SIAM J. SCI. COMPUT.
Vol. 34, No. 1, pp. A105–A136

DISCONTINUOUS GALERKIN APPROXIMATION OF RELAXATION MODELS FOR LINEAR AND NONLINEAR DIFFUSION EQUATIONS*

FAUSTO CAVALLI†, GIOVANNI NALDI‡, AND ILARIA PERUGIA§

Abstract. In this work we present finite element approximations of relaxed systems for nonlinear diffusion problems, which can also tackle the cases of degenerate and strongly degenerate diffusion equations. Relaxation schemes take advantage of the replacement of the original partial differential equation (PDE) with a semilinear hyperbolic system of equations, with a stiff source term, tuned by a relaxation parameter ε. When $\varepsilon \to 0^+$, the system relaxes onto the original PDE: in this way, a consistent discretization of the relaxation system for vanishing ε yields a consistent discretization of the original PDE. The numerical schemes obtained with this procedure do not require solving implicit nonlinear problems and possess the robustness of upwind discretizations. The proposed approximations are based on a discontinuous Galerkin method in space and on suitable implicit-explicit integration in time. Then, in principle, we can achieve any order of accuracy and obtain stable solutions, even when the diffusion equation becomes degenerate and solution singularities develop. Moreover, when needed, we can easily incorporate slope limiters within our schemes in order to handle spurious oscillatory phenomena. Some preliminary theoretical results are given, along with several numerical tests in one and two space dimensions, both for linear and nonlinear diffusion problems, including a degenerate diffusion equation, that provide numerical evidence of the properties of the presented approach.

Key words. discontinuous Galerkin method, relaxation models, nonlinear diffusion

AMS subject classifications. 65M60, 65M12, 35K55

DOI. 10.1137/110827752

1. Introduction.

Linear and nonlinear diffusion equations come from a variety of diffusion phenomena widely appearing in nature. They are suggested as mathematical models in many fields, such as filtration, phase transition, biochemistry, image analysis, and dynamics of biological groups. In the nonlinear case, the classical solutions to many of these PDEs fail to exist in finite time, even if the initial data are smooth. In such cases, suitable criteria have been introduced, which allow one to select physically relevant weak solutions beyond the singularity time.

Recently, relaxation approximations to such PDEs have been introduced. These methods are based on replacing the equation by a semilinear hyperbolic system with stiff relaxation terms, tuned by a relaxation parameter ε. When $\varepsilon \to 0^+$, the solution of this system "relaxes" onto the solution of the original PDE. Thus a consistent discretization of the relaxation system for $\varepsilon = 0$ yields a consistent discretization of the original PDE, as can be seen, for instance, in [29] and [2]. The advantage of this procedure is that the numerical scheme obtained in this fashion does not need approximate Riemann solvers for the convective term but possesses the robustness of upwind discretizations. Moreover, the complexity introduced by replacing the original

*Submitted to the journal's Methods and Algorithms for Scientific Computing section March 16, 2011; accepted for publication (in revised form) October 3, 2011; published electronically January 31, 2012.

http://www.siam.org/journals/sisc/34-1/82775.html

†Dipartimento di Matematica, Università di Brescia, Via Valotti 9, 25133 Brescia, Italy (fausto.cavalli@ing.unibs.it).

‡Dipartimento di Matematica, Università di Milano, Via Saldini 50, 20133 Milano, Italy (Giovanni.Naldi@mat.unimi.it).

§Dipartimento di Matematica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy (ilaria.perugia@unipv.it).

PDE with a stiff system of equations is only apparent because it is possible to manage the discretization in an efficient way.

Relaxation approximations for conservation laws were deeply investigated in [29, 38, 31] and extended to the diffusive case of parabolic equations in [28, 23, 37]; high order numerical schemes were introduced in [11, 12, 44]. Moreover, relaxation models based on the Bhatnagar–Gross–Krook (BGK) kinetic approach were developed in [30, 2]. We notice that the relaxation approximation is analogous to the regularization of the Euler equations by the Boltzmann or BGK kinetic equation [14, 18, 19, 22, 40, 7].

The aim of this paper is to analyze from both theoretical and computational viewpoints a finite element approximation of some relaxation systems for diffusion equations of the form

$$\begin{aligned} \frac{\partial u}{\partial t} - \Delta p(u) &= 0 && \text{in } \Omega \times (0,+\infty), \\ u &= g_D && \text{on } \Gamma_D \times (0,+\infty), \\ \nabla p(u) \cdot \boldsymbol{n}_\Omega &= g_N && \text{on } \Gamma_N \times (0,+\infty), \\ u_{|_{\{t=0\}}} &= u_0 && \text{in } \Omega, \end{aligned} \tag{1.1}$$

where Ω is a convex, polyhedral domain in $\mathbb{R}^d$, $d = 1, 2, 3$, with boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$. We denote by $\boldsymbol{n}_\Omega$ the unit normal vector to $\partial\Omega$ pointing outside Ω, and $g_D = g_D(\boldsymbol{x}, t)$, $g_N = g_N(\boldsymbol{x}, t)$, $u_0 = u_0(\boldsymbol{x})$. The considered time domain is $(0, +\infty)$. Moreover, $p : \mathbb{R} \to \mathbb{R}$ is a possibly nonlinear function. To our knowledge, this is the first finite element approximation proposed for diffusive relaxation models.

As a typical example of this model, we might consider a homogeneous, isotropic, rigid porous medium filled with a fluid. If absorption and chemical, osmotic, and thermal effects are ignored, and if we consider for horizontal flow, it is possible to deduce the equation

$$\frac{\partial u}{\partial t} - \Delta u^m = 0, \ m > 0, \tag{1.2}$$

where $u = u(\boldsymbol{x}, t)$ models the volumetric moisture content; when $p(u) = u^m$, with $m > 1$, (1.2) is usually called the porous medium equation.

In this work, $p : \mathbb{R} \to \mathbb{R}$ stands for a nondecreasing Lipschitz continuous function such that

$$0 \le l_p \le p'(s) \le L_p < +\infty \qquad \text{for a.e. } s \in \mathbb{R} \tag{1.3}$$

for given constants L_p and l_p, $p(0) = 0$, and there exists $s_0 > 0$ for which

$$p'(s) > 0 \qquad \text{for a.e. } s \ge s_0. \tag{1.4}$$

In the case $\Gamma_N = \varnothing$, the variational formulation of problem (1.1) reads as follows: find u with

$$u \in L^\infty(0, T; L^\infty(\Omega)) \cap H^1(0, T; H^{-1}), \ \ u(\cdot, 0) = u_0,$$

such that, for a.e. $t \in (0, T)$ and all $\phi \in H_0^1(\Omega)$, the following equation holds:

$$\int_\Omega u_t \, \phi \, d\boldsymbol{x} + \int_\Omega \nabla\theta \cdot \nabla\phi \, d\boldsymbol{x} = 0,$$

where $\theta(x, t) = p(u(x, t))$, a.e. $x \in \Omega$, $t \in (0, T)$. Well-posedness of this problem is discussed, for example, in [21, 27, 33], together with the additional regularity result

$$\theta \in H^1(0, T; L^2(\Omega)) \cap L^\infty(0, T; H_0^{-1}).$$

After a review of relaxation models in section 2, we introduce in section 3 our numerical schemes based on a time discretization by means of a particular family of Runge–Kutta schemes and on a discontinuous Galerkin (DG) space discretization. In section 4, we present some preliminary theoretical analysis, proving L^2-stability of the obtained methods in the case of the one-stage time discretization. Finally, in section 5, we present numerical tests in order to give evidence of the features of this approach.

2. The relaxation model. Jin and Xin first proposed in [29] a way to approximate nonlinear conservation laws through semilinear hyperbolic systems with stiff source terms. The idea was to linearize the differential operator by introducing an auxiliary variable and a small positive relaxation parameter ε such that when $\varepsilon \to 0^+$, the original equation be retrieved. For example, taking into account the scalar conservation law in one space dimension,

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \tag{2.1}$$

the following relaxation approximation was proposed:

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial v}{\partial x} = 0, \\ \dfrac{\partial v}{\partial t} + a^2 \dfrac{\partial u}{\partial t} = -\dfrac{1}{\varepsilon^2}\left(v - f(u)\right), \end{cases} \tag{2.2}$$

where a is a constant. Formally, if we let $\varepsilon \to 0^+$ in the second equation, we find $v = f(u)$, which substituted into the first equation gives back the original PDE. Moreover, it can be shown that (2.2) is a $O(\varepsilon^2)$ approximation of the scalar conservation law

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \varepsilon^2 \frac{\partial}{\partial x}\left((a^2 - f'(u)^2)\frac{\partial u}{\partial x}\right),$$

provided that a verifies the inequalities

$$-a \le f'(u) \le a \quad \forall u,$$

which is called the *subcharacteristic* condition (or Whitham condition). In [29] it was also shown how the above approach could be generalized to multidimensional systems of conservation laws in a natural way by adding further auxiliary variables and equations. Subsequently, the idea to approximate nonlinear PDEs by relaxation has been extended to diffusion and convection diffusion equations; see, for example, [1, 2, 28, 29, 32, 36, 37, 11, 13].

In particular, for nonlinear diffusion equations like (1.1), a relaxation system can be obtained introducing two auxiliary variables, as described in [36]. The first step consists in rewriting the second order differential equation as a first order system through the vector auxiliary variable $\boldsymbol{v}$ and the relaxation parameter ε, obtaining

$$\begin{cases} \dfrac{\partial u}{\partial t} + \nabla \cdot \boldsymbol{v} = 0, \\ \dfrac{\partial \boldsymbol{v}}{\partial t} + \dfrac{1}{\varepsilon^2}\nabla(p(u)) = -\dfrac{\boldsymbol{v}}{\varepsilon^2}. \end{cases} \tag{2.3}$$

Formally, in the small relaxation limit $\varepsilon \to 0^+$, the second equation of (2.3) reduces to $\boldsymbol{v} = -\nabla(p(u))$, which substituted in the first equation allows us to recover the leading order equation (1.1).

Since (2.3) is still nonlinear, we need to further relax the second equation. Introducing the scalar auxiliary variable w and a positive constant a, we obtain

$$\begin{cases} \dfrac{\partial u}{\partial t} + \nabla \cdot \boldsymbol{v} = 0, \\ \dfrac{\partial \boldsymbol{v}}{\partial t} + \dfrac{1}{\varepsilon^2} \nabla w = -\dfrac{\boldsymbol{v}}{\varepsilon^2}, \\ \dfrac{\partial w}{\partial t} + a^2 \nabla \cdot \boldsymbol{v} = -\dfrac{1}{\varepsilon^2} (p(u) - w). \end{cases} \tag{2.4}$$

It is easy to see that when $\varepsilon \to 0^+$ we formally retrieve (1.1), which is now approximated by a semilinear hyperbolic system. If, for small values of ε, a Chapman–Enskog expansion is performed, it is easy to see that the original equation (1.1) with a negative fourth order additional term of order $O(\varepsilon^2)$ is retrieved, which results in a stable perturbation of the diffusion equation. For more details on Chapman–Enskog expansion, see [15].

Appropriate boundary conditions for system (2.4) can be deduced from those of (1.1) and are

$$\begin{aligned} u &= g_D && \text{on } \Gamma_D \times (0, +\infty), \\ \boldsymbol{v} \cdot \boldsymbol{n}_\Omega &= -g_N && \text{on } \Gamma_N \times (0, +\infty), \\ w &= p(g_D) && \text{on } \Gamma_D \times (0, +\infty); \end{aligned}$$

similarly, suitable initial conditions are

$$\begin{aligned} u_{|_{\{t=0\}}} &= u_0 && \text{in } \Omega, \\ \boldsymbol{v}_{|_{\{t=0\}}} &= -\nabla u_0 && \text{in } \Omega, \\ w_{|_{\{t=0\}}} &= p(u_0) && \text{in } \Omega. \end{aligned}$$

We are interested in developing a numerical approximation for (2.4) in the relaxed limit, i.e., when $\varepsilon = 0$ (the so-called relaxation schemes), but the characteristic velocities of system (2.6) become stiff as $\varepsilon \to 0^+$. As described in [37], this numerical issue can be dealt with by introducing a constant (dimensional) vector $\boldsymbol{\alpha} = (\alpha_i)_{i=1,..,d}$ and the $d \times d$ diagonal matrix

$$A = \operatorname{diag}(\boldsymbol{\alpha}), \tag{2.5}$$

whose diagonal elements coincide with the components of $\boldsymbol{\alpha}$. The relaxation system can be rewritten as

$$\begin{cases} \dfrac{\partial u}{\partial t} + \nabla \cdot \boldsymbol{v} = 0, \\ \dfrac{\partial \boldsymbol{v}}{\partial t} + A^2 \nabla w = -\dfrac{1}{\varepsilon^2} \left(\boldsymbol{v} - (\varepsilon^2 A^2 - Id) \nabla w \right), \\ \dfrac{\partial w}{\partial t} + a^2 \nabla \cdot \boldsymbol{v} = -\dfrac{1}{\varepsilon^2} (p(u) - w), \end{cases} \tag{2.6}$$

where Id is the identity matrix. In the previous systems, the parameter ε^2 has the physical dimension of a time, while the dimension of w is equal to the dimension of u times length×length over time, and each component of $\boldsymbol{v}$ has the dimension of u times a velocity; finally, the dimension of the diagonal elements α_i^2 of A^2 is time^{-1}.

In the following, we will set $a = 1$ and consider $\alpha_i \geq 0$, $i = 1, \ldots, d$.

We notice that since for sufficiently small values of the relaxation parameter ε, the relaxation system (2.6) gives a good approximation of the original equation (1.1), integrating (2.6) becomes a convenient way to develop numerical approximation of (1.1). In fact, thanks to the simple linear structure of characteristic fields and the localized lower order term, one can easily develop numerical schemes that are simple and general and that deal with a wide class of nonlinearities. In previous works (see [11, 13]), high order methods both in time and space were developed using finite difference schemes, while in this work we investigate the possibility of using finite element methods in order to consider more general domains. In particular, since the solutions of degenerate parabolic equations can show some behaviors that are common to hyperbolic conservation laws, like the loss of regularity and the appearance of fronts that travel at finite speed, we choose DG finite elements in order to exploit the capabilities and the flexibility of this method in convection dominated problems; see, for example, [17]. *Continuous* finite elements could be used instead; in this case, suitable stabilization techniques need to be employed. We do not further enter into details; we only mention that in [8], for first order linear hyperbolic systems, a theoretical framework for the design and analysis of both stabilized continuous and DG finite element space discretizations has been studied.

For the time discretization, we use IMEX-RK (implicit-explicit Runge–Kutta) schemes. We use IMEX-type methods because of the presence of two different scales in (2.6), namely, a nonstiff one on the (linear) left-hand side, which can be safely treated by explicit methods, and a stiff one on the (nonlinear) right-hand side, which requires implicit methods. Since we consider the relaxed limit ($\varepsilon = 0$) like in [11, 13], the IMEX schemes reduce to explicit ones (see (3.6) below).

Explicit time stepping offers some advantages. First, we can avoid solving nonlinear systems; then, it is possible to recover desirable properties of the solutions (positivity and monotonicity), for instance, by including slope limiters (see the end of sections 3.2 and 5.2). A drawback is that stability requires the standard parabolic CFL condition, which constrains the time step to be proportional to the square of the mesh size.

3. Numerical schemes. High order numerical schemes for systems like (2.6) were developed in [11], first integrating in time and then approximating in space with finite differences. As mentioned, we follow these ideas and obtain first a semidiscrete scheme applying an IMEX-RK time integrator and then a fully discrete scheme by using a DG spatial discretization.

