SēMA

BOLETÍN NÚMERO 50

Marzo 2010

sumario

Boletín de la Sociedad Española de Matemática Aplicada S$\vec{e}$MA

Página web de S$\vec{e}$MA
`http://www.sema.org.es/`

e-mail
`info@sema.org.es`

Dirección Editorial: Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla - La Mancha. Avda. de Camilo José Cela s/n. 13071. Ciudad Real. `boletin.sema@uclm.es`

ISSN 1575-9822.
Depósito Legal: AS-1442-2002.

Imprime: Gráficas Lope. C/ Laguna Grande, parc. 79, Políg. El Montalvo II 37008. Salamanca.

Diseño de portada: Ernesto Aranda
Ilustración de portada: imágenes creadas por Ulrich Pinkall, Etienne Ghys y Jos Leys tomadas de la exposición *Imaginary* (`http://www.imaginary2008.de`)

Estimados socios:

El número que os ofrecemos en esta ocasión es especial, pues incluye algunas de las contribuciones del pasado CEDYA celebrado en Ciudad Real. Concretamente nos han llegado los artículos que presentaron los conferenciantes plenarios Claude Le Bris y Manuel J. Castro y algunos artículos correspondientes a las sesiones especiales de *Integración Geométrica* y *Recent Advances in Adaptative Finite Elements*.

Reconociendo que no siempre es fácil separar el tiempo necesario para escribir este tipo de trabajos, queremos agradecer sinceramente el esfuerzo a los autores que han enviado sus contribuciones para que sean publicadas en este número. Confiamos que sean de vuestro interés y agrado.

Finalmente, queremos hacer notar el cambio de algunos miembros del Comité Científico del boletín. Agradecemos haber formado parte del Comité estos dos últimos años a Alfredo Bermúdez, Olivier Pironneau y Juan Luis Vázquez, que a petición propia han causado baja en el mismo, y aprovechamos para dar la bienvenida a Gregoire Allaire de la École Polythechnique de Paris (Francia), M. Carme Calderer de la Universidad de Minnesota (EEUU), Francisco Guillén de la Universidad de Sevilla y Daniel B. Szyld de Temple University (EEUU). Vaya por adelantado nuestro agradecimiento por haber aceptado colaborar y nuestra confianza en que estas incorporaciones permitan mejorar la calidad científica de esta publicación.

Recibid un cordial saludo,

Grupo Editor
boletin.sema@uclm.es

Sesiones plenarias

Claude Le Bris

Manuel Castro

Bol. Soc. Esp. Mat. Apl.
n°50(2010), 9–27

VARIANCE REDUCTION IN STOCHASTIC HOMOGENIZATION: PROOF OF CONCEPT, USING ANTITHETIC VARIABLES

RONAN COSTAOUEC, CLAUDE LE BRIS, FRÉDÉRIC LEGOLL

École Nationale des Ponts et Chaussées, 6 & 8 avenue Blaise Pascal, 77455 Marne-La-Vallée Cedex 2 and INRIA Rocquencourt, MICMAC team-project Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France

{costaour,lebris}@cermics.enpc.fr, legoll@lami.enpc.fr

Abstract

We show that we can reduce the variance in a simple problem of stochastic homogenization using the classical technique of antithetic variables. The setting, and the presentation, are deliberately kept elementary. We point out the main issues, show some illustrative results, and demonstrate, both theoretically and numerically, the efficiency of the approach on simple cases.

1 Introduction

Several settings in homogenization require the solution of corrector problems posed on the entire space $\mathbb{R}^d$. In practice, truncations of these problems over bounded domains are considered and the homogenized coefficients are obtained in the limit of large domains. The question arises as to how such computations can be accelerated. In the deterministic case, acceleration techniques reminiscent of signal filtering have been introduced in [5]. The work has since then been significantly improved by A. Gloria in [12]. In [5], it was shown that acceleration techniques efficient for deterministic problems do not necessarily perform well in the stochastic framework. In the latter case, the main difficulty is related to the intrinsic noise present in the simulation. The challenge is consequently not that much to improve the rate of convergence, which is intrinsically that of the central limit theorem, but rather to reduce the variance, thereby improving the prefactor of the convergence given by the central limit theorem. Although very well investigated in other application fields such as financial mathematics, variance reduction techniques seem to have not been applied to the context of stochastic homogenization. The purpose of the present contribution is to present a first attempt to reduce the variance in stochastic homogenization. For this purpose, we consider a simple situation, and a simple variance reduction technique. The probability theoretic arguments we will make use of are elementary. The equation under consideration is a simple

elliptic equation in divergence form, with a scalar coefficient. The coefficient is assumed to consist of independent, identically distributed random variables set on a simple mesh (see (2) below). The technique used for variance reduction is that of antithetic variables. Our setting is academic in nature, somewhat far from physically relevant cases, and elementary. Many more difficult situations could be addressed: other types of stationary ergodic coefficients, matrix rather than scalar coefficients, other types of equations, other techniques for variance reduction, ... The present contribution is a proof of concept: variance reduction *can* be achieved in stochastic homogenization. Future works [3, 4, 11] will provide more details on the numerics and the theory, and also address some of the many possible extensions mentioned above. We also mention the related work [13] on stochastic homogenization of *discrete* elliptic equations.

2 Stochastic homogenization theory

Although we wish to keep the mathematical formalism as limited as possible in our exposition, we need to introduce the basic setting of stochastic homogenization (see [17] for a similar presentation and related issues). Throughout this article, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and we denote by $\mathbb{E}(X) = \int_\Omega X(\omega) d\mathbb{P}(\omega)$ the expectation value of any random variable $X \in L^1(\Omega, d\mathbb{P})$. We next fix $d \in \mathbb{N}^*$ (the ambient physical dimension), and assume that the group $(\mathbb{Z}^d, +)$ acts on Ω. We denote by $(\tau_k)_{k\in\mathbb{Z}^d}$ this action, and assume that it preserves the measure $\mathbb{P}$, that is, for all $k \in \mathbb{Z}^d$ and all $A \in \mathcal{F}$, $\mathbb{P}(\tau_k A) = \mathbb{P}(A)$. We assume that the action τ is *ergodic*, that is, if $A \in \mathcal{F}$ is such that $\tau_k A = A$ for any $k \in \mathbb{Z}^d$, then $\mathbb{P}(A) = 0$ or 1. In addition, we define the following notion of stationarity (see [7]): any $F \in L^1_{\rm loc}\left(\mathbb{R}^d, L^1(\Omega)\right)$ is said to be *stationary* if, for all $k \in \mathbb{Z}^d$,

$$F(x + k, \omega) = F(x, \tau_k \omega), \tag{1}$$

almost everywhere in x and almost surely. In this setting, the ergodic theorem [16, 18] can be stated as follows: *Let $F \in L^\infty\left(\mathbb{R}^d, L^1(\Omega)\right)$ be a stationary random variable in the above sense. For $k = (k_1, k_2, \dots k_d) \in \mathbb{Z}^d$, we set $|k|_\infty = \sup_{1\le i\le d} |k_i|$. Then*

$$\frac{1}{(2N+1)^d} \sum_{|k|_\infty \le N} F(x, \tau_k \omega) \underset{N\to\infty}{\longrightarrow} \mathbb{E}\left(F(x, \cdot)\right) \quad \text{in } L^\infty(\mathbb{R}^d), \text{ almost surely.}$$

This implies that (denoting by Q the unit cube in $\mathbb{R}^d$)

$$F\left(\frac{x}{\varepsilon}, \omega\right) \overset{*}{\underset{\varepsilon\to 0}{\rightharpoonup}} \mathbb{E}\left(\int_Q F(x, \cdot) dx\right) \quad \text{in } L^\infty(\mathbb{R}^d), \text{ almost surely.}$$

Besides technicalities, the purpose of the above setting is simply to formalize that, even though realizations may vary, the function F at point $x \in \mathbb{R}^d$ and the function F at point $x + k$, $k \in \mathbb{Z}^d$, share the same law. In the homogenization

context we now turn to, this means that the local, microscopic environment (encoded in the coefficient a, see (3) below) is everywhere the same *on average*. From this, homogenized, macroscopic properties will follow.

