{Fl}\frac{h_K}{p}$. Avoiding the $1/p$ term seems to lead to suboptimal efficiency analysis with respect to the polynomial degree p.

Bibliography

[1] N. Ahmed, G. Matthies, L. Tobiska, and H. Xie. Discontinuous Galerkin time stepping with local projection stabilization for transient convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Eng.*, 200(21-22):1747–1756, 2011.

[2] G Akrivis and Ch. Makridakis. Galerkin time-stepping methods for nonlinear parabolic equations. *M2AN, Math. Model. Numer. Anal.*, 38(2):261–289, 2004.

[3] G. Akrivis, Ch. Makridakis, and R. H. Nochetto. Galerkin and Runge-Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence. *Numer. Math.*, 118(3):429–456, 2011.

[4] I. Babuška and M. Suri. The h-p version of the finite element method with quasiuniform meshes. *RAIRO, Modélisation Math. Anal. Numér.*, 21:199–238, 1987.

[5] M. Balázsová and M. Feistauer. On the stability of the ALE space-time discontinuous Galerkin method for nonlinear convection-diffusion problems in time-dependent domains. *Appl. Math., Praha*, 60(5):501–526, 2015.

[6] M. Balázsová, M. Feistauer, and M. Vlasák. Stability of the ALE space-time discontinuous Galerkin method for nonlinear convection-diffusion problems in time-dependent domains. *ESAIM, Math. Model. Numer. Anal.*, 52(6):2327–2356, 2018.

[7] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*, volume 44 of *Springer Ser. Comput. Math.* Berlin: Springer, 2013.

[8] A. Bonito, I. Kyza, and R. H. Nochetto. Time-discrete higher order ALE formulations: a priori error analysis. *Numer. Math.*, 125(2):225–257, 2013.

[9] Rea Bonito, Irene Kyza, and Ricardo H. Nochetto. Time-discrete higher-order ALE formulations: stability. *SIAM J. Numer. Anal.*, 51(1):577–604, 2013.

[10] D. Braess, V. Pillwein, and J. Schöberl. Equilibrated residual error estimates are p-robust. *Comput. Methods Appl. Mech. Eng.*, 198(13-14):1189–1197, 2009.

[11] P. Brenner, M. Crouzeix, and V. Thomee. Single step methods for inhomogeneous linear differential equations in Banach space. *RAIRO, Anal. Numér.*, 16:5–26, 1982.

[12] K. Chrysafinos and N. J. Walkington. Error estimates for the discontinuous Galerkin methods for parabolic equations. *SIAM J. Numer. Anal.*, 44(1):349–366, 2006.

[13] K. Dekker. On the iteration error in algebraically stable runge-kutta methods. *Report NW 138/82*, Math. Centrum, Amsterdam, 1982.

[14] V. Dolejší, A. Ern, and M. Vohralík. A framework for robust a posteriori error control in unsteady nonlinear advection-diffusion problems. *SIAM J. Numer. Anal.*, 51(2):773–793, 2013.

[15] V. Dolejší and M. Feistauer. *Discontinuous Galerkin method. Analysis and applications to compressible flow*, volume 48 of *Springer Ser. Comput. Math.* Cham: Springer, 2015.

[16] V Dolejší and H.-G. Roos. Bdf-FEM for parabolic singularly perturbed problems with exponential layers on layer-adapted meshes in space. *Neural Parallel Sci. Comput.*, 18(2):221–235, 2010.

[17] V. Dolejší, F. Roskovec, and M. Vlasák. A posteriori error estimates for higher order space-time Galerkin discretizations of nonlinear parabolic problems. *SIAM J. Numer. Anal.*, 59(3):1486–1509, 2021.

[18] B. L. Ehle. On padé approximations to the exponential function and a-stable methods for the numerical solution of initial value problems. *Research report CSRR 2010,*, Dept. AACS, Univ. of Waterloo, Ontario, Canada, 1969.

[19] K. Eriksson, C. Johnson, and V. Thomée. Time discretization of parabolic problems by the discontinuous Galerkin method. *RAIRO, Modélisation Math. Anal. Numér.*, 19:611–643, 1985.

[20] A. Ern, I. Smears, and M. VohralÍk. Guaranteed, locally space-time efficient, and polynomial-degree robust a posteriori error estimates for high-order discretizations of parabolic problems. *SIAM J. Numer. Anal.*, 55(6):2811–2834, 2017.

[21] A. Ern and M. Vohralík. Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. *SIAM J. Numer. Anal.*, 53(2):1058–1081, 2015.

[22] M. Feistauer, J. Felcman, and I. Straškraba. *Mathematical and computational methods for compressible flow.* Numer. Math. Sci. Comput. Oxford: Oxford University Press, 2003.

[23] L. Formaggia and F. Nobile. A stability analysis for the arbitrary Lagrangian Eulerian formulation with finite elements. *East-West J. Numer. Math.*, 7(2):105–131, 1999.

[24] R. Frank, J. Schneid, and Ch. W. Ueberhuber. Order results for implicit Runge-Kutta methods applied to stiff systems. *SIAM J. Numer. Anal.*, 22:515–534, 1985.

[25] L. Gastaldi. A priori error estimates for the arbitrary Lagrangian Eulerian formulation with finite elements. *East-West J. Numer. Math.*, 9(2):123–156, 2001.

[26] C. W. Gear. *Numerical initial value problems in ordinary differential equations.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[27] E. H. Georgoulis, O. Lakkis, and J. M. Virtanen. A posteriori error control for discontinuous Galerkin methods for parabolic problems. *SIAM J. Numer. Anal.*, 49(2):427–458, 2011.

[28] A. Guillou and J. L. Soulé. La résolution numérique des problémes différentiels aux conditions initiales par des méthodes de collocation. *Rev. Franç. Inform. Rech. Opér.*, 3(R-3):17–44, 1969.

[29] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I: Nonstiff problems.*, volume 8 of *Springer Ser. Comput. Math.* Berlin: Springer, 2010.

[30] E. Hairer and G. Wanner. *Solving ordinary differential equations. II: Stiff and differential-algebraic problems.*, volume 14 of *Springer Ser. Comput. Math.* Berlin: Springer, 2010.

[31] B. L. Hulme. One-step piecewise polynomial Galerkin methods for initial value problems. *Math. Comput.*, 26:415–426, 1972.

[32] L. Kaland and H.-G. Roos. Parabolic singularly perturbed problems with exponential layers: robust discretizations using finite elements in space on Shishkin meshes. *Int. J. Numer. Anal. Model.*, 7(3):593–606, 2010.

[33] V. Kučera. On diffusion-uniform error estimates for the DG method applied to singularly perturbed problems. *IMA J. Numer. Anal.*, 34(2):820–861, 2014.

[34] V. Kučera and M. Vlasák. A priori diffusion-uniform error estimates for nonlinear singularly perturbed problems: BDF2, midpoint and time DG. *ESAIM, Math. Model. Numer. Anal.*, 51(2):537–563, 2017.

[35] J. Kuntzmann. Neuere Entwicklungen der Methode von Runge und Kutta. *Z. Angew. Math. Mech.*, 41:t28–t31, 1961.

[36] Ch. Makridakis and R. H. Nochetto. A posteriori error analysis for higher order dissipative methods for evolution problems. *Numer. Math.*, 104(4):489–514, 2006.

[37] J. M. Melenk and B. I. Wohlmuth. On residual-based a posteriori error estimation in hp-FEM. *Adv. Comput. Math.*, 15(1-4):311–331, 2001.

[38] L. E. Payne and H. F. Weinberger. An optimal Poincaré inequality for convex domains. *Arch. Ration. Mech. Anal.*, 5:286–292, 1960.

[39] W. Prager and J. L. Synge. Approximations in elasticity based on the concept of function space. *Q. Appl. Math.*, 5:241–269, 1947.

[40] S. Repin. Estimates of deviations from exact solutions of initial-boundary value problem for the heat equation. *Atti Accad. Naz. Lincei, Cl. Sci. Fis. Mat. Nat., IX. Ser., Rend. Lincei, Mat. Appl.*, 13(2):121–133, 2002.

[41] H.-G. Roos, M. Stynes, and L. Tobiska. *Robust numerical methods for singularly perturbed differential equations. Convection-diffusion-reaction and flow problems*, volume 24 of *Springer Ser. Comput. Math.* Berlin: Springer, 2008.

[42] D. Schötzau and T. P. Wihler. A posteriori error estimation for *hp*-version time-stepping methods for parabolic partial differential equations. *Numer. Math.*, 115(3):475–509, 2010.

[43] V. Thomée. *Galerkin finite element methods for parabolic problems.* Berlin: Springer, 2006.

[44] J. J. W. van der Vegt and H. van der Ven. Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows. I: General formulation. *J. Comput. Phys.*, 182(2):546–585, 2002.

[45] R. Verfürth. *A posteriori error estimation techniques for finite element methods.* Numer. Math. Sci. Comput. Oxford: Oxford University Press, 2013.

[46] M. Vlasák. Optimal spatial error estimates for DG time discretizations. *J. Numer. Math.*, 21(3):201–230, 2013.

[47] M. Vlasák. On polynomial robustness of flux reconstructions. *Appl. Math., Praha*, 65(2):153–172, 2020.

[48] M. Vlasak and H.-G. Roos. An optimal uniform a priori error estimate for an unsteady singularly perturbed problem. *Int. J. Numer. Anal. Model.*, 11(1):24–33, 2014.

[49] M. Vlasák and F. Roskovec. On Runge-Kutta, collocation and discontinuous Galerkin methods: mutual connections and resulting consequences to the analysis. In *Programs and algorithms of numerical mathematics 17. Proceedings of the 17th seminar (PANM), Dolní Maxov, Czech Republic, June 8–13, 2014*, pages 231–236. Prague: Academy of Sciences of the Czech Republic, Institute of Mathematics, 2015.

[50] K. Wright. Some relationships between implicit Runge-Kutta, collocation and Lanczos τ methods, and their stability properties. *BIT, Nord. Tidskr. Inf.-behandl.*, 10:217–227, 1970.

[51] Q. Zhang and Ch.-W. Shu. Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws. *SIAM J. Numer. Anal.*, 44(4):1703–1720, 2006.

Chapter 3

Linear unsteady singularly perturbed convection-diffusion problems

INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 11, Number 1, Pages 24–33

AN OPTIMAL UNIFORM A PRIORI ERROR ESTIMATE FOR AN UNSTEADY SINGULARLY PERTURBED PROBLEM

MILOSLAV VLASAK AND HANS–GÖRG ROOS

Abstract. A time–dependent convection–diffusion problem is discretized by the Galerkin finite element method in space with bilinear elements on a general layer adapted mesh and in time by discontinuous Galerkin method. We present optimal error estimates. The estimates hold true for consistent stabilization too.

Key words. discontinuous Galerkin, convection–diffusion, layer adapted mesh, error estimate

1. Introduction

We focus ourselves on the analysis of the solution of unsteady linear 2D singularly perturbed convection–diffusion equation. This type of equation can be considered as simplified model problem to many important problems, especially to Navier–Stokes equations.

The space discretization of such a problem is a difficult task and it stimulated development of many stabilization methods (e.g. streamline upwind Petrov–Galerkin (SUPG) method, local projection stabilization methods) and layer–adapting techniques (e.g. Shishkin meshes, Bakhvalov meshes). For the overview see [9] or [8].

In order to achieve optimal diffusion–uniform error estimates we employ layer adapted meshes. On these general layer adapted meshes we assume a general space discretization covering standard conforming finite element method (FEM) or consistent stabilization methods. The resulting system of ordinary differential equations is solved by discontinuous Galerkin (DG) method.

Considering the space discretization on Shishkin meshes, we will follow the theory for stationary singularly perturbed problems based on the solution decomposition, which enables us to derive a priori error estimates independent of the diffusion parameter even with respect to the norms (seminorms) of the exact solution, which can be also highly dependent on the diffusion parameter. For the details see [9].

The discontinuous Galerkin (DG) method is a very popular approach for solving ordinary differential equations arising from space discretization of parabolic problems, which is based on piecewise polynomial approximation in time. Among important advantages we should mention unconditional stability for arbitrary order, which allows us to solve stiff problems efficiently, and good smoothing property, which enables us to work with inexact or rough data. For introduction to DG time discretization see e.g. [11].

In [6] and [1] the authors study DG in time and DG and local projection stabilization method, respectively, in space on standard meshes for singularly perturbed problems. The error estimates in these papers contain norms of the exact solutions which go to infinity if diffusion parameter goes to zero.

There are only few papers dealing with finite elements in space on the special meshes combined with any discretization in time. While in [7] the θ–scheme as discretization in time is used, in [5] the authors study BDF time discretization.

Received by the editors November 23, 2012 and, in revised form, March 3, 2013.
2000 *Mathematics Subject Classification.* 65M15, 65N30, 65M50.

In [7] the authors also study DG time discretization and derive suboptimal error estimates.

Our aim is improving some results from [7] and proving optimal a priori diffusion–uniform error estimates for DG time discretization in $L^\infty(L^2)$ norm.

The main difficulty in proving optimal diffusion–uniform error estimates for DG time discretization is the fact that we cannot employ standard technique of the proof, which is based on the construction of a suitable projection, which enables us to eliminate discrete time derivative in the error equation, see e.g. [10]. This technique enforces us to do some upper bound of the projection error contained in stationary terms, which depends on a higher time derivative of the exact solution in H^1 seminorm, which depends on the diffusion parameter.

2. Continuous problem

Let $\Omega = (0,1)^2$ be a computational domain and $T > 0$. Then let us consider parabolic singularly perturbed problem

$$\begin{aligned} \frac{\partial u}{\partial t} - \varepsilon \Delta u + b \cdot \nabla u + cu &= f, \quad \forall x \in \Omega, t \in (0,T), \\ u &= 0, \quad \forall x \in \partial\Omega, t \in (0,T), \\ u(x,0) &= u^0(x), \quad \forall x \in \Omega, \end{aligned} \tag{1}$$

where function $u^0 \in L^2(\Omega)$, $0 < \varepsilon << 1$ and functions $f(x,t)$, $b(x)$ and $c(x)$ are sufficiently smooth with $b_1(x) > \beta_1 > 0$ and $b_2(x) > \beta_2 > 0$. By substitution in time variable we can achieve

$$c - \frac{1}{2}\nabla \cdot b \geq c_0 > 0. \tag{2}$$

To simplify the text we will use the following notation. $(.,.)$ and $\|.\|$ are $L^2(\Omega)$ scalar product and norm, $|.|_1$ and $\|.\|_1$ are $H^1(\Omega)$ seminorm and norm. Let us define bilinear form

$$a(u,v) = \varepsilon(\nabla u, \nabla v) + (b \cdot \nabla u + cu, v). \tag{3}$$

Definition 1. *We say that the function $u \in L^2(0,T,H_0^1(\Omega))$ with the time derivative $\frac{\partial u}{\partial t} \in L^2(0,T,H^{-1}(\Omega))$ is the weak solution of (1), if the following conditions are satisfied*

$$\begin{aligned} \left(\frac{\partial u(t)}{\partial t}, v\right) + a(u(t),v) &= (f(t),v) \quad \forall t \in (0,T), \forall v \in H_0^1(\Omega), \\ u(0) &= u^0. \end{aligned} \tag{4}$$

It is possible to show that the solution has in general boundary layer around the border of Ω at $x = 1$ and $y = 1$. Assuming sufficiently compatible data we can avoid the existence of interior layers, which enables us to concentrate on the boundary layers only, see [9] or [4]. Moreover, it is possible to guarantee the S–decomposition

of the solution: $u = S + V_1 + V_2 + V_{12}$, where

$$\left|\frac{\partial^{i+j+k} S(x_1,x_2,t)}{\partial x_1^i \partial x_2^j \partial t^k}\right| \leq C, \tag{5}$$

$$\left|\frac{\partial^{i+j+k} V_1(x_1,x_2,t)}{\partial x_1^i \partial x_2^j \partial t^k}\right| \leq C\varepsilon^{-i} e^{-\beta_1(1-x_1)/\varepsilon}, \tag{6}$$

$$\left|\frac{\partial^{i+j+k} V_2(x_1,x_2,t)}{\partial x_1^i \partial x_2^j \partial t^k}\right| \leq C\varepsilon^{-j} e^{-\beta_2(1-x_2)/\varepsilon}, \tag{7}$$

$$\left|\frac{\partial^{i+j+k} V_{12}(x_1,x_2,t)}{\partial x_1^i \partial x_2^j \partial t^k}\right| \leq C\varepsilon^{-i-j} \min\{e^{-\beta_1(1-x_1)/\varepsilon}, e^{-\beta_2(1-x_2)/\varepsilon}\}, \tag{8}$$

where i, j, k are nonnegative integers such that $i + j \leq 3$ and $k \leq q + 2$, where q denotes the degree of the intended polynomial approximation in time. S represents the smooth part of the solution, V_1 and V_2 represent boundary layers and V_{12} represents the corner layer. This result shows dependence of space derivatives on ε, which complicates deriving standard a priori error estimates.

2.1. Discretization. We want to discretize the problem (1) by either standard finite element method or some consistent stabilization method on general layer adapted meshes in space. This technique allows us to derive a priori error estimates that are independent of ε.

We will start with the construction of the general layer adapted mesh. To do this we will follow the approach described in [8] or [9]. Let us denote N, space mesh parameter, as an even number. Then let us set

$$0 = x_0 < x_1 < \ldots < x_N = 1, \quad 0 = y_0 < y_1 < \ldots < y_N = 1. \tag{9}$$

The final mesh arises as tensor product mesh with mesh points (x_i, y_j). Since the idea of distribution of mesh points is the same in both direction (using either parameter β_1 or β_2), we describe the idea only in x_1 direction. Let us introduce the mesh generating function ϕ satisfying $\phi(0) = 0$ and $\phi(1/2) = \ln(N)$, moreover we assume ϕ be continuous, increasing and differentiable. Let the mesh points are equally distributed in $[0, x_{N/2}]$ and graded according to the function ϕ in $[x_{N/2}, 1]$:

$$x_i = \frac{2i}{N}\left(1 - \frac{\sigma\varepsilon}{\beta_1}\phi\left(\frac{1}{2}\right)\right), \quad \forall i = 0, \ldots, N/2 \tag{10}$$

$$x_i = 1 - \frac{\sigma\varepsilon}{\beta_1}\phi\left(\frac{N-i}{N}\right), \quad \forall i = N/2, \ldots, N. \tag{11}$$

The parameter σ is chosen to satisfy $\sigma \geq 5/2$. These meshes can be called S–type meshes. For instance, the special choice of the function $\phi(s) = 2\ln(N)s$ leads to classical Shishikin mesh and the choice $\phi(s) = -\ln(1 - 2s(1 - N^{-1}))$ leads to Bakhvalov–type meshes.

Let us define the conforming bilinear finite element space V_N on our mesh. We denote $a_{st}(.,.)$ the space discretization bilinear form and f_{st} the corresponding right–hand side. In the case of classical finite element method the form $a_{st}(.,.)$ and the right–hand side f_{st} are identical to former bilinear form $a(.,.)$ and former right–hand side f, but they can differ in the case of stabilization methods. Moreover, we assume that the new bilinear form is consistent, i.e., the exact solution u satisfies

$$\left(\frac{\partial u}{\partial t}, v\right) + a_{st}(u, v) = (f_{st}, v), \quad \forall v \in V_N. \tag{12}$$

The semi–discrete problem reads: find $u_N \in C^1(0,T,V_N)$ satisfying

$$\begin{aligned} \left(\frac{\partial u_N(t)}{\partial t}, v\right) + a_{st}(u_N(t), v) &= (f_{st}(t), v), \quad \forall v \in V_N,\ \forall t \in (0,T), \\ (u_N(0), v) &= (u^0, v). \quad \forall v \in V_N \end{aligned} \tag{13}$$

To discretize this problem in time we assume time partition $0 = t_0 < t_1 < \ldots < t_r = T$ with time intervals $I_m = (t_{m-1}, t_m)$, time steps $\tau_m = |I_m| = t_m - t_{m-1}$ and $\tau = \max_{m=1,\ldots,r} \tau_m$. We denote the function values at the nodes as $v^m = v(t_m)$. To be able to use the Galerkin type of discretization we denote the space of piecewise polynomial functions

$$V_N^\tau = \{v \in L^2(0,T,V_N) : v|_{I_m} = \sum_{j=0}^{q} v_{j,m} t^j,\ v_{j,m} \in V_N\}. \tag{14}$$

For the functions from such a space we need to define the values at the nodes of time partition

$$v_\pm^m = v(t_m \pm) = \lim_{t \to t_m \pm} v(t) \tag{15}$$

and the jumps

$$\{v\}_m = v_+^m - v_-^m. \tag{16}$$

Definition 2. *We say that the function $U \in V_N^\tau$ is the approximate solution to the problem (1) if*

$$\begin{aligned} \int_{I_m} (U', v) + a_{st}(U, v)dt + (\{U\}_{m-1}, v_+^{m-1}) &= \int_{I_m} (f_{st}, v)dt, \\ &\forall v \in V_N^\tau,\ \forall m = 1, \ldots, r \\ (U_-^0, v) &= (u^0, v) \quad \forall v \in V_N. \end{aligned} \tag{17}$$

3. Error analysis

We define energy norm

$$|\!|\!| v |\!|\!|^2 = a_{st}(v, v), \quad \forall v \in H^1(\Omega). \tag{18}$$

3.1. Stationary problem. In this part we want to go through some well known results for the singularly perturbed problems (for the details see [9]). Let us assume related stationary problem

$$a_{st}(u, v) = (f_{st}^*, v), \quad \forall v \in H_0^1(\Omega), \tag{19}$$

with some $f_{st}^* \in L^2(\Omega)$, and corresponding discrete finite element problem on layer–adapted mesh. Let us define the Ritz projection $R : H_0^1(\Omega) \to V_N$ satisfying

$$a_{st}(u - Ru, v) = 0, \quad \forall v \in V_N. \tag{20}$$

We assume that on layer–adapted mesh following error estimates hold true:

$$|\!|\!| u - Ru |\!|\!| \leq C g_1(N), \tag{21}$$

$$\|u - Ru\| \leq C g_2(N), \tag{22}$$

$$\|u' - Ru'\| \leq C g_2(N), \tag{23}$$

with C independent of ε. In the case of classical finite element method on Shishkin mesh we obtain these results with $g_1(N) = N^{-1}\ln(N)$ and $g_2(N) = (N^{-1}\ln(N))^2$. The same situation with Bakhvalov mesh leads to the estimates $g_1(N) = N^{-1}$ and $g_2(N) = N^{-2}$. Remark that the estimates in L^2–norm are based on supercloseness

results, because it is not possible to use the Nitsche–duality trick. See [9] for more detailed informations. From this follows easily

Lemma 1. *Let u be the exact solution of (1). Then*

$$\|Ru(s_1) - u(s_1) - Ru(s_2) + u(s_2)\| \quad \leq \quad C|s_1 - s_2|g_2(N). \tag{24}$$

Proof.

$$\begin{aligned} \|Ru(s_1) - u(s_1) - Ru(s_2) + u(s_2)\| &= \|\int_{s_2}^{s_1} Ru'(t) - u'(t)dt\| \\ &\leq |s_1 - s_2| \sup_{I_m} \|Ru' - u'\| \\ &\leq C|s_1 - s_2|g_2(N) \end{aligned} \tag{25}$$

□

3.2. Radau quadrature. Let us define Radau quadrature on each interval I_m

$$\int_{I_m} f\,dt \approx Q[f] = \sum_{i=0}^{q} w_i f(t_{m,i}), \tag{26}$$

where $t_{m,i}$ are Radau quadrature nodes in I_m with $t_{m,0} = t_m$. Such a quadrature has algebraic order $2q$ and the coefficients of the quadrature satisfy $0 \leq w_i \leq \tau_m$ and

$$\sum_{i=0}^{q} w_i = \tau_m. \tag{27}$$

Let us assume for simplicity that right–hand side f (and therefore f_{st}) of our continuous problem (1) is polynomial up to the degree q. Otherwise, we will need to use additionally error estimate of following type

$$\int_{I_m} (f, v)dt - Q[(f, v)] \leq \tau_m C\tau^{q+1} \sup_{I_m} \|v\|, \quad \forall v \in V_N^\tau, \tag{28}$$

which holds true for f sufficiently smooth in time. Then it is possible to express our method (17) by

$$Q[(U', v)] + Q[a_{st}(U, v)] + (\{U\}_{m-1}, v_+^{m-1}) = Q[(f_{st}, v)], \quad \forall v \in V_N^\tau. \tag{29}$$

Since the equation for continuous solution (1) is defined at every point $t \in I_m$, we can see that the exact solution satisfy (29) too.

3.3. Projections. Let us set the space

$$V^\tau = \{v \in L^2(0, T, H_0^1(\Omega)) : v|_{I_m} = \sum_{j=0}^{q} v_{j,m} t^j,\ v_{j,m} \in H_0^1(\Omega)\}. \tag{30}$$

We define time projection $P : C([0, T], H_0^1(\Omega)) \to V^\tau$, such that

$$Pu(t) = \sum_{i=0}^{q} \ell_i(t) u(t_{m,i}), \tag{31}$$

where ℓ_i is Lagrange interpolation basis function for the quadrature node $t_{m,i}$. Since

$$RPu(t) = R\sum_{i=0}^{q} \ell_i(t) u(t_{m,i}) = \sum_{i=0}^{q} \ell_i(t) Ru(t_{m,i}) = PRu(t), \tag{32}$$

we can see that projections P and R commute. We define the space–time projection $\pi = PR : C(0, T, H_0^1(\Omega)) \to V_N^\tau$.

Now, we present some basic approximation properties of our projections P and π.

Lemma 2. *Let u be the exact solution of (1). Then*

$$\sup_{I_m} \|Pu - u\| \leq C\tau^{q+1}, \tag{33}$$

$$\sup_{I_m} \|Pu' - u'\| \leq C\tau^{q+1}, \tag{34}$$

where the constant C does not depend on τ.

Proof. The proof can be made by standard arguments. It is an analogy to e.g. [3, Theorem 3.1.4] in Bochner spaces. □

Lemma 3. *Let u be the exact solution of (1). Then*

$$\sup_{I_m} \|\pi u - u\| \leq C(\tau^{q+1} + g_2(N)), \tag{35}$$

where the constant C does not depend on τ or N.

Proof. Since $|\ell_i(t)| \leq C$, where the constant C depends only on q, we obtain

$$\begin{aligned} \sup_{I_m} \|\pi u - u\| &\leq \sup_{I_m} \|Pu - u\| + \sup_{I_m} \|PRu - Pu\| \\ &\leq C\tau^{q+1} + C\|\sum_{i=0}^{q} Ru(t_{m,i}) - u(t_{m,i})\| \\ &\leq C\tau^{q+1} + C(q+1) \max_{i=0,\dots,q} \|Ru(t_{m,i}) - u(t_{m,i})\| \\ &\leq C(\tau^{q+1} + g_2(N)). \end{aligned} \tag{36}$$

□

3.4. Auxiliary result. We subtract the equation for exact solution from (29) and divide the error into projection part $\eta = \pi u - u$ and $\xi = U - \pi u \in V_N^\tau$. We obtain

$$\begin{aligned} &\int_{I_m} (\xi', v) + a_{st}(\xi, v)dt + (\{\xi\}_{m-1}, v_+^{m-1}) \\ &= -Q[(\eta', v)] - (\{\eta\}_{m-1}, v_+^{m-1}) - Q[a_{st}(\eta, v)]. \end{aligned} \tag{37}$$

Since

$$Q[a_{st}(\eta, v)] = \sum_{i=0}^{q} w_i a_{st}(Ru(t_{m,i}) - u(t_{m,i}), v) = 0 \tag{38}$$

we need to estimate the rest of the right–hand side only.

Lemma 4. *Let u be an exact solution of (1). Then*

$$Q[(\eta', v)] + (\{\eta\}_{m-1}, v_+^{m-1}) \leq \tau_m C \left(\tau^{q+1} + g_2(N)\right) \sup_{I_m} \|v\|, \quad \forall v \in V_N^\tau. \tag{39}$$

Proof.

$$\begin{aligned} &Q[(\eta', v)] + (\{\eta\}_{m-1}, v_+^{m-1}) \\ &= \int_{I_m} ((\pi u)', v)dt - Q[(u', v)] + (\{\eta\}_{m-1}, v_+^{m-1}) \\ &= \int_{I_m} (\eta', v)dt + (\{\eta\}_{m-1}, v_+^{m-1}) + \int_{I_m} (u', v)dt - Q[(u', v)] \end{aligned} \tag{40}$$

We estimate first two terms and last two terms (quadrature error) individually.

$$\begin{aligned}(41)\qquad &\int_{I_m}(\eta',v)dt+(\{\eta\}_{m-1},v_+^{m-1})\\ &=-\int_{I_m}(\eta,v')+(\eta_-^m,v_-^m)-(\eta_-^{m-1},v_+^{m-1})\end{aligned}$$

We can see that $v_-^m=v_+^{m-1}+\int_{I_m}v'dt$. Using this fact we obtain

$$\begin{aligned}(42)\qquad &-\int_{I_m}(\eta,v')+(\eta_-^m,v_-^m)-(\eta_-^{m-1},v_+^{m-1})\\ &=\int_{I_m}(\eta_-^m-\eta,v')dt+(\eta_-^m-\eta_-^{m-1},v_+^{m-1})\end{aligned}$$

We estimate these terms individually. The first term we can rewrite in the following way

$$\begin{aligned}(43)\qquad &\int_{I_m}(\eta_-^m-\eta,v')dt=\int_{I_m}(Ru^m-u^m-RPu+u,v')dt\\ &=\int_{I_m}(Ru^m-u^m-RPu+Pu,v')dt+\int_{I_m}(u-Pu,v')dt\end{aligned}$$

Since all the terms in the first integral on the right–hand side are polynomials we can apply Radau quadrature exactly and using (27), Lemma 1 and inverse inequality we get

$$\begin{aligned}(44)\qquad &\int_{I_m}(Ru^m-u^m-RPu+Pu,v')dt\\ &=Q[(Ru^m-u^m-RPu+Pu,v')]\\ &\le\tau_m\sup_i\|Ru^m-u^m-Ru(t_{m,i})+u(t_{m,i})\|\sup_{I_m}\|v'\|\\ &\le\tau_mCg_2(N)\sup_{I_m}\|v\|\end{aligned}$$

We need to estimate $\int_{I_m}(u-Pu,v')dt$. To do this we define interpolation operator $\hat{P}$ such that $\hat{P}u$ is a polynomial of degree $q+1$ in time which interpolates u in Radau quadrature nodes $t_{m,i}$ and (in addition) t_{m-1}. Then we get

$$(45)\qquad \int_{I_m}(\hat{P}u,v')dt=\int_{I_m}(Pu,v')dt.$$

It is possible to show that $\sup_{I_m}\|u-\hat{P}u\|\le C\tau_m^{q+2}$ by the same arguments as for interpolation operator P. Then we get with the inverse inequality on the test function v

$$\begin{aligned}(46)\qquad &\int_{I_m}(u-Pu,v')dt=\int_{I_m}(u-\hat{P}u,v')dt\\ &\le\tau_mC\tau_m^{q+2}\sup_{I_m}\|v'\|\le\tau_mC\tau^{q+1}\sup_{I_m}\|v\|.\end{aligned}$$

The estimate for the second term follows directly from Lemma 1

$$\begin{aligned}(47)\qquad &(\eta_-^m-\eta_-^{m-1},v_+^{m-1})=(Ru^m-u^m-Ru^{m-1}+u^{m-1},v_+^{m-1})\\ &\le\tau_m\sup_{I_m}\|Ru'-u'\|\sup_{I_m}\|v\|\le\tau_mCg_2(N)\sup_{I_m}\|v\|.\end{aligned}$$

Finally, we need to estimate quadrature error.

$$\begin{aligned}\int_{I_m}(u',v)dt - Q[(u',v)] &= \int_{I_m}(u'-Pu',v)dt \\ &\le \tau_m C\tau^{q+1}\sup_{I_m}\|v\|\end{aligned} \tag{48}$$

□

Remark 1. *Lemma 4 can be easily generalized to any time projection that interpolates the end points of the intervals and to any space projection that commutes with the time projection. Then the result will take the following form*

$$\begin{aligned}&Q[(\eta',v)] + (\{\eta\}_{m-1}, v_+^{m-1}) \\ &\le \tau_m C\,(\text{'time error'} + \text{'space error'})\sup_{I_m}\|v\|, \quad \forall v \in V_N^\tau.\end{aligned} \tag{49}$$

For the estimates of supremum term we will need the following lemma.

Lemma 5. *Let $\xi \in V_N^\tau$ and*

$$\tilde{\xi} = P\left(\frac{\tau_m \xi(t)}{t-t_{m-1}}\right) \in V_N^\tau. \tag{50}$$

Then

$$\int_{I_m}(\xi',2\tilde{\xi})dt + (\xi_+^{m-1}, 2\tilde{\xi}_+^{m-1}) = \|\xi_-^m\|^2 + \frac{1}{\tau_m}\int_{I_m}\|\tilde{\xi}\|^2 dt. \tag{51}$$

Proof. The proof can be made as a simple extension of [2, Lemma 2.1], which describes the same result for scalar polynomials and on unit time interval. □

3.5. Main result. We are ready to present the main result.

Theorem 1. *Let u be an exact solution of (1) and $U \in V_N^\tau$ be its discrete approximation given by (17). Then*

$$\max_{m=1,\dots,r}\sup_{I_m}\|U-u\| \le C\left(g_2(N) + \tau^{q+1}\right). \tag{52}$$

Proof. We can estimate right–hand side of (37) by Lemma 4. Then we obtain

$$\begin{aligned}&\int_{I_m}(\xi',v) + a_{st}(\xi,v)dt + (\{\xi\}_{m-1}, v_+^{m-1}) \\ &\le \tau_m C\left(\tau^{q+1} + g_2(N)\right)\sup_{I_m}\|v\|.\end{aligned} \tag{53}$$

Setting $v = 2\xi$ we get

$$\begin{aligned}&\|\xi_-^m\|^2 - \|\xi_-^{m-1}\|^2 + \|\{\xi\}_{m-1}\|^2 + 2\int_{I_m}|\!|\!|\xi|\!|\!|^2 dt \\ &\le \tau_m C\left(\tau^{q+1} + g_2(N)\right)\sup_{I_m}\|\xi\| \\ &\le \tau_m C\left(\tau^{2q+2} + g_2(N)^2\right) + \frac{\tau_m}{2}\sup_{I_m}\|\xi\|^2\end{aligned} \tag{54}$$

We need to deal with the last term at the right–hand side.

It is simple to see that for $\tilde{\xi}$ defined by (50) we get

$$\begin{aligned}(55)\qquad &\int_{I_m} |\!|\!|\xi|\!|\!|^2 dt = \int_{I_m} a_{st}(\xi,\xi)dt \\ &= Q[a_{st}(\xi,\xi)] = \sum_{i=0}^{q} w_i a_{st}(\xi(t_{m,i}),\xi(t_{m,i})) \\ &\leq \sum_{i=0}^{q} w_i \frac{\tau_m}{t_{m,i}-t_{m-1}} a_{st}(\xi(t_{m,i}),\xi(t_{m,i})) \\ &= Q[a_{st}(\xi,\tilde{\xi})] = \int_{I_m} a_{st}(\xi,\tilde{\xi})dt,\end{aligned}$$

since $\tau_m/(t_{m,i}-t_{m-1}) \geq 1$.

Since the terms $\sup_{I_m}\|\xi\|^2$, $\frac{1}{\tau_m}\int_{I_m}\|\tilde{\xi}\|^2 dt$ and $\sup_{I_m}\|\tilde{\xi}\|^2$ are equivalent, we get by setting $v = 2\tilde{\xi}$ in (53) with the aid of Lemma 5

$$\begin{aligned}(56)\qquad \sup_{I_m}\|\xi\|^2 &\leq C\frac{1}{\tau_m}\int_{I_m}\|\tilde{\xi}\|^2 dt \\ &\leq C\left(\|\xi_-^m\|^2 + \frac{1}{\tau_m}\int_{I_m}\|\tilde{\xi}\|^2 dt + 2\int_{I_m}|\!|\!|\xi|\!|\!|^2 dt\right) \\ &\leq C\left((\xi_-^{m-1},\tilde{\xi}_+^{m-1}) + C\left(\tau^{q+1}+g_2(N)\right)\sup_{I_m}\|\xi\|\right) \\ &\leq C\left(\|\xi_-^{m-1}\|^2 + \tau^{2q+2} + g_2(N)^2\right) + \frac{1}{2}\sup_{I_m}\|\xi\|^2\end{aligned}$$

We can substitute this result into our error inequality (54) and we obtain

$$\|\xi_-^m\|^2 - \|\xi_-^{m-1}\|^2 \leq \tau_m C\left(\tau^{2q+2} + g_2(N)^2\right) + \tau_m C\|\xi_-^{m-1}\|^2.$$

Now, it is sufficient to employ the forward difference form of the discrete Gronwall lemma to obtain nodal error estimates. Estimates inside of intervals I_m follows from nodal estimates and from (56). □

Acknowledgments

The first author is a junior researcher of the University centre for mathematical modelling, applied analysis and computational mathematics (Math MAC).

References

[1] Ahmed, N., Matthies, G., Tobiska, L. and Xie, H. Discontinuous Galerkin time stepping with local projection stabilization for transient convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Eng.*, 200(21-22):1747–1756, 2011.

[2] Akrivis, G. and Makridakis, C. Galerkin time-stepping methods for nonlinear parabolic equations. 2004.

[3] Ciarlet, P.G. *The finite element methods for elliptic problems. Repr., unabridged republ. of the orig. 1978.* Classics in Applied Mathematics. 40. Philadelphia, PA: SIAM. xxiv, 530 p. , 2002.

[4] Clavero, C., Jorge, J.C., Lisbona, F. and Shishkin, G.I. A fractional step method on a special mesh for the resolution of multidimensional evolutionary convection-diffusion problems. *Appl. Numer. Math.*, 27(3):211–231, 1998.

[5] Dolejší, V. and Roos, H.-G. BDF-FEM for parabolic singularly perturbed problems with exponential layers on layer-adapted meshes in space. *Neural Parallel Sci. Comput.*, 18(2):221–235, 2010.

[6] Feistauer, M., Hájek, J. and Švadlenka, K. Space-time discontinuos Galerkin method for solving nonstationary convection-diffusion-reaction problems. *Appl. Math., Praha*, 52(3):197–233, 2007.

[7] Kaland, L. and Roos, H.-G. Parabolic singularly perturbed problems with exponential layers: robust discretizations using finite elements in space on Shishkin meshes. *Int. J. Numer. Anal. Model.*, 7(3):593–606, 2010.
[8] Linß, T. *Layer-adapted meshes for reaction-convection-diffusion problems.* Lecture Notes in Mathematics 1985. Berlin: Springer. xi, 320 p., 2010.
[9] Roos, H.-G., Stynes, M. and Tobiska, L.. *Robust numerical methods for singularly perturbed differential equations. Convection-diffusion-reaction and flow problems. 2nd ed.* Springer Series in Computational Mathematics 24. Berlin: Springer. xiv, 604 p. , 2008.
[10] Schötzau, D.. hp-DGFEM for parabolic evolution problems. Application to diffusion and viscous incompressible flow. *Ph.D. thesis, ETH Zürich*, 1999.
[11] Thomée, V. *Galerkin finite element methods for parabolic problems. 2nd revised and expanded ed.* Berlin: Springer. xii, 370 p., 2006.

Charles University in Prague, Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Prague, Czech Republic
E-mail: vlasak@karlin.mff.cuni.cz
URL: http://www.karlin.mff.cuni.cz/∼vlasak/

Technical University of Dresden, Institute of Numerical Mathematics, Helmholzstrasse 10, 01069 Dresden, Germany
E-mail: hans-goerg.roos@tu-dresden.de
URL: http://www.math.tu-dresden.de/∼roos/

Chapter 4

Semilinear unsteady singularly perturbed convection-diffusion problems

ESAIM: M2AN 51 (2017) 537–563
DOI: 10.1051/m2an/2016035

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

A PRIORI DIFFUSION-UNIFORM ERROR ESTIMATES FOR NONLINEAR SINGULARLY PERTURBED PROBLEMS: BDF2, MIDPOINT AND TIME DG *

Václav Kučera[1] and Miloslav Vlasák[1]

Abstract. This work deals with a nonlinear nonstationary semilinear singularly perturbed convection-diffusion problem. We discretize this problem by the discontinuous Galerkin method in space and by the midpoint rule, BDF2 and quadrature variant of discontinuous Galerkin in time. We present *a priori* error estimates for these three schemes that are uniform with respect to the diffusion coefficient going to zero and valid even in the purely convective case.

Mathematics Subject Classification. 65M12, 65M15, 65M60.

Received December 10, 2015. Revised April 19, 2016. Accepted May 16, 2016.

1. Introduction

The discontinuous Galerkin (DG) finite element method developed by Reed and Hill in [19] is a popular numerical method for the solution of advective and convective problems. The method uses high order piecewise polynomial approximations on a triangulation which are generally discontinuous between elements, unlike the standard conforming finite element method. The discontinuous nature of the approximation is natural for problems where discontinuities or sharp gradients and boundary layers occur in the solution, *e.g.* nonlinear convective problems or singular perturbations thereof.

Among the basic goals of numerical analysis is to prove *a priori* error estimates for the given problem and numerical method. For partial differential equations such techniques are usually based on some form of ellipticity/monotonicity in some part of the equation considered. The other terms are then dominated by this 'nice' part. In our case, for convection-diffusion problems, the convective terms are dominated by the elliptic diffusion terms, which, after the application of Gronwall's inequality leads to error estimates that blow up exponentially with respect to the diffusion parameter $\varepsilon \to 0$. Moreover this technique cannot be applied for purely convective problems, where the elliptic/monotone term is missing.

The fact that the DG scheme performs well for small or vanishing diffusion ε and even for the purely convective case is well known. When applied to smooth solutions, we know from practice that the error does not blow up exponentially, but rather stays bounded with respect to $\varepsilon \to 0$. Many numerical experiments confirming this

Keywords and phrases. Discontinuous Galerkin method, *a priori* error estimates, nonlinear convection-diffusion equation, diffusion-uniform error estimates.

* *The research is supported by the Grant No. P201/13/00522S of the Czech Science Foundation. The authors are junior researchers of the University centre for mathematical modelling, applied analysis and computational mathematics (Math MAC). V. Kučera is currently a Fulbright visiting scholar at Brown University, Providence, RI, USA, supported by the J. William Fulbright Commission in the Czech Republic.*

[1] Charles University in Prague, Faculty of Mathematics and Physics, Department of Numerical Mathematics, Sokolovská 83, 18675 Prague 8, Czech Republic. kucera@karlin.mff.cuni.cz; vlasak@karlin.mff.cuni.cz

can be found throughout the literature for various discretizations in time and varying ε, very small ε and $\varepsilon = 0$. For example, for the implicit-explicit (IMEX) variants of the backward difference formulas applied to the DG scheme, such results are contained in the papers [7,9]. In [9], the experiments are especially interesting as they are performed on general, very unusual grids, *e.g.* based on nonconvex quadrilaterals. For a combination of IMEX and time-DG, results are presented in the paper [22]. For explicit schemes and small ε, such results can be found in [8]. A comparison of small ε and $\varepsilon = 0$ and other similar numerical experiments, we refer to the recent book [6]. For purely convective problems, *i.e.* $\varepsilon = 0$, namely for the Euler equations, such results are obtained *e.g.* in [3,11]. Other works include for example [5,13]. In the presented paper we prove these observed results theoretically.

We will follow the ideas of Zhang and Shu [24], who developed a technique for *a priori* analysis of explicit time stepping DG schemes for convective problems. The technique is based on a specific estimate of the convective form which leads to the following: If the error is of the order $O(h^{(1+d)/2})$, where d is the spatial dimension of the computational domain, then we can prove the error estimate of the order $O(h^{p+1/2})$, where p is the spatial approximation order. A bootstrapping argument using mathematical induction is then applied to remove the *a priori* $O(h^{(1+d)/2})$ assumption. The argument works for explicit schemes under the assumption $p > (1+d)/2$.

In [16], the technique of Zhang and Shu was extended to the space semidiscretized DG and to the implicit Euler scheme. There it is proved that for implicit schemes, more information about the discrete solution is necessary to perform the bootstrapping argument. In [16], this difficulty is overcome by constructing a suitable continuation of the discrete solution with respect to time. The error analysis is then performed for the continued discrete solution, which implies error estimates for the original discrete solution. In the presented paper, we generalize these ideas to the BDF2, midpoint and quadrature version of time DG schemes. Specifically, we construct suitable continuations for these three schemes and then apply the induction argument presented in [16]. The quadrature time-DG scheme is especially interesting, since the continuation then depends on two variables, one of which is used for the induction argument, while the other represents the time variable of the original problem. In this case, the construction of the continuation is not complicated, however proving its necessary properties needed in the analysis is rather technical. Moreover, we were able to carry out the analysis only for the scheme where a quadrature formula in time is applied to the nonlinear terms. We do not view this as a limitation, since for practical computations, one must apply some form of quadrature to these terms anyway in order to evaluate them.

The structure of the paper is as follows. In Sections 2 and 3, we introduce the continuous problem, its spatial discretization by the DG method and the three considered time discretizations. In Section 4 we review the basic tools for our analysis, such as the basic estimate of the convective terms. Sections 5 and 6 deal with the analysis of the BDF2 and midpoint rules. We prove $O(h^{p+1/2} + \varepsilon h^p + \tau^2)$ error estimates in the $L^\infty(L^2)$-norm with the constant in the estimate independent of ε, h and τ. The estimates are derived under the $\tau = O(h)$ and $p > 1 + d/2$ conditions. For $\varepsilon = 0$, we obtain the weaker condition $p > (1+d)/2$ and the estimate of order $O(h^{p+1/2} + \tau^2)$.

Finally, Section 7 deals with the quadrature variant of the time-DG scheme. Under the same assumptions as for the BDF2 and midpoint schemes, we prove estimates of the order $O(h^{p+1/2} + \varepsilon h^p + \tau^{q+1})$, where q is the approximation order in time.

2. Continuous problem

Let $\Omega \subset \mathbb{R}^d$ be a bounded polyhedral domain and $T > 0$. We set $Q_T = \Omega \times (0,T)$. Let us consider the following problem: Find $u : Q_T \to \mathbb{R}$ such that

$$\begin{aligned} &\frac{\partial u}{\partial t} + \nabla \cdot f(u) - \varepsilon \Delta u = g \quad \text{in } Q_T, \\ &u\big|_{\partial\Omega\times(0,T)} = 0, \\ &u(x,0) = u^0(x), \quad x \in \Omega. \end{aligned} \tag{2.1}$$

Here $f = (f_1, \dots, f_d)$, $f_s \in C^2(\mathbb{R}) \cap W^{2,\infty}(\Omega)$, $f_s(0) = 0$, $s = 1, \dots, d$ represents convective terms, $\varepsilon \geq 0$, $g \in C([0,T]; L^2(\Omega))$ and $u^0 \in L^2(\Omega)$ is an initial condition. We assume that the weak solution of (2.1) is sufficiently regular and we will specify the exact assumptions on the smoothness of the weak solution for each time discretization method individually.

We note that in [16], mixed Dirichlet–Neumann boundary conditions are treated along with only locally Lipschitz nonlinearities $f_s \in C^2(\mathbb{R})$. This is also possible in our context, however we stay in the simpler setting to avoid too many technicalities.

To simplify the notation, we use $(\cdot, \cdot)$ to denote the L^2 scalar product and $\|\cdot\|$ for the L^2 norm. To further simplify notation, we shall drop the argument Ω in Sobolev norms, *e.g.* $\|\cdot\|_{H^{p+1}}$ denotes the $H^{p+1}(\Omega)$-norm. We will also denote the Bochner norms over the whole interval $(0,T)$ in concise form, *e.g.* $\|u\|_{L^\infty(H^{p+1})}$ denotes the $L^\infty(0,T; H^{p+1}(\Omega))$-norm.

3. Discrete problem

3.1. Space discretization

Let $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ be a system of partitions of $\overline{\Omega}$ into a finite number of closed d-dimensional simplices K with mutually disjoint interiors. Let $\mathcal{F}_h$ the system of all faces (edges in 2D) of $\mathcal{T}_h$ and let $\mathcal{F}_h^I$ be the set of interior edges and $\mathcal{F}_h^B$ the set of boundary edges. For each $\Gamma \in \mathcal{F}_h$ we fix a unit normal $\mathbf{n}_\Gamma$, which for $\Gamma \in \mathcal{F}_h^B$ has the same orientation as the outer normal to Ω. For each $\Gamma \in \mathcal{F}_h^I$ there exist two neighbours $K_\Gamma^{(L)}$, $K_\Gamma^{(R)} \in \mathcal{T}_h$ such that $\mathbf{n}_\Gamma$ is the outer normal to $K_\Gamma^{(L)}$. For v piecewise defined on $\mathcal{T}_h$ and $\Gamma \in \mathcal{F}_h^I$ we introduce $v|_\Gamma^{(L)}$ as the trace of $v|_{K_\Gamma^{(L)}}$ on Γ, $v|_\Gamma^{(R)}$ as the trace of $v|_{K_\Gamma^{(R)}}$ on Γ, $\langle v\rangle_\Gamma = \frac{1}{2}\big(v|_\Gamma^{(L)} + v|_\Gamma^{(R)}\big)$ and $[v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}$. On $\partial\Omega$, we define $v|_\Gamma^{(L)}$ as the trace of $v|_{K_\Gamma^{(L)}}$, *i.e.* on the element adjacent to Γ and $v|_\Gamma^{(R)} = 0$ corresponds to the homogeneous Dirichlet boundary conditions. If $[\cdot\,]_\Gamma, \langle\cdot\,\rangle_\Gamma, v|_\Gamma^{(L)}, v|_\Gamma^{(R)}$ appear in an integral over $\Gamma \in \mathcal{F}_h$, we omit the subscript Γ. Let

$$S_h = \{w;\ w|_K \in P_p(K),\ \forall K \in \mathcal{T}_h\}$$

denote the space of discontinuous piecewise polynomial functions of degree p on each $K \in \mathcal{T}_h$. We say that the function $u_h \in C^1(0,T; S_h)$ is the *semi-discrete approximate solution* of (2.1) if it satisfies the equation

$$\left(\frac{\partial u_h}{\partial t}(t), w\right) + \varepsilon A_h(u_h(t), w) + b_h(u_h(t), w) = \ell_h(w)(t) \quad \forall\, w \in S_h,\ \forall\, t \in [0,T],$$

and $(u_h(0), w) = (u^0, w)\ \forall\, w \in S_h$. Here the following forms are used: The *convective form*

$$b_h(v,\varphi) = -\sum_{K\in\mathcal{T}_h}\int_K f(v)\cdot\nabla\varphi \mathrm{d}x + \int_{\mathcal{F}_h^I} H(v^{(L)}, v^{(R)}, \mathbf{n})[\varphi]\mathrm{d}S + \int_{\mathcal{F}_h^B} H(v^{(L)}, v^{(R)}, \mathbf{n})\varphi^{(L)}\mathrm{d}S,$$

the diffusion terms are defined as

$$A_h(v,\varphi) = a_h(v,\varphi) + J_h(v,\varphi),$$

where the bilinear *diffusion form* corresponding to the *symmetric interior penalty Galerkin* (SIPG) is

$$a_h(v,\varphi) = \sum_{K\in\mathcal{T}_h}\int_K \nabla v\cdot\varphi \mathrm{d}x - \int_{\mathcal{F}_h^I}\langle\nabla v\rangle\cdot\mathbf{n}[\varphi]\mathrm{d}S - \int_{\mathcal{F}_h^I}\langle\nabla\varphi\rangle\cdot\mathbf{n}[v]\mathrm{d}S - \int_{\mathcal{F}_h^B}\nabla v\cdot\mathbf{n}\varphi\mathrm{d}S - \int_{\mathcal{F}_h^B}\nabla\varphi\cdot\mathbf{n}v\mathrm{d}S$$

and the *interior and boundary penalty jump terms* are defined by

$$J_h(v,\varphi) = \int_{\mathcal{F}_h^I}\sigma[v][\varphi]\mathrm{d}S + \int_{\mathcal{F}_h^B}\sigma v\varphi\mathrm{d}S.$$

Here the parameter σ is constant on every edge and defined by $\sigma|_\Gamma = C_W/|\Gamma|$ for all $\Gamma \in \mathcal{F}_h$, where $C_W > 0$ is a constant, which must be chosen large enough to ensure coercivity of the form A_h (*cf. e.g.* [10]).

Finally, we have the *right-hand side form*:

$$l_h(\varphi)(t) = \int_\Omega g(t)\varphi \mathrm{d}x.$$

As stated earlier, this is the case of homogeneous Dirichlet boundary conditions, for general mixed Dirichlet–Neumann conditions A_h has a more complicated form (*cf.* [16]), which we do not consider for simplicity.

We assume the numerical fluxes H in the convective form b_h to be Lipschitz continuous, conservative and consistent. Moreover, we assume that the numerical fluxes are E-fluxes:

$$(H(v,w,n) - f(q)\cdot n)(v-w) \geq 0, \quad \forall v,w \in \mathbb{R},\ \forall q \text{ between } v \text{ and } w,$$

where $n \in \mathbb{R}^d$ is an unit vector, *cf. e.g.* [2,18] for details.

We find that a sufficiently regular weak solution of (2.1) satisfies the identity

$$\left(\frac{\partial u}{\partial t}(t), w\right) + \varepsilon A_h(u(t),w) + b_h(u(t),w) = \ell_h(w)\,(t) \tag{3.1}$$

for all $w \in S_h$ and all $t \in (0,T)$.

Throughout this paper, we assume the mesh system $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ to be shape regular, satisfying the inverse assumption [4].

3.2. Time discretization

For simplicity we assume a uniform time partition $t_m = m\tau$, $m = 0,\dots,r$ with time intervals $I_m = (t_{m-1}, t_m)$ and with the time step $\tau = T/r = |I_m|$. To simplify the notation, we set $v^m = v(t_m)$.

3.2.1. BDF2

Definition 3.1. The set of functions $U^m \in S_h$, $m = 0,\dots,r$ is an approximate solution of problem (2.1) obtained by the BDF2-DG scheme if for all $w \in S_h$

$$\left(\frac{3}{2}U^m - 2U^{m-1} + \frac{1}{2}U^{m-2}, w\right) + \tau\varepsilon A_h(U^m,w) + \tau b_h(U^m,w) = \tau\ell_h(w)(t_m) \tag{3.2}$$

for $m \geq 2$. For $m = 1$ we define U^1 by

$$(U^1 - U^0, w) + \tau\varepsilon A_h(U^1,w) + \tau b_h(U^1,w) = \tau\ell_h(w)(t_m), \quad \forall w \in S_h. \tag{3.3}$$

The initial condition $U^0 \in S_h$ is the $L^2(\Omega)$-projection of u^0 onto S_h, *i.e.*

$$(U^0, w) = (u^0, w), \quad \forall w \in S_h. \tag{3.4}$$

Remark 3.2. Since the BDF2 is a 2-step method, we need to specify two initial values U^0 and U^1 to start the method. The value U^0 can be obtained by L^2 projection of initial condition u^0 and U^1 can be obtained by one step of the implicit Euler method. In this case the resulting scheme does not lose its second order of accuracy in time.

3.2.2. Midpoint rule

Definition 3.3. The set of functions $U^m \in S_h$, $m = 0,\dots,r$ is an approximate solution of problem (2.1) obtained by the midpoint-DG scheme if

$$(U^m - U^{m-1}, w) + \frac{\tau\varepsilon}{2}A_h(U^m + U^{m-1}, w) + \tau b_h\left(\frac{U^m + U^{m-1}}{2}, w\right) = \tau\ell_h(w)(t_{m-1} + \tau/2), \quad \forall w \in S_h, \tag{3.5}$$

where U^0 is the initial condition obtained by (3.4).

3.2.3. Discontinuous Galerkin method in time

We define the space

$$S_h^\tau = \{v \in L^2(0,T;S_h) : v|_{I_m} = \sum_{j=0}^{q} v_j^{(m)} t^j,\ v_j^{(m)} \in S_h,\ m = 1,\dots,r\},$$

which represents the space of piecewise polynomials up to degree p in space and up to degree q in time. For the functions from such a space we need to define one-sided values at nodes of the time partition:

$$v_\pm^m = v(t_m\pm) = \lim_{t\to t_m\pm} v(t)$$

and the jumps

$$\{v\}_m = v_+^m - v_-^m.$$

Definition 3.4. The function $U \in S_h^\tau$ is an approximate solution of problem (2.1) obtained by the space-time discontinuous Galerkin scheme if for all $w \in S_h^\tau$,

$$\int_{I_m} (U', w) + \varepsilon A_h(U,w) + b_h(U,w)\mathrm{d}t + \left(\{U\}_{m-1}, w_+^{m-1}\right) = \int_{I_m} \ell_h(w)(t)\mathrm{d}t,$$

for all $m = 1,\dots,r$. Here $U_-^0 := U^0$ is the initial condition obtained by (3.4).

Let us define the Radau quadrature on each interval I_m:

$$\int_{I_m} \Phi(t)\mathrm{d}t \approx Q_\tau^m[\Phi] := \tau \sum_{i=0}^{q} \omega_i \Phi(t_{m-1} + \tau\psi_i),$$

where ψ_i are Radau quadrature nodes in $[0,1]$ with $\psi_q = 1$. Such a quadrature has algebraic order $2q$ and the quadrature weights are positive and satisfy

$$\sum_{i=0}^{q} \omega_i = 1.$$

When we apply Radau quadrature to the integrals in Definition 3.4 we obtain the *quadrature version of the time-DG scheme*.