3.1. Time semidiscretization. System (2.6) can be recast into the form

$$\boldsymbol{s}_t + \nabla \cdot g(\boldsymbol{s}) = -\frac{1}{\varepsilon^2} h_\varepsilon(\boldsymbol{s}), \tag{3.1}$$

where $\boldsymbol{s} := (u, \boldsymbol{v}, w)^T$ and

$$g(\boldsymbol{s}) := \begin{pmatrix} v_1 & v_2 & \dots & v_d \\ \alpha_1^2 w & 0 & \dots & 0 \\ 0 & \alpha_2^2 w & \dots & 0 \\ & & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & \dots & \alpha_d^2 w \\ a^2 v_1 & a^2 v_2 & \dots & a^2 v_d \end{pmatrix}, \quad h_\varepsilon(\boldsymbol{s}) := \begin{pmatrix} 0 \\ \boldsymbol{v} - (A^2\varepsilon^2 - Id)\nabla w \\ p(u) - w \end{pmatrix},$$

and $\nabla \cdot g(\boldsymbol{s})$ is the vector of the divergences of the rows of $g(\boldsymbol{s})$. We highlight that g is a linear function, whereas h_ε is a nonlinear function.

A high order time semidiscretization of (3.1) can be achieved by using an IMEX scheme as described in [5, 39], which allows us to treat implicitly the stiff term on the right-hand side and to keep explicit the linear left-hand side of (3.1). Moreover, since the adopted IMEX scheme is only diagonally implicit, each implicit equation can be solved autonomously and does not match with the other equations.

For simplicity, we consider a uniform time step Δt: denoting with $\boldsymbol{s}^n$ the numerical approximation of the variable $\boldsymbol{s}$ at time $t^n = n\Delta t$, for $n = 0, 1, \ldots$, a ν-stages IMEX scheme for (3.1) has the form

$$\boldsymbol{s}^{(i)} = \boldsymbol{s}^n - \Delta t \sum_{k=1}^{i-1} a_{ik}^{\mathrm{EX}} \nabla \cdot g(\boldsymbol{s}^{(k)}) - \frac{\Delta t}{\varepsilon^2} \sum_{k=1}^{i} a_{ik}^{\mathrm{IM}} h_\varepsilon(\boldsymbol{s}^{(k)}), \qquad i = 1, \ldots, \nu, \tag{3.2a}$$

$$\boldsymbol{s}^{n+1} = \boldsymbol{s}^n - \Delta t \sum_{i=1}^{\nu} b_i^{\mathrm{EX}} \nabla \cdot g(\boldsymbol{s}^{(i)}) - \frac{\Delta t}{\varepsilon^2} \sum_{i=1}^{\nu} b_i^{\mathrm{IM}} h_\varepsilon(\boldsymbol{s}^{(i)}). \tag{3.2b}$$

The coefficients $(a_{ik}^{\mathrm{EX}}, b_i^{\mathrm{EX}})$ and $(a_{ik}^{\mathrm{IM}}, b_i^{\mathrm{IM}})$ represents the two Butcher's tableaux of, respectively, the explicit and the diagonally implicit parts of the IMEX pair. The time advancing is carried out by solving the implicit equations (3.2a) for $i = 1, \ldots, \nu$ and then updating $\boldsymbol{s}^n$ at time t^{n+1} with (3.2b).

Since we are looking for relaxed schemes, we let $\varepsilon \to 0^+$ in each equation of system (3.2), having

$$\begin{cases} \displaystyle\sum_{k=1}^{i} a_{ik}^{\mathrm{IM}} h_0(\boldsymbol{s}^{(k)}) = 0, \quad i = 1, \ldots, \nu, \\ \displaystyle\sum_{i=1}^{\nu} b_i^{\mathrm{IM}} h_0(\boldsymbol{s}^{(i)}) = 0. \end{cases}$$

So at each time stage k and for each component j, we have that

$$[h_0(\boldsymbol{s}^{(k)})]_j = 0, \tag{3.3}$$

namely,

$$\begin{aligned} w^{(k)} &= p(u^{(k)}), \\ \boldsymbol{v}^{(k)} &= -\nabla w^{(k)}. \end{aligned}$$

Substituting (3.3) into (3.2), we find

$$\boldsymbol{s}^{(i)} = \boldsymbol{s}^n - \Delta t \sum_{k=1}^{i-1} a_{ik}^{\mathrm{EX}} \nabla \cdot g(\boldsymbol{s}^{(k)}), \qquad i = 1 \ldots, \nu, \tag{3.4a}$$

$$\boldsymbol{s}^{n+1} = \boldsymbol{s}^n - \Delta t \sum_{i=1}^{\nu} b_i^{\mathrm{EX}} \nabla \cdot g(\boldsymbol{s}^{(i)}). \tag{3.4b}$$

Notice that in the limit $\varepsilon \to 0^+$, the original IMEX scheme actually reduces to an explicit scheme.

In the following, we will use the particular family of TVD-RK (total variation diminishing Runge–Kutta) methods introduced in [45]. This class of schemes is able to preserve some spatial stability properties, like TVD, TVDM (TVD in the mean), or TVB (total variation bounded), also after the time integration process, while a generic RK scheme can generate oscillations, even for a TVD spatial discretization (see [24] for details). Thus, we can rewrite the generic $(n+1)$th time step as follows:

$$\boldsymbol{s}^{(0)} = \boldsymbol{s}^n, \tag{3.5a}$$

$$\boldsymbol{s}^{(i)} = \sum_{k=0}^{i-1} \left[\widetilde{a}_{ik} \boldsymbol{s}^{(k)} + \Delta t \, \widetilde{b}_{ik} \nabla \cdot g(\boldsymbol{s}^{(k)}) \right], \qquad i = 1 \ldots, \nu, \tag{3.5b}$$

$$\boldsymbol{s}^{n+1} = \boldsymbol{s}^{(\nu)}, \tag{3.5c}$$

where, as explained in [24], the coefficients $\widetilde{a}_{ik}$ and $\widetilde{b}_{ik}$ must satisfy

$$\widetilde{a}_{ik} \geq 0, \qquad \widetilde{b}_{ik} \neq 0 \Rightarrow \widetilde{a}_{ik} \neq 0, \qquad \sum_{k=0}^{i-1} \widetilde{a}_{ik} = 1.$$

We point out that in order to recover a discretization for (1.1), we need to advance in time only the first component $[\boldsymbol{s}^{(i)}]_1 = u$ of (3.4), since it is the only physically relevant variable of the problem. Moreover, the components $[\boldsymbol{s}^{(i)}]_k$ for $k \geq 2$ do not need to be updated, since they would be immediately overridden in the next stage by the implicit computation. To sum up, at each time step, we can see a relaxation scheme as an iteration of the following steps:

1. *Initialization*:

$$u^{(0)} = u^n. \tag{3.6a}$$

2. For $i = 1, \ldots, \nu$,

Relaxation:

$$\begin{aligned} w^{(i-1)} &= p(u^{(i-1)}), \\ \boldsymbol{v}^{(i-1)} &= -\nabla w^{(i-1)}; \end{aligned} \tag{3.6b}$$

Transport:

$$u^{(i)} = \sum_{k=0}^{i-1} \left[\widetilde{a}_{ik} u^{(k)} + \Delta t \, \widetilde{b}_{ik} \nabla \cdot \boldsymbol{v}^{(k)} \right]. \tag{3.6c}$$

3. *Update*:

$$u^{n+1} = u^{(\nu)}. \tag{3.6d}$$

3.2. Space discretization. Having integrated the relaxation system in time, we need to detail the space discretization of (3.6), for which we propose a DG method. Let us introduce some notation.

Let $\mathcal{T}_h$ be a triangulation of Ω, where the mesh parameter h is defined by $h = \max_{K \in \mathcal{T}_h} h_K$, where $h_K = \operatorname{diam}(K)$. Define $\boldsymbol{n}_K$ as the unit normal vector to ∂K

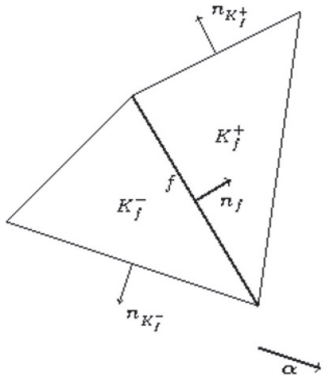

FIG. 3.1. *The vector $\boldsymbol{\alpha}$, a face f, and the two elements K_f^- and K_f^+ sharing f.*

pointing outside K. To fix the ideas, we assume the elements K to be tetrahedra in the three-dimensional case, triangles in two dimensions, and, obviously, intervals in the one-dimensional case. In two and three dimensions, we assume $\mathcal{T}_h$ to be shape-regular and, for simplicity, conformal, i.e., with no *hanging nodes*. Let $\mathcal{F}_h^I$ an $\mathcal{F}_h^B$ be the sets of internal and boundary "faces" (nodes in one dimension, edges in two dimensions, and faces in three dimensions), respectively, of $\mathcal{T}_h$; we denote by $\mathcal{F}_h$ the union $\mathcal{F}_h^I \cup \mathcal{F}_h^B$. We will also use the notation $\mathcal{F}_h^D$ and $\mathcal{F}_h^N$ to denote the subsets of $\mathcal{F}_h^B$ of faces contained in Γ_D and Γ_N, respectively. If D is a given bounded domain, we will denote by $\boldsymbol{n}_D$ the unit normal vector to ∂D pointing outside D.

Let f be a face of $\mathcal{F}_h$. We denote by $\boldsymbol{n}_f$ the unit normal vector to f such that $\boldsymbol{\alpha} \cdot \boldsymbol{n}_f > 0$, whenever $\boldsymbol{\alpha} = (\alpha_i)_{i=1,..,d}^T$ is not parallel to f, or either one of the two normal vectors (e.g., the one pointing from the element with smaller index to the one with larger index in the list of elements, if $f \in \mathcal{F}_h^I$, or the one pointing outside Ω, if $f \in \mathcal{F}_h^B$, to fix the ideas) whenever $\boldsymbol{\alpha}$ is parallel to f. Moreover, we name K_f^- and K_f^+ the two elements sharing f in such a way that $\boldsymbol{n}_f$ is directed from K_f^- to K_f^+ (see Figure 3.1).

Let ψ and $\boldsymbol{\varphi}$ be, respectively, a piecewise smooth function and a vector field on $\mathcal{T}_h$. On the face $\partial K^- \cap \partial K^+$, we define

$$\text{the averages:} \quad \{\!\{\psi\}\!\} := (\psi^+ + \psi^-)/2, \quad \{\!\{\boldsymbol{\varphi}\}\!\} := (\boldsymbol{\varphi}^+ + \boldsymbol{\varphi}^-)/2,$$
$$\text{the jumps:} \quad [\![\psi]\!]_N := \psi^+ \boldsymbol{n}^+ + \psi^- \boldsymbol{n}^-, \quad [\![\boldsymbol{\varphi}]\!]_N := \boldsymbol{\varphi}^+ \cdot \boldsymbol{n}^+ + \boldsymbol{\varphi}^- \cdot \boldsymbol{n}^-.$$

We consider trial and test functions in the following discontinuous finite element space:

$$\mathcal{V}_h = \{v \in L^2(\Omega) \,:\, v|_K \in \mathcal{P}^\ell(K),\ \forall K \in \mathcal{T}_h\}, \tag{3.7}$$

where $\mathcal{P}^\ell(K)$ is the space of polynomials of degree at most ℓ on K.

The problem consists in finding an approximation $u_h \in \mathcal{V}_h$ of the solution u of system (2.6) by the relaxation scheme (3.6a)–(3.6d). To that end, we introduce $\boldsymbol{v}_h \in \mathcal{V}_h^d$ and $w_h \in \mathcal{V}_h$ to approximate the auxiliary variables $\boldsymbol{v}, w$ and the test functions $\psi_h \in \mathcal{V}_h, \boldsymbol{\varphi}_h \in \mathcal{V}_h^d$. Now we can describe in detail the spatial approximations of a generic step n.

Relaxation step. We start by considering a single relaxation step in (3.6b). The discretization of the first equation of (3.6b) can be written as follows: find $w_h^{(i)} \in \mathcal{V}_h$

such that

$$\int_\Omega w_h^{(i)}\psi_h d\boldsymbol{x} = \int_\Omega p(u_h^{(i)})\psi_h d\boldsymbol{x} \tag{3.8}$$

for all $\psi_h \in \mathcal{V}_h$. For the second equation of (3.6b), since we are using a discontinuous finite element space, we cannot simply take the piecewise gradient of the approximating function on the right-hand side; otherwise the jumps at the interelement boundaries would be out of control. We use instead a standard DG gradient approximation, which also includes jump terms and which is constructed as follows. We multiply equation (3.8) by discrete test functions, integrate by parts, element by element, the right-hand side, and approximate the interelement traces by numerical fluxes, which include information from the neighboring elements and the Dirichlet boundary condition. In this way, we obtain

$$\int_\Omega \boldsymbol{v}_h^{(i)} \cdot \boldsymbol{\varphi}_h \, d\boldsymbol{x} = \int_\Omega w_h^{(i)} \nabla_h \cdot \boldsymbol{\varphi}_h \, d\boldsymbol{x} - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \widehat{w}_h^{(i)} \, \boldsymbol{\varphi}_h \cdot \boldsymbol{n}_K \, ds.$$

By defining the numerical fluxes $\widehat{w}_h^{(i)}$ on each face $f \in \mathcal{F}_h$ by

$$\widehat{w}_h^{(i)} = \begin{cases} \{\!\{ w_h^{(i)} \}\!\} & \text{if } f \in \mathcal{F}_h^I, \\ p(g_D(\cdot, t^{n,(i)})) & \text{if } f \in \mathcal{F}_h^D, \\ w_h^{(i)} & \text{if } f \in \mathcal{F}_h^N, \end{cases} \tag{3.9}$$

and integrating by parts element by element once more, we obtain the following discretization of the second equation of (3.6b): find $\boldsymbol{v}_h^{(i)} \in \mathcal{V}_h^d$ such that

$$\begin{aligned} \int_\Omega \boldsymbol{v}_h^{(i)} \cdot \boldsymbol{\varphi}_h \, d\boldsymbol{x} = &- \int_\Omega \nabla_h w_h^{(i)} \cdot \boldsymbol{\varphi}_h \, d\boldsymbol{x} + \int_{\mathcal{F}_h^I} [\![w_h^{(i)}]\!]_N \cdot \{\!\{ \boldsymbol{\varphi}_h \}\!\} \, ds \\ &+ \int_{\mathcal{F}_h^D} (w_h^{(i)} - p(g_D(\cdot, t^{n,(i)}))) \, \boldsymbol{\varphi}_h \cdot \boldsymbol{n}_\Omega \end{aligned} \tag{3.10}$$

for all $\boldsymbol{\varphi}_h \in \mathcal{V}_h^d$, where ∇_h denotes the elementwise application of the ∇ operator. We set $t^{n,(i)} := t^n + c_i \Delta t$ for some coefficients c_i, $0 \le i \le \nu$, with $c_0 = 0$, which are the elements of the vector $\boldsymbol{c}$ of the Butcher tableau (clearly, $t^{n,(0)} = t^n$).

This means that $w_h^{(i)}$ and $\boldsymbol{v}_h^{(i)}$ are computed from $u_h^{(i)}$ by solving the algebraic linear systems (3.8) and (3.10) with coefficient matrices given, respectively, by the mass matrices in $\mathcal{V}_h$ and $\mathcal{V}_h^d$. Moreover $w_h^{(i)}$ is simply the L_2-projection of $p(u_h^{(i)})$ on $\mathcal{V}_h$ and so it coincides with $u_h^{(i)}$ in the case of $p(u) = u$.

