An Analysis of Three Different Formulations of the Discontinuous Galerkin Method for Diffusion Equations

Mengping Zhang[1] and Chi-Wang Shu[2]

Dedicated to Professor Jim Douglas, Jr. on the occasion of his 75th birthday

Abstract

In this paper we present an analysis of three different formulations of the discontinuous Galerkin method for diffusion equations. The first formulation yields an inconsistent and weakly unstable scheme, while the other two formulations, the local discontinuous Galerkin approach and the Baumann-Oden approach, give stable and convergent results. When written as finite difference schemes, such a distinction among the three formulations cannot be easily analyzed by the usual truncation errors, because of the phenomena of supraconvergence and weak instability. We perform a Fourier type analysis and compare the results with numerical experiments. The results of the Fourier type analysis agree well with the numerical results.

Key Words: discontinuous Galerkin method, diffusion equation, stability, consistency, convergence, supraconvergence

[1] Department of Mathematics, University of Science and Technology of China, Hefei, Anhui 230026, P.R. China. E-mail: mpzhang@ustc.edu.cn. The research of this author is supported by NNSFC grant 10028103.

[2] Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. E-mail: shu@cfm.brown.edu. The research of this author is supported by NNSFC grant 10028103 while he is in residence at the Department of Mathematics, University of Science and Technology of China, Hefei, Anhui 230026, P.R. China. Additional support is provided by ARO grant DAAD19-00-1-0405 and NSF grant DMS-9804985.

1 Introduction

The discontinuous Galerkin method is a class of finite element methods using completely discontinuous piecewise polynomial space for the numerical solution and the test functions. A key ingredient of this method is the design of suitable inter-element boundary treatments (the so-called numerical fluxes) to obtain highly accurate and stable schemes in many difficult situations.

Until recently, the discontinuous Galerkin method was mainly used to solve first order linear or nonlinear hyperbolic problems, such as the two dimensional hyperbolic conservation law

$$u_t + f(u)_x + g(u)_y = 0. \tag{1.1}$$

We mention for example the first discontinuous Galerkin method introduced in 1973 by Reed and Hill [16], in the framework of neutron transport, i.e. equation (1.1) without the time dependent term u_t and with linear fluxes $f(u) = au$ and $g(u) = bu$ where a and b do not depend on u, and the work of Cockburn et al. in a series of papers [9, 8, 7, 5, 10], in which they have established a framework to easily solve *nonlinear* time dependent hyperbolic conservation laws (1.1) using explicit, nonlinearly stable high order Runge-Kutta time discretizations [20] and discontinuous Galerkin discretization in space with exact or approximate Riemann solvers as interface numerical fluxes and TVB (total variation bounded) nonlinear limiters [18] to achieve non-oscillatory properties for strong shocks.

The discontinuous Galerkin method for (1.1) has found rapid applications in many diverse areas. This method has several attractive properties, such as its easiness for any order of accuracy in space and time including the p-version or spectral elements, its easiness in handling adaptivity strategies since refinement or unrefinement of the mesh can be achieved without taking into account of the continuity restrictions typical of conforming finite element methods and its easiness in changing the degrees of the approximating polynomials from one element to the other, its explicit nature thus its efficiency for solving the hyperbolic problem (1.1) without any global linear or nonlinear system solvers, its combination of the flexibility of finite element methods in the easy handling of complicated geometry with the high resolution property for discontinuous solutions of finite difference and finite volume methods through monotone numerical fluxes or approximate Riemann solvers applied at the element interfaces and limiters, its nice stability properties including a local cell entropy inequality for the square entropy [13] for general triangulation for any scalar nonlinear conservation laws (1.1) in any spatial dimensions and for any order of accuracy, and finally, its highly compact structure allowing efficient parallel implementation of the method allowing a parallel efficiency of over 80% even in a dynamic load balancing setting for time dependent adaptive mesh calculations [3].

For more details of the discontinuous Galerkin method and its recent development and applications, we refer the readers to the survey article by Cockburn, Karniadakis and Shu [6], the lecture notes by Cockburn [4], and the review paper by Cockburn and Shu [12].

Recently, motivated by the successful numerical experiments of Bassi and Rebay [1], Cockburn and Shu developed the so-called local discontinuous Galerkin method in treating the second order viscous terms and proved stability and convergence with error estimates [11]. At about the same time, Baumann and Oden [2] introduced a new discontinuous Galerkin method for the discretization of the second order viscous terms, see also the paper by Oden, Babuška and Baumann [15]. In [19] Shu presented three different formulations of the discontinuous Galerkin method, the two mentioned above plus a third one which looks very natural (and has been used in the engineering literature!) but which turns out to be inconsistent, and used simple examples to illustrate the basic ideas of these approaches, to compare their performances, and to emphasize the possible "pitfalls" for using the discontinuous Galerkin method on the viscous terms, see also [12]. However, the mechanism of the success or failure of these approaches was not discussed. In this paper we again use simple examples to further analyze these three formulations of the discontinuous Galerkin method.

It turns out that, when written as finite difference schemes, these three formulations give misleading conclusions when analyzed by the usual truncation errors. The phenomenon is partly related to the so-called “supraconvergence”, in that a finite difference method, when measured in truncation errors, may have lower order accuracy or even be inconsistent, but nevertheless converges with the expected order of accuracy, see for example [14]. Thus the traditional truncation error analysis plus stability cannot be used to predict accurately the rate of convergence. Also, the formulation which leads to an inconsistent and weakly unstable scheme cannot be easily analyzed by truncation errors, and the instability is so mild that it cannot be easily observed numerically. We perform a Fourier type analysis to predict convergence and compare the results with numerical experiments. The results of the Fourier type analysis agree well with the numerical results.

2 Three different formulations of the discontinuous Galerkin method

In this paper, for the simplicity of presentation we present all the discontinuous Galerkin methods for the diffusion equations on the simple one dimensional linear heat equation

$$u_t - u_{xx} = 0, \tag{2.1}$$

for $x \in [0,2\pi]$ with periodic boundary conditions and with an initial condition $u(x,0) = \sin(x)$. We would like to point out, however, that the methods are actually designed and can be analyzed for much more general multidimensional nonlinear convection diffusion equations, see, e.g. [11]. The points we would like to make in this paper can be represented well by the simple case (2.1).

Before discussing the discontinuous Galerkin method for (2.1), let us first describe it very briefly for the first order conservation law

$$u_t - u_x = 0, \tag{2.2}$$

again for $x \in [0,2\pi]$ with periodic boundary conditions and with an initial condition $u(x,0) = \sin(x)$. This will also set up the notations to be used later.

Let us denote $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, for $j = 1, ..., N$, as a mesh for $[0,2\pi]$, where $x_{\frac{1}{2}} = 0$ and $x_{N+\frac{1}{2}} = 2\pi$. We denote the center of each cell by $x_j = \frac{1}{2}\left(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}}\right)$ and the size of each cell by $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$. The cells do not need to be uniform for the method, but for simplicity of analysis we will consider only uniform meshes in this paper and will denote the uniform mesh size by Δx.

