How to improve the decay of the numerical error for large times: the case of dissipative BGK systems

Maya Briani, Denise Aregba-Driollet, Roberto Natalini

Manuscript submitted to
AIMS' Journals
Volume **X**, Number **0X**, XX **200X**

Website: http://AIMsciences.org

pp. **X–XX**

HOW TO IMPROVE THE DECAY OF THE NUMERICAL ERROR FOR LARGE TIMES: THE CASE OF DISSIPATIVE BGK SYSTEMS

Maya Briani

Istituto per le Applicazioni del Calcolo Mauro Picone, Consiglio Nazionale delle Ricerche, Italy.

Denise Aregba-Driollet and Roberto Natalini

IMB, UMR CNRS 5251, Université de Bordeaux, France.
Istituto per le Applicazioni del Calcolo Mauro Picone, Consiglio Nazionale delle Ricerche, Italy.

Abstract. We introduce new finite differences schemes to approximate one dimensional dissipative semilinear hyperbolic systems with a BGK structure. Using accurate analytical time-decay properties of the local truncation error, it is possible to design schemes based on standard upwinding schemes, which are increasingly accurate for large times when computing small perturbations of constants asymptotic states.

1. **Introduction.** Consider the following class of one dimensional BGK systems:

$$\partial_t f^i + \lambda_i \partial_x f^i = M_i(u) - f^i, \quad i = 1, ..., m. \tag{1}$$

Here $f^i \in \mathbb{R}^k$, $u := \sum_{i=1}^m f^i$, $x \in \mathbb{R}$ and $t > 0$, and the functions $M_i = M_i(u) \in \mathbb{R}^k$ are smooth functions of u such that: $\sum_{i=1}^m M_i(u) = u$.

To obtain the time decay rates of these solutions, we need to rewrite the problem in more suitable coordinates. Following [3], we rewrite the BGK system in its *conservative-dissipative* form for the new unknowns

$$Z = (u, \tilde{Z})^T.$$

It is proved in [3] that, under some dissipativity conditions and for initial data which are small and smooth in some suitable norms, the time decay of the global solutions, for large times and in the L^∞-norm, is given by

$$\partial_x^l u = O(t^{-1/2-l/2}), \quad \partial_x^l \tilde{Z} = O(t^{-1-l/2}),$$

and similar estimates are available for their time derivatives. Notice that the improved estimate for $\tilde{Z}$ can only be obtained in these new coordinates.

The aim of this paper is to give a brief overview of the way it is possible to take advantage of these precise decay estimates to build up more accurate numerical schemes. Actually, we can see that, for standard numerical schemes, like for instance

2000 *Mathematics Subject Classification.* Primary: 65M12; Secondary: 35L65.

Key words and phrases. Finite differences methods, dissipative hyperbolic problems, BGK systems, asymptotic behavior, asymptotic high order schemes.

the upwind scheme with the source term approximated pointwise by the standard Euler scheme, the truncation error has the following decay as $t \to +\infty$:

$$\begin{array}{l} \mathcal{T}_u(x,t) = O(\Delta x\ t^{-3/2}) + O(\Delta t\ t^{-3/2}), \\ \mathcal{T}_{\tilde{Z}}(x,t) = O(\Delta x\ t^{-3/2}) + O(\Delta t\ t^{-3/2}). \end{array} \tag{2}$$

It can be seen numerically that the corresponding absolute error, for a fixed space step, decays as

$$e_u(t) = O(t^{-1/2}), e_z(t) = O(t^{-1}),$$

which implies that the relative error is essentially constant in time.