Transport step. The discretization of the transport step (3.6c) at the generic ith stage of the nth time step is performed in a DG fashion: find $u_h^{(i)} \in \mathcal{V}_h$ such that

$$\int_\Omega u_h^{(i)} \psi_h \, d\boldsymbol{x} = \sum_{k=0}^{i-1} \left[\tilde{a}_{ik} \int_\Omega u_h^{(k)} \psi_h \, d\boldsymbol{x} + \Delta t \, \tilde{b}_{ik} [\mathcal{B}_h(\boldsymbol{v}_h^{(k)}, \psi_h) - \mathcal{Q}_h^{n,(k)}(\psi_h)] \right] \tag{3.11}$$

for all $\psi_h \in \mathcal{V}_h$. The bilinear form $\mathcal{B}_h(\cdot,\cdot) : \mathcal{V}_h^d \times \mathcal{V}_h \to \mathbb{R}$ appearing in (3.11) is a DG bilinear form associated with the divergence operator and $\mathcal{Q}_h$ is a time-dependent

linear functional on $\mathcal{V}_h$ depending on the boundary data g_D and g_N:

$$\begin{aligned}
&\mathcal{B}_h(\boldsymbol{v}_h^{(k)},\psi_h) - \mathcal{Q}_h^{n,(k)}(\psi_h) \\
&\qquad = \int_\Omega \nabla_h \cdot \boldsymbol{v}_h^{(k)}\, \psi_h \, d\boldsymbol{x} - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \left(\boldsymbol{v}_h^{(k)} - \widehat{\boldsymbol{v}}_h^{(k)}\right) \cdot \boldsymbol{n}_K \, \psi_h \, ds \\
&\qquad = \sum_{j=1}^{d} \left[\int_\Omega \frac{\partial_h v_{h,j}^{(k)}}{\partial x_j}\, \psi_h \, d\boldsymbol{x} - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \left(v_{h,j}^{(k)} - \widehat{v}_{h,j}^{(k)}\right) n_{K,j}\, \psi_h \, ds \right],
\end{aligned} \tag{3.12}$$

where $\widehat{\boldsymbol{v}}_h^{(k)}$ are the *numerical fluxes*, which still have to be defined; the dependence on the (time-dependent) boundary data g_D and g_N is contained in the definition of the numerical fluxes on $\partial\Omega$.

In order to define the *numerical fluxes* for the $\boldsymbol{v}$ variable, we go back to the system (2.6) and perform the following steps:

1. Diagonalize the operator on the left-hand side of system (2.6) along each spatial direction and construct the characteristic variables.
2. Approximate the fluxes for the characteristic variables.
3. Reconstruct the fluxes of the $\boldsymbol{v}$ variable in terms of the original conservative variables $\boldsymbol{s}$.

The diagonalization of the operator on the left-hand side of system (2.6) along the direction x_j involves

$$\begin{bmatrix} \dfrac{\partial u}{\partial t} + \dfrac{\partial v_j}{\partial x_j} \\ \dfrac{\partial v_j}{\partial t} + \alpha_j^2 \dfrac{\partial w}{\partial x_j} \\ \dfrac{\partial w}{\partial t} + \dfrac{\partial v_j}{\partial x_j} \end{bmatrix}. \tag{3.13}$$

It is easy to see that by introducing the characteristic variables U_j, V_j, W_j defined by

$$\begin{aligned} U_j &= \frac{1}{2}\left(w + \frac{1}{\alpha_j}\, v_j\right), \\ V_j &= \frac{1}{2}\left(w - \frac{1}{\alpha_j}\, v_j\right), \\ W_j &= u - w \end{aligned} \tag{3.14}$$

for $j = 1, \ldots, d$, the expression in (3.13) can be diagonalized into

$$\begin{bmatrix} \dfrac{\partial U_j}{\partial t} + \alpha_j \dfrac{\partial U_j}{\partial x_j} \\ \dfrac{\partial V_j}{\partial t} - \alpha_j \dfrac{\partial V_j}{\partial x_j} \\ \dfrac{\partial W_j}{\partial t} \end{bmatrix}. \tag{3.15}$$

Since

$$v_j = \alpha_j\,(U_j - V_j),$$

the numerical fluxes for v_j can be directly derived from those of U_j and V_j; more precisely, in (3.12) we will set

$$\widehat{v}_{h,j}^{(k)} = \alpha_j \left(\widehat{U}_{h,j}^{(k)} - \widehat{V}_{h,j}^{(k)} \right); \tag{3.16}$$

thus, in order to complete the definition of the method, all we need to define are the fluxes for the characteristic variables U_j and V_j.

The first two components of (3.15) are linear transport operators with opposite transport directions. Assume, with no loss of generality, that $\alpha_j > 0$ (α_j enter the physical system (2.6) only through its square). It is therefore natural to use *upwind fluxes* from left to right for the U_j variables and from right to left for the V_j variables.

In order to define the numerical fluxes $\widehat{U}_{h,j}^{(k)}$ and $\widehat{V}_{h,j}^{(k)}$ on each face $f \in \mathcal{F}_h^I$, we need to consider a new orientation of the faces, depending on the component we are considering, defined as follows. Fix j: if f is a face in $\mathcal{F}_h^I$, if the jth component of $\boldsymbol{n}_f$ denoted by $(\boldsymbol{n}_f)_j$ is ≥ 0, we set $K_{f,j}^- = K_f^-$ and $K_{f,j}^+ = K_f^+$; otherwise we set $K_{f,j}^- = K_f^+$ and $K_{f,j}^+ = K_f^-$. With this notation, $\widehat{U}_{h,j}^{(k)}$ and $\widehat{V}_{h,j}^{(k)}$ are defined on each $f \in \mathcal{F}_h^I$ as follows:

$$\widehat{U}_{j,h}^{(k)} = (U_{j,h}^{(k)})^-{}_{|_f},$$

$$\widehat{V}_{j,h}^{(k)} = (V_{j,h}^{(k)})^+{}_{|_f},$$

where the superscripts $-$ and $+$ denote, respectively, the restrictions to $K_{f,j}^-$ and $K_{f,j}^+$.

Therefore, on interior faces, the numerical fluxes $\widehat{v}_{h,j}^{(k)}$, $j = 1, \ldots, d$, that appear (3.12) can be reconstructed from $\widehat{U}_{h,j}^{(k)}$ and $\widehat{V}_{h,j}^{(k)}$ according to (3.16):

$$\widehat{v}_{h,j}^{(k)} = \{\!\{ \widehat{v}_{h,j}^{(k)} \}\!\} + \frac{\alpha_j}{2} \frac{(\boldsymbol{n}_f)_j}{|(\boldsymbol{n}_f)_j|} \left((w_h^{(k)})^- - (w_h^{(k)})^+ \right), \tag{3.17}$$

where, with abuse of notation, we set $\frac{(\boldsymbol{n}_f)_j}{|(\boldsymbol{n}_f)_j|} = 0$, whenever $(\boldsymbol{n}_f)_j = 0$.

Since only the normal component of $\widehat{\boldsymbol{v}}_h^{(k)}$ is needed, we can write the vector-valued fluxes $\widehat{\boldsymbol{v}}_h^{(k)}$ in the following more standard way:

$$\widehat{\boldsymbol{v}}_h^{(k)} = \{\!\{ \boldsymbol{v}_h^{(k)} \}\!\} + \frac{\boldsymbol{\alpha} \cdot \boldsymbol{n}_f}{2} [\![w_h^{(k)}]\!]_N. \tag{3.18}$$

In fact, simple calculations show that on interior faces, the normal component of the flux defined in (3.17) coincides with the normal component of the flux in (3.18).

On Dirichlet boundary faces, $\widehat{\boldsymbol{v}}_h^{(k)}$ is defined in a similar way taking into account the boundary condition $w = p(g_{\mathcal{D}}(\cdot, t^{n,(k)}))$, while on Neumann boundary faces, we simply define the fluxes as the Neumann boundary condition.

Summarizing, the numerical fluxes $\widehat{\boldsymbol{v}}_h^{(k)}$ are defined on each face $f \in \mathcal{F}_h$ as follows:

$$\widehat{\boldsymbol{v}}_h^{(k)} = \begin{cases} \{\!\{ \boldsymbol{v}_h^{(k)} \}\!\} + \dfrac{\boldsymbol{\alpha} \cdot \boldsymbol{n}_f}{2} [\![w_h^{(k)}]\!]_N & \text{if } f \in \mathcal{F}_h^I, \\ \boldsymbol{v}_h^{(k)} + \dfrac{\boldsymbol{\alpha} \cdot \boldsymbol{n}_f}{2} \left(w_h^{(k)} - p(g_D(\cdot, t^{n,(k)})) \right) \boldsymbol{n}_\Omega & \text{if } f \in \mathcal{F}_h^D, \\ -g_N \boldsymbol{n}_\Omega & \text{if } f \in \mathcal{F}_h^N. \end{cases} \tag{3.19}$$

Remark 3.1. The numerical fluxes $\widehat{\boldsymbol{v}}_h^{(k)}$ on boundary faces correspond to the following boundary fluxes for the characteristic variables: if $f \in \mathcal{F}_h^B$,

$$\widehat{U}_{j,h}^{(k)} = \begin{cases} \dfrac{1}{2}\left(p(g_D(\cdot,t^{n,(k)})) + \dfrac{1}{\alpha_j}\, v_{j,h}^{(k)}\right) & \text{if } \boldsymbol{n}_f = -\boldsymbol{n}_\Omega,\ f \subset \Gamma_D, \\ \dfrac{1}{2}\left(w_h^{(k)} - \dfrac{1}{\alpha_j}\,\left(v_{j,h}^{(k)} + 2g_N(\cdot,t^{n,(k)})n_{\Omega,j}\right)\right) & \text{if } \boldsymbol{n}_f = -\boldsymbol{n}_\Omega,\ f \subset \Gamma_N, \\ (U_{j,h}^{(k)})^-{}_{|_f} & \text{if } \boldsymbol{n}_f = \boldsymbol{n}_\Omega, \end{cases}$$

$$\widehat{V}_{j,h}^{(k)} = \begin{cases} \dfrac{1}{2}\left(p(g_D(\cdot,t^{n,(k)})) - \dfrac{1}{\alpha_j}\, v_{j,h}^{(k)}\right) & \text{if } \boldsymbol{n}_f = \boldsymbol{n}_\Omega,\ f \subset \Gamma_D, \\ \dfrac{1}{2}\left(w_h^{(k)} + \dfrac{1}{\alpha_j}\,\left(v_{j,h}^{(k)} + 2g_N(\cdot,t^{n,(k)})n_{\Omega,j}\right)\right) & \text{if } \boldsymbol{n}_f = \boldsymbol{n}_\Omega,\ f \subset \Gamma_N, \\ (V_{j,h}^{(k)})^+{}_{|_f} & \text{if } \boldsymbol{n}_f = -\boldsymbol{n}_\Omega. \end{cases}$$

Notice that these boundary fluxes are consistent.

Taking into account the definition of $\widehat{\boldsymbol{v}}_h^{(k)} = \widehat{\boldsymbol{v}}_h^{(k)}(\boldsymbol{v}_h^{(k)}, w_h)$, the resulting expressions for $\mathcal{B}_h(\boldsymbol{v}_h^{(k)}, \psi_h)$ and $\mathcal{Q}_h^{n,(k)}(\psi_h)$ in (3.12) are the following:

$$\begin{aligned} \mathcal{B}_h(\boldsymbol{v}_h^{(k)}, \psi_h) = &\int_\Omega \nabla_h \cdot \boldsymbol{v}_h^{(k)}\, \psi_h\, d\boldsymbol{x} - \int_{\mathcal{F}_h^I} [\![\boldsymbol{v}_h^{(k)}]\!]_N \{\!\{\psi_h\}\!\}\, ds \\ &+ \int_{\mathcal{F}_h^I} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} [\![w_h^{(k)}]\!]_N \cdot [\![\psi_h]\!]_N\, ds \\ &+ \int_{\mathcal{F}_h^D} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} w_h^{(k)}\, \psi_h\, ds - \int_{\mathcal{F}_h^N} \boldsymbol{v}_h^{(k)} \cdot \boldsymbol{n}_\Omega\, \psi_h\, ds \end{aligned} \tag{3.20}$$

and

$$\mathcal{Q}_h^{n,(k)}(\psi_h) = \int_{\mathcal{F}_h^D} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} p(g_D(\cdot,t^{n,(k)}))\, \psi_h\, ds - \int_{\mathcal{F}_h^N} g_N(\cdot,t^{n,(k)})\, \psi_h\, ds. \tag{3.21}$$

Remark 3.2. Integrating by parts the first term on the right-hand side of (3.20), $\mathcal{B}_h(\boldsymbol{v}_h^{(k)}, \psi_h)$ can also be written as

$$\begin{aligned} \mathcal{B}_h(\boldsymbol{v}_h^{(k)}, \psi_h) = &-\int_\Omega \boldsymbol{v}_h^{(k)} \cdot \nabla_h \psi_h\, d\boldsymbol{x} + \int_{\mathcal{F}_h^I} \{\!\{\boldsymbol{v}_h^{(k)}\}\!\} \cdot [\![\psi_h]\!]_N\, ds \\ &+ \int_{\mathcal{F}_h^D} \boldsymbol{v}_h^{(k)} \cdot \boldsymbol{n}_\Omega\, \psi_h\, ds + \int_{\mathcal{F}_h^I} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} [\![w_h^{(k)}]\!]_N \cdot [\![\psi_h]\!]_N\, ds \\ &+ \int_{\mathcal{F}_h^D} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} w_h^{(k)}\, \psi_h\, ds. \end{aligned}$$

As we are going to prove in section 4, our scheme is L^2-stable: this kind of stability is not sufficient to grant that the scheme be, for example, monotonicity and positivity preserving. To try to regain these properties, one can make use of *slope limiting techniques*. There are several limiting approaches that can be found in the literature; here, we simply recall what a slope limiter is, after [17].

Given $u_h^{(k)} \in \mathcal{V}_h$, define $z_h^{ik} \in \mathcal{V}_h$ such that

$$\int_\Omega z_h^{ik}\,\psi_h\,d\boldsymbol{x} = \int_\Omega u_h^{(k)}\,\psi_h\,d\boldsymbol{x} + \frac{\tilde{b}_{ik}}{\tilde{a}_{ik}}\,\Delta t\,[\mathcal{B}_h(\boldsymbol{v}_h^{(k)},\psi_h) - \mathcal{Q}_h^{n,(k)}(\psi_h)] \qquad \forall \psi_h \in \mathcal{V}_h,$$

where $\boldsymbol{v}_h^{(k)}$ is related to $u_h^{(k)}$ through (3.8) and (3.10).

An "ideal" slope limiter projection $\Lambda\Pi_h : \mathcal{V}_h \to \mathcal{V}_h$ is a nonlinear operator devised in such a way that if $u_h^{(k)} = \Lambda\Pi_h v_h$ for some $v_h \in \mathcal{V}_h$, then

$$\left|z_h^{ik}\right|_{TV} \le \left|u_h^{(k)}\right|_{TV}, \tag{3.22}$$

where $|\cdot|_{TV}$ denotes the total variation seminorm. A slope limiter satisfying property (3.22) is said to be TVD. On the other hand, most of the limiters present in the literature are not TVD; in particular, the min-mod limiter described in [17], which we are going to use in our numerical simulations, is only TVDM.

For the way our scheme is written, inserting a slope limiter is straightforward: it is just a projection process that takes place after each transport step.

We do not further enter the discussion on slope limiters, since our aim here is only to point out how a slope limiter could be employed in our schemes in order to try to recover, at least partially, some lost properties of the continuous problem.