We now fix an open, regular, bounded subset $\mathcal{D}$ of $\mathbb{R}^d$, a L^2 function f on $\mathcal{D}$, and a random function a assumed stationary in the sense (1) defined above. We also assume a is bounded, positive and almost surely bounded away from zero. For simplicity, we take a random piecewise constant function of the form:

$$a(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbf{1}_{Q+k}(x) a_k(\omega), \tag{2}$$

where Q is the unit cube of $\mathbb{R}^d$ and $(a_k(\omega))_{k\in\mathbb{Z}^d}$ denotes a family of i.i.d. random variables. The standard results of stochastic homogenization [2, 15] apply to the boundary value problem

$$\left\{ \begin{array}{rcl} -\mathrm{div}\left(a\left(\dfrac{x}{\varepsilon},\omega\right)\nabla u^\varepsilon\right) &=& f \quad \text{in} \quad \mathcal{D}, \\ u^\varepsilon &=& 0 \quad \text{on} \quad \partial\mathcal{D}. \end{array} \right. \tag{3}$$

These results state that, in the limit $\varepsilon \longrightarrow 0$, the homogenized problem obtained from (3) reads:

$$\left\{ \begin{array}{rcl} -\mathrm{div}\left(A^\star \nabla u^\star\right) &=& f \quad \text{in} \quad \mathcal{D}, \\ u^\star &=& 0 \quad \text{on} \quad \partial\mathcal{D}. \end{array} \right. \tag{4}$$

The homogenized matrix $A^\star$ is defined as

$$[A^\star]_{ij} = \mathbb{E}\left(\int_Q \left(e_i + \nabla w_{e_i}(y,\cdot)\right)^T a\left(y,\cdot\right)\left(e_j + \nabla w_{e_j}(y,\cdot)\right)\, dy\right), \tag{5}$$

where, for any $p \in \mathbb{R}^d$, w_p is the solution (unique up to the addition of a (random) constant) in $\left\{w \in L^2_{\mathrm{loc}}(\mathbb{R}^d, L^2(\Omega)), \quad \nabla w \in L^2_{\mathrm{unif}}(\mathbb{R}^d, L^2(\Omega))\right\}$ to

$$\left\{ \begin{array}{l} -\mathrm{div}\left[a\left(y,\omega\right)\left(p + \nabla w_p(y,\omega)\right)\right] = 0 \quad \text{a.s. on } \mathbb{R}^d, \\ \nabla w_p \quad \text{is stationary in the sense of (1)}, \\ \mathbb{E}\left(\displaystyle\int_Q \nabla w_p(y,\cdot)\, dy\right) = 0, \end{array} \right. \tag{6}$$

where we have used the notation L^2_{unif} for the *uniform* L^2 space, that is the space of functions for which, say, the L^2 norm on a ball of unit size is bounded above independently from the center of the ball.

The solution u^ε to (3) is known to converge to the solution $u^\star$ to (4) in various appropriate senses. The tensor and function $A^\star$ and $u^\star$ are *deterministic quantities*, although they originate from a series of random problems. This is a consequence of the *ergodic* setting described above, which allows random microscopic quantities to average out in deterministic macroscopic quantities.

Note however that the computation of $A^\star$ requires the computation of the so-called corrector functions w_p, which are random.

The above result generalizes that of the classical periodic setting (see *e.g.* [2, 9]) where, instead of being stationary ergodic, the function a in (3) is periodic. Then, although the homogenized problem can be expressed similarly, the crucial difference is that (at least in this simple linear case) the corrector problem can, in the periodic case, be reduced to the equation $-\mathrm{div}\left[a(y)\left(p+\nabla w_p(y)\right)\right]=0$ set on the periodic cell $Q=[0,1]^d$, and not on the entire space $\mathbb{R}^d$ as in (6). Correspondingly, the terms of the homogenized tensor in (5) are simple deterministic integrals on Q. In the random case, the corrector problem (6) is intrinsically set on the entire space and the numerical approximation of its solution w_p is the main computational challenge. Problem (6) is in practice truncated on a bounded domain $Q_N=[-N,N]^d$ and usually supplied with periodic boundary conditions:

$$\left\{\begin{array}{l} -\mathrm{div}\left(a(\cdot,\omega)\left(p+\nabla w_p^N(\cdot,\omega)\right)\right)=0 \quad \text{on} \quad Q_N, \\ w_p^N \text{ is } Q_N\text{-periodic.} \end{array}\right. \tag{7}$$

Correspondingly, we set:

$$\left[A_N^\star\right]_{ij}(\omega)=\frac{1}{|Q_N|}\int_{Q_N}\left(e_i+\nabla w_{e_i}^N(y,\omega)\right)^T a(y,\omega)\left(e_j+\nabla w_{e_j}^N(y,\omega)\right)dy. \tag{8}$$

In the limit of large domains Q_N, the homogenized tensor (5) is recovered. In addition, the rate of convergence with which the truncated values approach the exact homogenized value $A^\star$ can be assessed theoretically. We refer to [8, 19] for the proof of all the above statements. As will be seen below, the variance of the random variables involved plays a role in the approximation procedure. Reducing this variance is the problem we now consider.

3 Variance reduction

3.1 Classical Monte Carlo method

As mentioned above, the large size (large N) limit of the coefficient (8) obtained using the solution of the truncated corrector problem (7) gives the value of the homogenized coefficient (5). Formally, this is a convergence of the type $A_N^\star(\omega)\longrightarrow A^\star$ as $N\longrightarrow+\infty$ almost surely. The practical approach to this problem is the Monte Carlo approach. We now briefly investigate the role of the variance in the problem.

To start with, we consider the one-dimensional setting. Although this setting is very particular (and sometimes misleading because oversimplified), it also allows to already understand the basic features of the problem and the bottom line of the approach, with the economy of many unnecessary technicalities.

In the one-dimensional setting, the definition (2) reads

$$a(x,\omega)=\sum_{k\in\mathbb{Z}}\mathbf{1}_{[k,k+1[}(x)a_k(\omega) \tag{9}$$

with $(a_k(\omega))_{k\in\mathbb{Z}}$ a family of i.i.d. random variables. It is easily seen that the truncated corrector problem (7) can be explicitly solved and leads to the value

$$a_N^\star(\omega) = \left(\frac{1}{2N}\sum_{k=-N}^{N-1}\frac{1}{a_k(\omega)}\right)^{-1} \tag{10}$$

of the approximation for the homogenized tensor (here, a scalar coefficient of course). In the limit of large N, it almost surely converges to the value of the *exact* homogenized coefficient

$$a^\star = \mathbb{E}\left(\frac{1}{a_0}\right)^{-1}. \tag{11}$$

This exact value is readily obtained explicitly solving (5)-(6). The simplest possible argument consists now in considering $(a_N^\star(\omega))^{-1} = \dfrac{1}{2N}\displaystyle\sum_{k=-N}^{N-1}\frac{1}{a_k(\omega)}$ and remark that the rate of convergence of this quantity to $(a^\star)^{-1}$ is evidently given by the central limit theorem, where the variance of the random variable $(a_k(\omega))^{-1}$ plays a crucial role. Although correct, this argument exploits too much the very peculiar nature of the one-dimensional setting (we have taken the inverse of the coefficient and recasted it as a sum, a fact that is not possible otherwise than in one dimension). An argument with slightly more generality consists in considering $a_N^\star(\omega)$ itself – and not its inverse–, and, using elementary calculus, showing that it also converges to $a^\star$ with a rate of convergence where the variance of $a_0(\omega)$ again plays the crucial role. Indeed, one may for instance remark that $\mathbb{E}\left(\left|\left(\frac{1}{2N}\sum_{k=-N}^{N-1}\frac{1}{a_k}\right)^{-1} - \mathbb{E}\left(\frac{1}{a_0}\right)^{-1}\right|^2\right)$ can be bounded from above (using a simple almost sure upper bound of $a_k(\omega)$) by $\mathbb{E}\left(\left|\left(\frac{1}{2N}\sum_{k=-N}^{N-1}\frac{1}{a_k}\right) - \mathbb{E}\left(\frac{1}{a_0}\right)\right|^2\right)$ up to an irrelevant multiplicative constant and that the latter quantity, once easily computed, is of the form $\dfrac{1}{2N}\,\mathbb{V}\mathrm{ar}\left(\dfrac{1}{a_0}\right)$. Again, the variance of the random coefficient plays a role.

In dimensions higher than one, the situation is considerably more intricate and the rate of convergence with which the coefficient arising from the truncated computation converges to its limit is not so simple to evaluate. This is the purpose, under appropriate conditions (called *mixing conditions* and which are indeed met in our present setting), of the work [8].

The numerical practice is as follows. A set of M independent realizations of the random coefficient a are considered. The corresponding truncated problems (7) are solved, and an empirical mean of the truncated coefficients (8) is inferred. This empirical mean only agrees with the theoretical mean value of the truncated coefficient within a margin of error which is given by the central limit theorem (in terms of M). The variance of the coefficients therefore

again plays a role, as a prefactor. For a sufficiently large truncation size N, this truncated value is admitted to be the exact value of the coefficient. The error made is controlled by the estimations of the theoretical work [8]. Of course, the overall computation described above is expensive, because each realization requires a new solution to the d-dimensional boundary value problem (7) of presumably large a size since N is taken large. There is therefore a huge interest in reducing the cost of the computation, or, otherwise stated, in reaching a better accuracy at a given computational cost. Since the variance of the truncated homogenized tensor is an important ingredient, reducing the variance becomes a challenging and sensitive issue.