Here, our main goal is to improve the decay estimates on the truncation order to achieve an effective decay in time of the relative error, both in u and $\tilde{Z}$. To obtain this result, we perform a detailed analysis of the behavior of the truncation error for a general class of schemes, called "Time Asymptotically High Order" (TAHO) schemes, which generalize those introduced in [2]. Thanks to this analysis, we are able to select some schemes such that the truncation order behaves as

$$\mathcal{T}_u(x,t) = O(\Delta x\ t^{-2}),\ \mathcal{T}_{\tilde{Z}}(x,t)) = O(\Delta x\ t^{-2}), \tag{3}$$

for a fixed CFL ratio and such that the numerical error observed in the practical tests improves of $t^{-1/2}$ on other schemes.

The plan of the paper is the following. In Section 2, we introduce our analytical framework. The main schemes are derived in Section 3, where we show how to improve the time decay of their local truncation error. Section 4 presents some numerical tests which show the nice behavior of our new schemes in two test cases.

2. The analytical framework.

Following [3], we rewrite system 1 in its *conservative-dissipative* form. This means that we assume that there exists an invertible matrix

$$D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}, \tag{4}$$

such that, setting $m_1 = k$, $m_2 = k(m-1)$, the new unknown

$$Z = Df = (u, \tilde{Z})^T \in \mathbb{R}^k \times \mathbb{R}^{m_2}, \tag{5}$$

solves the system

$$\begin{cases} \partial_t u + A_{11}\partial_x u + A_{12}\partial_x \tilde{Z} &= 0, \\ \partial_t \tilde{Z} + A_{21}\partial_x u + A_{22}\partial_x \tilde{Z} &= \tilde{Q}(u) - \tilde{Z}, \end{cases} \tag{6}$$

where A is symmetric and $\tilde{Q}(u)$ is quadratic in u, i.e.: $\tilde{Q}(0) = 0$ and $\tilde{Q}'(0) = 0$. Observe that, after this transformation, which a priori is not unique, the source term is zero in the first component and the second one is the sum of a quadratic term and of the dissipative term $-\tilde{Z}$.
Moreover, when transforming system 1 in system 7, we can always assume that blocks D_{11} and D_{12} have the special form

$$D_{11} = I_k, \quad D_{12} = (I_k I_k \cdots I_k) \in \mathbb{R}^{k \times m_2},$$

and, setting $\Lambda = diag(\lambda_1 I_k, ..., \lambda_m I_k)$, we have that

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = D\Lambda D^{-1}.$$

Therefore, we can rewrite our system in a more compact form:

$$\partial_t Z + A\partial_x Z = -Z + DM(u). \tag{7}$$

To guarantee the existence of the matrix D in 5, we can assume that our system is strictly entropy dissipative in the sense of [5] and verifies the Shizuta-Kawashima condition [6, 5, 3].
For instance, using Bouchut's Entropy dissipation condition [4], it is possible to prove the existence of a matrix D in 5, with all the above properties.

2.1. **The simplest example: the Jin-Xin 2×2 relaxation system.** Consider the following system

$$\begin{cases} \partial_t u + \partial_x v = 0, \\ \partial_t v + \lambda^2 \partial_x u = F(u) - v, \end{cases} \tag{8}$$

with $\lambda > 0$. The unknowns u and v are scalar and the function $F = F(u)$ is smooth, with $F(0) = 0$. This case is obtained from 1 for $k = 1$, $m = 2$, and $\lambda_2 = -\lambda_1 = \lambda$, by setting

$$u = f^1 + f^2,\ v = \lambda(f^2 - f^1),\ F(u) = \lambda(M_2 - M_1).$$

Under the condition

$$\lambda > |F'(0)|, \tag{9}$$

the problem is dissipative, at least in a small neighborhood of the origin, in the sense of [5] and the Shizuta-Kawashima condition is verified.
In this case the conservative-dissipative form is obtained by using

$$D = \begin{pmatrix} 1 & 1 \\ -\mu a_+ & \mu a_- \end{pmatrix},$$

where $a = F'(0)$, $\mu = (\lambda^2 - a^2)^{-1/2}$ is real and positive, $a_\pm = \lambda \pm a > 0$, from assumption 10.