If we multiply (2.2) by an arbitrary test function $v(x)$, integrate over the interval I_j, and integrate by parts, we get

$$\int_{I_j} u_t v dx + \int_{I_j} u v_x dx - u(x_{j+\frac{1}{2}}, t)v(x_{j+\frac{1}{2}}) + u(x_{j-\frac{1}{2}}, t)v(x_{j-\frac{1}{2}}) = 0. \tag{2.3}$$

This is the starting point for designing the discontinuous Galerkin method. We replace both the solution u and the test function v by piecewise polynomials of degree at most k but do not change their notations for simplicity. That is, $u, v \in V_{\Delta x}$ where

$$V_{\Delta x} = \{v : \ v \text{ is a polynomial of degree at most } k \text{ for } x \in I_j, \ \ j = 1, ..., N\}. \tag{2.4}$$

With this choice, there is an ambiguity in (2.3) in the last two terms involving the boundary values at $x_{j\pm\frac{1}{2}}$, as both the solution u and the test function v are *discontinuous* exactly at these boundary points. This is exactly the place where discontinuous Galerkin method has its flexibility over continuous finite element methods: one could cleverly design these terms so that the resulting numerical method is stable and accurate. To motivate the ideas, let us look at the simplest case $k = 0$. That is, the solution as well as the test functions are piecewise constants. If we denote by u_j the value of u (which is constant in each cell) in the cell I_j, (2.3) would become the familiar first order upwind finite volume scheme

$$\frac{d}{dt}u_j - \frac{1}{\Delta x}\left(u_{j+1} - u_j\right) = 0$$

if we perform the following in (2.3):

1. Replace the boundary terms $u(x_{j\pm\frac{1}{2}}, t)$ by single valued numerical fluxes $\hat{u}_{j\pm\frac{1}{2}} = \hat{u}(u^-_{j\pm\frac{1}{2}}, u^+_{j\pm\frac{1}{2}})$. This is crucial for conservation. These fluxes in general depend both on the left limit (e.g. $u^-_{j+\frac{1}{2}} = \lim_{x\to x^-_{j+\frac{1}{2}}} u(x,t)$) and on the right limit (e.g. $u^+_{j+\frac{1}{2}} = \lim_{x\to x^+_{j+\frac{1}{2}}} u(x,t)$). For the equation (2.2), the flux $\hat{u}_{j+\frac{1}{2}}$ is taken as $u^+_{j+\frac{1}{2}}$ according to upwinding, since information flows from right to left in this case.

2. Replace the test function v at the boundaries by the values taken from inside the cell I_j, namely $v^-_{j+\frac{1}{2}}$ and $v^+_{j-\frac{1}{2}}$.

The scheme now becomes: find $u \in V_{\Delta x}$ such that, for all test functions $v \in V_{\Delta x}$,

$$\int_{I_j} u_t v dx + \int_{I_j} u v_x dx - \hat{u}_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} + \hat{u}_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} = 0 \tag{2.5}$$

where the numerical flux $\hat{u}_{j+\frac{1}{2}} = u^+_{j+\frac{1}{2}}$.

After picking a local basis and inverting a local $(k+1) \times (k+1)$ mass matrix (which could be done by hand), the scheme (2.5) can be written as

$$\frac{d}{dt} u_j + \frac{1}{\Delta x} \left(A u_j + B u_{j+1} \right) = 0 \tag{2.6}$$

where u_j is a small vector of length $k+1$ containing the coefficients of the solution u in the local basis inside cell I_j, and A and B are $(k+1) \times (k+1)$ constant matrices which can be computed once and for all and stored at the beginning of the code. Different choices of basis could make A and / or B sparse to save computational cost, especially for higher order versions (e.g. the p-version).

Scheme (2.6) can then be easily discretized in time by the nonlinearly stable high order Runge-Kutta methods in [20]. We remark that the method (2.6) is extremely simple to code and easy to parallelize. This simplicity carries over to multi-dimensional linear systems such as the Maxwell equation: the structure of the scheme is still similar to (2.6)!

We now turn our attention to the heat equation (2.1) as an example of the general convection diffusion problems containing second derivatives and present three different formulations of discontinuous Galerkin methods for this equation.

2.1 First formulation

If we proceed as before we obtain the following equality similar to (2.3):

$$\int_{I_j} u_t v dx + \int_{I_j} u_x v_x dx - u_x(x_{j+\frac{1}{2}}, t)v(x_{j+\frac{1}{2}}) + u_x(x_{j-\frac{1}{2}}, t)v(x_{j-\frac{1}{2}}) = 0. \qquad (2.7)$$

The only difference between (2.3) and (2.7) is that, in all the terms except the first one, u in (2.3) is replaced by u_x in (2.7). A very natural way to extend the scheme (2.5) would be simply replacing u by u_x: find $u \in V_{\Delta x}$ such that, for all test functions $v \in V_{\Delta x}$,

$$\int_{I_j} u_t v dx + \int_{I_j} u_x v_x dx - \hat{u_x}_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} + \hat{u_x}_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} = 0 \qquad (2.8)$$

where, for the lack of an upwinding mechanism for the heat equation one naturally takes a central flux $\hat{u_x}_{j+\frac{1}{2}} = \frac{1}{2}\left((u_x)^-_{j+\frac{1}{2}} + (u_x)^+_{j+\frac{1}{2}}\right)$.

This is the first formulation of the discontinuous Galerkin method for solving (2.1).

We remark that, in the actual computation, the scheme is similar to (2.6) and takes the form

$$\frac{d}{dt}u_j + \frac{1}{\Delta x^2}\left(Au_{j-1} + Bu_j + Cu_{j+1}\right) = 0 \qquad (2.9)$$

where u_j is a small vector of length $k+1$ containing the coefficients of the solution u in the local basis inside cell I_j, and A, B, C are $(k+1) \times (k+1)$ constant matrices which can be computed once and for all and stored at the beginning of the code. Again, the third order Runge-Kutta method [20] can be used. Implicit time stepping can also be used if the small time step restriction for stability is a concern, however in practice the discontinuous Galerkin method is more useful for convection dominated convection diffusion problems, such as the Navier-Stokes equations with a high Reynolds number, hence explicit time stepping is usually preferred.

It is verified numerically in [19], see also [12], that this formulation leads to numerically stable but inconsistent solutions. In Fig. 2.1 we plot the numerical solution with 40 and 320 cells versus the exact solution, for the two cases $k = 1$ and 2 (piecewise linear and piecewise quadratic cases) at $t = 0.7$. We can see that the numerical solutions seem to converge with mesh refinements but have O(1) errors when comparing with the exact solution.