More explicitly, let $(a^{\mathbf{m}}(x,\omega))_{1\leq \mathbf{m}\leq M}$ denote M independent and identically distributed underlying random fields. We define a family $\left(A_N^{\star,\mathbf{m}}\right)_{1\leq \mathbf{m}\leq M}$ of i.i.d. homogenized matrices by, for any $1\leq i,j\leq d$,

$$\left[A_N^{\star,\mathbf{m}}\right]_{ij}(\omega) = \frac{1}{|Q_N|}\int_{Q_N} \left(e_i + \nabla w_{e_i}^{N,\mathbf{m}}(\cdot,\omega)\right)^T a^{\mathbf{m}}(\cdot,\omega)\ \left(e_j + \nabla w_{e_j}^{N,\mathbf{m}}(\cdot,\omega)\right),$$

where $w_{e_j}^{N,\mathbf{m}}$ is the solution of the corrector problem associated to $a^{\mathbf{m}}$. Then we define for each component of $A_N^\star$ the empirical mean and variance

$$\begin{aligned} \mu_M\left(\left[A_N^\star\right]_{ij}\right) &= \frac{1}{M}\sum_{\mathbf{m}=1}^{M}\left[A_N^{\star,\mathbf{m}}\right]_{ij}, \\ \sigma_M\left(\left[A_N^\star\right]_{ij}\right) &= \frac{1}{M-1}\sum_{\mathbf{m}=1}^{M}\left(\left[A_N^{\star,\mathbf{m}}\right]_{ij} - \mu_M\left(\left[A_N^\star\right]_{ij}\right)\right)^2. \end{aligned} \tag{12}$$

Since the matrices $A_N^{\star,\mathbf{m}}$ are i.i.d., the strong law of large numbers applies:

$$\mu_M\left(\left[A_N^\star\right]_{ij}\right)(\omega) \underset{M\to+\infty}{\longrightarrow} \mathbb{E}\left(\left[A_N^\star\right]_{ij}\right) \text{ almost surely.}$$

The central limit theorem then yields

$$\sqrt{M}\left(\mu_M\left(\left[A_N^\star\right]_{ij}\right) - \mathbb{E}\left(\left[A_N^\star\right]_{ij}\right)\right) \underset{M\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \sqrt{\mathbb{V}\mathrm{ar}\left(\left[A_N^\star\right]_{ij}\right)}\,\mathcal{N}(0,1), \tag{13}$$

where the convergence holds in law, and $\mathcal{N}(0,1)$ denotes the standard gaussian law. Introducing its 95 percent quantile, it is standard to consider that the exact mean $\mathbb{E}\left(\left[A_N^\star\right]_{ij}\right)$ is equal to $\mu_M\left(\left[A_N^\star\right]_{ij}\right)$ within a margin of error $1.96\dfrac{\sqrt{\mathbb{V}\mathrm{ar}\left(\left[A_N^\star\right]_{ij}\right)}}{\sqrt{M}}$. The exact variance $\mathbb{V}\mathrm{ar}\left(\left[A_N^\star\right]_{ij}\right)$ being unknown in practice, it is customary to replace it by the empirical variance given in (12) above. It is therefore considered that the expectation $\mathbb{E}\left(\left[A_N^\star\right]_{ij}\right)$ lies in the

interval

$$\left[\mu_M\left(\left[A_N^\star\right]_{ij}\right) - 1.96\frac{\sqrt{\sigma_M\left(\left[A_N^\star\right]_{ij}\right)}}{\sqrt{M}}\,,\ \mu_M\left(\left[A_N^\star\right]_{ij}\right) + 1.96\frac{\sqrt{\sigma_M\left(\left[A_N^\star\right]_{ij}\right)}}{\sqrt{M}}\right]. \tag{14}$$

The value $\mu_M\left(\left[A_N^\star\right]_{ij}\right)$ is thus, for both M and N sufficiently large, adopted as the approximation of the exact value $\left[A^\star\right]_{ij}$.

Of course, a tensorial argument could be applied here, not considering separately each entry of the matrix but treating the matrix as a whole. The approach developed above, component by component, is sufficient for the simple cases considered in the present work.

3.2 Antithetic variable for stochastic homogenization

We know from the previous section that constructing empirical means approximating $\mathbb{E}\left(A_N^\star\right)$ with a smaller variance at the same computational cost is of high interest. We now describe a possible approach to achieve this goal.

In generality, fix $M = 2\mathcal{M}$. Suppose that we give ourselves $\mathcal{M}$ i.i.d. copies $\left(a^{\mathbf{m}}(x,\omega)\right)_{1\leq\mathbf{m}\leq\mathcal{M}}$ of $a(x,\omega)$. Construct next $\mathcal{M}$ i.i.d. *antithetic* random fields

$$b^{\mathbf{m}}(x,\omega) = T\left(a^{\mathbf{m}}(x,\omega)\right), \quad 1 \leq \mathbf{m} \leq \mathcal{M},$$

from the $\left(a^{\mathbf{m}}(x,\omega)\right)_{1\leq\mathbf{m}\leq\mathcal{M}}$. The map T transforms the random field $a^{\mathbf{m}}$ into another, so-called *antithetic*, field $b^{\mathbf{m}}$. Explicit examples of such T are given in the sequel (see (20) and Section 4 below). The transformation is performed in such a way that, for each $\mathbf{m}$, $b^{\mathbf{m}}$ should have the same law as $a^{\mathbf{m}}$, namely the law of the coefficient a. Somewhat vaguely stated, if the coefficient a was obtained in a coin tossing game (using a fair coin), then $b^{\mathbf{m}}$ would be head each time $a^{\mathbf{m}}$ is tail and vice versa. We refer the reader to Figure 1 below for explicit illustrative examples of such a construction. Then, for each $1 \leq \mathbf{m} \leq \mathcal{M}$, we solve two corrector problems. One is associated to the original $a^{\mathbf{m}}$, the other one is associated to the antithetic field $b^{\mathbf{m}}$. Using its solution $v_p^{N,\mathbf{m}}$, we define the *antithetic homogenized matrix* $B_N^{\star,\mathbf{m}}$, whose elements read, for $1 \leq i,j \leq d$,

$$\left[B_N^{\star,\mathbf{m}}\right]_{ij}(\omega) = \frac{1}{|Q_N|}\int_{Q_N}\left(e_i + \nabla v_{e_i}^{N,\mathbf{m}}(\cdot,\omega)\right)^T b^{\mathbf{m}}(\cdot,\omega)\ \left(e_j + \nabla v_{e_j}^{N,\mathbf{m}}(\cdot,\omega)\right).$$

And finally we set, for any $1 \leq \mathbf{m} \leq \mathcal{M}$,

$$\widetilde{A}_N^{\star,\mathbf{m}}(\omega) := \frac{1}{2}\left(A_N^{\star,\mathbf{m}}(\omega) + B_N^{\star,\mathbf{m}}(\omega)\right). \tag{15}$$

Since $a^{\mathbf{m}}$ and $b^{\mathbf{m}}$ are identically distributed, so are $A_N^{\star,\mathbf{m}}$ and $B_N^{\star,\mathbf{m}}$. Thus, $\widetilde{A}_N^{\star,\mathbf{m}}$ is unbiased (that is, $\mathbb{E}\left(\widetilde{A}_N^{\star,\mathbf{m}}\right) = \mathbb{E}\left(A_N^{\star,\mathbf{m}}\right)$). In addition, it satisfies:

$$\widetilde{A}_N^{\star,\mathbf{m}} \underset{N\to+\infty}{\longrightarrow} A^\star \text{ almost surely},$$

because b is ergodic.

Let us define new estimators

$$\begin{aligned} \mu_{\mathcal{M}}\left(\left[\widetilde{A}_N^\star\right]_{ij}\right) &= \frac{1}{\mathcal{M}}\sum_{\mathbf{m}=1}^{\mathcal{M}}\left[\widetilde{A}_N^{\star,\mathbf{m}}\right]_{ij}, \\ \sigma_{\mathcal{M}}\left(\left[\widetilde{A}_N^\star\right]_{ij}\right) &= \frac{1}{\mathcal{M}-1}\sum_{\mathbf{m}=1}^{\mathcal{M}}\left(\left[\widetilde{A}_N^{\star,\mathbf{m}}\right]_{ij} - \mu_{\mathcal{M}}\left(\left[\widetilde{A}_N^\star\right]_{ij}\right)\right)^2, \end{aligned} \tag{16}$$

which require $2\mathcal{M}$ resolutions of corrector problems, i.e. as many as the classical estimators (12), since we choose $M = 2\mathcal{M}$. In addition, note that we have built a new random variable whose variance is

$$\mathbb{V}\mathrm{ar}\left(\left[\widetilde{A}_N^\star\right]_{ij}\right) = \frac{1}{2}\mathbb{V}\mathrm{ar}\left(\left[A_N^\star\right]_{ij}\right) + \frac{1}{2}\mathbb{C}\mathrm{ov}\left(\left[A_N^\star\right]_{ij}, \left[B_N^\star\right]_{ij}\right). \tag{17}$$

Applying the central limit theorem to $\widetilde{A}_N^\star$, we obtain

$$\sqrt{\mathcal{M}}\left(\mu_{\mathcal{M}}\left(\left[\widetilde{A}_N^\star\right]_{ij}\right) - \mathbb{E}\left(\left[A_N^\star\right]_{ij}\right)\right) \underset{\mathcal{M}\to+\infty}{\xrightarrow{\mathcal{L}}} \sqrt{\mathbb{V}\mathrm{ar}\left(\left[\widetilde{A}_N^\star\right]_{ij}\right)}\,\mathcal{N}(0,1). \tag{18}$$

Similarly to (14), we deduce a confidence interval from this convergence. The exact mean $\mathbb{E}\left(\left[\widetilde{A}_N^\star\right]_{ij}\right)$ is equal to $\mu_{\mathcal{M}}\left(\left[\widetilde{A}_N^\star\right]_{ij}\right)$ within a margin of error $1.96\dfrac{\sqrt{\mathbb{V}\mathrm{ar}\left(\left[\widetilde{A}_N^\star\right]_{ij}\right)}}{\sqrt{\mathcal{M}}}$. It results from (17) that, if

$$\mathbb{C}\mathrm{ov}\left(\left[A_N^\star\right]_{ij}, \left[B_N^\star\right]_{ij}\right) \leq 0, \tag{19}$$

then the width of this interval has been diminished by the new approach, and, correspondingly, the quality of approximation at given computational cost has increased.