Definition 3.5. The function $U \in S_h^\tau$ is an approximate solution of problem (2.1) obtained by the quadrature time discontinuous Galerkin (QT-DG) scheme if for all $w \in S_h^\tau$

$$\int_{I_m} (U', w) + \varepsilon A_h(U,w)\mathrm{d}t + Q_\tau^m[b_h(U,w)] + (\{U\}_{m-1}, w_+^{m-1}) = Q_\tau^m[\ell_h(w)(t)], \tag{3.6}$$

for all $m = 1,\dots,r$. Here $U_-^0 := U^0$, the initial condition obtained by (3.4).

Remark 3.6. We note that the first integral in (3.6) does not need to be approximated by quadrature. Due to the linearity of the terms (U', w) and $A_h(U,w)$ w.r.t. both arguments, these terms are a polynomial of degree at most $2q$ on each I_m and can therefore be integrated exactly by Radau quadrature. However, due to the nonlinearity of the convective fluxes f_s, the term $b_h(U,w)$ cannot be, in general, integrated analytically w.r.t. time and quadrature must be applied in practice. The same holds for the right-hand side form $l_h(w)(t)$ containing the general function g.

Remark 3.7. The numerical solution U from Definition 3.4 or 3.5 is constructed on each I_m independently, inductively for $m = 1,\dots,r$, with only U_-^{m-1} coming from the previous time interval I_{m-1} or the initial condition U^0.

4. Auxiliary results

We denote the energy norm $|||w|||^2 := A_h(w,w)$ for all $w \in S_h$. Note that the inverse inequality takes the following form $|||w||| \le Ch^{-1}\|w\|$ for $w \in S_h$. Let Π be the $L^2(\Omega)$-orthogonal projection on S_h.

Throughout this work we denote by C a generic constant independent of h, τ, t and the diffusion coefficient ε.

Lemma 4.1. *Let $u \in W^{1,\infty}(H^{p+1})$. Then*

$$\|\Pi u(t) - u(t)\| \le Ch^{p+1}|u(t)|_{H^{p+1}}, \tag{4.1}$$

$$\|\Pi u'(t) - u'(t)\| \le Ch^{p+1}|u'(t)|_{H^{p+1}}, \tag{4.2}$$

$$\big((\Pi u - u)(s_1) - (\Pi u - u)(s_2), w\big) \le C|s_1 - s_2|h^{p+1}\|u\|_{W^{1,\infty}(H^{p+1})}|||w|||, \tag{4.3}$$

for all $w \in S_h$ and $s_1, s_2, t \in [0,T]$.

Proof. Estimates (4.1) and (4.2) are a standard estimate for the $L^2(\Omega)$-projection approximation. Estimate (4.3) can be found *e.g.* in [9]. □

We summarize the properties of the forms A_h and b_h.

Lemma 4.2. *Let $u \in H^{p+1}(\Omega)$. Then*

$$A_h(v,w) \le C|||v|||\,|||w|||, \quad \forall v, w \in S_h, \tag{4.4}$$

$$A_h(\Pi u - u, w) \le Ch^p|||w|||, \quad \forall w \in S_h. \tag{4.5}$$

Proof. The proof of (4.4) and (4.5) can be done in a similar way as in ([8], Lem. 9). □

Lemma 4.3. *Let $u \in H^{p+1}(\Omega) \cap W^{1,\infty}(\Omega)$. Then*

$$b_h(v,w) - b_h(\bar v, w) \le C\|v - \bar v\|\,|||w|||, \quad \forall v, \bar v, w \in S_h, \tag{4.6}$$

$$b_h(v, v - \Pi u) - b_h(u, v - \Pi u) \le C\left(1 + \frac{\|v-u\|_\infty^2}{h^2}\right)(h^{2p+1} + \|v - \Pi u\|^2), \quad \forall v \in S_h. \tag{4.7}$$

Proof. The proof of (4.6) can be found in [8]. The proof of (4.7) is essentially the same as that of ([16], Lem. 5.1), however there the statement and proof are written for the specific choice $v := u_h, \xi := u_h - \Pi u$. □

In the following analyses, it will be important to eliminate the unpleasant term $\|e(t)\|_\infty^2/h^2$ in (4.7), where $e = u_h - u$. This is possible if we know *a priori* that $\|e(t)\|_\infty = O(h)$. Since we are concerned in $L^2(\Omega)$-estimates, we want to reformulate this in terms of the $L^2(\Omega)$-norm. The following result is proven in [16], we include the proof here for convenience:

Lemma 4.4. *Let $p \ge d/2$ and u satisfy the regularity assumptions (5.1). Then*

$$\|e(t)\| \le h^{1+d/2} \quad \Longrightarrow \quad \|e(t)\|_\infty \le Ch,$$

where C is independent of h, ε, t.

Proof. We write the error as $e(t) = \eta(t) + \xi(t)$, where $\eta = \Pi u - u$ and $\xi = u_h - \Pi u \in S_h$. Due to standard approximation properties of Π and the inverse inequality between the L^∞ and L^2-norms, we have

$$\begin{aligned}\|e(t)\|_\infty &\le \|\eta(t)\|_\infty + \|\xi(t)\|_\infty \le Ch|u(t)|_{W^{1,\infty}} + Ch^{-d/2}\|\xi(t)\| \\ &\le Ch + Ch^{-d/2}\|e(t)\| + Ch^{-d/2}\|\eta(t)\| \le Ch + Ch^{p+1-d/2}|u(t)|_{H^{p+1}} \le Ch.\end{aligned}$$

□

5. Error estimates for BDF2

We want to estimate the error $e_h^m = U^m - u^m$, where the values of U^m are obtained by the BDF2-DG method. To do so, we construct a suitable continuation $U(t)$ (*i.e.* continuous function with respect to time) such that $U(t_m) = U^m$. Then we can also generalize the error as $e_h = U - u$. Our aim is to investigate the generalized error at arbitrary time $t \in (0,T)$ and prove a suitable *a priori* error bound. Then the error bound for the BDF2-DG method is a trivial consequence of the more general error estimate. For the purpose of analysis of the BDF2-DG scheme we assume following regularity

$$u \in W^{1,\infty}(H^{p+1}) \cap L^\infty(W^{1,\infty}) \cap W^{3,\infty}(L^2). \tag{5.1}$$

Definition 5.1. We define the continued approximate solution $U : [0,T] \to S_h$ of problem (2.1) obtained by the BDF2-DG scheme in the following way: Let $m \geq 2$ and $s \in [0,\tau]$, we seek $U(t_{m-1}+s) \in S_h$ such that

$$\begin{aligned}\Big(\frac{\tau+2s}{\tau+s}U(t_{m-1}+s) - \frac{\tau+s}{\tau}U^{m-1} + \frac{s^2}{\tau^2+\tau s}U^{m-2}, w\Big) + s\varepsilon A_h(U(t_{m-1}+s),w) + sb_h(U(t_{m-1}+s),w)\\ = s\ell_h(w)(t_{m-1}+s), \quad \forall w \in S_h.\end{aligned} \tag{5.2}$$

This defines U on I_m for $m \geq 2$. For $m = 1$ we define U on I_1 by seeking $U(s) \in S_h$ such that

$$(U(s) - U^0, w) + s\varepsilon A_h(U(s),w) + sb_h(U(s),w) = s\ell(w)(s), \quad \forall w \in S_h. \tag{5.3}$$

Remark 5.2. Equation (5.3) was already used for general m in [16] to define the continuation of the implicit Euler scheme. It represents the implicit Euler method with a variable time step s. By taking $s = 0$, we get $U(0) = U^0$, while setting $s = \tau$, we get $U(\tau) = U(t_1) = U^1$ and it can be proven that between these two values, $U(\cdot)$ changes continuously.

The motivation for (5.2) is similar. This equation is in fact the backward difference formula with variable time step, *cf.* [14]. Setting $s = 0$, we get $U(t_{m-1}) = U^{m-1}$, while setting $s = \tau$, we recover the original BDF2-DG scheme (3.2), hence $U(t_m) = U^m$. Similarly as in [16], we shall prove that between the $s = 0$ and $s = \tau$, $U(\cdot)$ changes continuously.

Lemma 5.3. *There exist constants $C_1, C_2 > 0$ independent of h, τ, t, ε, such that the following holds. Let $h \in (0, h_0)$ and $\tau \in [0, \tau_0)$, where $\tau_0 = \max\{C_1\varepsilon, C_2 h\}$. Then U, the continued solution from Definition 5.1 exists, is uniquely determined, $\|U(t)\|$ is uniformly bounded with respect to $t \in [0,T]$, $U(t_m) = U^m$ for all $m = 0, \dots, r$ and $\|U(t)\|$ depends continuously on t.*

Proof. For $m = 1$, it is already proven in [16] that the resulting solution U is continuous on I_1 and $U(0) = U^0, U(t_1) = U(\tau) = U^1$. Therefore it is sufficient to consider the case $m \geq 2$.

(i) **Existence**: Let $m \geq 2$ and $s \in [0,\tau]$, we consider U on I_m. We denote the left- and right-hand sides from (5.2):

$$\begin{aligned}B_s(v,w) &= \frac{\tau+2s}{\tau+s}(v,w) + s\varepsilon A_h(v,w) + sb_h(v,w),\\ L_s^m(w) &= \left(\frac{\tau+s}{\tau}U^{m-1} - \frac{s^2}{\tau^2+\tau s}U^{m-2}, w\right) + s\ell_h(w)(t_{m-1}+s).\end{aligned}$$

We will show that B_s is strictly monotone and Lipschitz continuous on S_h equipped with the $L^2(\Omega)$-scalar product. Existence and uniqueness then follows from the nonlinear Lax–Milgram lemma, *cf.* [23].

Monotonicity: using the ellipticity of A_h, the boundedness of b_h and the inverse inequality, we get

$$\begin{aligned}B_s(v,v-w) - B_s(w,v-w) &\geq \frac{\tau+2s}{\tau+s}\|v-w\|^2 + s\varepsilon|\!|\!|v-w|\!|\!|^2 - Cs\|v-w\|\,|\!|\!|v-w|\!|\!|\\ &\geq \left(1 - \frac{Cs}{h}\right)\|v-w\|^2 = M\|v-w\|^2, \quad \forall v,w \in S_h,\end{aligned}$$

for s, τ sufficiently small with respect to h.

On the other hand, we may estimate using Young's inequality:

$$\begin{aligned} B_s(v,v-w)-B_s(w,v-w) &\geq \frac{\tau+2s}{\tau+s}\|v-w\|^2+s\varepsilon|\!|\!|v-w|\!|\!|^2-Cs\|v-w\||\!|\!|v-w|\!|\!| \\ &\geq \|v-w\|^2+s\varepsilon|\!|\!|v-w|\!|\!|^2-s\varepsilon|\!|\!|v-w|\!|\!|^2-\frac{C^2s}{4\varepsilon}\|v-w\|^2 \\ &\geq \left(1-\frac{C^2s}{4\varepsilon}\right)\|v-w\|^2=M\|v-w\|^2, \quad \forall v,w\in S_h, \end{aligned} \tag{5.4}$$

in this case we get the condition s,τ sufficiently small with respect to ε.

Lipschitz continuity: We shall show that B_s is Lipschitz continuous:

$$\begin{aligned} B_s(v,w)-B_s(\bar v,w) &\leq \frac{3}{2}\|v-\bar v\|\,\|w\|+Cs\varepsilon|\!|\!|v-\bar v|\!|\!|\,|\!|\!|w|\!|\!|+Cs\|v-\bar v\|\,|\!|\!|w|\!|\!| \\ &\leq \Big(\frac{3}{2}+\frac{Cs\varepsilon}{h^2}+\frac{Cs}{h}\Big)\|v-\bar v\|\,\|w\|=L\|v-\bar v\|\,\|w\|. \end{aligned}$$

Since the right-hand side L_s^m is a linear functional on the finite-dimensional space S_h, it is also bounded and by the nonlinear Lax–Milgram lemma we obtain the existence and uniqueness of the continued discrete solution and classical discrete solution, respectively. Finally, we obtain the uniform boundedness of $\|U(t)\|$ w.r.t. $t\in I_m$, since the nonlinear Lax–Milgram lemma gives us $\|U(t)\|\leq C\|L_s^m\|_{\mathcal{L}(L^2(\Omega),\mathbb{R})}$, which can be bounded independent of s similarly as in [16].

Since we have existence and uniqueness, we see that $U(t_m)=U^m$ by setting $s=\tau$ in (5.2).

(ii) **Continuity**: Now we show that the continued discrete solution is continuous with respect to time. Let $m>1$ and $t,\bar t\in(t_{m-1},t_m]$ and $s=t-t_{m-1},\ \bar s=\bar t-t_{m-1}$. Then by monotonicity,

$$\begin{aligned} M\|U(t)-U(\bar t)\|^2 &\leq B_t(U(t),U(t)-U(\bar t))-B_t(U(\bar t),U(t)-U(\bar t)) \\ &= L_t^m(U(t)-U(\bar t))-L_{\bar t}^m(U(t)-U(\bar t))+B_{\bar t}(U(\bar t),U(t)-U(\bar t))-B_t(U(\bar t),U(t)-U(\bar t)). \end{aligned} \tag{5.5}$$

We estimate the B and L terms individually.

$$\begin{aligned} &|B_{\bar t}(U(\bar t),U(t)-U(\bar t))-B_t(U(\bar t),U(t)-U(\bar t))| \\ &\quad\leq \Big|\frac{\tau+2\bar s}{\tau+\bar s}-\frac{\tau+2s}{\tau+s}\Big|\,|\!|\!|U(\bar t)|\!|\!|\,\|U(t)-U(\bar t)\|+|\bar s-s|\varepsilon A_h(U(\bar t),U(t)-U(\bar t))+|\bar s-s|b_h(U(\bar t),U(t)-U(\bar t)) \\ &\quad\leq |\bar s-s|\big(\tau+\frac{C\varepsilon}{h^2}++\frac{C}{h}\big)|\!|\!|U(\bar t)|\!|\!|\,\|U(t)-U(\bar t)\|. \end{aligned} \tag{5.6}$$

Similarly we get

$$\begin{aligned} &|L_t^m(U(t)-U(\bar t))-L_{\bar t}^m(U(t)-U(\bar t))| \leq \big(\Big|\frac{\tau+\bar s}{\tau}-\frac{\tau+s}{\tau}\Big|\,|\!|\!|U^{m-1}|\!|\!|+\Big|\frac{\bar s^2}{\tau^2+\tau\bar s}-\frac{s^2}{\tau^2+\tau s}\Big|\,|\!|\!|U^{m-2}|\!|\!|\big)\|U(t)-U(\bar t)\| \\ &\quad+|s\ell_h(U(t)-U(\bar t))(t)-\bar s\ell_h(U(t)-U(\bar t))(\bar t)| \\ &\leq |\bar s-s|\big(\tau^{-1}|\!|\!|U^{m-1}|\!|\!|+3\tau^{-1}|\!|\!|U^{m-2}|\!|\!|\big)\|U(t)-U(\bar t)\|+|s\ell_h(U(t)-U(\bar t))(t)-\bar s\ell_h(U(t)-U(\bar t))(\bar t)|. \end{aligned} \tag{5.7}$$

Assuming $|\bar s-s|=|\bar t-t|\to 0$, we get the limit for the terms on the last row

$$|s\ell_h(U(t)-U(\bar t))(t)-\bar s\ell_h(U(t)-U(\bar t))(\bar t)| \leq \underbrace{|s-\bar s|}_{\to 0}\,\underbrace{|\ell_h(U(t)-U(\bar t))(t)|}_{\text{bounded}}+\bar s|\underbrace{(g(t)-g(\bar t)}_{\to 0},\underbrace{U(t)-U(\bar t))}_{\text{bounded}}|\to 0,$$

since $\|U(t)\|,\|U(\bar t)\|$ are uniformly bounded with respect to $t,\bar t\in(t_{m-1},t_m]$.

From now it is possible to see that the terms in (5.6) and (5.7) tend to zero as $|t-\bar t|$ tends to zero. Together with (5.5) we get

$$\|U(t)-U(\bar t)\|\to 0 \text{ as } |\bar s-s|=|\bar t-t|\to 0.$$

Now, we prove the continuity at t_{m-1}, *i.e.* $U(t_{m-1}+s) \to U^{m-1}$ as s tends to $0+$. Since $\frac{\tau+2s}{\tau+s} \to 1$, $\frac{\tau+s}{\tau} \to 1$, $\frac{s^2}{\tau^2+\tau s} \to 0$ and the terms $A_h(U(t_{m-1}+s),w)$, $b_h(U(t_{m-1}+s),w)$ and $\ell_h(w)$ are bounded, we get from (5.2)

$$\underbrace{\frac{\tau+2s}{\tau+s}(U(t_{m-1}+s),w)}_{\to(U(t_{m-1}+s),w)} - \underbrace{\frac{\tau+s}{\tau}(U^{m-1},w)}_{\to(U^{m-1},w)} + \underbrace{\frac{s^2}{\tau^2+\tau s}(U^{m-2},w)}_{\to 0}$$
$$+ \underbrace{s\varepsilon A_h(U(t_{m-1}+s),w) + sb_h(U(t_{m-1}+s),w)}_{\to 0} = \underbrace{s\ell_h(w)(t_{m-1}+s)}_{\to 0},$$

i.e. continuity at t_{m-1}.

It remains to prove continuity of $U(\cdot)$ on I_1. In the case of computing initial condition by (3.3), we can continuate the solution on $I_1 = [0,\tau]$ by

$$(U(s)-U^0,w) + s\varepsilon A_h(U(s),w) + sb_h(U(s),w) = s\ell(w)(s).$$

It is already proved in [16] that such a continuation is continuous on $[0,\tau]$. □

Due to the regularity assumptions (5.1) the exact solution $u \in C([0,T];L^2(\Omega))$ and therefore uniformly continuous on the closed interval $[0,T]$. Therefore, by Lemma 5.3, the error $e_h = U(t) - u(t)$ is also uniformly continuous. We divide the error $e_h = \xi + \eta$, where $\xi = U - \Pi u$ and $\eta = \Pi u - u$.

Lemma 5.4. *Let u satisfy regularity assumptions (5.1). Let $s \in (0,\tau]$. Then*

$$\Big(\frac{\tau+2s}{\tau+s}u(t_{m-1}+s) - \frac{\tau+s}{\tau}u^{m-1} + \frac{s^2}{\tau^2+\tau s}u^{m-2} - su'(t_{m-1}+s), w\Big) \le Cs\tau^2\|u\|_{W^{3,\infty}(L^2)}\|w\|, \tag{5.8}$$

$$\big(u(s) - u^0 - su'(s), w\big) \le Cs\tau\|u\|_{W^{2,\infty}(L^2)}\|w\|, \tag{5.9}$$

$$\Big(\frac{\tau+2s}{\tau+s}\eta(t_{m-1}+s) - \frac{\tau+s}{\tau}\eta^{m-1} + \frac{s^2}{\tau^2+\tau s}\eta^{m-2}, w\Big) \le Csh^{p+1}\|u\|_{W^{1,\infty}(H^{p+1})}\|w\|. \tag{5.10}$$

Proof. Let us denote $y = t_{m-1}+s$. Since

$$\frac{\tau+s}{\tau} - \frac{s^2}{\tau^2+\tau s} = \frac{\tau+2s}{\tau+s}, \qquad \frac{\tau s+s^2}{\tau} - \frac{\tau s^2+s^3}{\tau^2+\tau s} = s, \qquad \frac{\tau s^2+s^3}{2\tau} - \frac{s^2(\tau+s)^2}{2\tau^2+2\tau s} = 0,$$

we can formally rewrite

$$\frac{\tau+2s}{\tau+s}u(y) - \frac{\tau+s}{\tau}u^{m-1} + \frac{s^2}{\tau^2+\tau s}u^{m-2} - su'(y) = \frac{\tau+s}{\tau}\left(u(y) - su'(y) + \frac{s^2}{2}u''(y) - u^{m-1}\right)$$
$$- \frac{s^2}{\tau^2+\tau s}\left(u(y) - (\tau+s)u'(y) + \frac{(\tau+s)^2}{2}u''(y) - u^{m-2}\right) \tag{5.11}$$

and

$$\frac{\tau+2s}{\tau+s}\eta(y) - \frac{\tau+s}{\tau}\eta^{m-1} + \frac{s^2}{\tau^2+\tau s}\eta^{m-2} = \frac{\tau+s}{\tau}(\eta(y)-\eta^{m-1}) - \frac{s^2}{\tau^2+\tau s}(\eta(y)-\eta^{m-2}). \tag{5.12}$$

Then it is simple to see

$$\frac{\tau+s}{\tau}\left(u(y) - su'(y) + \frac{s^2}{2}u''(y) - u^{m-1}, w\right) = \frac{\tau+s}{\tau}\int_{t_{m-1}}^{y}\int_{z_1}^{y}\int_{z_2}^{y}(u'''(z_3),w)\mathrm{d}z_3\mathrm{d}z_2\mathrm{d}z_1$$
$$\le \frac{\tau+s}{\tau}\|u\|_{W^{3,\infty}(L^2)}\|w\|\int_{t_{m-1}}^{y}\int_{z_1}^{y}\int_{z_2}^{y}1\mathrm{d}z_3\mathrm{d}z_2\mathrm{d}z_1 = \frac{\tau+s}{\tau}\frac{s^3}{6}\|u\|_{W^{3,\infty}(L^2)}\|w\| \le Cs\tau^2\|u\|_{W^{3,\infty}(L^2)}\|w\|$$

and

$$\begin{aligned}\frac{s^2}{\tau^2+\tau s}\left(u(y)-(\tau+s)u'(y)+\frac{(\tau+s)^2}{2}u''(y)-u^{m-2},w\right)&=\frac{s^2}{\tau^2+\tau s}\int_{t_{m-2}}^{y}\int_{z_1}^{y}\int_{z_2}^{y}(u'''(z_3),w)\mathrm{d}z_3\mathrm{d}z_2\mathrm{d}z_1\\&\le\frac{s^2}{\tau^2+\tau s}\|u\|_{W^{3,\infty}(L^2)}\|w\|\int_{t_{m-2}}^{y}\int_{z_1}^{y}\int_{z_2}^{y}1\mathrm{d}z_3\mathrm{d}z_2\mathrm{d}z_1\\&=\frac{s^2}{\tau(\tau+s)}\frac{(\tau+s)^3}{6}\|u\|_{W^{3,\infty}(L^2)}\|w\|\le Cs\tau^2\|u\|_{W^{3,\infty}(L^2)}\|w\|,\end{aligned}$$

which proves (5.8). The proof of (5.9) follows from

$$(u(s)-u^0-su'(s),w)=-\int_0^s\int_{z_1}^{s}(u''(z_2),w)\mathrm{d}z_2\mathrm{d}z_1\le\frac{1}{2}s^2\|u\|_{W^{2,\infty}(L^2)}\|w\|.$$

The proof of (5.10) follows directly from (5.12) and Lemma 4.1. □

In the proof of the error estimate of Lemma 5.7, we will need to estimate the BDF coefficients at $U(t_{m-1}+s)$, U^{m-1},U^{m-2} in (5.2). For this purpose we define the sequence $\{\gamma_j\}_{j=0}^{\infty}$ by

$$\begin{aligned}\gamma_0&=\frac{\tau+s}{\tau+2s},\\\frac{3}{2}\gamma_1-\frac{\tau+s}{\tau}\gamma_0&=0,\\\frac{3}{2}\gamma_2-2\gamma_1+\frac{s^2}{\tau^2+\tau s}\gamma_0&=0,\\\frac{3}{2}\gamma_{j+2}-2\gamma_{j+1}+\frac{1}{2}\gamma_j&=0,\quad\forall j=1,2,3,\ldots\end{aligned}\tag{5.13}$$

Lemma 5.5. *Let the sequence $\{\gamma_j\}_{j=0}^{\infty}$ be defined by (5.13). Then such a sequence is positive and bounded,* i.e. $0<\gamma_j<\gamma_\infty$ *for all $j=0,1,\ldots$ for some $\gamma_\infty\in\mathbb{R}$. Moreover,*

$$\gamma_1-2\frac{s^2}{\tau^2+\tau s}\gamma_0>0\tag{5.14}$$

and for $j\ge1$ the sequence γ_j is increasing.

Proof. Let us calculate the initial values for γ_j. γ_0 is defined already by (5.13).

$$\begin{aligned}\gamma_1&=\frac{2}{3}\frac{(\tau+s)^2}{\tau(\tau+2s)},\\\gamma_2&=\frac{8}{9}\frac{(\tau+s)^2}{\tau(\tau+2s)}-\frac{2}{3}\frac{s^2}{\tau(\tau+2s)}.\end{aligned}$$

From this (5.14) immediately follows. For $j=1,2,\ldots$, γ_j are defined by a difference equation with the initial condition γ_1 and γ_2 and with the solution

$$\gamma_j=\left(\frac{(\tau+s)^2}{\tau(\tau+2s)}-\frac{s^2}{\tau(\tau+2s)}\right)+\left(\frac{1}{3}\frac{s^2}{\tau(\tau+2s)}-\frac{1}{9}\frac{(\tau+s)^2}{\tau(\tau+2s)}\right)\left(\frac{1}{3}\right)^{j-2}.$$

Since

$$\frac{(\tau+s)^2}{\tau(\tau+2s)} - \frac{s^2}{\tau(\tau+2s)} > 0,$$
$$\frac{1}{3}\frac{s^2}{\tau(\tau+2s)} - \frac{1}{9}\frac{(\tau+s)^2}{\tau(\tau+2s)} < 0,$$

we can see that the sequence γ_j is increasing, positive and bounded. □

Let us start with the result on the initial condition defined by (3.3).

Lemma 5.6. *Let $p > d/2$. Let $s \in (0,\tau]$. If $\|e(t)\| \le h^{1+d/2}$ for $t \in [0,s]$, then*

$$\sup_{t\in[0,s]} \|e(t)\|^2 \le \tilde{C}_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^4),$$

where the constant $\tilde{C}_T$ is independent of h, τ, ε.

Proof. Since $U^0 = \Pi u^0$ we can see that $\|e^0\| \le Ch^{p+1}$. Multiplying (3.1) for $t = s$ by s, subtracting from (5.3) and adding several terms we get

$$\begin{aligned}\big(\xi(s) - \xi^0, w\big) + s\varepsilon A_h(\xi(s), w) = \big(su'(s) - u(s) + u^0, w\big) - \big(\eta(s) - \eta^0, w\big)\\ + s\big(b_h\left(u(s), w\right) - b_h\left(U(s), w\right)\big) - s\varepsilon A_h(\eta(s), w).\end{aligned}$$

Setting $w = 2\xi(s)$ and using Lemmas 4.1, 4.2, 4.3 and 5.4, we get

$$\begin{aligned}\|\xi(s)\|^2 - \|\xi^0\|^2 + \|\xi(s) - \xi^0\|^2 + s\varepsilon|\!|\!|\xi(s)|\!|\!|^2 \le C\tau\left(1 + \frac{\|e(s)\|_\infty^2}{h^2}\right)(h^{2p+1} + \|\xi(s)\|^2)\\ + C\tau^4 + Ch^{2p+2} + C\varepsilon h^{2p} + \frac{1}{2}\|\xi(s)\|^2.\end{aligned}$$

Using the assumptions and Lemma 4.4, we can get rid of the unpleasant term $\|e(s)\|_\infty^2/h^2$ and we get

$$\|\xi(s)\|^2 \le C(\|\xi^0\|^2 + h^{2p+1} + \varepsilon h^{2p} + \tau^4).$$

The proof is completed by taking similar estimates for η and the triangle inequality to estimate $e(s)$. □

Now, we extend Lemma 5.6 to the rest of $[0,T]$ by analyzing the BDF scheme (5.2).

Lemma 5.7. *Let $p > d/2$. Let $n > 0$ and $s \in (0,\tau]$. If $\|e(t)\| \le h^{1+d/2}$ for $t \in [0, t_{n-1}+s]$, then*

$$\sup_{t\in[0,t_{n-1}+s]} \|e(t)\|^2 \le C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^4),$$

where the constant C_T is independent of h, τ, ε.

Proof. To simplify the relations we set $y = t_{n-1} + s$. Multiplying (3.1) for $t = y$ by s, subtracting from (5.2) and adding several terms we get

$$\begin{aligned}&\left(\frac{\tau+2s}{\tau+s}\xi(y) - \frac{\tau+s}{\tau}\xi^{n-1} + \frac{s^2}{\tau^2+\tau s}\xi^{n-2}, w\right) + s\varepsilon A_h(\xi(y), w)\\ &= \left(su'(y) - \frac{\tau+2s}{\tau+s}u(y) + \frac{\tau+s}{\tau}u^{n-1} - \frac{s^2}{\tau^2+\tau s}u^{n-2}, w\right) - \left(\frac{\tau+2s}{\tau+s}\eta(y) - \frac{\tau+s}{\tau}\eta^{n-1} + \frac{s^2}{\tau^2+\tau s}\eta^{n-2}, w\right)\\ &\quad + s\left(b_h\left(u(y), w\right) - b_h\left(U(y), w\right)\right) - s\varepsilon A_h(\eta(y)), w).\end{aligned}$$

For lower time levels $m \leq n-1$ we obtain analogically

$$\left(\frac{3}{2}\xi^m - 2\xi^{m-1} + \frac{1}{2}\xi^{m-2}, w\right) + \tau\varepsilon A_h(\xi^m, w) = \left(\tau u'(t_m) - \frac{3}{2}u^m + 2u^{m-1} - \frac{1}{2}u^{m-2}, w\right)$$
$$- \left(\frac{3}{2}\eta^m - 2\eta^{m-1} + \frac{1}{2}\eta^{m-2}, w\right) + \tau\left(b_h\left(u^m, w\right) - b_h\left(U^m, w\right)\right) - \tau\varepsilon A_h(\eta^m, w).$$

Setting $w = 2\xi(y)$ we obtain on the left-hand side using the fact $s \in (0, \tau]$

$$\begin{aligned}
&2\left(\frac{\tau+2s}{\tau+s}\xi(y) - \frac{\tau+s}{\tau}\xi^{n-1} + \frac{s^2}{\tau^2+\tau s}\xi^{n-2}, \xi(y)\right) + 2s\varepsilon A_h(\xi(y), \xi(y)) \\
&= 2\frac{\tau+s}{\tau}(\xi(y) - \xi^{n-1}, \xi(y)) - 2\frac{s^2}{\tau^2+\tau s}(\xi(y) - \xi^{n-2}, \xi(y)) + 2s\varepsilon|||\xi(y)|||^2 \\
&= \frac{\tau+s}{\tau}(\|\xi(y)\|^2 - \|\xi^{n-1}\|^2 + \|\xi(y) - \xi^{n-1}\|^2) - \frac{s^2}{\tau^2+\tau s}(\|\xi(y)\|^2 - \|\xi^{n-2}\|^2 + \|\xi(y) - \xi^{n-2}\|^2) + 2s\varepsilon|||\xi(y)|||^2 \\
&\geq \frac{\tau+2s}{\tau+s}\|\xi(y)\|^2 - \frac{\tau+s}{\tau}\|\xi^{n-1}\|^2 + \frac{s^2}{\tau^2+\tau s}\|\xi^{n-2}\|^2 + \frac{\tau+s}{\tau}\|\xi(y) - \xi^{n-1}\|^2 \\
&\quad - 2\frac{s^2}{\tau^2+\tau s}\|\xi(y) - \xi^{n-1}\|^2 - 2\frac{s^2}{\tau^2+\tau s}\|\xi^{n-1} - \xi^{n-2}\|^2 + 2s\varepsilon|||\xi(y)|||^2 \\
&\geq \frac{\tau+2s}{\tau+s}\|\xi(y)\|^2 - \frac{\tau+s}{\tau}\|\xi^{n-1}\|^2 + \frac{s^2}{\tau^2+\tau s}\|\xi^{n-2}\|^2 + \frac{s}{\tau}\|\xi(y) - \xi^{n-1}\|^2 \\
&\quad - 2\frac{s^2}{\tau^2+\tau s}\|\xi^{n-1} - \xi^{n-2}\|^2 + 2s\varepsilon|||\xi(y)|||^2.
\end{aligned}$$

Setting $s = \tau$ (*i.e.* with $w = 2\xi^m$), the relations simplify to the usual

$$2\left(\frac{3}{2}\xi^m - 2\xi^{m-1} + \frac{1}{2}\xi^{m-2}, \xi^m\right) + 2\tau\varepsilon A_h(\xi^m, \xi^m)$$
$$\geq \frac{3}{2}\|\xi^m\|^2 - 2\|\xi^{m-1}\|^2 + \frac{1}{2}\|\xi^{m-2}\|^2 + \|\xi^m - \xi^{m-1}\|^2 - \|\xi^{m-1} - \xi^{m-2}\|^2 + 2\tau\varepsilon|||\xi^m|||^2.$$

Using Lemmas 4.1, 4.2, 4.3 and 5.4 to estimate the right-hand side terms, we get

$$\frac{\tau+2s}{\tau+s}\|\xi(y)\|^2 - \frac{\tau+s}{\tau}\|\xi^{n-1}\|^2 + \frac{s^2}{\tau^2+\tau s}\|\xi^{n-2}\|^2 - 2\frac{s^2}{\tau^2+\tau s}\|\xi^{n-1} - \xi^{n-2}\|^2$$
$$\leq Cs\left(1 + \frac{\|e(y)\|_\infty^2}{h^2}\right)(\varepsilon h^{2p} + h^{2p+1} + \tau^4 + \|\xi(y)\|^2)$$

and

$$\frac{3}{2}\|\xi^m\|^2 - 2\|\xi^{m-1}\|^2 + \frac{1}{2}\|\xi^{m-2}\|^2 + \|\xi^m - \xi^{m-1}\|^2 - \|\xi^{m-1} - \xi^{m-2}\|^2$$
$$\leq C\tau\left(1 + \frac{\|e^m\|_\infty^2}{h^2}\right)(\varepsilon h^{2p} + h^{2p+1} + \tau^4 + \|\xi^m\|^2).$$

Using the assumptions and Lemma 4.4, we can eliminate the terms $\|e(y)\|_\infty^2/h^2$ and $\|e^m\|_\infty^2/h^2$:

$$\frac{\tau+2s}{\tau+s}\|\xi(y)\|^2 - \frac{\tau+s}{\tau}\|\xi^{n-1}\|^2 + \frac{s^2}{\tau^2+\tau s}\|\xi^{n-2}\|^2 - 2\frac{s^2}{\tau^2+\tau s}\|\xi^{n-1} - \xi^{n-2}\|^2$$
$$\leq Cs(\varepsilon h^{2p} + h^{2p+1} + \tau^4 + \|\xi(y)\|^2), \tag{5.15}$$

$$\frac{3}{2}\|\xi^m\|^2 - 2\|\xi^{m-1}\|^2 + \frac{1}{2}\|\xi^{m-2}\|^2 + \|\xi^m - \xi^{m-1}\|^2 - \|\xi^{m-1} - \xi^{m-2}\|^2$$
$$\leq C\tau(\varepsilon h^{2p} + h^{2p+1} + \tau^4 + \|\xi^m\|^2). \tag{5.16}$$

Multiplying (5.15) by γ_0 and (5.16) by γ_{n-m} for $m = 2, \dots, n-1$, where the sequence $\{\gamma_j\}_{j=0}^{\infty}$ is defined by (5.13), and by summing all these inequalities together, we get by Lemma 5.5

$$\|\xi(y)\|^2 \le Cs\gamma_0\|\xi(y)\|^2 + C\gamma_\infty(\|\xi^1\|^2 + \|\xi^0\|^2) + C\tau\gamma_\infty \sum_{j=2}^{n-1} \|\xi^j\|^2 + C\gamma_\infty y(\varepsilon h^{2p} + h^{2p+1} + \tau^4).$$

Analogically we can obtain a similar result for $\|\xi^m\|$ for $m = 2, \dots, n-1$. Only this time the sequence for $\{\gamma_j\}_{j=0}^{\infty}$ used for multiplying the equations is modified by taking (5.13) with $s = \tau$:

$$\|\xi^m\|^2 \le Cs\gamma_0\|\xi^m\|^2 + C\gamma_\infty(\|\xi^1\|^2 + \|\xi^0\|^2) + C\tau\gamma_\infty \sum_{j=2}^{m-1} \|\xi^j\|^2 + C\gamma_\infty t_m(\varepsilon h^{2p} + h^{2p+1} + \tau^4).$$

Since $\|\xi^1\|^2$ and $\|\xi^0\|^2$ are bounded according to Lemma 5.6, we obtain the result using the discrete Gronwall lemma. □

Now we get rid of the *a priori* assumption $\|e(t)\| \le h^{1+d/2}$ from Lemmas 5.6 and 5.7.

Theorem 5.8. *Let $p > 1 + d/2$. Let τ_0 be defined as in Lemma 5.3. Let $h \in (0, h_0)$ and $\tau_1 \in (0, \tau_0)$ be such that*

$$C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^4) \le \frac{1}{4}h^{2+d}, \tag{5.17}$$

where C_T is the constant from Lemma 5.7 independent of h, τ, ε. Then the error of the BDF2-DG scheme satisfies

$$\sup_{t\in[0,T]} \|e(t)\|^2 \le C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^4). \tag{5.18}$$

Proof. We will follow the idea of continuous mathematical induction from [16]. Since the proof essentially follows the same pattern therein, we only give a brief description without details.

For time $t = 0$ it is easy to see that the error estimate holds, because the error is in fact the error of L^2 projection in initial data, which is sufficiently small under the assumptions of the theorem. Let us assume that the error estimate (5.18) holds on the interval $[0, s]$ for some $s \in [0, T]$. According to the assumption (5.17) we can see that the error can be estimated by $\|e(t)\| \le \frac{1}{2}h^{1+d/2}$, $t \in [0, s]$. Since the error $e(\cdot)$ is continuous (even uniformly continuous) with respect to time, we know that there exists some $\delta > 0$ such that $\|e(t)\| \le h^{1+d/2}$, $t \in [0, s+\delta]$ and we can see that it is possible to use Lemma 5.7 on the larger interval $[0, s+\delta]$, which guarantees the error estimate (5.18) on $[0, s+\delta]$. Since the error is *uniformly* continuous in time, we have a fixed $\delta > 0$ independent of s during the induction process and using the argument repeatedly we obtain the result up to $s = T$. □

Remark 5.9. The condition (5.17) can be essentially split into two parts, *e.g.* $C_T^2(h^{2p+1} + \varepsilon h^{2p}) \le \frac{1}{8}h^{2+d}$ and $C_T^2\tau^4 \le \frac{1}{8}h^{2+d}$. The first condition can be satisfied for sufficiently small h only if $p > 1 + d/2$. The second condition is satisfied only if the CFL-like condition $\tau = O(h^{1/2+d/4})$ holds. Of course, we still need the continued error $e(\cdot)$ to exist uniquely and be continuous in time, for which we need $\tau = O(\max\{\varepsilon, h\})$ by Lemma 5.3.

Remark 5.10. We note that if $\varepsilon = 0$, we obtain the improved estimate $O(h^{p+1/2} + \tau^2)$ under the weaker condition $p > (1 + d)/2$. This is also the case for Theorems 6.6 and 7.15 for the midpoint rule and QT-DG scheme.

The reader might ask why such an elaborate construction of the continuation as (5.2) is used, why not use *e.g.* some simple interpolation in time. In the proof of the estimates we proceed by induction from one time node to the next. Starting from the error of the initial condition, we want to prove that if the error e^{m-1}

at t_{m-1} is of the desired order, *e.g.* $O(h^{p+1/2})$, then so is e^m. The estimates of the convective terms allow us to do this if we know *a priori* that $\|e^m\| = O(h^{1+d/2})$. But in [16] it is proven, given the presented estimates, that the implication $\|e^{m-1}\| = O(h^{p+1/2}) \implies \|e^m\| = O(h^{1+d/2})$ does not hold for implicit schemes (the proof is for the backward Euler scheme, however exactly the same reasoning holds *e.g.* for the BDF2 scheme). The proposed solution is to work with a continuous in time variant of the error, not discrete, and the continuity will help us go from t_{m-1} to t_m *via* suitably small intermediate steps while satisfying the necessary assumption $\|e\| = O(h^{1+d/2})$ along the way simply by continuity. Therefore the three requirements on the continuation are that $e(t)$ is continuous w.r.t. t, that it coincides with e^{m-1} and e^m at t_{m-1}, t_m and that it has the same order of approximation in time as the analyzed scheme for all t. If we used *e.g.* a simple Lagrange interpolation of $e^m, e^{m-1}, \dots$ or $U^m, U^{m-1}, \dots$, in order to prove anything about this interpolation between t_{m-1}, t_m we would first need to know the behavior of the interpolated function at the interpolation nodes, including the last one: t_m. In other words, we would need to have estimates for e^m in advance, which we do not, we only have estimates for e^{m-1} and earlier. In our approach, for $t \in (t_{m-1}, t_m)$ the continuation is constructed only from U^{m-1}, U^{m-2} without any knowledge of U^m or e^m. We start from t_{m-1} and by varying the time step in a variable time step BDF2 scheme, we go continuously from t_{m-1} to t_m and obtain U^m at t_m in a natural way. The estimates for the continuation are therefore obtained only from estimates of e^{m-1}, e^{m-2} (which we have from the induction assumption) while having the advantage of continuity in time to help control the *a priori* assumption $\|e\| = O(h^{1+d/2})$. To work with the Lagrange interpolation of the error in time, we would need to know not only the behavior at e^{m-1}, but also at e^m as stated earlier. Another possibility, to use some form of extrapolation from $t_{m-1}, t_{m-2}, \dots$ would also not work, since at t_m we would not obtain U^m and therefore would not be estimating the BDF2 scheme but some different extrapolated solution.

Moreover, since our continuation is constructed using the BDF2 scheme itself (albeit with variable coefficients), the analysis of its properties is done using tools that would be used anyway. Perhaps a slightly simpler form than (5.2) could be possible, but given the presented reasoning, in the end it must be some variation on the BDF2 scheme itself, not simple interpolation.

6. Error estimates for the Midpoint rule

In this section, we investigate the error estimates of the approximate solution U^m, $m = 0, \dots, r$ obtained by the method (3.5). As in the case of the BDF2-DG scheme, we construct a continuous extension of the discrete solution similar to Definition 5.1. For the purpose of analysis of the midpoint-DG scheme we assume the following regularity

$$u \in W^{1,\infty}(H^{p+1}) \cap W^{2,\infty}\big(H^2 \cap W_0^{1,\infty}\big) \cap W^{3,\infty}(L^2) \tag{6.1}$$

Definition 6.1. We define the continued approximate solution $U : [0,T] \to S_h$ of problem (2.1) obtained by the midpoint-DG scheme in the following way: Let $m > 0$ and $s \in [0, \tau]$, we seek $U(t_{m-1} + s) \in S_h$ such that

$$\begin{aligned}\big(U(t_{m-1}+s) - U^{m-1}, w\big) + \frac{s\varepsilon}{2} A_h\big(U(t_{m-1}+s) + U^{m-1}, w\big) + s b_h\left(\frac{U(t_{m-1}+s) + U^{m-1}}{2}, w\right)\\ = s\ell_h(w)(t_{m-1} + s/2), \quad \forall w \in S_h.\end{aligned} \tag{6.2}$$

As in Definition 5.1, by setting $s := 0$, we obtain $U(t_{m-1}) = U^{m-1}$. By setting $s := \tau$, we obtain $U(t_m) = U^m$.

Similarly as for the BDF2 scheme we can prove existence, uniqueness and time-continuity of the continued midpoint-DG solution from Definition 6.1.

Lemma 6.2. *There exist constants $C_1, C_2 > 0$ independent of h, τ, t, ε, such that the following holds. Let $h \in (0, h_0)$ and $\tau \in [0, \tau_0)$, where $\tau_0 = \max\{C_1\varepsilon, C_2 h\}$. Then U, the continued solution from Definition 6.1 exists, is uniquely determined, $\|U(t)\|$ is uniformly bounded with respect to $t \in [0,T]$, $U(t_m) = U^m$ for all $m = 0, \dots, r$ and $\|U(t)\|$ depends continuously on t.*

Proof.
(i) **Existence**: We denote the left- and right-hand side from (6.2)

$$B_s^m(v,w) = (v - U^{m-1}, w) + \frac{s\varepsilon}{2} A_h(v + U^{m-1}, w) + s b_h\left(\frac{v + U^{m-1}}{2}, w\right),$$
$$L_s^m(w) = s\ell_h(w)(t_{m-1} + s/2).$$

Then B_s^m is strongly monotone on S_h:

$$\begin{aligned} B_s^m(v, v-w) - B_s^m(w, v-w) &\geq \|v-w\|^2 + \frac{s\varepsilon}{2}\|\!|v-w|\!\|^2 - Cs\|v-w\|\,\|\!|v-w|\!\| \\ &\geq \left(1 - \frac{Cs}{h}\right)\|v-w\|^2 = M\|v-w\|^2 \end{aligned}$$

for sufficiently small s, τ with respect to h. On the other hand, we may estimate using Young's inequality as in (5.4) to obtain monotonicity for s, τ sufficiently small with respect to ε.

Now, we show that B_s^m is Lipschitz continuous on S_h:

$$\begin{aligned} B_s^m(v,w) - B_s^m(\bar{v},w) &\leq \|v-w\|\,\|w\| + C\frac{s\varepsilon}{2}\|\!|v-\bar{v}|\!\|\,\|\!|w|\!\| + Cs\|v-\bar{v}\|\,\|\!|w|\!\| \\ &\leq \left(1 + \frac{Cs\varepsilon}{h^2} + \frac{Cs}{h}\right)\|v-\bar{v}\|\,\|w\| = C\|v-\bar{v}\|\,\|w\|. \end{aligned}$$

The right-hand side L_s^m is bounded, hence continuous, on S_h the nonlinear Lax-Milgram lemma gives us existence and uniqueness of the continued discrete solution and classical discrete solution, respectively.

(ii) **Continuity**: Continuity with respect to time can be proved in the same way as in the proof of Lemma 5.3. Again, we use the monotonicity of the form B_s^m and write

$$\begin{aligned} M\|U(t) - U(\bar{t})\|^2 &\leq B_t^m(U(t), U(t) - U(\bar{t})) - B_t^m(U(\bar{t}), U(t) - U(\bar{t})) \\ &= L_t^m(U(t) - U(\bar{t})) - L_{\bar{t}}^m(U(t) - U(\bar{t})) + B_{\bar{t}}^m(U(\bar{t}), U(t) - U(\bar{t})) - B_t^m(U(\bar{t}), U(t) - U(\bar{t})). \end{aligned}$$

Similarly as in BDF case we can estimate the terms on the second and third row and prove that they tend to zero as $|t - \bar{t}|$ tends to zero, therefore $\|U(t) - U(\bar{t})\|$ tends to zero as well. Analogically we can prove the continuity at $t_{m-1}+$. Since the exact solution u is continuous and since we have continuity on the closed interval $[0, T]$, we can see that the error $U(t) - u(t)$ is uniformly continuous. □

Lemma 6.3. *Let u satisfy regularity assumptions (6.1). Let $s \in (0, \tau]$. Then*

$$(u(t_{m-1} + s) - u^{m-1} - su'(t_{m-1} + s/2), w) \leq Cs\tau^2\|u\|_{W^{3,\infty}(L^2)}\|w\|$$

Proof. The proof is analogical to the proof of Lemma 5.4. We can formally rewrite

$$\begin{aligned} u(t_{m-1} + s) - u^{m-1} - su'(t_{m-1} + s/2) &= u(t_{m-1} + s) - u^{m-1} - su'(t_{m-1}) \\ &\quad - \frac{s^2}{2}u''(t_{m-1}) - su'(t_{m-1} + s/2) + su'(t_{m-1}) + \frac{s^2}{2}u''(t_{m-1}). \end{aligned}$$

Then it is easy to see that

$$\begin{aligned} \big(u(t_{m-1} + s) - u^{m-1} - su'(t_{m-1} + s/2), w\big) &= \int_{t_{m-1}}^{t_{m-1}+s}\int_{t_{m-1}}^{z_1}\int_{t_{m-1}}^{z_2}(u'''(z_3), w)\mathrm{d}z_3\mathrm{d}z_2\mathrm{d}z_1 \\ &\quad - s\int_{t_{m-1}}^{t_{m-1}+s/2}\int_{t_{m-1}}^{z_1}(u'''(z_2), w)\mathrm{d}z_2\mathrm{d}z_1 \\ &\leq \left(\frac{s^3}{6} + \frac{s^3}{8}\right)\|u\|_{W^{3,\infty}(L^2)}\|w\|. \end{aligned}$$

□

Lemma 6.4. *Let u satisfy regularity assumptions (6.1). Let $s \in (0,\tau]$. Then*

$$A_h\left(u(t_{m-1}+s/2) - \frac{u(t_{m-1}+s)+u^{m-1}}{2}, w\right) \le C\tau^2 \|u\|_{W^{2,\infty}(H^2)} |||w||| \tag{6.3}$$

$$b_h\big(u(t_{m-1}+s/2), w\big) - b_h\left(\frac{u(t_{m-1}+s)+u^{m-1}}{2}, w\right) \le C\tau^2 \|u\|_{W^{2,\infty}(H^2)} \|w\| \tag{6.4}$$

Proof. Let us denote $u_1 = u(t_{m-1}+s/2)$ and $u_2 = \frac{u(t_{m-1}+s)+u^{m-1}}{2}$. Moreover, it is possible to see that

$$\begin{aligned} u_1 - u_2 &= u(t_{m-1}+s/2) - \frac{u(t_{m-1}+s)+u^{m-1}}{2} \\ &= \frac{1}{2}u(t_{m-1}+s/2) - \frac{1}{2}u(t_{m-1}+s) + \frac{s}{2}u'(t_{m-1}+s/2) + \frac{1}{2}u(t_{m-1}+s/2) - \frac{1}{2}u^{m-1} - \frac{s}{2}u'(t_{m-1}+s/2) \\ &= -\frac{1}{2}\int_{t_{m-1}+s/2}^{t_{m-1}+s}\int_{t_{m-1}+s/2}^{z_1} u''(z_2)\mathrm{d}z_2\mathrm{d}z_1 - \frac{1}{2}\int_{t_{m-1}}^{t_{m-1}+s/2}\int_{z_1}^{t_{m-1}+s/2} u''(z_2)\mathrm{d}z_2\mathrm{d}z_1. \end{aligned} \tag{6.5}$$

Following the proof of ([8], Lem. 9) it can be shown

$$A_h(u_1-u_2,w) \le C(|||u_1-u_2||| + |u_1-u_2|_{H^2})|||w|||. \tag{6.6}$$

Since $u_1, u_2 \in H_0^1(\Omega)$, we can simplify (6.6) to

$$A_h(u_1-u_2,w) \le C\|u_1-u_2\|_{H^2}|||w|||.$$

Using (6.5) we get

$$A_h(u_1-u_2,w) \le C\frac{s^2}{2}\|u\|_{W^{2,\infty}(H^2)}|||w|||,$$

which implies (6.3). Since u_1 and u_2 are smooth enough, it implies

$$b_h(u_1,w) - b_h(u_2,w) = \int_\Omega \nabla\cdot(f(u_1)-f(u_2))w\mathrm{d}x \le \|\nabla\cdot(f(u_1)-f(u_2))\|\,\|w\|.$$

To prove (6.4) it is sufficient to estimate $\|\nabla\cdot(f(u_1)-f(u_2))\|$.

$$\begin{aligned} \|\nabla\cdot(f(u_1)-f(u_2))\| &\le \sum_{i=1}^d \left\| f_i'(u_1)\frac{\partial u_1}{\partial x_i} - f_i'(u_2)\frac{\partial u_2}{\partial x_i}\right\| \\ &\le \sum_{i=1}^d \left(\left\| f_i'(u_1)\left(\frac{\partial u_1}{\partial x_i} - \frac{\partial u_2}{\partial x_i}\right)\right\| + \left\|(f_i'(u_1)-f_i'(u_2))\frac{\partial u_2}{\partial x_i}\right\|\right) \\ &\le d\max_i \|f_i'(u_1)\|_{L^\infty}\,|u_1-u_2|_{H^1} + d\max_i\|f_i'(u_1)-f_i'(u_2)\|_{L^\infty}\,|u_2|_{H^1}. \end{aligned}$$

Then (6.4) is a consequence of (6.5) and $\|f_i'(u_1)-f_i'(u_2)\|_{L^\infty} \le C\|u_1-u_2\|_{L^\infty}$. □

Now, we shall derive the error estimate of the continued solution at arbitrary time $t \in [0,T]$ which immediately implies the error estimate for the original midpoint scheme (3.5).

Lemma 6.5. *Let $p > d/2$. Let $m > 0$ and $s \in (0,\tau]$. If $\|e(t)\| \le h^{1+d/2}$ for $t \in [0, t_{m-1}+s]$, then*

$$\sup_{t\in[0,t_{m-1}+s]} \|e(t)\|^2 \le C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^4),$$

where the constant C_T is independent of h, τ, ε.

Proof. We set $y = t_{m-1} + s$. Multiplying (3.1) for $t = t_{m-1} + s/2$ by s, subtracting from (6.2) and adding several terms we get

$$\begin{aligned}(\xi(y) - \xi^{m-1}, w) + \frac{s\varepsilon}{2} A_h\big(\xi(y) + \xi^{m-1}), w\big) \le{}& \left(s\frac{\partial u}{\partial t}(t_{m-1} + s/2) - u(y) + u^{m-1}, w \right) \\ &+ s\left(b_h\big(u(t_{m-1} + s/2), w\big) - b_h\Big(\frac{u(y) + u^{m-1}}{2}, w\Big)\right) + \big(\eta(y) - \eta^{m-1}, w\big) \\ &+ s\left(b_h\Big(\frac{u(y) + u^{m-1}}{2}, w\Big) - b_h\Big(\frac{U(y) + U^{m-1}}{2}, w\Big)\right) - \frac{s\varepsilon}{2} A_h\big(\eta(y) + \eta^{m-1}), w\big) \\ &+ s\left(A_h(u(t_{m-1} + s/2), w) - A_h\Big(\frac{u(y) + u^{m-1}}{2}, w\Big)\right).\end{aligned}$$

Setting $w = \xi(y) + \xi^{m-1}$ and using Lemmas 4.1–4.3 and Lemma 6.4 to estimate the right-hand side, we get

$$\|\xi(y)\|^2 - \|\xi^{m-1}\|^2 \le Cs\left(1 + \frac{\|e(y) + e^{m-1}\|_\infty^2}{h^2}\right)(\varepsilon h^{2p} + h^{2p+1} + \tau^4 + \|\xi(y)\|^2 + \|\xi^{m-1}\|^2).$$

Using the assumptions we can get rid of the unpleasant term $\|e(s) + e^{m-1}\|_\infty^2/h^2$. Finally, by taking $s := \tau$ and $m = 1, \dots$, we obtain a similar estimate for $\|\xi^m\|^2 - \|\xi^{m-1}\|^2$. By the discrete Gronwall lemma we can finish the proof. □

Theorem 6.6. *Let $p > 1 + d/2$. Let τ_0 be defined as in Lemma 6.2. Let $h \in (0, h_0)$ and $\tau_1 \in (0, \tau_0)$ be such that*

$$C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^4) \le \frac{1}{4} h^{2+d}, \tag{6.7}$$

where C_T is the constant from Lemma 6.5 independent of h, τ, ε. Then the error of the midpoint–DG scheme satisfies

$$\sup_{t\in[0,T]} \|e(t)\|^2 \le C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^4).$$

Proof. The proof is essentially identical to that of Theorem 5.8. We have the desired estimate for $t = 0$ and due to continuity and Lemma 6.5, we can extend its validity to time T by induction. □

Remark 6.7. Similarly as in Remark 6.7, the condition (6.7) can be essentially split into two parts: $p > 1 + d/2$ and $\tau = O(h^{1/2+d/4})$. The latter condition is weaker than for the backward Euler method, where we needed $\tau = O(h^{1+d/2})$.

7. Quadrature variant of time-DG

In this section, we will prove error estimates for the quadrature variant of the QT-DG. As in the previous sections, we will construct a suitable continuation of U from Definition 3.5. While for the BDF2 and midpoint schemes, the discrete solution is defined only in the nodes of the partition t_m and the continuation "fills in the gaps" between these points, the DG solution U is already inherently defined on the whole interval $(0, T)$. It is therefore a question how to define a continuation w.r.t. time for such an object. In our approach, we construct the continuation U_y with respect to an auxiliary parameter y. Then U_y will be a piecewise polynomial function defined on $(0, y)$ which will depend continuously on y in the $L^\infty(L^2)$-norm. Again, we will use an induction argument to pass with y from 0 to T. For our analysis we will need the following regularity

$$u \in W^{1,\infty}(H^{p+1}) \cap L^\infty(W^{1,\infty}) \cap W^{q+1,\infty}(H^1). \tag{7.1}$$

7.1. Construction of the continuation

Throughout this section, let $s \in (0,\tau]$ and $m \in \{1,\dots,r\}$. We denote $y = t_{m-1} + s$, the continuation parameter and define $I_m(s) = (t_{m-1}, y)$. Let us generalize the quadrature Q_τ^m to Q_s^m:

$$\int_{I_m(s)} \Phi(t)\mathrm{d}t \approx Q_s^m[\Phi] = s\sum_{i=0}^{q} \omega_i \Phi(t_{m-1} + s\psi_i).$$

We define the space of piecewise polynomials up to degree p in space and up to degree q in time defined on I_m:

$$S_h^m = \Big\{ v \in L^2(I_m; S_h) : v = \sum_{j=0}^{q} v_j t^j,\ v_j \in S_h \Big\}.$$

Definition 7.1. Let $y \in I_m \cup \{t_m\}$. We say that the function $U_y \in L^2(0,t_m;S_h)$ is a *continued approximate solution* of problem (2.1) obtained by the QT-DG scheme if $U_y|_{I_l} = U|_{I_l}$ for $l = 0,\dots,m-1$, where U is the space-time DG solution from Definition 3.5 and $U_y|_{I_m} \in S_h^m$ satisfies

$$\int_{I_m(s)} (U_y', w) + \varepsilon A_h(U_y, w)\mathrm{d}t + Q_s^m[b_h(U_y,w)] + (\{U_y\}_{m-1}, w_+^{m-1}) = Q_s^m[\ell_h(w)] \quad \forall w \in S_h^m. \tag{7.2}$$

Remark 7.2. We note that by taking $s = \tau$, or equivalently $y = t_m$, we get $U_y|_{(0,t_m)} = U|_{(0,t_m)}$, *i.e.* we obtain the original space-time DG solution on $(0,t_m)$. Specifically, by taking $y = T$, we get $U_T = U$ on the whole interval $(0,T)$. We note also that relation (7.2) provides naturally the definition of U_y on $I_m(s)$. Since $U_y|_{I_m}$ is a polynomial with respect to time, U_y is uniquely defined on the remaining part of I_m and corresponds to the natural prolongation of $U_y|_{I_m(s)}$.