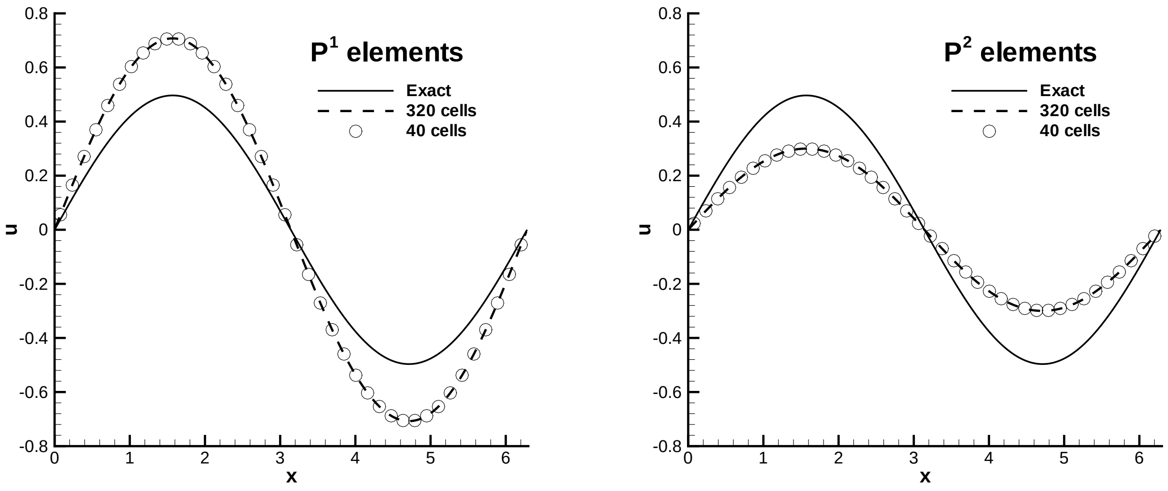

Figure 2.1: The inconsistent discontinuous Galerkin method (2.8) applied to the heat equation (2.1) with an initial condition $u(x, 0) = \sin(x)$. $t = 0.7$. Third order Runge-Kutta in time with small Δt so that time error can be ignored. Numerical solutions with 40 cells (circles) and 320 cells (dashed lines), versus the exact solution (solid line). Left: $k = 1$; Right: $k = 2$.

We remark that this is indeed a "pitfall" for the discontinuous Galerkin method applied to diffusion equations. It is very dangerous that the scheme (2.8) produces numerically stable but completely incorrect solution. If one does not know the exact solution, even if one does a mesh refinement study, one could still conclude incorrectly that the method is convergent. If the method is used to solve the complicated Navier-Stokes equations and produces beautiful color pictures, one would not be able to tell that the result is actually wrong (that is why this incorrect method was used in the engineering literature)!

2.2 Second formulation

If we rewrite the heat equation (2.1) as a first order system

$$u_t - q_x = 0, \qquad q - u_x = 0, \tag{2.10}$$

we can then *formally* use the same discontinuous Galerkin method for the convection equation to solve (2.10), resulting in the following scheme: find $u, q \in V_{\Delta x}$ such that, for all test functions $v, w \in V_{\Delta x}$,

$$\begin{aligned} \int_{I_j} u_t v dx + \int_{I_j} q v_x dx - \hat{q}_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} + \hat{q}_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} &= 0 \\ \int_{I_j} q w dx + \int_{I_j} u w_x dx - \hat{u}_{j+\frac{1}{2}} w^-_{j+\frac{1}{2}} + \hat{u}_{j-\frac{1}{2}} w^+_{j-\frac{1}{2}} &= 0, \end{aligned} \tag{2.11}$$

where, again for the lack of upwinding mechanism in a heat equation one could try the central fluxes (an arithmetic mean between the left and right values), but it turns out that a better choice for the fluxes, both in accuracy and in compactness of the eventual stencil, is

$$\hat{u}_{j+\frac{1}{2}} = u^+_{j+\frac{1}{2}}, \qquad \hat{q}_{j+\frac{1}{2}} = q^-_{j+\frac{1}{2}}, \tag{2.12}$$

i.e. we alternatively take the left and right limits for the fluxes in u and q (we could of course also take the pair $u^-_{j+\frac{1}{2}}$ and $q^+_{j+\frac{1}{2}}$ as the fluxes).

This is the second formulation of the discontinuous Galerkin method for solving (2.1). It was designed and analyzed by Cockburn and Shu [11], motivated by the numerical results of Bassi and Rebay [1] for the compressible Navier-Stokes equations.

We remark that the appearance of the auxiliary variable q is superficial: when a local basis is chosen in cell I_j then q is eliminated and the actual scheme for u, (2.11) with the fluxes (2.12), takes the identical simple form (2.9), of course with different matrices A, B and C.

For illustration purpose we show in Table 2.1 the L^2 and L^∞ errors and numerically observed orders of accuracy, for both u and q, for the two cases $k = 1$ and 2 (piecewise linear and piecewise quadratic cases) to $t = 1$. Clearly $(k + 1)$-th order of accuracy is achieved for both odd and even k and also the same order of accuracy is achieved for q which approximates u_x. We thus obtain the advantage of mixed finite element methods in approximating the derivatives of the exact solution to the same order of accuracy as the solution themselves, yet without additional storage or computational costs for the auxiliary variable q.

Table 2.1: L^2 and L^∞ errors and orders of accuracy for the local discontinuous Galerkin method (2.11) with fluxes (2.12) applied to the heat equation (2.1) with an initial condition $u(x,0) = \sin(x)$, $t = 1$. Third order Runge-Kutta in time with a small Δt so that time error can be ignored.

	$k=1$				$k=2$			
Δx	L^2 error	order	L^∞ error	order	L^2 error	order	L^∞ error	order
$2\pi/20$, u	1.58E-03	—	6.01E-03	—	3.98E-05	—	1.89E-04	—
$2\pi/20$, q	1.58E-03	—	6.01E-03	—	3.98E-05	—	1.88E-04	—
$2\pi/40$, u	3.93E-04	2.00	1.51E-03	1.99	4.98E-06	3.00	2.37E-05	2.99
$2\pi/40$, q	3.94E-04	2.00	1.51E-03	1.99	4.98E-06	3.00	2.37E-05	2.99
$2\pi/80$, u	9.83E-05	2.00	3.78E-04	2.00	6.22E-07	3.00	2.97E-06	3.00
$2\pi/80$, q	9.83E-05	2.00	3.78E-04	2.00	6.22E-07	3.00	2.97E-06	3.00
$2\pi/160$, u	2.46E-05	2.00	9.45E-05	2.00	7.78E-08	3.00	3.71E-07	3.00
$2\pi/160$, q	2.46E-05	2.00	9.45E-05	2.00	7.78E-08	3.00	3.71E-07	3.00

2.3 Third formulation

Another possible modification to the inconsistent scheme (2.8) is given by Baumann and Oden [2], see also Oden, Babuška, and Baumann [15]. Basically, extra penalty terms are added to the inter-element boundaries such that, when one takes $v = u$ and sums over all cells, the boundary contribution disappears and one gets a nice L^2 norm stability control. The scheme now becomes: find $u \in V_{\Delta x}$ such that, for all test functions $v \in V_{\Delta x}$,

$$\begin{aligned} \int_{I_j} u_t v dx + \int_{I_j} u_x v_x dx - \hat{u_x}_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} + \hat{u_x}_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} & \\ -\frac{1}{2}(v_x)^-_{j+\frac{1}{2}} \left(u^+_{j+\frac{1}{2}} - u^-_{j+\frac{1}{2}}\right) - \frac{1}{2}(v_x)^+_{j-\frac{1}{2}} \left(u^+_{j-\frac{1}{2}} - u^-_{j-\frac{1}{2}}\right) &= 0 \end{aligned} \tag{2.13}$$

where, again for the lack of upwinding mechanism in a heat equation one naturally takes a central flux $\hat{u_x}_{j+\frac{1}{2}} = \frac{1}{2}\left((u_x)^-_{j+\frac{1}{2}} + (u_x)^+_{j+\frac{1}{2}}\right)$. Notice that the extra terms added makes the system unsymmetric.