To understand slightly more in details at the theoretical level why the approach is likely to perform well, we again consider the one-dimensional setting (9) for which we recall the explicit expressions (10) and (11) for the truncated and the exact homogenized coefficients, respectively.

Suppose as a first illustration that a_0 is a Bernoulli distributed random variable $a_0 \sim \mathcal{B}(1/2)$:

$$\mathbb{P}(a_0 = \alpha) = 1/2 \quad \text{and} \quad \mathbb{P}(a_0 = \beta) = 1/2,$$

for some $0 < \alpha < \beta$. Defining the antithetic variable

$$b_k(\omega) = \alpha + \beta - a_k(\omega)$$

and next the antithetic field

$$b(x,\omega) = \sum_{k\in\mathbb{Z}} \mathbf{1}_{[k,k+1[}(x)\, b_k(\omega) = \sum_{k\in\mathbb{Z}} \mathbf{1}_{[k,k+1[}(x)\, (\alpha+\beta-a_k(\omega))\,, \tag{20}$$

it is immediately seen that

$$\frac{1}{2}\left(\frac{1}{a_N^\star(\omega)} + \frac{1}{b_N^\star(\omega)}\right) = \mathbb{E}\left(\frac{1}{a_0}\right).$$

The variance of the inverse of the truncated coefficient has vanished. This example might seem oversimplified because we are indeed making use of two peculiarities of the problem: the set $\{\alpha,\beta\}$ of values taken by the coefficient a has cardinality two, and the explicit expression (10) allows us to explicitly manipulate the inverse of the homogenized coefficient. The situation, although oversimplified, is yet a first good indicator of the interest of the approach. As in the previous section, we can be slightly more general, by considering for instance that the random coefficient a is now uniformly distributed over a given interval, say $a_0 \sim \mathcal{U}([\alpha,\beta])$. Then,

$$\frac{1}{2}\left(\frac{1}{a_N^\star(\omega)} + \frac{1}{b_N^\star(\omega)}\right) = \frac{1}{2N}\sum_{k=-N}^{N-1} \frac{1}{2}\left(\frac{1}{a_k(\omega)} + \frac{1}{b_k(\omega)}\right). \tag{21}$$

It is a simple matter to show that, because the function $x \mapsto 1/x$ is decreasing, we have

$$\mathbb{C}\mathrm{ov}\left(\frac{1}{a_0}, \frac{1}{b_0}\right) \le 0. \tag{22}$$

Consider indeed a decreasing function f, and X and Y two independent random variables, identically distributed according to $\mathcal{U}([\alpha,\beta])$. Since $x \mapsto f(\alpha+\beta-x)$ is increasing, we observe that

$$(f(X)-f(Y))\ (f(\alpha+\beta-X)-f(\alpha+\beta-Y)) \le 0,$$

hence

$$\mathbb{E}[f(X)\ f(\alpha+\beta-X)] \le \mathbb{E}[f(X)]\ \mathbb{E}[f(\alpha+\beta-X)],$$

which reads $\mathbb{C}\mathrm{ov}[f(X), f(\alpha+\beta-X)] \le 0$. Choosing $f(x) = 1/x$ yields (22).

Since

$$\mathbb{V}\mathrm{ar}\left(\frac{1}{2}\left(\frac{1}{a_N^\star} + \frac{1}{b_N^\star}\right)\right) = \frac{1}{4N}\mathbb{V}\mathrm{ar}\left(\frac{1}{a_0}\right) + \frac{1}{4N}\mathbb{C}\mathrm{ov}\left(\frac{1}{a_0}, \frac{1}{b_0}\right),$$

we conclude that

$$\mathbb{V}\mathrm{ar}\left(\frac{1}{2}\left(\frac{1}{a_N^\star} + \frac{1}{b_N^\star}\right)\right) \le \mathbb{V}\mathrm{ar}\left(\frac{1}{a_{2N}^\star}\right).$$

Therefore, $\mathbb{E}(1/a_0)$ can be approximated either by (21) or by $1/a_{2N}^\star$, with an equal cost (i.e. an equal number of random variables in both sums), but the former has a smaller variance than the latter. It is hence of better quality.

As mentioned above, the practice in dimensions higher than one is to generate a set of identically distributed coefficients for each truncated corrector problem, and to use (15). The appropriate analogous one-dimensional approach is to consider $\mathcal{M} = \dfrac{M}{2}$ independent copies of $a(x,\omega)$ and set

$$\begin{aligned} \widetilde{a}_N^{\star,\mathbf{m}}(\omega) &:= \frac{1}{2}\left(a_N^{\star,\mathbf{m}}(\omega) + b_N^{\star,\mathbf{m}}(\omega)\right) \\ &= \frac{1}{2}\left(\frac{1}{2N}\sum_{k=-N}^{N-1}\frac{1}{a_k^{\mathbf{m}}(\omega)}\right)^{-1} + \frac{1}{2}\left(\frac{1}{2N}\sum_{k=-N}^{N-1}\frac{1}{b_k^{\mathbf{m}}(\omega)}\right)^{-1} \end{aligned}$$

with empirical mean

$$\mu_{\mathcal{M}}\left(\widetilde{a}_N^\star\right)(\omega) = \frac{1}{\mathcal{M}}\sum_{\mathbf{m}=1}^{\mathcal{M}} \widetilde{a}_N^{\star,\mathbf{m}}(\omega).$$

We approach more generality since

$$\mu_{\mathcal{M}}\left(\widetilde{a}_N^\star\right)(\omega) \underset{\mathcal{M}\to+\infty}{\longrightarrow} \mathbb{E}\left(\widetilde{a}_N^\star\right) = \mathbb{E}\left(a_N^\star\right) \quad \text{almost surely,}$$

but $\mathbb{E}\left(a_N^\star\right) \neq a^\star$. It can again be remarked that $a_N^\star(\omega)$ is an increasing function of the uniform variables $(a_k(\omega))_{k\in\mathbb{Z}}$. From this observation, it is possible to show that $\mathbb{C}\mathrm{ov}\left(a_N^\star, b_N^\star\right) \leq 0$, and to conclude that the variance of $\mu_{\mathcal{M}}\left(\widetilde{a}_N^\star\right)$ is smaller than that of $\mu_{2\mathcal{M}}\left(a_N^\star\right)$. For this proof on a model by analogy, as well as for proofs that variance reduction is indeed achieved for some *actual* settings in dimensions higher than one (such as for instance those from [1, 10, 6]), we refer to [3, 11]. The above simplified arguments were only meant to have pedagogic value.

4 Numerical experiments

The previous section provides some elementary ingredients for a theoretical analysis of the efficiency of the approach. The one-dimensional setting is however too particular. More convincing theoretical arguments have to be developed. As announced, this will be the purpose of future publications. Meanwhile, it is possible to test the approach on actual two-dimensional cases, and it is the purpose of this section to report on such tests. As above, we only consider random coefficients that are piecewise constant and of the form (2). The test cases we choose correspond to three different laws for a_0:

- case (i): a Bernoulli law of parameter 1/2, namely $a_0 \sim \mathcal{B}(1/2)$, $\mathbb{P}(a_0 = \alpha) = 1/2$ and $\mathbb{P}(a_0 = \beta) = 1/2$;
- case (ii): a Bernoulli law of parameter 1/3, namely $a_0 \sim \mathcal{B}(1/3)$, $\mathbb{P}(a_0 = \alpha) = 1/3$ and $\mathbb{P}(a_0 = \beta) = 2/3$;
- case (iii): a uniform law, namely $a_0 \sim \mathcal{U}([\alpha,\beta])$.

We take the specific values $\alpha = 3$ and $\beta = 20$, just to fix the ideas. Similar qualitative conclusions would be reached with other generic values. Figure 1 shows a realization of a and its antithetic field b in cases (i) and (iii).