In order to prove existence, uniqueness and continuous dependence on y, we first need to establish monotonicity and Lipschitz continuity of the corresponding forms in (7.2). The same results can then be derived for (3.6) by taking $s := \tau$. Let us denote the left- and right-hand side of (7.2) by

$$\begin{aligned} B_s^m(v,w) &= \int_{I_m(s)} (v',w) + \varepsilon A_h(v,w)\mathrm{d}t + Q_s^m[b_h(v,w)] + (v_+^{m-1}, w_+^{m-1}),\\ L_s^m(w) &= Q_s^m[\ell_h(w)] + (U_-^{m-1}, w_+^{m-1}). \end{aligned}$$

Definition 7.3. We define the projection $P_s^m : C(\overline{I_m(s)}; L^2(\Omega)) \to S_h^m$ by

$$(P_s^m v)(t_{m-1} + s\psi_i) = v(t_{m-1} + s\psi_i), \quad \forall i = 0,\dots,q. \tag{7.3}$$

Furthermore, for any function $v \in S_h^m$ we denote

$$\tilde v(t) = P_s^m\Big(\frac{s}{t - t_{m-1}} v(t)\Big). \tag{7.4}$$

We point out that the relevant factors $\frac{s}{t_{m-1}+s\psi_i - t_{m-1}} = \frac{1}{\psi_i} \ge 1$. We have the following approximation properties of P_s^m:

Lemma 7.4. *Let $u \in W^{q+1,\infty}(H^1)$. Then*

$$\begin{aligned} \sup_{I_m(s)} \|P_s^m u - u\| &\le C s^{q+1} \sup_{I_m(s)} \|u^{(q+1)}\|,\\ \sup_{I_m(s)} |\!|\!| P_s^m u - u |\!|\!| &\le C s^{q+1} \sup_{I_m(s)} |\!|\!| u^{(q+1)} |\!|\!|, \end{aligned}$$

where the constant C does not depend on s.

Proof. The proof is an analogy to (*e.g.* [4], Thm. 3.1.5) for Bochner spaces. The result is also derived in the Appendix of [20]. □

We shall use following technical lemmas.

Lemma 7.5. *For any $v \in S_h^m$ the following terms are equivalent with the equivalence constants depending only on q:*

$$\sup_{I_m(s)} \|v\|^2, \qquad \sup_{I_m(s)} \|\tilde{v}\|^2, \qquad \frac{1}{s}\int_{I_m(s)} \|\tilde{v}\|^2 \mathrm{d}t.$$

Proof. The proof follows immediately from the fact that S_h^m has finite dimension. □

Lemma 7.6. *Let $v \in S_h^m$ and $\tilde{v}$ defined by (7.4). Then*

$$\int_{I_m(s)} (v', 2\tilde{v})\mathrm{d}t + (v_+^{m-1}, 2\tilde{v}_+^{m-1}) = \|v(y)\|^2 + \frac{1}{s}\int_{I_m(s)} \|\tilde{v}\|^2 \mathrm{d}t.$$

Proof. The proof can be made as a simple extension of ([1], Lem. 2.1), which describes the same result for scalar polynomials and on the unit time interval. □

Lemma 7.7. *Let $v \in S_h^m$ and $\tilde{v}$ be defined by (7.4), then*

$$0 \le \int_{I_m(s)} A_h(v, v)\mathrm{d}t \le \int_{I_m(s)} A_h(v, \tilde{v})\mathrm{d}t.$$

Proof.

$$0 \le \int_{I_m(s)} A_h(v, v)\mathrm{d}t = Q_s^m[A_h(v, v)] = s\sum_{i=0}^{q} \omega_i A_h(v(t_{m-1} + s\psi_i), v(t_{m-1} + s\psi_i))$$
$$\le s\sum_{i=0}^{q} \omega_i \frac{1}{\psi_i} A_h(v(t_{m-1} + s\psi_i), v(t_{m-1} + s\psi_i)) = Q_s^m[A_h(v, \tilde{v})] = \int_{I_m(s)} A_h(v, \tilde{v})\mathrm{d}t,$$

since $1/\psi_i \ge 1$. □

Now we are ready to prove fundamental properties of the forms B_s^m and L_s^m. We note that the mapping $v \to \tilde{v}$ is a bijection on S_h^m, therefore we can reformulate problem (7.2), *i.e.* $B_s^m(U_s, w) = L_s^m(w)$, for all $w \in S_h^m$ to the equivalent problem $B_s^m(U_s, \tilde{w}) = L_s^m(\tilde{w})$ for all $w \in S_h^m$. Hence for the purpose of proving existence and uniqueness of U_y, we can deal either with $B_s^m(.,.)$ or $B_s^m(.,\tilde{.})$ and similarly for L_s^m.

Lemma 7.8. *Let $s \le \tau \le C_1 h$, where C_1 is a suitable constant. Then the form $B_s^m(.,\tilde{.})$ is strongly monotone and Lipschitz continuous on S_h^m with respect to the $L^2(\Omega)$-norm, with the monotonicity and Lipschitz constants independent of s. Furthermore, L_s^m is bounded on this space, with norm uniformly bounded with respect to s but depending on $\|U_-^{m-1}\|$.*

Proof. To simplify the notation, all of the suprema in this proof are over the relevant interval $I_m(s)$.

(i) **Monotonicity of** B_s^m: Let $v, w \in S_h^m$, then

$$\begin{aligned} B_s^m(v, \tilde{v} - \tilde{w}) - B_s^m(w, \tilde{v} - \tilde{w}) &= \int_{I_m(s)} (v' - w', \tilde{v} - \tilde{w}) + \varepsilon A_h(v - w, \tilde{v} - \tilde{w}) \mathrm{d}t \\ &\quad + Q_s^m[b_h(v, \tilde{v} - \tilde{w}) - b_h(w, \tilde{v} - \tilde{w})] + (v_+^0 - w_+^0, \tilde{v}_+^0 - \tilde{w}_+^0) \\ &\geq \frac{1}{2}\|v(y) - w(y)\|^2 + \frac{1}{2s}\int_{I_m(s)} \|\tilde{v} - \tilde{w}\|^2 \mathrm{d}t + \varepsilon \int_{I_m(s)} A_h(v - w, v - w) \mathrm{d}t \\ &\quad - Cs \sup \|v - w\| \sup |\!|\!| \tilde{v} - \tilde{w} |\!|\!| \\ &\geq c \sup \|v - w\|^2 - \frac{C}{h} s \sup \|v - w\| \sup \|\tilde{v} - \tilde{w}\| \\ &\geq \left(c - \frac{Cs}{h}\right) \sup \|v - w\|^2, \end{aligned}$$

where the constant c comes from Lemma 7.5 and the generic constant C comes from Lemmas 4.3 and 7.5. If $s \leq \tau \leq C_1 h$ with a sufficiently small constant C_1, we obtain strong monotonicity with the monotonicity constant $M = c - \frac{C\tau}{h}$.

(ii) **Lipschitz continuity of** B_s^m: Let $v, \bar{v}, w \in S_h^m$. We estimate individual terms in B_s^m:

$$\begin{aligned} \int_{I_m(s)} (v' - \bar{v}', w) \mathrm{d}t + (v_+^{m-1} - \bar{v}_+^{m-1}, w_+^{m-1}) &\leq s \sup \|v' - \bar{v}'\| \sup \|w\| + \sup \|v - \bar{v}\| \sup \|w\| \\ &\leq C \sup \|v - \bar{v}\| \sup \|w\|, \\ \varepsilon \int_{I_m(s)} A(v - \bar{v}, w) \mathrm{d}t \leq C\varepsilon \int_{I_m(s)} |\!|\!| v - \bar{v} |\!|\!| \, |\!|\!| w |\!|\!| \mathrm{d}t &\leq Ch^{-2} s\varepsilon \sup \|v - \bar{v}\| \sup \|w\|, \\ Q_s^m[b(v, w) - b(\bar{v}, w)] \leq CQ_s^m[\|v - \bar{v}\| \, |\!|\!| w |\!|\!|] &\leq Csh^{-1} \sup \|v - \bar{v}\| \sup \|w\|. \end{aligned}$$

Hence, we have

$$\begin{aligned} B_s^m(v, w) - B_s^m(\bar{v}, w) &\leq C \sup \|v - \bar{v}\| \sup \|w\|, \\ B_s^m(v, \tilde{w}) - B_s^m(\bar{v}, \tilde{w}) &\leq C \sup \|v - \bar{v}\| \sup \|\tilde{w}\| \leq C \sup \|v - \bar{v}\| \sup \|w\|. \end{aligned}$$

Here the resulting constants C depend also on ε, h, s, however, for the sake of the existence and uniqueness proof, these may be considered as fixed quantities. Elsewhere, we can bound $s \leq \tau$ to obtain s-independence of the Lipschitz constant.

(iii) **Boundedness of** L_s^m:

$$\begin{aligned} L_s^m(v) &= Q_s^m[\ell(v)] + (U_-^{m-1}, v_+^{m-1}) \leq s \sup \|g\| \sup \|v\| + \|U_-^{m-1}\| \sup \|v\| \leq C \sup \|v\|, \\ L_s^m(\tilde{v}) &\leq C \sup \|\tilde{v}\| \leq C \sup \|v\|. \end{aligned}$$

The constant C in the resulting estimate depends also on $\|U_-^{m-1}\|$ and s, however by bounding $s \leq \tau$, we obtain s-independence of the boundedness constant. □

Existence and uniqueness of the continued solution U_y follows immediately from Lemma 7.8. We will also need uniform boundedness of $\|U_y\|$ and $\|U_y'\|$ with respect to $t \in [0, y]$. The resulting boundedness constants depend on ε and negative powers of h, however since the main goal is to prove continuous dependence of U_y on y, this is not a problem.

Lemma 7.9. *There exist constants $C_1, C_2 > 0$ independent of h, τ, t, ε, such that the following holds. Let $h \in (0, h_0)$ and $\tau \in [0, \tau_0)$, where $\tau_0 = \max\{C_1\varepsilon, C_2 h\}$. Then U_y, the continued solution from Definition 7.1 exists, is uniquely determined and $\|U_y(t)\|$, $\|U_y'(t)\|$ are uniformly bounded with respect to $t \in [0, y]$. Furthermore, for fixed $\|U_-^{m-1}\|$, the norms $\sup_{t \in I_m(s)} \|U_y(t)\|$, $\sup_{t \in I_m(s)} \|U_y'(t)\|$ are uniformly bounded with respect to $y \in I_m$.*

Proof. Remark 3.7 holds for U_y as well, therefore, we can prove unique existence and boundedness of U_y on each interval independently. From Lemma 7.8, we obtain existence and uniqueness of U_y.

(i) **Boundedness of** U_y: Due to Lemma 7.8,

$$M \sup_{I_m(s)} \|U_y\|^2 = M \sup_{I_m(s)} \|U_y - 0\|^2 \le B_s^m(U_y, \tilde{U}_y) - B_s^m(0, \tilde{U}_y) = B_s^m(U_y, \tilde{U}_y) = L_s^m(\tilde{U}_y) \le C \sup_{I_m(s)} \|U_y\|.$$

Due to Lemma 7.8, all the constants involved are independent of s, hence y.

(ii) **Boundedness of** U_y': setting $v(t) = (t - t_{m-1})U_y'(t) \in S_h^m$

$$\int_{I_m(s)} (t - t_{m-1})\|U_y'\|^2 + (t - t_{m-1})\varepsilon A_h(U_y, U_y')\mathrm{d}t + Q_s^m[(t - t_{m-1})b_h(U_y, U_y')] = Q_s^m[(t - t_{m-1})\ell_h(U_y')].$$

From this follows

$$\begin{aligned} cs \int_{I_m(s)} \|U_y'\|^2 \mathrm{d}t &\le \int_{I_m(s)} (t - t_{m-1})\|U_y'\|^2 \mathrm{d}t \\ &\le -\int_{I_m(s)} (t - t_{m-1})\varepsilon A_h(U_y, U_y')\mathrm{d}t - Q_s^m\left[(t - t_{m-1})b_h(U_y, U_y') - (t - t_{m-1})\ell_h(U_y')\right] \\ &\le s \int_{I_m(s)} C\varepsilon h^{-2}\|U_y\|\|U_y'\|\mathrm{d}t + sQ_s^m\left[(Ch^{-1}\|U_y\| + C)\|U_y'\|\right] \\ &= s \int_{I_m(s)} \|U_y'\|(C\varepsilon h^{-2}\|U_y\| + Ch^{-1}\|U_y\| + C)\mathrm{d}t \\ &\le \frac{cs}{2} \int_{I_m(s)} \|U_y'\|^2 \mathrm{d}t + C(\varepsilon, h)s^2, \end{aligned}$$

where we have used Hölder's and Young's inequality in the last step. Since

$$s \sup_{I_m(s)} \|U_y'\|^2 \le C \int_{t_{m-1}}^{y} \|U_y'\|^2 \mathrm{d}t,$$

we get the boundedness of $\|U_y'\|$. Moreover, after cancellation of the term s^2 from the resulting estimate, we obtain s-independence of the upper bound. □

Before we prove the main property of U_y, continuous dependence on y, we need one more technical lemma concerning the estimation of quadratures.

Lemma 7.10. *Let $s, \bar{s} \in (0, \tau]$ and $m \in 1, \dots, r$. Let $v, w \in S_h^m$. Then $|s - \bar{s}| \to 0$ implies*

$$\begin{aligned} Q_s^m[b_h(v, w)] - Q_{\bar{s}}^m[b_h(v, w)] \to 0, \\ Q_s^m[\ell_h(v)] - Q_{\bar{s}}^m[\ell_h(v)] \to 0. \end{aligned}$$

Proof. Let us assume $|s - \bar{s}| \to 0$. In order to simplify the notation of quadrature points, we shall set $s_i := t_{m-1} + s\psi_i$ and $\bar{s}_i := t_{m-1} + \bar{s}\psi_i$. Then

$$\begin{aligned} Q_s^m[b_h(v, w)] - Q_{\bar{s}}^m[b_h(v, w)] &= \sum_{i=0}^{q} (s\omega_i b_h(v, w)|_{s_i} - \bar{s}\omega_i b_h(v, w)|_{\bar{s}_i}) \\ &= s \sum_{i=0}^{q} \omega_i \left(b_h(v, w)|_{s_i} - b_h(v, w)|_{\bar{s}_i}\right) + \sum_{i=0}^{q} (s - \bar{s})\, \omega_i b_h(v, w)|_{\bar{s}_i} \end{aligned}$$

and

$$Q_s^m[\ell_h(v)] - Q_{\bar s}^m[\ell_h(v)] = \sum_{i=0}^{q} (s\omega_i \ell_h(v)|_{s_i} - \bar s \omega_i \ell_h(v)|_{\bar s_i}) = s\sum_{i=0}^{q} \omega_i \left(\ell_h(v)|_{s_i} - \ell_h(v)|_{\bar s_i}\right) + \sum_{i=0}^{q} (s-\bar s)\,\omega_i \ell_h(v)|_{\bar s_i}.$$

From continuity of $b_h(.,.)$ and v, w we get

$$b_h(v,w)|_{s_i} - b_h(v,w)|_{\bar s_i} \to 0$$

and since $\ell_h(v) = (g, v)$, where g is continuous with respect to time, we get

$$\ell_h(v)|_{s_i} - \ell_h(v)|_{\bar s_i} \to 0.$$

From boundedness of $b_h(v,w)|_{\bar s_i}$ and $\ell_h(v)|_{\bar s_i}$ we obtain $(s-\bar s)\omega_i b_h(v,w)|_{\bar s_i} \to 0$ and $(s-\bar s)\omega_i \ell_h(v,w)|_{\bar s_i} \to 0$. □

Lemma 7.11. *Let the assumptions of Lemma 7.9 hold. Then $U_{t_m} = U|_{(0,t_m)}$ for all $m = 0, \dots, r$ and U_y depends continuously on the parameter y in the following sense:*

$$\sup_{(0,\min(y,\bar y))} \|U_y - U_{\bar y}\| \to 0, \text{ as } |y-\bar y| \to 0,$$
$$\sup_{(t_{m-1},y)} \|U_y - U_-^{m-1}\| \to 0, \text{ as } y \to t_{m-1}+ . \tag{7.5}$$

Proof. Let $y = t_{m-1} + s, \bar y = t_{m-1} + \bar s$ for some m. Without loss of generality, let $0 < s < \bar s \le \tau$. Since $U_y = U_{\bar y} = U$ on $(0, t_{m-1})$, it is sufficient to prove the first relation only on (t_{m-1}, y). Let us denote $w = U_y - U_{\bar y}$. Due to monotonicity of $B_s^m(.,\tilde .)$ and Lemma 7.10, we have

$$M \sup_{(t_{m-1},y)} \|U_y - U_{\bar y}\|^2 \le B_s^m(U_y, \tilde w) - B_s^m(U_{\bar y}, \tilde w) = L_s^m(\tilde w) - L_{\bar s}^m(\tilde w) + B_{\bar s}^m(U_{\bar y}, \tilde w) - B_s^m(U_{\bar y}, \tilde w)$$
$$= \int_{t_{m-1}+s}^{t_{m-1}+\bar s} (U'_{\bar y}, \tilde w) + \varepsilon A_h(U_{\bar y}, \tilde w)\mathrm{d}t + Q_{\bar s}^m[b_h(U_{\bar y}, \tilde w)] - Q_s^m[b_h(U_{\bar y}, \tilde w)] - Q_{\bar s}^m[\ell_h(\tilde w)] + Q_s^m[\ell_h(\tilde w)].$$

Since the terms in the integral are bounded, the integral tends to zero as $|s - \bar s| \to 0$. According to Lemma 7.10 the quadrature terms tend to zero as well. From this it follows that $\sup_{(t_{m-1},y)} \|U_s - U_{\bar s}\| \to 0$ for $|s - \bar s| \to 0$.

It remains to prove the second formula in (7.5). Since U_y is continuous on (t_{m-1}, y), it is sufficient to prove ${U_y}_+^{m-1} \to U_-^{m-1}$ as $y \to t_{m-1}+$, *i.e.* $s \to 0+$: Testing (7.2) with $w \equiv {U_y}_+^{m-1} - U_-^{m-1}$, we get

$$\int_{t_{m-1}}^{y} (U'_y, {U_y}_+^{m-1} - U_-^{m-1}) + \varepsilon A_h(U_y, {U_y}_+^{m-1} - U_-^{m-1})\mathrm{d}t + Q_s^m[b(U_y, {U_y}_+^{m-1} - U_-^{m-1})] + \|{U_y}_+^{m-1} - U_-^{m-1}\|^2$$
$$= Q_s^m[\ell_h({U_y}_+^{m-1} - U_-^{m-1})].$$

Except for the last left-hand side term $\|{U_y}_+^{m-1} - U_-^{m-1}\|^2$, all remaining terms tend to zero as $s \to 0+$, therefore $\|{U_y}_+^{m-1} - U_-^{m-1}\|^2$ tends to zero as well. □

7.2. Error estimates

As the final step we shall derive the error estimate of the continued solution at arbitrary time $t \in [0, T]$ which immediately implies the error estimate for the classical method.

As usual, we shall split the error $e_y(t) = U_y(t) - u(t)$ into two parts $e_y(t) = \xi_y(t) + \eta_y(t)$, where we define:

$$\eta_y|_{I_i} = \begin{cases} \pi_\tau^i u|_{I_i} - u|_{I_i}, & i = 0, \dots, m-1, \\ \pi_s^m u|_{I_m} - u|_{I_m}, & i = m, \end{cases}$$
$$\xi_y|_{I_i} = \begin{cases} U_y|_{I_i} - \pi_\tau^i u|_{I_i}, & i = 0, \dots, m-1, \\ U_y|_{I_m} - \pi_s^m u|_{I_m}, & i = m, \end{cases}$$

where $\pi_s^i = P_s^i \Pi$. We have the following estimates for η_y and ξ_y:

Lemma 7.12. *Let u satisfy regularity assumptions (7.1). Then for all $v \in S_h^m$*

$$\sup_{I_m(s)} \|\eta_y\| \le C(h^{p+1} + s^{q+1}), \tag{7.6}$$

$$Q_s^m[(\eta_y', v)] + (\{\eta_y\}_{m-1}, v_+^{m-1}) \le sC(h^{p+1} + s^{q+1}) \sup_{I_m(s)} \|v\|. \tag{7.7}$$

Proof. The estimate (7.6) follows directly from Lemmas 4.1 and 7.4. The estimate (7.7) is proved in ([21], Lem. 4). □

Lemma 7.13. *Let u satisfy regularity assumptions (7.1). Then*

$$Q_s^m[b_h(u,\xi_y) - b_h(U_y,\xi_y)] \le Cs\left(1 + \frac{\sup_{I_m(s)} \|U_y - u\|^2}{h^2}\right)(h^{2p+1} + \sup_{I_m(s)} \|\xi_y\|^2),$$

$$Q_s^m[b_h(u,\tilde\xi_y) - b_h(U_y,\tilde\xi_y)] \le Cs\left(1 + \frac{\sup_{I_m(s)} \|U_y - u\|^2}{h^2}\right)(h^{2p+1} + \sup_{I_m(s)} \|\xi_y\|^2).$$

Proof. The proof is analogical for both of these inequalities, so we will prove only the second (more difficult) one.

$$\begin{aligned} Q_s^m[b(u,\tilde\xi_y) - b(U_y,\tilde\xi_y)] &= s\sum_{i=0}^{q} \omega_i \left(b_h(u,\tilde\xi_y) - b_h(U_y,\tilde\xi_y)\right)|_{t=t_{m-1}+s\psi_i} \\ &= s\sum_{i=0}^{q} \omega_i \frac{1}{\psi_i}\big(b_h(u, U_y - \Pi_s^m u) - b_h(U_y, U_y - \Pi_s^m u)\big)|_{t=t_{m-1}+s\psi_i} \\ &\le s\frac{1}{\psi_0} \sup_{I_m(s)} \big(b_h(u, U_u - \Pi_s^m u) - b_h(U_y, U_u - \Pi_s^m u)\big). \end{aligned}$$

Now it is sufficient to apply Lemma 4.3. □

Now, we shall prove the analogy to Lemmas 5.7 and 6.5.

Lemma 7.14. *Let $p > d/2$. Let $s \in (0,\tau]$ and $y = t_{m-1} + s$. If $\|e_y(t)\| \le h^{1+d/2}$ for $t \in [0,y]$, then*

$$\sup_{t\in[0,y]} \|e_y(t)\|^2 \le C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^{2q+2}),$$

where the constant C_T is independent of h, τ, ε.

Proof. Again, it is sufficient to estimate the error only on the last time interval $I_m(s)$, the previous ones are treated similarly. The error equation reads

$$\begin{aligned} \int_{I_m(s)} (\xi_y', v) + \varepsilon A_h(\xi_y, v)\mathrm{d}t + (\{\xi_y\}_{m-1}, v_+^{m-1}) &= Q_s^m[\varepsilon A_h(\eta_y, v)] - Q_s^m[(\eta_y', v)] \\ &\quad - (\{\eta_y\}_{m-1}, v_+^{m-1}) + Q_s^m[b_h(u,v) - b_h(U_y,v)]. \end{aligned}$$

By setting $v = 2\xi_y$ we get

$$\begin{aligned} &\|\xi_y(y)\|^2 - \|\xi_{y-}^{m-1}\|^2 + \|\{\xi_y\}_{m-1}\|^2 + 2\varepsilon \int_{I_m(s)} |\!|\!|\xi_y|\!|\!|^2 \mathrm{d}t \\ &\quad \le Cs\varepsilon(h^{2p} + s^{2q+2}) + \varepsilon \int_{I_m(s)} |\!|\!|\xi_y|\!|\!|^2 \mathrm{d}t + sC(h^{p+1} + s^{q+1}) \sup_{I_m(s)} \|\xi_y\| \\ &\qquad + Cs\Big(1 + \frac{\sup_{I_m(s)} \|U_y - u\|_\infty^2}{h^2}\Big)(h^{2p+1} + \sup_{I_m(s)} \|\xi_y\|^2). \end{aligned}$$

Therefore

$$\|\xi_y(y)\|^2 - \|\xi_{y-}^{m-1}\|^2 + \varepsilon \int_{I_m(s)} \|\!|\xi_y|\!\|^2 \mathrm{d}t \le Cs\left(1 + \frac{\sup_{I_m(s)} \|U_y - u\|_\infty^2}{h^2}\right)(h^{2p+1} + \varepsilon h^{2p} + s^{2q+2} + \sup_{I_m(s)} \|\xi_y\|^2). \tag{7.8}$$

With the aid of Lemmas 7.5–7.7 we get

$$\begin{aligned} c \sup_{I_m(s)} \|\xi_y\|^2 &\le \frac{1}{s} \int_{I_m(s)} \|\xi_y\|^2 \le \int_{I_m(s)} (\xi_y', 2\tilde{\xi}_y)\mathrm{d}t + (\xi_{y+}^{m-1}, 2\tilde{\xi}_{y+}^{m-1}) \\ &\le \int_{I_m(s)} (\xi_y', 2\tilde{\xi}_y) + 2\varepsilon A_h(\xi_y, \tilde{\xi}_y)\mathrm{d}t + (\xi_{y+}^{m-1}, 2\tilde{\xi}_{y+}^{m-1}). \end{aligned} \tag{7.9}$$

By setting $v = 2\tilde{\xi}_y$ in the error equation we get

$$\begin{aligned} &\int_{I_m(s)} (\xi_y', 2\tilde{\xi}_y) + 2\varepsilon A_h(\xi_y, \tilde{\xi}_y)\mathrm{d}t + (\xi_{y+}^{m-1}, 2\tilde{\xi}_{y+}^{m-1}) \\ &= Q_s^m[\varepsilon A_h(\eta_y, 2\tilde{\xi}_y)] - Q_s^m[(\eta_y', 2\tilde{\xi}_y)] - (\{\eta_y\}_{m-1}, 2\tilde{\xi}_{y+}^{m-1}) + (\xi_{y-}^{m-1}, 2\tilde{\xi}_{y+}^{m-1}) + Q_s^m[b(u, 2\tilde{\xi}_y) - b(U_y, 2\tilde{\xi}_y)] \\ &\le Cs\varepsilon(h^{2p} + s^{2q+2}) + (\xi_{y-}^{m-1}, 2\tilde{\xi}_{y+}^{m-1}) + \varepsilon \int_{I_m(s)} \|\!|\xi_y|\!\|^2 \mathrm{d}t + sC(h^{p+1} + s^{q+1}) \sup_{I_m(s)} \|\xi_y\| \\ &\quad + Cs\left(1 + \frac{\sup_{I_m(s)} \|U_y - u\|_\infty^2}{h^2}\right)(h^{2p+1} + \sup_{I_m(s)} \|\xi_y\|^2) \\ &\le Cs\left(1 + \frac{\sup_{I_m(s)} \|U_y - u\|_\infty^2}{h^2}\right)(h^{2p+1} + \varepsilon h^{2p} + s^{2q+2} + \sup_{I_m(s)} \|\xi_y\|^2) + \frac{2C}{c}\|\xi_{y-}^{m-1}\|^2 + \frac{c}{4} \sup_{I_m(s)} \|\xi_y\|^2, \end{aligned} \tag{7.10}$$

where $\varepsilon \int_{I_m(s)} \|\!|\xi_y|\!\|^2 \mathrm{d}t$ is estimated with the aid of (7.8).

Under the assumption $\|e(t)\| \le h^{1+d/2}$ inequality (7.8) can be simplified to

$$\|\xi_y(y)\|^2 - \|\xi_{y-}^{m-1}\|^2 \le Cs(h^{2p+1} + \varepsilon h^{2p} + s^{2q+2} + \sup_{I_m(s)} \|\xi_y\|^2) \tag{7.11}$$

and inequalities (7.9) and (7.10) give

$$\sup_{I_m(s)} \|\xi_y\|^2 \le \frac{C}{c} s(h^{2p+1} + \varepsilon h^{2p} + s^{2q+2} + \sup_{I_m(s)} \|\xi_y\|^2) + \frac{2C}{c^2}\|\xi_{y-}^{m-1}\|^2 + \frac{1}{4} \sup_{I_m(s)} \|\xi_y\|^2.$$

If $s \le \tau \le c/4C$ than the last inequality can be simplified to

$$\sup_{I_m(s)} \|\xi_y\|^2 \le h^{2p+1} + \varepsilon h^{2p} + s^{2q+2} + \frac{4C}{c^2}\|\xi_{y-}^{m-1}\|^2$$

Substituting this estimate into (7.11), we get

$$\|\xi(y)\|^2 - \|\xi_{y-}^{m-1}\|^2 \le Cs(h^{2p+1} + \varepsilon h^{2p} + s^{2q+2} + \|\xi_{y-}^{m-1}\|^2).$$

Similar estimates can be obtained on all previous time intervals. By application of the discrete Gronwall lemma we finish the proof. □

Theorem 7.15. *Let $p > 1 + d/2$. Let τ_0 be defined as in Lemma 7.9. Let $h \in (0, h_0)$ and $\tau_1 \in (0, \tau_0)$ be such that*

$$C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^{2q+2}) \le \frac{1}{16} h^{2+d},$$

where C_T is the constant from Lemma 7.14 independent of h, τ, ε. Then the error of the QT-DG scheme satisfies

$$\sup_{t\in[0,T]} \|e(t)\|^2 \le C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^{2q+2}).$$

Proof. Since the continuation $U_y(t)$ now depends on two variables, y and t, we proceed more carefully. We define the propositional function φ by

$$\varphi(y) \equiv \Big\{ \max_{t\in[0,y]} \|e_y(t)\|^2 \le C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^{2q+2}) \Big\}.$$

Due to the approximation of the initial condition, $\varphi(0)$ holds trivially. We want to prove $\varphi(T)$. We will proceed by continuous induction, *cf.* [17]. For this we need to prove that

$$\begin{aligned} &(A) \quad \forall y \in [0,T)\ \exists \delta(y) > 0 :\ \varphi(y) \text{ implies } \varphi(y+\delta),\ \forall \delta \in [0, \delta(y)] : y + \delta \in [0,T] \\ &(B) \quad \forall y_1, y_2 \in [0,T], y_1 < y_2 : \text{If } \varphi \text{ holds on } (y_1, y_2) \text{ then } \varphi(y_2) \text{ holds.} \end{aligned} \tag{7.12}$$

First we note, that due to the construction of U, U_y, it is sufficient to assume $y, y+\delta \in [t_{m-1}, t_m]$ and then proceed by induction with respect to $m = 1, \dots, r$. Our main tools will be the continuity of U_y with respect to y, *cf.* Lemma 7.11, the uniform boundedness of $\|U_y'(t)\|$ with respect to t and y, *cf.* Lemma 7.9 and uniform continuity of u from $[t_{m-1}, t_m]$ to $L^2(\Omega)$. Specifically, if $y \in [t_{m-1}, t_m)$ there exists $\delta(y) > 0$ such that

$$\begin{aligned} \delta \in [0, \delta(y)], t \in [y, y+\delta] &\implies \|u(y) - u(t)\| \le \frac{1}{4} h^{1+d/2}, \\ \delta \in [0, \delta(y)] &\implies \sup_{(t_{m-1}, y)} \|U_{y+\delta} - U_y\| \le \frac{1}{4} h^{1+d/2}. \end{aligned}$$

Without loss of generality, $\delta(y)$ can be taken small enough so that $C\delta(y) \le \frac{1}{4} h^{1+d/2}$, where C is the uniform bound for $\|U_y'(t)\|$ from Lemma 7.9.

Induction step (A)*:* Let us assume that $\varphi(y)$ holds. We want to prove that $\varphi(y+\delta)$ holds, where $\delta \in [0, \delta(y)]$. In other words, we want to estimate

$$\max_{t\in[0,y+\delta]} \|e_{y+\delta}(t)\| = \max\{ \max_{t\in[0,y]} \|e_{y+\delta}(t)\|, \max_{t\in[y,y+\delta]} \|e_{y+\delta}(t)\| \}. \tag{7.13}$$

We estimate the first right-hand side term in (7.13) by

$$\begin{aligned} \max_{t\in[0,y]} \|e_{y+\delta}(t)\| &= \max_{t\in[0,y]} \|U_{y+\delta}(t) - u(t)\| \le \max_{t\in[0,y]} \|U_{y+\delta}(t) - U_y(t)\| + \max_{t\in[0,y]} \|U_y(t) - u(t)\| \\ &= \max_{t\in[t_{m-1},y]} \|U_{y+\delta}(t) - U_y(t)\| + \max_{t\in[0,y]} \|e_y(t)\| \le \frac{1}{4} h^{1+d/2} + C_T \sqrt{h^{2p+1} + \varepsilon h^{2p} + \tau^{2q+2}} \le \frac{1}{2} h^{1+d/2} \end{aligned} \tag{7.14}$$

by Lemma 7.11 and the induction assumption. As for the second right-hand side term in (7.13), we have

$$\begin{aligned} \max_{t\in[y,y+\delta]} \|e_{y+\delta}(t)\| &= \max_{t\in[y,y+\delta]} \|U_{y+\delta}(t) - u(t)\| \\ &\le \max_{t\in[y,y+\delta]} \|U_{y+\delta}(t) - U_{y+\delta}(y)\| + \|U_{y+\delta}(y) - U_y(y)\| + \|U_y(y) - u(y)\| + \max_{t\in[y,y+\delta]} \|u(y) - u(t)\| \\ &\le \delta \max_{t\in[t_{m-1},y]} \|U_{y+\delta}'(t)\| + \max_{t\in[0,y]} \|U_{y+\delta}(t) - U_y(t)\| + \max_{t\in[0,y]} \|e_y(t)\| + \frac{1}{4} h^{1+d/2} \\ &\le C\delta + \frac{1}{4} h^{1+d/2} + C_T \sqrt{h^{2p+1} + \varepsilon h^{2p} + \tau^{2q+2}} + \frac{1}{4} h^{1+d/2} \le h^{1+d/2}, \end{aligned} \tag{7.15}$$

due to Lemmas 7.9, 7.11 and the induction assumption. Collecting (7.13)–(7.15) gives us

$$\max_{t\in[0,y+\delta]} \|e_{y+\delta}(t)\| \le h^{1+d/2}. \tag{7.16}$$

Lemma 7.14 then gives us $\varphi(y+\delta)$.

Induction step (B)*:* We prove (B) in (7.12) by contradiction. Fix $y_1, y_2 \in [0,T]$. Assume that for all $y \in (y_1, y_2)$ the statement $\varphi(y)$ holds, but $\varphi(y_2)$ is false. In other words assume that

$$\max_{t\in[0,y]} \|e_y(t)\|^2 \le C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^{2q+2}) \qquad \text{and} \qquad \max_{t\in[0,y_2]} \|e_{y_2}(t)\|^2 > C_T^2(h^{2p+1} + \varepsilon h^{2p} + \tau^{2q+2}). \tag{7.17}$$

Therefore, after taking the square root,

$$\max_{t\in[0,y_2]} \|e_{y_2}(t)\| - \max_{t\in[0,y]} \|e_y(t)\| \ge c_0 > 0, \qquad \text{for all } y \in (y_1, y_2), \tag{7.18}$$

where $c_0 > 0$ is an appropriate constant independent of $y \in (y_1, y_2)$.

We can estimate by the triangle inequality

$$\max_{t\in[y,y_2]} \|e_{y_2}(t)\| \le \|e_{y_2}(y)\| + \max_{t\in[y,y_2]} \|e_{y_2}(t) - e_{y_2}(y)\| \le \max_{t\in[0,y]} \|e_{y_2}(t)\| + C|y_2 - y|,$$

since u is uniformly continuous and $U_y'(t)$ is uniformly bounded with respect to y, t. Therefore,

$$\max_{t\in[0,y_2]} \|e_{y_2}(t)\| \le \max\{\max_{t\in[0,y]} \|e_{y_2}(t)\|, \max_{t\in[y,y_2]} \|e_{y_2}(t)\|\} \le \max_{t\in[0,y]} \|e_{y_2}(t)\| + C|y_2 - y|.$$

Hence, the left-hand side of (7.18) can be estimated as

$$\begin{aligned}
\max_{t\in[0,y_2]} \|e_{y_2}(t)\| - \max_{t\in[0,y]} \|e_y(t)\| &\le \max_{t\in[0,y]} \|e_{y_2}(t)\| + C|y_2 - y| - \max_{t\in[0,y]} \|e_y(t)\| \\
&\le \max_{t\in[0,y]} \|U_{y_2}(t) - U_y(t)\| + \max_{t\in[0,y]} \|U_y(t) - u(t)\| + C|y_2 - y| - \max_{t\in[0,y]} \|e_y(t)\| \\
&= \max_{t\in[0,y]} \|U_{y_2}(t) - U_y(t)\| + C|y_2 - y| \longrightarrow 0, \qquad \text{as } y \to y_2,
\end{aligned} \tag{7.19}$$

which is a contradiction with (7.18), *i.e.* (7.17). Thus (B) is proved, which completes the proof. □

8. Conclusions

We have proved *a priori* error estimates for the discontinuous Galerkin method applied to a nonlinear time-dependent singularly perturbed, convection-diffusion problem. The BDF-2, midpoint and quadrature version of the space-time DG scheme were analyzed. The main contribution of the paper is that $L^\infty(L^2)$-estimates are derived that are uniform with respect to the diffusion parameter $\varepsilon \to 0$ and valid even in the purely convective case $\varepsilon = 0$. The paper extends the work [16], where similar estimates were derived for the space-semidiscretization and implicit Euler scheme as well as the paper [17], where similar estimates are obtained for the conforming finite element method. The basis of the technique is the idea of [24], where the analysis is carried out for an explicit Runge–Kutta scheme in time.

Similarly as in [16], the presented error analysis is based on construction of suitable continuations of the discrete solution with respect to time and performing, *via* induction. The resulting estimates are of the order $O(h^{p+1/2} + \varepsilon h^p + \tau^4)$ for the BDF-2 and midpoint schemes and $O(h^{p+1/2} + \varepsilon h^p + \tau^{q+1})$ for q-order quadrature time-DG. The estimates are derived under the CFL-like $\tau = O(h)$ condition guaranteeing the unique existence and continuity of the continuation. Furthermore, the estimates are derived under the order condition $p > 1 + d/2$, or $p > (1+d)/2$ for $\varepsilon = 0$, where d is the spatial dimension of the problem.

Future work includes removing of the CFL and order conditions and extension to more difficult equations, *e.g.* nonlinear diffusion as in [15], derivation of optimal order $L^\infty(L^2)$-error estimates and analysis of other temporal discretizations, especially the space-time DG scheme without quadratures.

References

[1] G. Akrivis and C. Makridakis, Galerkin time-stepping methods for nonlinear parabolic equations. *ESAIM: M2AN* **38** (2004) 261–289.

[2] T. Barth and M. Ohlberger, Finite Volume Methods: Foundation and Analysis. Vol. 1 of *Encyclopedia of Computational Mechanics*. John Wiley & Sons, Chichester, New York, Brisbane (2004) 439–474.

[3] F. Bassi and S. Rebay, High-Order Accurate Discontinuous Finite Element Solution of the 2D Euler Equations. *J. Comput. Phys.* **138** (1997) 251–285.

[4] P.G. Ciarlet, *The finite element methods for elliptic problems*. North-Holland, Amsterdam (1978).

[5] B. Cockburn and C.-W. Shu, The Local Discontinuous Galerkin Method for Time-Dependent Convection-Diffusion Systems. *SIAM J. Numer. Anal.* **35** (1998) 2440–2463.

[6] V. Dolejší and M. Feistauer, *Discontinuous Galerkin Method: Analysis and Applications to Compressible Flow*. Springer (2015).

[7] V. Dolejší and M. Vlasák, Analysis of a BDF-DG scheme for nonlinear convection-diffusion problems. *Numer. Math.* **110** (2008) 405–447.

[8] V. Dolejší, M. Feistauer and V. Sobotíková, Analysis of the discontinuous Galerkin method for nonlinear convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.* **194** (2005) 2709–2733.

[9] V. Dolejší, M. Feistauer and J. Hozman, Analysis of semi-implicit DGFEM for nonlinear convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.* **196** (2007) 2813–2827.

[10] V. Dolejší, M. Feistauer, V. Kučera and V. Sobotíková, An optimal $L^\infty(L^2)$-error estimate for the discontinuous Galerkin approximation of a nonlinear non-stationary convection-diffusion problem. *IMA J. Numer. Anal.* **28** (2008) 496–521.

[11] M. Feistauer and V. Kučera, On a robust discontinuous Galerkin technique for the solution of compressible flow. *J. Comput. Phys.* **224** (2007) 208–221.

[12] M. Feistauer, J. Hájek and K. Švadlenka, Space-time discontinuous Galerkin method for solving nonstationary convection-diffusion-reaction problems. *Appl. Math.* **52** (2007) 197–233.

[13] S. Gianni, D. Schötzau and L. Zhu, An a posteriori error estimate for hp-adaptive DG methods for convection-diffusion problems on anisotropically refined meshes. *Comp. Math. Appl.* **67** (2014) 869–887.

[14] E. Hairer, S.P. Norsett and G. Wanner, *Solving ordinary differential equations I, Nonstiff problems*. Springer Verlag (2000).

[15] V. Kučera, Optimal $L^\infty(L^2)$-error Estimates for the DG Method Applied to Nonlinear Convection-Diffusion Problems with Nonlinear Diffusion. *Numer. Func. Anal. Opt.* **31** (2010) 285–312.

[16] V. Kučera, On diffusion-uniform error estimates for the DG method applied to singularly perturbed problems. *IMA J. Numer. Anal.* **32** (2014) 820–861.

[17] V. Kučera, Finite element error estimates for nonlinear convective problems. *J. Numer. Math.* **24** (2016) 143–165.

[18] S. Osher, Riemann solvers, the entropy condition, and difference approximations. *SIAM. J. Numer. Anal.* **21** (1984) 217–235.

[19] W.H. Reed and T.R. Hill, *Triangular mesh methods for the neutron transport equation*. Technical Report LA–UR–73–479, Los Alamos Scientific Laboratory (1973).

[20] M. Vlasák, Optimal spatial error estimates for DG time discretizations. *J. Numer. Math.* **21** (2013) 201–230.

[21] M. Vlasák and H.G. Roos, An optimal uniform *a priori* error estimate for an unsteady singularly perturbed problem. *Int. J. Numer. Anal. Model.* **11** (2014) 24–33.

[22] M. Vlasák, V. Dolejší and J. Hájek, *A priori* error estimates of an extrapolated space-time discontinuous Galerkin method for nonlinear convection-diffusion problems. *Numer. Methods Partial Differ. Eqs.* **27** (2011) 1456–1482.

[23] E. Zeidler, *Nonlinear Functional Analysis and Its Applications II/B: Nonlinear Monotone Operators*. Springer, Heidelberg (1986).

[24] Q. Zhang and C.W. Shu, Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws. *SIAM J. Numer. Anal.* **44** (2006) 1703–1720.

Chapter 5

Nonlinear unsteady convection-diffusion problems in time-dependent domains

ESAIM: M2AN 52 (2018) 2327–2356
https://doi.org/10.1051/m2an/2018062

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

STABILITY OF THE ALE SPACE-TIME DISCONTINUOUS GALERKIN METHOD FOR NONLINEAR CONVECTION-DIFFUSION PROBLEMS IN TIME-DEPENDENT DOMAINS†

Monika Balázsová[1], Miloslav Feistauer[1,*] and Miloslav Vlasák[1]

Abstract. The paper is concerned with the analysis of the space-time discontinuous Galerkin method (STDGM) applied to the numerical solution of nonstationary nonlinear convection-diffusion initial-boundary value problem in a time-dependent domain. The problem is reformulated using the arbitrary Lagrangian–Eulerian (ALE) method, which replaces the classical partial time derivative by the so-called ALE derivative and an additional convective term. The problem is discretized with the use of the ALE-space time discontinuous Galerkin method (ALE-STDGM). In the formulation of the numerical scheme we use the nonsymmetric, symmetric and incomplete versions of the space discretization of diffusion terms and interior and boundary penalty. The nonlinear convection terms are discretized with the aid of a numerical flux. The main attention is paid to the proof of the unconditional stability of the method. An important step is the generalization of a discrete characteristic function associated with the approximate solution and the derivation of its properties.

Mathematics Subject Classification. 65M60, 65M99.

Received December 22, 2017. Accepted October 21, 2018.

1. Introduction

Most of the results on the solvability and numerical analysis of nonstationary partial differential equations (PDEs) are obtained under the assumption that a space domain Ω is independent of time t. However, problems in time-dependent domains Ω_t are important in a number of areas of science and technology. We can mention, for example, problems with moving boundaries, when the motion of the boundary $\partial\Omega_t$ is prescribed, or free boundary problems, when the motion of the boundary $\partial\Omega_t$ should be determined together with the solution of the PDEs in consideration. This is particularly the case of fluid-structure interaction (FSI), when the flow is solved in a domain deformed due to the coupling with an elastic structure.

There are various approaches to the solution of problems in time-dependent domains as, for example, fictitious domain method, see [43], or immersed boundary method, see [10]. A very popular technique is the arbitrary Lagrangian–Eulerian (ALE) method based on a suitable one-to-one ALE mapping of the reference domain Ω_{ref} onto the current configuration Ω_t. It is usually applied in connection with conforming finite element space

Keywords and phrases. nonlinear convection-diffusion equation, time-dependent domain, ALE method, space-time discontinuous Galerkin method, discrete characteristic function, unconditional stability in space and time.

[1] Charles University, Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Praha 8, Czech Republic.
*Corresponding author: feist@karlin.mff.cuni.cz
†Dedicated to Professor Chi-Wang Shu on the occasion of his 60th birthday.

discretization and combined with the time discretization by the use of a backward difference formula (BDF). From a wide literature we mention, *e.g.*, the works [22, 39, 41, 42]. This method is analyzed theoretically for linear parabolic convection-diffusion initial-boundary value problems. Paper [35] investigates the stability of the ALE-conforming finite element method. In [4,36] error estimates for the ALE-conforming finite element method are derived.

In the numerical solution of compressible flow, it is suitable to apply the discontinuous Galerkin method (DGM) for the space discretization. It is based on piecewise polynomial approximations over finite element meshes, in general discontinuous on interfaces between neighbouring elements. This method was applied to the solution of compressible flow first in [8] and then in [9]. It enables us to get a good resolution of boundary and internal layers (including shock waves and contact discontinuities) and has been used for the solution of various types of flow problems, see [19,26,32]. Theory of the space discretization by the discontinuous Galerkin method is a subject of a number of works. We cite only some of them: [2, 3, 13, 18, 21, 34, 38, 40, 46, 47, 52]. It is also possible to refer to the monograph [20] containing a number of references.

In the cited works, the time discretization is carried out with the aid of the BDF of the first or second order. One possibility to construct a higher order method in time is the application of the DGM in time. This technique uses a piecewise polynomial approximation in time, in general discontinuous at discrete time instants that form a partition in a time interval. This method was used for time discretization combined with conforming finite elements for the space discretization of linear parabolic equations in [1, 17, 23–25, 48–50].

By the combination of the DGM in space and time we get the space-time discontinuous Galerkin method (STDGM). This method was theoretically analyzed in [7, 14, 20, 29, 33, 53]. In [28, 44], the BDF-DGM and STDGM is applied to linear and nonlinear dynamic elasticity problems. One of the advantages of the STDGM is the possibility to use different meshes on different time levels.

The mentioned methods have also been extended to the numerical solution of initial-boundary value problems in time-dependent domains using the ALE method. The ALE method combined with the space DGM and BDF in time (ALE-DGM-BDF) was applied with success to interaction of compressible flow with elastic structures in [15, 30, 37, 44]. In [16], the ALE-STDGM is applied to the simulation of flow induced airfoil vibrations and the results are compared with the ALE-DGM-BDF approach. It appears that the ALE-STDGM is more robust and accurate. Here we can cite the important work [51] dealing with the space-time DGM to the solution of inviscid compressible flows. The approach in this paper consideres the time variable equivalent to the space variables and uses meshes formed by space-time four-dimensional elements. It allows to use different meshes in different time slabs. This paper also discusses the relation of the presented technique with the ALE method. The method analyzed in the following parts of our paper consideres time and space variables separately in contrast to [51]. Moreover, we deal with a problem containing diffusion, which should be analogy to the compressible Navier-Stokes equations.

The ALE-time discontinuous Galerkin semidiscretization of a linear parabolic convection-diffusion problem is analyzed in [11,12]. Both papers assume that the transport velocity is divergence free and consider homogeneous Dirichlet boundary condition. In [11], the stability of the ALE-time DGM is proved and [12] is devoted to the error estimation. Papers [5, 6] are concerned with the stability analysis of the ALE-STDGM applied to a linear convection-diffusion initial-boundary value problem, and to the case with nonlinear convection and diffusion, respectively. In both cases nonhomogeneous Dirichlet boundary conditions and piecewise linear DG time discretization are used.

In the present paper we extend the results from [5]. We deal with the stability analysis of the ALE-STDGM with arbitrary polynomial degree in space as well as in time, applied to a scalar nonstationary nonlinear convection-diffusion problem equipped with initial condition and nonhomogeneous Dirichlet boundary condition. This problem can be considered as a simplified prototype of the compressible Navier-Stokes system. The ALE-STDGM analyzed here corresponds to the technique used in [16,28] for the numerical simulation of airfoil vibrations induced by compressible flow. (The construction of the ALE mapping is described very briefly. It is hidden in the computer program.)

We present here a new formulation of the problem and technique of theoretical analysis in contrast to [5]. In [5] we proved the unconditional stability of the ALE-STDGM with arbitrary polynomial degree in space, but only linear approximation in time. Moreover, in [5] the standard ALE method prescribed globally in the whole time interval was used (see also [11, 12, 22, 35, 36, 39, 41]). In the present paper we apply a different ALE technique that can use different meshes with different numbers of elements in different time levels. We assume that the ALE mapping is prescribed for each time slab separately.

In the analysis presented in this paper it was necessary to overcome a number of various difficult obstacles. An important tool in our theory is the concept of the discrete characteristic function introduced in [17] in the framework of the time DGM applied to a linear parabolic problem. In [7, 14] the discrete characteristic function was generalized in connection with the STDGM for nonlinear parabolic problems in fixed domains. An important new and original result contained in the present paper is the extension of the discrete characteristic function and the proof of its properties in the case of the ALE-STDGM in time-dependent domains. On the basis of a technical analysis we obtain an unconditional stability of this method represented by a bound of the approximate solution in terms of data without any limitation of the time step in dependence on the size of the triangulations.

In Section 2 we formulate the continuous problem. Section 3 is devoted to the ALE space-time discretization. We describe here triangulations, ALE mappings and introduce important function spaces and concepts. Then an approximate solution is defined. Section 4 deals with the stability analysis. First some auxiliary results are presented. Then we introduce important estimates and the generalized concept of the discrete characteristic function. An important part is devoted to the derivation of its properties. Finally, the last part presents the proof of unconditional stability of the ALE-STDGM.

2. Formulation of the continuous problem

In what follows, we shall use the standard notation $L^2(\omega)$ for the Lebesgue space, $W^{k,p}(\omega)$, $H^k(\omega) = W^{k,2}(\omega)$ for the Sobolev spaces over a bounded domain $\omega \subset \mathbb{R}^d$, $d = 2, 3$, and the Bochner spaces $L^\infty(0,T;X)$ with a Banach space X and

$$W^{1,\infty}(0,T;W^{1,\infty}(\Omega_t)) = \left\{f \in L^\infty(0,T;W^{1,\infty}(\Omega_t));\ \mathrm{d}f/\mathrm{d}t \in L^\infty(0,T;W^{1,\infty}(\Omega_t))\right\},$$

where $\mathrm{d}f/\mathrm{d}t$ denotes here the distributional derivative.

If X is a Banach (Hilbert) space, then its norm (scalar product) will be denoted by $\|\cdot\|_X$ $((\cdot,\cdot)_X)$. By $|\cdot|_X$ we denote a seminorm in X. For simplicity we use the notation $\|\cdot\|_{L^2(\omega)} = \|\cdot\|_\omega$, $(\cdot,\cdot)_{L^2(\omega)} = (\cdot,\cdot)_\omega$ and $\|\cdot\|_{L^2(\partial\omega)} = \|\cdot\|_{\partial\omega}$.

We shall be concerned with an initial-boundary value nonlinear convection-diffusion problem in a time-dependent bounded domain $\Omega_t \subset \mathbb{R}^d$, where $t \in [0,T]$, $T > 0$: Find a function $u = u(x,t)$ with $x \in \Omega_t$, $t \in (0,T)$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} - \operatorname{div}(\beta(u)\nabla u) = g \quad \text{in } \Omega_t,\ t \in (0,T), \tag{2.1}$$

$$u = u_D \quad \text{on } \partial\Omega_t,\ t \in (0,T), \tag{2.2}$$

$$u(x,0) = u^0(x), \quad x \in \Omega_0. \tag{2.3}$$

We assume that $f_s \in C^1(\mathbb{R})$, $f_s(0) = 0$,

$$|f_s'| \le L_f, \quad s = 1, \dots, d, \tag{2.4}$$

where the constant L_f does not depend on u. Moreover we assume that function β is bounded and Lipschitz-continuous:

$$\beta : \mathbb{R} \to [\beta_0, \beta_1], \quad 0 < \beta_0 < \beta_1 < \infty, \tag{2.5}$$

$$|\beta(u_1) - \beta(u_2)| \le L_\beta |u_1 - u_2| \quad \forall u_1, u_2 \in \mathbb{R}. \tag{2.6}$$

Problem (2.1)–(2.3) can be reformulated with the aid of the so-called arbitrary Lagrangian–Eulerian (ALE) method. A standard ALE formulation is based on an ALE mapping prescribed globally in the whole time interval $[0,T]$. It is based on a regular one-to-one ALE mapping of the reference configuration $\Omega_{\rm ref}$ onto the current configuration Ω_t:

$$\mathcal{A}_t : \overline{\Omega}_{\rm ref} \to \overline{\Omega}_t, \quad X \in \overline{\Omega}_{\rm ref} \to x = \mathcal{A}_t(X) \in \overline{\Omega}_t, \quad t \in [0,T]. \tag{2.7}$$

Usually it is assumed that $\Omega_{\rm ref} = \Omega_0$, as in *cf.*, *e.g.*, [5, 11, 12, 22, 35, 36, 39, 41]. However, in this case it is impossible to use different space partitions in different time slabs, which allows the STDGM. Therefore, we shall proceed as is described in the next section.

The transformation of the partial differential equation (2.1) into the ALE form is based on the following concepts. We introduce the domain velocity

$$\tilde{\boldsymbol{z}}(X,t) = \frac{\partial}{\partial t}\mathcal{A}_t(X), \ \boldsymbol{z}(x,t) = \tilde{\boldsymbol{z}}(\mathcal{A}_t^{-1}(x),t), \ t \in [0,T], \ X \in \Omega_{\rm ref}, \ x \in \Omega_t, \tag{2.8}$$

and define the ALE derivative $D_t f = Df/Dt$ of a function $f = f(x,t)$ for $x \in \Omega_t$ and $t \in [0,T]$ as

$$D_t f(x,t) = \frac{D}{Dt} f(x,t) = \frac{\partial \tilde{f}}{\partial t}(X,t), \tag{2.9}$$

where $\tilde{f}(X,t) = f(\mathcal{A}_t(X),t)$, $X \in \Omega_{\rm ref}$, and $x = \mathcal{A}_t(X) \in \Omega_t$. The use of the chain rule yields the relation

$$\frac{Df}{Dt} = \frac{\partial f}{\partial t} + \boldsymbol{z} \cdot \nabla f, \tag{2.10}$$

which allows us to reformulate problem (2.1)–(2.3) in the ALE form: Find $u = u(x,t)$ with $x \in \Omega_t$, $t \in (0,T)$ such that

$$\frac{Du}{Dt} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} - \boldsymbol{z} \cdot \nabla u - \operatorname{div}(\beta(u)\nabla u) = g \quad \text{in } \Omega_t, \ t \in (0,T), \tag{2.11}$$

$$u = u_D \quad \text{on } \partial\Omega_t, \ t \in (0,T), \tag{2.12}$$

$$u(x,0) = u^0(x), \quad x \in \Omega_0. \tag{2.13}$$

3. ALE-space time discretization

In the time interval $[0,T]$ we consider a partition $0 = t_0 < t_1 < \cdots < t_M = T$ and set $\tau_m = t_m - t_{m-1}$, $I_m = (t_{m-1}, t_m)$, $\overline{I}_m = [t_{m-1}, t_m]$ for $m = 1, \ldots, M$, $\tau = \max_{m=1,\ldots,M} \tau_m$. We assume that $\tau \in (0, \overline{\tau})$, where $\overline{\tau} > 0$. The space-time discontinuous Galerkin method (STDGM) has an advantage that on every time interval $\overline{I}_m = [t_{m-1}, t_m]$ it is possible to consider a different space partition (*i.e.* triangulation) – see, *e.g.* [14, 20]. Here we also use this possibility for the application of the STDGM in the framework of the ALE method. It allows us to consider an ALE mapping separately on each time interval $[t_{m-1}, t_m)$ for $m = 1, \ldots, M$ and the resulting ALE mapping in $[0,T]$ may be discontinuous at time instants t_m, $m = 1, \ldots, M-1$. This means that one-sided limits $\mathcal{A}_{(t_m-)} \neq \mathcal{A}_{(t_m+)}$ in general. Similarly the same may hold for the approximate solution. This means that we deal with a new generalized ALE technique based on the STDGM. To this end, we introduce the following notation.