For coding purpose (2.13) is the most convenient form, however it might be more illustrative if we rewrite (2.13) into a global form: find $u \in V_{\Delta x}$ such that, for all test functions $v \in V_{\Delta x}$,

$$\int_0^{2\pi} u_t v dx + \sum_{j=1}^{N} \left(\int_{I_j} u_x v_x dx + \hat{u_x}_{j+\frac{1}{2}} [v]_{j+\frac{1}{2}} - \hat{v_x}_{j+\frac{1}{2}} [u]_{j+\frac{1}{2}} \right) = 0 \tag{2.14}$$

Table 2.2: L^2 and L^∞ errors and orders of accuracy for the Baumann-Oden discontinuous Galerkin method (2.13) applied to the heat equation (2.1) with an initial condition $u(x,0) = \sin(x)$, $t = 1$. Third order Runge-Kutta in time with a small Δt so that time error can be ignored.

	$k=1$				$k=2$			
Δx	L^2 error	order	L^∞ error	order	L^2 error	order	L^∞ error	order
$2\pi/20$	6.53E-03	—	1.21E-02	—	4.09E-03	—	5.77E-03	—
$2\pi/40$	1.63E-03	2.01	3.03E-03	2.00	1.06E-03	1.95	1.49E-03	1.95
$2\pi/80$	4.06E-04	2.00	7.57E-04	2.00	2.67E-04	1.99	3.77E-04	1.99
$2\pi/160$	1.02E-04	2.00	1.89E-04	2.00	6.68E-05	2.00	9.45E-05	2.00

where $[w] \equiv w^+ - w^-$ denotes the jump of the function w at the interface and the flux for v_x is also a central flux $\hat{v_x}_{j+\frac{1}{2}} = \frac{1}{2}\left((v_x)^-_{j+\frac{1}{2}} + (v_x)^+_{j+\frac{1}{2}}\right)$. The anti-symmetry nature of the boundary terms (which disappear when one takes $v = u$) is clearly seen in the global formulation (2.14).

We remark that once again we recover exactly the scheme in the form of (2.9) (of course with different constant matrices A, B and C) when a local basis is chosen. Hence the computational cost and storage requirement of the scheme (2.13) is the same as that of the inconsistent scheme (2.8) or as that of the local discontinuous Galerkin method (2.11)-(2.12). There is no saving in the computational cost here over the method (2.11)-(2.12) even though the latter has nominally an additional auxiliary variable q. This statement is valid when a linear PDE is solved. For nonlinear problems the computational cost of the Baumann-Oden method (2.13) may be smaller than that of the local discontinuous Galerkin method (2.11)-(2.12).

The order of accuracy for the scheme (2.13) is k for even k (sub-optimal) and $k+1$ for odd k (optimal).

For illustration purpose we show in Table 2.2 the L^2 and L^∞ errors and numerically observed orders of accuracy, for the two cases $k = 1$ and 2 (piecewise linear and piecewise quadratic cases) to $t = 1$. Clearly $(k+1)$-th order of accuracy is achieved for the odd $k = 1$ and k-th order of accuracy is achieved for the even $k = 2$.

3 An analysis for the three formulations

In this section we attempt to give an analysis to explain the dramatically different behaviors of the first and the other two discontinuous Galerkin formulations for the equation (2.1) given in the previous section.

In [11] and [17], standard finite element techniques are used to prove the convergence and error estimates for the second and third formulations (the proof for the third formulation is given only for the steady state version in [17]). However, it does not seem easy to use such techniques to prove the inconsistency and weak instability of the first formulation.

A natural thought is to rewrite the three formulations of the discontinuous Galerkin method as finite difference schemes, and then use finite difference techniques to analyze their stability, consistency and convergence. This should be particularly helpful to reveal the nature of inconsistency and/or instability for the first formulation. Towards this goal we choose the degrees of freedom for the k-th degree polynomial inside the cell I_j as the point values of the solution, denoted by

$$u_{j+\frac{2i-k}{2(k+1)}}, \qquad i = 0, ..., k,$$

at the k+1 equally spaced points

$$\left(j + \frac{2i-k}{2(k+1)}\right) \Delta x, \qquad i = 0, ..., k.$$

The schemes written in terms of these degrees of freedom become finite difference schemes on a globally uniform mesh (with a mesh size $\Delta x/(k+1)$), however they are not standard finite difference schemes because each point in the group of k+1 points belonging to the cell I_j obeys a different form of the finite difference scheme.

To be more specific, we concentrate on the piecewise linear $k = 1$ case. We have also carried out analysis for the piecewise quadratic $k = 2$ case obtaining similar results, but we will not present those results to save space. For the piecewise linear $k = 1$ case, we choose the degrees of freedom as the point values at the $2N$ uniformly spaced points

$$u_{j-\frac{1}{4}},\ u_{j+\frac{1}{4}}, \qquad j = 1, ..., N.$$

The solution inside the cell I_j is then represented by

$$u(x) = u_{j-\frac{1}{4}}\phi_{j-\frac{1}{4}}(x) + u_{j+\frac{1}{4}}\phi_{j+\frac{1}{4}}(x)$$

where $\phi_{j-\frac{1}{4}}(x)$ is the linear polynomial which equals 1 at the point $(j-\frac{1}{4})\Delta x$ and equals 0 at the point $(j+\frac{1}{4})\Delta x$, and similarly $\phi_{j+\frac{1}{4}}(x)$ is the linear polynomial which equals 0 at the point $(j-\frac{1}{4})\Delta x$ and equals 1 at the point $(j+\frac{1}{4})\Delta x$. With this representation, taking the test functions v also as $\phi_{j-\frac{1}{4}}(x)$ and $\phi_{j+\frac{1}{4}}(x)$, respectively, and inverting the small 2×2 mass matrix by hand, we obtain easily the three finite difference schemes corresponding to the three different formulations.

3.1 First formulation

For the first formulation (2.8) we obtain the scheme

$$\begin{aligned} u'_{j-\frac{1}{4}} &= \frac{1}{2\Delta x^2}\left(5u_{j-\frac{5}{4}} - 5u_{j-\frac{3}{4}} - 6u_{j-\frac{1}{4}} + 6u_{j+\frac{1}{4}} + u_{j+\frac{3}{4}} - u_{j+\frac{5}{4}}\right) \\ u'_{j+\frac{1}{4}} &= \frac{1}{2\Delta x^2}\left(-u_{j-\frac{5}{4}} + u_{j-\frac{3}{4}} + 6u_{j-\frac{1}{4}} - 6u_{j+\frac{1}{4}} - 5u_{j+\frac{3}{4}} + 5u_{j+\frac{5}{4}}\right) \end{aligned} \tag{3.1}$$

for $j = 1, ..., N$. Here u' denotes the time derivative of u. The scheme can be rewritten into a more compact form

$$\left(\begin{array}{c} u'_{j-\frac{1}{4}} \\ u'_{j+\frac{1}{4}} \end{array}\right) = \frac{1}{2\Delta x^2}\left[A\left(\begin{array}{c} u_{j-\frac{5}{4}} \\ u_{j-\frac{3}{4}} \end{array}\right) + B\left(\begin{array}{c} u_{j-\frac{1}{4}} \\ u_{j+\frac{1}{4}} \end{array}\right) + C\left(\begin{array}{c} u_{j+\frac{3}{4}} \\ u_{j+\frac{5}{4}} \end{array}\right)\right]. \tag{3.2}$$

with

$$A = \left(\begin{array}{cc} 5 & -5 \\ -1 & 1 \end{array}\right), \qquad B = \left(\begin{array}{cc} -6 & 6 \\ 6 & -6 \end{array}\right), \qquad C = \left(\begin{array}{cc} 1 & -1 \\ -5 & 5 \end{array}\right). \tag{3.3}$$