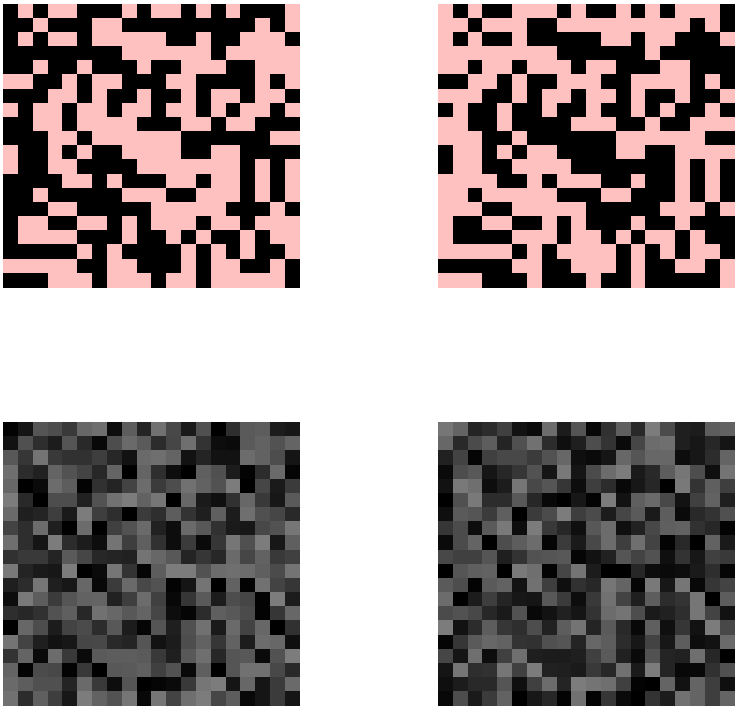


Figure 1: Realization of $a(x,\omega)$ given by (2) (left) and the associated antithetic field $b(x,\omega)$ (right). Top figures: $a_0 \sim \mathcal{B}(1/2)$; bottom figures: $a_0 \sim \mathcal{U}([\alpha,\beta])$.

Our numerical tests have been performed using the finite elements software FreeFem++ developed by F. Hecht (Paris VI, see [14]). The discretization of the corrector problem is performed using $\mathbb{P}1$ Lagrange finite elements, and a regular Q-periodic mesh of Q_N. The discretization meshsize is fixed and has value $h = 0.2$.

It is worth mentioning how we practically proceed to generate an antithetic variable. This may indeed be delicate. We have taken random coefficients that can all originally be expressed in terms of a uniformly distributed random variable (with a view, notably, to be consistent with the way a random variable is practically generated on a computer). We then build the antithetic variable precisely using the 'mother' uniform random variable. The technique

is best explained on case (ii). Write the variable $a_0 \sim \mathcal{B}(1/3)$ as $a_0 \sim \alpha + (\beta - \alpha)\mathbf{1}_{\{1/3 \leq U_0 \leq 1\}}$ where $U_0 \sim \mathcal{U}([0,1])$ denotes a random variable that has uniform law on the interval $[0,1]$. The antithetic variable is then taken as $b_0 \sim \alpha + (\beta - \alpha)\mathbf{1}_{\{0 \leq U_0 \leq 2/3\}}$ and the correspondence is made realization by realization using the actual realization of U_0.

In cases (i) and (ii), in dimension 2, the *exact* homogenized tensor is known to be *isotropic*, $A^\star = a^\star \mathbb{I}_2$ (see [15, Chap. 7, pp. 234-237] for a proof). Of course, for N finite, $A_N^\star$ is a generic matrix, but our numerical experiments consistently show that, for N sufficiently large, the off-diagonal terms are very small on average compared to the diagonal terms, in the three cases we have considered. Table 1 summarizes, in case (iii), the estimated means and standard deviations of the components of $A_N^\star$ for different values of N. It confirms that the main sources of variance are the diagonal terms. The same conclusion holds in cases (i) and (ii).

N	$[A_N^\star]_{11}$	$[A_N^\star]_{22}$	$[A_N^\star]_{12}$
5	10.42 (0.608)	10.39 (0.620)	0.00391 (0.074)
10	10.39 (0.269)	10.39 (0.273)	0.00369 (0.033)
20	10.37 (0.171)	10.37 (0.162)	0.00089 (0.017)
40	10.39 (0.069)	10.39 (0.070)	-0.00219 (0.0095)
60	10.38 (0.045)	10.38 (0.045)	0.00059 (0.0069)
80	10.38 (0.033)	10.38 (0.034)	0.00013 (0.0047)
100	10.38 (0.028)	10.38 (0.028)	0.00010 (0.0033)

Table 1: For each entry of $A_N^\star$, empirical mean $\mu_{100}\left([A_N^\star]_{ij}\right)$ (and empirical standard deviation $\sigma_{100}^{1/2}\left([A_N^\star]_{ij}\right)$, in brackets), in the case (iii).

In our three test cases, we have compared for different values of N the estimated variance of $\left[\widetilde{A}_N^\star\right]_{11}$ with that of $[A_N^\star]_{11}$. In order to quantitatively assess the efficiency of the antithetic variables method, we introduce the effectivity ratio

$$R\left([A_N^\star]_{11}\right) = \frac{\sigma_{100}\left([A_N^\star]_{11}\right)}{2\sigma_{50}\left(\left[\widetilde{A}_N^\star\right]_{11}\right)}.$$

The factor 2 at the denominator accounts for the number of realizations associated to the classical and antithetic Monte Carlo methods, given that we wish to work at fixed computational cost. Indeed, after solving $M = 2\mathcal{M}$ corrector problems (7), one can either build a confidence interval of size $1.96\sqrt{\sigma_M\left([A_N^\star]_{11}\right)/M}$ following (13) and (14), or a confidence interval of size $1.96\sqrt{\sigma_{\mathcal{M}}\left(\left[\widetilde{A}_N^\star\right]_{11}\right)/\mathcal{M}}$ following (18).

Our next table, Table 2, contains the values of this representative ratio for each test case. We have also plotted on Figure 2 the curves of estimated

means (12) and (16), with their confidence intervals, for the three cases under study here.

If we admit that the theory developed in the previous section applies to the two-dimensional case, another manner to check variance reduction is to compute the empirical covariance between $[A_N^\star]_{11}$ and $[B_N^\star]_{11}$ (recall (19)). This is the reason why we have also plotted on Figure 2 the normalized empirical value of this covariance,

$$\frac{\mathbb{C}\mathrm{ov}\left([A_N^\star]_{11}, [B_N^\star]_{11}\right)}{\sqrt{\mathbb{V}\mathrm{ar}\left([A_N^\star]_{11}\right)\ \mathbb{V}\mathrm{ar}\left([B_N^\star]_{11}\right)}}, \tag{23}$$

for test case (iii) (similar results have been obtained for the two other test cases).

N	$a_0 \sim \mathcal{B}(1/2)$	$a_0 \sim \mathcal{B}(1/3)$	$a_0 \sim \mathcal{U}([\alpha, \beta])$
5	5.34	2.06	6.31
10	3.91	1.56	6.46
20	5.41	2.92	10.2
40	3.07	2.31	6.67
60	4.41	2.47	6.16
80	4.49	1.95	5.68
100	4.28	2.99	7.89

Table 2: Representative effectivity ratios $R\left([A_N^\star]_{11}\right)$ for test cases (i), (ii) and (iii). The number shown gives the gain in computational time or, equivalently, at given computational cost, the square of the gain in the width of the confidence interval.

The results are self-explanatory: the variance is reduced. The reduction is not spectacular, but it is definite, and, equally importantly, systematic. Considering that the approach induces no additional computational cost at all, this is very good. Other more adapted, but also more delicate to design and implement, variance reduction approaches will be tested in the future [4, 11], and one may expect even more significant reductions.

5 Variance reduction for the solution $u^\star$

We conclude this article examining the problem of variance reduction from a slightly different perspective. We have so far investigated the question of variance reduction for the homogenized tensor $A^\star$. This is the question typically relevant in Mechanics, where for instance determining the homogenized tensor is an important issue because it allows to define, say, the Young modulus or the Poisson ratio of the homogenized material. In some contexts however, the focus is more on the solution of the homogenized problem, rather than on the coefficients of the homogenized equation. For a given right-hand side f in (3) (or for a *set* of such right-hand sides), one wishes to know the behaviour of the solution u^ε for small ε. Now, reducing the variance on the solution u is

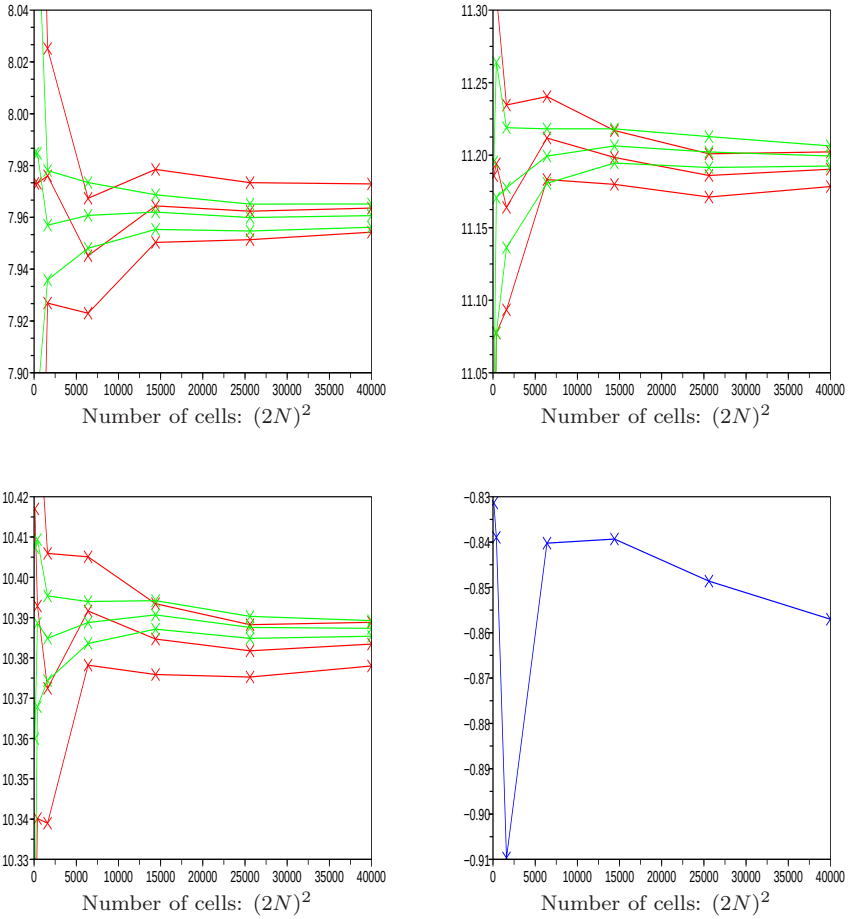


Figure 2: Estimated means (with confidence intervals) for $[A_N^\star]_{11}$ (red) and $\left[\widetilde{A}_N^\star\right]_{11}$ (green), in the cases $a_0 \sim \mathcal{B}(1/2)$ (top left), $a_0 \sim \mathcal{B}(1/3)$ (top right) and $a_0 \sim \mathcal{U}([\alpha, \beta])$ (bottom left). In the latter case, we also plot the estimator (23) of the normalized covariance between $[A_N^\star]_{11}$ and $[B_N^\star]_{11}$ (bottom right).