3.1. ALE mappings and triangulations

For every $m = 1, \dots, M$ we consider a standard conforming triangulation $\hat{\mathcal{T}}_{h,t_{m-1}}$ in $\Omega_{t_{m-1}}$, where $h \in (0, \overline{h})$ and $\overline{h} > 0$. This triangulation is formed by a finite number of closed triangles ($d = 2$) or tetrahedra ($d = 3$) with disjoint interiors. We assume that the domain $\Omega_{t_{m-1}}$ is polygonal (polyhedral). Further, for each $m = 1, \dots, M$ we introduce a one-to-one ALE mapping

$$\mathcal{A}_{h,t}^{m-1} : \overline{\Omega}_{t_{m-1}} \overset{\text{onto}}{\longrightarrow} \overline{\Omega}_t \text{ for } t \in [t_{m-1}, t_m),\ h \in (0, \overline{h}). \tag{3.1}$$

We assume that $\mathcal{A}_{h,t}^{m-1}$ is in space a piecewise affine mapping on the triangulation $\hat{\mathcal{T}}_{h,t_{m-1}}$, continuous in space variable $X \in \Omega_{t_{m-1}}$ and in time $t \in [t_{m-1}, t_m)$ and $\mathcal{A}_{h,t_{m-1}}^{m-1} = \mathrm{Id}$ (identical mapping). Hence, we assume that all domains Ω_t are polygonal (polyhedral). For every $t \in [t_{m-1}, t_m)$ we define the conforming triangulation

$$\mathcal{T}_{h,t} = \left\{ K = \mathcal{A}_{h,t}^{m-1}(\hat{K});\ \hat{K} \in \hat{\mathcal{T}}_{h,t_{m-1}} \right\} \text{ in } \Omega_t. \tag{3.2}$$

This means that every domain Ω_{t_m-1} represents a reference configuration for the ALE mapping $\mathcal{A}_t^{m-1}$ with $t \in I_m$. It is important that this mapping is not an approximation of some regular mapping of Ω_0 onto Ω_t, as is standard in other works.

At $t = t_m$ we define the one-sided limit $\mathcal{A}_{h,t_m-}^{m-1}$, introduce the triangulation

$$\mathcal{T}_{h,t_m-} = \{ \mathcal{A}_{h,t_m-}^{m-1}(\hat{K});\ \hat{K} \in \hat{\mathcal{T}}_{h,t_{m-1}} \} \text{ in } \overline{\Omega}_{t_m}$$

and suppose that

$$\mathcal{A}_{h,t_m}^{m-1} \left(\overline{\Omega}_{t_{m-1}} \right) = \overline{\Omega}_{t_m}. \tag{3.3}$$

We have $\mathcal{T}_{h,t_{m-1}} = \hat{\mathcal{T}}_{h,t_{m-1}}$, but in general, $\mathcal{T}_{h,t_m-} \neq \hat{\mathcal{T}}_{h,t_m}$.

As we see, for every $t \in [0, T]$ we may have a family $\{\mathcal{T}_{h,t}\}_{h \in (0,\overline{h})}$ of triangulations of the domain Ω_t. Triangulations $\hat{\mathcal{T}}_{h,t_{m-1}}$ and $\hat{\mathcal{T}}_{h,t_m}$ have different structure and, in general, different number of cells. Triangulations $\mathcal{T}_{h,t}$ and $\mathcal{T}_{h,t_m-}$ have the same structure as $\hat{\mathcal{T}}_{h,t_{m-1}}$ for $t \in [t_{m-1}, t_m]$, but starting from $\hat{\mathcal{T}}_{h,t_m}$ the structure of $\mathcal{T}_{h,t}$ for $t \in [t_m, t_{m+1}]$, may be different from the structure of $\mathcal{T}_{h,t}$ for $t \in [t_{m-1}, t_m]$.

In what follows, for the sake of simplicity, we use the notation $\mathcal{A}_t$ for the ALE mapping defined in $\bigcup_{m=1}^{M} I_m$ so that

$$\mathcal{A}_t(X) = \mathcal{A}_{h,t}^{m-1}(X) \quad \text{for } X \in \overline{\Omega}_{t_{m-1}},\ t \in \overline{I}_m,\ m = 1, \dots, M,\ h \in (0, \overline{h}). \tag{3.4}$$

The symbol $\mathcal{A}_t^{-1}$ will denote the inverse to $\mathcal{A}_t$. This means that $\mathcal{A}_t^{-1} : \overline{\Omega}_t \overset{\text{onto}}{\longrightarrow} \overline{\Omega}_{t_{m-1}}$ for $t \in \overline{I}_m,\ m = 1, \dots, M$.

3.2. Discrete function spaces

In what follows, for every $m = 1, \dots, M$ we consider the space

$$S_h^{p,m-1} = \left\{ \varphi \in L^2(\Omega_{t_{m-1}});\ \varphi|_{\hat{K}} \in P^p(\hat{K})\ \forall\, \hat{K} \in \hat{\mathcal{T}}_{h,t_{m-1}} \right\}, \tag{3.5}$$

where $p \geq 1$ is an integer and $P^p(\hat{K})$ is the space of all polynomials on $\hat{K}$ of degree $\leq p$. Now for every $t \in \overline{I}_m$ we define the space

$$S_h^{t,p,m-1} = \left\{ \varphi \in L^2(\Omega_t);\ \varphi \circ \mathcal{A}_{h,t}^{m-1} \in S_h^{p,m-1} \right\}. \tag{3.6}$$

It is possible to see that

$$S_h^{t,p,m-1} = \left\{ \varphi \in L^2(\Omega_t);\ \varphi|_K \in P^p(K)\ \forall\, K \in \mathcal{T}_{h,t} \right\}. \tag{3.7}$$

Of course, $S_h^{t_m,p,m-1} \neq S_h^{p,m}$ in general.

Further, let p, $q \geq 1$ be integers. By $P^q(I_m; S_h^{p,m-1})$ we denote the space of mappings of the time interval I_m into the space $S_h^{p,m-1}$ which are polynomials of degree $\leq q$ in time. We set

$$S_{h,\tau}^{p,q} = \left\{ \varphi;\ \varphi\left(\mathcal{A}_{h,t}^{m-1}(X), t\right) = \sum_{i=0}^{q} \vartheta_i(X)\, t^i, \quad \vartheta_i \in S_h^{p,m-1},\ X \in \Omega_{t_{m-1}},\ t \in \overline{I}_m,\ m = 1, \ldots, M \right\}. \tag{3.8}$$

An approximate solution of problem (2.11)–(2.13) and test functions will be elements of the space $S_{h,\tau}^{p,q}$. By D_t we denote the ALE derivative defined by (2.9) for $t \in \bigcup_{m=1}^{M} I_m$.

3.3. Some notation and important concepts

Over a triangulation $\mathcal{T}_{h,t}$, for each positive integer k, we define the broken Sobolev space

$$H^k(\Omega_t, \mathcal{T}_{h,t}) = \{v;\ v|_K \in H^k(K) \quad \forall K \in \mathcal{T}_{h,t}\},$$

equipped with the seminorm

$$|v|_{H^k(\Omega_t, \mathcal{T}_{h,t})} = \left(\sum_{K \in \mathcal{T}_{h,t}} |v|^2_{H^k(K)} \right)^{1/2},$$

where $|\cdot|_{H^k(K)}$ denotes the seminorm in the space $H^k(K)$.

By $\mathcal{F}_{h,t}$ we denote the system of all faces of all elements $K \in \mathcal{T}_{h,t}$. It consists of the set of all inner faces $\mathcal{F}_{h,t}^I$ and the set of all boundary faces $\mathcal{F}_{h,t}^B$: $\mathcal{F}_{h,t} = \mathcal{F}_{h,t}^I \cup \mathcal{F}_{h,t}^B$. Each $\Gamma \in \mathcal{F}_{h,t}$ will be associated with a unit normal vector $\mathbf{n}_\Gamma$. By $K_\Gamma^{(L)}$ and $K_\Gamma^{(R)} \in \mathcal{T}_{h,t}$ we denote the elements adjacent to the face $\Gamma \in \mathcal{F}_{h,t}^I$. Moreover, for $\Gamma \in \mathcal{F}_{h,t}^B$ the element adjacent to this face will be denoted by $K_\Gamma^{(L)}$. We shall use the convention, that $\mathbf{n}_\Gamma$ is the outer normal to $\partial K_\Gamma^{(L)}$.

If $v \in H^1(\Omega_t, \mathcal{T}_{h,t})$ and $\Gamma \in \mathcal{F}_{h,t}$, then $v_\Gamma^{(L)}$ and $v_\Gamma^{(R)}$ will denote the traces of v on Γ from the side of elements $K_\Gamma^{(L)}$ and $K_\Gamma^{(R)}$, respectively. We set $h_K = \operatorname{diam} K$ for $K \in \mathcal{T}_{h,t}$, $h(\Gamma) = \operatorname{diam} \Gamma$ for $\Gamma \in \mathcal{F}_{h,t}$ and $\langle v \rangle_\Gamma = \frac{1}{2}\left(v_\Gamma^{(L)} + v_\Gamma^{(R)}\right)$, $[v]_\Gamma = v_\Gamma^{(L)} - v_\Gamma^{(R)}$, for $\Gamma \in \mathcal{F}_{h,t}^I$. Moreover, by ρ_K we denote the diameter of the largest ball inscribed into $K \in \mathcal{T}_{h,t}$.

3.4. Discretization

First we introduce the space semidiscretization of problem (2.11)–(2.13). We assume that u is a sufficiently smooth solution of our problem. If we choose an arbitrary but fixed $t \in (0, T)$, multiply equation (2.11) by a test function $\varphi \in H^2(\Omega_t, \mathcal{T}_{h,t})$, integrate over any element K and finally sum over all elements $K \in \mathcal{T}_{h,t}$, then for $t \in I_m$ we get

$$\begin{aligned} &\sum_{K \in \mathcal{T}_{h,t}} \int_K \frac{Du}{Dt} \varphi \,\mathrm{d}x + \sum_{K \in \mathcal{T}_{h,t}} \int_K \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} \varphi \,\mathrm{d}x \\ &\quad - \sum_{K \in \mathcal{T}_{h,t}} \int_K \sum_{s=1}^{d} z_s \frac{\partial u}{\partial x_s} \varphi \,\mathrm{d}x - \sum_{K \in \mathcal{T}_{h,t}} \int_K \operatorname{div}(\beta(u) \nabla u) \varphi \,\mathrm{d}x = \sum_{K \in \mathcal{T}_{h,t}} \int_K g \varphi \,\mathrm{d}x. \end{aligned} \tag{3.9}$$

Applying Green's theorem to the convection and diffusion terms, introducing the concept of a numerical flux and suitable expressions mutually vanishing, after some manipulation we arrive at the identity

$$(D_t u, \varphi) + A_h(u, \varphi, t) + b_h(u, \varphi, t) + d_h(u, \varphi, t) = l_h(\varphi, t), \tag{3.10}$$

where the forms appearing here are defined for $u, \varphi \in H^2(\Omega_t, \mathcal{T}_{h,t})$, $\theta \in I\!R$ and $c_W > 0$ in the following way

$$
\begin{aligned}
a_h(u,\varphi,t) := & \sum_{K\in\mathcal{T}_{h,t}} \int_K \beta(u)\nabla u\cdot\nabla\varphi \,\mathrm{d}x \\
& - \sum_{\Gamma\in\mathcal{F}^I_{h,t}} \int_\Gamma \left(\langle\beta(u)\nabla u\rangle\cdot\mathbf{n}_\Gamma\,[\varphi] + \theta\,\langle\beta(u)\nabla\varphi\rangle\cdot\mathbf{n}_\Gamma\,[u]\right)\,\mathrm{d}S \\
& - \sum_{\Gamma\in\mathcal{F}^B_{h,t}} \int_\Gamma \left(\beta(u)\nabla u\cdot\mathbf{n}_\Gamma\,\varphi + \theta\beta(u)\nabla\varphi\cdot\mathbf{n}_\Gamma\,u - \theta\beta(u)\nabla\varphi\cdot\mathbf{n}_\Gamma\,u_D\right)\,\mathrm{d}S,
\end{aligned} \tag{3.11}
$$

$$
J_h(u,\varphi,t) := c_W \sum_{\Gamma\in\mathcal{F}^I_{h,t}} h(\Gamma)^{-1}\int_\Gamma [u]\,[\varphi]\,\mathrm{d}S + c_W \sum_{\Gamma\in\mathcal{F}^B_{h,t}} h(\Gamma)^{-1}\int_\Gamma u\,\varphi\,\mathrm{d}S, \tag{3.12}
$$

$$
J^B_h(u,\varphi,t) := c_W \sum_{\Gamma\in\mathcal{F}^B_{h,t}} h(\Gamma)^{-1}\int_\Gamma u\,\varphi\,\mathrm{d}S, \tag{3.13}
$$

$$
A_h(u,\varphi,t) := a_h(u,\varphi,t) + \beta_0\, J_h(u,\varphi,t), \tag{3.14}
$$

$$
\begin{aligned}
b_h(u,\varphi,t) := & -\sum_{K\in\mathcal{T}_{h,t}} \int_K \sum_{s=1}^d f_s(u)\frac{\partial\varphi}{\partial x_s}\,\mathrm{d}x \\
& + \sum_{\Gamma\in\mathcal{F}^I_{h,t}} \int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(R)}, \mathbf{n}_\Gamma)\,[\varphi]\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}^B_{h,t}} \int_\Gamma H(u_\Gamma^{(L)}, u_\Gamma^{(L)}, \mathbf{n}_\Gamma)\,\varphi\,\mathrm{d}S,
\end{aligned} \tag{3.15}
$$

$$
d_h(u,\varphi,t) := -\sum_{K\in\mathcal{T}_{h,t}} \int_K \sum_{s=1}^d z_s\frac{\partial u}{\partial x_s}\varphi\,\mathrm{d}x = -\sum_{K\in\mathcal{T}_{h,t}} \int_K (\boldsymbol{z}\cdot\nabla u)\varphi\,\mathrm{d}x, \tag{3.16}
$$

$$
l_h(\varphi,t) := \sum_{K\in\mathcal{T}_{h,t}} \int_K g\varphi\,\mathrm{d}x + \beta_0\, c_W \sum_{\Gamma\in\mathcal{F}^B_{h,t}} h(\Gamma)^{-1}\int_\Gamma u_D\,\varphi\,\mathrm{d}S. \tag{3.17}
$$

Let us note that in integrals over faces we omit the subscript Γ of $\langle\cdot\rangle$ and $[\cdot]$. We consider $\theta = 1$, $\theta = 0$ and $\theta = -1$ and get the symmetric (SIPG), incomplete (IIPG) and nonsymmetric (NIPG) variants of the approximation of the diffusion terms, respectively.

In (3.15), H is a numerical flux with the following properties:

(H1) $H(u,v,\mathbf{n})$ is defined in $\mathbb{R}^2\times B_1$, where $B_1 = \{\mathbf{n} = (n_1,\dots,n_d)\in\mathbb{R}^d;\ |\mathbf{n}| = 1\}$, and is Lipschitz-continuous with respect to u, v: there exists $L_H > 0$ such that

$$
|H(u,v,\mathbf{n}) - H(u^*,v^*,\mathbf{n})| \le L_H(|u-u^*| + |v-v^*|), \text{ for all } u,v,u^*,v^*\in\mathbb{R}.
$$

(H2) H is consistent: $H(u,u,\mathbf{n}) = \sum_{s=1}^d f_s(u)n_s, \quad u\in\mathbb{R}$, $\mathbf{n}\in B_1$,

(H3) H is conservative: $H(u,v,\mathbf{n}) = -H(v,u,-\mathbf{n}), \quad u,v\in\mathbb{R}$, $\mathbf{n}\in B_1$.

In what follows, in the stability analysis we shall use the properties **(H1)** and **(H2)**. (Assumption **(H3)** is important for error estimation, but here it is not necessary.)

For a function φ defined in $\bigcup_{m=1}^M I_m$ we denote

$$
\varphi_m^\pm = \varphi(t_m\pm) = \lim_{t\to t_m\pm}\varphi(t), \quad \{\varphi\}_m = \varphi(t_m+) - \varphi(t_m-), \tag{3.18}
$$

if the one-sided limits $\varphi_m^\pm$ exist.

Now we define an ALE-STDG approximate solution of problem (2.11)–(2.13).

Definition 3.1. A function U is an approximate solution of problem (2.11)–(2.13), if $U \in S_{h,\tau}^{p,q}$ and

$$\int_{I_m} \Big((D_t U, \varphi)_{\Omega_t} + A_h(U,\varphi,t) + b_h(U,\varphi,t) + d_h(U,\varphi,t) \Big)\,\mathrm{d}t \tag{3.19}$$
$$+ (\{U\}_{m-1}, \varphi_{m-1}^+)_{\Omega_{t_{m-1}}} = \int_{I_m} l_h(\varphi,t)\,\mathrm{d}t \quad \forall \varphi \in S_{h,\tau}^{p,q}, \quad m = 1,\dots,M,$$
$$U_0^- \in S_h^{p,0}, \quad (U_0^- - u^0, v_h) = 0 \quad \forall v_h \in S_h^{p,0}. \tag{3.20}$$

(For $m = 1$ we set $\{U\}_{m-1} = \{U\}_0 := U_0^+ - U_0^-$ with U_0^- given by (3.20)).

The ALE-STDG numerical method (3.19)–(3.20) was applied in [16, 44] to the numerical simulation of a compressible flow in time-dependent domains and fluid-structure interaction.

4. Analysis of the stability

In what follows we shall be concerned with the numerical solution of the ALE problem (2.11)–(2.13) by the space-time discontinuous Galerkin method. In the theoretical analysis a number of various constants will appear. Some important constants in main assertions will be denoted by C_{L1}, C_{L1}^*, C_{L1}^{**}, etc. in Lemma 4.1, C_{L2}, etc. in Lemma 4.2, etc. and C_{T1}, C_{T1}^*, C_{T2}, C_{T2}^*, etc. in Theorems 4.1, 4.2, etc. Further, we use special notation of constants appearing in properties of various structures, *e.g.* L_f, L_β, L_H, c_R, etc. Inside proofs, constants are denoted locally by c, c_1, c_2, c^* etc. The aim of this notation is to increase the readability of the paper and to show the relations between individual theorems and lemmas.

4.1. Some auxiliary results

As was mentioned in Section 3.1, for each $t \in [0,T]$ we consider a system of triangulations $\{\mathcal{T}_{h,t}\}_{h\in(0,\overline{h})}$. We assume that these systems are uniformly shape regular. This means that there exists a positive constant c_R, independent of K, t and h such that

$$\frac{h_K}{\rho_K} \le c_R \qquad \text{for all} \quad K \in \mathcal{T}_{h,t},\ h \in (0,\overline{h}), t \in [t_{m-1}, t_m], \tag{4.1}$$
$$\tau_m \le \tau \in (0,\overline{\tau}), \quad m = 1,\dots,M.$$

By $(\mathcal{A}_{h,t}^{m-1})^{-1}$ we denote the inverse to the mapping $\mathcal{A}_{h,t}^{m-1}$. The symbols $\frac{\mathrm{d}\mathcal{A}_{h,t}^{m-1}}{\mathrm{d}X}$ and $\frac{\mathrm{d}(\mathcal{A}_{h,t}^{m-1})^{-1}}{\mathrm{d}x}$ denote the Jacobian matrices of $\mathcal{A}_{h,t}^{m-1}$ and $(\mathcal{A}_{h,t}^{m-1})^{-1}$, respectively. The entries of $\frac{\mathrm{d}\mathcal{A}_{h,t}^{m-1}}{\mathrm{d}X}$ and $\frac{\mathrm{d}(\mathcal{A}_{h,t}^{m-1})^{-1}}{\mathrm{d}x}$ are constant on every element $\hat{K} \in \hat{\mathcal{T}}_{h,t_{m-1}}$ and $K \in \mathcal{T}_{h,t}$, respectively. Moreover, we define the Jacobians $J(X,t) = \det\frac{\mathrm{d}\mathcal{A}_{h,t}^{m-1}(X)}{\mathrm{d}X}$, $X \in \Omega_{t_{m-1}}$, and $J^{-1}(x,t) = \det\frac{\mathrm{d}(\mathcal{A}_{h,t}^{m-1}(x))^{-1}}{\mathrm{d}x}$, $x \in \Omega_t$. The Jacobians J and J^{-1} are piecewise constant over $\hat{\mathcal{T}}_{h,t_{m-1}}$ and $\mathcal{T}_{h,t}$, respectively. The constant value of J on $\hat{K} \in \hat{\mathcal{T}}_{h,t_{m-1}}$ and of J^{-1} on $K \in \mathcal{T}_{h,t}$ will be denoted by $J_{\hat{K}}$ and J_K^{-1}, respectively. Of course, these terms depend on t and, hence, $J_{\hat{K}} = J_{\hat{K}}(t)$ and $J_K^{-1} = J_K^{-1}(t)$.

In what follows, we assume that

$$\mathcal{A}_{h,t}^{m-1} \in W^{1,\infty}(I_m; W^{1,\infty}(\Omega_{t_{m-1}})), \quad m = 1,\dots,M,\ h \in (0,\overline{h}) \tag{4.2}$$

and

$$(A_{h,t}^{m-1})^{-1} \in W^{1,\infty}(I_m; W^{1,\infty}(\Omega_t)), \quad m = 1,\dots M,\ h \in (0,\overline{h}). \tag{4.3}$$

Obviously, we have $J \in W^{1,\infty}(I_m; L^\infty(\Omega_{t_{m-1}}))$, $J^{-1} \in W^{1,\infty}(I_m; L^\infty(\Omega_t))$. Since $\mathcal{A}_{h,t_{m-1}}^{m-1}$ is the identical mapping and, hence, $J(X, t_{m-1}) = 1$, we assume that there exist constants C_J^-, $C_J^+ > 0$ such that the Jacobians

satisfy the conditions

$$C_J^- \le J(X,t) \le C_J^+, \quad X \in \overline{\Omega}_{t_{m-1}},\ t \in \overline{I}_m,\ m = 1,\dots,M,\ h \in (0,\overline{h}), \tag{4.4}$$
$$(C_J^+)^{-1} \le J^{-1}(x,t) \le (C_J^-)^{-1}, \quad x \in \overline{\Omega}_t,\ t \in \overline{I}_m,\ m = 1,\dots,M,\ h \in (0,\overline{h}).$$

Finally, there exist constants $C_A^-, C_A^+ > 0$ such that

$$\left\| \frac{\mathrm{d}\mathcal{A}_{h,t}^{m-1}(X)}{\mathrm{d}X} \right\| \le C_A^+,\ X \in \overline{\Omega}_{t_{m-1}},\ t \in \overline{I}_m,\ m = 1,\dots,M,\ h \in (0,\overline{h}), \tag{4.5}$$
$$\left\| \frac{\mathrm{d}(\mathcal{A}_{h,t}^{m-1})^{-1}(x)}{\mathrm{d}x} \right\| \le C_A^-,\ x \in \overline{\Omega}_t,\ t \in \overline{I}_m,\ m = 1,\dots,M,\ h \in (0,\overline{h}), \tag{4.6}$$

where $\|\cdot\|$ is the matrix norm induced by the Euclidean norm $|\cdot|$ in $\mathbb{R}^d$.

The above assumptions imply the following properties of the domain velocity: There exists a constant $c_z > 0$ such that

$$|\boldsymbol{z}(x,t)|,\ |\mathrm{div}\boldsymbol{z}(x,t)| \le c_z \quad \text{for } x \in \Omega_t,\ t \in (0,T). \tag{4.7}$$

Under assumption (4.1), the multiplicative trace inequality and the inverse inequality hold: There exist constants $c_M, c_I > 0$ independent of v, h, t and K such that

$$\|v\|_{L^2(\partial K)}^2 \le c_M \left(\|v\|_{L^2(K)} |v|_{H^1(K)} + h_K^{-1} \|v\|_{L^2(K)}^2 \right), \tag{4.8}$$
$$v \in H^1(K),\ K \in \mathcal{T}_{h,t},\ h \in (0,\overline{h}),\ t \in [0,T],$$

and

$$|v|_{H^1(K)} \le c_I\, h_K^{-1} \|v\|_{L^2(K)}, \quad v \in P^p(K),\ K \in \mathcal{T}_{h,t},\ h \in (0,\overline{h}),\ t \in [0,T]. \tag{4.9}$$

In the space $H^1(\Omega_t, \mathcal{T}_{h,t})$ we define the norm

$$\|\varphi\|_{\mathrm{DG},t} = \left(\sum_{K \in \mathcal{T}_{h,t}} |\varphi|_{H^1(K)}^2 + J_h(\varphi,\varphi,t) \right)^{1/2}. \tag{4.10}$$

Moreover, over $\partial\Omega$ we define the norm

$$\|u_D\|_{\mathrm{DGB},t} = \left(c_W \sum_{\Gamma \in \mathcal{F}_{h,t}^B} h(\Gamma)^{-1} \int_\Gamma |u_D|^2 \,\mathrm{d}S \right)^{1/2} = \left(J_h^B(u_D,u_D,t) \right)^{1/2}. \tag{4.11}$$

If we use $\varphi := U$ as a test function in (3.19), we get the basic identity

$$\int_{I_m} \Big((D_t U, U)_{\Omega_t} + A_h(U,U,t) + b_h(U,U,t) + d_h(U,U,t) \Big) \mathrm{d}t \tag{4.12}$$
$$+ (\{U\}_{m-1}, U_{m-1}^+)_{\Omega_{t_{m-1}}} = \int_{I_m} l_h(U,t)\,\mathrm{d}t.$$

In what follows we need to estimate each term in (4.12). These estimates are summarized in Section 4.2. The skipped proofs can be found in [5]. They are based on the multiplicative trace inequality (4.8), inverse inequality (4.9), Young's inequality and assumptions (2.5) on the function β.

These estimates, apart from another, produce a problematic term $\int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t$, which we need to estimate in terms of data. To overcome this difficulty we generalize the concept of discrete characteristic function in time-dependent domains. In Theorem 4.1 we prove the continuity of the previously defined discrete characteristic function in $\|\cdot\|_{\Omega_t}$ and $\|\cdot\|_{\mathrm{DG},t}$ norms.

Then, in Theorems 4.2 and 4.3 we apply estimates from Section 4.2 to the basic identity (4.12). In Lemmas 4.6–4.10 we estimate similar terms in Section 4.2, but the test function (second variable) is replaced by the discrete characteristic function. Using these lemmas and properties of the discrete characteristic function proved in Theorem 4.1, we finally estimate the problematic term $\int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t$ in terms of data in Theorem 4.4.

Using this key result from Theorem 4.4 and the discrete Gronwall inequality from Lemma 4.11, the unconditional stability of the method is proved in Theorem 4.5.

4.2. Important estimates

Here we estimate the forms from (4.12). The proofs can be carried out in a similar way as in [5]. For a sufficiently large constant c_W we obtain the coercivity of the diffusion and penalty terms.

Lemma 4.1. *Let*

$$c_W \geq \frac{\beta_1^2}{\beta_0^2} c_M (c_I + 1) \quad \text{for} \quad \theta = -1 \text{ (NIPG)}, \tag{4.13}$$

$$c_W \geq \frac{\beta_1^2}{\beta_0^2} c_M (c_I + 1) \quad \text{for} \quad \theta = 0 \text{ (IIPG)}, \tag{4.14}$$

$$c_W \geq \frac{16\beta_1^2}{\beta_0^2} c_M (c_I + 1) \quad \text{for} \quad \theta = 1 \text{ (SIPG)}. \tag{4.15}$$

Then

$$\int_{I_m} \left(a_h(U,U,t) + \beta_0 J_h(U,U,t)\right) \mathrm{d}t \geq \frac{\beta_0}{2} \int_{I_m} \|U\|^2_{\mathrm{DG},t} \,\mathrm{d}t - \frac{\beta_0}{2} \int_{I_m} \|u_D\|^2_{\mathrm{DGB},t} \,\mathrm{d}t. \tag{4.16}$$

Further, we estimate the convection terms and the right-hand side form:

Lemma 4.2. *For each $k_1, k_2, k_3 > 0$ there exists a constant $c_b, c_d > 0$ such that we have*

$$\int_{I_m} |b_h(U,U,t)| \mathrm{d}t \leq \frac{\beta_0}{2k_1} \int_{I_m} \|U\|^2_{\mathrm{DG},t} \mathrm{d}t + c_b \int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t, \tag{4.17}$$

$$\int_{I_m} |d_h(U,U,t)| \,\mathrm{d}t \leq \frac{\beta_0}{2k_2} \int_{I_m} \|U\|^2_{\mathrm{DG},t} \,\mathrm{d}t + \frac{c_d}{2\beta_0} \int_{I_m} \|U\|^2_{\Omega_t} \,\mathrm{d}t, \tag{4.18}$$

$$\begin{aligned} \int_{I_m} |l_h(U,t)| \,\mathrm{d}t \leq{}& \frac{1}{2} \int_{I_m} \left(\|g\|^2_{\Omega_t} + \|U\|^2_{\Omega_t}\right) \mathrm{d}t \\ &+ \frac{\beta_0 k_3}{2} \int_{I_m} \|u_D\|^2_{\mathrm{DGB},t} \,\mathrm{d}t + \frac{\beta_0}{2k_3} \int_{I_m} \|U\|^2_{\mathrm{DG},t} \,\mathrm{d}t. \end{aligned} \tag{4.19}$$

Finally we need to estimate the term with the ALE derivative:

Lemma 4.3. *It holds that*

$$\int_{I_m} (D_t U, U)_{\Omega_t} \,\mathrm{d}t \geq \frac{1}{2}\left(\|U_m^-\|^2_{\Omega_{t_m}} - \|U_{m-1}^+\|^2_{\Omega_{t_{m-1}}} - c_z \int_{I_m} \|U\|^2_{\Omega_t} \,\mathrm{d}t \right), \tag{4.20}$$

$$\left(\{U\}_{m-1}, U_{m-1}^+\right)_{\Omega_{t_{m-1}}} = \frac{1}{2}\left(\|U_{m-1}^+\|^2_{\Omega_{t_{m-1}}} + \|\{U\}_{m-1}\|^2_{\Omega_{t_{m-1}}} - \|U_{m-1}^-\|^2_{\Omega_{t_{m-1}}} \right), \tag{4.21}$$

$$\begin{aligned} &\int_{I_m} (D_t U, U)_{\Omega_t} \,\mathrm{d}t + \left(\{U\}_{m-1}, U_{m-1}^+\right)_{\Omega_{t_{m-1}}} \\ &\qquad \geq \frac{1}{2}\|U_m^-\|^2_{\Omega_{t_m}} + \frac{1}{2}\|U_{m-1}^+\|^2_{\Omega_{t_{m-1}}} - \frac{c_z}{2}\int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t - \left(U_{m-1}^-, U_{m-1}^+\right)_{\Omega_{t_{m-1}}}. \end{aligned} \tag{4.22}$$

Proof. We start with the first inequality. We have

$$\int_{I_m} (D_t U, U)_{\Omega_t} \,\mathrm{d}t = \int_{I_m} \sum_{K \in \mathcal{T}_{h,t}} (D_t U, U)_K \,\mathrm{d}t. \tag{4.23}$$

By virtue of relation (3.2), the Reynolds transport theorem (see, *e.g.* [27] or [1]) and relation (2.10), we get

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\int_K U^2(x,t)\,\mathrm{d}x &= \int_K \left(\frac{\partial U^2(x,t)}{\partial t} + \boldsymbol{z}(x,t)\cdot\nabla(U^2(x,t)) + U^2(x,t)\mathrm{div}\,\boldsymbol{z}(x,t) \right) \mathrm{d}x \\ &= \int_K \left(2U(x,t)\left(\frac{\partial U(x,t)}{\partial t} + \boldsymbol{z}(x,t)\cdot\nabla U(x,t) \right) + U^2(x,t)\mathrm{div}\,\boldsymbol{z}(x,t) \right) \mathrm{d}x \\ &= 2(D_t U, U)_K + (U^2, \mathrm{div}\,\boldsymbol{z})_K. \end{aligned} \tag{4.24}$$

Expressing $(D_t U, U)_K$, summing over $K \in \mathcal{T}_{h,t}$ and integrating over I_m together with assumption (4.7) yield

$$\begin{aligned} \int_{I_m} (D_t U, U)_{\Omega_t} \,\mathrm{d}t &= \frac{1}{2}\int_{I_m} \frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega_t} U^2 \,\mathrm{d}x\,\mathrm{d}t - \frac{1}{2}\int_{I_m} (U^2, \mathrm{div}\,\boldsymbol{z})_{\Omega_t}\,\mathrm{d}t \\ &\geq \frac{1}{2}\|U_m^-\|^2_{\Omega_{t_m}} - \frac{1}{2}\|U_{m-1}^+\|^2_{\Omega_{t_{m-1}}} - \frac{c_z}{2}\int_{I_m} \|U\|^2_{\Omega_t}\,\mathrm{d}t, \end{aligned} \tag{4.25}$$

which gives (4.20).

Further, by a simple manipulation we find that

$$2(U_{m-1}^+ - U_{m-1}^-, U_{m-1}^+)_{\Omega_{t_{m-1}}} = \|U_{m-1}^+\|^2_{\Omega_{t_{m-1}}} + \|\{U\}_{m-1}\|^2_{\Omega_{t_{m-1}}} - \|U_{m-1}^-\|^2_{\Omega_{t_{m-1}}},$$

which immediately implies (4.21).

Concerning inequality (4.22), from (4.25) we get

$$\begin{aligned} &\int_{I_m} (D_t U, U)_{\Omega_t} \mathrm{d}t + \left(\{U\}_{m-1}, U_{m-1}^+\right)_{\Omega_{t_{m-1}}} \\ &\qquad = \frac{1}{2}\|U_m^-\|^2_{\Omega_{t_m}} - \frac{1}{2}\|U_{m-1}^+\|^2_{\Omega_{t_{m-1}}} - \frac{1}{2}\int_{I_m} (U^2, \mathrm{div}\,\boldsymbol{z})_{\Omega_t}\mathrm{d}t + \|U_{m-1}^+\|_{\Omega_{t_{m-1}}} - (U_{m-1}^-, U_{m-1}^+)_{\Omega_{t_{m-1}}} \\ &\qquad \geq \frac{1}{2}\left(\|U_m^-\|^2_{\Omega_{t_m}} + \|U_{m-1}^+\|^2_{\Omega_{t_{m-1}}} - c_z\int_{I_m} \|U\|^2_{\Omega_t}\mathrm{d}t \right) - \left(U_{m-1}^-, U_{m-1}^+\right)_{\Omega_{t_{m-1}}}, \end{aligned}$$

which proves the lemma. □

4.3. Discrete characteristic function

In our further considerations, the concept of a discrete characteristic function will play an important role, which is generalized to time-dependent domains.

For $m = 1, \dots, M$ we use the following notation: $U = U(x,t)$, $x \in \Omega_t$, $t \in I_m$ will denote the approximate solution in Ω_t, and $\tilde{U} = \tilde{U}(X,t) = U(\mathcal{A}_t(X),t)$, $X \in \Omega_{t_{m-1}}$ $t \in I_m$ denotes the approximate solution transformed to the reference domain $\Omega_{t_{m-1}}$.

For $s \in I_m$ we denote $\tilde{\mathcal{U}}_s = \tilde{\mathcal{U}}_s(X,t)$, $X \in \Omega_{t_{m-1}}$, $t \in I_m$, the discrete characteristic function to $\tilde{U}$ at a point $s \in I_m$. It is defined as $\tilde{\mathcal{U}}_s \in P^q(I_m; S_h^{p,m-1})$ such that

$$\int_{I_m} (\tilde{\mathcal{U}}_s, \varphi)_{\Omega_{t_{m-1}}} \,\mathrm{d}t = \int_{t_{m-1}}^{s} (\tilde{U}, \varphi)_{\Omega_{t_{m-1}}} \,\mathrm{d}t \quad \forall \varphi \in P^{q-1}(I_m; S_h^{p,m-1}), \tag{4.26}$$

$$\tilde{\mathcal{U}}_s(X, t_{m-1}{}^+) = \tilde{U}(X, t_{m-1}{}^+), \ X \in \Omega_{t_{m-1}}. \tag{4.27}$$

The existence and uniqueness of the discrete characteristic function satisfying (4.26) and (4.27) is proved in the monograph [20]. Further, we introduce the discrete characteristic function $\mathcal{U}_s = \mathcal{U}_s(x,t)$, $x \in \Omega_t, t \in I_m$ to $U \in S_{h,\tau}^{p,q}$ at a point $s \in I_m$:

$$\mathcal{U}_s(x,t) = \tilde{\mathcal{U}}_s(\mathcal{A}_t^{-1}(x), t), \ x \in \Omega_t, \ t \in I_m. \tag{4.28}$$

Hence, in view of (3.8), $\mathcal{U}_s \in S_{h,\tau}^{p,q}$ and for $X \in \Omega_{t_{m-1}}$ we have

$$\mathcal{U}_s(X, t_{m-1}+) = U(X, t_{m-1}+). \tag{4.29}$$

In what follows, we prove some important properties of the discrete characteristic function. Namely, we prove that the discrete characteristic function mapping $U \to \mathcal{U}_s$ is continuous with respect of the norms $\|\cdot\|_{L^2(\Omega_t)}$ and $\|\cdot\|_{\mathrm{DG},t}$. In the proof we use a result from [7] for the discrete characteristic function on a reference domain: There exists a constant $\tilde{c}_{CH}^{(1)} > 0$ depending on q only such that

$$\int_{I_m} \|\tilde{\mathcal{U}}_s\|_{\Omega_{t_{m-1}}}^2 \,\mathrm{d}t \le \tilde{c}_{CH}^{(1)} \int_{I_m} \|\tilde{U}\|_{\Omega_{t_{m-1}}}^2 \,\mathrm{d}t, \tag{4.30}$$

for all $m = 1, \dots, M$ and $h \in (0, \overline{h})$.

Lemma 4.4. *There exist constants C_{L4}^*, $C_{L4}^{**} > 0$ such that*

$$C_{L4}^* \, h(\hat{\Gamma})^{-1} \le h(\Gamma)^{-1} \le C_{L4}^{**} \, h(\hat{\Gamma})^{-1} \tag{4.31}$$

for all $\hat{\Gamma} \in \mathcal{F}_{h,t_{m-1}}, \Gamma = \mathcal{A}_t(\hat{\Gamma}) \in \mathcal{F}_{h,t}$ and all $t \in \overline{I}_m$, $m = 1, \dots, M$, $h \in (0, \overline{h})$.

Proof. We use the relation between Γ and $\hat{\Gamma}$ and the properties (4.5) and (4.6) of the mappings $\mathcal{A}_t$ and $\mathcal{A}_t^{-1}$. We also take into account that $\hat{\Gamma} \subset \hat{K}$ for some $\hat{K} \in \hat{\mathcal{T}}_{h,t_{m-1}}$, $\Gamma \subset K = \mathcal{A}_t(\hat{K}) \in \mathcal{T}_{h,t}$ and that the Jacobian matrices $\frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}$ and $\frac{\mathrm{d}\mathcal{A}_t^{-1}}{\mathrm{d}x}$ are constant on $\hat{K}$ and K, respectively. Then we can write

$$\begin{aligned} h(\Gamma) = \operatorname{diam}(\Gamma) &= \max_{x,x^* \in \Gamma} |x - x^*| = \max_{X, X^* \in \hat{\Gamma}} |\mathcal{A}_t(X) - \mathcal{A}_t(X^*)| \\ &\le \max_{X \in \hat{\Gamma}} \left\| \frac{\mathrm{d}\mathcal{A}_t(X)}{\mathrm{d}X} \right\| \max_{X, X^* \in \hat{\Gamma}} |X - X^*| \le C_A^+ \max_{X, X^* \in \hat{\Gamma}} |X - X^*| = C_A^+ \, h(\hat{\Gamma}). \end{aligned}$$

Similarly, we get $h(\hat{\Gamma}) \le C_A^- \, h(\Gamma)$. These inequalities immediately imply (4.31) with $C_{L4}^* = (C_A^+)^{-1}$ and $C_{L4}^{**} = C_A^-$. □

Theorem 4.1. *There exist constants $C^*_{T1}, C^{**}_{T1} > 0$ such that*

$$\int_{I_m} \|\mathcal{U}_s\|^2_{\Omega_t} \,\mathrm{d}t \le C^*_{T1} \int_{I_m} \|U\|^2_{\Omega_t} \,\mathrm{d}t \tag{4.32}$$

$$\int_{I_m} \|\mathcal{U}_s\|^2_{\mathrm{DG},t} \,\mathrm{d}t \le C^{**}_{T1} \int_{I_m} \|U\|^2_{\mathrm{DG},t} \,\mathrm{d}t \tag{4.33}$$

for all $s \in I_m$, $m = 1, \dots, M$ and $h \in (0, \overline{h})$.

Proof. We begin with the proof of the first inequality. We have

$$\begin{aligned}
\|\mathcal{U}_s(t)\|^2_{\Omega_t} &= \int_{\Omega_t} |\mathcal{U}_s(x,t)|^2 \,\mathrm{d}x = \int_{\Omega_t} |\tilde{\mathcal{U}}_s(\mathcal{A}_t^{-1}(x),t)|^2 \,\mathrm{d}x \\
&= \int_{\Omega_{t_{m-1}}} |\tilde{\mathcal{U}}_s(X,t)|^2 J(X,t) \,\mathrm{d}X \le C_J^+ \int_{\Omega_{t_{m-1}}} |\tilde{\mathcal{U}}_s(X,t)|^2 \,\mathrm{d}X \\
&= C_J^+ \|\tilde{\mathcal{U}}_s(t)\|^2_{\Omega_{t_{m-1}}}
\end{aligned}$$

Integrating over I_m and using (4.30) and (4.4), we obtain

$$\begin{aligned}
\int_{I_m} \|\mathcal{U}_s(t)\|^2_{\Omega_t} \,\mathrm{d}t &\le C_J^+ \int_{I_m} \|\tilde{\mathcal{U}}_s(t)\|^2_{\Omega_{t_{m-1}}} \,\mathrm{d}t \\
&\le C_J^+ \tilde{c}^{(1)}_{CH} \int_{I_m} \|\tilde{U}(t)\|^2_{\Omega_{t_{m-1}}} \,\mathrm{d}t \\
&= C_J^+ \tilde{c}^{(1)}_{CH} \int_{I_m} \left(\int_{\Omega_{t_{m-1}}} |\tilde{U}(X,t)|^2 \,\mathrm{d}X \right) \mathrm{d}t \\
&= C_J^+ \tilde{c}^{(1)}_{CH} \int_{I_m} \left(\int_{\Omega_{t_{m-1}}} |U(\mathcal{A}_t(X),t)|^2 \,\mathrm{d}X \right) \mathrm{d}t \\
&= C_J^+ \tilde{c}^{(1)}_{CH} \int_{I_m} \left(\int_{\Omega_t} |U(x,t)|^2 J^{-1}(x,t) \,\mathrm{d}x \right) \mathrm{d}t \\
&\le C_J^+ \tilde{c}^{(1)}_{CH} (C_J^-)^{-1} \int_{I_m} \left(\int_{\Omega_t} |U(x,t)|^2 \,\mathrm{d}x \right) \mathrm{d}t \\
&= C_J^+ \tilde{c}^{(1)}_{CH} (C_J^-)^{-1} \int_{I_m} \|U(t)\|^2_{\Omega_t} \,\mathrm{d}t.
\end{aligned}$$

Setting $C^*_{T1} = C_J^+ \tilde{c}^{(1)}_{CH} (C_J^-)^{-1}$, we get (4.32).

Now we pay our attention to the proof of the second inequality in the theorem. From the definition of the DG-norm we have

$$\begin{aligned}
\int_{I_m} ||\mathcal{U}_s||^2_{\mathrm{DG},t} \,\mathrm{d}t = \int_{I_m} \sum_{K \in \mathcal{T}_{h,t}} |\mathcal{U}_s|^2_{H^1(K)} \,\mathrm{d}t + \int_{I_m} \left(\sum_{\Gamma \in \mathcal{F}^I_{h,t}} \frac{c_W}{h(\Gamma)} \int_\Gamma [\mathcal{U}_s]^2 \,\mathrm{d}S \right) \mathrm{d}t \\
+ \int_{I_m} \left(\sum_{\Gamma \in \mathcal{F}^B_{h,t}} \frac{c_W}{h(\Gamma)} \int_\Gamma |\mathcal{U}_s|^2 \,\mathrm{d}S \right) \mathrm{d}t,
\end{aligned} \tag{4.34}$$

where $\mathcal{F}^I_{h,t} = \{\mathcal{A}^{m-1}_{h,t}(\hat{\Gamma});\, \hat{\Gamma} \in \mathcal{F}^I_{h,t_{m-1}}\}$ and similarly $\mathcal{F}^B_{h,t} = \{\mathcal{A}^{m-1}_{h,t}(\hat{\Gamma});\, \hat{\Gamma} \in \mathcal{F}^B_{h,t_{m-1}}\}$.

Further, we estimate each term on the right-hand side of (4.34). From [20], relation (6.161), it follows that

$$\sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}}\int_{I_m}|\tilde{\mathcal{U}}_s(t)|^2_{H^1(\hat{K})}\,\mathrm{d}t\le\tilde{c}^{(2)}_{CH}\sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}}\int_{I_m}|\tilde{U}(t)|^2_{H^1(\hat{K})}\,\mathrm{d}t, \tag{4.35}$$

with a constant $\tilde{c}^{(2)}_{CH}>0$ depending on q only. For simplicity let us denote

$$B_t=B_t(X)=\frac{\mathrm{d}\mathcal{A}^{m-1}_{h,t}(X)}{\mathrm{d}X},\quad B_t^{-1}=B_t^{-1}(x)=\frac{\mathrm{d}(\mathcal{A}^{m-1}_{h,t})^{-1}(x)}{\mathrm{d}x}.$$

Then it follows from (4.5) and (4.6) that $\|B_t\|\le C_A^+$ and $\|B_t^{-1}\|\le C_A^-$.
Now, for $K\in\mathcal{T}_{h,t}$, $K=\mathcal{A}_t(\hat{K})$ with $\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}$, using that $\|B_t|_{\hat{K}}\|$ and $\|B_t^{-1}|_{\hat{K}}\|$ are constant, we have

$$\begin{aligned}|\mathcal{U}_s(t)|^2_{H^1(K)}&=\int_K|\nabla\mathcal{U}_s(x,t)|^2\,\mathrm{d}x=\int_K\left|\nabla\tilde{\mathcal{U}}_s(\mathcal{A}_t^{-1}(x),t)\right|^2\mathrm{d}x\\&\le\int_{\hat{K}}\left|B_t^{-1}|_K\nabla\tilde{\mathcal{U}}_s(X,t)\right|^2J(X,t)\,\mathrm{d}X\le(C_A^-)^2C_J^+\,|\tilde{\mathcal{U}}_s(t)|^2_{H^1(\hat{K})}.\end{aligned} \tag{4.36}$$

The summation over all $K\in\mathcal{T}_{h,t}$, integration over I_m, the use of (4.35), (4.4), the Fubini and the substitution theorem imply that

$$\begin{aligned}\int_{I_m}\sum_{K\in\mathcal{T}_{h,t}}|\mathcal{U}_s(t)|^2_{H^1(K)}\,\mathrm{d}t&\le(C_A^-)^2C_J^+\int_{I_m}\sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}}|\tilde{\mathcal{U}}_s(t)|^2_{H^1(\hat{K})}\,\mathrm{d}t\\&\le(C_A^-)^2C_J^+\tilde{c}^{(2)}_{CH}\int_{I_m}\left(\sum_{K\in\mathcal{T}_{h,t}}\int_K|\nabla U(t)|^2\|B_t\|^2J_K^{-1}\,\mathrm{d}x\right)\mathrm{d}t\\&\le c_1\int_{I_m}\sum_{K\in\mathcal{T}_{h,t}}|U(t)|^2_{H^1(K)}\,\mathrm{d}t\\&=c_1\int_{I_m}|U(t)|^2_{H^1(\Omega_t,\mathcal{T}_{h,t})}\,\mathrm{d}t,\end{aligned} \tag{4.37}$$

where $c_1:=(C_A^-)^2C_J^+(C_J^-)^{-1}\tilde{c}^{(2)}_{CH}(C_A^+)^2$.

Now we turn our attention to the term

$$\int_{I_m}\left(\sum_{\Gamma\in\mathcal{F}^I_{h,t}}\frac{c_W}{h(\Gamma)}\int_\Gamma[\mathcal{U}_s]^2\,\mathrm{d}S\right)\mathrm{d}t.$$

For simplicity we assume that $d=2$. In Appendix A we briefly describe the proof for $d=3$. We use estimate (6.162) from [20], which implies that

$$\int_{I_m}\left(\sum_{\hat{\Gamma}\in\mathcal{F}^I_{h,t_{m-1}}}\frac{c_W}{h(\hat{\Gamma})}\int_{\hat{\Gamma}}[\tilde{\mathcal{U}}_s]^2\,\mathrm{d}S^{\hat{\Gamma}}\right)\mathrm{d}t\le c_2\int_{I_m}\left(\sum_{\hat{\Gamma}\in\mathcal{F}^I_{h,t_{m-1}}}\frac{c_W}{h(\hat{\Gamma})}\int_{\hat{\Gamma}}[\tilde{U}]^2\,\mathrm{d}S^{\hat{\Gamma}}\right)\mathrm{d}t. \tag{4.38}$$

(Here $\mathrm{d}S^{\hat{\Gamma}}$ denotes the element of the arc $\hat{\Gamma}$. Similarly we use the notation $\mathrm{d}S^{\Gamma}$.)

Now we consider the relation $\Gamma=\mathcal{A}_t(\hat{\Gamma})$, $\hat{\Gamma}\in\mathcal{F}^I_{h,t_{m-1}}$, and introduce a parametrization of $\hat{\Gamma}$:

$$\hat{\Gamma}=\mathcal{B}^{\hat{\Gamma}}_{m-1}([0,1])=\{X=\mathcal{B}^{\hat{\Gamma}}_{m-1}(v);v\in[0,1]\}.$$

Then an element of $\hat{\Gamma}$ can be expressed as

$$\mathrm{d}S^{\hat{\Gamma}} = |(\mathcal{B}_{m-1}^{\hat{\Gamma}})'(\upsilon)|\,\mathrm{d}\upsilon, \quad \upsilon \in [0,1].$$

These relations imply that

$$\Gamma = \{x = \mathcal{A}_t(\mathcal{B}_{m-1}^{\hat{\Gamma}}(\upsilon)); \upsilon \in [0,1]\}$$
$$\mathrm{d}S^{\Gamma} = \left|\frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}(\mathcal{B}_{m-1}^{\hat{\Gamma}}(\upsilon))(\mathcal{B}_{m-1}^{\hat{\Gamma}})'(\upsilon)\right| \mathrm{d}\upsilon, \quad \upsilon \in [0,1].$$

The term $(\mathcal{B}_{m-1}^{\hat{\Gamma}})'(\upsilon)$ is a tangent vector to $\hat{\Gamma}$ at the point $\mathcal{B}_{m-1}^{\hat{\Gamma}}(\upsilon)$. It follows from the properties of the mapping $\mathcal{A}_t$ that the values of

$$\frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}(\mathcal{B}_{m-1}^{\hat{\Gamma}}(\upsilon))(\mathcal{B}_{m-1}^{\hat{\Gamma}})'(\upsilon)$$

are identical from the sides of both elements $K_{\hat{\Gamma}}^{(L)}$ and $K_{\hat{\Gamma}}^{(R)}$ adjacent to $\hat{\Gamma}$. Then we can use the above relations, inequalities (4.31), (4.5), and write

$$\begin{aligned}\int_\Gamma \frac{1}{h(\Gamma)}[\mathcal{U}_s]^2\mathrm{d}S^\Gamma &= \int_0^1 \frac{1}{h(\Gamma)}[\mathcal{U}_s(\mathcal{A}_t(\mathcal{B}_{m-1}^{\hat{\Gamma}}(\upsilon)))]^2 \left|\frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}(\mathcal{B}_{m-1}^{\hat{\Gamma}}(\upsilon))(\mathcal{B}_{m-1}^{\hat{\Gamma}})'(\upsilon)\right| \mathrm{d}\upsilon \\ &\le \int_0^1 \frac{1}{h(\Gamma)}[\tilde{\mathcal{U}}_s(\mathcal{B}_{m-1}^{\hat{\Gamma}}(\upsilon))]^2 \underbrace{\left\|\frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}(\mathcal{B}_{m-1}^{\hat{\Gamma}}(\upsilon))\right\|}_{\le C_A^+} \left|(\mathcal{B}_{m-1}^{\hat{\Gamma}})'(\upsilon)\right| \mathrm{d}\upsilon \\ &\le C_A^+ \int_{\hat{\Gamma}} \frac{C_{L4}^{**}}{h(\hat{\Gamma})}[\tilde{\mathcal{U}}_s]^2\mathrm{d}S^{\hat{\Gamma}}.\end{aligned} \tag{4.39}$$

From (4.38) and (4.39) we get

$$\int_{I_m}\left(\sum_{\Gamma\in\mathcal{F}_{h,t}^I}\frac{c_W}{h(\Gamma)}\int_\Gamma [\mathcal{U}_s]^2\,\mathrm{d}S^\Gamma\right)\mathrm{d}t \le c_2 C_A^+ C_{L4}^{**}\int_{I_m}\left(\sum_{\hat{\Gamma}\in\mathcal{F}_{h,t_{m-1}}^I}\frac{c_W}{h(\hat{\Gamma})}\int_{\hat{\Gamma}}[\tilde{U}]^2\,\mathrm{d}S^{\hat{\Gamma}}\right)\mathrm{d}t. \tag{4.40}$$

Further, for $\Gamma = \mathcal{A}_t(\hat{\Gamma})$, where $\hat{\Gamma} \in \mathcal{F}_{h,t_{m-1}}^I$, we consider the parametrization

$$\Gamma = \{x = \mathcal{B}_t^\Gamma(\upsilon); \upsilon \in [0,1]\},$$
$$\hat{\Gamma} = \{X = \mathcal{A}_t^{-1}(\mathcal{B}_t^\Gamma(\upsilon)); \upsilon \in [0,1]\},$$
$$\mathrm{d}S^{\hat{\Gamma}} = \left|\frac{\mathrm{d}\mathcal{A}_t^{-1}}{\mathrm{d}x}(\mathcal{B}_t^\Gamma(\upsilon))(\mathcal{B}_t^\Gamma)'(\upsilon)\right| \mathrm{d}\upsilon.$$

Then, by (4.6),

$$\begin{aligned}\int_{\hat{\Gamma}}[\tilde{U}]^2\,\mathrm{d}S^{\hat{\Gamma}} &= \int_0^1 \underbrace{[\tilde{U}(\mathcal{A}_t^{-1}(\mathcal{B}_t^\Gamma(\upsilon)))]^2}_{[U(\mathcal{B}_t^\Gamma(\upsilon))]^2} \left|\frac{\mathrm{d}\mathcal{A}_t^{-1}}{\mathrm{d}x}(\mathcal{B}_t^\Gamma(\upsilon))(\mathcal{B}_t^\Gamma)'(\upsilon)\right| \mathrm{d}\upsilon \\ &\le \int_0^1 [U(\mathcal{B}_t^\Gamma(\upsilon))]^2 \underbrace{\left\|\frac{\mathrm{d}\mathcal{A}_t^{-1}}{\mathrm{d}x}(\mathcal{B}_t^\Gamma(\upsilon))\right\|}_{\le C_A^-} \left|(\mathcal{B}_t^\Gamma)'(\upsilon)\right| \mathrm{d}\upsilon \\ &\le C_A^- \int_0^1 [U(\mathcal{B}_t^\Gamma(\upsilon))]^2|(\mathcal{B}_t^\Gamma)'(\upsilon)|\,\mathrm{d}\upsilon \\ &= C_A^- \int_\Gamma [U]^2\,\mathrm{d}S^\Gamma.\end{aligned}$$

Substituting back to (4.40) and using (4.31), we find that

$$\int_{I_m}\left(\sum_{\Gamma\in\mathcal{F}_{h,t}^I}\frac{c_W}{h(\Gamma)}\int_\Gamma[\mathcal{U}_s]^2\,\mathrm{d}S^\Gamma\right)\mathrm{d}t\le c_3\int_{I_m}\left(\sum_{\Gamma\in\mathcal{F}_{h,t}^I}\frac{c_W}{h(\Gamma)}\int_\Gamma[U]^2\,\mathrm{d}S\right)\mathrm{d}t, \tag{4.41}$$

where $c_3=c_2C_A^+C_{L4}^{**}(C_{L4}^*)^{-1}C_A^-$.