3.3. Fully discrete relaxation scheme. In order to write the complete IMEX RKDG-relaxation method for (1.1), we assume we have the functions

$$\begin{aligned}
w_h^{(i)} &= \texttt{relax_w}(u_h^{(i)}),\\
\boldsymbol{v}_h^{(i)} &= \texttt{relax_v}(w_h^{(i)}),\\
u_h^{(i)} &= \texttt{transport}((u_h^{(k)})_{k=0}^{i-1},(\boldsymbol{v}_h^{(k)})_{k=0}^{i-1}),\\
\widetilde{u}_h^{(i)} &= \texttt{sl_projection}(u_h^{(i)}),
\end{aligned}$$

the first two representing the relaxation step, the third the transport step (see (3.6b) and (3.6c)), and the last the application of a slope limiter, i.e.,

- $\texttt{relax\_w} : u_h^{(i)} \in \mathcal{V}_h \mapsto w_h^{(i)} \in \mathcal{V}_h$ such that

$$\int_\Omega w_h^{(i)}\psi_h d\boldsymbol{x} = \int_\Omega p(u_h^{(i)})\psi_h d\boldsymbol{x}$$

  for all $\psi_h \in \mathcal{V}_h$;
- $\texttt{relax\_v} : w_h^{(i)} \in \mathcal{V}_h \mapsto \boldsymbol{v}_h^{(i)} \in \mathcal{V}_h^d$ such that

$$\begin{aligned}
\int_\Omega \boldsymbol{v}_h^{(i)}\cdot\boldsymbol{\varphi}_h\,d\boldsymbol{x} = &-\int_\Omega \nabla_h w_h^{(i)}\cdot\boldsymbol{\varphi}_h\,d\boldsymbol{x} + \int_{\mathcal{F}_h^I}[\![w_h^{(i)}]\!]_N\cdot\{\!\{\boldsymbol{\varphi}_h\}\!\}\,ds\\
&+\int_{\mathcal{F}_h^D}(w_h^{(i)} - p(g_D(\cdot,t^{n,(k)})))\,\boldsymbol{\varphi}_h\cdot\boldsymbol{n}_\Omega
\end{aligned}$$

  for all $\boldsymbol{\varphi}_h \in \mathcal{V}_h^d$;
- $\texttt{transport} : ((u_h^{(k)})_{k=0}^{i-1},(\boldsymbol{v}_h^{(k)})_{k=0}^{i-1}) \in \mathcal{V}_h^i \times (\mathcal{V}_h^d)^i \mapsto u_h^{(i)} \in \mathcal{V}_h$ such that

$$\int_\Omega u_h^{(i)}\psi_h\,d\boldsymbol{x} = \sum_{k=0}^{i-1}\left[\tilde{a}_{ik}\int_\Omega u_h^{(k)}\psi_h\,d\boldsymbol{x} + \Delta t\,\tilde{b}_{ik}[\mathcal{B}_h(\boldsymbol{v}_h^{(k)},\psi_h) - \mathcal{Q}_h^{n,(k)}(\psi_h)]\right]$$

for all $\psi_h \in \mathcal{V}_h$, with $\mathcal{B}_h(\boldsymbol{v}_h^{(k)}, \psi_h)$ and $\mathcal{Q}_h^{n,(k)}(\psi_h)$ defined by (3.20) and (3.21), respectively;

- `sl_projection`: $u_h^{(i)} \in \mathcal{V}_h \to \widetilde{u}_h^{(i)} \in \mathcal{V}_h$ as in [17].

The complete IMEX RKDG-relaxation method for (1.1) reads as follows.

Initialize: Define u_h^0 as the L^2-projection of the the initial datum u_0 onto $\mathcal{V}_h$: find $u_h^0 \in \mathcal{V}_h$ such that

$$\int_\Omega u_h^0 \psi_h \, d\boldsymbol{x} = \int_\Omega u_0 \psi_h \, d\boldsymbol{x} \tag{3.23}$$

for all $\psi_h \in \mathcal{V}_h$; compute $u_h^0 = \texttt{sl\_projection}(u_h^0)$.

Time stepping: For $n = 0, 1, \ldots,$

(a) Set $u_h^{(0)} = u_h^n$.

(b) For $i = 1, \ldots, \nu$ (time stages),

i. Relaxation: $w_h^{(i-1)} = \texttt{relax\_w}(u_h^{(i-1)})$ and $\boldsymbol{v}_h^{(i-1)} = \texttt{relax\_v}(w_h^{(i-1)})$;

ii. Transport: $u_h^{(i)} = \texttt{transport}((u_h^{(k)})_{k=0}^{i-1}, (\boldsymbol{v}_h^{(k)})_{k=0}^{i-1})$;

iii. Slope limiter projection: $u_h^{(i)} = \texttt{sl\_projection}(u_h^{(i)})$.

(c) Update: $u_h^{n+1} = u_h^{(\nu)}$.

(d) Slope limiter projection: $u_h^{n+1} = \texttt{sl\_projection}(u_h^{n+1})$.

4. Stability analysis. In this section, we perform the stability analysis of the relaxation scheme described in section 3.3 in its basic version, where `sl_projection` is the identity function. We will also take

$$\boldsymbol{\alpha} = \ell^2 h^{-1} \boldsymbol{a} \tag{4.1}$$

with $\boldsymbol{a}$ independent of the mesh size h and the polynomial approximation degree ℓ (see Remark 4.1).

We start in section 4.1 by reformulating the method in a more compact form, eliminating the $\boldsymbol{v}$ unknown from the system; then we proceed in section 4.2 by stating some preliminary results needed in the proof of the L^2-stability which is developed in section 4.3.

4.1. Reformulation. In order to perform the stability analysis, it is convenient to rewrite the relaxation scheme described in section 3.3 in a more compact form by eliminating the unknown $\boldsymbol{v}$ from the final system: in order to do that, we need to introduce the so-called lifting operators (see [3]).

For w piecewise smooth on $\mathcal{T}_h$, we introduce the *lifting* $\mathcal{L}(w) \in \mathcal{V}_h^d$ defined by

$$\int_\Omega \mathcal{L}(w) \cdot \boldsymbol{\varphi}_h \, d\boldsymbol{x} = \int_{\mathcal{F}_h^I} [\![w]\!]_N \cdot \{\!\{\boldsymbol{\varphi}_h\}\!\} \, ds + \int_{\mathcal{F}_h^D} w \, \boldsymbol{\varphi}_h \cdot \boldsymbol{n}_\Omega \, ds \tag{4.2}$$

for all $\boldsymbol{\varphi}_h \in \mathcal{V}_h^d$.

We also need to define the lifting $\mathcal{G}_D(t) \in \mathcal{V}_h^d$ of the Dirichlet boundary condition $g_D(\cdot, t)$:

$$\int_\Omega \mathcal{G}_D(t) \cdot \boldsymbol{\varphi}_h \, d\boldsymbol{x} = \int_{\mathcal{F}_h^D} p(g_D(\cdot, t)) \, \boldsymbol{\varphi}_h \cdot \boldsymbol{n}_\Omega \, ds \tag{4.3}$$

for all $\boldsymbol{\varphi}_h \in \mathcal{V}_h^d$. With these definitions, we can write the relaxation steps as

$$\int_\Omega \boldsymbol{v}_h^{(k)} \cdot \boldsymbol{\varphi}_h \, d\boldsymbol{x} = -\int_\Omega \left(\nabla_h w_h^{(k)} - \mathcal{L}(w_h^{(k)}) + \mathcal{G}_D(t^{n,(k)}) \right) \cdot \boldsymbol{\varphi}_h \, d\boldsymbol{x}$$

for all $\boldsymbol{\varphi}_h \in \mathcal{V}_h^d$, i.e.,

$$\boldsymbol{v}_h^{(k)} = -(\nabla_h w_h^{(k)} - \mathcal{L}(w_h^{(k)}) + \mathcal{G}_D(t^{n,(k)})), \tag{4.4}$$

and the form $\mathcal{B}_h$(see Remark 3.2) as

$$\begin{aligned}\mathcal{B}_h(\boldsymbol{v}_h^{(k)}, \psi_h) = &- \int_\Omega \boldsymbol{v}_h^{(k)} \cdot (\nabla_h \psi_h - \mathcal{L}(\psi_h))\, d\boldsymbol{x} \\ &+ \int_{\mathcal{F}_h^I} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} [\![w_h^{(k)}]\!]_N \cdot [\![\psi_h]\!]_N \, ds + \int_{\mathcal{F}_h^D} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} w_h^{(k)}\, \psi_h \, ds.\end{aligned} \tag{4.5}$$

By inserting (4.4) into (4.5), we obtain

$$\mathcal{B}_h\left(\boldsymbol{v}_h^{(k)}\left(w_h^{(k)}, p(g_D(\cdot, t^{n,(k)}))\right), \psi_h\right) - \mathcal{Q}_h^{n,(k)}(\psi_h) = \mathcal{A}_h(w_h^{(k)}, \psi_h) - \mathcal{P}_h^{(k)}(\psi_h),$$

where

$$\begin{aligned}\mathcal{A}_h(w_h^{(k)}, \psi_h) = &\int_\Omega (\nabla_h w_h^{(k)} - \mathcal{L}(w_h^{(k)})) \cdot (\nabla_h \psi_h - \mathcal{L}(\psi_h))\, d\boldsymbol{x} \\ &+ \int_{\mathcal{F}_h^I} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} [\![w_h^{(k)}]\!]_N \cdot [\![\psi_h]\!]_N \, ds + \int_{\mathcal{F}_h^D} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} w_h^{(k)}\, \psi_h \, ds\end{aligned} \tag{4.6}$$

and

$$\begin{aligned}\mathcal{P}_h^{n,(k)}(\psi_h) = &\int_{\mathcal{F}_h^D} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} p(g_D(\cdot, t^{n,(k)}))\, \psi_h \, ds + \int_{\mathcal{F}_h^N} g_N(\cdot, t^{n,(k)})\, \psi_h \, ds \\ &+ \int_\Omega \mathcal{G}_D(t^{n,(k)}) \cdot (\nabla_h \psi_h - \mathcal{L}(\psi_h))\, d\boldsymbol{x}.\end{aligned} \tag{4.7}$$

The method then reads as follows:

Initialize: Define u_h^0 as the L^2-projection of the the initial datum u_0 onto $\mathcal{V}_h$: find $u_h^0 \in \mathcal{V}_h$ such that

$$\int_\Omega u_h^0 \psi_h \, d\boldsymbol{x} = \int_\Omega u_0 \psi_h \, d\boldsymbol{x}$$

for all $\psi_h \in \mathcal{V}_h$.

Time stepping: For $n = 0, 1, \ldots,$

(a) Set $u_h^{(0)} = u_h^n$;

(b) For $i = 1, \ldots, \nu$ (time stages),

i. Compute $w_h^{(i-1)} =$ `relax_w`$(u_h^{(i-1)})$;

ii. Find $u_h^{(i)} \in \mathcal{V}_h$ such that

$$\begin{aligned}&\int_\Omega u_h^{(i)} \psi_h \, d\boldsymbol{x} \\ &\quad = \sum_{k=0}^{i-1} \left[\tilde{a}_{ik} \int_\Omega u_h^{(k)} \psi_h \, d\boldsymbol{x} + \Delta t\, \tilde{b}_{ik} [\mathcal{A}_h(w_h^{(k)}, \psi_h) - \mathcal{P}_h^{n,(k)}(\psi_h)] \right]\end{aligned}$$

for all $\psi_h \in \mathcal{V}_h$, with $\mathcal{A}_h(w_h^{(k)}, \psi_h)$ and $\mathcal{P}_h^{n,(k)}(\psi_h)$ defined by (4.6) and (4.7);

(c) Update: $u_h^{n+1} = u_h^{(\nu)}$.

Remark 4.1. In the linear case, it is immediate to see that the IMEX RKDG-relaxation method coincides with an RKDG method (see [17]) where the semidiscretization in space is performed by a modified local discontinuous Galerkin (LDG) method (see [3] and [10]).

Comparing our definition of numerical fluxes (see (3.9) and (3.19)) with that of [10], it is clear that our space discretization in the linear case is an LDG-type method with

$$\boldsymbol{C}_{12} = \mathbf{0},\ C_{22} = 0,\ (C_{11})_{|_f} = \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2}.$$

(In the notation of [10], $\boldsymbol{C}_{12}$ is only defined on interior faces, C_{11} on interior and Dirichlet boundary faces, and C_{22} on interior and Neumann boundary faces.) For stability reasons, the stabilization parameter C_{11} needs to match with an inverse inequality; this leads to the standard choice

$$C_{11} = \frac{\mathtt{a}\,\ell^2}{h}$$

with the constant $\mathtt{a}$ independent of the mesh size h and the polynomial approximation degree ℓ (see, e.g., [26, 41]).

4.2. Preliminary results. Let us define

$$\|\psi\|_{DG}^2 = \mathcal{A}_h(\psi,\psi) \tag{4.8}$$

for all $\psi \in H^1(\mathcal{T}_h)$, where $H^1(\mathcal{T}_h)$ denotes the space of functions in Ω whose restrictions to K belong to $H^1(K)$ for all $K \in \mathcal{T}_h$. In order to prove that expression (4.8) actually defines a norm, we will make use of [16, Lemma 3.2]. Even if [16] was focused on the multidimensional case, the proof of [16, Lemma 3.2] clearly also covers the case $d = 1$. We report this result for completeness.

LEMMA 4.2. *Let $K \subset \mathbb{R}^d$, $d \geq 1$, be an element and let $f_1, \ldots, f_d$ be d faces of K. Given $\boldsymbol{\sigma} \in L^2(K)^d$ and $\zeta_i \in L^2(f_i)$, $i = 1, \ldots, d$, there exists a unique function $\boldsymbol{Z} \in \mathcal{P}^\ell(K)^d$ such that*

$$\int_K (\boldsymbol{Z} - \boldsymbol{\sigma})\cdot\boldsymbol{v}\,d\boldsymbol{x} = 0 \qquad \forall \boldsymbol{v} \in \mathcal{P}^{\ell-1}(K)^d,$$
$$\int_{f_i} (\boldsymbol{Z}\cdot\boldsymbol{n}_i - \zeta_i)\,\omega\,ds = 0 \qquad \forall \omega \in \mathcal{P}^\ell(f_i),\ i = 1, \ldots, d,$$

where $\boldsymbol{n}_i$ is the outward normal unit vector of f_i.

PROPOSITION 4.3. *The expression* (4.8) *defines a norm in $\mathcal{V}_h$, and thus*

$$|\mathcal{A}_h(u,\psi)| \leq \|u\|_{DG}\|\psi\|_{DG} \qquad \forall u, \psi \in \mathcal{V}_h. \tag{4.9}$$

Proof. It is clear that $\|\cdot\|_{DG}$ is a seminorm. In order to show that it is a norm, we only need to prove that $\|\psi\|_{DG} = 0$ implies $\psi = 0$.

In order to do that, notice that

$$0 = \mathcal{A}_h(\psi,\psi) = \|\nabla_h\psi - \mathcal{L}(\psi)\|_{0,\Omega}^2 + \int_{\mathcal{F}_h^I} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2}[\![\psi]\!]_N^2\,ds + \int_{\mathcal{F}_h^D} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2}\psi^2\,ds$$

implies $\nabla_h\psi = \mathcal{L}(\psi)$, as well as

$$[\![\psi]\!]_N = 0 \quad \text{on each face } f \in \mathcal{F}_h^I \text{ such that } \boldsymbol{\alpha}\cdot\boldsymbol{n}_f \neq 0, \tag{4.10}$$
$$\psi = 0 \quad \text{on each face } f \text{ in } \mathcal{F}_h^D \text{ such that } \boldsymbol{\alpha}\cdot\boldsymbol{n}_f \neq 0. \tag{4.11}$$

Since $\boldsymbol{\alpha}$ is constant, then $\boldsymbol{\alpha}\cdot\boldsymbol{n}_f \neq 0$ on at least one face on each element K, say, f_K; we denote by $\mathcal{F}_h^\alpha$ the set of all f_K.