not exactly the same question as reducing the variance on the coefficients of the equation (because the map that associates the solution to the coefficients of the equation is a highly nonlinear nonlocal map). Note also that a systematic way to investigate the question would of course be to study the variance of the homogenized operator itself (or of its eigenelements) and it is indeed on our agenda to do so in a more extensive article [4, 11]. But for the time being, we briefly mention here a possible variance reduction approach on the solution $u^\star$, for a given representative right-hand side f.

In principle, one may think of several possible ways for computing the solution $u^\star$ to the homogenized problem (4). A first approach, which we denote by **(M1)**, consists in the following schematic sequence of computations

$$(a^{\mathbf{m}}(x,\omega))_{1\leq \mathbf{m}\leq M} \overset{\text{corrector pb}}{\longrightarrow} \left(A_N^{\star,\mathbf{m}}(\omega)\right)_{1\leq \mathbf{m}\leq M} \overset{\frac{1}{M}\sum}{\longrightarrow} \mu_M\left(A_N^\star\right) \overset{(24)}{\longrightarrow} u_{N,M}^\star,$$

where $u_{N,M}^\star$ solves the boundary value problem

$$\left\{ \begin{array}{rcl} -\mathrm{div}\left(\mu_M\left(A_N^\star\right)(\omega)\,\nabla u_{N,M}^\star(x,\omega)\right) &=& f \quad \text{in} \quad \mathcal{D}, \\ u_{N,M}^\star(x,\omega) &=& 0 \quad \text{on} \quad \partial\mathcal{D}. \end{array} \right. \tag{24}$$

In short, **(M1)** consists in *first* approximating $A^\star$ using the Monte Carlo approach and its outcome $\mu_M\left(A_N^\star\right)$, and *next* to solve for $u_{N,M}^\star$.

A second approach, **(M2)**, consists in the sequence

$$(a^{\mathbf{m}}(x,\omega))_{1\leq \mathbf{m}\leq M} \overset{\text{corrector pb}}{\longrightarrow} \left(A_N^{\star,\mathbf{m}}(\omega)\right)_{1\leq \mathbf{m}\leq M} \overset{(25)}{\longrightarrow} \left(u_N^{\star,\mathbf{m}}(\cdot,\omega)\right)_{1\leq \mathbf{m}\leq M}.$$

Otherwise stated, for each $1 \leq \mathbf{m} \leq M$, the problem

$$\left\{ \begin{array}{rcl} -\mathrm{div}\left(A_N^{\star,\mathbf{m}}\nabla u_N^{\star,\mathbf{m}}\right) &=& f \quad \text{in} \quad \mathcal{D}, \\ u_N^{\star,\mathbf{m}} &=& 0 \quad \text{on} \quad \partial\mathcal{D}, \end{array} \right. \tag{25}$$

is first solved, and the empirical mean and variance of the corresponding solutions are next constructed:

$$\begin{array}{rcl} \mu_M\left(u_N^\star\right)(x,\omega) &=& \dfrac{1}{M}\displaystyle\sum_{\mathbf{m}=1}^{M} u_N^{\star,\mathbf{m}}(x,\omega), \\ \sigma_M\left(u_N^\star\right)(x,\omega) &=& \dfrac{1}{M-1}\displaystyle\sum_{\mathbf{m}=1}^{M}\left(u_N^{\star,\mathbf{m}}(x,\omega)-\mu_M\left(u_N^\star\right)(x,\omega)\right)^2. \end{array} \tag{26}$$

The empirical mean is then taken as the approximation of our seeked solution $u^\star$.

Of course, it is immediately seen that a set of approaches, intermediate between **(M1)** and **(M2)**, can be designed. This is the set of approaches **(M3)**. For each $1 \leq \mathbf{m} \leq M$, we first solve the corrector problem, and thus obtain $A_N^{\star,\mathbf{m}}(\omega)$. We next set $M = PR$, and define, for each $1 \leq \mathbf{r} \leq R$,

$$\mu_P^{\mathbf{r}}\left(A_N^\star\right)(\omega) = \frac{1}{P}\sum_{\mathbf{p}=1}^{P} A_N^{\star,\mathbf{p}+(\mathbf{r}-1)P}(\omega),$$

which is an empirical mean computed with P realizations among the M available realizations. For each $1 \leq \mathbf{r} \leq R$, we next solve the boundary value problem

$$\left\{ \begin{array}{rcl} -\mathrm{div}\left(\mu_P^{\mathbf{r}}\left(A_N^\star\right)\nabla u_N^{\star,\mathbf{r}}\right) &=& f \quad \text{in} \quad \mathcal{D}, \\ u_N^{\star,\mathbf{r}} &=& 0 \quad \text{on} \quad \partial\mathcal{D}. \end{array} \right.$$

The estimators for $u^\star$ then are

$$\begin{aligned} \mu_{R,P}\left(u_N^\star\right)(x,\omega) &= \frac{1}{R}\sum_{\mathbf{r}=1}^{R} u_N^{\star,\mathbf{r}}(x,\omega), \\ \sigma_{R,P}\left(u_N^\star\right)(x,\omega) &= \frac{1}{R-1}\sum_{\mathbf{r}=1}^{R}\left(u_N^{\star,\mathbf{r}}(x,\omega) - \mu_{R,P}\left(u_N^\star\right)(x,\omega)\right)^2. \end{aligned}$$

We observe that, in dimension one, the solution of (25) satisfies

$$\left(u_N^{\star,\mathbf{m}}\right)'(x,\omega) = -\frac{1}{a_N^{\star,\mathbf{m}}(\omega)}\left(F(x) - \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} F\right),$$

where $F(x)$ is such that $F'(x) = f(x)$. Hence, in view of (10) and (11), we have

$$\mathbb{E}\left[\left(u_N^{\star,\mathbf{m}}\right)'\right] = -\frac{1}{a^\star}\left(F(x) - \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} F\right) = \mathbb{E}\left[(u^\star)'\right].$$

As a consequence, the empirical mean built following approach **(M2)**, namely $\mu_M\left(u_N^\star\right)(x,\omega)$ defined by (26), is an unbiased estimator of $u^\star(x)$, for any finite N and M, in the one-dimensional case. The estimators built following approaches **(M1)** and **(M3)** do not share this property.

In the present work, we only consider approach **(M2)**, leaving the study of the other approaches for future works. We apply the exact same technique as above, considering antithetic variables to reduce the variance. The variance under consideration is however now that of the approximation of $u^\star$.

We consider the test case (iii) defined in the previous section. We choose the right-hand side $f(x,y) = (x-0.5)^2 + (y-0.5)^2$ on the domain $\mathcal{D} = (0,1)^2$ (similar results have been obtained with other right-hand sides). The efficiency of the antithetic variable technique is assessed using the following ratio

$$R\left(u_N^\star\right) = \operatorname*{Ess\,Inf}_{x\in\mathcal{D}} \frac{\sigma_{100}\left(u_N^\star\right)}{2\sigma_{50}\left(\widetilde{u}_N^\star\right)}. \tag{27}$$

We have also checked that the technique does not introduce any bias by monitoring the estimator

$$\operatorname*{Ess\,Sup}_{x\in\mathcal{D}} \left|\frac{\mu_{100}\left(u_N^\star\right) - \mu_{50}\left(\widetilde{u}_N^\star\right)}{\mu_{100}\left(u_N^\star\right)}\right|. \tag{28}$$

Numerical results are gathered in Table 3. We observe that the technique does not introduce any bias, and that, again, a significant variance reduction, at fixed computational cost, is obtained.

N	Estimator (28)	Estimator (27)
5	4.20×10^{-4}	10.1
10	3.80×10^{-4}	10.9
20	1.56×10^{-3}	14.6
40	4.05×10^{-4}	11.8
80	5.21×10^{-4}	9.10
100	3.24×10^{-4}	9.02

Table 3: Estimator (28) of the bias, and estimator (27) of the variance reduction, in the case $a_0 \sim \mathcal{U}([\alpha, \beta])$ (the equation (25) has been solved on a mesh of size $h = 0.1$).