Notice that (3.1) or (3.2)-(3.3) is a finite difference scheme defined on a uniform mesh with mesh size $\Delta x/2$, however the even points and the odd points obey different forms of the scheme. Such finite difference schemes are non-standard. If we perform the usual truncation error analysis, namely substituting the exact solution u of the PDE (2.1) into the scheme (3.1) and performing Taylor expansions, we obtain the leading terms of the local truncation errors (LTE) as

$$\begin{aligned} LTE_{j-\frac{1}{4}} &= u_t(x_{j-\frac{1}{4}},t) - \frac{1}{2\Delta x^2}\Big(5u(x_{j-\frac{5}{4}},t) - 5u(x_{j-\frac{3}{4}},t) - 6u(x_{j-\frac{1}{4}},t) \\ &\quad +6u(x_{j+\frac{1}{4}},t) + u(x_{j+\frac{3}{4}},t) - u(x_{j+\frac{5}{4}},t)\Big) \\ &= O(\Delta x), \qquad (3.4) \\ LTE_{j+\frac{1}{4}} &= u_t(x_{j+\frac{1}{4}},t) - \frac{1}{2\Delta x^2}\Big(-u(x_{j-\frac{5}{4}},t) + u(x_{j-\frac{3}{4}},t) + 6u(x_{j-\frac{1}{4}},t) \\ &\quad -6u(x_{j+\frac{1}{4}},t) - 5u(x_{j+\frac{3}{4}},t) + 5u(x_{j+\frac{5}{4}},t)\Big) \\ &= O(\Delta x). \end{aligned}$$

This seems to indicate that the scheme is consistent and is (at least) first order accurate. But apparently the scheme is *not* consistent by the numerical experiments indicated in the previous section. We will address this apparent contradiction later.

Instead let us now perform the following standard Fourier analysis. This analysis depends heavily on the assumption of uniform mesh sizes and periodic boundary conditions. We make an ansatz of the form

$$\begin{pmatrix} u_{j-\frac{1}{4}}(t) \\ u_{j+\frac{1}{4}}(t) \end{pmatrix} = \begin{pmatrix} \hat{u}_{k,-\frac{1}{4}}(t) \\ \hat{u}_{k,\frac{1}{4}}(t) \end{pmatrix} e^{ikx_j} \qquad (3.5)$$

and substitute this into the scheme (3.2)-(3.3) to find the evolution equation for the coefficient vector as

$$\begin{pmatrix} \hat{u}'_{k,-\frac{1}{4}}(t) \\ \hat{u}'_{k,\frac{1}{4}}(t) \end{pmatrix} = G(k,\Delta x) \begin{pmatrix} \hat{u}_{k,-\frac{1}{4}}(t) \\ \hat{u}_{k,\frac{1}{4}}(t) \end{pmatrix} \qquad (3.6)$$

where the amplification matrix $G(k,\Delta x)$ is given by

$$G(k,\Delta x) = \frac{1}{2\Delta x^2}\left(A\,e^{-ik\Delta x} + B + C\,e^{ik\Delta x}\right). \qquad (3.7)$$

with the matrices A, B, C defined by (3.3). The two eigenvalues of the amplification matrix $G(k,\Delta x)$ are

$$\lambda_1 = -\frac{6}{\Delta x^2}\left(1-\cos(k\Delta x)\right), \qquad \lambda_2 = 0. \qquad (3.8)$$

We notice that both eigenvalues are non-positive. However, we will see later that there is still a very weak instability for this semi-discrete system. The general solution of the ODE

(3.6) is given by

$$\begin{pmatrix} \hat{u}_{k,-\frac{1}{4}}(t) \\ \hat{u}_{k,\frac{1}{4}}(t) \end{pmatrix} = a\, e^{\lambda_1 t}\, V_1 \, + b\, e^{\lambda_2 t}\, V_2, \tag{3.9}$$

where the eigenvalues λ_1 and λ_2 are given by (3.8), and V_1 and V_2 are the corresponding eigenvectors given by

$$V_1 = \begin{pmatrix} 3(1-\cos(\xi)) + 2\,i\,\sin(\xi) \\ -3(1-\cos(\xi)) + 2\,i\,\sin(\xi) \end{pmatrix}, \qquad V_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{3.10}$$

with $\xi = k\Delta x$. Our emphasis now is on consistency. Thus we look at the low modes, in particular for $k = 1$. To fit the given initial condition

$$u_{j\pm\frac{1}{4}}(0) = e^{ix_{j\pm\frac{1}{4}}}, \tag{3.11}$$

whose imaginary part is our initial condition for (2.1), we require, at $t = 0$,

$$\begin{pmatrix} \hat{u}_{1,-\frac{1}{4}}(0) \\ \hat{u}_{1,\frac{1}{4}}(0) \end{pmatrix} = \begin{pmatrix} e^{-i\,\frac{\Delta x}{4}} \\ e^{i\,\frac{\Delta x}{4}} \end{pmatrix},$$

hence we obtain the coefficients a and b in (3.9) as

$$a = -i\,\frac{\sin\left(\frac{\Delta x}{4}\right)}{3\,(1-\cos(\Delta x))}, \qquad b = \cos\left(\frac{\Delta x}{4}\right) - \frac{2\sin\left(\frac{\Delta x}{4}\right)\sin(\Delta x)}{3\,(1-\cos(\Delta x))}. \tag{3.12}$$

We remark that the usual way of taking initial conditions in a finite element method is via an L^2 projection, not by a point value collocation (3.11), however we have verified that this does not affect the final results in the analysis in this paper. We thus have the explicit solutions of the scheme (3.2)-(3.3) with the initial condition (3.11), for example

$$u_{j-\frac{1}{4}}(t) = ae^{ix_j+\lambda_1 t}(3(1-\cos(\Delta x)) + 2i\,\sin(\Delta x)) + be^{ix_j+\lambda_2 t} \tag{3.13}$$

with the eigenvalues λ_1, λ_2 given by (3.8) with $k = 1$ and the coefficients a, b given by (3.12). By a simple Taylor expansion, we obtain the imaginary part of $u_{j-\frac{1}{4}}(t)$ to be

$$Im\{u_{j-\frac{1}{4}}(t)\} = \frac{2+e^{-3t}}{3}\,\sin(x_{j-\frac{1}{4}}) + O(\Delta x).$$

This is about $0.7075\sin(x_{j-\frac{1}{4}})$ when $t = 0.7$, which matches very well with the numerical results in the previous section (see Fig. 2.1, left). We also clearly see that the scheme is *not* consistent, i.e. the numerical solution does not converge to the solution of the PDE (which equals to $\sin(x)\, e^{-t}$). Similar analysis can be done for the P^2 (piecewise quadratic) case, leading to the solution

$$Im\{u_j(t)\} = (1-t)\,\sin(x_j) + O(\Delta x^2),$$

which is about 0.3 $\sin(x_j)$ when $t = 0.7$, again matching very well with the numerical results in the previous section (see Fig. 2.1, right) and is inconsistent with the PDE.