Similarly we can prove the inequality

$$\int_{I_m}\left(\sum_{\Gamma\in\mathcal{F}_{h,t}^B}\frac{c_W}{h(\Gamma)}\int_\Gamma|\mathcal{U}_s|^2\,\mathrm{d}S^\Gamma\right)\mathrm{d}t\le c_4\int_{I_m}\left(\sum_{\Gamma\in\mathcal{F}_{h,t}^B}\frac{c_W}{h(\Gamma)}\int_\Gamma|U|^2\,\mathrm{d}S\right)\mathrm{d}t. \tag{4.42}$$

Finally, (4.37), (4.41) and (4.42) imply (4.33) with $C_{T1}^{**}=\max\{c_1,c_3,c_4\}$. □

4.4. Proof of the unconditional stability

Theorem 4.2. *There exists a constant $C_{T2}>0$ such that*

$$\begin{aligned}\|U_m^-\|_{\Omega_{t_m}}^2-\|U_{m-1}^-\|_{\Omega_{t_{m-1}}}^2+\|\{U\}_{m-1}\|_{\Omega_{t_{m-1}}}^2+\frac{\beta_0}{2}\int_{I_m}\|U\|_{\mathrm{DG},t}^2\mathrm{d}t\\ \le C_{T2}\left(\int_{I_m}\|g\|_{\Omega_t}^2\mathrm{d}t+\int_{I_m}\|u_D\|_{\mathrm{DGB},t}^2\mathrm{d}t+\int_{I_m}\|U\|_{\Omega_t}^2\mathrm{d}t\right).\end{aligned} \tag{4.43}$$

Proof. From (4.12), by virtue of (4.20), (4.16), (4.17), (4.18), (4.21) and (4.19), after some manipulation we get

$$\begin{aligned}&\|U_m^-\|_{\Omega_{t_m}}^2-\|U_{m-1}^-\|_{\Omega_{t_{m-1}}}^2+\|\{U\}_{m-1}\|_{\Omega_{t_{m-1}}}^2+\beta_0\left(1-\frac{1}{k_1}-\frac{1}{k_2}-\frac{1}{k_3}\right)\int_{I_m}\|U\|_{\mathrm{DG},t}^2\mathrm{d}t\\ &\quad\le\int_{I_m}\|g\|_{\Omega_t}^2\mathrm{d}t+\beta_0(1+k_3)\int_{I_m}\|u_D\|_{\mathrm{DGB},t}^2\mathrm{d}t+\left(c_z+1+\frac{c_d}{\beta_0}+2c_b\right)\int_{I_m}\|U\|_{\Omega_t}^2\mathrm{d}t.\end{aligned}$$

Hence, choosing $k_1=k_2=k_3=6$, we get (4.43) with $C_{T2}=\max\{1,7\beta_0,c_z+1+c_d/\beta_0+2c_b\}$. □

Theorem 4.3. *There exist constants $C_{T3}^*,C_{T3}^{**}>0$ such that for any $\delta_1>0$ we have*

$$\begin{aligned}&\|U_m^-\|_{\Omega_{t_m}}^2+\|U_{m-1}^+\|_{\Omega_{t_{m-1}}}^2+\frac{\beta_0}{2}\int_{I_m}\|U\|_{\mathrm{DG},t}^2\mathrm{d}t\\ &\quad\le C_{T3}^*\int_{I_m}\|U\|_{\Omega_t}^2\mathrm{d}t+C_{T3}^{**}\int_{I_m}\big(\|g\|_{\Omega_t}^2+\|u_D\|_{\mathrm{DGB},t}^2\big)\mathrm{d}t+\frac{2}{\delta_1}\|U_{m-1}^-\|_{\Omega_{t_{m-1}}}^2+4\delta_1\|U_{m-1}^+\|_{\Omega_{t_{m-1}}}^2.\end{aligned} \tag{4.44}$$

Proof. From (3.19), by virtue of (4.22), (4.16), (4.17), (4.18), (4.21) and (4.19), we get

$$\begin{aligned}&\|U_m^-\|_{\Omega_{t_m}}^2+\|U_{m-1}^+\|_{\Omega_{t_{m-1}}}^2+\beta_0\left(1-\frac{1}{k_1}-\frac{1}{k_2}-\frac{1}{k_3}\right)\int_{I_m}\|U\|_{\mathrm{DG},t}^2\mathrm{d}t\\ &\quad\le\int_{I_m}\|g\|_{\Omega_t}^2\mathrm{d}t+\beta_0(1+k_3)\int_{I_m}\|u_D\|_{\mathrm{DGB},t}^2\mathrm{d}t\\ &\qquad+\left(1+c_z+2c_b+\frac{c_d}{\beta_0}\right)\int_{I_m}\|U\|_{\Omega_t}^2\mathrm{d}t+2\big(U_{m-1}^-,U_{m-1}^+\big)_{\Omega_{t_{m-1}}}.\end{aligned}$$

Using Young's inequality for the term $2(U_{m-1}^-,U_{m-1}^+)$ and setting $k_1=k_2=k_3=6$, we get (4.44), where $C_{T3}^*=1+c_z+2c_b+c_d/\beta_0$ and $C_{T3}^{**}=\max\{1,7\beta_0\}$. □

We introduce the following notation:

$$t_{m-1+l/q} = t_{m-1} + \tau_m \frac{l}{q},$$
$$U_{m-1+l/q} = U(t_{m-1+l/q}), \quad l = 0, \dots, q.$$

Lemma 4.5. *There exist constants $C^*_{L5}, C^{**}_{L5} > 0$ such that for $m = 1, \dots, M$ we have*

$$\sum_{l=0}^{q} \|U_{m-1+l/q}\|^2_{\Omega_{t_{m-1+l/q}}} \geq \frac{C^*_{L5}}{\tau_m} \int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t, \tag{4.45}$$

$$\|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}} \leq \frac{C^{**}_{L5}}{\tau_m} \int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t. \tag{4.46}$$

Proof. Using the equivalence of norms in the space of polynomials of degree $\leq q$, for $p(t) = \tilde{U}(X,t)$, $t \in I_m$, and any fixed $X \in \Omega_{t_{m-1}}$, we have

$$\sum_{l=0}^{q} \tilde{U}^2\left(X, t_{m-1+l/q}\right) \geq \frac{L_q}{\tau_m} \int_{I_m} \tilde{U}^2(X,t)\,\mathrm{d}t,$$
$$\tilde{U}^2\big(X, t^+_{m-1}\big) \leq \frac{M_q}{\tau_m} \int_{I_m} \tilde{U}^2(X,t)\,\mathrm{d}t,$$

where the constants L_q, $M_q > 0$ were introduced in [20], Section 6.2.3.2. Integrating over $\Omega_{t_{m-1}}$ and using Fubini's theorem, we get

$$\sum_{l=0}^{q} \int_{\Omega_{t_{m-1}}} |\tilde{U}\left(X, t_{m-1+l/q}\right)|^2 \mathrm{d}X \geq \frac{L_q}{\tau_m} \int_{\Omega_{t_{m-1}}} \left(\int_{I_m} |\tilde{U}(X,t)|^2 \mathrm{d}t \right) \mathrm{d}X$$
$$= \frac{L_q}{\tau_m} \int_{I_m} \left(\int_{\Omega_{t_{m-1}}} |\tilde{U}(X,t)|^2 \mathrm{d}X \right) \mathrm{d}t.$$

Analogously we find that

$$\int_{\Omega_{t_{m-1}}} |\tilde{U}\big(X, t^+_{m-1}\big)|^2 \,\mathrm{d}X \leq \frac{M_q}{\tau_m} \int_{I_m} \left(\int_{\Omega_{t_{m-1}}} |\tilde{U}(X,t)|^2 \mathrm{d}X \right) \mathrm{d}t.$$

Now the substitution $X = \mathcal{A}_t^{-1}(x)$, where $X \in \Omega_{t_{m-1}}$, $x \in \Omega_t$, relation $\tilde{U}(\mathcal{A}_t^{-1}(x), t) = U(x,t)$ and (4.4) imply that

$$\sum_{l=0}^{q} \|U_{m-1+l/q}\|^2_{\Omega_{t_{m-1+l/q}}}$$
$$\geq C_J^- \sum_{l=0}^{q} \int_{\Omega_{t_{m-1+l/q}}} |U(x, t_{m-1+l/q})|^2 J^{-1}(x, t_{m-1+l/q})\,\mathrm{d}x$$
$$\geq \frac{L_q}{\tau_m} C_J^- \int_{I_m} \left(\int_{\Omega_{t_{m-1}}} |\tilde{U}(X,t)|^2 \mathrm{d}X \right) \mathrm{d}t$$
$$= \frac{L_q}{\tau_m} C_J^- \int_{I_m} \left(\int_{\Omega_t} |\tilde{U}(\mathcal{A}_t^{-1}(x), t)|^2 J^{-1}(x,t)\,\mathrm{d}x \right) \mathrm{d}t$$
$$\geq \frac{L_q}{\tau_m} (C_J^+)^{-1} C_J^- \int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t.$$

Hence, we get (4.45) with $C^*_{L5} = L_q(C^+_J)^{-1}C^-_J$.

Further, since $x = \mathcal{A}_{t_{m-1}}(X) = X$ and, thus, $\tilde{U}(X, t^+_{m-1}) = U(x, t^+_{m-1})$, using the substitution theorem and (4.4), we obtain

$$\begin{aligned}\|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}} &= \int_{\Omega_{t_{m-1}}} |\tilde{U}(X, t^+_{m-1})|^2 \mathrm{d}X \\ &\le \frac{M_q}{\tau_m}\int_{I_m}\left(\int_{\Omega_{t_{m-1}}} |\tilde{U}(X,t)|^2 \mathrm{d}X\right)\mathrm{d}t \\ &\le \frac{C^{**}_{L5}}{\tau_m}\int_{I_m}\|U\|^2_{\Omega_t}\mathrm{d}t,\end{aligned}$$

where $C^{**}_{L5} = M_q(C^-_J)^{-1}$. □

In what follows, because of simplicity, we use the notation $\tilde{U}' = \frac{\partial \tilde{U}}{\partial t}$ and do not write the arguments X and t in integrals.

Lemma 4.6. *There exists a constant $C_{L6} > 0$ such that*

$$\begin{aligned}&\int_{I_m}(D_t U, \mathcal{U}_s)_{\Omega_t}\mathrm{d}t + (\{U\}_{m-1}, \mathcal{U}_s(t^+_{m-1}))_{\Omega_{t_{m-1}}} \\ &\qquad \ge \frac{1}{2}\left(\|U(s-)\|^2_{\Omega_s} + \|U(t^+_{m-1})\|^2_{\Omega_{t_{m-1}}}\right) - C_{L6}\int_{I_m}\|U\|^2_{\Omega_t}\mathrm{d}t - (U^+_{m-1}, U^-_{m-1})_{\Omega_{t_{m-1}}}.\end{aligned} \tag{4.47}$$

for any $s \in I_m$, $m = 1, \dots, M$ and $h \in (0, \overline{h})$.

Proof. By virtue of the definition of the ALE derivative (2.9), the definitions of $\tilde{U}, \tilde{\mathcal{U}}_s, \mathcal{U}_s$, the fact that $\tilde{U}'$ is a polynomial of degree $\le q-1$ in time and the substitution theorem we can write

$$\begin{aligned}\int_{I_m}(D_t U, \mathcal{U}_s)_{\Omega_t}\mathrm{d}t &= \int_{I_m}\left(\tilde{U}', \tilde{\mathcal{U}}_s J\right)_{\Omega_{t_{m-1}}}\mathrm{d}t \\ &= \int_{I_m}\left(\tilde{U}', \tilde{\mathcal{U}}_s\right)_{\Omega_{t_{m-1}}}\mathrm{d}t + \int_{I_m}\left(\tilde{U}', \tilde{\mathcal{U}}_s(J-1)\right)_{\Omega_{t_{m-1}}}\mathrm{d}t \\ &= \int_{t_{m-1}}^{s}\left(\tilde{U}', \tilde{U}\right)_{\Omega_{t_{m-1}}}\mathrm{d}t + \int_{I_m}\left(\tilde{U}', \tilde{\mathcal{U}}_s(J-1)\right)_{\Omega_{t_{m-1}}}\mathrm{d}t \\ &= \int_{t_{m-1}}^{s}\left(\tilde{U}', \tilde{U}J\right)_{\Omega_{t_{m-1}}}\mathrm{d}t + \int_{t_{m-1}}^{s}\left(\tilde{U}', \tilde{U}(1-J)\right)_{\Omega_{t_{m-1}}}\mathrm{d}t + \int_{I_m}\left(\tilde{U}', \tilde{\mathcal{U}}_s(J-1)\right)_{\Omega_{t_{m-1}}}\mathrm{d}t \\ &= \int_{t_{m-1}}^{s}(D_t U, U)_{\Omega_t}\mathrm{d}t + \int_{t_{m-1}}^{s}\left(\tilde{U}', \tilde{U}(1-J)\right)_{\Omega_{t_{m-1}}}\mathrm{d}t + \int_{I_m}\left(\tilde{U}', \tilde{\mathcal{U}}_s(J-1)\right)_{\Omega_{t_{m-1}}}\mathrm{d}t.\end{aligned} \tag{4.48}$$

Now we estimate the second and third term on the right-hand side. We begin with the third term. The fact that J is constant on each $\hat{K} \in \hat{\mathcal{T}}_{h,t_{m-1}}$ and the substitution theorem imply that

$$\begin{aligned}\left|\int_{I_m}\left(\tilde{U}', \tilde{\mathcal{U}}_s(J-1)\right)_{\Omega_{t_{m-1}}}\mathrm{d}t\right| &= \left|\sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}}\int_{I_m}(J_{\hat{K}}-1)\left(\int_{\hat{K}}\tilde{U}'\tilde{\mathcal{U}}_s\,\mathrm{d}X\right)\mathrm{d}t\right| \\ &\le \sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}}\max_{t\in I_m}|J_{\hat{K}}-1|\int_{I_m}\left(\int_{\hat{K}}|\tilde{U}'\tilde{\mathcal{U}}_s|\,\mathrm{d}X\right)\mathrm{d}t.\end{aligned}$$

Using the relation $J_{\hat{K}}(t_{m-1}) = 1$, we have

$$\max_{t\in I_m} |J_{\hat{K}} - 1| \le \int_{t_{m-1}}^{t_m} |J'_{\hat{K}}| \,\mathrm{d}t \le c_J \tau_m,$$

where $c_J > 0$ is a constant independent of h, τ_m, m. Then we find that

$$\begin{aligned}
&\sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}} \max_{t\in I_m} |J_{\hat{K}} - 1| \int_{I_m}\int_{\hat{K}} |\tilde{U}'\tilde{\mathcal{U}}_s| \,\mathrm{d}X\,\mathrm{d}t \\
&\qquad \le c_J\tau_m \sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}} \int_{\hat{K}} \left(\left(\int_{I_m} |\tilde{U}'|^2 \,\mathrm{d}t \right)^{1/2} \left(\int_{I_m} |\tilde{\mathcal{U}}_s|^2 \,\mathrm{d}t \right)^{1/2} \right) \mathrm{d}X.
\end{aligned}$$

Now we apply the inverse inequality in time: There exists a constant $\hat{c}_I$ such that

$$\left(\int_{I_m} |\tilde{U}'(X,t)|^2 \,\mathrm{d}t \right)^{1/2} \le \frac{\hat{c}_I}{\tau_m} \left(\int_{I_m} |\tilde{U}(X,t)|^2 \,\mathrm{d}t \right)^{1/2} \tag{4.49}$$

holds for every $X \in \Omega_{t_{m-1}}$, $\tau_m \in (0, \overline{\tau})$ and $m = 1, \dots, M$.

This inequality, Young's inequality, Fubini's theorem, (4.30), substitution theorem and (4.4) imply that

$$\begin{aligned}
&\tau_m \sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}} \int_{\hat{K}} \left(\left(\int_{I_m} |\tilde{U}'|^2 \,\mathrm{d}t \right)^{1/2} \left(\int_{I_m} |\tilde{\mathcal{U}}_s|^2 \,\mathrm{d}t \right)^{1/2} \right) \mathrm{d}X \\
&\qquad \le \hat{c}_I \sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}} \int_{\hat{K}} \left(\int_{I_m} |\tilde{U}|^2 \mathrm{d}t \right)^{1/2} \left(\int_{I_m} |\tilde{\mathcal{U}}_s|^2 \mathrm{d}t \right)^{1/2} \mathrm{d}X \\
&\qquad \le \frac{\hat{c}_I}{2} \sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}} \int_{\hat{K}} \left(\int_{I_m} \left(|\tilde{U}|^2 + |\tilde{\mathcal{U}}_s|^2 \right) \mathrm{d}t \right) \mathrm{d}X \\
&\qquad = \frac{\hat{c}_I}{2} \sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}} \int_{I_m} \left(\int_{\hat{K}} \left(|\tilde{U}|^2 + |\tilde{\mathcal{U}}_s|^2 \right) \mathrm{d}X \right) \mathrm{d}t \\
&\qquad \le \frac{\hat{c}_I}{2} (1 + \tilde{c}_{CH}^{(1)}) \int_{I_m} \|\tilde{U}\|^2_{\Omega_{t_{m-1}}} \mathrm{d}t \\
&\qquad \le c^* \int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t,
\end{aligned}$$

where $c^* = (C_J^-)^{-1} \hat{c}_I (1 + \tilde{c}_{CH}^{(1)})/2$. Summarizing the obtained results, we see that we have proved the inequality

$$\left| \int_{I_m} \left(\tilde{U}', \tilde{\mathcal{U}}_s (J-1) \right)_{\Omega_{t_{m-1}}} \mathrm{d}t \right| \le c^* c_J \int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t. \tag{4.50}$$

Similarly as above we can estimate the second term on the right-hand side of (4.48):

$$\begin{aligned}\left|\int_{t_{m-1}}^{s}\left(\tilde{U}',\tilde{U}(1-J)\right)_{\Omega_{t_{m-1}}}\mathrm{d}t\right| &\le \int_{I_m}\left|(\tilde{U}',\tilde{U}(1-J))_{\Omega_{t_{m-1}}}\right|\mathrm{d}t\\ &\le \sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}}\max_{t\in I_m}|1-J_{\hat{K}}|\int_{I_m}\int_{\hat{K}}|\tilde{U}'\tilde{U}|\,\mathrm{d}X\mathrm{d}t\\ &\le c_J\tau_m\sum_{\hat{K}\in\hat{\mathcal{T}}_{h,t_{m-1}}}\int_{\hat{K}}\left(\left(\int_{I_m}|\tilde{U}'|^2\,\mathrm{d}t\right)^{1/2}\left(\int_{I_m}|\tilde{U}|^2\,\mathrm{d}t\right)^{1/2}\right)\mathrm{d}X.\end{aligned}$$

Now the inverse inequality in time, Young's inequality, Fubini's theorem, (4.30) and (4.4) yield the inequality

$$\left|\int_{t_{m-1}}^{s}\left(\tilde{U}',\tilde{U}(1-J)\right)_{\Omega_{t_{m-1}}}\mathrm{d}t\right| \le c_1\int_{I_m}\|U\|^2_{\Omega_t}\mathrm{d}t. \tag{4.51}$$

with $c_1 = c_J(C_J^-)^{-1}\hat{c}_I/2$.

Finally, from (4.48), (4.50), (4.51) and analogy to (4.22), (4.29) putting $c_2 = c^*c_J + c_1$ we find that

$$\begin{aligned}&\int_{I_m}(D_tU,\mathcal{U}_s)_{\Omega_t}\mathrm{d}t + (\{U\}_{m-1},\mathcal{U}_s(t_{m-1}+))_{\Omega_{t_{m-1}}}\\ &\quad\ge \int_{t_{m-1}}^{s}(D_tU,U)_{\Omega_t}\mathrm{d}t + \|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}} - (U^-_{m-1},U^+_{m-1})_{\Omega_{t_{m-1}}} - c_2\int_{I_m}\|U\|^2_{\Omega_t}\mathrm{d}t\\ &\quad= \frac{1}{2}\int_{t_{m-1}}^{s}\left(\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega_t}U^2(x,t)\mathrm{d}x\right)\mathrm{d}t - \frac{1}{2}\int_{t_{m-1}}^{s}\left(U^2\mathrm{div},\boldsymbol{z}\right)_{\Omega_t}\mathrm{d}t\\ &\qquad+ \|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}} - (U^-_{m-1},U^+_{m-1})_{\Omega_{t_{m-1}}} - c_2\int_{I_m}\|U\|^2_{\Omega_t}\mathrm{d}t\\ &\quad= \frac{1}{2}\left(\|U(s-)\|^2_{\Omega_s} + \|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}}\right) - \frac{c_z}{2}\int_{t_{m-1}}^{s}\|U\|^2_{\Omega_t}\mathrm{d}t\\ &\qquad- c_2\int_{I_m}\|U\|^2_{\Omega_t}\mathrm{d}t - (U^-_{m-1},U^+_{m-1})_{\Omega_{t_{m-1}}},\end{aligned}$$

which implies (4.47) with $C_{L6} = c_z/2 + c_2$. □

In the following lemmas, for simplicity we use the notation $\mathcal{U}_l^*$ and $\tilde{\mathcal{U}}_l^*$ for the discrete characteristic functions to U and $\tilde{U}$, respectively at the time instant $t_{m-1+l/q}$.

Lemma 4.7. *There exists a constant $C_{L7} > 0$ such that*

$$|a_h(U,\mathcal{U}_l^*,t) + \beta_0 J_h(U,\mathcal{U}_l^*,t)| \le C_{L7}\left(\|U\|^2_{\mathrm{DG},t} + \|\mathcal{U}_l^*\|^2_{\mathrm{DG},t} + \|u_D\|^2_{\mathrm{DGB},t}\right) \tag{4.52}$$

for all $t, l \in I_m$, $m = 1,\dots,M$, $h \in (0,\overline{h})$.

Proof. Using the definition of the form a_h, the property of the function β, the Cauchy inequality and Young's inequality, we get

$$
\begin{aligned}
|a_h(U,\mathcal{U}_l^*,t)| \le \beta_1 &\sum_{K\in\mathcal{T}_{h,t}} \int_K \left(|\nabla U|^2 + |\nabla \mathcal{U}_l^*|^2\right) \mathrm{d}x \\
&+ \beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^I} \int_\Gamma \left(\frac{h(\Gamma)}{c_W} \left(|\nabla U_\Gamma^{(L)}|^2 + |\nabla U_\Gamma^{(R)}|^2\right) + \frac{c_W}{h(\Gamma)} [\mathcal{U}_l^*]^2 \right) \mathrm{d}S \\
&+ \beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^I} \int_\Gamma \left(\frac{h(\Gamma)}{c_W} \left(|\nabla (\mathcal{U}_l^*)_\Gamma^{(L)}|^2 + |\nabla (\mathcal{U}_l^*)_\Gamma^{(R)}|^2\right) + \frac{c_W}{h(\Gamma)} [U]^2 \right) \mathrm{d}S \\
&+ \beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^B} \int_\Gamma \left(\frac{h(\Gamma)}{c_W} |\nabla U|^2 + \frac{c_W}{h(\Gamma)} |\mathcal{U}_l^*|^2 \right) \mathrm{d}S \\
&+ \beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^B} \int_\Gamma \left(\frac{h(\Gamma)}{c_W} |\nabla \mathcal{U}_l^*|^2 + \frac{c_W}{h(\Gamma)} |U|^2 \right) \mathrm{d}S \\
&+ \beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^B} \int_\Gamma |\nabla \mathcal{U}_l^*|\, |u_D| \,\mathrm{d}S.
\end{aligned} \tag{4.53}
$$

The last term can be estimated using Young's inequality and the relation $h(\Gamma) \le h_{K_\Gamma^{(L)}}$, for each $\varepsilon > 0$ the last term can be estimated in the following way:

$$
\beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^B} \int_\Gamma |\nabla \mathcal{U}_l^*|\, |u_D| \,\mathrm{d}S \le \frac{\beta_1 \varepsilon}{2c_W} J_h^B(u_D,u_D) + \frac{\beta_1}{2\varepsilon} \sum_{\Gamma\in\mathcal{F}_{h,t}^B} \int_{\partial K_\Gamma^{(L)}} h_{K_\Gamma^{(L)}} |\nabla \mathcal{U}_l^*|^2 \,\mathrm{d}S.
$$

Now we express the first term on the right-hand side of this inequality with the aid of the definition of the $\|\cdot\|_{\mathrm{DGB},t}$-norm and to the second term we apply the multiplicative trace inequality (4.8) and the inverse inequality (4.9). We get

$$
\beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^B} \int_\Gamma |\nabla \mathcal{U}_l^*|\, |u_D| \,\mathrm{d}S \le \frac{\beta_1 \varepsilon}{2c_W} \|u_D\|_{\mathrm{DGB},t}^2 + \frac{\beta_1}{2\varepsilon} c_M (c_I+1) \|\mathcal{U}_l^*\|_{\mathrm{DG},t}^2. \tag{4.54}
$$

Setting $\varepsilon := \frac{\beta_1}{\beta_0} c_M(c_I+1)$ in (4.54) and substituting back to (4.53) we get

$$
\begin{aligned}
|a_h(U,\mathcal{U}_l^*,t)| \le \beta_1 &\sum_{K\in\mathcal{T}_{h,t}} \int_K \left(|\nabla U|^2 + |\nabla \mathcal{U}_l^*|^2\right) \mathrm{d}x \\
&+ \beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^I} \int_\Gamma \frac{h(\Gamma)}{c_W} \left(|\nabla U_\Gamma^{(L)}|^2 + |\nabla U_\Gamma^{(R)}|^2\right) \mathrm{d}S + \beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^B} \int_\Gamma \frac{h(\Gamma)}{c_W} |\nabla U|^2 \,\mathrm{d}S \\
&+ \beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^I} \int_\Gamma \frac{h(\Gamma)}{c_W} \left(|\nabla (\mathcal{U}_l^*)_\Gamma^{(L)}|^2 + |\nabla (\mathcal{U}_l^*)_\Gamma^{(R)}|^2\right) \mathrm{d}S \\
&+ \beta_1 \sum_{\Gamma\in\mathcal{F}_{h,t}^B} \int_\Gamma \frac{h(\Gamma)}{c_W} |\nabla \mathcal{U}_l^*|^2 \,\mathrm{d}S + \frac{\beta_1^2}{2\beta_0 c_W} c_M (c_I+1) \|u_D\|_{\mathrm{DGB},t}^2 \\
&+ \frac{\beta_0}{2} \|\mathcal{U}_l^*\|_{\mathrm{DG},t}^2 + \beta_1\, J_h(\mathcal{U}_l^*,\mathcal{U}_l^*,t) + \beta_1\, J_h(U,U,t).
\end{aligned}
$$

Using the inequality $h(\Gamma) \le h_K$ for $\Gamma \subset \partial K$, we have

$$|a_h(U,\mathcal{U}_l^*,t)| \le \beta_1 \sum_{K\in\mathcal{T}_{h,t}} \int_K \left(|\nabla U|^2 + |\nabla \mathcal{U}_l^*|^2\right) \mathrm{d}x + \frac{\beta_1}{c_W} \sum_{K\in\mathcal{T}_{h,t}} \int_{\partial K} h_K \left(|\nabla U|^2 + |\nabla \mathcal{U}_l^*|^2\right) \mathrm{d}S \tag{4.55}$$
$$+ \frac{\beta_1^2}{2\beta_0 c_W} c_M(c_I+1)\|u_D\|^2_{\mathrm{DGB},t} + \frac{\beta_0}{2}\|\mathcal{U}_l^*\|^2_{\mathrm{DG},t}$$
$$+ \beta_1\, J_h(\mathcal{U}_l^*,\mathcal{U}_l^*,t) + \beta_1\, J_h(U,U,t).$$

Now, applying the multiplicative inequality and the inverse inequality, we can obtain the estimate

$$\sum_{K\in\mathcal{T}_{h,t}} \int_{\partial K} h_K \left(|\nabla U|^2 + |\nabla \mathcal{U}_l^*|^2\right) \mathrm{d}S \le c_M(c_I+1) \sum_{K\in\mathcal{T}_{h,t}} \left(|U|^2_{H^1(\Omega)} + |\mathcal{U}_l^*|^2_{H^1(\Omega)}\right). \tag{4.56}$$

From (4.55) and (4.56), the definition of the $\|\cdot\|_{\mathrm{DG},t}$-norm, using the inequality

$$J_{h,}(U,\mathcal{U}_l^*,t) \le J_h(U,U,t) + J_h(\mathcal{U}_l^*,\mathcal{U}_l^*,t)$$

and putting $C_{L7} = \max\{\beta_0 + \beta_1 + \beta_1 c_M(c_I+1)/c_W, \beta_1^2 c_M(c_I+1)/(2\beta_0 c_W)\}$, we finally get

$$\begin{aligned}|a_h(U,\mathcal{U}_l^*,t) + \beta_0\, J_h(U,\mathcal{U}_l^*,t)| &\le \left(\beta_1 + \frac{\beta_1}{c_W} c_M(c_I+1)\right) |U|^2_{H^1(\Omega_t,\mathcal{T}_{h,t})} \\ &\quad + (\beta_0+\beta_1) J_h(U,U,t) + \left(\beta_1 + \frac{\beta_0}{2} + \frac{\beta_1}{c_W} c_M(c_I+1)\right) |\mathcal{U}_l^*|^2_{H^1(\Omega_t,\mathcal{T}_{h,t})} \\ &\quad + (\beta_0+\beta_1) J_h(\mathcal{U}_l^*,\mathcal{U}_l^*,t) + \frac{\beta_1^2}{2\beta_0 c_W} c_M(c_I+1)\|u_D\|^2_{\mathrm{DGB},t} \\ &\le C_{L7}\left(\|U\|^2_{\mathrm{DG},t} + \|\mathcal{U}_l^*\|^2_{\mathrm{DG},t} + \|u_D\|^2_{\mathrm{DGB},t}\right).\end{aligned}$$

□

Lemma 4.8. *For each $k_1 > 0$ there exists a constant $c_b > 0$ such that for the approximate solution U and the discrete characteristic function $\mathcal{U}_l^*$ we have the inequality*

$$\int_{I_m} |b_h(U,\mathcal{U}_l^*,t)|\mathrm{d}t \le \frac{\beta_0}{2k_1} \int_{I_m} \|\mathcal{U}_l^*\|^2_{\mathrm{DG},t}\mathrm{d}t + c_b \int_{I_m} \|U\|^2_{\Omega_t}\mathrm{d}t. \tag{4.57}$$

Proof. It can be proved in a similar way as in the proof of inequality (5.18) from [7]. □

Lemma 4.9. *For each $k_2 > 0$ there exists a constant $c_d > 0$ such that the approximate solution U and the discrete characteristic function $\mathcal{U}_l^*$ satisfy the inequality*

$$\int_{I_m} |d_h(U,\mathcal{U}_l^*,t)|\,\mathrm{d}t \le \frac{\beta_0}{2k_2} \int_{I_m} \|U\|^2_{\mathrm{DG},t}\,\mathrm{d}t + \frac{c_d}{2\beta_0} \int_{I_m} \|\mathcal{U}_l^*\|^2_{\Omega_t}\,\mathrm{d}t. \tag{4.58}$$

Proof. By (3.16), (4.7) and the Cauchy and Young's inequalities,

$$\int_{I_m} |d_h(U,\mathcal{U}_l^*,t)|\,\mathrm{d}t \le \frac{\beta_0}{2k_2} \int_{I_m} \|U\|^2_{\mathrm{DG},t}\,\mathrm{d}t + \frac{c_z^2 k_2}{2\beta_0} \int_{I_m} \|\mathcal{U}_l^*\|^2_{\Omega_t}\,\mathrm{d}t,$$

which is (4.58) with $c_d = c_z^2 k_2$. □

Lemma 4.10. *For the approximate solution U, the discrete characteristic function $\mathcal{U}_l^*$ and any $k_3 > 0$ we have*

$$\int_{I_m} |l_h(\mathcal{U}_l^*,t)|\,\mathrm{d}t \le \frac{1}{2}\int_{I_m} \left(\|g\|_{\Omega_t}^2 + \|\mathcal{U}_l^*\|_{\Omega_t}^2\right)\,\mathrm{d}t + \frac{\beta_0 k_3}{2}\int_{I_m} \|u_D\|_{\mathrm{DGB},t}^2\,\mathrm{d}t + \frac{\beta_0}{2k_3}\int_{I_m} \|\mathcal{U}_l^*\|_{\mathrm{DG},t}^2\,\mathrm{d}t. \tag{4.59}$$

Proof. From (3.17), using the Cauchy and Young's inequality with $k_3 > 0$, we find that

$$\begin{aligned}
&|(g,\mathcal{U}_l^*) + \beta_0 c_W \sum_{\Gamma\in\mathcal{F}_{h,t}^B} h(\Gamma)^{-1}\int_\Gamma u_D\,\mathcal{U}_l^*\,\mathrm{d}S| \\
&\quad\le \frac{1}{2}(\|g\|_{\Omega_t}^2 + \|\mathcal{U}_l^*\|_{\Omega_t}^2) + \frac{\beta_0 k_3}{2}\underbrace{c_W \sum_{\Gamma\in\mathcal{F}_{h,t}^B} h(\Gamma)^{-1}\int_\Gamma |u_D|^2\,\mathrm{d}S}_{=\|u_D\|_{\mathrm{DGB},t}^2} \\
&\qquad + \frac{\beta_0}{2k_3}\underbrace{c_W \sum_{\Gamma\in\mathcal{F}_{h,t}^B} h(\Gamma)^{-1}\int_\Gamma |\mathcal{U}_l^*|^2\,\mathrm{d}S}_{\le J_h(\mathcal{U}_l^*,\mathcal{U}_l^*,t)\le \|\mathcal{U}_l^*\|_{\mathrm{DG},t}^2},
\end{aligned}$$

from which we get (4.59) by integrating both sides over the interval I_m. □

Now we prove an important estimate regarding the problematic term $\int_{I_m}\|U\|_{\Omega_t}^2\,\mathrm{d}t$.

Theorem 4.4. *There exist constants $C_{T4}, C_{T4}^* > 0$ such that*

$$\int_{I_m}\|U\|_{\Omega_t}^2\mathrm{d}t \le C_{T4}\tau_m\left(\|U_{m-1}^-\|_{\Omega_{t_{m-1}}}^2 + \int_{I_m}\left(\|g\|_{\Omega_t}^2 + \|u_D\|_{\mathrm{DGB},t}^2\right)\mathrm{d}t\right) \tag{4.60}$$

provided $0 < \tau_m < C_{T4}^*$.

Proof. For $q = 1$, the proof can be carried out similarly as in [5]. Let us assume that $q \ge 2$, $l \in \{1,\dots,q-1\}$. From the definition of the approximate solution (3.19) and (3.20) for $\varphi := \mathcal{U}_l^*$ we get

$$\begin{aligned}
&\int_{I_m}(D_t U,\mathcal{U}_l^*)_{\Omega_t}\,\mathrm{d}t + \left(\{U\}_{m-1},\{\mathcal{U}_l^*\}_{m-1}^+\right)_{\Omega_{t_{m-1}}} \\
&\qquad = \int_{I_m}\left(-a_h(U,\mathcal{U}_l^*,t) - \beta_0 J_h(U,\mathcal{U}_l^*,t) - b_h(U,\mathcal{U}_l^*,t)\right)\mathrm{d}t \\
&\qquad\quad + \int_{I_m}\left(-d_h(U,\mathcal{U}_l^*,t) + l_h(\mathcal{U}_l^*,t)\right)\mathrm{d}t.
\end{aligned} \tag{4.61}$$

This relation and Lemma 4.6 imply that

$$\begin{aligned}
&\frac{1}{2}\left(\left\|U_{m-1+l/q}\right\|_{\Omega_{t_{m-1+l/q}}}^2 + \|U_{m-1}^+\|_{\Omega_{t_{m-1}}}^2\right) \\
&\qquad \le \int_{I_m}|a_h(U,\mathcal{U}_l^*,t) + \beta_0 J_h(U,\mathcal{U}_l^*,t)|\,\mathrm{d}t + \int_{I_m}|b_h(U,\mathcal{U}_l^*,t)|\,\mathrm{d}t \\
&\qquad\quad + \int_{I_m}|d_h(U,\mathcal{U}_l^*,t)|\,\mathrm{d}t + \int_{I_m}|l_h(\mathcal{U}_l^*,t)|\,\mathrm{d}t + \left(U_{m-1}^-,U_{m-1}^+\right)_{\Omega_{t_{m-1}}} \\
&\qquad\quad + C_{L6}\int_{I_m}\|U\|_{\Omega_t}^2\,\mathrm{d}t \equiv \mathrm{RHS}.
\end{aligned} \tag{4.62}$$

Now we need to estimate the right-hand side of (4.62) from above. Using (4.52), (4.57), (4.58),(4.59) with $k_1 = k_2 = k_3 = 1$, (4.47) and Young's inequality with any $\delta_2 > 0$, we get

$$\begin{aligned}\text{RHS} \le c_1 \int_{I_m} \left(\|U\|^2_{\mathrm{DG},t} + \|\mathcal{U}_l^*\|^2_{\mathrm{DG},t} + \|\mathcal{U}_l^*\|^2_{\Omega_t} + \|U\|^2_{\Omega_t} + \|g\|^2_{\Omega_t} + \|u_D\|^2_{\mathrm{DGB},t}\right) \mathrm{d}t \\ + \frac{\|U^-_{m-1}\|^2_{\Omega_{t_{m-1}}}}{\delta_2} + \delta_2 \|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}},\end{aligned}$$

where $c_1 = \max\{C_{L9} + \beta_0 + c_d/(2\beta_0) + 1/2, c_b + C_{L6}\}$. Now we apply Theorem 4.1 on the continuity of the discrete characteristic function:

$$\int_{I_m} \|\mathcal{U}_l^*\|^2_{\Omega_t} \mathrm{d}t \le C^*_{T1} \int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t, \quad \int_{I_m} \|\mathcal{U}_l^*\|^2_{\mathrm{DG},t} \mathrm{d}t \le C^{**}_{T1} \int_{I_m} \|U\|^2_{\mathrm{DG},t} \mathrm{d}t.$$

Hence,

$$\begin{aligned}\text{RHS} \le c_2 \int_{I_m} \left(\|U\|^2_{\mathrm{DG},t} + \|U\|^2_{\Omega_t} + \|g\|^2_{\Omega_t} + \|u_D\|^2_{\mathrm{DGB},t}\right) \mathrm{d}t \\ + \frac{\|U^-_{m-1}\|^2_{\Omega_{t_{m-1}}}}{\delta_2} + \delta_2 \|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}},\end{aligned}$$

with $c_2 = c_1 \max\{1 + C^*_{T1}, 1 + C^{**}_{T1}\}$. Then it follows from (4.62) that

$$\begin{aligned}&\frac{1}{2}\left(\left\|U^-_{m-1+l/q}\right\|^2_{\Omega_{t_{m-1+l/q}}} + \|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}}\right) \\ &\quad \le c_2 \int_{I_m} \left(\|U\|^2_{\mathrm{DG},t} + \|U\|^2_{\Omega_t} + \|g\|^2_{\Omega_t} + \|u_D\|^2_{\mathrm{DGB},t}\right) \mathrm{d}t + \frac{\|U^-_{m-1}\|^2_{\Omega_{t_{m-1}}}}{\delta_2} + \delta_2 \|U^+_{m-1}\|_{\Omega_{t_{m-1}}}.\end{aligned} \tag{4.63}$$

Further, multiplying (4.63) by $\frac{\beta_0}{4c_2(q-1)}$, summing over $l = 1, \dots, q-1$ and adding to (4.44), we find that

$$\begin{aligned}&\|U^-_m\|^2_{\Omega_{t_m}} + \frac{\beta_0}{8c_2(q-1)} \sum_{l=1}^{q-1} \|U\|^2_{\Omega_{t_{m-1+l/q}}} + \left(\frac{\beta_0}{8c_2} + 1\right) \|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}} + \frac{\beta_0}{2} \int_{I_m} \|U\|^2_{\mathrm{DG},t} \mathrm{d}t \\ &\quad \le \frac{\beta_0}{4} \int_{I_m} \|U\|^2_{\mathrm{DG},t} \mathrm{d}t + \left(\frac{\beta_0}{4} + C^*_{T3}\right) \int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t \\ &\qquad + \left(\frac{\beta_0}{4} + C^{**}_{T3}\right) \int_{I_m} \left(\|g\|^2_{\Omega_t} + \|u_D\|^2_{\mathrm{DGB},t}\right) \mathrm{d}t \\ &\qquad + \left(\frac{\beta_0}{4c_2\delta_2} + \frac{2}{\delta_1}\right) \|U^-_{m-1}\|^2_{\Omega_{t_{m-1}}} + \left(\frac{\beta_0\delta_2}{4c_2} + 4\delta_1\right) \|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}}.\end{aligned}$$

Setting $c_3 := \min\left\{\frac{\beta_0}{8c_2(q-1)}, \frac{\beta_0}{8c_2} + 1\right\}$ and rearranging, we get

$$\begin{aligned}&c_3 \underbrace{\left(\|U^-_m\|_{\Omega^2_{t_m}} + \sum_{l=1}^{q-1} \|U^2_{m-1+l/q}\|^2_{\Omega_{t_{m-1+l/q}}} + \|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}}\right)}_{=\sum_{l=0}^{q} \|U_{m-1+l/q}\|^2_{\Omega_{t_{m-1+l/q}}}} + \frac{\beta_0}{4} \int_{I_m} \|U\|^2_{\mathrm{DG},t} \mathrm{d}t \\ &\le \left(\frac{\beta_0}{4} + C^*_{T3}\right) \int_{I_m} \|U\|^2_{\Omega_t} \mathrm{d}t + \left(\frac{\beta_0}{4} + C^{**}_{T3}\right) \int_{I_m} \left(\|g\|^2_{\Omega_t} + \|u_D\|^2_{\mathrm{DGB},t}\right) \mathrm{d}t \\ &\quad + \left(\frac{\beta_0}{4c_2\delta_2} + \frac{2}{\delta_1}\right) \|U^-_{m-1}\|^2_{\Omega_{t_{m-1}}} + \left(\frac{\beta_0\delta_2}{4c_2} + 4\delta_1\right) \|U^+_{m-1}\|^2_{\Omega_{t_{m-1}}}.\end{aligned}$$

It follows from inequalities (4.45) and (4.46) that

$$\begin{aligned}
&\frac{c_3 L_q^*}{\tau_m}\int_{I_m}\|U\|^2_{\Omega_t}\mathrm{d}t+\frac{\beta_0}{4}\int_{I_m}\|U\|^2_{\mathrm{DG},t}\mathrm{d}t\\
&\quad\le\left(\frac{\beta_0\delta_2 M_q^*}{4c_2\tau_m}+\frac{4\delta_1 M_q^*}{\tau_m}+\frac{\beta_0}{4}+C_{T3}^*\right)\int_{I_m}\|U\|^2_{\Omega_t}\mathrm{d}t\\
&\qquad+\left(\frac{\beta_0}{4}+C_{T3}^{**}\right)\int_{I_m}\left(\|g\|^2_{\Omega_t}+\|u_D\|^2_{\mathrm{DG},t}\right)\mathrm{d}t+\left(\frac{\beta_0}{4c_2\delta_2}+\frac{2}{\delta_1}\right)\|U_{m-1}^-\|^2_{\Omega_{t_{m-1}}}.
\end{aligned}$$

Setting $\delta_1=\frac{c_3L_q^*}{16M_q^*}$, $\delta_2=\frac{c_3c_2L_q^*}{\beta_0M_q^*}$, $c_4:=\frac{\beta_0}{4c_2\delta_2}+\frac{2}{\delta_1}$, $c_5:=\frac{\beta_0}{4}+C_{T3}^{**}$ we get

$$\begin{aligned}
&\left(\frac{c_3L_q^*}{2\tau_m}-\frac{\beta_0}{4}-C_{T3}^*\right)\int_{I_m}\|U\|^2_{\Omega_t}\mathrm{d}t+\frac{\beta_0}{4}\int_{I_m}\|U\|^2_{\mathrm{DG},t}\mathrm{d}t\\
&\qquad\le c_5\int_{I_m}\left(\|g\|^2_{\Omega_t}+\|u_D\|^2_{\mathrm{DGB},t}\right)\mathrm{d}t+c_4\|U_{m-1}^-\|^2_{\Omega_{t_{m-1}}}.
\end{aligned}\tag{4.64}$$

If the condition $0<\tau_m\le C_{T4}^*:=\frac{c_3L_q^*}{4(\frac{\beta_0}{4}+C_{T3}^*)}$ is satisfied, then $\frac{\beta_0}{4}+C_{T3}^*\ge\frac{c_3L_q^*}{4\tau_m}$ and from (4.64) we obtain the estimate

$$\frac{c_3L_q^*}{4\tau_m}\int_{I_m}\|U\|^2_{\Omega_t}\mathrm{d}t+\frac{\beta_0}{4}\int_{I_m}\|U\|^2_{\mathrm{DG},t}\,\mathrm{d}t\le c_5\int_{I_m}\left(\|g\|^2_{\Omega_t}+\|u_D\|^2_{\mathrm{DGB},t}\right)\mathrm{d}t+c_4\|U_{m-1}^-\|^2_{\Omega_{t_{m-1}}},$$

which implies (4.60). □

The stability analysis will be finished by the application of the following auxiliary lemma.

Lemma 4.11. *(Discrete Gronwall inequality) Let x_m, a_m, b_m and y_m, where $m=1,2,\ldots$, be non-negative sequences and let the sequence a_m be nondecreasing. Then, if*

$$\begin{aligned}
x_0+y_0&\le a_0,\\
x_m+y_m&\le a_m+\sum_{j=0}^{m-1}b_jx_j\quad \textit{for}\quad m\ge 1,
\end{aligned}$$

we have

$$x_m+y_m\le a_m\prod_{j=0}^{m-1}(1+b_j)\quad \textit{for}\quad m\ge 0.$$

The proof can be carried out by induction, see [20].

Now, if (4.60) is substituted into (4.43), an inequality is obtained, which is a basis of the proof of our main result about the stability:

$$\begin{aligned}
&\|U_m\|^2_{\Omega_{t_m}}-\|U_{m-1}^-\|^2_{\Omega_{t_{m-1}}}+\|\{U\}_{m-1}\|^2_{\Omega_{t_{m-1}}}+\frac{\beta_0}{2}\int_{I_m}\|U\|^2_{\mathrm{DG},m}\,\mathrm{d}t\\
&\qquad\le(C_{T2}+C_{T4}\,\tau_m)\int_{I_m}(\|g\|^2_{\Omega_t}+\|u_D\|^2_{\mathrm{DGB},t})\,\mathrm{d}t+C_{T2}C_{T4}\,\tau_m\|U_{m-1}^-\|^2_{\Omega_{t_{m-1}}}.
\end{aligned}\tag{4.65}$$

Theorem 4.5. *Let $0 < \tau_m \le C^*_{T4}$ for $m = 1, \dots, M$. Then there exists a constant $C_{T5} > 0$ such that*

$$\|U_m^-\|^2_{\Omega_{t_m}} + \sum_{j=1}^{m} \|\{U_{j-1}\}\|^2_{\Omega_{t_{j-1}}} + \frac{\beta_0}{2} \sum_{j=1}^{m} \int_{I_j} \|U\|^2_{\mathrm{DG},t} \,\mathrm{d}t \tag{4.66}$$
$$\le C_{T5} \left(\|U_0^-\|^2_{\Omega_{t_0}} + \sum_{j=1}^{m} \int_{I_j} R_{t,j} \,\mathrm{d}t \right), \quad m = 1, \dots, M,\ h \in (0, \overline{h}),$$

where $R_{t,j} = (C_{T2} + C_{T4}\,\tau_j)\,(\|g\|^2_{\Omega_t} + \|u_D\|^2_{\mathrm{DGB},t})$ for $t \in I_j$.

Proof. Writing j instead of m in (4.65),we obtain

$$\|U_j^-\|^2_{\Omega_{t_j}} - \|U_{j-1}^-\|^2_{\Omega_{t_{j-1}}} + \|\{U\}_{j-1}\|^2_{\Omega_{t_{m-1}}} + \frac{\beta_0}{2} \int_{I_j} \|U\|^2_{\mathrm{DG},t} \,\mathrm{d}t$$
$$\le \int_{I_j} R_{t,j} \,\mathrm{d}t + C_{T2} C_{T4}\,\tau_j \|U_{j-1}^-\|^2_{\Omega_{t_{j-1}}}.$$

Let $m \ge 1$. The summation over all $j = 1, \dots, m$ yields the inequality

$$\|U_m^-\|^2_{\Omega_{t_m}} + \sum_{j=1}^{m} \|\{U\}_{j-1}\|^2_{\Omega_{t_{j-1}}} + \frac{\beta_0}{2} \sum_{j=1}^{m} \int_{I_j} \|U\|^2_{\mathrm{DG},t} \,\mathrm{d}t$$
$$\le \|U_0^-\|^2_{\Omega_0} + C_{T2} C_{T4} \sum_{j=0}^{m} \tau_{j+1} \|U_j^-\|^2_{\Omega_{t_j}} + \sum_{j=1}^{m} \int_{I_j} R_{t,j} \,\mathrm{d}t.$$

The use of the discrete Gronwall inequality with setting

$$\begin{aligned}
x_0 &= a_0 = \|U_0^-\|^2_{\Omega_{t_0}}, \quad c_0 = 0,\\
x_m &= \|U_m^-\|^2_{\Omega_{t_m}},\\
y_m &= \sum_{j=1}^{m} \|\{U_{j-1}\}\|^2_{\Omega_{t_{j-1}}} + \frac{\beta_0}{2} \sum_{j=1}^{m} \int_{I_j} \|U\|^2_{\mathrm{DG},t} \,\mathrm{d}t,\\
a_m &= \|U_0^-\|^2_{\Omega_{t_0}} + \sum_{j=1}^{m} \int_{I_j} R_{t,j} \,\mathrm{d}t,\\
b_j &= C_{T2} C_{T4}\,\tau_{j+1}, \quad j = 0, 1, \dots, m,
\end{aligned}$$

yield

$$\|U_m^-\|^2_{\Omega_{t_m}} + \sum_{j=1}^{m} \|\{U_{j-1}\}\|^2_{\Omega_{t_{j-1}}} + \frac{\beta_0}{2} \sum_{j=1}^{m} \int_{I_j} \|U\|^2_{\mathrm{DG},t} \,\mathrm{d}t \tag{4.67}$$
$$\le \left(\|U_0^-\|^2_{\Omega_{t_0}} + \sum_{j=1}^{m} \int_{j} R_{t,j} \,\mathrm{d}t \right) \prod_{j=0}^{m-1} \left(1 + C_{T2} C_{T4}\,\tau_{j+1}\right).$$

Finally (4.67) and the inequality $1 + \sigma < \exp(\sigma)$ valid for any $\sigma > 0$ immediately yield (4.66) with the constant $C_{T5} := \exp(C_{T2} C_{T4} T)$. □

5. Conclusion

This paper is devoted to the stability analysis of the space-time discontinuous Galerkin method applied to the numerical solution of an initial-boundary value problem for a nonlinear convection-diffusion equation in a time-dependent domain. The problem is formulated with the aid of a new version of the arbitrary Lagrangian–Eulerian (ALE) method allowing to use different meshes in different time slabs. In the numerical scheme we use the nonsymmetric, symmetric and incomplete versions of the space discretization of diffusion terms and interior and boundary penalty. The nonlinear convection terms are discretized with the aid of a numerical flux. The space discretization uses piecewise polynomial approximations of degree $\le p$ with an integer $p \ge 1$. For the discontinuous Galerkin discretization in time we use polynomials of degree $\le q$ with $q \ge 1$. (We are not concerned with the case $q = 0$, which yields the simple backward Euler time discretization.) Main attention is paid here to the situation when $q \ge 2$, which is much more complicated and a special technique based on the ALE-generalization of the concept of the discrete characteristic function has been applied. This approach combined with a number of various estimates results in the proof of unconditional stability of the method. The obtained results represent a theoretical support of the ALE-STDGM developed in [16] for the numerical solution of compressible Navier-Stokes equations in time-dependent domains and interaction of compressible flow with elastic structures.

Further step will be the application of derived results to the analysis of error estimates of the ALE-STDGM in time-dependent domains. Interesting, but very difficult would be the analysis of the ALE-STDGM applied to singularly perturbed nonlinear problems, generalizing results of papers [45, 54].

Appendix A. Proof of estimates (4.41) and (4.42) from the proof of Theorem 4.1 in the 3D case (by Z. Vlasáková)

We introduce a parametrization of $\hat\Gamma$. Let Δ^2 be a reference simplex in $\mathbb{R}^2$ (with one vertex being the origin and all of the other vertices have only one non-zero coordinate equal to 1). Now

$$
\begin{aligned}
&\Gamma = \mathcal{A}_t(\hat\Gamma), \quad \hat\Gamma \in \mathcal{F}^I_{h,t_{m-1}},\\
&\hat\Gamma = \mathcal{B}^{\hat\Gamma}_{m-1}(\Delta^2) = \{X = \mathcal{B}^{\hat\Gamma}_{m-1}(v);\, v \in \Delta^2\},\\
&\mathrm{d}S^{\hat\Gamma} = \left\| \frac{\partial \mathcal{B}^{\hat\Gamma}_{m-1}}{\partial x^1}(v) \times \frac{\partial \mathcal{B}^{\hat\Gamma}_{m-1}}{\partial x^2}(v) \right\| \mathrm{d}x^1\mathrm{d}x^2, \quad v \in \Delta^2,\\
&\Gamma = \{x = \mathcal{A}_t(\mathcal{B}^{\hat\Gamma}_{m-1}(v));\, v \in \Delta^2\},\\
&\mathrm{d}S^{\Gamma} = \left\| \frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}(\mathcal{B}^{\hat\Gamma}_{m-1}(v)) \frac{\partial \mathcal{B}^{\hat\Gamma}_{m-1}}{\partial x^1}(v) \times \frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}(\mathcal{B}^{\hat\Gamma}_{m-1}(v)) \frac{\partial \mathcal{B}^{\hat\Gamma}_{m-1}}{\partial x^2}(v) \right\| \mathrm{d}x^1\mathrm{d}x^2, \quad v \in \Delta^2.
\end{aligned}
$$

By the symbol $\times$ we denote the vector product. The terms $\frac{\partial \mathcal{B}^{\hat\Gamma}_{m-1}}{\partial x^i}(v)$ are tangent vectors to $\hat\Gamma$ at the point $\mathcal{B}^{\hat\Gamma}_{m-1}(v)$. It follows from the properties of the mapping $\mathcal{A}_t$ that the values of $\frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}(\mathcal{B}^{\hat\Gamma}_{m-1}(v)) \frac{\partial \mathcal{B}^{\hat\Gamma}_{m-1}}{\partial x^i}(v)$ are identical from the sides of both elements $\hat K^{\hat\Gamma}_L$ and $\hat K^{\hat\Gamma}_R$ adjacent to $\hat\Gamma$.