Acknowledgments

The authors thank Xavier Blanc for several stimulating discussions. Part of this work was initiated while the second author was visiting the Institute for Mathematics and its Applications and the Department of Mathematics of the University of Minnesota. The hospitality of these institutions is gratefully acknowledged. The work of the authors is partially supported by ONR under contract Grant 00014-09-1-0470.

References

[1] A. Anantharaman and C. Le Bris, Homogenization of a weakly randomly perturbed periodic material, *C. R. Acad. Sci. Série I*, 2009, submitted.

[2] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, **Asymptotic analysis for periodic structures**, Studies in Mathematics and its Applications, 5. North-Holland Publishing Co., Amsterdam-New York, 1978.

[3] X. Blanc, R. Costaouec, C. Le Bris, and F. Legoll, Variance reduction in stochastic homogenization using antithetic variables, in preparation.

[4] X. Blanc, R. Costaouec, C. Le Bris, and F. Legoll, in preparation.

[5] X. Blanc and C. Le Bris, Improving on computation of homogenized coefficients in the periodic and quasi-periodic settings, *Netw. Heterog. Media*, 5(1): 1–29, 2010.

[6] X. Blanc, C. Le Bris, and P.-L. Lions, Une variante de la théorie de l'homogénéisation stochastique des opérateurs elliptiques [A variant of stochastic homogenization theory for elliptic operators], *C. R. Acad. Sci. Série I*, 343(11-12):717–724, 2006.

[7] X. Blanc, C. Le Bris, and P.-L. Lions, Stochastic homogenization and random lattices, *J. Math. Pures Appl.*, 88(1):34–63, 2007.

[8] A. Bourgeat and A. Piatnitski, Approximation of effective coefficients in stochastic homogenization, *Ann I. H. Poincaré - PR*, 40(2):153–165, 2004.

[9] D. Cioranescu and P. Donato, **An introduction to homogenization**, Oxford Lecture Series in Mathematics and its Applications, 17. Oxford University Press, New York, 1999.

[10] R. Costaouec, C. Le Bris, and F. Legoll, Approximation numérique d'une classe de problèmes en homogénéisation stochastique [Numerical approximation of a class of problems in stochastic homogenization], *C. R. Acad. Sci. Série I*, 348(1-2):99–103, 2010.

[11] R. Costaouec, Université Paris Est, Ph.D thesis, in preparation.

[12] A. Gloria, Reduction of the resonance error. Part 1: Approximation of homogenized coefficients, preprint available at `http://hal.archives-ouvertes.fr/inria-00457159/en/`.

[13] A. Gloria and F. Otto, An optimal error estimate in stochastic homogenization of discrete elliptic equations, preprint available at `http://hal.inria.fr/hal-00383953`.

[14] `FreeFEM, http://www.freefem.org`

[15] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, **Homogenization of differential operators and integral functionals**, Springer-Verlag, 1994.

[16] U. Krengel, **Ergodic theorems**, de Gruyter Studies in Mathematics, vol. 6, de Gruyter, 1985.

[17] C. Le Bris, Some numerical approaches for "weakly" random homogenization, *Proceedings of ENUMATH 2009*, Lect. Notes Comput. Sci. Eng., Springer, in press.

[18] A. N. Shiryaev, **Probability**, Graduate Texts in Mathematics, vol. 95, Springer, 1984.

[19] V. V. Yurinskii, Averaging of symmetric diffusion in random medium, *Sibirskii Mat. Zh.*, 27(4):167–180, 1986.

Bol. Soc. Esp. Mat. Apl.
n°50(2010), 27–45

ON THE BENEFITS OF USING GPUS TO SIMULATE SHALLOW FLOWS WITH FINITE VOLUME SCHEMES.

MANUEL J. CASTRO*, SERGIO ORTEGA*,
MARC DE LA ASUNCIÓN†, JOSÉ M. MANTAS†

*Dpto. de Análisis Matemático, Universidad de Málaga. Campus de Teatinos s/n, 29071-Málaga, Spain.
†Dpto. Lenguajes y Sistemas Informáticos. Facultad de Informática. Universidad de Granada.

{castro,sergio}@anamat.cie.uma.es,marc@correo.ugr.es, jmmantas@ugr.es

Abstract

In this paper, we focus on the efficient implementation of path conservative Roe type high order finite volume schemes to simulate shallow flows. The motion of a layer of homogeneous non-viscous fluid is supposed to be governed by the shallow-water system, formulated under the form of a conservation law with source terms. The implementation of the scheme is carried out on Graphics Processing Units (GPUs), thus achieving a substantial improvement of the speedup with respect to normal CPUs. Finally, some numerical experiments are presented.

Key words: *GPUs, Finite volume methods, shallow-water, high-order schemes.*

AMS subject classifications: *65N06, 76B15, 76M20, 76N99*

1 Introduction

In this work, we show the benefits of using Graphics Processing Units (GPUs) to simulate shallow flows with finite volume schemes. The motion of a layer of homogeneous non-viscous fluid is supposed here to be governed by the shallow water system, formulated under the form of a conservation law with source terms or balance law:

This research has been partially supported by the Spanish Government Research projects MTM09-11923, TIN2007-29664-E, MTM2008-06349-C03-03, and P06-RNM-01594. The numerical computations have been performed at the Laboratory of Numerical Methods of the University of Málaga.

$$\begin{cases} \dfrac{\partial h}{\partial t} + \dfrac{\partial q_x}{\partial x} + \dfrac{\partial q_y}{\partial y} = 0, \\ \dfrac{\partial q_x}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{q_x^2}{h} + \dfrac{g}{2}h^2\right) + \dfrac{\partial}{\partial y}\left(\dfrac{q_x q_y}{h}\right) = gh\dfrac{\partial H}{\partial x}, \\ \dfrac{\partial q_y}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{q_x q_y}{h}\right) + \dfrac{\partial}{\partial y}\left(\dfrac{q_y^2}{h} + \dfrac{g}{2}h^2\right) = gh\dfrac{\partial H}{\partial y}. \end{cases} \tag{1}$$

Here $h(\mathbf{x}, t)$ denotes the thickness of the water layer, $q_\alpha(\mathbf{x}, t)$, $\alpha = x, y$, are the mass-flows in the coordinate directions, $H(\mathbf{x})$ represents the depth function (bathymetry) and g is the gravity constant.

Let us denote $U = [h,\, q_x,\, q_y]^T$ and

$$F_1(U) = \left[q_x, \frac{q_x^2}{h} + \frac{1}{2}gh^2, \frac{q_x q_y}{h}\right]^T, \quad F_2(U) = \left[q_y, \frac{q_x q_y}{h}, \frac{q_y^2}{h} + \frac{1}{2}gh^2\right]^T,$$
$$S_1(U) = [0, gh, 0]^T, \quad S_2(U) = [0, 0, gh]^T.$$

Let $J_i(U) = \dfrac{\partial F_i}{\partial U}(U)$ be the Jacobian of the flux F_i, for $i = 1, 2$. Given an unit vector $\boldsymbol{\eta} = (\eta_x, \eta_y) \in \mathbb{R}^2$, we define the matrix $A(U, \boldsymbol{\eta}) = J_1(U)\eta_x + J_2(U)\eta_y$, and the vectors $\boldsymbol{F_\eta}(U) = F_1(U)\eta_x + F_2(U)\eta_y$ and $\boldsymbol{S_\eta}(U) = \eta_x S_1(U) + \eta_y S_2(U)$.

Equation (1) can be rewritten as

$$W_t + \mathcal{A}_1(W)W_x + \mathcal{A}_2(W)W_y = 0, \tag{2}$$

by considering $W = [U, H]^T$ and

$$\mathcal{A}_i(W) = \begin{pmatrix} J_i(U) & -S_i(U) \\ 0 & 0 \end{pmatrix}, \; i = 1, 2.$$

The numerical solution of this model is useful for several applications related to geophysical flows: simulation of rivers, channels, dambreak problems, floods, etc. These simulations impose a great demand of computing power due to the dimensions of the domain (space and time). As a consequence, extremely efficient high performance solvers are required to solve and analyse these problems in reasonable execution times.

The nonconservative products involved in (2) do not make sense in general within the framework of distributions. Here, we follow the theory developed by Dal Maso, LeFloch and Murat in [4] to give a sense to these products as *Borel measures*. This theory is based on the choice of a family of paths.

In [1, 2] and [3] path conservative first and high order well-balanced Roe type schemes for solving balance laws and non-conservative systems were presented. An efficient parallel implementation of the numerical schemes for a PC cluster has been presented in [1]. This parallel implementation has been improved by using SSE-optimized software modules in order to accelerate small matrix computations at each processing node of the cluster (see [2]). Although these

improvements have made it possible to obtain results in lower computational times, the simulations still require too much runtime despite of using efficiently all the resources of a powerful PC cluster.

Currently, a cost effective emerging architecture exists which is specially indicated to accelerate considerably computationally intensive tasks like the one considered in this paper. Modern Graphics Processing Units (GPUs) are not only used to render 3D graphics but can also be a cost effective way to speedup the numerical solution of several mathematical models in Science and Engineering (see [16, 18, 17] for a revision of the topic). Modern GPUs offer over 100-200 processing units optimized for performing massively floating point operations in parallel [12]. As a consequence, for several algorithmic structures, these architectures are able to obtain a substantially higher performance than a powerful CPU.