By definition of $\mathcal{L}$, $\nabla_h\psi = \mathcal{L}(\psi)$ can be written as

$$\int_\Omega \nabla_h\psi\cdot\boldsymbol{\varphi}\,d\boldsymbol{x} - \int_{\mathcal{F}_h^I}[\![\psi]\!]_N\cdot\{\!\{\boldsymbol{\varphi}\}\!\}\,ds - \int_{\mathcal{F}_h^D}\psi\,\boldsymbol{\varphi}\cdot\boldsymbol{n}_\Omega\,ds = 0 \qquad \forall\boldsymbol{\varphi}\in\mathcal{V}_h^d.$$

In each $K\in\mathcal{T}_h$, we choose $\boldsymbol{\varphi}=\boldsymbol{Z}$ given by Lemma 4.2 with

$$\begin{aligned}
&\boldsymbol{\sigma} := \nabla\psi,\\
&\zeta_i := -[\![\psi]\!]_N\cdot\boldsymbol{n}_i && \text{on } f_i\in(\partial K\cap\mathcal{F}_h^I),\\
&\zeta_i := -\psi && \text{on } f_i\in(\partial K\cap\mathcal{F}_h^B),
\end{aligned}$$

where $f_i\neq f_K$ for $i=1,\dots,d$, and we obtain

$$\|\nabla_h\psi\|_{0,\Omega}^2 + \int_{\mathcal{F}_h^I\setminus\mathcal{F}_h^\alpha}[\![\psi]\!]_N^2\,ds + \int_{\mathcal{F}_h^D\setminus\mathcal{F}_h^\alpha}\psi^2\,ds = 0,$$

which, together with (4.10) and (4.11), implies that ψ is constant in every $K\in\mathcal{T}_h$, $[\![\psi]\!]_N=0$ on every $f\in\mathcal{F}_h^I$ and $\psi=0$ on every $f\in\mathcal{F}_h^D$, and thus $\psi=0$.

The continuity property (4.9) is straightforward. ∎

We define the L^2-norm on the set of faces $\mathcal{F}_h$ as

$$\|\psi\|_{0,\mathcal{F}_h}^2 = \sum_{f\in\mathcal{F}_h}\|\psi\|_{0,f}^2$$

and the L^2-norm on subsets of $\mathcal{F}_h$ similarly.

The following inverse inequality holds true.

LEMMA 4.4. *There exists constant $C_{\rm inv}>0$ independent of h and ℓ such that*

$$\|\psi\|_{DG} \le C_{\rm inv}\ell^2 h^{-1}\|\psi\|_{0,\Omega} \qquad \forall\psi\in\mathcal{V}_h.$$

Proof. First, we recall from [41, Proposition 3.2] and [42, Proposition 4.2] that the lifting operator $\mathcal{L}$ satisfies the following stability bound: there exists a constant $C_{\rm lift}>0$ only dependent on the shape regularity of the mesh such that

$$\|\mathcal{L}(\psi)\|_{0,\Omega} \le C_{\rm lift}\left(\|\ell h^{-1/2}[\![\psi]\!]_N\|_{0,\mathcal{F}_h^I} + \|\ell h^{-1/2}\psi\|_{0,\mathcal{F}_h^D}\right) \qquad \forall\psi\in\mathcal{V}_h. \tag{4.12}$$

Using (4.12) and the standard *hp*-version inverse estimates (see, e.g., [43] and [26]), for all $\psi\in\mathcal{V}_h$ we have

$$\begin{aligned}
\|\psi\|_{DG}^2 &\le \|\nabla_h\psi\|_{0,\Omega}^2 + \|\mathcal{L}(\psi)\|_{0,\Omega}^2 + \frac{|\boldsymbol{a}|}{2}\left(\|\ell h^{-1/2}[\![\psi]\!]_N\|_{0,\mathcal{F}_h^I}^2 + \|\ell h^{-1/2}\psi\|_{0,\mathcal{F}_h^D}^2\right)\\
&\le \|\nabla_h\psi\|_{0,\Omega}^2 + \left(2C_{\rm lift}^2 + \frac{|\boldsymbol{a}|}{2}\right)\left(\|\ell h^{-1/2}[\![\psi]\!]_N\|_{0,\mathcal{F}_h^I}^2 + \|\ell h^{-1/2}\psi\|_{0,\mathcal{F}_h^D}^2\right)\\
&\le C\ell^4 h^{-2}\|\psi\|_{0,\Omega}
\end{aligned}$$

with C only depending on the shape regularity of the mesh and on $|\boldsymbol{a}|$; taking $C_{\rm lift}=\sqrt{C}$ completes the proof. ∎

We set, for brevity,

$$\mathcal{N}_g(t) := \|\ell\,h^{-1/2}g_D(\cdot,t)\|_{0,\mathcal{F}_h^D} + \|g_N(\cdot,t)\|_{0,\Gamma_N}.$$

PROPOSITION 4.5. *There exists a positive constant $C_{\rm rhs}$ independent of h and ℓ such that*

$$\left|\mathcal{P}_h^{n,(k)}(\psi)\right| \le C_{\rm rhs}\,\mathcal{N}_g(t^{n,(k)})\|\psi\|_{DG} \qquad \forall \psi \in \mathcal{V}_h.$$

Proof. We estimate the three terms of $|\mathcal{P}_h^n(\psi)|$ (see (4.7)) separately.

For the first one, the weighted Cauchy–Schwarz inequality immediately gives

(4.13)
$$\begin{aligned}\left|\int_{\mathcal{F}_h^D} \frac{\boldsymbol{\alpha}\cdot\boldsymbol{n}_f}{2} p(g_D(\cdot,t^{n,(k)}))\,\psi\,ds\right| &\le C\,\|\ell\,h^{-1/2} g_D(\cdot,t^{n,(k)})\|_{0,\mathcal{F}_h^D}\|\ell\,h^{-1/2}\psi\|_{0,\mathcal{F}_h^D}\\ &\le C\,\|\ell\,h^{-1/2} g_D(\cdot,t^{n,(k)})\|_{0,\mathcal{F}_h^D}\|\psi\|_{DG},\end{aligned}$$

where $C>0$ depends on the continuity constant of the nonlinearity p and on $|\boldsymbol{a}|$ but is independent of h and ℓ.

We consider now the second term. We observe that since $g_N(\cdot,t)\in L^2(\Gamma_N)$ for all t, there exists $\widetilde{\boldsymbol{g}}(\cdot,t)\in L^2(\Omega)^d$ with $\nabla\cdot\widetilde{\boldsymbol{g}}(\boldsymbol{x},t)=0$ in Ω for all t such that for all t,

$$\widetilde{\boldsymbol{g}}(\boldsymbol{x},t)\cdot\boldsymbol{n}_\Omega = g_N \text{ on } \Gamma_N, \qquad \|\widetilde{\boldsymbol{g}}(\cdot,t)\|_{0,\Omega}\le C_\Omega\|g_N(\cdot,t)\|_{0,\Gamma_N},$$

with the constant $C_\Omega>0$ only depending on Ω. This can be easily seen by considering the auxiliary problem

$$\begin{aligned} -\Delta\phi &= 0 && \text{in } \Omega,\\ \phi &= 0 && \text{on } \Gamma_D,\\ \nabla\phi\cdot\boldsymbol{n}_\Omega &= g_N && \text{on } \Gamma_N,\end{aligned}$$

which admits a unique solution ϕ satisfying $\|\nabla\phi\|_{0,\Omega}\le C_\Omega\|g_N(\cdot,t)\|_{0,\Gamma_N}$, and setting $\widetilde{\boldsymbol{g}}(\cdot,t):=\nabla\phi$ for all t.

Thus, we can rewrite the second term of $|\mathcal{P}_h^{n,(k)}(\psi)|$ as follows. We omit, for brevity, the dependence on t in the intermediate steps. Using the definition of $\mathcal{L}$, integrating by parts, and recalling that $\nabla\cdot\widetilde{\boldsymbol{g}}=0$ and $[\![\widetilde{\boldsymbol{g}}]\!]_N=0$ on all faces in $\mathcal{F}_h^I$, we have

$$\begin{aligned}\int_\Omega \widetilde{\boldsymbol{g}}\cdot(\nabla_h\psi-\mathcal{L}(\psi))\,d\boldsymbol{x} &= \int_\Omega \widetilde{\boldsymbol{g}}\cdot\nabla_h\psi\,d\boldsymbol{x} - \int_{\mathcal{F}_h^I}\{\!\{\widetilde{\boldsymbol{g}}\}\!\}\cdot[\![\psi]\!]_N\,ds - \int_{\mathcal{F}_h^D}\widetilde{\boldsymbol{g}}\cdot\boldsymbol{n}_\Omega\psi\,ds\\ &= -\int_\Omega \nabla\cdot\widetilde{\boldsymbol{g}}\psi\,d\boldsymbol{x} + \int_{\mathcal{F}_h^I}\left([\![\widetilde{\boldsymbol{g}}]\!]_N\{\!\{\psi\}\!\} + \{\!\{\widetilde{\boldsymbol{g}}\}\!\}\cdot[\![\psi]\!]_N\right)ds\\ &\quad + \int_{\mathcal{F}_h^D}\widetilde{\boldsymbol{g}}\cdot\boldsymbol{n}_\Omega\psi\,ds + \int_{\mathcal{F}_h^N}\widetilde{\boldsymbol{g}}\cdot\boldsymbol{n}_\Omega\psi\,ds\\ &\quad - \int_{\mathcal{F}_h^I}\{\!\{\widetilde{\boldsymbol{g}}\}\!\}\cdot[\![\psi]\!]_N\,ds - \int_{\mathcal{F}_h^D}\widetilde{\boldsymbol{g}}\cdot\boldsymbol{n}_\Omega\psi\,ds\\ &= \int_{\mathcal{F}_h^N}\widetilde{\boldsymbol{g}}\cdot\boldsymbol{n}_\Omega\psi\,ds,\end{aligned}$$

and thus we have the bound

(4.14)
$$\begin{aligned}\left|\int_{\mathcal{F}_h^N} g_N(\cdot,t^{n,(k)})\psi\,ds\right| &\le \|\widetilde{\boldsymbol{g}}(\cdot,t^{n,(k)})\|_{0,\Omega}\|\nabla_h\psi-\mathcal{L}(\psi)\|_{0,\Omega}\\ &\le C_\Omega\|g_N(\cdot,t^{n,(k)})\|_{0,\Gamma_N}\|\psi\|_{DG}.\end{aligned}$$

The lifting $\mathcal{G}_D^{n,(k)}$ of the Dirichlet boundary condition satisfies a bound similar to the bound (4.12) for the lifting $\mathcal{L}$ and which can be proved in a similar way:

$$\|\mathcal{G}_D^{n,(k)}\|_{0,\Omega} \le C_{\text{lift}} \|\ell\, h^{-1/2} g_D(\cdot, t^{n,(k)})\|_{0,\mathcal{F}_h^D} \|\psi\|_{DG}.$$

Therefore, for the third term, we have

$$\begin{aligned} \int_\Omega \mathcal{G}_D^{n,(k)} \cdot (\nabla_h \psi - \mathcal{L}(\psi))\, d\boldsymbol{x} &\le \|\mathcal{G}_D^{n,(k)}\|_{0,\Omega} \|(\nabla_h \psi - \mathcal{L}(\psi))\|_{0,\Omega} \\ &\le C_{\text{lift}} \|\ell\, h^{-1/2} g_D(\cdot, t^{n,(k)})\|_{0,\mathcal{F}_h^D} \|\psi\|_{DG}. \end{aligned} \tag{4.15}$$

Considering together (4.13), (4.14), and (4.15) completes the proof. ∎

4.3. Proof of L^2-stability. In the following proof, we restrict ourselves to the case of the one-stage time discretization (forward Euler method). The general case is not a straightforward consequence and is still an open issue.

The method in this case reads as follows:

Initialize: Set $u_h^0 = L^2$–projection of u_0 onto $\mathcal{V}_h$ (see (3.23)).

Time stepping: For $n = 0, 1, \ldots,$

(a) Compute

$$w_h^n = \texttt{relax\_w}(u_h^n); \tag{4.16}$$

(b) Compute $u_h^{n+1} \in \mathcal{V}_h$:

$$\int_\Omega \frac{u_h^{n+1} - u_h^n}{\Delta t}\, \psi_h\, d\boldsymbol{x} + \mathcal{A}_h(w_h^n, \psi_h) = \mathcal{P}_h^n(\psi_h) \qquad \forall \psi_h \in \mathcal{V}_h. \tag{4.17}$$

For the stability analysis, it is useful to rewrite the method as follows:

Initialize: Set $u_h^0 = L^2$–projection of u_0 onto $\mathcal{V}_h$ (see (3.23)).

Time stepping: For $n = 0, 1, \ldots,$

(a) Compute

$$q_h^n = \texttt{relax\_w}(u_h^n); \tag{4.18}$$

(b) Compute $q_h^{n+1} \in \mathcal{V}_h$:

$$\mu \int_\Omega \frac{q_h^{n+1} - q_h^n}{\Delta t}\, \psi_h\, d\boldsymbol{x} + \mathcal{A}_h(q_h^n, \psi_h) = \mathcal{P}_h^n(\psi_h) \qquad \forall \psi_h \in \mathcal{V}_h, \tag{4.19}$$

where μ is an arbitrary parameter such that $0 < \mu \le L_p^{-1}$, with L_p as in (1.3);

(c) Compute $u_h^{n+1} \in \mathcal{V}_h$:

$$u_h^{n+1} = u_h^n + \mu(q_h^{n+1} - q_h^n). \tag{4.20}$$

We notice that these two algorithms are equivalent; in fact, q_h^n in (4.18) is equal to w_h^n in (4.16), and combining this with (4.20) and (4.19) gives (4.17).

Remark 4.6. The only restriction on the choice of the positive constant μ in (4.19) and (4.20) depends on the nonlinearity p of the original problem (1.1), namely, μ has to satisfy $0 < \mu \le L_p^{-1}$, where L_p is an upper bound for p'. The stability properties of the method are related to the possible choices of μ as indicated in

Theorem 4.7 and Corollary 4.8: larger values of μ give a less-restrictive CFL condition and corresponding smaller stability constants. Practically, the knowledge of a sharper Lipschitz constant of p allows us to determine a less-restrictive CFL condition.

We prove our stability result following [34]. First, we introduce some notation.

Given an absolutely continuous function $\lambda : \mathbb{R} \to \mathbb{R}$ such that $\lambda(0) = 0$ and $0 \le \lambda' \le \Lambda < +\infty$, define the convex function

$$\Phi_\lambda := \int_0^s \lambda(z)\, dz \qquad \forall s \in \mathbb{R}.$$

Then Φ_λ satisfies

$$\frac{1}{2\Lambda}\lambda^2(s) \le \Phi_\lambda(s) \le \frac{\Lambda}{2}s^2 \qquad \forall s \in \mathbb{R}. \tag{4.21}$$

The function Φ_p thus satisfies (4.21). Moreover, setting

$$\beta := I - \mu p,$$

we have that

$$0 \le \beta'(s) \le 1 \qquad \text{for a.e. } s \in \mathbb{R},$$

and thus Φ_β also satisfies (4.21) with $\Lambda = 1$.