2.2. **A 3×3 BGK example.** Let us now compute the conservative-dissipative form for the following 3×3 BGK model,

$$\begin{cases} \partial_t f_1 - \lambda \partial_x f_1 = M_1(u) - f_1, \\ \partial_t f_2 = M_2(u) - f_2, \\ \partial_t f_3 + \lambda \partial_x f_3 = M_3(u) - f_3. \end{cases}$$

Let $F = F(u)$ be a smooth scalar function such that $F(0) = 0$ and let γ be such that $\gamma'(u) = |F'(u)|$, with $\gamma(0) = 0$. We choose our three maxwellian functions as follows, for $\beta \in]0, 1[$ and $\lambda > 0$

$$M_1(u) = \frac{1}{2}\left(\frac{\gamma(u) - F(u)}{\lambda} + \beta u\right), M_3(u) = \frac{1}{2}\left(\frac{\gamma(u) + F(u)}{\lambda} + \beta u\right),$$

$$M_2(u) = u - M_1(u) - M_3(u) = (1-\beta)u - \frac{\gamma(u)}{\lambda}.$$

The functions M_i, $i = 1, 2, 3$, are strictly increasing if for any u under consideration

$$\lambda > \frac{|F'(u)|}{1-\beta},$$

and so the entropy dissipation condition [4] is verified. Let $a = F'(0)$ and $\alpha = |a| + \beta\lambda$, following the results in [4, 5, 3], the matrix D for the transformation in

the conservative–dissipative form 6 is given by

$$D = \begin{pmatrix} 1 & 1 & 1 \\ \frac{\alpha+a}{\alpha-a}\sqrt{\frac{\lambda(\alpha-a)}{\alpha(\alpha+a)}} & 0 & -\sqrt{\frac{\lambda(\alpha-a)}{\alpha(\alpha+a)}} \\ -\sqrt{\frac{\lambda-\alpha}{\alpha}} & -\sqrt{\frac{\lambda-\alpha}{\alpha}} + \frac{\lambda}{\sqrt{\alpha(\lambda-\alpha)}} & -\sqrt{\frac{\lambda-\alpha}{\alpha}} \end{pmatrix}.$$

3. **The numerical approximation.** In this section we first introduce general finite difference approximations for system 1. Then, we compute the local truncation error of these schemes and we discuss its decays properties. The main result is given in Theorem 3.1, where a class of Time Asymptotically High Order (TAHO) schemes is fully characterized. First, we approximate the differential part following the direction of the characteristic velocities, so we study the methods for the system in diagonal form 1.

We denote by $f = (f^1, ..., f^m)$ the exact solution. Let Δx the uniform mesh-length and $x_j = j\,\Delta x$ the spatial grid points for all $j \in \mathbb{Z}$. The time levels t_n, with $t_0 = 0$, are also spaced uniformly with mesh-length $\Delta t = t_{n+1} - t_n$ for $n \in \mathbb{N}$. We denote by ρ the CFL ratio $\rho = \Delta t/\Delta x$, which is taken constant through all the paper.

We consider the Cauchy problem for system 1 possibly subjected to some stability conditions. The initial data f^0 is supposed to be smooth and approximated by its node values. The approximate solution $(f^1_{j,n}, ..., f^m_{j,n})^T$, $f^i_{j,n} \in \mathbb{R}^k$, $i = 1, ..., m$, for $j \in \mathbb{Z}$ and $n \in \mathbb{N}$, is given by

$$\begin{aligned} &\frac{f^i_{j,n+1} - f^i_{j,n}}{\Delta t} + \frac{\lambda_i}{2\Delta x}\left(f^i_{j+1,n} - f^i_{j-1,n}\right) - \frac{q_i}{2\Delta x}\delta_x^2 f^i_{j,n} \\ &= \sum_{l=-1,0,1} \left(\mathcal{B}^i_l(u_{j+l,n}) - \beta^i_l f^i_{j+l,n}\right), \end{aligned} \tag{10}$$