The apparent contradiction with the traditional truncation error analysis can be explained by a very weak instability of this scheme. For this stability analysis we look at (3.6) for the high modes (large k). We still denote $\xi = k\Delta x$. When $\cos(\xi) = 1$, we clearly have the amplification matrix $G(k, \Delta x) = 0$, hence the solution to (3.6) remains to be the initial condition. When $\cos(\xi) \neq 1$, the amplification matrix $G(k, \Delta x)$ is diagonalizable. With eigenvalues of $G(k, \Delta x)$ given by (3.8), and the matrix consisting of the eigenvectors (3.10) of $G(k, \Delta x)$ as columns given by

$$R = \begin{pmatrix} 3(1-\cos(\xi)) + 2i\,\sin(\xi) & 1 \\ -3(1-\cos(\xi)) + 2i\,\sin(\xi) & 1 \end{pmatrix} \tag{3.14}$$

which has an inverse when $\cos(\xi) \neq 1$

$$R^{-1} = \frac{1}{6(1-\cos(\xi))} \begin{pmatrix} 1 & -1 \\ 3(1-\cos(\xi)) - 2i\,\sin(\xi) & 3(1-\cos(\xi)) + 2i\,\sin(\xi) \end{pmatrix} \tag{3.15}$$

hence we obtain explicitly the solution to (3.6) as

$$\begin{pmatrix} \hat{u}_{k,-\frac{1}{4}}(t) \\ \hat{u}_{k,\frac{1}{4}}(t) \end{pmatrix} = e^{G(k,\Delta x)t} \begin{pmatrix} \hat{u}_{k,-\frac{1}{4}}(0) \\ \hat{u}_{k,\frac{1}{4}}(0) \end{pmatrix}$$

with

$$e^{G(k,\Delta x)t} = R \begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & 1 \end{pmatrix} R^{-1}.$$

It is now possible, using the explicit formulas (3.8), (3.14) and (3.15), to explicitly write out $e^{G(k,\Delta x)t}$, and compute its L^2 norm, namely, the square root of the spectral radius of the symmetric matrix $\left(e^{G(k,\Delta x)t}\right)^* \left(e^{G(k,\Delta x)t}\right)$. This L^2 norm is given by

$$||e^{G(k,\Delta x)t}|| =$$
$$\sqrt{\frac{1}{18}\left(5+8\alpha+5\alpha^2+\frac{1-\alpha}{\beta}\left[8(1-\alpha)+\sqrt{(8+5\beta)\left[8(1-\alpha)^2+\beta(5+26\alpha+5\alpha^2)\right]}\right]\right)},$$

where

$$\alpha = e^{\lambda_1 t}, \qquad \beta = 1-\cos(\xi).$$

For the stability of (3.6) we would need $||e^{G(k,\Delta x)t}||$ to be uniformly bounded with respect to the two parameters k and Δx. However, if we take $\beta = \frac{\Delta x^2}{t}$, then it is easy to see that

$$||e^{G(k,\Delta x)t}|| = O\left(\frac{1}{\Delta x}\right)$$

which is unbounded when $\Delta x \to 0$. Hence the semi-discrete system (3.6) is not stable.

This instability is however very mild, and it grows at most linearly with a mesh refinement. Also, further analysis, by looking at the eigenvectors, shows that this instability only occurs when the initial condition is chosen so that the slope of the linear function in each cell is of order $O\left(\frac{1}{\Delta x}\right)$. Since such initial conditions are not physical, they can only occur at the round-off level and they grow slower than linearly with the number of time steps. This explains why we have never seen such instability in the numerical experiments: our meshes are simply not refined enough. However, this instability accounts for the apparent contradiction of a consistent local truncation error (3.4) and a global O(1) error of the numerical solution.

We remark that standard finite element type energy estimate can also partially reveal the weak instability of this scheme[3]. However, the Fourier type analysis given here pinpoints more accurately the source and growth of this instability.

3.2 Second formulation

We now use the same method to analyze the second formulation, namely the local discontinuous Galerkin method of Cockburn and Shu [11] given by (2.11)-(2.12). Instead of (3.2)-(3.3),

[3] B. Cockburn, private communications

we now obtain the scheme (3.2) with

$$A = \begin{pmatrix} 0 & 20 \\ 0 & -4 \end{pmatrix}, \qquad B = \begin{pmatrix} -39 & 17 \\ 15 & -25 \end{pmatrix}, \qquad C = \begin{pmatrix} 3 & -1 \\ 21 & -7 \end{pmatrix}. \tag{3.16}$$

We can repeat the truncation error analysis and obtain, instead of (3.4),

$$\begin{aligned} LTE_{j-\frac{1}{4}} &= u_t(x_{j-\frac{1}{4}},t) - \frac{1}{2\Delta x^2}\Big(20u(x_{j-\frac{3}{4}},t) - 39u(x_{j-\frac{1}{4}},t) + 17u(x_{j+\frac{1}{4}},t) \\ &\quad +3u(x_{j+\frac{3}{4}},t) - u(x_{j+\frac{5}{4}},t)\Big) \\ &= \frac{3}{2}u_{xx}(x_{j-\frac{1}{4}},t) + O(\Delta x), \qquad\qquad (3.17) \\ LTE_{j+\frac{1}{4}} &= u_t(x_{j+\frac{1}{4}},t) - \frac{1}{2\Delta x^2}\Big(-4u(x_{j-\frac{3}{4}},t) + 15u(x_{j-\frac{1}{4}},t) - 25u(x_{j+\frac{1}{4}},t) \\ &\quad +21u(x_{j+\frac{3}{4}},t) - 7u(x_{j+\frac{5}{4}},t)\Big) \\ &= -\frac{3}{2}u_{xx}(x_{j+\frac{1}{4}},t) + O(\Delta x). \end{aligned}$$

Now it looks like the scheme is inconsistent as the local truncation errors are O(1)! This is related to the phenomenon called *supraconvergence*, namely the local truncation error predicts a convergence rate lower than the actual convergence rate, or the local truncation error could even be O(1) or blowing up for a convergent scheme, see, e.g. [14]. In [11], standard finite element techniques are used to prove the stability and convergence rate of this method. Here we follow the Fourier type analysis in the previous subsection to give an alternative proof.