THEOREM 4.7. *Consider the method detailed at the beginning of section* 4.3. *Let* ξ *be any value in* $(0, 1)$. *Provided that*

$$\Delta t \le \frac{4\mu\, h^2}{3C_{\rm inv}^2 \ell^4}(1 - \xi), \tag{4.22}$$

for every time step m *we have*

$$\mu\, \|p(u_h^m)\|_{0,\Omega}^2 \le \frac{2}{\mu}\|u_h^0\|_{0,\Omega}^2 + C_{\rm rhs}^2(m\,\Delta t)\left[\frac{1}{\xi^2} + \frac{4(1-\xi^2)}{\xi}\right](\mathcal{N}_g^m)^2, \tag{4.23}$$

where we set

$$\mathcal{N}_g^m := \max_{0 \le n \le m-1} \mathcal{N}_g(t^n).$$

Before proceeding with the proof of Theorem 4.7, we highlight that the constant ξ in (4.22) and (4.23) can be chosen arbitrarily in $(0, 1)$. Clearly, if we choose a smaller ξ, the condition (4.22) on Δt is less restrictive, but the price to pay is a larger stability constant in (4.23). For $\xi = 0.5$, for instance, we have both reasonable CFL conditions and stability constants.

Proof of Theorem 4.7. We denote by $(\cdot, \cdot)$ the $L^2(\Omega)$-inner product, for simplicity, and by Π_h the $L^2(\Omega)$-projection onto $\mathcal{V}_h$.

Step 1. Take $\psi_h = \Delta t\, q_h^n$ in (4.19). Using (4.20), we obtain

$$\begin{aligned}
\Delta t\, \mathcal{P}_h^n(q_h^n) &= \mu(q_h^{n+1} - q_h^n, q_h^n) + \Delta t\, \mathcal{A}_h(q_h^n, q_h^n) \\
&= \mu(q_h^{n+1} - q_h^n, q_h^{n+1}) - \mu(q_h^{n+1} - q_h^n, q_h^{n+1} - q_h^n) + \Delta t\, \|q_h^n\|_{DG}^2 \\
&= (u_h^{n+1} - u_h^n, q_h^{n+1}) - \frac{1}{\mu}\|u_h^{n+1} - u_h^n\|_{0,\Omega}^2 + \Delta t\, \|q_h^n\|_{DG}^2.
\end{aligned}$$

Adding over $n = 0, \dots, (m-1)$ gives

$$\begin{aligned}\Delta t \sum_{n=0}^{m-1} \mathcal{P}_h^n(q_h^n) \; &= \sum_{n=0}^{m-1} (u_h^{n+1} - u_h^n, q_h^{n+1}) \\ &\quad - \frac{1}{\mu} \sum_{n=0}^{m-1} \|u_h^{n+1} - u_h^n\|_{0,\Omega}^2 + \Delta t \sum_{n=0}^{m-1} \|q_h^n\|_{DG}^2.\end{aligned} \tag{4.24}$$

We bound $\sum_{n=0}^{m-1}(u_h^{n+1} - u_h^n, q_h^{n+1})$ following [34]. For completeness, we report here the complete proof.

From (4.20) and $q_h^n = \Pi_h(p(u_h^n))$, we have

$$q_h^{n+1} = \frac{1}{\mu}(u_h^{n+1} - u_h^n) + \Pi_h(p(u_h^n)),$$

and, taking into account the definition $\beta = I - \mu p$, we can write

$$\begin{aligned}q_h^{n+1} = \frac{1}{2}\Pi_h(p(u_h^{n+1})) + \frac{1}{2\mu}[\Pi_h(\beta(u_h^{n+1})) - \Pi_h(\beta(u_h^n))] \\ + \frac{1}{2\mu}u_h^{n+1} - \frac{1}{2\mu}\Pi_h(\beta(u_h^n)).\end{aligned}$$

Therefore, using $\Phi_\lambda' = \lambda$ and the convexity of Φ_λ, for both $\lambda = p$ and $\lambda = \beta$, the fact that $\beta' \geq 0$, and the identity

$$2a(a-b) = a^2 - b^2 + (a-b)^2 \qquad \forall a, b \in \mathbb{R},$$

we obtain

$$\begin{aligned}(u_h^{n+1} - u_h^n, q_h^{n+1}) \geq \frac{1}{2}\int_\Omega [\Phi_p(u_h^{n+1}) - \Phi_p(u_h^n)]\, d\boldsymbol{x} + 0 \\ + \frac{1}{4\mu}\left(\|u_h^{n+1}\|_{0,\Omega}^2 - \|u_h^n\|_{0,\Omega}^2 + \|u_h^{n+1} - u_h^n\|_{0,\Omega}^2\right) \\ + \frac{1}{2\mu}\int_\Omega [\Phi_\beta(u_h^n) - \Phi_\beta(u_h^{n+1})]\, d\boldsymbol{x}.\end{aligned}$$

Adding over $n = 0, \dots, (m-1)$ gives

$$\begin{aligned}\sum_{n=0}^{m-1}(u_h^{n+1} - u_h^n, q_h^{n+1}) \geq \frac{1}{2}\int_\Omega [\Phi_p(u_h^m) - \Phi_p(u_h^0)]\, d\boldsymbol{x} \\ + \frac{1}{4\mu}\left[\|u_h^m\|_{0,\Omega}^2 - \|u_h^0\|_{0,\Omega}^2 + \sum_{n=0}^{m-1}\|u_h^{n+1} - u_h^n\|_{0,\Omega}^2\right] \\ + \frac{1}{2\mu}\int_\Omega [\Phi_\beta(u_h^0) - \Phi_\beta(u_h^m)]\, d\boldsymbol{x},\end{aligned}$$

and using (4.21) with $\lambda = p$ and $\lambda = \beta$, where $\Lambda = \mu^{-1}$ if $\lambda = p$, and $\Lambda = 1$ if $\lambda = \beta$, we get

$$\sum_{n=0}^{m-1}(u_h^{n+1} - u_h^n, q_h^{n+1}) \geq \frac{\mu}{4}\|p(u_h^m)\|_{0,\Omega}^2 - \frac{1}{2\mu}\|u_h^0\|_{0,\Omega}^2 + \frac{1}{4\mu}\sum_{n=0}^{m-1}\|u_h^{n+1} - u_h^n\|_{0,\Omega}^2. \tag{4.25}$$

For the term $\Delta t \sum_{n=0}^{m-1} \mathcal{P}_h^n(q_h^n)$, Proposition 4.5 and the Young inequality give

$$\mathcal{P}_h^n(q_h^n) \le \frac{C_{\rm rhs}^2}{2\eta}\mathcal{N}_g(t^n)^2 + \frac{\eta}{2}\|q_h^n\|_{DG}^2$$

with $\eta > 0$, and thus

$$\Delta t \sum_{n=0}^{m-1} \mathcal{P}_h^n(q_h^n) \le \frac{C_{\rm rhs}^2 m\,\Delta t}{2\eta}(\mathcal{N}_g^m)^2 + \frac{\eta\,\Delta t}{2}\sum_{n=0}^{m-1}\|q_h^n\|_{DG}^2. \tag{4.26}$$

Therefore, inserting (4.25) and (4.26) into (4.24), we obtain

$$\begin{aligned}\Delta t \sum_{n=0}^{m-1} \|q_h^n\|_{DG}^2 + \frac{\mu}{4}\|p(u_h^m)\|_{0,\Omega}^2 \le{}& \frac{1}{2\mu}\|u_h^0\|_{0,\Omega}^2 + \frac{3}{4\mu}\sum_{n=0}^{m-1}\|u_h^{n+1}-u_h^n\|_{0,\Omega}^2 \\ &+ \frac{C_{\rm rhs}^2 m\,\Delta t}{2\eta}(\mathcal{N}_g^m)^2 + \frac{\eta\,\Delta t}{2}\sum_{n=0}^{m-1}\|q_h^n\|_{DG}^2.\end{aligned} \tag{4.27}$$

Step 2. Take $\psi_h = \frac{\Delta t}{\mu}(q_h^{n+1} - q_n^n)$ in (4.19). By applying the continuity of $\mathcal{A}_h(\cdot,\cdot)$, Proposition 4.5, and the inverse inequality of Lemma 4.4, we obtain

$$\begin{aligned}\|q_h^{n+1}-q_n^n\|_{0,\Omega}^2 \le{}& \frac{\Delta t}{\mu}\|q_h^n\|_{DG}\|q_h^{n+1}-q_n^n\|_{DG} \\ &+ \frac{\Delta t}{\mu}C_{\rm rhs}\,\mathcal{N}_g(t^n)\|q_h^{n+1}-q_n^n\|_{DG} \\ \le{}& \frac{\Delta t}{\mu}C_{\rm inv}\ell^2 h^{-1}[\|q_h^n\|_{DG} + C_{\rm rhs}\,\mathcal{N}_g(t^n)]\|q_h^{n+1}-q_n^n\|_{0,\Omega},\end{aligned}$$

and, due to (4.20),

$$\|u_h^{n+1}-u_n^n\|_{0,\Omega} \le \Delta t C_{\rm inv}\ell^2 h^{-1}[\|q_h^n\|_{DG} + C_{\rm rhs}\,\mathcal{N}_g(t^n)].$$

Let $\xi \in (0,1)$; since

$$(a+b)^2 \le (1+\xi)a^2 + (1+\xi^{-1})b^2 \qquad \forall a,b\in\mathbb{R},\ a,b>0,$$

we have

$$\|u_h^{n+1}-u_n^n\|_{0,\Omega}^2 \le \Delta t^2 C_{\rm inv}^2\ell^4 h^{-2}[(1+\xi)\|q_h^n\|_{DG}^2 + (1+\xi^{-1})C_{\rm rhs}^2\mathcal{N}_g(t^n)^2]. \tag{4.28}$$

Step 3. Insert (4.28) into (4.27). We get

$$\begin{aligned}\Delta t \sum_{n=0}^{m-1} \|q_h^n\|_{DG}^2 + \frac{\mu}{4}\|p(u_h^m)\|_{0,\Omega}^2 \le{}& \frac{1}{2\mu}\|u_h^0\|_{0,\Omega}^2 + \frac{C_{\rm rhs}^2 m\,\Delta t}{2\eta}(\mathcal{N}_g^m)^2 \\ &+ \frac{3m\,\Delta t^2 C_{\rm inv}^2 C_{\rm rhs}^2\ell^4}{4\mu\,h^2}(1+\xi^{-1})(\mathcal{N}_g^m)^2 \\ &+ \left(\frac{3\Delta t C_{\rm inv}^2\ell^4}{4\mu\,h^2}(1+\xi) + \frac{\eta}{2}\right)\Delta t\sum_{n=0}^{m-1}\|q_h^n\|_{DG}^2.\end{aligned}$$

Take $\eta = 2\xi^2$. Then, if (4.22) is satisfied, we have

$$(1-\xi^2) - \frac{3\Delta t C_{\rm inv}^2 \ell^4}{4\mu\, h^2}(1+\xi) \geq 0,$$

and thus

$$\frac{\mu}{4}\|p(u_h^m)\|_{0,\Omega}^2 \leq \frac{1}{2\mu}\|u_h^0\|_{0,\Omega}^2 + \frac{C_{\rm rhs}^2 m\,\Delta t}{4\xi^2}(\mathcal{N}_g^m)^2 + \frac{3m\,\Delta t^2 C_{\rm inv}^2 C_{\rm rhs}^2 \ell^4}{4\mu\, h^2}(1+\xi^{-1})(\mathcal{N}_g^m)^2,$$

which, taking into account (4.22), gives

$$\frac{\mu}{4}\|p(u_h^m)\|_{0,\Omega}^2 \leq \frac{1}{2\mu}\|u_h^0\|_{0,\Omega}^2 + C_{\rm rhs}^2 m\,\Delta t\left[\frac{1}{4\xi^2} + (1+\xi^{-1})(1-\xi)\right](\mathcal{N}_g^m)^2,$$

which concludes the proof. ∎

In the following corollary, we transfer to $\{u_h^m\}$ the stability result proved for $\{p(u_h^m)\}$. A strict monotonicity of the nonlinearity p toward $+\infty$ is needed in the proof; this is guaranteed by assumption (1.4).

COROLLARY 4.8. *Provided that condition* (4.22) *is satisfied, for every time step* m *we have*

$$\|u_h^m\|_{0,\Omega}^2 \leq \frac{2C_p^2}{\mu^2}\|u_h^0\|_{0,\Omega}^2 + \frac{C_p^2 C_{\rm rhs}^2}{\mu}(m\,\Delta t)\left[\frac{1}{\xi^2} + \frac{4(1-\xi^2)}{\xi}\right](\mathcal{N}_g^m)^2 + s_0^2\,|\Omega|^2,$$

where we have denoted by C_p *the continuity constant of* p^{-1} *in* $[s_0, +\infty)$.

Proof. Set $S_m := \{\boldsymbol{x} \in \Omega :\ |u_h^m(\boldsymbol{x})| \geq s_0\}$. Then, we have

$$\|u_h^m\|_{0,\Omega}^2 = \|u_h^m\|_{0,S_m}^2 + \|u_h^m\|_{0,\Omega\setminus S_m}^2 \leq C_p\|p(u_h^m)\|_{0,\Omega}^2 + s_0^2\,|\Omega|^2,$$

and Theorem 4.7 allows us to conclude. ∎

5. Numerical results. In this section, we show the results of several numerical simulations for both the linear and the nonlinear diffusion equation (1.1), in one and two space dimensions. In the matter of coupling spatial and temporal approximation orders, we notice that if the solution of the continuous problem is smooth enough, when we use polynomial reconstructions of degree ℓ, we expect a $O(h^{\ell+1})$ error (measured in the L^2-norm), while when we integrate in time using a Runge–Kutta scheme of order r, we expect a $O(\Delta t^r)$ error. Since the time integration step Δt is subdued to the parabolic stability constraint (4.22), i.e., $\Delta t = O(h^2)$, the order of convergence of the time discretization is given by $O(\Delta t^r) = O(h^{2r})$. In accordance with this, to match in an optimal way the spatial and the temporal reconstructions, when the solution is regular it is natural to couple spatial elements of degree ℓ together with Runge–Kutta schemes of order $r = \lceil(\ell+1)/2\rceil$. In particular, for the time integration, the coefficients of the explicit parts of the IMEX schemes we use are

$$A_1 = \begin{pmatrix}1\end{pmatrix},\ B_1 = \begin{pmatrix}1\end{pmatrix}, \qquad A_2 = \begin{pmatrix}1 & 0\\ \frac{1}{2} & \frac{1}{2}\end{pmatrix},\ B_2 = \begin{pmatrix}1 & 0\\ 0 & \frac{1}{2}\end{pmatrix},$$

$$A_3 = \begin{pmatrix}1 & 0 & 0\\ \frac{3}{4} & \frac{1}{4} & 0\\ \frac{1}{3} & 0 & \frac{2}{3}\end{pmatrix},\ B_3 = \begin{pmatrix}1 & 0 & 0\\ 0 & \frac{1}{4} & 0\\ 0 & 0 & \frac{2}{3}\end{pmatrix},$$

where A_j and B_j refer to a scheme of order j written in the form (3.5).