with $f^i_{j,0} = f^i_0(x_j)$ and $\delta_x^2 f_{j,n} = (f_{j+1,n} - 2f_{j,n} + f_{j+1,n})$, for all $i = 1, ..., m$. The artificial diffusion terms q_i are diagonal matrices in $\mathbb{R}^{k\times k}_+$. The source term approximation is defined, for $l = -1, 0, 1$, by the diagonal matrices $\beta^i_l \in \mathbb{R}^{k\times k}$ and by the vectors of functions $\mathcal{B}^i_l(\cdot) \in \mathbb{R}^k$.

We assume the scheme 11 is consistent with system 1, i.e, for all $i = 1, ..., m$

$$\beta^i_{-1} + \beta^i_0 + \beta^i_1 = I_k + \Delta x C^i,$$
$$\mathcal{B}^i_{-1}(u) + \mathcal{B}^i_0(u) + \mathcal{B}^i_1(u) = M_i(u) + \Delta x \mathcal{C}_i(u),$$

where $C^i = diag(c^i_1, ..., c^i_k) \in \mathbb{R}^{k\times k}$ and $\mathcal{C}_i(u)$ are k functions to be defined.

3.1. **Decay properties of the local truncation error.** In this section we focus on the local truncation error for the general scheme 11. By applying the time decay properties given in [3], we will show how it is possible to build up numerical schemes which are more accurate for large times.

Set, for $i = 1, ..., m$,

$$C = diag(C^i), \quad \bar{C} = DCD^{-1}, \quad \mathcal{C}(u) = (\mathcal{C}_i(u))^T, \quad \gamma^i = (\beta^i_1 - \beta^i_{-1}).$$

Scheme 11 is clearly consistent. Now, using the time decay estimates in [3] and similar estimates for their time derivatives, we obtain for a general approximation

the following estimates for the local truncation error, as $t \to +\infty$:

$$\mathcal{T}_u(x,t) = O(\Delta x\ t^{-3/2}) + O(\Delta t\ t^{-3/2}), \mathcal{T}_z(x,t) = O(\Delta x\ t^{-3/2}) + O(\Delta t\ t^{-3/2}).$$

We would like to improve the decay property of this local truncation error to build up more accurate numerical schemes. The main idea is to chose the free parameters of the scheme to delete the terms that decay more slowly in the Taylor expansion of the local truncation error (see [1] for details).

Let $g_i = diag(\gamma_{(i-1)k+1}, ..., \gamma_{ik})$ for $i = 1, ..., m$ and $G = diag(g_1, ..., g_m)$.

Theorem 3.1 (Local Truncation Error). *Let $\Delta t/\Delta x = \rho$ be fixed and let $H = diag(h_1, \dots, h_m)$ be the block diagonal matrix given by $H = D^T D$ and set $P = \sum_{i=1}^m \lambda_i^2 h_i^{-1}$. Assume $A_{11} \neq 0$ and that the following condition holds:*

$$\text{the matrix } (\lambda_i I_k - A_{11}) \text{ is invertible for } i = 1, ..., m.$$

If we make the following choice for the coefficients of the scheme 11,

$$C = -\frac{\rho}{2} I_{km}, \quad \mathcal{C} = CM(u) = -\frac{\rho}{2} M(u), \tag{11}$$

$$g_i = -\left(\frac{1}{2} q_i h_i^{-1} - \frac{\rho}{2} h_i^{-1} \left(P - (\lambda_i I_k - A_{11})^2\right)\right) (\lambda_i I_k - A_{11})^{-1} h_i \tag{12}$$

and

$$\Gamma_i'(u) = g_i M_i'(u) + \frac{\rho}{2}\left((h_i^{-1} - M_i'(u))A_{11} + \lambda_i M_i'(u) - h_i^{-1} \sum_{j=1}^m \lambda_j M_j'(u)\right), \tag{13}$$

both for $i = 1, ..., m$, then the scheme 11 is TAHO and the local truncation error reads

$$\mathcal{T}_u(x,t) = O\left(\Delta x\ t^{-2}\right) + O\left(\Delta x^2\ t^{-3/2}\right), \mathcal{T}_z(x,t) = O\left(\Delta x\ t^{-2}\right) + O\left(\Delta x^2\ t^{-3/2}\right).$$

For the proof and further considerations in case $A_{11} = 0$ we refer to [1].