In what follows, for the sake of simplicity, by c we denote a generic positive constant independent of h, with different values at different places. Then we can write

$$
\begin{aligned}
\int_\Gamma \frac{1}{h(\Gamma)} [U_s]^2 \mathrm{d}S^\Gamma = \int_{\Delta^2} \frac{1}{h(\Gamma)} &[U_s(\mathcal{A}_t(\mathcal{B}^{\hat\Gamma}_{m-1}(v)))]^2 \\
&\left\| \frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}(\mathcal{B}^{\hat\Gamma}_{m-1}(v)) \frac{\partial \mathcal{B}^{\hat\Gamma}_{m-1}}{\partial x^1}(v) \times \frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}(\mathcal{B}^{\hat\Gamma}_{m-1}(v)) \frac{\partial \mathcal{B}^{\hat\Gamma}_{m-1}}{\partial x^2}(v) \right\| \mathrm{d}x^1\mathrm{d}x^2
\end{aligned}
$$

$$\leq \int_{\Delta^2} \frac{1}{h(\Gamma)} [\tilde{U}_s(\mathcal{B}^{\hat{\Gamma}}_{m-1}(v))]^2 \left\| \frac{\mathrm{d}\mathcal{A}_t}{\mathrm{d}X}(\mathcal{B}^{\hat{\Gamma}}_{m-1}(v)) \right\|^2 \left\| \frac{\partial \mathcal{B}^{\hat{\Gamma}}_{m-1}}{\partial x^1}(v) \times \frac{\partial \mathcal{B}^{\hat{\Gamma}}_{m-1}}{\partial x^2}(v) \right\| \mathrm{d}x^1 \mathrm{d}x^2$$
$$\leq \int_{\hat{\Gamma}} \frac{c}{h(\hat{\Gamma})} [\tilde{U}_s]^2 \mathrm{d}S^{\hat{\Gamma}}.$$

Hence,

$$\int_{I_m} \left(\sum_{\Gamma \in \mathcal{F}^I_{h,t}} \frac{c_W}{h(\Gamma)} \int_\Gamma [U_s]^2 \mathrm{d}S^\Gamma \right) \mathrm{d}t \leq c \int_{I_m} \left(\sum_{\hat{\Gamma} \in \mathcal{F}^I_{h,t_{m-1}}} \frac{c_W}{h(\hat{\Gamma})} \int_{\hat{\Gamma}} [\tilde{U}]^2 \mathrm{d}S^{\hat{\Gamma}} \right) \mathrm{d}t.$$

Further for $\Gamma = \mathcal{A}_t(\hat{\Gamma})$, $\hat{\Gamma} \in \mathcal{F}^I_{h,t_{m-1}}$, we consider the parametrization

$$\Gamma = \{x = \mathcal{B}^\Gamma_t(v);\ v \in \Delta^2\},$$
$$\hat{\Gamma} = \{X = \mathcal{A}^{-1}_t(\mathcal{B}^\Gamma_t(v));\ v \in \Delta^2\},$$
$$\mathrm{d}S^\Gamma = \left\| \frac{\partial \mathcal{B}^\Gamma_{m-1}}{\partial x^1}(v) \times \frac{\partial \mathcal{B}^\Gamma_{m-1}}{\partial x^2}(v) \right\| \mathrm{d}v, \quad v \in \Delta^2$$
$$\mathrm{d}S^{\hat{\Gamma}} = \left\| \frac{\mathrm{d}\mathcal{A}^{-1}_t}{\mathrm{d}x}(\mathcal{B}^\Gamma_t(v)) \frac{\partial \mathcal{B}^\Gamma_t}{\partial x^1}(v) \times \frac{\mathrm{d}\mathcal{A}^{-1}_t}{\mathrm{d}x}(\mathcal{B}^\Gamma_t(v)) \frac{\partial \mathcal{B}^\Gamma_t}{\partial x^2}(v) \right\| \mathrm{d}v, \quad v \in \Delta^2.$$

Then

$$\int_{\hat{\Gamma}} [\tilde{U}]^2 \mathrm{d}S^{\hat{\Gamma}} = \int_{\Delta^2} [\tilde{U}(\mathcal{A}^{-1}_t(\mathcal{B}^\Gamma_t(v)))]^2 \left\| \frac{\mathrm{d}\mathcal{A}^{-1}_t}{\mathrm{d}x}(\mathcal{B}^\Gamma_t(v)) \frac{\partial \mathcal{B}^\Gamma_t}{\partial x^1}(v) \times \frac{\mathrm{d}\mathcal{A}^{-1}_t}{\mathrm{d}x}(\mathcal{B}^\Gamma_t(v)) \frac{\partial \mathcal{B}^\Gamma_t}{\partial x^2}(v) \right\| \mathrm{d}x^1 \mathrm{d}x^2$$
$$\leq \int_{\Delta^2} [U(\mathcal{B}^\Gamma_t(v))]^2 \left\| \frac{\mathrm{d}\mathcal{A}^{-1}_t}{\mathrm{d}x}(\mathcal{B}^\Gamma_t(v)) \right\|^2 \left\| \frac{\partial \mathcal{B}^\Gamma_{m-1}}{\partial x^1}(v) \times \frac{\partial \mathcal{B}^\Gamma_{m-1}}{\partial x^2}(v) \right\| \mathrm{d}x^1 \mathrm{d}x^2$$
$$\leq c \int_{\Delta^2} [U]^2 \mathrm{d}S^\Gamma.$$

Together we get

$$\int_{I_m} \left(\sum_{\Gamma \in \mathcal{F}^I_{h,t}} \frac{c_W}{h(\Gamma)} \int_\Gamma [U_s]^2 \mathrm{d}S^\Gamma \right) \mathrm{d}t \leq c \int_{I_m} \left(\sum_{\Gamma \in \mathcal{F}^I_{h,t}} \frac{c_W}{h(\Gamma)} \int_\Gamma [U]^2 \mathrm{d}S^\Gamma \right) \mathrm{d}t,$$

which is the 3D version of (4.41). Similarly we prove (4.42) in the 3D case.

Acknowledgements. This research was supported by the grant 17-01747S of the Czech Science Foundation and the research of M. Balázsová was also supported by the Charles University in Prague, project GA UK No. 127615. M. Vlasák is a junior member of the University centre for mathematical modeling, applied analysis and computational mathematics (MathMAC). We are grateful to Z. Vlasáková for stimulating suggestions in the analysis of the discrete characteristic functions. We also acknowledge our membership in the Nečas Center of Mathematical Modeling (http://ncmm.karlin.mff.cuni.cz).

References

[1] G. Akrivis and C. Makridakis, Galerkin time-stepping methods for nonlinear parabolic equations. *ESAIM: M2AN* **38** (2004) 261–289.

[2] D.N. Arnold, F. Brezzi, B. Cockburn and D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.* **39** (2002) 1749–1779.

[3] I. Babuška, C.E. Baumann and T. J. Oden, A discontinuous *hp* finite element method for diffusion problems, 1D analysis. *Comput. Math. Appl.* **37** (1999) 103–122.

[4] S. Badia and R. Codina, Analysis of a stabilized finite element approximation of the transient convection-diffusion equation using an ALE framework. *SIAM J. Numer. Anal.* **44** (2006) 2159–2197.

[5] M. Balázsová and M. Feistauer, On the stability of the space-time discontinuous Galerkin method for nonlinear convection-diffusion problems in time-dependent domains. *Appl. Math.* **60** (2015) 501–526.

[6] M. Balázsová and M. Feistauer, On the uniform stability of the space-time discontinuous Galerkin method for nonstationary problems in time-dependent domains. In: *Proc. Conf. ALGORITMY* (2016) 84–92.

[7] M. Balázsová, M. Feistauer, M. Hadrava and A. Kosík, On the stability of the space-time discontinuous Galerkin method for the numerical solution of nonstationary nonlinear convection-diffusion problems. *J. Numer. Math.* **23** (2015) 211–233.

[8] F. Bassi and S. Rebay, A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations. *J. Comput. Phys.* **131** (1997) 267–279.

[9] C.E. Baumann and T.J. Oden, A discontinuous *hp* finite element method for the Euler and Navier–Stokes equations. *Int. J. Numer. Methods Fluids* **31** (1999) 79–95.

[10] D. Boffi, L. Gastaldi and L. Heltai, Numerical stability of the finite element immersed boundary method. *Math. Models Methods Appl. Sci.* **17** (2007) 1479–1505.

[11] A. Bonito, I. Kyza and R.H. Nochetto, Time-discrete higher-order ALE formulations: stability. *SIAM J. Numer. Anal.* **51** (2013) 577–604.

[12] A. Bonito, I. Kyza and R.H. Nochetto, Time-discrete higher order ALE formulations: a priori error analysis. *Numer. Math.* **125** (2013) 225–257.

[13] F. Brezzi, G. Manzini, D. Marini, P. Pietra and A. Russo, Discontinuous Galerkin approximations for elliptic problems. *Numer. Methods Partial Differ. Equ.* **16** (2000) 365–378.

[14] J. Česenek and M. Feistauer, Theory of the space-time discontinuous Galerkin method for nonstationary parabolic problems with nonlinear convection and diffusion. *SIAM J. Numer. Anal.* **50** (2012) 1181–1206.

[15] J. Česenek, M. Feistauer, J. Horáček, V. Kučera and J. Prokopová, Simulation of compressible viscous flow in time-dependent domains. *Appl. Math. Comput.* **219** (2013) 7139–7150.

[16] J. Česenek, M. Feistauer and A. Kosík, DGFEM for the analysis of airfoil vibrations induced by compressible flow. *ZAMM Z. Angew. Math. Mech.* **93** (2013) 387–402.

[17] K. Chrysafinos and N.J. Walkington, Error estimates for the discontinuous Galerkin methods for parabolic equations. *SIAM J. Numer. Anal.* **44** (2006) 349–366.

[18] B. Cockburn and C.-W. Shu, Runge–Kutta discontinuous Galerkin methods for convection-dominated problems. Review article. *J. Sci. Comput.* **16** (2001) 173–261.

[19] V. Dolejší, On the discontinuous Galerkin method for the numerical solution of the Navier–Stokes equations. *Int. J. Numer. Methods Fluids* **45** (2004) 1083–1106.

[20] V. Dolejší and M. Feistauer, Discontinuous Galerkin Method – Analysis and Applications to Compressible Flow. Springer, Berlin (2015).

[21] V. Dolejší, M. Feistauer and J. Hozman, Analysis of semi-implicit DGFEM for nonlinear convection-diffusion problems on nonconforming meshes. *Comput. Methods Appl. Mech. Eng.* **196** (2007) 2813–2827.

[22] J. Donéa, S. Giuliani and J. Halleux, An arbitrary Lagrangian–Eulerian finite element method for transient dynamic fluid-structure interactions. *Comput. Methods Appl. Mech. Eng.* **33** (1982) 689–723.

[23] K. Eriksson, D. Estep, P. Hansbo and C. Johnson, Computational Differential Equations. Cambridge University Press, Cambridge (1996)

[24] K. Eriksson and C. Johnson, Adaptive finite element methods for parabolic problems I: a linear model problem. *SIAM J. Numer. Anal.* **28** (1991) 43–77.

[25] D. Estep and S. Larsson, The discontinuous Galerkin method for semilinear parabolic problems. *ESAIM: M2AN* **27** (1993) 35–54.

[26] M. Feistauer, V. Dolejší and V. Kučera, On the discontinuous Galerkin method for the simulation of compressible flow with wide range of Mach numbers. *Comput. Visual. Sci.* **10** (2007) 17–27.

[27] M. Feistauer, J. Felcman and I. Straškraba, Mathematical and Computational Methods for Compressible Flow. Clarendon Press, Oxford (2003)

[28] M. Feistauer, M. Hadrava, J. Horáček and A. Kosík, Numerical solution of fluid-structure interaction by the space-time discontinuous Galerkin method, edited by J. Fuhrmann, M. Ohlberger and C. Rohde. In: *Proc. of the conf. FVCA7 (Finite volumes for complex applications VII elliptic, parabolic and hyperbolic problems), Berlin, June 16–20, 2014*. Springer, Cham (2014) 567–575.

[29] M. Feistauer, J. Hájek and K. Švadlenka, Space-time discontinuous Galerkin method for solving nonstationary linear convection-diffusion-reaction problems. *Appl. Math.* **52** (2007) 197–233.

[30] M. Feistauer, J. Hasnedlová-Prokopová, J. Horáček, A. Kosík and V. Kučera, DGFEM for dynamical systems describing interaction of compressible fluid and structures. *J. Comput. Appl. Math.* **254** (2013) 17–30.

[31] M. Feistauer, J. Horáček, V. Kučera and J. Prokopová, On the numerical solution of compressible flow in time-dependent domains. *Math. Bohem.* **137** (2012) 1–16.

[32] M. Feistauer and V. Kučera, On a robust discontinuous Galerkin technique for the solution of compressible flow. *J. Comput. Phys.* **224** (2007) 208–221.

[33] M. Feistauer, V. Kučera, K. Najzar and J. Prokopová, Analysis of space-time discontinuous Galerkin method for nonlinear convection-diffusion problems. *Numer. Math.* **117** (2011) 251–288.

[34] M. Feistauer, V. Kučera and J. Prokopová, Discontinuous Galerkin solution of compressible flow in time-dependent domains. *Math. Comput. Simul.* **80** (2010) 1612–1623.

[35] L. Formaggia and F. Nobile, A stability analysis for the arbitrary Lagrangian Eulerian formulation with finite elements. *East–West J. Numer. Math.* **7** (1999) 105–131.

[36] L. Gastaldi, A priori error estimates for the Arbitrary Lagrangian Eulerian formulation with finite elements. *East–West J. Numer. Math.* **9** (2001) 123–156.

[37] J. Hasnedlová, M. Feistauer, J. Horáček, A. Kosík and V. Kučera, Numerical simulation of fluid-structure interaction of compressible flow and elastic structure. *Computing* **95** (2013) 343–361.

[38] O. Havle, V. Dolejší and M. Feistauer, Discontinuous Galerkin method for nonlinear convection-diffusion problems with mixed Dirichlet–Neumann boundary conditions. *Appl. Math.* **55** (2010) 353–372.

[39] C.W. Hirt, A.A. Amsdem and J.L. Cook, An arbitrary Lagrangian–Eulerian computing method for all flow speeds. *J. Comput. Phys.* **135** (1997) 198–216.

[40] P. Houston, C. Schwab and E. Süli, Discontinuous *hp*-finite element methods for advection-diffusion problems. *SIAM J. Numer. Anal.* **39** (2002) 2133–2163.

[41] T.J.R. Hughes, W.K. Liu and T.K. Zimmermann, Lagrangian–Eulerian finite element formulation for incompressible viscous flows. *Comput. Methods Appl. Mech. Eng.* **29** (1981) 329–349.

[42] T. Nomura and T.J.R. Hughes, An arbitrary Lagrangian–Eulerian finite element method for interaction of fluid and a rigid body. *Comput. Methods Appl. Mech. Eng.* **95** (1992) 115–138.

[43] K. Khadra, P. Angot, S. Parneix and J.-P. Caltagirone, Fictitious domain approach for numerical modelling of Navier–Stokes equations. *Int. J. Numer. Methods Fluids* **34** (2000) 651–684.

[44] A. Kosík, M. Feistauer, M. Hadrava and J. Horáček, Numerical simulation of the interaction between a nonlinear elastic structure and compressible flow by the discontinuous Galerkin method. *Appl. Math. Comput.* **267** (2015) 382–396.

[45] V. Kučera and M. Vlasák, A priori diffusion-uniform error estimates for nonlinear singularly perturbed problems: BDF2, midpoint and time DG. *ESAIM: M2AN* **51** (2017) 537–563.

[46] J.T. Oden, I. Babuška and C.E. Baumann, A discontinuous *hp* finite element method for diffusion problems. *J. Comput. Phys.* **146** (1998) 491–519.

[47] C.-W. Shu, Discontinuous Galerkin method for time dependent problems: survey and recent developments, edited by X. Feng *et al.* Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations. Springer, Cham (2014) 25–62.

[48] D. Schötzau, *hp-DGFEM for parabolic evolution problems. Applications to diffusion and viscous incompressible fluid flow.* Ph.D. thesis, ETH No. 13041, Zürich (1999).

[49] D. Schötzau and C. Schwab, An *hp* a priori error analysis of the Discontinuous Galerkin time-stepping method for initial value problems. *Calcolo* **37** (2000) 207–232.

[50] V. Thomée, Galerkin Finite Element Methods for Parabolic Problems. Springer, Berlin (2006).

[51] J.J.W. van der Vegt and H. van der Ven, Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows. Part I. General formulation. *J. Comput. Phys.* **182** (2002) 546–585.

[52] M. Vlasák, Optimal spatial error estimates for DG time discretizations. *J. Numer. Math.* **21** (2013) 201–230.

[53] M. Vlasák, V. Dolejší and J. Hájek, A Priori error estimates of an extrapolated space-time discontinuous Galerkin method for nonlinear convection-diffusion problems. *Numer. Methods Partial Differ. Equ.* **27** (2011) 1456–1482.

[54] Q. Zhang and C.-W. Shu, Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws. *SIAM J. Numer. Anal.* **42** (2004) 641–666.

Chapter 6

A posteriori error estimates for nonlinear parabolic problems

SIAM J. NUMER. ANAL.
Vol. 59, No. 3, pp. 1486–1509

A POSTERIORI ERROR ESTIMATES FOR HIGHER ORDER SPACE-TIME GALERKIN DISCRETIZATIONS OF NONLINEAR PARABOLIC PROBLEMS*

VÍT DOLEJŠÍ†, FILIP ROSKOVEC†, AND MILOSLAV VLASÁK‡

Abstract. We deal with the numerical solution of nonlinear time-dependent convection-diffusion-reaction equations with the aid of continuous and discontinuous Galerkin discretization of an arbitrary polynomial approximation degree. We derive a posteriori error estimates in the space-time mesh-dependent dual norm. The estimates are based on the equilibrated flux reconstruction techniques which are locally computable. We prove the upper and lower bounds and present several numerical experiments justifying the theoretical results.

Key words. a posteriori error estimates, equilibrated flux reconstruction, nonlinear convection-diffusion, reliability, efficiency

AMS subject classifications. 65M60, 65M15

DOI. 10.1137/18M117594X

1. Introduction.

We deal with an a posteriori error analysis of a scalar nonlinear time-dependent convection-diffusion-reaction problem which is discretized by continuous and/or discontinuous Galerkin approximation of an arbitrary polynomial degree with respect to the space and time. The aim is to derive error estimates which are *reliable* (the upper bounds do not depend on unknown constants), *locally computable* and *locally efficient* (local lower bounds are valid).

There are a number of results devoted to a posteriori error estimates of parabolic problems. For an introduction to the topic with an overview of the concepts, we refer to [46]. The upper bound for the heat equation problem is derived in, e.g., [20], [32], [36], or [38] and the applications to adaptivity can be found in, e.g., [8], [29]. The upper and lower bounds for linear parabolic problems are derived in, e.g., [5], [17], or [44]. For an extension to nonlinear problems, see, e.g., [19], [28], [43], and [45].

A posteriori error analysis for nonlinear parabolic problems is addressed in [11], where a special dual norm of the residual is constructed. This norm enables one to derive a reliable upper bound that is efficient locally in time and globally in space. However, all the papers mentioned above are devoted to a *low* (first or second) order time discretizations only. The analysis of linear problems discretized by the higher order methods in time can be found, e.g., in [2], [14], [15], [41]. The analysis of the higher order time discretizations for nonlinear problems can be found, e.g., in [33], where the upper bound consists of dual norms and it is not directly computable.

The aim of this paper is to generalize the results from [11] to higher order time discretizations, namely to continuous Galerkin (cG) and discontinuous Galerkin (dG) time discretization methods in the combination with the classical conforming finite

*Received by the editors March 16, 2018; accepted for publication (in revised form) February 22, 2021; published electronically June 3, 2021.
https://doi.org/10.1137/18M117594X
Funding: This research was financially supported by the Czech Science Foundation grants 20-01074S (first and second authors) and 20-14736S (third author). Moreover, the second author's research was supported by Charles University grant SVV-2018-260455.
†Faculty of Mathematics and Physics, Charles University, Sokolovska 83, 186 75 Prague 8, Czech Republic (dolejsi@karlin.mff.cuni.cz, roskovec@gmail.com).
‡Faculty of Civil Engineering, Czech Technical University, Thakurova 7, 166 29 Prague 6, Czech Republic (vlasamil@cvut.cz).

element method (FEM) and discontinuous Galerkin finite element method (DGFEM) in space. Since Galerkin time discretizations use the same principles in time as FEM or DGFEM do in space, it is possible to develop a posteriori error analysis of the time discretization in a similar way as for the space discretizations. Then the reconstruction with respect to time can be made with the aid of interpolation on right Gauss–Radau quadrature nodes. This type of reconstruction is well known from the analysis connecting dG, Radau collocation method, and Radau IIA Runge-Kutta method; see, e.g., [2], [22], [25], and namely the seminal paper [13].

Throughout this paper we deal with four methods for which we use the following abbreviations denoting the combination of time-space discretization:

- cG–FEM: the continuous Galerkin method in time and the conforming finite element method in space,
- dG–FEM: the discontinuous Galerkin method in time and the conforming finite element method in space,
- cG–DGFEM: the continuous Galerkin method in time and the discontinuous Galerkin method in space,
- dG–DGFEM: the discontinuous Galerkin method in time as well as in space.

An important contribution of this paper lies in theoretical justifications of unified analysis of cG (conforming) and dG (nonconforming) time discretizations. This is possible due to the reformulation of the original parabolic problem in such a way that the exact solution of the original problem is also the solution of the new one and that both discretization methods behave as conforming with respect to the new reformulated problem. These properties enable us to derive a posteriori upper bound, where the penalization of nonconformity is naturally included into the estimator and no additional artificial penalization term is needed. We employ the *equilibrated flux reconstruction* technique, which is close to the hypercircle method; see [37]. For an overview of the equilibrated flux reconstruction technique applied to the stationary problems, see, e.g., [1], [6], [18], [16], [31] and references cited therein.

The theoretical analysis in [6] and [18] also shows the independence of the constants arising in the efficiency estimates on the polynomial degree for certain types of flux reconstructions. See also [34] and [47], where the analysis of polynomial dependence of efficiency estimates is presented too. Nevertheless, this topic is beyond the scope of this paper and will be possibly addressed in future works of the authors.

For time-dependent problems, it is natural to consider discretization on changing (space) meshes in time. In order to simplify the exposition, we start with the cG–FEM and dG–FEM approach on fixed meshes. Then we proceed with the cG–DGFEM and dG–DGFEM techniques on meshes changing in time. An extension of cG–FEM and dG–FEM on varying meshes is straightforward.

The contents of the rest of this paper is the following. In section 2, we introduce the continuous problem, whose (space-continuous) Galerkin discretization is treated in section 3. In section 4, we describe the error measure as dual norm of residual with respect to the reformulated problem for the cG–FEM and dG–FEM techniques, reconstruct the solution with respect to time and space, and derive an a posteriori upper bound. The corresponding local lower bounds are presented in section 5. In section 6, we briefly extend the results to discontinuous space discretization (cG-DGFEM and dG–DGFEM) including the generalization for meshes changing in time. Finally, section 7 contains several numerical experiments.

2. Continuous problem. Let $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ be a bounded polyhedral domain with Lipschitz continuous boundary $\partial\Omega$ and $T > 0$. We use standard notation

for Lebesgue, Sobolev, and Bochner spaces. Let $(.,.)$ and $\|.\|$ be the $L^2(\Omega)$ scalar product and norm, respectively. Let us consider the following initial-boundary value problem: find $u : \Omega \times (0,T) \to \mathbb{R}$ such that

$$\begin{aligned} \frac{\partial u}{\partial t} - \nabla \cdot \sigma(u, \nabla u) + c(u) &= 0 \quad \text{in } \Omega \times (0,T), \\ u &= 0 \quad \text{in } \partial\Omega \times (0,T), \\ u(0) &= u^0 \quad \text{in } \Omega, \end{aligned} \tag{1}$$

where the initial condition $u^0 \in H_0^1(\Omega)$, the reaction term $c : L^2(0,T,L^2(\Omega)) \to L^2(0,T,L^2(\Omega))$, and the flux term $\sigma : L^2(0,T,L^2(\Omega))^{d+1} \to L^2(0,T,L^2(\Omega)^d)$. We assume that $\sigma(u,\nabla u) \in L^2(0,T,H(\mathrm{div},\Omega))$ for the sufficiently regular exact solution u (cf. Definition 1), where $H(\mathrm{div},\Omega) = \{v \in L^2(\Omega)^d : \mathrm{div}\, v \in L^2(\Omega)\}$. Moreover, we assume that the complete spatial operator is monotone, i.e.,

$$(\sigma(u,\nabla u) - \sigma(v,\nabla v), \nabla u - \nabla v) + (c(u) - c(v), u - v) \geq 0 \quad \forall u, v \in H_0^1(\Omega). \tag{2}$$

Let us note that (1) also covers the case when a source term $g(x,t)$ appears on the right-hand side of (1). Then the source term can be absorbed by the nonlinear terms, e.g., by setting $\tilde{c}(u) := c(u) - g(x,t)$. The assumption (2) remains preserved. For simplicity, we consider the homogeneous Dirichlet boundary condition which can be relaxed.

Let us denote the weak time derivative $u' = \frac{\partial u}{\partial t}$ and define spaces

$$\begin{aligned} X &= L^2(0,T,H_0^1(\Omega)), \\ Y &= \{v \in X : v' \in L^2(0,T,L^2(\Omega))\} \subset C([0,T],L^2(\Omega)), \\ Y^0 &= \{v \in Y : v(0) = u^0\}. \end{aligned} \tag{3}$$

DEFINITION 1. *We say that the function $u \in Y^0$ is the weak solution of* (1) *if*

$$\int_0^T (u', v) + (\sigma(u,\nabla u), \nabla v) + (c(u), v)\mathrm{d}t = 0 \quad \forall v \in X. \tag{4}$$

In this paper we assume that there exists a solution of problem (4). The possible proof of the existence of the solution can be carried out under the assumption of the Lipschitz continuity and coercivity of data; see, e.g., [30, Part II, section 2]. The uniqueness of the solution of problem (4) follows from (2). Assumption (2) can be relaxed, e.g., by Gårding-like inequality.

3. cG–FEM, dG–FEM discretizations. We consider a space partition $\mathcal{T}_h$ consisting of a finite number of closed, d-dimensional simplices with mutually disjoint interiors and covering $\overline{\Omega}$, i.e., $\overline{\Omega} = \cup_{K \in \mathcal{T}_h} K$. We assume conforming properties of the mesh, i.e., neighboring elements share an entire edge or face. In the rest of this paper we speak only about edges, but we mean edges or faces depending on the dimension d. We denote the vertices of the mesh by a and edges by e. We set $h_e = \mathrm{diam}(e)$, $h_K = \mathrm{diam}(K)$ and $h = \max_K h_K$. We assume shape regularity of elements, i.e., $h_K/\rho_K \leq C$ for all $K \in \mathcal{T}_h$, where ρ_K is the radius of the largest d-dimensional ball inscribed into K and constant C does not depend on $\mathcal{T}_h$ for $h \in (0,h_0)$. Moreover, we assume the local quasi-uniformity of the mesh, i.e., $h_K \leq C h_{K'}$ for neighboring elements K and K', where constant C does not depend on $\mathcal{T}_h$ for $h \in (0,h_0)$ again. For each edge e, let $n = n_e$ denote an unit normal vector to e with arbitrary but

fixed direction for the inner edges and with outer direction on $\partial\Omega$. Moreover, for each $K \in \mathcal{T}_h$, n_K is the unit outer normal vector to K.

For the purpose of the classical FEM, we define the space

$$V_h = \{v \in H_0^1(\Omega) : v|_K \in P^p(K), K \in \mathcal{T}_h\}, \tag{5}$$

where the space $P^p(K)$ denotes the space of polynomials on K up to the degree $p \geq 1$ (fixed for all $K \in \mathcal{T}_h$). We denote by Π the L^2-orthogonal projection on V_h.

In order to discretize problem (4) in time, we consider a time partition $0 = t_0 < t_1 < \cdots < t_r = T$ with time intervals $I_m = (t_{m-1}, t_m)$, time steps $\tau_m = |I_m| = t_m - t_{m-1}$ and $\tau = \max_{m=1,\ldots,r} \tau_m$. In the following, we consider two variants of the time discretization, conforming continuous Galerkin (cG) method and nonconforming discontinuous Galerkin (dG) method. For the conforming case, we seek the discrete solution in the affine subspace of Y, i.e., in

$$Y_h^\tau = \{v \in Y : v|_{I_m} \in P^{q+1}(I_m, V_h), v(0) = \Pi u^0\}, \tag{6}$$

where $P^q(I_m, B)$ is the space of polynomial functions in time of degree less than or equal to $q \geq 0$ with values in the Hilbert space B. It is important to notice that Y_h^τ is a subspace of Y and not a subspace of Y^0 in general, because Y_h^τ violates the initial condition. Since $Y \subset C([0,T], L^2(\Omega))$, this space consists of piecewise polynomials that are continuous in space and time and satisfy the approximation of the initial condition Πu^0 and homogeneous boundary conditions.

For the nonconforming case, we define the space of piecewise polynomial functions

$$X_h^\tau = \{v \in L^2(0,T,V_h) : v|_{I_m} \in P^q(I_m, V_h)\}. \tag{7}$$

Let us note that the dimension of Y_h^τ and X_h^τ is the same $(= r(q+1)\dim(V_h))$ since in X_h^τ we consider polynomials of degree q only, but the discontinuities at t_m, $m = 1, \ldots, r-1$, give the additional degrees of freedom. The spaces Y_h^τ and X_h^τ represent natural discrete variants of spaces Y^0 and X, respectively.

Since space X_h^τ consists of functions that can be discontinuous in time, we need to define the one-sided limits and jumps

$$\begin{aligned} &v_+^m = \lim_{t \to t_m +} v(t), \quad m = 0, \ldots, r-1, \qquad v_-^m = \lim_{t \to t_m -} v(t), \quad m = 1, \ldots, r, \\ &\{v\}_m = v_+^m - v_-^m, \quad m = 1, \ldots, r-1, \qquad v_-^0 = u^0, \qquad \{v\}_0 = v_+^0 - u^0, \end{aligned} \tag{8}$$

where u^0 is the initial condition. We omit the subscript $\pm$ for continuous functions.

In order to simplify the notation, we set the local L^2-scalar products

$$(u,v)_M = \int_M uv \,\mathrm{d}x, \qquad (u,v)_{M,m} = \int_{M \times I_m} uv \,\mathrm{d}x\mathrm{d}t \quad \forall m = 1, \ldots, r, \tag{9}$$

where $M \subset \overline{\Omega}$ is some collection of elements K from $\mathcal{T}_h$, and the corresponding norms $\|.\|_M$, $\|.\|_{M,m}$. Similarly, for $K \in \mathcal{T}_h$, $m = 1, \ldots, r$, we define

$$a_{K,m}(u,v) = (\sigma(u, \nabla u), \nabla v)_{K,m} + (c(u), v)_{K,m}, \quad u, v \in X \tag{10}$$

to describe localized version of the spatial operator on $K \times I_m$. By $\sum_{K,m}$ we will denote a sum over all space-time elements $K \times I_m$, where $K \in \mathcal{T}_h$ and $m = 1, \ldots, r$.

Now we are able to define the cG–FEM and dG–FEM discretizations. The cG–FEM discretization is based on the original weak formulation (4), where the original spaces Y^0 and X are naturally approximated by spaces Y_h^τ and X_h^τ, respectively.

DEFINITION 2. *We say that the function $u_h^\tau \in Y_h^\tau$ is the approximate solution of* (4) *obtained by the time continuous Galerkin–FEM (cG–FEM) if*

$$\sum_{K,m} (((u_h^\tau)', v)_{K,m} + a_{K,m}(u_h^\tau, v)) = 0 \quad \forall v \in X_h^\tau. \tag{11}$$

On the other hand, the dG–FEM discretization approximates the solution space Y^0 by X_h^τ. This violation of the nature of space Y^0 is compensated by the additional jump terms.

DEFINITION 3. *We say that the function $u_h^\tau \in X_h^\tau$ is the approximate solution of* (4) *obtained by the time discontinuous Galerkin–FEM (dG–FEM) if*

$$\sum_{K,m} \left(((u_h^\tau)', v)_{K,m} + a_{K,m}(u_h^\tau, v) + (\{u_h^\tau\}_{m-1}, v_+^{m-1})_K\right) = 0 \quad \forall v \in X_h^\tau. \tag{12}$$

The methods (11) and (12) can be viewed as a generalization of classical one-step methods for parabolic problems. It is possible to show that setting $q = 0$, i.e., piecewise linear continuous approximation in time for cG–FEM or piecewise constant approximation in time for dG–FEM, are equivalent (up to a suitable quadrature of the time integral) to the Crank–Nicolson method or backward Euler method, respectively, in time combined with FEM in space. These methods for $q \geq 1$ (with corresponding Gauss or right Gauss–Radau quadrature) lead to certain well-known implicit Runge–Kutta methods (Kunzmann–Butchder method (also known as Gauss–Legendre method) or Radau IIA method). For details about relations between Runge–Kutta methods, collocation methods, and Galerkin methods, see, e.g., [2], [22], [25], and [48] for the details about implicit Runge–Kutta methods; see, e.g., [7], [23], and [24].

In virtue of (12), for $K \in \mathcal{T}_h$, $m = 1, \ldots, r$, we set

$$b_{K,m}(u, v) = (u', v)_{K,m} + a_{K,m}(u, v) + (\{u\}_{m-1}, v_+^{m-1})_K. \tag{13}$$

The form $b_{K,m}$ is not linear, but it is affine, since the first jump term includes the initial condition; see (8). This enables us to work with $b_{K,m}$ on the first time level in the same way as on the other levels. This unified notation naturally covers the nonconformity in the deviation of the initial condition in the same way as the usual nonconformity coming from the discontinuity of the discrete solution; see section 4.2.

4. A posteriori error analysis of cG–FEM and dG–FEM discretizations.

4.1. Error measure. Inspired by [11, section 2.3.1], we define a parameter associated to space-time element $K \times I_m$ by $d_{K,m}$. The forthcoming analysis will be independent of the choice of this parameter. We only assume for the analysis that $d_{K,m} > 0$ and that $d_{K,m}$ is locally quasi-uniform, i.e., $d_{K,m} \leq C d_{K',m}$ for neigboring elements K and K', where constant C does not depend on $\mathcal{T}_h$ for $h \in (0, h_0)$ and $m = 1, \ldots, r$. Expected possible choices are $d_{K,m}^2 = h_K^2 + \tau_m^2$ or $d_{K,m} = 1$ or $d_{K,m} = h_K$. The assumption of local quasi-uniformity of $d_{K,m}$ (as can be seen from expected choices for $d_{K,m}$) brings most often no additional restrictions with respect to the adaptivity of the mesh, since the local quasi-uniformity of the parameter (or space elements, respectively) is the direct consequence of the shape regularity assumption that is typically assumed for most discretizations already.

Let us define the space

$$Y^\tau = \{v \in X : v'|_{I_m} \in L^2(I_m, L^2(\Omega))\} \tag{14}$$

of piecewise continuous functions with respect to time. In fact, Y^τ is a broken Sobolev space with respect to time. From the definition of the space Y^τ we can see that

$$Y^0 \subset Y \subset Y^\tau \subset X, \qquad Y_h^\tau,\, X_h^\tau \subset Y^\tau. \tag{15}$$

This space is suitable for redefinition of the weak formulation (4), which is replaced by the following one: find $u \in Y^\tau$ such that

$$\sum_{K,m} b_{K,m}(u,v) = 0 \quad \forall v \in Y^\tau. \tag{16}$$

LEMMA 4. *Let $u \in Y^0$ be the unique solution of problem* (4). *Then u is the unique solution of problem* (16).

Proof. From (15) we can see that the unique exact solution of former problem (4) is also the solution of problem (16). It is sufficient to prove the uniqueness. We follow the idea of energy estimates from the seminal work [13, Lemma 4]. Let $u \in Y^0$ be the original solution of problem (4) that satisfies problem (16), and let $u_1 \in Y^\tau$ be another solution of (16). Then subtracting the relations (16) for u and u_1 with $v = 2(u - u_1)$ and using monotonicity (2), we gain

$$\begin{aligned} 0 &= \sum_{K,m} \left(2((u-u_1)', u-u_1)_{K,m} + 2a_{K,m}(u, u-u_1) - 2a_{K,m}(u_1, u-u_1)\right) \\ &\quad + \sum_{K,m} 2(\{u-u_1\}_{m-1}, (u-u_1)_+^{m-1})_K \\ &\geq \sum_{K,m} \left(2((u-u_1)', u-u_1)_{K,m} + 2(\{u-u_1\}_{m-1}, (u-u_1)_+^{m-1})_K\right) \\ &= \|(u-u_1)_-^r\|^2 + \sum_{K,m} \|\{u-u_1\}_{m-1}\|_K^2. \end{aligned} \tag{17}$$

Since the original solution u is continuous ($u \in Y^0$), we get that $\{u\}_{m-1} = 0$. Then relation (17) implies that $\{u_1\}_{m-1} = 0$ too. From this follows that $u_1 \in Y^0$ and satisfies (4). ∎

The reason for the redefinition of the standard formulation (4) is that (16) is closer to the nonconforming dG–FEM discretization which can be handled as a conforming discretization with respect to (16), i.e., discretization on subspaces; see (15).

We define a norm on Y^τ,

$$\|v\|_{Y^\tau}^2 = \sum_{K,m} \|v\|_{Y^\tau,K,m}^2 \text{ with } \|v\|_{Y^\tau,K,m}^2 = d_{K,m}^{-2} \left(h_K^2 \|\nabla v\|_{K,m}^2 + \tau_m^2 \|v'\|_{K,m}^2\right). \tag{18}$$

The same concept of norms is used in [11, section 2.3.1]. Following formulation (16) and using norm $\|.\|_{Y^\tau}$, we define the error measure $\mathcal{E} = \mathcal{E}(u_h^\tau)$ as a dual norm of the residual

$$\mathcal{E} = \mathrm{Res}(u_h^\tau) = \sup_{0 \neq v \in Y^\tau} \frac{\sum_{K,m} b_{K,m}(u_h^\tau, v)}{\|v\|_{Y^\tau}}. \tag{19}$$

Let $u_h^\tau \in X_h^\tau$ be an arbitrary function. Then $\mathrm{Res}(u_h^\tau)$ represents a natural error measure for $u - u_h^\tau$. Moreover, an arbitrary Hilbert norm on Y^τ of $u - u_h^\tau$ can be associated with $\mathrm{Res}(u_h^\tau)$ by a duality argument. Furthermore, it is possible to see that the uniqueness of problem (16) implies $\mathrm{Res}(u_h^\tau) = 0$ iff u_h^τ is equal to the exact solution u. Our aim is to estimate $\mathrm{Res}(u_h^\tau)$ for u_h^τ being the solution of (11) or (12).

4.2. Reconstruction of the approximate solution with respect to time. The exact solution u of problem (4) or (16) belongs to Y^0. It is possible to see that a function w from either the space X_h^τ or Y_h^τ belongs to the space Y^0 iff w is continuous in time and satisfies the initial condition u^0. This is, in particular, not guaranteed for cG–FEM solution $u_h^\tau \in Y_h^\tau$, since it only satisfies the approximation of the initial condition Πu^0; see the definition of Y_h^τ (6). To be able to produce a posteriori error estimates we need to reconstruct the solution u_h^τ from either X_h^τ or Y_h^τ in such a way that the resulting reconstruction R_h^τ is conforming, i.e., $R_h^\tau \in Y^0$, and that $u_h^\tau \approx R_h^\tau$.

Let $r_m \in P^{q+1}$ be the right Radau polynomial on I_m, i.e., $r_m(t_{m-1}) = 1$, $r_m(t_m) = 0$, and r_m is orthogonal to $P^{q-1}(I_m)$ with respect to the $L^2(I_m)$-inner product. Alternative equivalent definition for $r_m \in P^{q+1}$ is that the zeros of r_m lie in the right Gauss–Radau quadrature nodes on I_m and the polynomial is scaled such that $r_m(t_{m-1}) = 1$. Then the polynomial reconstruction R_h^τ for both time discretizations can be determined on each interval I_m by

$$R_h^\tau(x,t) = u_h^\tau(x,t) - \{u_h^\tau\}_{m-1}(x) r_m(t), \quad x \in \Omega,\ t \in I_m. \tag{20}$$

The resulting function R_h^τ is continuous in time and satisfies the initial condition

$$R_h^\tau(x,0) = u_h^\tau(x,0) - \{u_h^\tau\}_0(x) r_1(0) = u_h^\tau(x,0) - (u_h^\tau(x,0) - u^0(x)) = u^0(x); \tag{21}$$

cf. (8). Moreover, using the integration by parts and the properties of the Radau polynomial r_m, we get

$$\begin{aligned}((R_h^\tau)', v)_{m,K} =& ((u_h^\tau)', v)_{m,K} - (r_m' \{u_h^\tau\}_{m-1}, v)_{m,K} \\ =& ((u_h^\tau)', v)_{m,K} + (r_m \{u_h^\tau\}_{m-1}, v')_{m,K} \\ & - r_m(t_m)(\{u_h^\tau\}_{m-1}, v_-^m)_K + r_m(t_{m-1})(\{u_h^\tau\}_{m-1}, v_+^{m-1})_K \\ =& ((u_h^\tau)', v)_{m,K} + (\{u_h^\tau\}_{m-1}, v_+^{m-1})_K \quad \forall v \in P^q(I_m, L^2(K)).\end{aligned} \tag{22}$$

Since the reconstruction R_h^τ is locally defined and explicit, its computation is very cheap and easy to implement. This reconstruction is also used to show equivalence among Radau IIA Runge–Kutta method, Radau collocation method, and discontinuous Galerkin method; see [2], [14], [21], [33], [41].

4.3. Reconstruction of the approximate solution with respect to space. Since $\sigma(u_h^\tau, \nabla u_h^\tau) \notin L^2(0,T,H(\mathrm{div},\Omega))$ in general, we also reconstruct for similar reasons as above the spatial fluxes of the solution in such a way that $\sigma_h^\tau \in L^2(0,T,H(\mathrm{div},\Omega))$ and $\sigma(u_h^\tau, \nabla u_h^\tau) \approx \sigma_h^\tau$. Following reconstruction is a generalization of the reconstruction from [18, Construction 3.4]. Let $\mathrm{RTN}_p(K)$ be the Raviart–Thomas–Nedelec space of order p for element $K \in \mathcal{T}_h$, i.e., $\mathrm{RTN}_p(K) = P_p(K)^d + xP_p(K)$. Let us denote the patch $\mathcal{T}_a = \bigcup_{a \in K} K$ associated with a vertex a of a mesh $\mathcal{T}_h$. Moreover, we define RTN spaces on $\mathcal{T}_a$ with zero Neumann boundary condition:

$$\begin{aligned}W_{\mathrm{RTN},p}(\mathcal{T}_a) &= \{v \in H(\mathrm{div},\mathcal{T}_a) : v|_K \in \mathrm{RTN}_p(K),\ v \cdot n = 0\ \forall e \in \partial\mathcal{T}_a\},\ a \notin \partial\Omega, \\ W_{\mathrm{RTN},p}(\mathcal{T}_a) &= \{v \in H(\mathrm{div},\mathcal{T}_a) : v|_K \in \mathrm{RTN}_p(K),\ v \cdot n = 0\ \forall e \in \partial\mathcal{T}_a \setminus \partial\Omega\},\ a \in \partial\Omega.\end{aligned}$$

Let us denote $P_*^p(\mathcal{T}_a)$ elementwise polynomial functions (possibly discontinuous) of order $\le p$ and with zero mean value for $a \notin \partial\Omega$. We denote as ψ_a the piecewise linear "hat" function associated to vertex a with $\psi_a(a) = 1$ and $\psi_a(\bar{a}) = 0$ for other vertices $\bar{a}$. Obviously, $\sum_{a \in K} \psi_a|_K = 1|_K$.

Moreover, we put

$$\begin{aligned} \xi_a^1 &= \psi_a \sigma(u_h^\tau, \nabla u_h^\tau), \\ \xi_a^2 &= \psi_a (R_h^\tau)' + \psi_a c(u_h^\tau) + \nabla \psi_a \cdot \sigma(u_h^\tau, \nabla u_h^\tau). \end{aligned} \tag{23}$$

Then we seek $\sigma_a^\tau \in P^q(I_m, W_{\mathrm{RTN},p}(\mathcal{T}_a))$ and $r_a^\tau \in P^q(I_m, P_*^p(\mathcal{T}_a))$ as solutions of the following local mixed finite element problem:

$$\begin{aligned} (\sigma_a^\tau, v)_{\mathcal{T}_a,m} - (r_a^\tau, \nabla \cdot v)_{\mathcal{T}_a,m} &= (\xi_a^1, v)_{\mathcal{T}_a,m} \quad \forall v \in P^q(I_m, W_{\mathrm{RTN},p}(\mathcal{T}_a)), \\ (\nabla \cdot \sigma_a^\tau, \varphi)_{\mathcal{T}_a,m} &= (\xi_a^2, \varphi)_{\mathcal{T}_a,m} \quad \forall \varphi \in P^q(I_m, P_*^p(\mathcal{T}_a)). \end{aligned} \tag{24}$$

It is possible to observe that the second relation in (24) is satisfied for test function φ constant in space, since then both sides of the relation are equal to zero due to the special choice of the right-hand side. Therefore, the second relation in (24) holds elementwise for arbitrary test function $\varphi \in P^q(I_m, P^p(K))$. Problem (24) represents a local Neumann mixed finite element problem and the existence and uniqueness of problem (24) follows from [39, Theorem 10.1 and section 14]. Finally, we define σ_h^τ separately for each I_m by

$$\sigma_h^\tau|_{K \times I_m} = \sum_{a \in K} \sigma_a^\tau|_K. \tag{25}$$

Summing the second relation from (24) over all vertices $a \in K$, we get

$$(\nabla \cdot \sigma_h^\tau, \varphi)_{K,m} = ((R_h^\tau)' + c(u_h^\tau), \varphi)_{K,m} \quad \forall \varphi \in P^q(I_m, P^p(K)). \tag{26}$$

This relation is known as *local flux equilibration* and represents an important tool for deriving a posteriori error upper bound.

4.4. Upper bound. In this section we prove an a posteriori upper bound for $\mathrm{Res}(u_h^\tau)$; cf. (19). We estimate $\mathrm{Res}(u_h^\tau)$ in terms of data σ and c, discrete solution u_h^τ, and functions R_h^τ and σ_h^τ that are cheaply (locally) computable from the discrete solution u_h^τ. Let us assume $v \in Y^\tau$. We divide the numerator of $\mathrm{Res}(u_h^\tau)$ as

$$\begin{aligned} &\sum_{K,m} \left(((u_h^\tau)' + c(u_h^\tau), v)_{K,m} + (\sigma(u_h^\tau, \nabla u_h^\tau), \nabla v) + (\{u_h^\tau\}_{m-1}, v_+^{m-1})_K \right) \\ &= \underbrace{\sum_{K,m} ((R_h^\tau)' + c(u_h^\tau) - \nabla \cdot \sigma_h^\tau, v)_{K,m}}_{\chi_1} - \underbrace{\sum_{K,m} (\sigma_h^\tau - \sigma(u_h^\tau, \nabla u_h^\tau), \nabla v)_{K,m}}_{\chi_2} \\ &\quad \underbrace{- \sum_{K,m} \left((R_h^\tau - u_h^\tau)', v)_{K,m} - (\{u_h^\tau\}_{m-1}, v_+^{m-1})_K \right)}_{\chi_3}. \end{aligned} \tag{27}$$

We estimate individual terms χ_1, χ_2 and χ_3. Let $v_{K,m}$ be the L^2-orthogonal projection of $v \in Y^\tau$ on $P^0(K \times I_m)$. Using (26) and the space-time scaled Poincaré

inequality from [11, Lemma 2.2], we obtain

$$
\begin{aligned}
\chi_1 &= \sum_{K,m} ((R_h^\tau)' + c(u_h^\tau) - \nabla\cdot\sigma_h^\tau, v)_{K,m} \\
&= \sum_{K,m} ((R_h^\tau)' + c(u_h^\tau) - \nabla\cdot\sigma_h^\tau, v - v_{K,m})_{K,m} \\
&\le \sum_{K,m} C_P \|(R_h^\tau)' + c(u_h^\tau) - \nabla\cdot\sigma_h^\tau\|_{K,m} (h_K^2 \|\nabla v\|_{K,m}^2 + \tau_m^2 \|v'\|_{K,m}^2)^{1/2} \\
&= \sum_{K,m} C_P d_{K,m} \|(R_h^\tau)' + c(u_h^\tau) - \nabla\cdot\sigma_h^\tau\|_{K,m} \|v\|_{Y^\tau,K,m},
\end{aligned} \tag{28}
$$

where C_P is the constant from the Poincaré inequality. In our case, where the space mesh consists of simplices, which are convex, we can bound $C_P \le 1/\pi$; see, e.g., [35]. For more detailed discussion on the estimates for the Poincaré constant, see, e.g., [42].

For the second term we get

$$
\begin{aligned}
\chi_2 &= \sum_{K,m} (\sigma_h^\tau - \sigma(u_h^\tau, \nabla u_h^\tau), \nabla v)_{K,m} \\
&\le \sum_{K,m} \frac{d_{K,m}}{h_K} \|\sigma_h^\tau - \sigma(u_h^\tau, \nabla u_h^\tau)\|_{K,m} \frac{h_K}{d_{K,m}} \|\nabla v\|_{K,m}.
\end{aligned} \tag{29}
$$

For the last term we apply the integration by parts and (20) to get

$$
\begin{aligned}
\chi_3 &= \sum_{K,m} \left(((R_h^\tau - u_h^\tau)', v)_{K,m} - (\{u_h^\tau\}_{m-1}, v_+^{m-1})_K \right) \\
&= -\sum_{K,m} (R_h^\tau - u_h^\tau, v')_{K,m} \le \sum_{K,m} \frac{d_{K,m}}{\tau_m} \|R_h^\tau - u_h^\tau\|_{K,m} \frac{\tau_m}{d_{K,m}} \|v'\|_{K,m}.
\end{aligned} \tag{30}
$$

Combining the partial results (28), (29), and (30) we estimate the numerator of

$\mathrm{Res}(u_h^\tau)$ by the Cauchy inequalities

$$
\begin{aligned}
(31)\qquad & \sum_{K,m} \left(((u_h^\tau)', v)_{K,m} + a_{K,m}(u_h^\tau, v)_{K,m} + (\{u_h^\tau\}_{m-1}, v_+^{m-1})_K \right) \\
& \le \sum_{K,m} \Bigg(C_P d_{K,m} \|(R_h^\tau)' + c(u_h^\tau) - \nabla \cdot \sigma_h^\tau\|_{K,m} \|v\|_{Y^\tau,K,m} \\
& \quad + \frac{d_{K,m}}{h_K} \|\sigma_h^\tau - \sigma(u_h^\tau, \nabla u_h^\tau)\|_{K,m} \frac{h_K}{d_{K,m}} \|\nabla v\|_{K,m} \\
& \quad + \frac{d_{K,m}}{\tau_m} \|R_h^\tau - u_h^\tau\|_{K,m} \frac{\tau_m}{d_{K,m}} \|v'\|_{K,m} \Bigg) \\
& \le \sum_{K,m} \Bigg(C_P d_{K,m} \|(R_h^\tau)' + c(u_h^\tau) - \nabla \cdot \sigma_h^\tau\|_{K,m} \\
& \quad + \sqrt{\frac{d_{K,m}^2}{h_K^2} \|\sigma_h^\tau - \sigma(u_h^\tau, \nabla u_h^\tau)\|_{K,m}^2 + \frac{d_{K,m}^2}{\tau_m^2} \|R_h^\tau - u_h^\tau\|_{K,m}^2} \Bigg) \|v\|_{Y^\tau,K,m} \\
& \le \Bigg[\sum_{K,m} \Bigg(C_P d_{K,m} \|(R_h^\tau)' + c(u_h^\tau) - \nabla \cdot \sigma_h^\tau\|_{K,m} \\
& \quad + \sqrt{\frac{d_{K,m}^2}{h_K^2} \|\sigma_h^\tau - \sigma(u_h^\tau, \nabla u_h^\tau)\|_{K,m}^2 + \frac{d_{K,m}^2}{\tau_m^2} \|R_h^\tau - u_h^\tau\|_{K,m}^2} \Bigg)^2 \Bigg]^{1/2} \|v\|_{Y^\tau}.
\end{aligned}
$$

Let us denote partial estimators

$$
\begin{aligned}
(32)\qquad \eta_{R,K,m} &= d_{K,m} \|(R_h^\tau)' + c(u_h^\tau) - \nabla \cdot \sigma_h^\tau\|_{K,m}, \\
\eta_{S,K,m} &= \frac{d_{K,m}}{h_K} \|\sigma_h^\tau - \sigma(u_h^\tau, \nabla u_h^\tau)\|_{K,m}, \\
\eta_{T,K,m} &= \frac{d_{K,m}}{\tau_m} \|R_h^\tau - u_h^\tau\|_{K,m}, \quad K \in \mathcal{T}_h,\ m = 1, \dots, r.
\end{aligned}
$$

The forthcoming theorem describing the upper bound to $\mathcal{E} = \mathrm{Res}(u_h^\tau)$ is a direct consequence of (31).

THEOREM 5. *Let $u \in Y$ be the solution of* (4)*, and let u_h^τ be either the cG–FEM solution defined by* (11) *or the dG–FEM solution defined by* (12)*. Let R_h^τ and σ_h^τ be reconstructions obtained from u_h^τ by* (20)*, respectively,* (24) *and* (25)*. Then*

$$
(33)\qquad \mathcal{E}^2 \le \eta^2 := \sum_{K,m} \left(C_P\, \eta_{R,K,m} + (\eta_{S,K,m}^2 + \eta_{T,K,m}^2)^{1/2} \right)^2,
$$

where the constant $C_P \le 1/\pi$.

5. Local efficiency. The goal of this section is to show that local individual terms $\eta_{R,K,m}$, $\eta_{S,K,m}$, and $\eta_{T,K,m}$ from a posteriori estimate (33) are locally efficient. It means that they provide local lower bounds to the error measure up to some generic constant $C > 0$ that may depend on constants coming from the original continuous problem (the size of the domain Ω, etc.) or on the constants coming from the discretization (mesh shape regularity constant, polynomial degrees p and q, etc.). However, this constant is independent of the exact solution u, discrete solution u_h^τ,

and space-time mesh sizes h_K and τ_m. We will denote dependence of the estimate up to this generic constant by $\lesssim$.

To be able to apply the result in a local way, we need the following notation. Let $\mathcal{T}_e$ be a patch of elements sharing common face e. Let $\mathcal{T}_K$ be a patch consisting of elements sharing at least a vertex with K and $\mathcal{T}_K^2$ be a union of patches $\mathcal{T}_{K'}$, where $K' \subset \mathcal{T}_K$, i.e.,

$$\mathcal{T}_e = \bigcup_{e \cap K' \neq \emptyset} K', \qquad \mathcal{T}_K = \bigcup_{K \cap K' \neq \emptyset} K', \qquad \mathcal{T}_K^2 = \bigcup_{K' \subset \mathcal{T}_K} \mathcal{T}_{K'}.$$

Let $M \subset \overline{\Omega}$, e.g., $M = K$ or $M = \mathcal{T}_K$. We define a local version of the space Y^τ by

$$Y^\tau_{M,m} = \{v \in Y^\tau : \operatorname{supp}(v) \subset \overline{M \times I_m}\}, \tag{34}$$

and a local version of $\operatorname{Res}(w)$,

$$\operatorname{Res}_{M,m}(w) = \sup_{0 \neq v \in Y^\tau_{M,m}} \frac{1}{\|v\|_{Y^\tau}} \sum_{K,m} b_{K,m}(w, v). \tag{35}$$

Typically, we use $\operatorname{Res}_{K,m}(u_h^\tau)$ or $\operatorname{Res}_{\mathcal{T}_K,m}(u_h^\tau)$. Using the shape regularity of the mesh it is possible to see that

$$\sum_{K,m} \operatorname{Res}_{K,m}(u_h^\tau) \leq \sum_{K,m} \operatorname{Res}_{\mathcal{T}_K,m}(u_h^\tau) \leq \sum_{K,m} \operatorname{Res}_{\mathcal{T}_K^2,m}(u_h^\tau) \lesssim \operatorname{Res}(u_h^\tau). \tag{36}$$

For the purpose of the effectivity analysis we approximate $\sigma(u_h^\tau, \nabla u_h^\tau)$ and $c(u_h^\tau)$, since these terms are not polynomials in general even if u_h^τ is. We define $\bar{c} = \bar{c}(u_h^\tau) \in P^q(I_m, P^p(K))$ on $K \times I_m$ by

$$(\bar{c}, v)_{K,m} = (c(u_h^\tau), v)_{K,m} \quad \forall v \in P^q(I_m, P^p(K)). \tag{37}$$

For vertex $a \in K$ we define $\bar{\sigma}_a = \bar{\sigma}_a(u_h^\tau, \nabla u_h^\tau) \in P^q(I_m, \mathrm{RTN}_p(K))$ by

$$\begin{aligned} (\bar{\sigma}_a \cdot n, v)_{e,m} &= (\psi_a \langle \sigma(u_h^\tau, \nabla u_h^\tau) \rangle \cdot n, v)_{e,m} \quad \forall v \in P^q(I_m, P^p(e)),\ e \subset K, \\ (\bar{\sigma}_a, v)_{K,m} &= (\psi_a \sigma(u_h^\tau, \nabla u_h^\tau), v)_{K,m} \quad \forall v \in P^q(I_m, P^{p-1}(K)^d), \end{aligned} \tag{38}$$

where $\langle \sigma(u_h^\tau, \nabla u_h^\tau) \rangle$ denotes the arithmetic mean value of $\sigma(u_h^\tau, \nabla u_h^\tau)$ on the edge, and finally $\bar{\sigma}|_{K \times I_m} = \sum_{a \in K} \bar{\sigma}_a$.

We often encounter $\bar{\sigma} - \sigma(u_h^\tau, \nabla u_h^\tau)$ or $\bar{c} - c(u_h^\tau)$ in the forthcoming analysis. These terms are a generalization to the classical oscillation term that is usually considered as negligible. For simplicity, we will cover these terms into the efficiency analysis under following assumption for $\bar{\sigma}_a$:

$$\begin{aligned} &d_{K,m} \|\nabla \cdot \bar{\sigma}_a - \nabla \cdot (\psi_a \sigma(u_h^\tau, \nabla u_h^\tau))\|_{K,m} + \frac{d_{K,m}}{h_K} \|\bar{\sigma}_a - \psi_a \sigma(u_h^\tau, \nabla u_h^\tau)\|_{K,m} \\ &\lesssim \operatorname{Res}_{\mathcal{T}_K,m}(u_h^\tau), \quad K \in \mathcal{T}_h,\ m = 1, \dots, r, \end{aligned} \tag{39}$$

and for $\bar{c}$

$$d_{K,m} \|\bar{c} - c(u_h^\tau)\|_{K,m} \lesssim \operatorname{Res}_{K,m}(u_h^\tau), \quad K \in \mathcal{T}_h,\ m = 1, \dots, r. \tag{40}$$

A direct consequence of (39) and triangle inequality is

$$d_{K,m}\|\nabla\cdot\bar\sigma-\nabla\cdot\sigma(u_h^\tau,\nabla u_h^\tau)\|_{K,m}+\frac{d_{K,m}}{h_K}\|\bar\sigma-\sigma(u_h^\tau,\nabla u_h^\tau)\|_{K,m}\\ \lesssim \mathrm{Res}_{\mathcal{T}_K,m}(u_h^\tau),\quad K\in\mathcal{T}_h,\ m=1,\dots,r. \tag{41}$$

These assumptions can be justified in a similar way as in [11, Assumption 4.1].