5.1. Linear diffusion tests. In the following simulations we consider the two-dimensional heat equation in the domain $\Omega = [-1, 1]^2$, coupled with Dirichlet boundary conditions:

$$\begin{aligned} &\frac{\partial u}{\partial t} - \Delta u = 0 && \text{in } \Omega \times (0, T), \\ &u(x, t) = x && \text{on } \partial\Omega \times (0, T), \\ &u(x, 0) = x + \sin(\pi x) \sin(\pi y) && \text{in } \Omega, \end{aligned} \tag{5.1}$$

with analytical solution

$$u(x, y, t) = x + \exp(-2\pi^2 t) \sin(\pi x) \sin(\pi y).$$

We compute the errors and test the convergence rates of the scheme described in section 3 as we vary the polynomial degree ℓ of the spatial discretization and the order r of the Runge–Kutta scheme used for the time integration.

In section 4, we proved L^2-stability in the case of $\boldsymbol{\alpha}$ defined by (4.1). This analysis was restricted to the case of one-stage time discretization, but the numerical results reported below seem to indicate that the scheme is stable also when several time stages are used.

In the following simulations, we are also going to test the effects of the choice of the stabilization parameter $\boldsymbol{\alpha}$ independent of the mesh size h (i.e., $|\boldsymbol{\alpha}| = O(1)$ instead of $|\boldsymbol{\alpha}| = O(h^{-1})$).

We performed the simulations using six subsequently refined unstructured meshes with mesh size $h = 1.70, 0.85, 0.47, 0.26, 0.13, 0.06$, respectively, avoiding further refining the mesh when the computed error has already reached machine precision. The obtained results are shown in Tables 5.1 and 5.2, which gather the errors in the L^2-norm and the estimated convergence rates, respectively.

For $\ell \geq 1$ and $|\boldsymbol{\alpha}| = O(h^{-1})$, these experiments confirm that optimal expected rates $\ell + 1$ are achieved. Moreover, they show that the choice $|\boldsymbol{\alpha}| = O(1)$ is effective. In fact, if we compare the corresponding convergence rates and the errors for $|\boldsymbol{\alpha}| = O(h^{-1})$ and $|\boldsymbol{\alpha}| = O(1)$, we can see that in most cases, we have similar results, even if we can highlight a more regular behavior for the simulations with $|\boldsymbol{\alpha}| = O(h^{-1})$. Despite the lack of theoretical result for the case of piecewise constant elements, we also tested our scheme for $\ell = 0$; it seems that in this case, first order convergence is achieved, at least with the choice $|\boldsymbol{\alpha}| = O(1)$. Very similar results were also obtained on structured grids constructed by dividing each element of Cartesian grids into two triangles.

In the next simulation, we investigate the effect of the stabilization coefficient $\boldsymbol{\alpha} \cdot \boldsymbol{n}_f$, which is a nonstandard DG-stabilization (there is no stabilization on edges parallel to $\boldsymbol{\alpha}$). In order to do this, we consider the following linear heat equation in the domain $\Omega = [0, 4]^2$ with a Dirichlet condition on the boundary $\Gamma_D = [0, 4] \times \{0, 4\}$ and a Neumann condition on the boundary $\Gamma_N = \{0, 4\} \times [0, 4]$:

$$\begin{aligned} &\frac{\partial u}{\partial t} - \Delta u = 0 && \text{in } \Omega \times (0, T), \\ &u(x, t) = x && \text{on } \Gamma_D \times (0, T), \\ &\nabla u \cdot \boldsymbol{n}_\Omega = 0 && \text{on } \Gamma_N \times (0, T), \\ &u(x, 0) = x + \sin\left(\frac{\pi}{2} x\right) && \text{in } \Omega \end{aligned} \tag{5.2}$$

TABLE 5.1

L^2-norm of the absolute errors of the numerical solution for the linear test problem (5.1) for several spatial and temporal approximations and for two different choices of the stabilization parameter $\boldsymbol{\alpha}$.

	$h = 1.70$	$h = 0.85$	$h = 0.47$	$h = 0.26$	$h = 0.13$	$h = 0.06$
$\ell = 0,\ r = 1$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	5.95e-01	3.77e-01	1.93e-01	1.07e-01	5.56e-02	3.03e-02
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	5.94e-01	3.75e-01	2.01e-01	1.22e-01	6.78e-02	4.52e-02
$\ell = 1,\ r = 1$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	2.95e-01	1.01e-01	3.08e-02	8.21e-03	2.12e-03	5.36e-04
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	2.95e-01	1.03e-01	3.13e-02	8.38e-03	2.18e-03	5.51e-04
$\ell = 2,\ r = 2$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	8.57e-02	2.40e-02	2.93e-03	3.76e-04	4.66e-05	5.84e-06
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	8.61e-02	2.38e-02	2.93e-03	3.79e-04	4.69e-05	5.85e-06
$\ell = 3,\ r = 2$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	4.70e-02	3.03e-03	2.34e-04	1.52e-05	9.70e-07	6.09e-08
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	4.78e-02	3.11e-03	2.41e-04	1.55e-05	9.81e-07	6.12e-08
$\ell = 4,\ r = 3$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	8.53e-03	5.09e-04	1.65e-05	5.89e-07	1.81e-08	5.65e-10
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	8.54e-03	5.26e-04	1.74e-05	6.26e-07	1.91e-08	5.93e-10
$\ell = 5,\ r = 3$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	4.31e-03	5.75e-05	1.18e-06	1.94e-08	3.10e-10	
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	4.33e-03	6.00e-05	1.26e-06	2.10e-08	3.35e-10	

TABLE 5.2

Computational convergence rates of the numerical solution for the linear test problem (5.1) for several spatial and temporal approximations, and for two different choices of the stabilization parameter $\boldsymbol{\alpha}$.

	$h = 1.70$	$h = 0.85$	$h = 0.47$	$h = 0.26$	$h = 0.13$	$h = 0.06$
$\ell = 0,\ r = 1$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	–	0.658	1.120	0.994	0.916	0.918
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	–	0.663	1.042	0.843	0.848	0.592
$\ell = 1,\ r = 1$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	–	1.539	1.859	2.077	1.922	2.044
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	–	1.521	1.857	2.067	1.915	2.040
$\ell = 2,\ r = 2$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	–	1.833	3.290	3.218	2.969	3.085
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	–	1.857	3.272	3.207	2.970	3.091
$\ell = 3,\ r = 2$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	–	3.955	4.003	4.287	3.910	4.112
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	–	3.941	3.997	4.306	3.920	4.120
$\ell = 4,\ r = 3$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	–	4.068	5.352	5.234	4.948	5.002
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	–	4.020	5.330	5.212	4.957	5.00
$\ell = 5,\ r = 3$						
$\lvert\boldsymbol{\alpha}\rvert = O(1)$	–	6.231	6.069	6.442	5.881	
$\lvert\boldsymbol{\alpha}\rvert = O(h^{-1})$	–	6.174	6.033	6.427	5.876	

with analytical solution

$$u(x,t) = x + \exp\left(-\frac{\pi^2}{4}t\right)\sin\left(\frac{\pi}{2}x\right).$$

If in this situation we choose $\boldsymbol{\alpha} = (1,0)^T$, $\boldsymbol{\alpha}\cdot\boldsymbol{n}_f$ vanishes on the whole Γ_D. Since the stabilization coefficient is not present on the Neumann boundary (see (3.19), (3.20), and (3.21)), no boundary edge is stabilized. In order to reduce as much as possible the number of interior edges on which the stabilization acts, the domain Ω is discretized

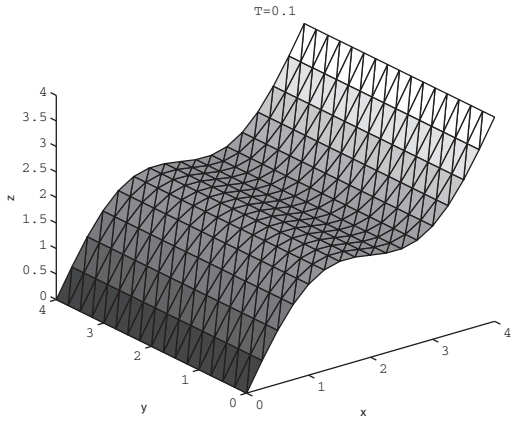

FIG. 5.1. *Solution of problem* (5.2) *at* $T = 0.1$: *even if the stabilization coefficient* $\boldsymbol{\alpha} \cdot \boldsymbol{n}_f$ *vanishes on all the boundary edges, the scheme is still stable.*

through structured meshes obtained from Cartesian grids as before. The proof we gave in section 4 stated that in this situation the scheme is stable. We tested numerically the evidence of this fact, and in Figure 5.1 we can see the stable numerical solution obtained for (5.2) at time $T = 0.1$ on a grid with 512 triangles, using a piecewise discontinuous quadratic elements and the second order Runge–Kutta method.

As mentioned in Remark 4.1, in the linear case our space discretization is a modified LDG method. More precisely, it is an LDG method with reduced stabilization, again with the notation of [10], while in the standard LDG method the stabilization parameter C_{11} is strictly positive in all the interior and Dirichlet boundary faces, in our scheme $(C_{11})_{|_f} = 0$ whenever $\boldsymbol{\alpha} \cdot \boldsymbol{n}_f = 0$. There are other LDG methods with reduced stabilization in the literature. We mention here the one proposed in [16], where an artificial wind is introduced, and choosing the coefficient $\boldsymbol{C}_{12}$ in a suitable way allows for removing the stabilization from interior and inflow boundary faces, without altering the convergence properties. A similar approach is the one studied in [35]: on structured quadrilateral meshes, it is proved that it is enough to stabilize only on the part of the Dirichlet boundary which is inflow with respect to the vector-valued coefficient $\boldsymbol{C}_{12}$. (Also in this case, the role of $\boldsymbol{C}_{12}$ is important.) A different philosophy is applied in [9]. There, the LDG method with $\boldsymbol{C}_{12} = \boldsymbol{0}$ is considered, like in our case, and the stabilization reduction consists in penalizing only some of the jump modes (either the lower or the higher ones), but on all faces. Provided that the number of the penalized modes is suitably chosen, this method possesses the same convergence properties as the original LDG method.

5.2. Nonlinear diffusion tests. We consider in $\Omega = [-4, 4]^2$ the porous media equation

$$\frac{\partial u}{\partial t} - \Delta u^m = 0 \qquad \text{in } \Omega \times (0, T) \tag{5.3}$$

with homogeneous Dirichlet boundary conditions. Equation (5.3) degenerates for $u = 0$, since $p'(u) = 0$; thus, compactly supported initial data give rise to solutions with interfaces that travel with finite speeds, as the well-known similarity solution studied by Barenblatt (see, for example, [6]).

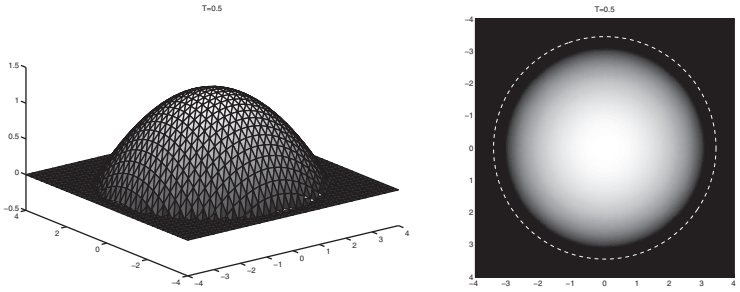

FIG. 5.2. *Numerical solution of problem* (5.3) *with Barenblatt initial datum at final time* $T = 0.5$ *seen in isometric perspective (left plot) and from above (right plot). In the right plot, the white line represents the contour of the support of the analytical solution.*

Simulations show that our scheme applied to problem (5.3) is stable and accurate. In Figure 5.2, we report the solution of (5.3) with $m = 2$ obtained evolving the Barenblatt solution until final time $T = 0.5$. The simulation uses piecewise discontinuous cubic elements and forward Euler for the time integration; we choose the stability parameter $\boldsymbol{\alpha} = h^{-1}(1, 0)$. As we can see, the shape and the symmetry of the solution are correctly represented, and the speed of the traveling front is correctly approximated by the numerical scheme. We have observed that these properties are independent from the choice of the direction of the parameter $\boldsymbol{\alpha}$.

Another interesting nonlinear test also involves (5.3), but in this case we consider $m = 3$ and we take as initial datum the C^1 function

$$u(x, y) = \begin{cases} \frac{1}{2}\cos^2(\pi\sqrt{x^2 + y^2}/2), & \sqrt{x^2 + y^2} \le 1, \\ 0, & \sqrt{x^2 + y^2} > 1. \end{cases} \tag{5.4}$$

This test is a two-dimensional version of that proposed in [25]. As shown in [4] for the one-dimensional case, the solution with initial condition (5.4) develops a discontinuity in ∇u at some finite time $T_0 > 0$ and has a front that initially stands still and then starts expanding at a certain finite time T_1. In Figure 5.3, we can see four different situations of the evolution of (5.4): first the solution is regular, then it starts losing regularity, and finally the front expands.

We remark that since the exact solutions of the tests we performed on (5.3) present discontinuities in both ∇u and $\partial_t u$, we cannot expect to increase the convergence rate by raising the approximation orders: in fact, all the simulation we performed achieved the 1.5 limit convergence rate. However, we observe that in some applications, for example, in image processing methods based on PDEs, we have to handle a fixed computational grid; then one could take advantage of a higher order method to get a lower error.

We consider now problem (5.3) with $m = 2$ in one space dimension with initial condition given by the one-dimensional Barenblatt function. In Table 5.3, we compare the errors of the simulations at the same final time $t = 2$ for two different fixed numbers of spatial elements ($N = 20$ and $N = 640$) and for several spatial and temporal approximations with stabilization parameter $\boldsymbol{\alpha} = h^{-1}(1, 0)$. Also in this case, rotating $\boldsymbol{\alpha}$ does not significantly affect the simulation results. We can see that increasing the degree of the spatial reconstruction considerably decreases the error, while passing from explicit Euler to higher order Runge–Kutta schemes seems not to give much benefit: we remark that explicit Runge Kutta schemes are very sensitive with respect to the regularity of the solution.