4. **Numerical tests.** In this Section we show how, for large time simulations, TAHO schemes give better numerical results than standard approximations for both examples considered in Sections 2.1 and 2.2.

Specifically, we shall compare our TAHO scheme with two numerical approximations: i) a source pointwise approximation, denoted by STD and defined by 11 with $\Gamma_i = 0$, $\gamma^i = 0$, $C^i = 0$ and $\mathcal{C}_i = 0$, for $i = 1 =, ..., m$; ii) a source upwinding approximation, denoted by ROE and defined by

$$\begin{aligned} &\frac{f^i_{j,n+1} - f^i_{j,n}}{\Delta t} + \frac{\lambda_i}{2\Delta x}\left(f^i_{j+1,n} - f^i_{j-1,n}\right) - \frac{|\lambda_i|}{2\Delta x}\delta_x^2 f^i_{j,n} \\ =&\frac{M_i(u^n_{j-1}) + 2M_i(u^n_j) + M_i(u^n_{j+1})}{4} + \frac{sgn(\lambda_i)}{4}(M_i(u^n_{j-1}) - M_i(u^n_{j+1})) \\ &- \frac{f^i_{j-1,n} + 2f^i_{j,n} + f^i_{j+1,n}}{4} - \frac{sgn(\lambda_i)}{4}(f^i_{j-1,n} - f^i_{j+1,n}). \end{aligned} \tag{14}$$

To complete the definition of scheme 11 coupled with conditions 12-14 it is still necessary to choose some free parameters, such as for the 2×2 case $\mathcal{B}_0^{1,2}(\cdot)$ and $\beta_0^{1,2}$. For both cases considered, such parameters can be defined by applying monotonicity conditions to the scheme. We refer to [1] for more details.

For all tests, we focus our attention on the numerical error as a function of time: we plot the error $e(t) = \|(u^H - U^h)(t)\|_{L^\infty}$ as the time $t = n\Delta t$ increases, where u^H is the reference solution obtained by the ROE scheme 15, with $\Delta x = \mathcal{O}(10^{-4})$.

Then, given different numerical approximations U^h, we look for a constant C and γ which best fit the equality

$$e(t) = \|(u^H - U^h)(t)\|_{L^\infty} = Ct^{-\gamma}. \tag{15}$$

Given N data points $(t_i, e(t_i))_{i=1,N}$, we shall define γ and C as the solution of the following least squares problem,

$$\min_{C,\gamma} \sum_{i=1}^{N} |\ln(e(t_i)) - \ln(Ct^{-\gamma})|^2.$$

For all schemes, we fix the steps ratio ρ to verify all the CFL conditions. Since all schemes are of first order approximation, to emphasize the good behaviour of TAHO compared to the others schemes, we compute the numerical solutions U^h by using a quite big grid step $\Delta x = \mathcal{O}(10^{-1})$.

All numerical results we present show that for standard approximations, such as STD and ROE, the absolute error $e(t)$, for a fixed space step, decays as

$$e_u(t) = O(t^{-1/2}),\ e_z(t) = O(t^{-1}),$$

while for the TAHO scheme, it improves of $t^{-1/2}$ on the previous schemes.