We make the same ansatz as in (3.5) and substitute it into the scheme (3.2)-(3.16) to obtain the evolution equation for the coefficient vector (3.6) with the amplification matrix (3.7), where the matrices A, B, C are defined by (3.16). The two eigenvalues of the amplification matrix $G(k, \Delta x)$ are

$$\lambda_{1,2} = -\frac{2}{\Delta x^2}\left(8 + \cos(\xi) \pm \sqrt{(8+\cos(\xi))^2 - 18(1-\cos(\xi))}\right), \tag{3.18}$$

where as before $\xi = k\Delta x$. Clearly both eigenvalues are real and non-positive. The general solution of the ODE (3.6) is again given by (3.9) where the eigenvalues λ_1 and λ_2 are given by (3.18), and V_1 and V_2 are the corresponding eigenvectors given by

$$V_1 = \begin{pmatrix} \alpha - 4\beta \\ 3\gamma \end{pmatrix}, \qquad V_2 = \begin{pmatrix} \alpha + 4\beta \\ 3\gamma \end{pmatrix}, \tag{3.19}$$

where

$$\begin{aligned}\alpha &= -7(1-\cos(\xi)) + 3\,i\,\sin(\xi), \qquad \beta = \sqrt{(8+\cos(\xi))^2 - 18(1-\cos(\xi))},\\ \gamma &= 5 + 7\cos(\xi) + 7\,i\,\sin(\xi). \end{aligned} \tag{3.20}$$

To study consistency, we look at the low mode case $k = 1$. To fit the initial condition (3.11), we obtain the coefficients a and b in (3.9) as

$$\begin{aligned} a &= \frac{e^{i\frac{\Delta x}{4}}\left(\alpha + 4\beta - 36\cos\left(\frac{\Delta x}{2}\right) - 6\,i\,\sin\left(\frac{\Delta x}{2}\right)\right)}{24\beta\gamma}, \qquad (3.21)\\ b &= \frac{e^{i\frac{\Delta x}{4}}\left(-\alpha + 4\beta + 36\cos\left(\frac{\Delta x}{2}\right) + 6\,i\,\sin\left(\frac{\Delta x}{2}\right)\right)}{24\beta\gamma}, \end{aligned}$$

where α, β and γ are given by (3.20) with $\xi = \Delta x$. We thus again have the explicit solutions of the scheme (3.2)-(3.16) with the initial condition (3.11), for example

$$u_{j-\frac{1}{4}}(t) = a\,e^{ix_j + \lambda_1 t}\,(\alpha - 4\beta) + b\,e^{ix_j+\lambda_2 t}\,(\alpha + 4\beta), \tag{3.22}$$

where α, β and γ are given by (3.20) and the eigenvalues λ_1, λ_2 are given by (3.18) with $\xi = \Delta x$, and the coefficients a, b are given by (3.21). By a simple Taylor expansion, we obtain the imaginary part of $u_{j-\frac{1}{4}}(t)$ to be

$$Im\{u_{j-\frac{1}{4}}(t)\} = \sin(x_{j-\frac{1}{4}})\,e^{-t} + O(\Delta x^2).$$

This is clearly consistent with the exact solution to second order accuracy.

It is also easy to establish stability of the semi-discrete scheme (3.6) in this case. The matrix consisting of the eigenvectors (3.19) of $G(k, \Delta x)$ as columns is given by

$$R = \begin{pmatrix} \alpha - 4\beta & \alpha + 4\beta \\ 3\gamma & 3\gamma \end{pmatrix} \tag{3.23}$$

with its inverse given by

$$R^{-1} = \frac{1}{24\beta\gamma}\begin{pmatrix} -3\gamma & \alpha + 4\beta \\ 3\gamma & -\alpha + 4\beta \end{pmatrix} \tag{3.24}$$

where α, β and γ are again given by (3.20). We can now explicitly compute the L^2 norms of R and R^{-1}, namely the square roots of the spectral radii of the symmetric matrices R^*R

and $(R^{-1})^* (R^{-1})$:

$$||R|| = 2\sqrt{365 + 269\cos(\xi) + 14\cos^2(\xi) + \sqrt{(1-\cos(\xi))(10681 + 555\cos(\xi) - 5404\cos^2(\xi))}}$$

and

$$||R^{-1}|| = \frac{1}{12}\sqrt{\frac{365 + 269\cos(\xi) + 14\cos^2(\xi) + \sqrt{(1-\cos(\xi))(10681 + 555\cos(\xi) - 5404\cos^2(\xi))}}{2(1702 + 2868\cos(\xi) + 1227\cos^2(\xi) + 35\cos^3(\xi))}}.$$

It is easy to see that both $||R||$ and $||R^{-1}||$ are uniformly bounded with respect to the parameter ξ. Thus the stability of the semi-discrete scheme (3.6) in this case is established.

3.3 Third formulation

Finally we turn to the analysis of the third formulation, namely the Baumann and Oden method [2] given by (2.13). We have therefore the scheme (3.2) with

$$A = \frac{1}{2}\begin{pmatrix} 7 & -1 \\ 1 & -7 \end{pmatrix}, \qquad B = \begin{pmatrix} -12 & 12 \\ 12 & -12 \end{pmatrix}, \qquad C = \frac{1}{2}\begin{pmatrix} -7 & 1 \\ -1 & 7 \end{pmatrix}. \tag{3.25}$$

We can repeat a truncation error analysis and obtain, instead of (3.4) or (3.17),

$$\begin{aligned}
LTE_{j-\frac{1}{4}} &= u_t(x_{j-\frac{1}{4}},t) - \frac{1}{4\Delta x^2}\Big(7u(x_{j-\frac{5}{4}},t) - u(x_{j-\frac{3}{4}},t) - 24u(x_{j-\frac{1}{4}},t) \\
&\qquad +24u(x_{j+\frac{1}{4}},t) - 7u(x_{j+\frac{3}{4}},t) + u(x_{j+\frac{5}{4}},t)\Big) \\
&= O(\Delta x), \qquad\qquad (3.26)\\
LTE_{j+\frac{1}{4}} &= u_t(x_{j+\frac{1}{4}},t) - \frac{1}{4\Delta x^2}\Big(u(x_{j-\frac{5}{4}},t) - 7u(x_{j-\frac{3}{4}},t) + 24u(x_{j-\frac{1}{4}},t) \\
&\qquad -24u(x_{j+\frac{1}{4}},t) - u(x_{j+\frac{3}{4}},t) + 7u(x_{j+\frac{5}{4}},t)\Big) \\
&= O(\Delta x).
\end{aligned}$$

For this case the truncation error analysis indicates that the scheme is consistent, but fails to indicate the correct second order convergence rate. This is again related to the phenomenon of supraconvergence. In [17], standard finite element techniques are used to prove the stability and convergence rate of this method in the steady state case. Here we follow the Fourier type analysis in the previous subsections to give an alternative proof.