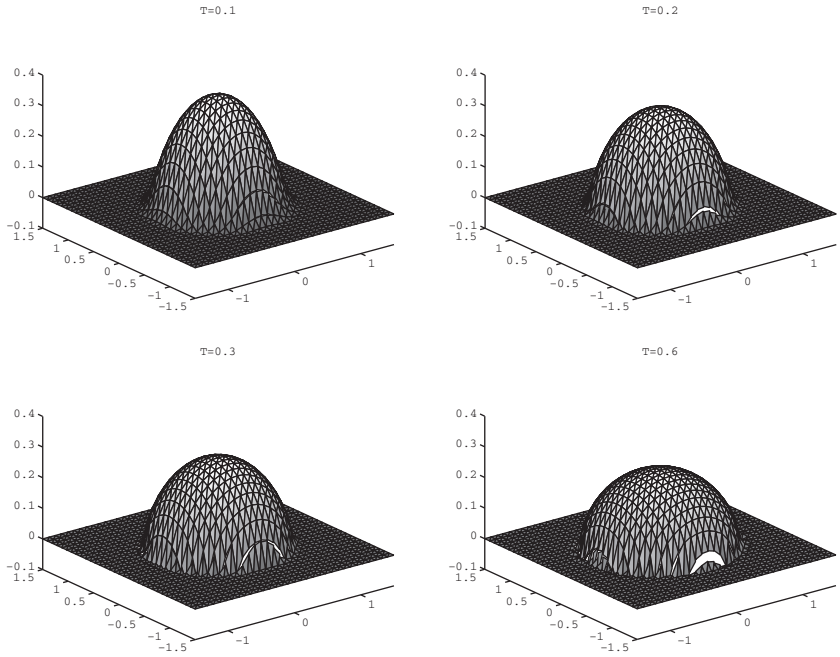

FIG. 5.3. *Numerical solution of problem* (5.3) *with* $m = 3$ *and with initial datum* (5.4). *Upper left figure: initially the solution is regular and the front does not evolve. Upper right figure: the solution starts to develop the discontinuity in* ∇u. *Lower left figure: the solution has lost regularity but the front still does not evolve. Lower right figure: the front is moving.*

TABLE 5.3

L^2-norm errors for the solution of the problem (5.3) *in one dimension for two different numbers N of elements with different degrees ℓ of polynomial reconstructions and orders r of Runge–Kutta methods. Increasing the spatial reconstruction degree allows appreciably more accurate solutions, while a higher order time integration method has little affect on the accuracy of the solution.*

$N = 20$			
	$\ell = 1$	$\ell = 2$	$\ell = 3$
$r = 1$	1.265e-02	9.350e-03	2.273e-03
$r = 2$	1.185e-02	8.817e-03	2.224e-03
$r = 3$	1.187e-02	8.818e-03	2.224e-03
$N = 640$			
	$\ell = 1$	$\ell = 2$	$\ell = 3$
$r = 1$	9.5488e-05	5.4739e-05	9.981e-06
$r = 2$	9.4351e-05	5.4349e-05	9.9148e-06
$r = 3$	9.4351e-05	5.4349e-05	9.9148e-06

As pointed out at the end of section 3.2, the L^2-stability does not grant that the discrete solution reproduce some properties of the continuous one, like positivity or monotonicity preserving. All the solutions we computed in the nonlinear tests showed oscillations near the fronts and the positivity of the solution was not preserved. This can be recovered at least partially with the use of suitable limiting techniques, as described in section 3.2. In Figure 5.4, we compare the solutions obtained for the above one-dimensional problem, with piecewise linear polynomial approximation in space and explicit Euler, with and without slope limiter (we have used the min-mod

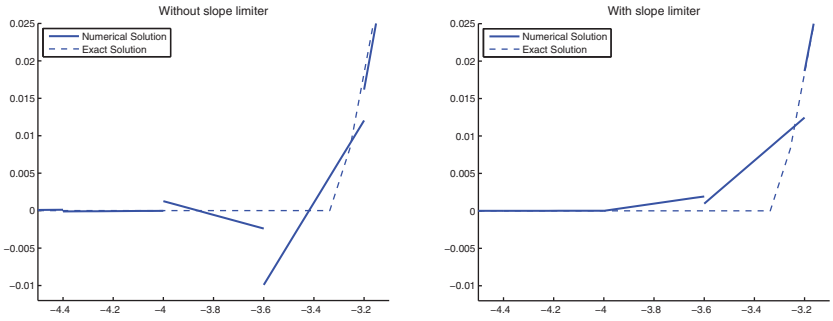

FIG. 5.4. *A detail of the solution of problem* (5.3) *at* $t = 2$: *in the left plot the solution is obtained with no slope limiting techniques and presents oscillations; in the right plot we can see the solution obtained with the scheme corrected by the min-mod slope limiter.*

slope limiter described in [17]): the oscillations of the nonlimited solution are smeared out by the slope limiter.

We remark that the slope limiter we used in this simulations does not grant that the resulting scheme be positivity preserving, since, as shown in [17], it only grants that the scheme be total variation diminishing in means. We do not further elaborate here on slope limiters, and thus we do not test slope limiters in two dimensions nor different limiting approaches in one dimension.

Finally we present the following test:

$$\begin{aligned} \frac{\partial u}{\partial t} - \frac{\partial^2 g(u)}{\partial x^2} &= 0 && \text{in } [-1,1]\times(0,T), \\ u(-1,t) = u(1,t) &= 0 && \text{for } t \geq 0, \\ u(x,0) &= \chi_{|x|\leq 1/2} && \text{in } [-1,1], \end{aligned} \tag{5.5a}$$

where

$$g(x) = \begin{cases} \dfrac{5}{4}x^2 - \dfrac{5}{4}x + \dfrac{5}{16} & \text{if } 0.5 < x < 0.6, \\ \dfrac{1}{4}x - \dfrac{11}{80} & \text{if } x \geq 0.6, \\ 0 & \text{otherwise}, \end{cases} \tag{5.5b}$$

is a strongly degenerate diffusion function vanishing over $[0, 0.5]$. The previous equation is the adaption of a similar test for convection-diffusion equations which can be found in [20].

In our analysis, we did not address the possibility for a strongly degenerate equation to present more than a weak solution, among which one physical solution has to be selected by introducing an entropy function, as in the case of hyperbolic conservation laws. For a detailed presentation of the entropy issue for the parabolic equations, see [20] and references therein.

The entropic solution of problem (5.5) is showed in Figure 5.5, left, and is computed with an entropic scheme on a very fine grid. In Figure 5.5, right, the solution obtained with a nonentropic finite difference scheme is reported. With our symmetric scheme, using piecewise linear elements on a mesh of 80 elements, and forward Euler in time, either with or with no slope limiter, the right entropy solution is obtained, as reported in Figure 5.6.

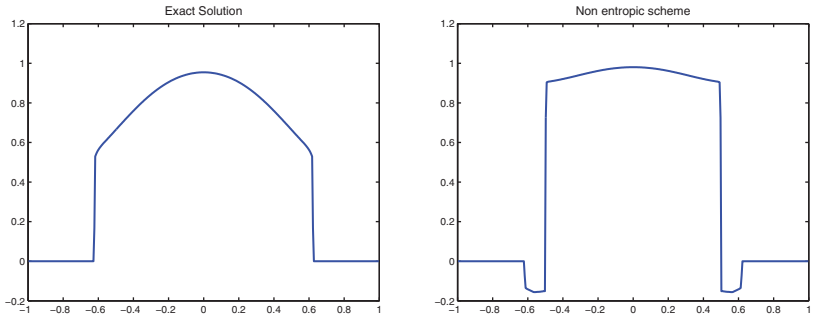

FIG. 5.5. *Solutions of* (5.5)*: the left plot represents the exact solution obtained with an entropic scheme on a very fine grid, and the right plot is obtained with a finite difference non entropic scheme.*

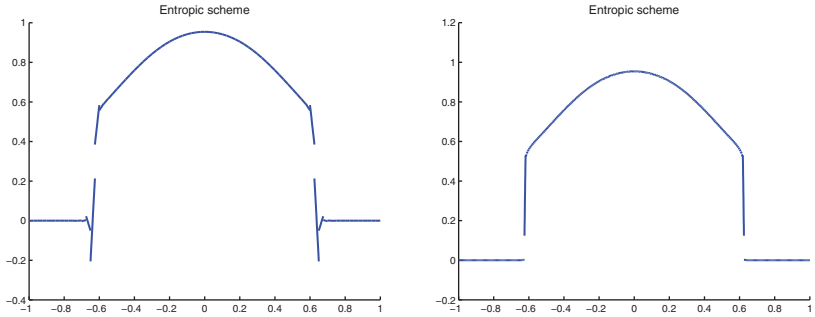

FIG. 5.6. *Numerical solutions of* (5.5a)*: the left plot refers to the scheme without slope limiters; the right one is obtained with the min-mod slope limiter. Both simulations provide the right entropic solution.*

6. Conclusions. In this work we have introduced the first finite element schemes for the approximation of linear and nonlinear diffusion problems, based on diffusive relaxation framework. Our methods couple explicit Runge–Kutta time stepping and DG space discretizations. Preliminary theoretical results are provided and some comparisons with other DG methods are discussed. Several one- and two-dimensional numerical tests are discussed in order to point out the main properties and the effectiveness of these schemes for linear and nonlinear (degenerate) diffusion.

For nonlinear problems, some of the advantages of the presented method are that we do not need to solve nonlinear systems and that the numerical solutions inherit positivity and monotonicity (at least in the means) properties of the analytical solutions. On the other hand, its stability requires the standard parabolic CFL condition, which constrains the time step to be proportional to the square of the mesh size.

In future work, we plan to further analyze the nonlinear stability and the convergence of these schemes to investigate different slope limiting techniques, also in the multidimensional case, and to look for the possibility of implicit time discretizations, in order to avoid the time step restriction given by the parabolic CFL condition. This approach can be extended to the discretization of convection-diffusion equations and to nonlinear fourth order diffusion equations, like the thin film equation, in the presence of degenerate and strongly degenerate diffusion.

Acknowledgments. The authors wish to thank the reviewers and the editor for their careful reading of the manuscript and their helpful comments and suggestions.

They are also extremely grateful to Paola Antonietti for providing her DG-FEM code for elliptic problems, which constituted the starting point for their implementation.

REFERENCES

[1] D. Aregba-Driollet and R. Natalini, *Discrete kinetic schemes for multidimensional systems of conservation laws*, SIAM J. Numer. Anal., 37 (2000), pp. 1973–2004.

[2] D. Aregba-Driollet, R. Natalini, and S. Tang, *Explicit diffusive kinetic schemes for nonlinear degenerate parabolic systems*, Math. Comp., 73 (2004), pp. 63–94.

[3] 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 (2001/02), pp. 1749–1779.

[4] D. G. Aronson, *Regularity properties of flows through porous media: A counterexample*, SIAM J. Appl. Math., 19 (1970), pp. 299–307.

[5] U. Asher, S. Ruuth, and R. J. Spiteri, *Implicit-explicit Runge-Kutta methods for time dependent partial differential equations*, Appl. Numer. Math., 25 (1997), pp. 151–167.

[6] G. I. Barenblatt, *Scaling, Self-Similarity, and Intermediate Asymptotics*, Cambridge Texts Appl. Math. 14, Cambridge University Press, Cambridge, UK, 1996.

[7] F. Bouchut, *Entropy satisfying flux vector splittings and kinetic BGK models*, Numer. Math., 94 (2003), pp. 623–672.

[8] E. Burman, A. Ern, and M. A. Fernández, *Explicit Runge-Kutta schemes and finite elements with symmetric stabilization for first-order linear PDE systems*, SIAM J. Numer. Anal., 48 (2010), pp. 2019–2042.

[9] E. Burman and B. Stamm, *Local discontinuous Galerkin method with reduced stabilization for diffusion equations*, Commun. Comput. Phys., 5 (2009), pp. 498–514.

[10] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[11] F. Cavalli, G. Naldi, G. Puppo, and M. Semplice, *High-order relaxation schemes for nonlinear degenerate diffusion problems*, SIAM J. Numer. Anal., 45 (2007), pp. 2098–2119.

[12] F. Cavalli, G. Naldi, G. Puppo, and M. Semplice, *A family of relaxation schemes for nonlinear convection diffusion problems*, Commun. Comput. Phys., 5 (2009), pp. 532–545.

[13] F. Cavalli, G. Naldi, G. Puppo, and M. Semplice, *Relaxed Schemes Based on Diffusive Relaxation for Hyperbolic-Parabolic Problems: Some New Developments*, in Numer. Methods Balance Laws 24, P. Gabriella and R. Giovanni, eds, Aracne editrice, Rome, 2010.

[14] C. Cercignani, *The Boltzmann Equation and Its Applications*, Appl. Math. Sci. 67, Springer-Verlag, New York, 1988.

[15] S. Chapman and T. G. Cowling, *The Mathematical Theory of Nonuniform Gases*, 3rd ed., Cambridge University Press, Cambridge, UK, 1970.

[16] B. Cockburn and B. Dong, *An analysis of the minimal dissipation local discontinuous Galerkin method for convection-diffusion problems*, J. Sci. Comput., 32 (2007), pp. 233–262.

[17] B. Cockburn and C.-W. Shu, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.

[18] F. Coron and B. Perthame, *Numerical passage from kinetic to fluid equations*, SIAM J. Numer. Anal., 28 (1991), pp. 26–42.

[19] S. M. Deshpande, P. S. Kulkarni, and A. K. Ghosh, *New developments in kinetic schemes*, Comput. Math. Appl., 35 (1998), pp. 75–93.

[20] S. Evje and K. H. Karlsen, *Monotone difference approximations of BV solutions to degenerate convection-diffusion equations*, SIAM J. Numer. Anal., 37 (2000), pp. 1838–1860.

[21] A. Friedman, *The Stefan problem in several space variable*, Trans. Amer. Math. Soc., 133 (1968), pp. 51–87.

[22] E. Gabetta, L. Pareschi, and G. Toscani, *Relaxation schemes for nonlinear kinetic equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2168–2194.

[23] L. Gosse and G. Toscani, *An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations*, C. R. Math. Acad. Sci. Paris, 334 (2002), pp. 337–342.

[24] S. Gottlieb and C.-W. Shu, *Total variation diminishing Runge-Kutta schemes*, Math. Comp., 67 (1998), pp. 73–85.

[25] J. L. Graveleau and P. Jamet, *A finite difference approach to some degenerate nonlinear parabolic equations*, SIAM J. Appl. Math., 20 (1971), pp. 199–223.

[26] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.
[27] J. W. Jerome and M. Rose, *Error estimates for the multidimensional two-phase stefan problem*, Math. Comp., 39 (1982), pp. 377–414.
[28] S. Jin, L. Pareschi, and G. Toscani, *Diffusive relaxation schemes for multiscale discrete velocity kinetic equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2405–2439.
[29] S. Jin and Z. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimension*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.
[30] C. Lattanzio and R. Natalini, *Convergence of diffusive BGK approximations for nonlinear strongly parabolic systems*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 341–358.
[31] R. J. LeVeque and M. Pelanti, *A class of approximate Riemann solvers and their relation to relaxation schemes*, J. Comput. Phys., 172 (2001), pp. 572–591.
[32] P. L. Lions and G. Toscani, *Diffusive limit for two-velocity Boltzmann kinetic models*, Rev. Mat. Iberoamericana, 13 (1997), pp. 473–513.
[33] E. Magenes, *Problemi di stefan bifase in piti variabili spaziali*, Matematiche, 36 (1981), pp. 65–108.
[34] E. Magenes, R. H. Nochetto, and C. Verdi, *Energy error estimates for a linear scheme to approximate nonlinear parabolic problems*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 655–678.
[35] D. Marazzina, *Stability properties of discontinuous Galerkin methods for 2D elliptic problems*, IMA J. Numer. Anal., 28 (2008), pp. 552–579.
[36] G. Naldi and L. Pareschi, *Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation*, SIAM J. Numer. Anal., 37 (2000), pp. 1246–1270.
[37] G. Naldi, L. Pareschi, and G. Toscani, *Relaxation schemes for partial differential equations and applications to degenerate diffusion problems*, Surveys Math. Indust., 10 (2002), pp. 315–343.
[38] R. Natalini, *Convergence to equilibrium for the relaxation approximations of conservation laws*, Comm. Pure Appl. Math., 49 (1996), pp. 795–823.
[39] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation*, J. Sci. Comput., 25 (2005), pp. 129–155.
[40] B. Perthame, *Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions*, SIAM J. Numer. Anal., 29 (1992), pp. 1–19.
[41] I. Perugia and D. Schötzau, *An hp-analysis of the local discontinuous Galerkin method for diffusion problems*, in Proceedings of the Fifth International Conference on Spectral and High Order Methods (ICOSAHOM-01) Uppsala, Sweden, vol. 17, 2002, pp. 561–571.
[42] I. Perugia and D. Schötzau, *The hp-local discontinuous Galerkin method for low-frequency time-harmonic Maxwell equations*, Math. Comp., 72 (2003), pp. 1179–1214.
[43] C. Schwab, *p- and hp-finite element methods*, in Numerical Mathematics and Scientific Computation, Theory Appl. Solid Fluid Mech., Oxford University Press, New York, 1998.
[44] M. Seaïd, *High-resolution relaxation scheme for the two-dimensional Riemann problems in gas dynamics*, Numer. Methods Partial Differential Equations, 22 (2006), pp. 397–413.
[45] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 1073–1084.