4.1. **Results for the Jin–Xin 2×2 system.** We fix $q = \lambda$ and we compare for the 2×2 case the TAHO scheme coupled with monotonicity assumptions, with ROE and STD scheme. We shall consider as initial datum the function

$$u_0 = \chi_{[-1,1]}\left(-x^2 + 1\right), \quad z_0 = \frac{1}{\lambda} F(u_0(x)),$$

and we fix

$$F(u) = a\left(u - u^2\right).$$

The numerical results 1 show a better performance of the TAHO scheme; for both conservative and dissipative variable, the numerical solution obtained by TAHO fit better the benchmark curve. Again, the decay of the errors $e_u(t)$ and $e_z(t)$, Figure 1-(c)-(d), goes faster for the TAHO scheme, as confirmed by Table 1. There the decay parameter γ is numerically computed for all three schemes. The value obtained for the TAHO scheme improve of $t^{-1/2}$ on the others. We stress on that the numerical solutions are computed with quite big step $\Delta x = 0.1$.

4.2. **Results for the 3×3 system.** As initial data, we take the smooth function u_0 defined by

$$u_0(x) = \chi_{[-1,1]} \exp\left(1 - \frac{1}{1 - x^2}\right).$$

Then we set $f_0(x) = M(u_0(x))$. We choose $a = 1$, $\lambda = 2.1$, $\beta = (\alpha - a)/\lambda = 0.1$. The discretization parameters are $\Delta x = 0.1$, $\rho = \frac{1}{2\lambda}$, which satisfy all monotonicity requirements, see [1].

The numerical results show as in the 2×2 case a better performance of the TAHO scheme. In Figure 2-(a)-(b), we plot the time evolution of the l^∞ errors $e_u(t)$ and $e_z(t)$. They show how for the TAHO scheme both errors decay as time increases more quickly than other. This result is also confirmed by Table 2, where the values of γ and C are computed. Looking at the different values of γ, it is clear that for the TAHO approximation the decay velocity of the absolute error improves of $t^{-1/2}$ on the previous schemes.

scheme	C_u	γ_u	C_z	γ_z
STD	0.013797	0.374708	0.010744	0.341554
ROE	0.004874	0.333634	0.007850	0.439996
TAHO	0.111380	1.151517	0.495480	1.451030

TABLE 1. The 2×2 Test, see subsection 4.1. Evaluation of constants γ and C for $e_u(t) = C_u t^{-\gamma_u}$ and $e_z(t) = C_z t^{-\gamma_z}$ defined in 16. For standard approximation STD and ROE, the absolute error decays as $e_{u,z}(t) \approx O(t^{-1/2})$; while, for the TAHO scheme it improves of $t^{-1/2}$.