We make the same ansatz as in (3.5) and substitute it into the scheme (3.2)-(3.25) to obtain the evolution equation for the coefficient vector (3.6) with the amplification matrix (3.7), where the matrices A, B, C are defined by (3.25). The two eigenvalues of the amplification matrix $G(k, \Delta x)$ are

$$\lambda_{1,2} = \frac{6}{\Delta x^2}\left(-1 \pm \sqrt{1 - \frac{1}{3}\sin^2(\xi)}\right), \tag{3.27}$$

where as before $\xi = k\Delta x$. Clearly both eigenvalues are real and non-positive. The general solution of the ODE (3.6) is again given by (3.9) where the eigenvalues λ_1 and λ_2 are given by (3.27), and V_1 and V_2 are the corresponding eigenvectors given by

$$V_1 = \begin{pmatrix} -12\alpha + 7\,i\,\sin(\xi) \\ \beta \end{pmatrix}, \qquad V_2 = \begin{pmatrix} 12\alpha + 7\,i\,\sin(\xi) \\ \beta \end{pmatrix}, \tag{3.28}$$

where

$$\alpha = \sqrt{1 - \frac{1}{3}\sin^2(\xi)}, \qquad \beta = -12 + \,i\,\sin(\xi). \tag{3.29}$$

Consistency can again be studied by looking at the low mode case $k = 1$. To fit the initial condition (3.11), we obtain the coefficients a and b in (3.9) as

$$\begin{aligned} a &= \frac{6e^{-i\frac{\Delta x}{4}} + 6\alpha e^{i\frac{\Delta x}{4}} - \sin(\Delta x)\left(4\sin\left(\frac{\Delta x}{4}\right) - 3\,i\,\cos\left(\frac{\Delta x}{4}\right)\right)}{12\alpha\beta}, \\ b &= \frac{-6e^{-i\frac{\Delta x}{4}} + 6\alpha e^{i\frac{\Delta x}{4}} + \sin(\Delta x)\left(4\sin\left(\frac{\Delta x}{4}\right) - 3\,i\,\cos\left(\frac{\Delta x}{4}\right)\right)}{12\alpha\beta}, \end{aligned} \tag{3.30}$$

where α and β are given by (3.29) with $\xi = \Delta x$. We thus obtain the explicit solutions of the scheme (3.2)-(3.25) with the initial condition (3.11), for example

$$u_{j-\frac{1}{4}}(t) = a\,e^{ix_j+\lambda_1 t}\left(-12\alpha + 7\,i\,\sin(\Delta x)\right) + b\,e^{ix_j+\lambda_2 t}\left(12\alpha + 7\,i\,\sin(\Delta x)\right) \tag{3.31}$$

where α is given by (3.29) and the eigenvalues λ_1, λ_2 are given by (3.27) with $\xi = \Delta x$, and the coefficients a, b are given by (3.30). By a simple Taylor expansion, we obtain the imaginary part of $u_{j-\frac{1}{4}}(t)$ to be

$$Im\{u_{j-\frac{1}{4}}(t)\} = \sin(x_{j-\frac{1}{4}})\,e^{-t} + O(\Delta x^2),$$

thus establishing second order accuracy.

It is also easy to establish stability of the semi-discrete scheme (3.6) in this case. The matrix consisting of the eigenvectors (3.28) of $G(k, \Delta x)$ as columns is given by

$$R = \begin{pmatrix} -12 + 7\,i\,\sin(\xi) & 12\alpha + 7\,i\,\sin(\xi) \\ \beta & \beta \end{pmatrix} \tag{3.32}$$

with its inverse given by

$$R^{-1} = \frac{1}{24\alpha\beta} \begin{pmatrix} -\beta & 12\alpha + 7\,i\,\sin(\xi) \\ \beta & 12\alpha - 7\,i\,\sin(\xi) \end{pmatrix}, \tag{3.33}$$

where α and β are again given by (3.29). We can now explicitly compute the L^2 norms of R and R^{-1},

$$||R|| = \sqrt{288 + 2\sin^2(\xi) + 14\sqrt{\sin^2(\xi)(144 + \sin^2(\xi))}}$$

and

$$||R^{-1}|| = \frac{1}{8}\sqrt{\frac{144 + \sin^2(\xi) + 7\sqrt{\sin^2(\xi)(144 + \sin^2(\xi))}}{3(290 + 143\cos^2(\xi) - \cos^4(\xi))}}.$$

It is easy to see that both $||R||$ and $||R^{-1}||$ are uniformly bounded with respect to the parameter ξ, thus establishing the stability of the semi-discrete scheme (3.6) in this case.

4 Concluding remarks

We have presented three different formulations of the discontinuous Galerkin method for solving the diffusion equations. Using the one dimensional heat equation as an example, we have written out these schemes in the finite difference format by choosing the point values at equally spaced points as the degrees of freedom, and performed stability and consistency analysis for these different formulations. The results of the analysis match well with numerical experiments. The first conclusion of this paper is that, when using finite difference techniques to analyze discontinuous Galerkin methods, one must pay special attention to the phenomenon of supraconvergence, namely the truncation errors might be too large and do not faithfully represent the accuracy of the scheme. The second conclusion is that one must be very careful in designing discontinuous Galerkin methods for PDEs involving higher derivatives, as one might obtain inconsistent and weakly unstable approximations which numerically might look like converging to a function that is however not the exact solution of the PDE.

Acknowledgment: The authors would like to thank Bernardo Cockburn for many helpful discussions regarding the problems in this paper.

References

[1] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[2] C. E. Baumann and J. T. Oden, *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 311–341.

[3] R. Biswas, K. D. Devine and J. Flaherty, *Parallel, adaptive finite element methods for conservation laws*, Appl. Numer. Math., 14 (1994), pp. 255–283.

[4] B. Cockburn, *Discontinuous Galerkin methods for convection-dominated problems*, in *High-Order Methods for Computational Physics*, T.J. Barth and H. Deconinck, editors, Lecture Notes in Computational Science and Engineering, volume 9, Springer, 1999, pp. 69–224.

[5] B. Cockburn, S. Hou, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[6] B. Cockburn, G. Karniadakis and C.-W. Shu, *The development of discontinuous Galerkin methods*, in *Discontinuous Galerkin Methods: Theory, Computation and Applications*, B. Cockburn, G. Karniadakis and C.-W. Shu, editors, Lecture Notes in

Computational Science and Engineering, volume 11, Springer, 2000, Part I: Overview, pp. 3-50.

[7] B. Cockburn, S.-Y. Lin and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.

[8] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II: general framework*, Math. Comp., 52 (1989), pp. 411–435.

[9] B. Cockburn and C.-W. Shu, *The Runge-Kutta local projection P^1-discontinuous-Galerkin finite element method for scalar conservation laws*, Math. Model. Numer. Anal. (M^2AN), v25 (1991), pp. 337-361.

[10] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws V: multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199-224.

[11] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[12] B. Cockburn and C.-W. Shu, *Runge-Kutta Discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., v16 (2001), pp. 173-261.

[13] G. Jiang and C.-W. Shu, *On cell entropy inequality for discontinuous Galerkin methods*, Math. Comp., 62 (1994), pp. 531–538.

[14] H.-O. Kreiss, T.A. Manteuffel, B. Swartz, B. Wendroff and A.B. White, Jr., *Supra-convergent schemes on irregular grids*, Math. Comp., 47 (1986), pp. 537–554.

[15] J. T. Oden, Ivo Babuška, and C. E. Baumann, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–519.

[16] W. H. Reed and T. R. Hill, *Triangular mesh methods for the neutron transport equation*, Tech. Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.

[17] B. Riviere, M. Wheeler and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902-931.

[18] C.-W. Shu, *TVB uniformly high-order schemes for conservation laws*, Math. Comp., 49 (1987), pp. 105-121.

[19] C.-W. Shu, *Different formulations of the discontinuous Galerkin method for the viscous terms*, in *Advances in Scientific Computing*, Z.-C. Shi, M. Mu, W. Xue and J. Zou, editors, Science Press, 2001, pp.144-155.

[20] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.