We will divide the proof of the local efficiency of the individual partial estimators $\eta_{R,K,m}$, $\eta_{S,K,m}$, and $\eta_{T,K,m}$ into the next auxiliary lemmas.

LEMMA 6. *Let R_h^τ be defined by* (20). *Let $\bar\sigma$ be defined by* (38) *and $\bar c$ by* (37) *satisfying* (39) *and* (40), *respectively. Then*

$$\eta_{T,K,m}\lesssim \mathrm{Res}_{\mathcal{T}_K,m}(u_h^\tau),\quad K\in\mathcal{T}_h,\ m=1,\dots,r. \tag{42}$$

Proof. Let us construct a suitable test function $w\in Y^\tau$ associated with $K\times I_m$. Due to the piecewise continuous nature in time of the space Y^τ (see (14)), it is sufficient to define the function w on time layer $\Omega\times I_m$ only

$$w(x,t)=\frac{d_{K,m}^2}{\tau_m}\{u_h^\tau\}_{m-1}(x)\chi_K(x)\Phi_m(t),\quad x\in\Omega,\ t\in I_m, \tag{43}$$

where χ_K is a polynomial bubble function on the element K such that $\|\chi_K\|_{L^\infty(\Omega)}=1$ and Φ_m is a Legendre polynomial of degree $q+1$ orthogonal to polynomials of degree q on I_m such that $\Phi_m(t_{m-1})=1$. We shall point out that the resulting function w (after the extension by zero to the other time layers) satisfies $\mathrm{supp}(w)=K\times I_m$. Then

$$\begin{aligned}\mathrm{Res}_{K,m}&\ge\frac{\sum_{K,m}((u_h^\tau)'+c(u_h^\tau),w)_{K,m}+(\sigma(u_h^\tau,\nabla u_h^\tau),\nabla w)_{K,m}+(\{u_h^\tau\}_{m-1},w_+^{m-1})_K}{\|w\|_{Y^\tau,K,m}}\\ &\ge\frac{((u_h^\tau)'+\bar c,w)_{K,m}+(\bar\sigma,\nabla w)_{K,m}+(\{u_h^\tau\}_{m-1},w_+^{m-1})_K}{\|w\|_{Y^\tau,K,m}}\\ &\quad+\frac{(c(u_h^\tau)-\bar c,w)_{K,m}+(\sigma(u_h^\tau,\nabla u_h^\tau)-\bar\sigma,\nabla w)_{K,m}}{\|w\|_{Y^\tau,K,m}}.\end{aligned} \tag{44}$$

We can estimate these terms individually. At first let us estimate $\|w\|_{Y^\tau,K,m}$. Using inverse inequality and (43) we get

$$\|w\|_{Y^\tau,K,m}^2=\frac{h_K^2\|\nabla w\|_{K,m}^2+\tau_m^2\|w'\|_{K,m}^2}{d_{K,m}^2}\lesssim\frac{\|w\|_{K,m}^2}{d_{K,m}^2}\\ \le\frac{d_{K,m}^2}{\tau_m^2}\|\{u_h^\tau\}_{m-1}\|_K^2\int_{I_m}\Phi_m^2(t)\mathrm{d}t\lesssim\frac{d_{K,m}^2}{\tau_m}\|\{u_h^\tau\}_{m-1}\|_K^2. \tag{45}$$

The term containing $c(u_h^\tau)-\bar c$ can be estimated by the Poincaré inequality and (40),

$$\frac{(c(u_h^\tau)-\bar c,w)_{K,m}}{\|w\|_{Y^\tau,K,m}}=\frac{(c(u_h^\tau)-\bar c,w-w_{K,m})_{K,m}}{\|w\|_{Y^\tau,K,m}}\\ \lesssim d_{K,m}\|c(u_h^\tau)-\bar c\|_{K,m}\frac{\sqrt{h_K^2\|\nabla w\|_{K,m}^2+\tau_m^2\|w'\|_{K,m}^2}}{d_{K,m}\|w\|_{Y^\tau,K,m}}\lesssim\mathrm{Res}_{K,m}(u_h^\tau), \tag{46}$$

where $w_{K,m}$ is projection of w on constants $P^0(K \times I_m)$. The term containing $\sigma(u_h^\tau, \nabla u_h^\tau) - \bar{\sigma}$ can be estimated with the aid of (41),

$$\begin{aligned}(47)\qquad &\frac{(\sigma(u_h^\tau, \nabla u_h^\tau) - \bar{\sigma}, \nabla w)_{K,m}}{\|w\|_{Y^\tau,K,m}} \\ &\le \frac{d_{K,m}}{h_K}\|\sigma(u_h^\tau, \nabla u_h^\tau) - \bar{\sigma}\|_{K,m}\frac{h_K\|\nabla w\|_{K,m}}{d_{K,m}\|w\|_{Y^\tau,K,m}} \lesssim \mathrm{Res}_{\mathcal{T}_K,m}(u_h^\tau).\end{aligned}$$

Using orthogonality of w to polynomials of degree q, equivalence of norms on finite-dimensional spaces, and (45), we get

$$\begin{aligned}(48)\qquad &\frac{((u_h^\tau)' + \bar{c}, w)_{K,m} + (\bar{\sigma}, \nabla w)_{K,m} + (\{u_h^\tau\}_{m-1}, w_+^{m-1})_K}{\|w\|_{Y^\tau,K,m}} \\ &= \frac{(\{u_h^\tau\}_{m-1}, w_+^{m-1})_K}{\|w\|_{Y^\tau,K,m}} = \frac{d_{K,m}^2}{\tau_m}\frac{(\{u_h^\tau\}_{m-1}, \chi_K\{u_h^\tau\}_{m-1})_K}{\|w\|_{Y^\tau,K,m}} \\ &\gtrsim \frac{d_{K,m}^2}{\tau_m}\frac{\|\{u_h^\tau\}_{m-1}\|_K^2}{\|w\|_{Y^\tau,K,m}} \gtrsim \frac{d_{K,m}}{\sqrt{\tau_m}}\|\{u_h^\tau\}_{m-1}\|_K.\end{aligned}$$

Finally, (20), (44), (46), (47), and (48) give

$$\begin{aligned}(49)\qquad &\frac{d_{K,m}}{\tau_m}\|R_h^\tau - u_h^\tau\|_{K,m} = \frac{d_{K,m}}{\tau_m}\|\{u_h^\tau\}_{m-1}r_m\|_{K,m} \\ &= \frac{d_{K,m}}{\tau_m}\|\{u_h^\tau\}_{m-1}\|_K\sqrt{\int_{I_m} r_m(t)^2\mathrm{d}t} \lesssim \frac{d_{K,m}}{\sqrt{\tau_m}}\|\{u_h^\tau\}_{m-1}\|_K \\ &\lesssim \mathrm{Res}_{K,m}(u_h^\tau) + \frac{((\bar{c} - c(u_h^\tau), w)_{K,m})}{\|w\|_{Y^\tau,K,m}} + \frac{(\bar{\sigma} - \sigma(u_h^\tau, \nabla u_h^\tau), \nabla w)_{K,m}}{\|w\|_{Y^\tau,K,m}} \\ &\lesssim \mathrm{Res}_{\mathcal{T}_K,m}(u_h^\tau). \qquad \square\end{aligned}$$

LEMMA 7. *Let σ_h^τ be defined by* (24) *and* (25)*, R_h^τ be defined by* (20)*. Let $\bar{\sigma}$ be defined by* (38) *and $\bar{c}$ by* (37) *satisfying* (39) *and* (40)*, respectively. Then for $K \in \mathcal{T}_h$, $m = 1, \dots, r$,*

$$\eta_{R,K,m} \lesssim d_{K,m}\|(R_h^\tau)' + \bar{c} - \nabla\cdot\bar{\sigma}\|_{K,m} + \eta_{S,K,m} + \mathrm{Res}_{\mathcal{T}_K,m}(u_h^\tau).$$

Proof. We apply triangle inequality, inverse inequality, (41) and (40) and we get

(50)

$$\begin{aligned}&d_{K,m}\|(R_h^\tau)' + c(u_h^\tau) - \nabla\cdot\sigma_h^\tau\|_{K,m} \\ &\le d_{K,m}\|(R_h^\tau)' + \bar{c} - \nabla\cdot\bar{\sigma}\|_{K,m} + d_{K,m}\|\bar{c} - c(u_h^\tau)\|_{K,m} + d_{K,m}\|\nabla\cdot\sigma_h^\tau - \nabla\cdot\bar{\sigma}\|_{K,m} \\ &\lesssim d_{K,m}\|(R_h^\tau)' + \bar{c} - \nabla\cdot\bar{\sigma}\|_{K,m} + d_{K,m}\|\bar{c} - c(u_h^\tau)\|_{K,m} + \frac{d_{K,m}}{h_K}\|\sigma_h^\tau - \bar{\sigma}\|_{K,m} \\ &\lesssim d_{K,m}\|(R_h^\tau)' + \bar{c} - \nabla\cdot\bar{\sigma}\|_{K,m} + \frac{d_{K,m}}{h_K}\|\sigma_h^\tau - \sigma(u_h^\tau, \nabla u_h^\tau)\|_{K,m} + \mathrm{Res}_{\mathcal{T}_K,m}(u_h^\tau). \qquad \square\end{aligned}$$

We recall that the estimates in two previous lemmas depend on the polynomial approximation degree due to the use of the inverse inequality.

LEMMA 8. *Let R_h^τ be defined by* (20)*. Let $\bar{\sigma}$ be defined by* (38) *and $\bar{c}$ by* (37) *satisfying* (39) *and* (40)*, respectively. Then*

$$(51)\qquad d_{K,m}\|(R_h^\tau)' + \bar{c} - \nabla\cdot\bar{\sigma}\|_{K,m} \lesssim \mathrm{Res}_{\mathcal{T}_K,m}(u_h^\tau), \quad K \in \mathcal{T}_h,\ m = 1, \dots, r.$$

Proof. The proof is essentially similar to the proof in [46, Theorem 4.10] for stationary problems, where the resulting oscillation terms need to be estimated by (41) and (40). ∎

For the choice of σ_h^τ defined by (24) and (25) it is possible to show the following lemma.

LEMMA 9. *Let σ_h^τ is defined by* (24) *and* (25). *Let $\bar{\sigma}$ be defined by* (38) *and $\bar{c}$ by* (37) *satisfying* (39) *and* (40), *respectively. Then*

$$\eta_{S,K,m} \lesssim \mathrm{Res}_{\mathcal{T}_K^2,m}(u_h^\tau), \quad K \in \mathcal{T}_h, m = 1, \dots, r. \tag{52}$$

Proof. The proof mimics the stationary version proved in [18, Theorem 3.12]. It is possible to divide the estimate by triangle inequality

$$\|\sigma_h^\tau - \sigma(u_h^\tau, \nabla u_h^\tau)\|_{K,m} \le \|\sigma_h^\tau - \bar{\sigma}\|_{K,m} + \|\bar{\sigma} - \sigma(u_h^\tau, \nabla u_h^\tau)\|_{K,m}. \tag{53}$$

According to the assumptions on $\bar{\sigma}$ we can estimate the last term on the right-hand side by (41). Since

$$\|\sigma_h^\tau - \bar{\sigma}\|_{K,m} \le \sum_{a \in K} \|\sigma_a^\tau - \bar{\sigma}_a\|_{K,m} \le \sum_{a \in K} \|\sigma_a^\tau - \bar{\sigma}_a\|_{\mathcal{T}_a,m}, \tag{54}$$

it remains to estimate $\|\sigma_a^\tau - \bar{\sigma}_a\|_{\mathcal{T}_a,m}$. Let us denote $\check{\sigma}_a = \sigma_a^\tau - \bar{\sigma}_a$that satisfies

$$\begin{aligned} (\check{\sigma}_a, v)_{\mathcal{T}_a,m} - (\bar{r}_a, \nabla \cdot v)_{\mathcal{T}_a,m} &= (\xi_a^1 - \bar{\sigma}_a, v)_{\mathcal{T}_a,m} \quad \forall v \in P^q(I_m, W_{\mathrm{RTN},p}(\mathcal{T}_a)), \\ (\nabla \cdot \check{\sigma}_a, \varphi)_{\mathcal{T}_a,m} &= (\xi_a^2 - \nabla \cdot \bar{\sigma}_a, \varphi)_{\mathcal{T}_a,m} \quad \forall \varphi \in P^q(I_m, P_*^p(\mathcal{T}_a)). \end{aligned} \tag{55}$$

It is possible to see that the right-hand side terms can be replaced by $\Pi_1(\xi_a^1 - \bar{\sigma}_a)$ and $\Pi_2(\xi_a^2 - \nabla \cdot \bar{\sigma}_a)$, where projections Π_1 and Π_2 are L^2 orthogonal projections on $P^q(I_m, W_{\mathrm{RTN},p}(\mathcal{T}_a))$ and $P^q(I_m, P_*^p(\mathcal{T}_a))$, respectively. According to [39, section 2] it is possible to show

$$\|\check{\sigma}_a\|_{\mathcal{T}_a,m} \le \|v\|_{\mathcal{T}_a,m} \quad \forall v \in P^q(I_m, W_{\mathrm{RTN},p}(\mathcal{T}_a)),\ \nabla \cdot v = \Pi_2(\xi_a^2 - \nabla \cdot \bar{\sigma}_a). \tag{56}$$

Let us define space $H_*^1(\mathcal{T}_a)$ as

$$\begin{aligned} H_*^1(\mathcal{T}_a) &= \{v \in H^1(\mathcal{T}_a) : (v,1)_{\mathcal{T}_a} = 0\} \quad a \text{ interior vertex}, \\ H_*^1(\mathcal{T}_a) &= \{v \in H^1(\mathcal{T}_a) : v = 0 \text{ on } \partial\mathcal{T}_a \cap \partial\Omega\} \quad a \text{ boundary vertex}. \end{aligned} \tag{57}$$

Then we define $r_a^N \in P^q(I_m, H_*^1(\mathcal{T}_a))$ by

$$(\nabla r_a^N, \nabla v)_{\mathcal{T}_a,m} = (\Pi_2(\xi_a^2 - \nabla \cdot \bar{\sigma}_a), v)_{\mathcal{T}_a,m} + (\Pi_1(\xi_a^1 - \bar{\sigma}_a), \nabla v)_{\mathcal{T}_a,m}. \tag{58}$$

According to [6, Theorem 7], there exists $v_a^\tau \in P^q(I_m, W_{\mathrm{RTN},p}(\mathcal{T}_a))$ such that

$$\|v_a^\tau\|_{\mathcal{T}_a,m} \lesssim \|\nabla r_a^N\|_{\mathcal{T}_a,m}, \qquad \nabla \cdot v_a^\tau = \Pi_2(\xi_a^2 - \nabla \cdot \bar{\sigma}_a); \tag{59}$$

see also the proof of [18, Corollary 3.16]. From this follows

$$\|\check{\sigma}_a\|_{\mathcal{T}_a,m} \le \|v_a^\tau\|_{\mathcal{T}_a,m} \lesssim \|\nabla r_a^N\|_{\mathcal{T}_a,m}. \tag{60}$$

It remains to estimate $\|\nabla r_a^N\|_{\mathcal{T}_a,m}$. We employ either Poincaré or Friedrichs inequality depending on whether a is an interior or boundary vertex. Then

$$\begin{aligned}\|\nabla r_a^N\|^2_{\mathcal{T}_a,m} \leq & \|\nabla\psi_a\cdot\sigma(u_h^\tau,\nabla u_h^\tau)+\psi_a(R_h^\tau)'+\psi_a c(u_h^\tau)-\nabla\cdot\bar\sigma_a\|_{\mathcal{T}_a,m}\|r_a^N\|_{\mathcal{T}_a,m}\\ &+\|\psi_a\sigma(u_h^\tau,\nabla u_h^\tau)-\bar\sigma_a\|_{\mathcal{T}_a,m}\|\nabla r_a^N\|_{\mathcal{T}_a,m}\\ &\lesssim (h_K\|\nabla\cdot(\psi_a\sigma(u_h^\tau,\nabla u_h^\tau))-\nabla\cdot\bar\sigma_a\|_{\mathcal{T}_a,m}+\|\psi_a\sigma(u_h^\tau,\nabla u_h^\tau)-\bar\sigma_a\|_{\mathcal{T}_a,m}\\ &+h_K\|\psi_a((R_h^\tau)'+c(u_h^\tau)-\nabla\cdot\sigma(u_h^\tau,\nabla u_h^\tau))\|_{\mathcal{T}_a,m})\|\nabla r_a^N\|_{\mathcal{T}_a,m}.\end{aligned}$$

Applying (39) and (40) and Lemma 8, we get

$$\begin{aligned}\|\nabla r_a^N\|_{\mathcal{T}_a,m} &\lesssim h_K\|(R_h^\tau)'+\bar c+\nabla\cdot\bar\sigma\|_{\mathcal{T}_a,m}+h_K\|\nabla\cdot\sigma(u_h^\tau,\nabla u_h^\tau)-\nabla\cdot\bar\sigma\|_{\mathcal{T}_a,m}\\ &\quad+h_K\|\nabla\cdot(\psi_a\sigma(u_h^\tau,\nabla u_h^\tau))-\nabla\cdot\bar\sigma_a\|_{\mathcal{T}_a,m}+\|\psi_a\sigma(u_h^\tau,\nabla u_h^\tau)-\bar\sigma_a\|_{\mathcal{T}_a,m}\\ &\lesssim \frac{h_K}{d_{K,m}}\operatorname{Res}_{\mathcal{T}_K^2,m}(u_h^\tau).\end{aligned}$$

∎

The direct application of Lemmas 6–9 gives us local efficiency estimates of individual terms from a posteriori error estimate (33).

THEOREM 10. *Let σ_h^τ be defined by* (24) *and* (25) *and R_h^τ be defined by* (20)*. Let $\bar\sigma$ be defined by* (38) *and $\bar c$ by* (37) *satisfying* (39) *and* (40)*, respectively. Then*

$$\begin{aligned}\eta_{R,K,m}&=d_{K,m}\|(R_h^\tau)'+c(u_h^\tau)-\nabla\cdot\sigma_h^\tau\|_{K,m}\lesssim\operatorname{Res}_{\mathcal{T}_K^2,m}(u_h^\tau),\\ \eta_{S,K,m}&=\frac{d_{K,m}}{h_K}\|\sigma_h^\tau-\sigma(u_h^\tau,\nabla u_h^\tau)\|_{K,m}\lesssim\operatorname{Res}_{\mathcal{T}_K^2,m}(u_h^\tau),\\ \eta_{T,K,m}&=\frac{d_{K,m}}{\tau_m}\|R_h^\tau-u_h^\tau\|_{K,m}\lesssim\operatorname{Res}_{\mathcal{T}_K,m}(u_h^\tau),\quad K\in\mathcal{T}_h, m=1,\dots,r.\end{aligned}\tag{61}$$

Combining Theorem 10 with (36) implies global efficiency of a posteriori error estimate (33).

Let us note that for linear problems, where no approximation by $\bar\sigma$ and $\bar c$ is necessary, the localness can be improved and efficiency estimates from Theorem 10 depend only on elements K or patches $\mathcal{T}_K$ (basically they are one layer better).

6. A posteriori error analysis of cG–DGFEM and dG–DGFEM discretizations. We extend the previous results to discontinuous Galerkin space discretization and generalize the discretization to the situation of varying meshes for each time interval I_m.

6.1. DGFEM space discretization. For the purpose of the space discontinuous Galerkin discretization we assume similarly as in [11]

$$\sigma(u,\nabla u)=\mathcal{K}(u)\nabla u-F(u),\tag{62}$$

i.e., the complete flux σ contains a diffusive flux $\mathcal{K}(u)\nabla u$, where $\mathcal{K}\in L^\infty(\mathbb{R})^{d\times d}$, and a convective flux $F(u)$, where $F\in C^1(\mathbb{R})^d$. We restrict ourselves here to form (62) since it allows for a simple derivation of the primal DGFEM formulation from the mixed formulation based on the numerical fluxes; see [4].

We use the same notation for the triangulation $\mathcal{T}_h$, elements K, edges e, local mesh size h_K, normals with fixed orientation n, normals n_K oriented outward with respect to element K as in the conforming space discretization; see section 3.

Moreover, we assume that the mesh $\mathcal{T}_h$ may differ for different time intervals I_m. We denote the dependence of the mesh on the time interval by the superscript, i.e., we

use $\mathcal{T}_h^m$ for the mesh in interval I_m. Similarly as for FEM, we define the discontinuous Galerkin version of V_h

$$V_{h,\mathrm{disc}}^m = \{v \in L^2(\Omega) : v|_K \in P^p(K), K \in \mathcal{T}_h^m\}, \tag{63}$$

and the corresponding version of X_h^τ,

$$X_{h,\mathrm{disc}}^\tau = \{v : v|_{I_m} \in P^q(I_m, V_{h,\mathrm{disc}}^m)\}. \tag{64}$$

Since the space $V_{h,\mathrm{disc}}^m$ is generally different for each time slab, it is not possible to assume continuity in time for the generalization of space Y_h^τ. We use the same idea for each transition between time slabs as in the definition of the original space Y_h^τ for the transition at the initial time t_0,

$$Y_{h,\mathrm{disc}}^\tau = \{v : v|_{I_m} \in P^{q+1}(I_m, V_{h,\mathrm{disc}}^m), v_+^{m-1} = \Pi^m v_-^{m-1}\}, \tag{65}$$

where Π^m is the L^2-orthogonal projection on $V_{h,\mathrm{disc}}^m$.

We define one-sided values, jumps, and mean values for $v \in V_{h,\mathrm{disc}}^m$ on the inner edges with the unit normal n,

$$\begin{aligned} v_L(x) &= \lim_{s\to 0+} v(x - ns), \qquad v_R(x) = \lim_{s\to 0+} v(x + ns), \\ [v] &= (v_L - v_R)n, \qquad \langle v\rangle = (v_L + v_R)/2. \end{aligned} \tag{66}$$

The value $[v]$ is independent on the orientation of n. For the boundary edges we define $v_R = 0$, $\langle v\rangle = v_L$, and $[v] = v_L n$, where n is the unit outer normal to Ω.

We define the discontinuous Galerkin version of $a_{K,m}$ for $K \in \mathcal{T}_h^m$, $m = 1, \dots, r$ by

$$A_{K,m}(u, v) = a_{K,m}(u, v) - (\hat{\sigma} \cdot n_K, v)_{\partial K,m} + ((\hat{\mathcal{K}} - \mathcal{K}(u)u)n_K, \nabla v)_{\partial K,m}, \tag{67}$$

where the numerical fluxes $\hat{\sigma} = \hat{\sigma}(u, \nabla u)$ and $\hat{\mathcal{K}} = \widehat{\mathcal{K}(u)u}$ approximate $\sigma(u, \nabla u)$ and $\mathcal{K}(u)u$ on ∂K, respectively, and n_K is the unit outer normal to $K \in \mathcal{T}_h^m$. We shall point out that the volume terms containing $\sigma(u, \nabla u)$ and $c(u)$ are included in $a_{K,m}$ form; cf. (10). We assume that both of these fluxes are consistent and that the numerical flux $\hat{\sigma}$ is conservative; see, e.g., [4, section 3.1]. A possible definition of these numerical fluxes on ∂K is the following:

$$\begin{aligned} \hat{\mathcal{K}} &= \langle \mathcal{K}(u)u\rangle + \theta[\mathcal{K}(u)u] \cdot n_K, \quad e \not\subset \partial\Omega, \\ \hat{\mathcal{K}} &= 2\theta[\mathcal{K}(u)u] \cdot n_K, \quad e \subset \partial\Omega, \end{aligned} \tag{68}$$

$$\begin{aligned} \hat{\sigma} &= \langle \mathcal{K}(u)\nabla u\rangle - \alpha h_e^{-1}[u] - F(u_L) \quad \text{if } F'(\langle u\rangle) \cdot n > 0, \\ \hat{\sigma} &= \langle \mathcal{K}(u)\nabla u\rangle - \alpha h_e^{-1}[u] - F(u_R) \quad \text{if } F'(\langle u\rangle) \cdot n \le 0, \end{aligned} \tag{69}$$

where $\alpha > 0$ is a penalization parameter large enough to ensure the ellipticity of the discretization of the elliptic term $\mathcal{K}(u)\nabla u$. This choice of numerical fluxes corresponds to the interior penalty discretization (SIPG with $\theta = 0$, IIPG with $\theta = 1/2$ and NIPG with $\theta = 1$) of the diffusion term and upwind numerical discretization of the convective term; see, e.g., [12, Chapter 2]. We note that (68) and (69) are independent of the orientation of the unit normals n to interior edges.

DEFINITION 11. *We say that the function $u_h^\tau \in Y_{h,\mathrm{disc}}^\tau$ is the approximate solution of* (4) *obtained by the time continuous Galerkin–DGFEM (cG–DGFEM) if*

$$\sum_{K,m} (((u_h^\tau)', v)_{K,m} + A_{K,m}(u_h^\tau, v)) = 0 \quad \forall v \in X_{h,\mathrm{disc}}^\tau, \tag{70}$$

and we say that the function $u_h^\tau \in X_{h,\mathrm{disc}}^\tau$ is the approximate solution of (4) *obtained by the time discontinuous Galerkin–DGFEM (dG–DGFEM) if*

$$\sum_{K,m} \left(((u_h^\tau)', v)_{K,m} + A_{K,m}(u_h^\tau, v) + (\{u_h^\tau\}_{m-1}, v_+^{m-1})_K\right) = 0 \quad \forall v \in X_{h,\mathrm{disc}}^\tau, \tag{71}$$

where $A_{K,m}$ is given by (67).

6.2. Error measure. Again, we use the notation $d_{K,m}$ for the local parameter; see section 4.1. Unfortunately, it is no longer possible to take only $\mathrm{Res}(u_h^\tau)$ as the error measure (see (19)) since this error measure is suited for functions from Y^τ and u_h^τ is no longer in Y^τ in general. We overcome this problem by enhancing the error measure with an additional term

$$J(v) = \sum_{K,m} J_{K,m}(v), \quad J_{K,m}(v) = \frac{d_{K,m}^2}{h_K^2} C_{K,\mathcal{K},F,\alpha} \|[v]\|_{\partial K,m}^2, \tag{72}$$

where

$$C_{K,\mathcal{K},F,\alpha} = \alpha^2 h_K^{-1} + \|\mathcal{K}(u_h^\tau)\|_{L^\infty(K\times I_m)}^2 h_K^{-1} + \|F'(u_h^\tau)\|_{L^\infty(K\times I_m)} h_K. \tag{73}$$

This choice of constant $C_{K,\mathcal{K},F,\alpha}$ improves the robustness of the a posteriori error estimate with respect to scaling of $\mathcal{K}$ and F', when the problem is *singularly perturbed*; see [11]. For details about singularly perturbed problems, see [40]. For nonhomogeneous Dirichlet boundary condition the form $J(v)$ has to be modified. The purpose of $J(u_h^\tau)$ is to measure the distance of function u_h^τ from Y^τ. The resulting error measure $\mathcal{E} = \mathcal{E}(u_h^\tau)$ is defined by

$$\mathcal{E}^2 = \mathrm{Res}(u_h^\tau)^2 + J(u_h^\tau), \tag{74}$$

where $\mathrm{Res}(u_h^\tau)$ is given by (19) and $J(u_h^\tau)$ by (72). Since $J(u_h^\tau) = 0$ for $u_h^\tau \in X_h^\tau$, we can see that this choice of additional error measure term is consistent with the error measure designed for FEM based discretizations.

6.3. Reconstruction of the solution. Since the time reconstruction from section 4.2 is independent of space discretization, we employ reconstruction R_h^τ (cf. (20)) in the DGFEM setting as well. For cG–DGFEM solution u_h^τ, we have $(\{u_h^\tau\}_{m-1}, v)_K = 0$ for all $v \in P^p(K)$, $K \in \mathcal{T}_h^m$; cf. (65). Then relation (22) implies that u_h^τ (either cG–DGFEM solution defined by (70) or dG–DGFEM solution defined by (71)) satisfies

$$\begin{aligned}((R_h^\tau)', v)_{K,m} &= ((u_h^\tau)', v)_{K,m} + (\{u_h^\tau\}_{m-1}, v_+^{m-1})_K \\ &= -A_{K,m}(u_h^\tau, v) \quad \forall v \in P^q(I_m, P^p(K)).\end{aligned} \tag{75}$$

Moreover, although it is possible to use the space reconstruction for $\sigma(u_h^\tau, \nabla u_h^\tau)$ from section 4.3 also for dG space discretization, we employ another one which is computationally cheaper and less sensitive to rounding errors; see [16]. The reconstruction

is defined elementwise: $\sigma_h^\tau|_{K\times I_m} \in P^q(I_m, \mathrm{RTN}_p(K))$ such that

$$(\sigma_h^\tau \cdot n, v)_{e,m} = (\hat{\sigma}\cdot n, v)_{e,m} \quad \forall v \in P^q(I_m, P^p(e)),\ e \subset K, \tag{76}$$
$$(\sigma_h^\tau, v)_{K,m} = (\sigma(u_h^\tau, \nabla u_h^\tau), v)_{K,m} + ((\hat{\mathcal{K}} - \mathcal{K}(u_h^\tau)u_h^\tau)\cdot n_K, v)_{\partial K,m}, \quad \forall v \in P^q(I_m, P^{p-1}(K)^d).$$

Here, the conservativity of the numerical flux $\hat{\sigma}$ implies that the resulting reconstruction $\sigma_h^\tau \in L^2(0,T,H(\mathrm{div},\Omega))$ globally. From the definition of $A_{K,m}$ (67), the local definition (76), and from (75) it is possible to see that such a reconstruction exists and satisfies (26), since

$$\begin{aligned}
&(\nabla\cdot\sigma_h^\tau, v)_{K,m} = -(\sigma_h^\tau, \nabla v)_{K,m} + (\sigma_h^\tau\cdot n_K, v)_{\partial K,m}\\
&= -(\sigma(u_h^\tau,\nabla u_h^\tau),\nabla v)_{K,m} - ((\hat{\mathcal{K}} - \mathcal{K}(u_h^\tau)u_h^\tau)\cdot n_K, v)_{\partial K,m} + (\hat{\sigma}\cdot n_K, v)_{\partial K,m}\\
&= -(\sigma(u_h^\tau,\nabla u_h^\tau),\nabla v)_{K,m} - (c(u_h^\tau), v)_{K,m} - ((\hat{\mathcal{K}} - \mathcal{K}(u_h^\tau)u_h^\tau)\cdot n_K, v)_{\partial K,m}\\
&\quad + (\hat{\sigma}\cdot n_K, v)_{\partial K,m} + (c(u_h^\tau), v)_{K,m}\\
&= -A_{K,m}(u_h^\tau, v) + (c(u_h^\tau), v)_{K,m} = ((R_h^\tau)' + c(u_h^\tau), v)_{K,m} \quad \forall v \in P^q(I_m, P^p(K)).
\end{aligned} \tag{77}$$

6.4. Upper bound. The aim of this section is to show the upper bound to the complete error measure (74). Obviously, $J(u_h^\tau)$ is fully computable from the discrete solution, and it is necessary to estimate $\mathrm{Res}(u_h^\tau)$ only. Since reconstructions R_h^τ and σ_h^τ defined by (20) and (76), respectively, satisfy (26), we can provide the bound to $\mathrm{Res}(u_h^\tau)$ in the same way as for conforming space discretization in section 4.4.

THEOREM 12. *Let $u \in Y$ be the solution of (4), where the flux σ satisfies (62), and let u_h^τ be either the cG–DGFEM solution defined by (70) or the dG–DGFEM solution defined by (71). Let σ_h^τ be defined by (76), R_h^τ be defined by (20), and η be defined in (33). Then*

$$\mathcal{E} \le \sqrt{\eta^2 + J(u_h^\tau)}, \tag{78}$$

where the constant $C_P \le 1/\pi$. Let us note that η depends on C_P; see (33).

6.5. Local efficiency. Here we briefly show similar local efficiency results for (78) as in section 5. We use the same notation for $\mathrm{Res}_{M,m}(u_h^\tau)$, $\bar{c}$, and $\bar{\sigma}$ as in section 5 (see (35), (37), and (38)), and we assume that estimates (39) and (40) hold. Since Lemmas 6–8 are independent of the choice of σ_h^τ, they are valid for DGFEM discretization as well. For the specific choice of physical flux $\sigma(.,.)$ defined by (62) and for σ_h^τ defined by (76), we get following lemma.

LEMMA 13. *Let the numerical fluxes $\hat{\sigma}$ and $\hat{\mathcal{K}}$ be defined by (69) and (68), respectively. Let σ_h^τ be defined by (76). Let $\bar{\sigma}$ be defined by (38) and $\bar{c}$ by (37) satisfying (39) and (40), respectively. Then for $K \in \mathcal{T}_h^m$, $m = 1,\dots,r$,*

$$\eta_{S,K,m} = \frac{d_{K,m}}{h_K}\|\sigma_h^\tau - \sigma(u_h^\tau,\nabla u_h^\tau)\|_{K,m} \lesssim \mathrm{Res}_{\mathcal{T}_K^2,m}(u_h^\tau) + J_{K,m}(u_h^\tau)^{1/2}. \tag{79}$$

Proof. The proof follows the ideas from the seminal work [26, Theorem 3.2]. The triangle inequality gives

$$\frac{d_{K,m}}{h_K}\|\sigma_h^\tau - \sigma(u_h^\tau,\nabla u_h^\tau)\|_{K,m} \le \frac{d_{K,m}}{h_K}\|\sigma_h^\tau - \bar{\sigma}\|_{K,m} + \frac{d_{K,m}}{h_K}\|\bar{\sigma} - \sigma(u_h^\tau,\nabla u_h^\tau)\|_{K,m},$$

where the second term on the right-hand side can be estimated by (41). The estimate of $\sigma_h^\tau - \bar{\sigma}$ can be done in the same way as in [11, Lemma 7.5], where the final relation must be integrated over I_m. ∎

The direct application of Lemmas 6–8 and 13 gives local efficiency estimates of individual terms from a posteriori error estimate (78).

THEOREM 14. *Let $u \in Y$ be the solution of* (4)*, where the flux σ satisfies* (62)*, and let u_h^τ be either the solution of* (70) *or* (71)*, where the numerical fluxes $\hat{\sigma}$ and $\hat{\mathcal{K}}$ are defined by* (69) *and* (68)*, respectively. Let σ_h^τ be defined by* (76) *and R_h^τ be defined by* (20)*. Let $\bar{\sigma}$ be defined by* (38) *and $\bar{c}$ by* (37) *satisfying* (39) *and* (40)*, respectively. Then*

$$\begin{aligned}
\eta_{R,K,m} &= d_{K,m}\|(R_h^\tau)' + c(u_h^\tau) - \nabla\cdot\sigma_h^\tau\|_{K,m} \lesssim \mathrm{Res}_{\mathcal{T}_K^2,m}(u_h^\tau),\\
\eta_{S,K,m} &= \frac{d_{K,m}}{h_K}\|\sigma_h^\tau - \sigma(u_h^\tau,\nabla u_h^\tau)\|_{K,m} \lesssim \mathrm{Res}_{\mathcal{T}_K^2,m}(u_h^\tau) + J_{K,m}(u_h^\tau)^{1/2},\\
\eta_{T,K,m} &= \frac{d_{K,m}}{\tau_m}\|R_h^\tau - u_h^\tau\|_{K,m} \lesssim \mathrm{Res}_{\mathcal{T}_K,m}(u_h^\tau), \quad K\in\mathcal{T}_h^m, m=1,\dots,r.
\end{aligned} \tag{80}$$

Combining Theorem 14 with (36) implies global efficiency of a posteriori error estimate (33).

7. Numerical experiments. In this section, we present two numerical experiments for dG-DGFEM, which illustrate the presented error estimates. The computations are made with the aid of our in-house code ADGFEM [10]. We focus on the verification of the upper bound (78) and local efficiency (80).

Unfortunately, the residual part Res of the error measure $\mathcal{E}$ is not practically computable even for problems with known exact solution u, since the supremum in (19) is taken over the infinite dimensional space Y^τ. Its computation can be rewritten as the following dual problem: Find $\psi \in Y^\tau$ such that

$$(\psi,\varphi)_{Y^\tau} = \sum_{K,m} b_{K,m}(u_h^\tau,\varphi) \quad \forall\varphi\in Y^\tau. \tag{81}$$

Then it holds that $\|\psi\|_{Y^\tau} = \mathrm{Res}(u_h^\tau)$. Unlike the approach in [11, section 8.1], there is no requirement on continuity of $\psi \in Y^\tau$ in time, and hence (81) can be solved independently in each interval I_m, $m = 1,\dots,r$, i.e., find $\psi_m \in Y_{\Omega,m}^\tau$ such that

$$(\psi_m,\varphi)_{Y^\tau} = \sum_{K,m} b_{K,m}(u_h^\tau,\varphi) \quad \forall\varphi\in Y_{\Omega,m}^\tau, \tag{82}$$

and then $\mathrm{Res}(u_h^\tau)^2 = \sum_{m=1}^r \|\psi_m\|_{Y^\tau}^2$.

We approximate ψ_m by linear FEM using the FEniCS software [3] on three-dimensional simplicial meshes. These meshes are obtained by global refinement of the space-time mesh which was proportional to the polynomial degrees p and q. For similar approach, see, e.g., [9]. This FEM approximation of Res is denoted by $\widetilde{\mathrm{Res}}$ and we put $\tilde{\mathcal{E}}^2 = \widetilde{\mathrm{Res}}^2 + J(u_h^\tau)$; cf. (74).

Unfortunately, this procedure is very time consuming and its reliability decreases with increasing p and q, see last row in Table 3.

7.1. Setting. For DGFEM the error measure $\mathcal{E}$ is enhanced by $J(u_h^\tau)$ in (74). Since this term is contained in the error estimate, it improves the effectivity indices.

TABLE 1

Linear diffusion problem, comparison of the approximate error measure $\tilde{\mathcal{E}}$ with the residual estimator η, and the penalization term $J(u_h^\tau)^{1/2}$.

	h	τ	$\tilde{\mathcal{E}}$	η	EOC_η	$J(u_h^\tau)^{1/2}$	EOC_J	i_{eff}	$i_{\mathrm{eff}}^{\mathrm{tot}}$
	0.354	0.100	8.49×10^{-2}	1.16×10^{-1}	–	5.82×10^{-2}	–	1.368	1.218
$p=1$	0.177	0.050	5.67×10^{-2}	7.45×10^{-2}	(0.64)	3.12×10^{-2}	(0.90)	1.314	1.202
$q=0$	0.088	0.025	3.55×10^{-2}	4.51×10^{-2}	(0.72)	1.72×10^{-2}	(0.86)	1.271	1.183
	0.044	0.013	1.99×10^{-2}	2.27×10^{-2}	(0.99)	5.96×10^{-3}	(1.53)	1.143	1.110
	0.354	0.100	1.75×10^{-2}	3.44×10^{-2}	–	7.30×10^{-2}	–	1.960	1.186
$p=2$	0.177	0.050	9.14×10^{-3}	1.83×10^{-2}	(0.91)	2.72×10^{-2}	(1.42)	2.002	1.252
$q=1$	0.088	0.025	3.41×10^{-3}	7.51×10^{-3}	(1.28)	9.18×10^{-3}	(1.57)	2.203	1.326
	0.044	0.013	6.15×10^{-4}	1.17×10^{-3}	(2.68)	1.47×10^{-3}	(2.64)	1.909	1.268
	0.354	0.100	1.48×10^{-2}	3.16×10^{-2}	–	5.12×10^{-2}	–	2.135	1.254
$p=3$	0.177	0.035	4.16×10^{-3}	9.33×10^{-3}	(1.76)	1.11×10^{-2}	(2.21)	2.244	1.340
$q=1$	0.088	0.013	7.96×10^{-4}	1.91×10^{-3}	(2.29)	2.04×10^{-3}	(2.44)	2.397	1.393
	0.044	0.004	7.83×10^{-5}	1.87×10^{-4}	(3.35)	7.93×10^{-5}	(4.68)	2.391	1.691
	0.354	0.100	1.13×10^{-2}	1.92×10^{-2}	–	5.41×10^{-2}	–	1.698	1.121
$p=3$	0.177	0.050	4.62×10^{-3}	8.54×10^{-3}	(1.17)	1.51×10^{-2}	(1.84)	1.850	1.199
$q=2$	0.088	0.025	1.24×10^{-3}	2.57×10^{-3}	(1.73)	3.78×10^{-3}	(2.00)	2.079	1.266
	0.044	0.013	3.90×10^{-5}	1.95×10^{-4}	(3.72)	2.08×10^{-4}	(4.18)	4.996	1.631

Hence in Tables 1 and 3 we present two effectivity indices,

$$i_{\mathrm{eff}}^{\mathrm{tot}\,2} = \frac{\eta^2 + J(u_h^\tau)}{\tilde{\mathcal{E}}^2}, \qquad i_{\mathrm{eff}} = \frac{\eta}{\widetilde{\mathrm{Res}}(u_h^\tau)}. \tag{83}$$

We employ the dG time discretization with the NIPG method ($\theta = 1$) and the upwind numerical fluxes; cf (69) and (68). For linearization of the discrete problem we use the damped Newton-like method.

We consider square domains Ω and a family of uniformly refined space-time meshes. We choose initial parameters (h_0, τ_0) such that the spatial and temporal parts of the error are of comparable size. Then, motivated by the theoretical assumption that the error asymptotically behaves like $O(h^p + \tau^{q+1})$ (cf. [12, section 6.2]), we set $h_n := h_0 2^{-n}$ and $\tau_n := \tau_0 2^{\frac{-np}{q+1}}$ for $n \in \mathbb{N}$ and we chose $d_{K,m} = \sqrt{h_K^2 + \tau_m^2}$.

We evaluate the experimental order of convergence

$$\mathrm{EOC} = \frac{\log(E_n/E_{n-1})}{\log(h_n/h_{n-1})},\ n = 2, \ldots, \tag{84}$$

where E_n may be either an error estimator η or the penalization term $J(u_h^\tau)^{1/2}$.

7.2. Linear diffusion. First, we consider a convection-diffusion problem from [11, § 8.2]. We set $\Omega = (-1, 1)^2$ and $T = 1$ with $\sigma(u, \nabla u) = \varepsilon \nabla u - \frac{u^2}{2}(1, 1)^{\mathrm{T}}$ and $c(u) = 0$. The initial and boundary conditions are chosen such that

$$u = \left(1 + \exp\left(\frac{x + y - t + 1}{2\varepsilon}\right)\right)^{-1}, \tag{85}$$

which forms an inner layer moving in the diagonal direction. The steepness of this layer is increasing as ε decreases to zero. We set $\varepsilon = 10^{-2}$.

We consider space polynomial degrees $p \in \{1, 2, 3\}$ and $q \in \{0, 1, 2\}$ with respect to time. In Table 1 error estimates η and $J(u_h^\tau)$ are compared to the numerical

TABLE 2

Linear diffusion problem with $p = 2$ and $q = 1$ and comparison of the individual components of the estimator η.

h	τ	η_R	η_S	η_T	$J(u_h^\tau)^{1/2}$
0.354	0.100	2.37×10^{-6}	2.31×10^{-2}	1.15×10^{-2}	7.30×10^{-2}
0.177	0.050	1.23×10^{-5}	1.31×10^{-2}	5.37×10^{-3}	2.72×10^{-2}
0.088	0.025	3.22×10^{-5}	5.53×10^{-3}	2.04×10^{-3}	9.18×10^{-3}
0.044	0.013	7.11×10^{-5}	5.45×10^{-4}	6.35×10^{-4}	1.47×10^{-3}

TABLE 3

Forchheimer flow problem, comparison of the approximate error measure $\tilde{\mathcal{E}}$ with the residual estimator η, and the penalization term $J(u_h^\tau)^{1/2}$.

	h	τ	$\tilde{\mathcal{E}}$	η	EOC$_\eta$	$J(u_h^\tau)^{1/2}$	EOC$_J$	i_{eff}	$i_{\text{eff}}^{\text{tot}}$
$p=1$ $q=0$	0.177	0.100	3.12×10^{-3}	6.92×10^{-3}	–	8.51×10^{-3}	–	2.221	1.327
	0.088	0.050	1.18×10^{-3}	2.62×10^{-3}	(1.40)	3.16×10^{-3}	(1.43)	2.211	1.330
	0.044	0.025	4.37×10^{-4}	9.63×10^{-4}	(1.44)	1.15×10^{-3}	(1.46)	2.204	1.333
	0.022	0.013	1.09×10^{-4}	3.49×10^{-4}	(1.47)	4.11×10^{-4}	(1.48)	3.210	1.462
$p=1$ $q=1$	0.177	0.200	3.28×10^{-3}	6.93×10^{-3}	–	9.69×10^{-3}	–	2.116	1.282
	0.088	0.141	1.19×10^{-3}	2.53×10^{-3}	(1.45)	3.43×10^{-3}	(1.50)	2.120	1.289
	0.044	0.100	4.27×10^{-4}	9.04×10^{-4}	(1.48)	1.20×10^{-3}	(1.52)	2.119	1.294
	0.022	0.071	1.78×10^{-4}	3.29×10^{-4}	(1.46)	4.22×10^{-4}	(1.51)	1.852	1.252
$p=2$ $q=1$	0.177	0.100	1.92×10^{-4}	4.49×10^{-4}	–	8.29×10^{-4}	–	2.335	1.252
	0.088	0.050	3.86×10^{-5}	8.86×10^{-5}	(2.34)	1.64×10^{-4}	(2.34)	2.291	1.246
	0.044	0.025	5.00×10^{-5}	1.94×10^{-5}	(2.19)	3.09×10^{-5}	(2.41)	0.388	0.622

approximation of the dual error $\tilde{\mathcal{E}}$. In Table 2 the individual components η_R, η_S, and η_T are presented, where $\eta_R^2 = \sum_{K,m} \eta_{R,K,m}^2$ and the other components are defined analogously.

7.3. Forchheimer flow. In this experiment we test our algorithm with the Forchheimer two-term law from [27] which is a nonlinear diffusion problem with

$$\sigma(u, \nabla u) = \frac{2\varepsilon}{1 + \sqrt{1 + 4|\nabla u|}} \nabla u. \tag{86}$$

We set $\Omega = (0,1)^2$, $T = 1$, and the initial condition, the boundary condition and $c(u)$ are chosen such that

$$u = e^{-2t} x(1-x) y(1-y). \tag{87}$$

In Table 3 we present comparison of the error estimates η and $J(u_h^\tau)$ with the numerically computed approximation of the dual error $\tilde{\mathcal{E}}$.

Both experiments confirm that a posteriori error estimate (78) is a guaranteed upper bound. Moreover, the overestimation of the error is reasonable; see effectivity indices in Tables 1 and 3.

Although the effectivity indices sometimes worsen during refinements, experimental order of convergence of η remains at almost constant level. Hence we suppose the ambiguous results, e.g., in last row of Table 3, are caused by inaccuracies in the numerical computation of the dual norm, which is especially for higher polynomial degrees a very delicate task, rather than by inaccuracy of the error estimates η.

8. Conclusion. We presented a posteriori error estimate (33) for nonlinear parabolic problem (4), where the discretization in time is based on continuous (conforming) or discontinuous (nonconforming) Galerkin method of arbitrary order in time and by conforming FEM in space; see (11) or (12). This estimate is a guaranteed upper bound and locally (cheaply) computable. Moreover, we derived local efficiency estimates (61).

The technique allowing us to produce uniform estimate for conforming discretization either as for nonconforming one is based on the reformulation of the original continuous problem (4) into the new artificial problem (16) with the same solution in such a way that both discretizations are conforming with respect to the new problem. This enables us to naturally include the *penalization* term into the error estimate.

These results are then briefly extended to a general DGFEM discretization in space, where the numerical fluxes involved in the discretization are consistent and the numerical flux for approximation of the physical flux $\sigma(.,.)$ is conservative.

The theoretical investigation of the constants in the efficiency estimates on the polynomial degree is not covered in this paper. The authors expect that this dependence is rather low.

Finally, we present two numerical experiments showing efficiency and reliability of the derived estimates. The main problem with the verification of the error estimate lies in evaluation of the error measure that is difficult to compute even if the exact solution is known.

There are several items for future work.

- The natural inclusion of the nonconformity in space discretizations into the error estimate similarly as it is made for time discretizations.
- Deriving a posteriori error estimates for other error measures, e.g. error in $L^\infty(0,T,L^2(\Omega))$ norm.
- Deriving a posteriori error estimates for quadrature versions of Galerkin time discretizations, i.e., for certain implicit Runge–Kutta methods.
- Investigating the dependence of the constants in efficiency estimates on the polynomial degree.
- Extension of the technique to *hp*-methods.

Acknowledgment. The authors are very grateful to Martin Vohralik for a fruitful discussion about the equilibrated flux reconstruction technique for a posteriori error estimates.

REFERENCES

[1] M. Ainsworth, *A framework for obtaining guaranteed error bounds for finite element approximations*, J. Comput. Appl. Math., 234 (2010), pp. 2618–2632, https://doi.org/10.1016/j.cam.2010.01.037.

[2] G. Akrivis, C. Makridakis, and R. H. Nochetto, *Galerkin and Runge-Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence*, Numer. Math., 118 (2011), pp. 429–456, https://doi.org/10.1007/s00211-011-0363-6.

[3] M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells, *The FEniCS project version* 1.5, Arch. Num. Soft., 3 (2015), 100, https://doi.org/10.11588/ans.2015.100.20553.

[4] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779, https://doi.org/10.1137/S0036142901384162.

[5] A. Bergam, C. Bernardi, and Z. Mghazli, *A posteriori analysis of the finite element discretization of some parabolic equations*, Math. Comp., 74 (2005), pp. 1117–1138, https://doi.org/10.1090/S0025-5718-04-01697-7.

[6] D. Braess, V. Pillwein, and J. Schöberl, *Equilibrated residual error estimates are p-robust*, Comput. Methods Appl. Mech. Eng., 198 (2009), pp. 1189–1197, https://doi.org/10.1016/j.cma.2008.12.010.

[7] J. Butcher, *Implicit Runge-Kutta processes*, Math. Comp., 18 (1964), pp. 50–64, https://doi.org/10.2307/2003405.

[8] Z. Chen and J. Feng, *An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems*, Math. Comp., 73 (2004), pp. 1167–1193.

[9] L. Demkowicz and J. Gopalakrishnan, *A class of discontinuous Petrov-Galerkin methods.* II. *Optimal test functions*, Numer. Methods Partial Differential Equations, 27 (2011), pp. 70–105, https://doi.org/10.1002/num.20640.

[10] V. Dolejší, *ADGFEM – software package*, Charles University Prague, Faculty of Mathematics and Physics, 2014, https://www2.karlin.mff.cuni.cz/~dolejsi/adgfem/index.html.

[11] V. Dolejší, A. Ern, and M. Vohralík, *A framework for robust a posteriori error control in unsteady nonlinear advection-diffusion problems*, SIAM J. Numer. Anal., 51 (2013), pp. 773–793, https://doi.org/10.1137/110859282.

[12] V. Dolejší and M. Feistauer, *Discontinuous Galerkin Method: Analysis and Applications to Compressible Flow*, Springer, Cham, 2015, https://doi.org/10.1007/978-3-319-19267-3.

[13] K. Eriksson, C. Johnson, and V. Thomée, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO, Modél. Math. Anal. Numér., 19 (1985), pp. 611–643.

[14] A. Ern, I. Smears, and M. Vohralík, *Guaranteed, locally space-time efficient, and polynomial-degree robust a posteriori error estimates for high-order discretizations of parabolic problems*, SIAM J. Numer. Anal., 55 (2017), pp. 2811–2834, https://doi.org/10.1137/16M1097626.

[15] A. Ern, I. Smears, and M. Vohralík, *Equilibrated flux a posteriori error estimates in $L^2(H^1)$-norms for high-order discretizations of parabolic problems*, IMA J. Numer. Anal., 39 (2019), pp. 1158–1179.

[16] A. Ern, A. F. Stephansen, and M. Vohralík, *Guaranteed and robust discontinuous Galerkin a posteriori error estimates for convection-diffusion-reaction problems*, J. Comput. Appl. Math., 234 (2010), pp. 114–130.

[17] A. Ern and M. Vohralík, *A posteriori error estimation based on potential and flux reconstruction for the heat equation*, SIAM J. Numer. Anal., 48 (2010), pp. 198–223, https://doi.org/10.1137/090759008.

[18] A. Ern and M. Vohralík, *Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations*, SIAM J. Numer. Anal., 53 (2015), pp. 1058–1081, https://doi.org/10.1137/130950100.

[19] E. H. Georgoulis and O. Lakkis, *A posteriori error bounds for discontinuous Galerkin methods for quasilinear parabolic problems*, in Numerical Mathematics and Advanced Applications 2009, Proceedings of ENUMATH 2009, the 8th European Conference on Numerical Mathematics and Advanced Applications, Uppsala, Sweden, 2009, Springer, Berlin, 2010, pp. 351–358.

[20] E. H. Georgoulis, O. Lakkis, and J. M. Virtanen, *A posteriori error control for discontinuous Galerkin methods for parabolic problems*, SIAM J. Numer. Anal., 49 (2011), pp. 427–458, https://doi.org/10.1137/080722461.

[21] E. H. Georgoulis, O. Lakkis, and T. P. Wihler, *A Posteriori Error Bounds for Fully-discrete hp-discontinuous Galerkin Timestepping Methods for Parabolic Problems*, https://arxiv.org/abs/1708.05832, 2017.

[22] A. Guillou and J. Soulé, *La résolution numérique des problémes différentiels aux conditions initiales par des méthodes de collocation*, Rev. Franç. Inform. Rech. Opér., 3 (1969), pp. 17–44.

[23] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations* I. *Non-stiff Problems*, Springer Ser. Comput. Math. 8, Springer-Verlag, Berlin, Heidelberg, 2000.

[24] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential-algebraic Problems*, Springer Ser. Comput. Math. 14, Springer-Verlag, Berlin, Heidelberg, 2002.

[25] B. L. Hulme, *One-step piecewise polynomial Galerkin methods for initial value problems*, Math. Comput., 26 (1972), pp. 415–426.

[26] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399, https://doi.org/10.1137/S0036142902405217.

[27] T. Kieu, *Numerical analysis for generalized Forchheimer flows of slightly compressible fluids*, Numer. Methods Partial Differential Equations, 34 (2018), pp. 228–256, https://doi.org/

10.1002/num.22194.
[28] C. Kreuzer, *Reliable and efficient a posteriori error estimates for finite element approximations of the parabolic p-Laplacian*, Calcolo, 50 (2013), pp. 79–110.
[29] C. Kreuzer, C. A. Möller, A. Schmidt, and K. G. Siebert, *Design and convergence analysis for an adaptive discretization of the heat equation*, IMA J. Numer. Anal., 32 (2012), pp. 1375–1403.
[30] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Gauthier-Villars, Paris, 1969.
[31] R. Luce and B. Wohlmuth, *A local a posteriori error estimator based on equilibrated fluxes*, SIAM J. Numer. Anal., 42 (2004), pp. 1394–1414, https://doi.org/10.1137/S0036142903433790.
[32] C. Makridakis and R. H. Nochetto, *Elliptic reconstruction and a posteriori error estimates for parabolic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 1585–1594, https://doi.org/10.1137/S0036142902406314.
[33] C. Makridakis and R. H. Nochetto, *A posteriori error analysis for higher order dissipative methods for evolution problems*, Numer. Math., 104 (2006), pp. 489–514, https://doi.org/10.1007/s00211-006-0013-6.
[34] J. M. Melenk and B. I. Wohlmuth, *On residual-based a posteriori error estimation in hp-FEM*, Adv. Comput. Math., 15 (2001), pp. 311–331.
[35] L. Payne and H. Weinberger, *An optimal Poincaré inequality for convex domains*, Arch. Ration. Mech. Anal., 5 (1960), pp. 286–292, https://doi.org/10.1007/BF00252910.
[36] M. Picasso, *Adaptive finite elements for a linear parabolic problem*, Comput. Methods Appl. Mech. Eng., 167 (1998), pp. 223–237, https://doi.org/10.1016/S0045-7825(98)00121-2.
[37] W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of function space*, Quart. Appl. Math., 5 (1947), pp. 241–269.
[38] S. Repin, *Estimates of deviations from exact solutions of initial-boundary value problem for the heat equation*, Atti Accad. Naz. Lincei, Cl. Sci. Fis. Mat. Nat., IX. Ser., Rend. Lincei, Mat. Appl., 13 (2002), pp. 121–133.
[39] J. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of numerical analysis. Volume II: Finite element methods (Part 1), North-Holland, Amsterdam, 1991, pp. 523–639.
[40] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equation*, Springer Ser. Comput. Math. 24, Springer-Verlag, Berlin, 1996.
[41] D. Schötzau and T. P. Wihler, *A posteriori error estimation for hp-version time-stepping methods for parabolic partial differential equations*, Numer. Math., 115 (2010), pp. 475–509.
[42] I. Šebestová and T. Vejchodský, *Two-sided bounds for eigenvalues of differential operators with applications to Friedrichs, Poincaré, trace, and similar constants*, SIAM J. Numer. Anal., 52 (2014), pp. 308–329, https://doi.org/10.1137/13091467X.
[43] R. Verfürth, *A posteriori error estimates for nonlinear problems: $L^r(0,T;W^{1,\rho}(\Omega))$-error estimates for finite element discretizations of parabolic equations*, Numer. Methods Partial Differ. Equations, 14 (1998), pp. 487–518, https://doi.org/10.1002/(SICI)1098-2426(199807)14:4⟨487::AID-NUM4⟩3.0.CO;2-G.
[44] R. Verfürth, *A posteriori error estimates for finite element discretizations of the heat equation*, Calcolo, 40 (2003), pp. 195–212, https://doi.org/10.1007/s10092-003-0073-2.
[45] R. Verfürth, *Robust a posteriori error estimates for nonstationary convection-diffusion equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1783–1802, https://doi.org/10.1137/040604273.
[46] R. Verfürth, *A Posteriori Error Estimation Techniques for Finite Element Methods*, Oxford University Press, Oxford, 2013.
[47] M. Vlasák, *On polynomial robustness of flux reconstructions*, Appl. Math., 65 (2020), pp. 153–172.
[48] K. Wright, *Some relationship between implicit Runge-Kutta collocation and Lanczos τ methods and their stability properties*, Nordisk Tidskr. Informationsbehandling (BIT), 10 (1969), pp. 217–227.