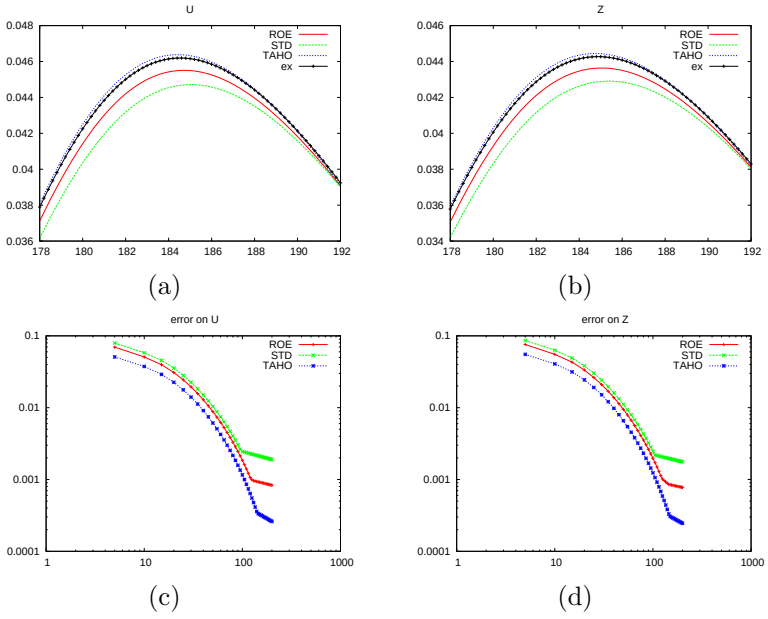

FIGURE 1. The 2×2 Test, see subsection 4.1. (a)-(b) Zoom on the solutions u and z respectively obtained by the different schemes at final time T. The plot show that TAHO scheme gives better results than others with a quite big step $\Delta x = 0.1$. (c)-(d) Time evolution of the l^∞ errors $e_u(t)$ and $e_z(t)$ defined in 16 for the different schemes. As expected by our asymptotic analysis, for the TAHO scheme the absolute errors $e_{u,z}(t)$ decay faster as the time increases. This result is confirmed in Table 1, where we compute the decay parameters γ of absolute errors previously plotted.

REFERENCES

[1] D. Aregba-Driollet, M. Briani and R. Natalini, *Time Asymptotic High Order schemes for dissipative BGK hyperbolic systems*, preprint, arXiv:math.NA/1207.6279v1

[2] D. Aregba-Driollet, M. Briani and R. Natalini, *Asymptotic high-order schemes for 2X2 dissipative hyperbolic systems*, SIAM J. Numer. Anal. **46** (2008), no.2, 869–894

scheme	C_u	γ_u	C_z	γ_z
STD	0.0052	0.54	0.0064	1.1
ROE	0.0027	0.66	0.0036	1.2
TAHO	0.006	1	0.012	1.62

TABLE 2. The 3×3 Test, see section 4.2. Evaluation of constants γ and C for $e_u(t) = C_u t^{-\gamma_u}$ and $e_z(t) = C_z t^{-\gamma_z}$ defined in 16. For STD and ROE approximations, the numerical results show that the absolute error decays as $e_u(t) = O(t^{-1/2})$ and $e_z(t) = O(t^{-1})$; while, for TAHO's it improves of $t^{-1/2}$.

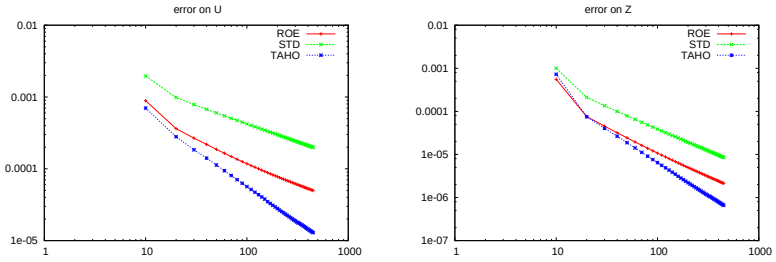

FIGURE 2. The 3×3 Test, see section 4.2. Absolute errors for u component (left) and Z components (right) with respect to time.

[3] S. BIANCHINI, B. HANOUZET AND R. NATALINI, *Asymptotic behavior of smooth solutions for partially Dissipative hyperbolic systems with a convex entropy*, Communications Pure Appl. Math. **60** (2007), 1559–1622.

[4] F. Bouchut, *Construction of BGK models with a family of kinetic entropies for a given system of conservation law*, J. Statist. Phys. **95** (1999), 113–170.

[5] B. Hanouzet, R. Natalini, *Global existence of smooth solutions for partially dissipative hyperbolic systems with a convex entropy*, Arch. Ration. Mech. Anal. **169** (2003), no. 2, 89–117.

[6] Y. Shizuta, S. Kawashima, *Systems of equations of hyperbolic–parabolic type with applications to the discrete Boltzmann equation*, Hokkaido Math. J. 14 (1984) 435–457.

Received xxxx 20xx; revised xxxx 20xx.

E-mail address: m.briani@iac.cnr.it
E-mail address: Denise.Aregba@math.u-bordeaux1.fr
E-mail address: r.natalini@iac.cnr.it