Chapter 7

Polynomial robustness of efficiency estimates

65 (2020) APPLICATIONS OF MATHEMATICS No. 2, 153–172

ON POLYNOMIAL ROBUSTNESS OF FLUX RECONSTRUCTIONS

Miloslav Vlasák, Praha

Received June 15, 2019. Published online February 26, 2020.

Abstract. We deal with the numerical solution of elliptic not necessarily self-adjoint problems. We derive a posteriori upper bound based on the flux reconstruction that can be directly and cheaply evaluated from the original fluxes and we show for one-dimensional problems that local efficiency of the resulting a posteriori error estimators depends on $p^{1/2}$ only, where p is the discretization polynomial degree. The theoretical results are verified by numerical experiments.

Keywords: a posteriori error estimate; p-robustness; elliptic problem

MSC 2020: 65N15, 65N30

1. Introduction

A posteriori error estimates are important and practical tools in numerical mathematics. They serve two main purposes in numerical discretization of PDEs: to provide information about the discretization error for the current choice of discretization parameters and to provide the localization of the sources of high errors for upcoming possible adaptive procedures. For the survey of main a posteriori techniques for PDE discretizations see e.g. [2], [4], [9], [17], [21] and references cited therein. The applications and comparisons of a posteriori error estimates can be found in e.g. [13].

Since higher order methods and *hp*-adaptive techniques start to be more and more popular, the question of robustness with respect to the discretization polynomial degree becomes very important. On the other hand and in contrast to the number of existing results devoted to the robustness with respect to the mesh-size, there are not many theoretical results devoted to the robustness with respect to the polynomial degree. A posteriori error techniques based on the local Neumann problem for

The work was supported from European Regional Development Fund-Project "Center for Advanced Applied Science" (No. CZ.02.1.01/0.0/0.0/16_019/0000778).

DOI: 10.21136/AM.2020.0152-19

hp-adaptive discretizations are discussed e.g. in [1] and [3]. For the analysis of the polynomial dependence of the technique based on the local residual estimators see e.g. [14]. It shall be pointed out that the efficiency of individual estimators proved in [14] behaves as p^1, where p is the underlying polynomial degree used in the finite element method (FEM) discretization.

Important class of approaches for deriving guaranteed a posteriori upper bounds is based on the hypercircle theorem, see [15], where the reconstruction of fluxes should be fully equilibrated, i.e. they should satisfy exactly certain differential equation. By the residual splitting using the dual variable, the restrictive condition of exact solution of full equilibration of the fluxes can be replaced by a milder assumption that the fluxes should be in $H(\mathrm{div})$ only, see e.g. [16]. The extension of these ideas to nonconforming discretizations can be found in e.g. [8], [20]. The quality of the resulting error estimate depends heavily on the choice of the flux reconstruction. Among many approaches for flux reconstructions, the local mixed finite element technique is very popular, since it enables to reconstruct the fluxes based on local relatively cheap problems and since the resulting reconstruction is completely polynomially robust, i.e. the resulting estimators are efficient independently of the polynomial degree. The core of the proof of the polynomial robustness can be found in [7]. The extension of these ideas to wide class of discretization methods can be found in [11].

We assume in this paper even more simple and cheaper reconstruction following the ideas from [10] that can be easily evaluated directly, i.e. without the necessity to solve any local problems. The main aim of this paper is to show its practical usefulness by proving that the resulting local estimators for one-dimensional problems are efficient up to extremely mild polynomial dependence $p^{1/2}$.

This paper is organized as follows: Section 2 contains the continuous problem setting and the corresponding FEM discretization. Auxiliary results are presented in Section 3. A posteriori error upper bound is derived in Section 4 and corresponding efficiency results are proved in Section 5. Finally, Section 6 contains the numerical experiments illustrating the results derived in Section 5.

2. Continuous problem and its discretization

2.1. Continuous problem. Let $\Omega \subset \mathbb{R}^d$ be a bounded polyhedral domain with Lipschitz continuous boundary $\partial\Omega$. We use standard notation for Lebesgue and Sobolev spaces. Let us consider the following boundary value problem: find $u\colon \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -\Delta u + b \cdot \nabla u + cu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{in } \partial\Omega, \end{aligned} \tag{2.1}$$

where $f \in L^2(\Omega)$ and $b \in \mathbb{R}^d$, $c \in \mathbb{R}$ are constants such that $c \geqslant 0$. Moreover, we assume that the convective constant b is of mediocre size at most, i.e. at most $|b| \sim 1$, to prevent the problem becoming convection dominated. Convection dominated problems represent a very challenging task, see e.g. [18] and the references cited therein, and they are beyond the scope of this paper. Let us denote weak space derivative of u by u' for $d = 1$.

Let $(\cdot, \cdot)$ and $\|\cdot\|$ be the $L^2(\Omega)$ scalar product and norm, respectively. Let us denote the function space $V = H_0^1(\Omega)$.

Definition 2.1. We say that the function $u \in V$ is a weak solution of (2.1) if

$$(\nabla u, \nabla v) + (b \cdot \nabla u + cu, v) = (f, v) \quad \forall\, v \in V. \tag{2.2}$$

According to the Lax-Milgram lemma, there exists a unique solution of problem (2.2).

2.2. Discrete problem. We consider a space partition $\mathcal{T}_h$ consisting of a finite number of closed, d-dimensional simplices K with mutually disjoint interiors and covering $\overline{\Omega}$, i.e. $\overline{\Omega} = \bigcup\limits_{K \in \mathcal{T}_h} K$. We denote the vertices of the mesh by a and edges (or faces) by e. In the rest of the paper we talk about boundary objects of co-dimension 1 as about edges, but we mean vertices, edges or faces depending on the dimension d. For each edge e, let $n = n_e$ denote a unit normal vector to e with arbitrary but fixed direction for the inner edges and with outer direction on $\partial\Omega$. We assume conforming properties of the mesh, i.e. neighbouring elements share an entire edge. We set $h_K = \operatorname{diam}(K)$ and $h = \max_K h_K$. We assume shape regularity of elements, i.e. $h_K/\varrho_K \leqslant C$ for all $K \in \mathcal{T}_h$, where ϱ_K is the radius of the largest d-dimensional ball inscribed into K and constant the C does not depend on $\mathcal{T}_h$ for $h \in (0, h_0)$. Moreover, we assume the local quasi-uniformity of the mesh, i.e. we assume $h_K \leqslant C h_{K'}$ for neighbouring elements K and K' and constant the C does not depend on $\mathcal{T}_h$ for $h \in (0, h_0)$ again.

In order to simplify the notation, we set $(\cdot, \cdot)_M$ and $\|\cdot\|_M$ the local $L^2(M)$-scalar products and norms, respectively, where $M \subset \overline{\Omega}$ is a union of elements $K \in \mathcal{T}_h$.

We define classical finite element space

$$V_h = \{v \in H_0^1(\Omega) \colon\ v|_K \in P_p(K)\}, \tag{2.3}$$

where the space $P_p(K)$ denotes the space of polynomials up to the degree $p \geqslant 1$.

Now we are able to define finite element solution of problem (2.2).

Definition 2.2. We say that the function $u_h \in V_h$ is a discrete solution of (2.2) if

$$(\nabla u_h, \nabla v_h) + (b \cdot \nabla u_h + c u_h, v_h) = (f, v_h) \quad \forall\, v_h \in V_h. \tag{2.4}$$

The existence and uniqueness of the discrete solution follows again from the Lax-Milgram lemma.

Although the functions from V_h are globally continuous, we will need to work with piece-wise continuous functions as well. We define one-sided values, jumps and mean values on the inner edges respectively as

$$\begin{aligned} v(x-) &= \lim_{s\to 0+} v(x-ns), \quad v(x+) = \lim_{s\to 0+} v(x+ns), \\ [v](x) &= v(x-) - v(x+), \quad \langle v\rangle(x) = \frac{1}{2}(v(x-) + v(x+)). \end{aligned} \tag{2.5}$$

For the boundary edges we define

$$v(x-) = \langle v\rangle(x) = \lim_{s\to 0+} v(x-ns), \quad [v](x) = 0. \tag{2.6}$$

3. Auxiliary results

Let $\{\widehat{\phi}_s \in P_s(-1,1)\}_{s=0}^{\infty}$ be Legendre orthogonal polynomials, i.e. $\widehat{\phi}_s \perp P_{s-1}(-1,1)$ with respect to $L^2(-1,1)$-scalar product, normalized by $\widehat{\phi}_s(1) = 1$. The lowest degree examples are $\widehat{\phi}_0(x) = 1$ and $\widehat{\phi}_1(x) = x$. Let $\{\widehat{\chi}_s \in P_s(-1,1)\}_{s=1}^{\infty}$ be Radau polynomials defined by

$$\widehat{\chi}_s = \frac{\widehat{\phi}_s + \widehat{\phi}_{s-1}}{2} \tag{3.1}$$

and $\{\widehat{\psi}_s \in P_s(-1,1)\}_{s=2}^{\infty}$ be Lobatto polynomials defined by

$$\widehat{\psi}_s = \widehat{\phi}_s - \widehat{\phi}_{s-2}. \tag{3.2}$$

Lemma 3.1. *The Legendre polynomials satisfy*

$$\|\widehat{\phi}_s\|^2_{L^2(-1,1)} = \frac{2}{2s+1}, \quad \widehat{\phi}_s'(1) = \frac{s(s+1)}{2}. \tag{3.3}$$

The Radau polynomials defined by (3.1) *satisfy*

$$\widehat{\chi}_s(-1) = 0, \quad \widehat{\chi}_s(1) = 1, \quad \widehat{\chi}_s \perp P_{s-2}(-1,1), \quad \|\widehat{\chi}_s\|^2_{L^2(-1,1)} = \frac{2s}{4s^2-1}. \tag{3.4}$$

The Lobatto polynomials defined by (3.2) *satisfy*

$$\widehat{\psi}_s(1) = 0,\ \widehat{\psi}_s(-1) = 0,\ \widehat{\psi}_s \perp P_{s-3}(-1,1),\ \|\widehat{\psi}_s\|^2_{L^2(-1,1)} = \frac{8s-4}{(2s+1)(2s-3)}, \tag{3.5}$$

and

$$\widehat{\psi}'_s = (2s-1)\widehat{\phi}_{s-1}, \quad \|\widehat{\psi}'_s\|^2_{L^2(-1,1)} = 4s-2. \tag{3.6}$$

P r o o f. The relation for the norm of Legendre polynomials can be found in e.g. [19]. Moreover, the Legendre polynomials satisfy the three-term recurrence

$$(s+1)\widehat{\phi}_{s+1}(x) = (2s+1)x\widehat{\phi}_s(x) - s\widehat{\phi}_{s-1}(x), \tag{3.7}$$

see e.g. [19]. Differentiating the three-term recurrence, inserting $x = 1$ and using $\widehat{\phi}_s(1) = 1$, we obtain

$$(s+1)\widehat{\phi}'_{s+1}(1) = (2s+1) + (2s+1)\widehat{\phi}'_s(1) - s\widehat{\phi}'_{s-1}(1). \tag{3.8}$$

Then the relation for $\widehat{\phi}'_s(1)$ follows by induction. Relations (3.4) and (3.5) can be directly verified from (3.1) and (3.2), respectively, and from the properties of Legendre polynomials. Now, let us show that $\widehat{\psi}'_s = C\widehat{\phi}_{s-1}$, where $C = C(s)$ is a constant. Since $\widehat{\psi}'_s \in P_{s-1}(-1,1)$, it is sufficient to show that $\widehat{\psi}'_s \perp P_{s-2}(-1,1)$. Using (3.5), we get

$$\int_{-1}^{1} \widehat{\psi}'_s w \,\mathrm{d}x = -\int_{-1}^{1} \widehat{\psi}_s w' \,\mathrm{d}x - \widehat{\psi}_s(-1)w(-1) + \widehat{\psi}_s(1)w(1) = 0 \quad \forall\, w \in P_{s-2}(-1,1). \tag{3.9}$$

From this it follows that

$$C\widehat{\phi}_{s-1}(1) = \widehat{\psi}'_s(1) = \widehat{\phi}'_s(1) - \widehat{\phi}'_{s-2}(1). \tag{3.10}$$

Applying (3.3), we arrive at $C = 2s-1$. The relation for the norm of $\widehat{\psi}'_s$ then follows from the relation for the norm of Legendre polynomials. □

The Lobatto polynomials ψ_s on $K = [a_L, a_R]$ are defined by transformation of $\widehat{\psi}_s$ from the reference interval $[-1,1]$,

$$\psi_s(x) = \widehat{\psi}_s\Big(\frac{2(x-a_L)}{h_K} - 1\Big), \quad x \in K. \tag{3.11}$$

The Legendre polynomials ϕ_s and the Radau polynomials χ_s are defined on $K \in \mathcal{T}_h$ analogously.

Lemma 3.2. *Let $v \in V$. Then there exists $v_h \in V_h$ and constant $C_{\mathrm{Fl}} > 0$ independent of local mesh-size h_K and polynomial degree $p \geqslant 1$ such that*

$$\|v - v_h\|_K \leqslant C_{\mathrm{Fl}} \frac{h_K}{p} \|\nabla v\|_K. \tag{3.12}$$

P r o o f. The result can be found in [5]. □

For some cases, the value of the constant C_{Fl} from Lemma 3.2 can be determined exactly. We will show the value of C_{Fl} for $d = 1$.

Lemma 3.3. *Let $d = 1$ and $v \in V$. Then there exists $v_h \in V_h$ such that estimate* (3.12) *holds with*

$$C_{\mathrm{Fl}} = \frac{p}{\sqrt{(2p+3)(2p-1)}}. \tag{3.13}$$

P r o o f. Let us decompose $v|_K \in H^1(K)$ as

$$v|_K = \varphi + \sum_{s=2}^{\infty} \alpha_s \psi_s, \tag{3.14}$$

where $\{\alpha_s\}_{s=2}^{\infty} \subset \mathbb{R}$, $\varphi \in P_1(K)$ is the linear interpolation at the end points of K and $\psi_s \in P_s(K)$ are Lobatto basis (bubble) function defined on $K \in \mathcal{T}_h$ by (3.11). Let us construct suitable v_h element-wise as

$$v_h|_K = \varphi + \sum_{s=2}^{p} \alpha_s \psi_s. \tag{3.15}$$

Applying (3.2) and the orthogonality of Legendre polynomials ϕ_s, we get

$$\begin{aligned} \|v - v_h\|_K^2 &= \left\| \sum_{s=p+1}^{\infty} \alpha_s \psi_s \right\|_K^2 = \left\| \sum_{s=p+1}^{\infty} \alpha_s (\phi_s - \phi_{s-2}) \right\|_K^2 \\ &= \sum_{s=p+1}^{\infty} \alpha_s^2 (\|\phi_s\|_K^2 + \|\phi_{s-2}\|_K^2) - 2 \sum_{s=p+1}^{\infty} \alpha_s \alpha_{s+2} \|\phi_s\|_K^2 \end{aligned} \tag{3.16}$$

$$\leqslant \sum_{s=p+1}^{\infty} \alpha_s^2(\|\phi_s\|_K^2 + \|\phi_{s-2}\|_K^2) + \sum_{s=p+1}^{\infty} \alpha_s^2\|\phi_s\|_K^2$$
$$+ \sum_{s=p+1}^{\infty} \alpha_{s+2}^2\|\phi_s\|_K^2 \leqslant 2 \sum_{s=p+1}^{\infty} \alpha_s^2(\|\phi_s\|_K^2 + \|\phi_{s-2}\|_K^2)$$
$$= 2 \sum_{s=p+1}^{\infty} \alpha_s^2\|\psi_s\|_K^2.$$

From Lemma 3.1, it follows for Lobatto polynomials scaled to $[-1, 1]$ that

$$\|\widehat{\psi}_s\|_{(-1,1)}^2 = \frac{2}{(2s+1)(2s-3)}\|\widehat{\psi}_s'\|_{(-1,1)}^2. \tag{3.17}$$

Since the ratio between the original element K and the reference domain $[-1, 1]$ is $h_K/2$, we get after transformation from $[-1, 1]$ to K that

$$\|\psi_s\|_K^2 = \frac{h_K^2}{2(2s+1)(2s-3)}\|\psi_s'\|_K^2. \tag{3.18}$$

Inserting this relation into (3.16), we obtain

$$\|v - v_h\|_K^2 \leqslant 2 \sum_{s=p+1}^{\infty} \alpha_s^2\|\psi_s\|_K^2 = 2 \sum_{s=p+1}^{\infty} \alpha_s^2 \frac{h_K^2}{2(2s+1)(2s-3)}\|\psi_s'\|_K^2 \tag{3.19}$$
$$\leqslant \frac{h_K^2}{(2p+3)(2p-1)} \sum_{s=p+1}^{\infty} \alpha_s^2\|\psi_s'\|_K^2.$$

Since $\widehat{\psi}_s' = (2s-1)\widehat{\phi}_{s-1}$, $s \geqslant 2$, the derivatives of Lobatto basis and constants are mutually orthogonal. Then we get

$$\sum_{s=p+1}^{\infty} \alpha_s^2\|\psi_s'\|_K^2 \leqslant \|\varphi'\|_K^2 + \sum_{s=2}^{\infty} \alpha_s^2\|\psi_s'\|_K^2 = \left\|\varphi' + \sum_{s=2}^{\infty} \alpha_s\psi_s'\right\|_K^2 = \|v'\|_K^2. \tag{3.20}$$

□

4. Flux reconstruction, error measure and its upper bound

4.1. Flux reconstruction. Since the discretization by FEM is conforming, the exact solution u as well as the discrete solution u_h belong to common space $V = H_0^1(\Omega)$. This quality, i.e. the exact and the discrete solutions belong to common space, does not hold for the gradient of the solution, since $\nabla u \in H(\mathrm{div}, \Omega)$ and $\nabla u_h \notin H(\mathrm{div}, \Omega)$ in general. Our aim is to find suitable reconstruction $\sigma_h = \sigma_h(\nabla u_h) \in H(\mathrm{div}, \Omega)$ such that $\sigma_h \approx \nabla u_h$.

Let $RT_p(K)$ be the local Raviart-Thomas space of order p for element $K \in \mathcal{T}_h$, i.e. $RT_p(K) = P_p(K)^d + x\overline{P}_p(K)$, where $\overline{P}_p(K)$ is a subspace of $P_p(K)$ containing only the polynomial terms of degree p. For $d = 1$, $RT_p(K)$ space is simplified to $P_{p+1}(K)$. The details about Raviart-Thomas spaces and about FEM-like spaces for approximation $H(\mathrm{div}, \Omega)$ in general can be found in [6]. We define the reconstruction σ_h element-wise. We seek $\sigma_h|_K \in RT_p(K)$ such that

$$\begin{aligned} \sigma_h|_e \cdot n &= \langle \nabla u_h \rangle|_e \cdot n \quad \forall\, e \subset K, \\ (\sigma_h, z_h)_K &= (\nabla u_h, z_h)_K \quad \forall\, z_h \in P_{p-1}(K)^d. \end{aligned} \tag{4.1}$$

The conditions in (4.1) represent the natural degrees of freedom for $RT_p(K)$, see [6], Proposition 2.3.4. Applying basis corresponding to these degrees of freedom enables to assemble σ_h directly without the necessity to solve any local linear problems, which results in extremely cheap evaluation of the reconstruction σ_h. This property will be demonstrated later in Lemma 5.1 for $d = 1$.

We should point out that the resulting function σ_h has continuous normal components on inter-element edges and therefore the composition of local contributions of σ_h is in $H(\mathrm{div}, \Omega)$, see e.g. [6].

Important property of σ_h is the orthogonality of $f + \mathrm{div}\, \sigma_h - b \cdot \nabla u_h - cu_h$ on V_h that follows from the discrete problem formulation (2.4) and from (4.1)

$$\begin{aligned} (f + \mathrm{div}\, \sigma_h - b \cdot \nabla u_h - cu_h, v_h) &= (f, v_h) - (b \cdot \nabla u_h + cu_h, v_h) - (\sigma_h, \nabla v_h) \\ &= (f, v_h) - (b \cdot \nabla u_h + cu_h, v_h) - (\nabla u_h, \nabla v_h) = 0 \quad \forall\, v_h \in V_h. \end{aligned} \tag{4.2}$$

R e m a r k 4.1. Relation (4.2) represents a weaker version of the equilibrated flux property

$$(f + \mathrm{div}\, \sigma_h - b \cdot \nabla u_h - cu_h, v_h)_K = 0 \quad \forall\, v_h \in P_p(K), \tag{4.3}$$

used in e.g. [11].

R e m a r k 4.2. The important ingredient for relation (4.2) is that u_h is the exact solution of the discrete problem (2.4). Such a solution is not available for the reconstruction in practical computations, since many other sources of errors come into play (algebraic errors, quadrature errors, rounding errors, etc.). Including these sources of errors will result in the necessity to enhance relation (4.2) by corresponding remainders, e.g. the algebraic error could be represented by the additional term corresponding to the algebraic residuum. A posteriori error estimate including algebraic error can be found in e.g. [12]. For simplicity, we assume in this paper that the exact solution u_h of problem (2.4) is available.

4.2. Upper bound. We define the error measure for $w \in V$ as the dual norm of residual

$$\mathrm{Err}(w) = \sup_{0 \neq v \in V} \frac{(f,v) - (\nabla w, \nabla v) - (b \cdot \nabla w + cw, v)}{\|\nabla v\|}. \tag{4.4}$$

R e m a r k 4.3. For the most simple case $b = 0$, $c = 0$, the error measure is equivalent to H^1-seminorm, i.e. $\mathrm{Err}(w) = \|\nabla u - \nabla w\|$.

The aim of this section is to bound the error measure $\mathrm{Err}(u_h)$ from above. Let $v \in V$ be arbitrary, let $u_h \in V_h$ be the discrete solution given by (2.4) and let σ_h be the reconstruction obtained from u_h by (4.1). Then

$$\begin{aligned}(f,v) - (\nabla u_h, \nabla v) - (b \cdot \nabla u_h + cu_h, v) \\ = (f + \operatorname{div} \sigma_h - b \cdot \nabla u_h - cu_h, v) + (\sigma_h - \nabla u_h, \nabla v).\end{aligned} \tag{4.5}$$

We estimate the terms on the right-hand side individually. We apply (4.2) and Lemma 3.2 on the first term and we get

$$\begin{aligned}(f + \operatorname{div} \sigma_h - b \cdot \nabla u_h - cu_h, v) &= \inf_{v_h \in V_h} (f + \operatorname{div} \sigma_h - b \cdot \nabla u_h - cu_h, v - v_h) \\ &\leqslant \sum_K C_{\mathrm{Fl}} \frac{h_K}{p} \|f + \operatorname{div} \sigma_h - b \cdot \nabla u_h - cu_h\|_K \|\nabla v\|_K.\end{aligned} \tag{4.6}$$

The second term can be estimated by the Cauchy inequality

$$(\sigma_h - \nabla u_h, \nabla v) \leqslant \sum_K \|\sigma_h - \nabla u_h\|_K \|\nabla v\|_K. \tag{4.7}$$

Applying these individual estimates together, we get

$$\begin{aligned}&((f - b \cdot \nabla u_h - cu_h, v) - (\nabla u_h, \nabla v))^2 \\ &\quad \leqslant \sum_K \Big(C_{\mathrm{Fl}} \frac{h_K}{p} \|f + \operatorname{div} \sigma_h - b \cdot \nabla u_h - cu_h\|_K + \|\sigma_h - \nabla u_h\|_K \Big)^2 \|\nabla v\|^2.\end{aligned} \tag{4.8}$$

Let us denote partial estimators

$$\eta_{R,K} = C_{\mathrm{Fl}} \frac{h_K}{p} \| f + \operatorname{div} \sigma_h - b \cdot \nabla u_h - c u_h \|_K, \tag{4.9}$$
$$\eta_{F,K} = \| \sigma_h - \nabla u_h \|_K.$$

From these considerations follows the upper a posteriori error estimate.

Theorem 4.1. *Let $u_h \in V_h$ be the discrete solution obtained by* (2.4) *and σ_h be the reconstruction obtained from u_h by* (4.1). *Then*

$$\operatorname{Err}(u_h)^2 \leqslant \eta^2 = \sum_K (\eta_{R,K} + \eta_{F,K})^2. \tag{4.10}$$

R e m a r k 4.4. The constant C_{Fl} contained in $\eta_{R,K}$ is unknown in general. This constant can be determined in some special cases, e.g. the application of Lemma 3.3 instead of Lemma 3.2 gives the modification of the estimator $\eta_{R,K}$ for $d = 1$

$$\eta_{R,K} = \frac{h_K}{\sqrt{(2p+3)(2p-1)}} \| f + \sigma_h' - b u_h' - c u_h \|_K, \tag{4.11}$$

where all the terms in (4.11) are known. Then both the estimators $\eta_{R,K}$ and $\eta_{F,K}$ are fully computable.

5. Local error measures and its lower bound in one dimension

In this section we assume $d = 1$. The aim of this section is to show that the local individual estimators $\eta_{R,K}$ and $\eta_{F,K}$ from a posteriori estimate (4.10) are locally efficient and how this efficiency depends on the polynomial degree p. It means that these local estimators provide local lower bounds to the local error measure up to some powers of p and some generic constant $C > 0$ that may depend on constants coming from the original continuous problem (the size of the domain Ω, etc.) or on the constants coming from the discretization (mesh shape regularity constant, etc.). However, this constant should be independent of the exact solution u, discrete solution u_h, local mesh sizes h_K, and polynomial degree p. Dependence of the estimate up to this generic constant will be denoted by $\lesssim$.

For the purpose of the efficiency analysis we suppose a traditional assumption that $f \in V_h$. Otherwise, classical oscillation term

$$\sup_{0 \neq v \in V} \frac{(f - f_h, v)}{\| v' \|} \tag{5.1}$$

appears additionally in the efficiency results, where f_h is L^2-orthogonal projection of f on V_h.

To be able to apply the result in a local way, we need the following notation. Let ω_a be a patch consisting of elements sharing common vertex a and ω_K be a patch consisting of elements sharing at least a vertex with K. Let $M \subset \overline{\Omega}$, e.g. $M = K$ or $M = \omega_K$. We define a local version of the space V by

$$V_M = \{v \in V\colon \operatorname{supp}(v) \subset M\} \tag{5.2}$$

and a corresponding local version of Err

$$\operatorname{Err}_M(w) = \sup_{0 \neq v \in V_M} \frac{(f,v) - (w',v') - (bw' + cw, v)}{\|v'\|}. \tag{5.3}$$

Typically, we use $\operatorname{Err}_K(u_h)$, $\operatorname{Err}_{\omega_a}(u_h)$ or $\operatorname{Err}_{\omega_K}(u_h)$. Since the patch ω_K is composed from three elements at most, it is possible to see that

$$\sum_K \operatorname{Err}_K(u_h)^2 \leqslant \sum_K \operatorname{Err}_{\omega_K}(u_h)^2 \lesssim \operatorname{Err}(u_h)^2. \tag{5.4}$$

We divide the proof of the local efficiency of the individual partial estimators $\eta_{R,K}$ and $\eta_{F,K}$ into next auxiliary lemmas.

Lemma 5.1. *Let $d = 1$. Let us denote a polynomial $r_L \in P_{p+1}(K)$ such that $r_L(a_L) = 1$, $r_L(a_R) = 0$ and $r_L \perp P_{p-1}(K)$ for the element $K = [a_L, a_R]$. The polynomial $r_R \in P_{p+1}(K)$ associated with a_R instead of a_L is defined analogically, i.e. $r_R(a_R) = 1$, $r_R(a_L) = 0$ and $r_R \perp P_{p-1}(K)$. Then the reconstruction σ_h defined by* (4.1) *can be expressed by*

$$\sigma_h|_K = u_h'|_K + \frac{1}{2} n[u_h'](a_L) r_L - \frac{1}{2} n[u_h'](a_R) r_R. \tag{5.5}$$

P r o o f. Inserting a_L and a_R into (5.5), we obtain $\sigma_h(a_L) = \langle u_h' \rangle(a_L)$ and $\sigma_h(a_R) = \langle u_h' \rangle(a_R)$, respectively. That corresponds to the first condition in (4.1). Using the orthogonality of polynomials r_L and r_R on $P_{p-1}(K)$, we gain the second condition in (4.1). □

R e m a r k 5.1. The polynomials r_L and r_R are known as Radau polynomials, e.g. $r_R = \chi_{p+1}$, where χ_{p+1} is transformation of the reference Radau polynomial $\widehat{\chi}_{p+1}$ defined in Section 3. They can be alternatively defined as polynomials with zeros in the Radau quadrature nodes. They represent natural basis functions associated with *edge* degrees of freedom in (4.1) for $d = 1$.

Lemma 5.2. *Let $d = 1$, $f \in V_h$, $u_h \in V_h$ and let σ_h be the reconstruction obtained from u_h' by* (4.1). *Then*

$$\eta_{F,K} = \|\sigma_h - u_h'\|_K \lesssim p^{1/2}\,\mathrm{Err}_{\omega_K}(u_h). \tag{5.6}$$

P r o o f. Let us denote the end points of K as a_L and a_R, i.e. $K = [a_L, a_R]$. Then applying Lemma 5.1 and Lemma 3.1 and scaling between reference interval $[-1, 1]$ and K, we get

$$\begin{aligned}\|\sigma_h - u_h'\|_K &\leqslant \frac{1}{2}(|[u_h'](a_L)|\|r_L\|_K + |[u_h'](a_R)|\|r_R\|_K)\\ &= \frac{1}{4}\frac{\sqrt{h_K}}{\sqrt{2}}(|[u_h'](a_L)| + |[u_h'](a_R)|)\|\widehat{\chi}_{p+1}\|_{(-1,1)}\\ &= \frac{1}{4}\frac{\sqrt{h_K(p+1)}}{\sqrt{4(p+1)^2-1}}(|[u_h'](a_L)| + |[u_h'](a_R)|)\\ &\lesssim \frac{\sqrt{h_K}}{\sqrt{p}}(|[u_h'](a_L)| + |[u_h'](a_R)|).\end{aligned} \tag{5.7}$$

Now, let us show the relation between $|[u_h'](a)|$ for $a = a_L, a_R$ and $\mathrm{Err}_{\omega_K}(u_h)$. The case $a = a_R$ is very similar to the case $a = a_L$. Therefore, we discuss only the version with $a = a_L$. Let φ_{a_L} be piece-wise linear function associated with vertex a_L such that $\varphi_{a_L}(a_L) = 1$ and $\varphi_{a_L}(a) = 0$ for other vertices a. Let us define ϕ_{a_L} a piece-wise polynomial function of degree at most $p+2$ satisfying $\mathrm{supp}(\phi_{a_L}) \subset \omega_{a_L}$, $\phi_{a_L}(a_L) = 1$ and ϕ_{a_L} be orthogonal to piece-wise polynomials up to degree $p+1$. Now, we are able to design a suitable test function $w_{a_L} \in V_{\omega_{a_L}}$

$$w_{a_L} = -\mathrm{sgn}([u_h'](a_L))\varphi_{a_L}\phi_{a_L}. \tag{5.8}$$

Then

$$\begin{aligned}\mathrm{Err}_{\omega_{a_L}}(u_h) &= \sup_{0\neq v\in V_{\omega_{a_L}}} \frac{(f,v) - (u_h', v') - (bu_h' + cu_h, v)}{\|v'\|}\\ &\geqslant \frac{(f, w_{a_L}) - (u_h', w_{a_L}') - (bu_h' + cu_h, w_{a_L})}{\|w_{a_L}'\|}\\ &= \frac{\sum_K (f + u_h'' - bu_h' - cu_h, w_{a_L})_K - \sum_a [u_h'](a) w_{a_L}(a)}{\|w_{a_L}'\|}\\ &= \frac{|[u_h'](a_L)|}{\|w_{a_L}'\|}.\end{aligned} \tag{5.9}$$

We shall investigate $\|w'_{a_L}\|^2 = \|w'_{a_L}\|_K^2 + \|w'_{a_L}\|_{K'}^2$, where $K' \subset \omega_a$ is the neighbouring element of K. The forthcoming analysis is very similar for both elements. Therefore, we focus only on $\|w'_{a_L}\|_K^2$. From (5.8) it follows that

$$\|w'_{a_L}\|_K^2 = \int_{a_L}^{a_R} (w'_{a_L})^2 \,\mathrm{d}x = \int_{a_L}^{a_R} (\varphi'_{a_L}\phi_{a_L} + \varphi_{a_L}\phi'_{a_L})^2 \,\mathrm{d}x \tag{5.10}$$
$$\lesssim \int_{a_L}^{a_R} (\varphi'_{a_L})^2 \phi_{a_L}^2 \,\mathrm{d}x + \int_{a_L}^{a_R} \varphi_{a_L}^2 (\phi'_{a_L})^2 \,\mathrm{d}x.$$

We estimate the final integrals individually. Since $(\varphi'_{a_L})^2|_K = 1/h_K^2$, we obtain by Lemma 3.1 and by scaling between $[-1,1]$ and K

$$\int_{a_L}^{a_R} (\varphi'_{a_L})^2 \phi_{a_L}^2 \,\mathrm{d}x = \frac{1}{h_K^2} \int_{a_L}^{a_R} \phi_{a_L}^2 \,\mathrm{d}x = \frac{1}{2h_K} \|\widehat{\phi}_{p+2}\|_{(-1,1)}^2 = \frac{1}{h_K(2p+5)}. \tag{5.11}$$

Since $0 \leqslant \varphi_{a_L} \leqslant 1$, we get

$$\begin{aligned} \int_{a_L}^{a_R} \varphi_{a_L}^2 (\phi'_{a_L})^2 \,\mathrm{d}x &\leqslant \int_{a_L}^{a_R} \varphi_{a_L}(x)(\phi'_{a_L})^2 \,\mathrm{d}x \\ &= \varphi_{a_L}(a_R)\phi'_{a_L}(a_R)\phi_{a_L}(a_R) - \varphi_{a_L}(a_L)\phi'_{a_L}(a_L)\phi_{a_L}(a_L) \\ &\quad - \int_{a_L}^{a_R} (\varphi'_{a_L}\phi'_{a_L} + \varphi_{a_L}\phi''_{a_L})\phi_{a_L} \,\mathrm{d}x \\ &= -\phi'_{a_L}(a_L). \end{aligned} \tag{5.12}$$

We get by Lemma 3.1 and by scaling between $[-1,1]$ and K

$$-\phi'_{a_L}(a_L) = \frac{2}{h_K}\widehat{\phi}'_{p+2}(1) = \frac{(p+2)(p+3)}{h_K}. \tag{5.13}$$

Putting these individual estimates together and applying the local quasi-uniformity of the mesh, we obtain

$$\|w'_{a_L}\|^2 = \|w'_{a_L}\|_K^2 + \|w'_{a_L}\|_{K'}^2 \lesssim \frac{p^2}{h_K} + \frac{p^2}{h_{K'}} \lesssim \frac{p^2}{h_K}. \tag{5.14}$$

Then estimates (5.7), (5.9), and (5.14) give

$$\|\sigma_h - u'_h\|_K^2 \lesssim \frac{h_K}{p}(|[u'_h](a_L)|^2 + |[u'_h](a_R)|^2) \tag{5.15}$$
$$\leqslant \frac{h_K}{p} \operatorname{Err}_{\omega_K}(u_h)^2 (\|w'_{a_L}\|^2 + \|w'_{a_R}\|^2) \lesssim p \operatorname{Err}_{\omega_K}(u_h)^2.$$

□

Lemma 5.3. *Let $d = 1$, $f \in V_h$, $u_h \in V_h$ and let σ_h be the reconstruction obtained from u_h by (4.1). Then*

$$\eta_{R,K} = \frac{h_K}{\sqrt{(2p+3)(2p-1)}} \|f + \sigma_h' - bu_h' - cu_h\|_K \lesssim p^{1/2} \operatorname{Err}_{\omega_K}(u_h). \tag{5.16}$$

P r o o f. Let us denote $w = f + \sigma_h' - bu_h' - cu_h$. Let us represent $v \in V_K$ as

$$v = \sum_{s=2}^{\infty} \alpha_s \psi_s, \tag{5.17}$$

where ψ_s are Lobatto polynomials defined by (3.2) and transformed from the reference element $[-1,1]$ to K and $\{\alpha_s\}_{s=2}^{\infty} \subset \mathbb{R}$ are the corresponding coefficients. Let us show that

$$\sum_{s=2}^{\infty} \alpha_s^2 \|\psi_s\|_K^2 \lesssim \left\| \sum_{s=2}^{\infty} \alpha_s \psi_s \right\|_K^2 = \|v\|_K^2. \tag{5.18}$$

It is possible to show it equivalently on the reference element $[-1,1]$ instead of K. Applying Lemma 3.1, we can see that

$$\begin{aligned} \left\| \sum_{s=2}^{\infty} \alpha_s \widehat{\psi}_s \right\|_{(-1,1)}^2 &= \sum_{s=2}^{\infty} \alpha_s^2 \|\widehat{\psi}_s\|_{(-1,1)}^2 - \sum_{s=4}^{\infty} \alpha_s \alpha_{s-2} \|\widehat{\phi}_{s-2}\|_{(-1,1)}^2 \\ &\geqslant \sum_{s=2}^{\infty} \alpha_s^2 \|\widehat{\psi}_s\|_{(-1,1)}^2 - \frac{1}{2} \sum_{s=4}^{\infty} (\alpha_s^2 + \alpha_{s-2}^2) \|\widehat{\phi}_{s-2}\|_{(-1,1)}^2 \\ &\geqslant \sum_{s=2}^{\infty} \alpha_s^2 (\|\widehat{\phi}_s\|_{(-1,1)}^2 + \|\widehat{\phi}_{s-2}\|_{(-1,1)}^2) \\ &\quad - \frac{1}{2} \sum_{s=2}^{\infty} \alpha_s^2 (\|\widehat{\phi}_s\|_{(-1,1)}^2 + \|\widehat{\phi}_{s-2}\|_{(-1,1)}^2) = \frac{1}{2} \sum_{s=2}^{\infty} \alpha_s^2 \|\widehat{\psi}_s\|_{(-1,1)}^2. \end{aligned} \tag{5.19}$$

Using density of $H_0^1(K)$ in $L^2(K)$ and (5.18), we get

$$\|w\|_K^2 = \sup_{v \in V_K} \frac{(w,v)^2}{\|v\|^2} \lesssim \sup_{v \in V_K} \frac{(w,v)^2}{\sum_{s=2}^{\infty} \alpha_s^2 \|\psi_s\|_K^2}. \tag{5.20}$$

Since $\psi_s \perp P_{s-3}(K)$ and $w \in P_p(K)$ and since $w \perp \psi_s$, $s = 2, \ldots, p$ according to (4.2), we can see that it is possible to take supremum in (5.20) over $v \in V_{K,p+1}$ only, where

$$V_{K,p+1} = \operatorname{span}\{\psi_{p+1}, \psi_{p+2}\} \subset V_K. \tag{5.21}$$

From this follows

$$\frac{h_K^2}{p^2}\|w\|_K^2 \lesssim \sup_{v\in V_{K,p+1}} \frac{(w,v)^2}{\alpha_{p+1}^2\|\psi_{p+1}\|_K^2+\alpha_{p+2}^2\|\psi_{p+2}\|_K^2}\frac{h_K^2}{p^2} \tag{5.22}$$
$$= \sup_{v\in V_{K,p+1}} \frac{(w,v)^2}{\|v'\|^2}\frac{h_K^2}{p^2}\frac{\|v'\|^2}{\alpha_{p+1}^2\|\psi_{p+1}\|_K^2+\alpha_{p+2}^2\|\psi_{p+2}\|_K^2}.$$

According to Lemma 5.2,

$$\sup_{v\in V_{K,p+1}} \frac{(w,v)}{\|v'\|} = \sup_{v\in V_{K,p+1}} \frac{(f+\sigma_h'-bu_h'-cu_h,v)}{\|v'\|} \tag{5.23}$$
$$\leqslant \sup_{v\in V_{K,p+1}} \frac{(f-bu_h'-cu_h,v)-(u_h',v')}{\|v'\|} + \sup_{v\in V_{K,p+1}} \frac{(u_h'-\sigma_h,v')}{\|v'\|}$$
$$\leqslant \operatorname{Err}_{\omega_K}(u_h) + \|u_h'-\sigma_h\|_K \lesssim p^{1/2}\operatorname{Err}_{\omega_K}(u_h).$$

Then it is sufficient to prove that

$$h_K^2\|v'\|_K^2 \lesssim p^2(\alpha_{p+1}^2\|\psi_{p+1}\|_K^2+\alpha_{p+2}^2\|\psi_{p+2}\|_K^2) \quad \forall\, v\in V_{K,p+1} \tag{5.24}$$

to finish the proof. We can show (3.18) in the same way as in the proof of Lemma 3.3. Since ψ_s' are othogonal, see Lemma 3.1, we get with the aid of (3.18)

$$h_K^2\|v'\|_K^2 = h_K^2(\alpha_{p+1}^2\|\psi_{p+1}'\|_K^2+\alpha_{p+2}^2\|\psi_{p+2}'\|_K^2) \tag{5.25}$$
$$= h_K^2\Big(\alpha_{p+1}^2\frac{2(2p+3)(2p-1)}{h_K^2}\|\psi_{p+1}\|_K^2 + \alpha_{p+2}^2\frac{2(2p+5)(2p+1)}{h_K^2}\|\psi_{p+2}\|_K^2\Big)$$
$$\lesssim p^2(\alpha_{p+1}^2\|\psi_{p+1}\|_K^2+\alpha_{p+2}^2\|\psi_{p+2}\|_K^2).$$

□

We summarize the results from Lemma 5.2 and Lemma 5.3 in the following theorem.

Theorem 5.1. *Let $d=1$, $f\in V_h$, $u_h\in V_h$, and let σ_h be the reconstruction obtained from u_h by* (4.1). *Then*

$$\eta_{R,K} \lesssim p^{1/2}\operatorname{Err}_{\omega_K}(u_h), \tag{5.26}$$
$$\eta_{F,K} \lesssim p^{1/2}\operatorname{Err}_{\omega_K}(u_h).$$

Global efficiency estimate is a direct consequence of Theorem 5.1 and (5.4).

Theorem 5.2. *Let $d = 1$, $f \in V_h$, $u_h \in V_h$, and let σ_h be the reconstruction obtained from u_h by* (4.1). *Then*

$$\sum_K (\eta_{R,K} + \eta_{F,K})^2 \lesssim p\,\mathrm{Err}(u_h)^2. \tag{5.27}$$

6. Numerical experiments

The aim of this section is to show the reliability, robustness and efficiency of the estimate from Theorem 4.1 for $d = 1$.

The computation of the individual a posteriori error estimators can be made directly according to (4.9) or (4.11). On the other hand, the computation of the error measures $\mathrm{Err}(u_h)$ or $\mathrm{Err}_{\omega_K}(u_h)$ is difficult even if the exact solution is known, since these error measures are defined as suprema over infinite dimensional spaces. We approximate these error measures by computing these suprema over space $V_h^+ \subset V$ that is richer than the original FEM space V_h, but still finite dimensional. We use four times denser mesh than V_h and polynomial degree $p + 2$ instead of p for the construction of V_h^+. We construct spaces $V_{h,M}^+ \subset V_M$ as subspaces of V_h^+ containing functions with supports restricted to $M \subset \overline{\Omega}$. We compute the approximation of the Riesz representative of residual $z \in V_h^+$ satisfying

$$(z, v_h) = (f - b \cdot \nabla u_h - cu_h, v_h) - (\nabla u_h, \nabla v_h) \quad \forall\, v_h \in V_h^+. \tag{6.1}$$

Then $\mathrm{Err}(u_h) \approx \mathrm{Err}_h^+(u_h) = \|\nabla z\|$. The localized versions $\mathrm{Err}_M(u_h)$ are approximated analogically with the aid of $V_{h,M}^+$ instead of V_h^+.

Let us denote approximate effectivity index

$$\mathrm{Eff} = \frac{\eta}{\mathrm{Err}_h^+(u_h)} \tag{6.2}$$

and its local counterparts for element K

$$\mathrm{Eff}_{R,K} = \frac{\eta_{R,K}}{\mathrm{Err}_{h,\omega_K}^+(u_h)}, \quad \mathrm{Eff}_{F,K} = \frac{\eta_{F,K}}{\mathrm{Err}_{h,\omega_K}^+(u_h)}. \tag{6.3}$$

6.1. Problem settings. We restrict ourselves to $d = 1$ and $\Omega = (0, 1)$. We assume two problems: purely elliptical problem (PEP), where $b, c = 0$, and convection-diffusion-reaction problem (CDRP), where $b = 2$ and $c = 1$. We set the right-hand side $f = \pi^2 \sin(\pi x)$ for PEP and $f = 1$ for CDRP.

6.2. Global h-performance. We test the error estimate (4.10) with respect to the mesh refinement. The polynomial degree is set as $p = 3$. We assume a sequence of successively refined equidistant meshes started with $h = 1/10$ and halved in each step.

We can see from Table 1 that the effectivity indices are tending to one for decreasing h.

	PEP			CDRP		
$1/h$	$\mathrm{Err}_h^+(u_h)$	η	Eff	$\mathrm{Err}_h^+(u_h)$	η	Eff
10	$2.1672-4$	$2.6869-4$	1.24	$1.7478-5$	$2.9540-5$	1.69
20	$2.7111-5$	$3.0187-5$	1.11	$2.1903-6$	$3.1610-3$	1.44
40	$3.3896-7$	$3.5760-6$	1.06	$2.7397-7$	$3.4520-7$	1.26
80	$4.2372-7$	$4.3520-7$	1.03	$3.4251-8$	$3.9150-8$	1.14
160	$5.2966-8$	$5.3678-8$	1.01	$4.2816-9$	$4.6046-9$	1.08
320	$6.6214-9$	$6.6650-9$	1.01	$5.3521-10$	$5.5598-10$	1.04

Table 1. Global h-performance for PEP and CDRP, $p = 3$.

6.3. Global p-performance. We test the error estimate (4.10) with respect to the changing polynomial degree p. We assume equidistant mesh with $h = 1/10$ and $p = 1, \ldots, 7$.

We can observe from Table 2 that two regimes for odd and even polynomial degrees appear. For both regimes the efficiency indices very mildly (sublinearly) increase with increasing p.

	PEP			CDRP		
p	$\mathrm{Err}_h^+(u_h)$	η	Eff	$\mathrm{Err}_h^+(u_h)$	η	Eff
1	$2.0113-1$	$2.4015-1$	1.19	$3.0604-2$	$4.9461-2$	1.62
2	$8.1594-3$	$1.4489-2$	1.78	$8.3845-4$	$1.4924-3$	1.78
3	$2.1669-4$	$2.6883-4$	1.24	$1.7478-5$	$2.9540-5$	1.69
4	$4.2891-6$	$9.6339-6$	2.25	$2.6469-7$	$5.9576-7$	2.25
5	$6.7722-8$	$8.7754-8$	1.30	$3.2125-9$	$5.9543-9$	1.85
6	$8.8966-10$	$2.3607-9$	2.65	$3.2419-11$	$8.6252-11$	2.66
7	$9.9930-12$	$1.3472-11$	1.35	$3.4397-13$	$5.6761-13$	1.65

Table 2. Global p-performance for PEP and CDRP, $h = 1/10$.

6.4. Local efficiency, h-performance. We test the robustness of efficiency estimates (5.26) with respect to decreasing h. The polynomial degree is set as $p = 3$. We assume a sequence of successively refined equidistant meshes started with $h = 1/10$ and halved in each step. For each mesh we take element $K = [0.4, 0.4 + h]$ and we investigate local efficiency on this element.

We can see that the efficiency indices in Table 3 and Table 4 are uniformly bounded for decreasing h.

$1/h$	$\mathrm{Err}^+_{h,\omega_K}(u_h)$	$\eta_{R,K}$	$\mathrm{Eff}_{R,K}$	$\eta_{F,K}$	$\mathrm{Eff}_{F,K}$
10	$1.6053-4$	$6.3399-6$	0.04	$9.3672-5$	0.58
20	$1.4327-5$	$3.4539-7$	0.02	$8.2921-6$	0.58
40	$1.2611-6$	$1.7357-8$	0.01	$7.2852-7$	0.58
80	$1.1099-7$	$8.1726-10$	0.01	$6.4090-8$	0.58
160	$9.7842-9$	$3.7263-11$	0.00	$5.6491-9$	0.58
320	$8.6359-10$	$1.6483-12$	0.00	$4.9857-10$	0.58

Table 3. Local h-performance for PEP, $p=3$, $K=[0.4, 0.4+h]$.

$1/h$	$\mathrm{Err}^+_{h,\omega_K}(u_h)$	$\eta_{R,K}$	$\mathrm{Eff}_{R,K}$	$\eta_{F,K}$	$\mathrm{Eff}_{F,K}$
10	$5.8645-6$	$9.6390-7$	0.16	$3.3251-6$	0.57
20	$4.7421-7$	$3.9647-8$	0.08	$2.7252-7$	0.58
40	$4.0379-8$	$1.6952-9$	0.04	$2.3286-8$	0.58
80	$3.5096-9$	$7.3741-11$	0.02	$2.0257-9$	0.58
160	$3.0776-10$	$3.2330-12$	0.01	$1.7767-10$	0.58
320	$2.7096-11$	$1.4142-13$	0.01	$1.5645-11$	0.58

Table 4. Local h-performance for CDRP, $p=3$, $K=[0.4, 0.4+h]$.

6.5. Local efficiency, p-performance. We test the robustness of efficiency estimates (5.26) with respect to the changing polynomial degree p. We assume equidistant mesh with $h=1/10$ and $p=1,\dots,7$. Similarly as in the previous tests, we take $K=[0.4, 0.5]$ and we investigate local efficiency on this element.

We can observe again in Table 5 and Table 6 two regimes for odd and even polynomial degrees, where the dominating estimator is $\eta_{F,K}$ for odd degrees and $\eta_{R,K}$ for even degrees. The efficiency indices stagnates or very mildly (sublinearly) increase with increasing p.

p	$\mathrm{Err}^+_{h,\omega_K}(u_h)$	$\eta_{R,K}$	$\mathrm{Eff}_{R,K}$	$\eta_{F,K}$	$\mathrm{Eff}_{F,K}$
1	$1.4891-1$	$4.5229-3$	0.03	$8.7188-2$	0.59
2	$1.8492-3$	$1.6367-3$	0.89	$3.7698-4$	0.20
3	$1.6053-4$	$6.3399-6$	0.04	$9.3672-5$	0.58
4	$9.6990-7$	$1.0032-6$	1.03	$2.3560-7$	0.24
5	$5.0174-8$	$2.4500-9$	0.05	$2.9248-8$	0.58
6	$2.0106-10$	$2.3903-10$	1.19	$5.2231-11$	0.26
7	$7.4029-12$	$4.2582-13$	0.06	$4.3191-12$	0.58

Table 5. Local p-performance for PEP, $h=1/10$, $K=[0.4, 0.5]$.

p	$\mathrm{Err}^{+}_{h,\omega_K}(u_h)$	$\eta_{R,K}$	$\mathrm{Eff}_{R,K}$	$\eta_{F,K}$	$\mathrm{Eff}_{F,K}$
1	$1.1293-2$	$1.6445-3$	0.15	$6.4324-3$	0.57
2	$2.7437-4$	$2.6495-4$	0.97	$1.5933-5$	0.06
3	$5.8645-6$	$9.6390-7$	0.16	$3.3251-6$	0.57
4	$8.8444-8$	$1.0897-7$	1.23	$5.3687-9$	0.06
5	$1.0742-9$	$2.0228-10$	0.19	$6.0930-10$	0.57
6	$1.0838-11$	$1.5873-11$	1.46	$6.8052-13$	0.06
7	$1.1321-13$	$2.4244-14$	0.21	$5.4811-14$	0.48

Table 6. Local p-performance for CDRP, $h = 1/10$, $K = [0.4, 0.5]$.

7. Conclusion

We derived a posteriori upper bound for not necessarily self-adjoint elliptic problems based on the cheap direct evaluation. We showed that this reconstruction is efficient up to $p^{1/2}$ for one-dimensional problems, where p is the underlying polynomial degree given by the finite element approximation. The robustness with respect to the mesh-size h and to the polynomial degree p was verified by numerical experiments.

Since the majority of the techniques applied in the efficiency proofs in this paper are extendable to multi-dimensional problems, the author hopes that the proof of the efficiency up to $p^{1/2}$ of this direct reconstruction for multi-dimensional problems will be possible in the future.

References

[1] *M. Ainsworth, J. T. Oden*: A procedure for a posteriori error estimation for h-p finite element methods. Comput. Methods Appl. Mech. Eng. *101* (1992), 73–96. zbl MR doi

[2] *M. Ainsworth, J. T. Oden*: A Posteriori Error Estimation in Finite Element Analysis. Pure and Applied Mathematics, Wiley-Interscience, New York, 2000. zbl MR doi

[3] *M. Ainsworth, B. Senior*: An adaptive refinement strategy for hp-finite element computations. Appl. Numer. Math. *26* (1998), 165–178. zbl MR doi

[4] *I. Babuška, T. Strouboulis*: The Finite Element Method and Its Reliability. Numerical Mathematics and Scientific Computation, Clarendon Press, Oxford, 2001. zbl MR

[5] *I. Babuška, M. Suri*: The h-p version of the finite element method with quasi-uniform meshes. RAIRO, Modélisation Math. Anal. Numér. *21* (1987), 199–238. zbl MR doi

[6] *D. Boffi, F. Brezzi, M. Fortin*: Mixed Finite Element Methods and Applications. Springer Series in Computational Mathematics 44, Springer, Berlin, 2013. zbl MR doi

[7] *D. Braess, V. Pillwein, J. Schöberl*: Equilibrated residual error estimates are p-robust. Comput. Methods Appl. Mech. Eng. *198* (2009), 1189–1197. zbl MR doi

[8] *S. Cochez-Dhondt, S. Nicaise, S. Repin*: A posteriori error estimates for finite volume approximations. Math. Model. Nat. Phenom. *4* (2009), 106–122. zbl MR doi

[9] *K. Eriksson, D. Estep, P. Hansbo, C. Johnson*: Computational Differential Equations. Cambridge University Press, Cambridge, 1996. zbl MR

[10] *A. Ern, A. F. Stephansen, M. Vohralík*: Guaranteed and robust discontinuous Galerkin a posteriori error estimates for convection-diffusion-reaction problems. J. Comput. Appl. Math. *234* (2010), 114–130. zbl MR doi

[11] *A. Ern, M. Vohralík*: Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. SIAM J. Numer. Anal. *53* (2015), 1058–1081. zbl MR doi

[12] *P. Jiránek, Z. Strakoš, M. Vohralík*: A posteriori error estimates including algebraic error and stopping criteria for iterative solvers. SIAM J. Sci. Comput. *32* (2010), 1567–1590. zbl MR doi

[13] *O. Mali, P. Neittaanmäki, S. I. Repin*: Accuracy Verification Methods. Theory and Algorithms. Computational Methods in Applied Sciences 32, Springer, Dordrecht, 2014. zbl MR doi

[14] *J. M. Melenk, B. Wohlmuth*: On residual-based a posteriori error estimation in *hp*-FEM. Adv. Comput. Math. *150* (2001), 311–331. zbl MR doi

[15] *W. Prager, J. L. Synge*: Approximations in elasticity based on the concept of function space. Q. Appl. Math. *5* (1947), 241–269. zbl MR doi

[16] *S. I. Repin*: A posteriori error estimation for variational problems with uniformly convex functionals. Math. Comput. *69* (2000), 481–500. zbl MR doi

[17] *S. I. Repin*: A Posteriori Estimates for Partial Differential Equations. Radon Series on Computational and Applied Mathematics 4, de Gruyter, Berlin, 2008. zbl MR doi

[18] *H.-G. Roos, M. Stynes, L. Tobiska*: Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion-Reaction and Flow Problems. Springer Series in Computational Mathematics 24, Springer, Berlin, 2008. zbl MR doi

[19] *G. Szegő*: Orthogonal Polynomials. Colloquium Publication 23, American Mathematical Society, New York, 1939. zbl MR doi

[20] *S. K. Tomar, S. I. Repin*: Efficient computable error bounds for discontinuous Galerkin approximations of elliptic problems. J. Comput. Appl. Math. *226* (2009), 358–369. zbl MR doi

[21] *R. Verfürth*: A Posteriori Error Estimation Techniques for Finite Element Methods. Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013. zbl MR doi

Author's address: *Miloslav Vlasák*, Czech Technical University, Faculty of Civil Engineering, Thákurova 7, 166 29 Praha 6, Czech Republic, e-mail: `miloslav.vlasak@cvut.cz`.