In [6], an explicit central-upwind scheme is implemented on a NVIDIA GeForce 7800 GTX card to simulate the one-layer shallow-water system and a speedup from 15 to 30 is achieved with respect to an implementation on an Intel Xeon processor. In [8, 9], a first order path conservative Roe type solver has been implemented on several NVIDIA GeForce cards to simulate the one-layer shallow water system and a speedup of two orders of magnitude faster than a SSE-optimized CPU version of the solver for medium-size problems is achieved. In [5], a third order path conservative Roe type solver has also been implemented and a speedup of two orders is also achieved.

Here, we summarise the results obtained in [8, 9] and [5] and new results over non-structured triangular meshes are also presented.

The structure of the paper is as follows: in Section 3, the high-order finite volume scheme developed in [3] is reviewed. Next, Section 4 is dedicated to give the main ideas of the implementation of the numerical scheme in GPUs. Finally, some numerical experiments are presented and, some conclusions are drawn in Section 6.

2 Weak solutions

Let us consider the system (2) where $W(\mathbf{x},t)$ takes values on a convex domain Ω of $\mathbb{R}^N$ and $\mathcal{A}_i$, $i=1,2$ are two smooth and locally bounded matrix-valued functions from Ω to $\mathcal{M}_{N\times N}(\mathbb{R})$.

We assume that (2) is strictly hyperbolic, i.e. for all $W \in \Omega$ and $\forall\ \boldsymbol{\eta}=(\eta_x,\eta_y)\in\mathbb{R}^2$, the matrix $\mathcal{A}(W,\boldsymbol{\eta})$

$$\mathcal{A}(W,\boldsymbol{\eta}) = \mathcal{A}_1(W)\eta_x + \mathcal{A}_2(W)\eta_y$$

has N real and distinct eigenvalues

$$\lambda_1(W,\boldsymbol{\eta}) < \ldots < \lambda_N(W,\boldsymbol{\eta}).$$

$\mathcal{A}(W,\boldsymbol{\eta})$ is thus diagonalizable:

$$\mathcal{A}(W,\eta) = \mathcal{K}(W,\boldsymbol{\eta})\Lambda(W,\boldsymbol{\eta})\mathcal{K}^{-1}(W,\boldsymbol{\eta}),$$

being $\Lambda(W,\boldsymbol{\eta})$ the diagonal matrix whose coefficients are the eigenvalues of $\mathcal{A}(W,\boldsymbol{\eta})$ and $\mathcal{K}(W,\boldsymbol{\eta})$ is a matrix whose j-th column is an eigenvector $R_j(W,\boldsymbol{\eta})$ associated to the eigenvalue $\lambda_j(W,\boldsymbol{\eta})$, $j=1,\dots,N$.

For discontinuous solutions W, the nonconservative products $\mathcal{A}_1(W)W_x$ and $\mathcal{A}_2(W)W_y$ do not make sense as distributions. However, the theory developed by Dal Maso, LeFloch and Murat in [4] allows to give a rigorous definition of nonconservative products, associated to the choice of a family of paths in Ω.

Definition 1 *A* family of paths *in* $\Omega\subset\mathbb{R}^N$ *is a locally Lipschitz map*

$$\Psi\colon [0,1]\times\Omega\times\Omega\times\mathcal{S}^1\to\Omega,$$

where $\mathcal{S}^1\subset\mathbb{R}^2$ *denotes the unit sphere, that satisfies the following properties:*

1. $\Psi(0;W_L,W_R,\boldsymbol{\eta})=W_L$ *and* $\Psi(1;W_L,W_R,\boldsymbol{\eta})=W_R$*, for any* $W_L,W_R\in\Omega$*,* $\boldsymbol{\eta}\in\mathcal{S}^1$*.*

2. $\Psi(s;W_L,W_R,\boldsymbol{\eta})=\Psi(1-s;W_R,W_L,-\boldsymbol{\eta})$*, for any* $W_L,W_R\in\Omega$*,* $s\in[0,1]$*,* $\boldsymbol{\eta}\in\mathcal{S}^1$*.*

3. *Given an arbitrary bounded set* $\mathcal{B}\subset\Omega$*, there exists a constant* k *such that*

$$\left|\frac{\partial\Psi}{\partial s}(s;W_L,W_R,\boldsymbol{\eta})\right|\le k|W_L-W_R|,$$

   for any $W_L,W_R\in\mathcal{B}$*,* $s\in[0,1]$*,* $\boldsymbol{\eta}\in\mathcal{S}^1$*.*

4. *For every bounded set* $\mathcal{B}\subset\Omega$*, there exists a constant* K *such that*

$$\left|\frac{\partial\Psi}{\partial s}(s;W_L^1,W_R^1,\boldsymbol{\eta})-\frac{\partial\Psi}{\partial s}(s;W_L^2,W_R^2,\boldsymbol{\eta})\right|\le K(|W_L^1-W_L^2|+|W_R^1-W_R^2|),$$

   for each $W_L^1,W_R^1,W_L^2,W_R^2\in\mathcal{B}$*,* $s\in[0,1]$*,* $\boldsymbol{\eta}\in\mathcal{S}^1$*.*

Remark 1 *The dependency of the family of paths on* $\boldsymbol{\eta}$ *can be dropped for rotationally invariant systems. In fact , in [4] the families of path introduced to define the nonconservative products in the multidimensional case do not depend on* $\boldsymbol{\eta}$*.*

Suppose that a family of paths Ψ in Ω has been chosen. Then, the nonconservative products in (2) can be interpreted as a Borel measure and a rigorous definition of weak solution can be given (see [4] for details). According to this definition, a piecewise regular function W is a weak solution if and only if the two following conditions are satisfied:

(i) W is a classical solution where it is smooth.

(ii) At every point of a discontinuity W satisfies the jump condition

$$\int_0^1 \big(\sigma\mathcal{I} - \mathcal{A}(\Psi(s;W^-,W^+,\boldsymbol{\eta}),\boldsymbol{\eta})\big)\frac{\partial\Psi}{\partial s}(s;W^-,W^+,\boldsymbol{\eta})\,ds = 0, \quad (3)$$

where $\mathcal{I}$ is the identity matrix; σ, the speed of propagation of the discontinuity; $\boldsymbol{\eta}$ a unit vector normal to the discontinuity at the considered point; and W^-, W^+, the lateral limits of the solution at the discontinuity.

Together with the definition of weak solutions, a notion of entropy has to be chosen. We will assume here that the system can be endowed with an entropy pair $(\eta, \mathbf{G})$, i.e. a pair of regular functions $\eta : \Omega \to \mathbb{R}$ and $\mathbf{G} = (G_1, G_2) : \Omega \to \mathbb{R}^2$ such that:

$$\nabla G_i(W) = \nabla\eta(W)\cdot\mathcal{A}_i(W), \quad \forall\, W \in \Omega, \quad i = 1, 2.$$

Definition 2 *A weak solution is said to be an entropy solution if it satisfies the inequality*

$$\partial_t\eta(W) + \partial_{x_1}G_1(W) + \partial_{x_2}G_2(W) \le 0,$$

in the distributions sense.

The choice of the family of paths is important because it determines the speed of propagation of discontinuities. The simplest choice is given by the family of segments:

$$\Psi(s;W_L,W_R,\boldsymbol{\eta}) = W_L + s(W_R - W_L), \quad (4)$$

that corresponds to the definition of nonconservative products proposed by Volpert (see [20]). In practical applications, it has to be based on the physical background of the problem. In [10] a clear motivation for the selection of the family of paths is provided when a physical regularization by diffusion, dispersion, etc is available. Nevertheless, it is natural from the mathematical point of view to require this family to satisfy some hypotheses concerning the relation of the paths with the integral curves of the characteristic fields (see [3] for more details).

3 High-order finite volume schemes

3.1 Roe method

To discretize (2) the computational domain D is decomposed into subsets with a simple geometry, called cells or finite volumes: $V_i \subset \mathbb{R}^2$. It is assumed that the cells are closed convex polygons whose intersections are either empty, a complete edge or a vertex. Denote by $\mathcal{T}$ the mesh, i.e., the set of cells, and by NV the number of cells. Here, we consider rectangular structured meshes or triangular non-structured ones.

Given a finite volume V_i, $|V_i|$ will represent its area; $N_i \in \mathbb{R}^2$ its center; $\mathcal{N}_i$ the set of indexes j such that V_j is a neighbor of V_i; E_{ij} the common edge of

two neighboring cells V_i and V_j, and $|E_{ij}|$ its length; d_{ij} the distance from N_i to E_{ij}; $\boldsymbol{\eta}_{ij} = (\eta_{ij,1}, \eta_{ij,2})$ the normal unit vector at the edge E_{ij} pointing towards the cell V_j (see figure 1); Δ the maximum of the diameters of the cells; W_i^n the constant approximation to the average of the solution in the cell V_i at time t^n provided by the numerical scheme:

$$W_i^n \cong \frac{1}{|V_i|} \int_{V_i} W(\mathbf{x}, t^n)\, d\mathbf{x}.$$

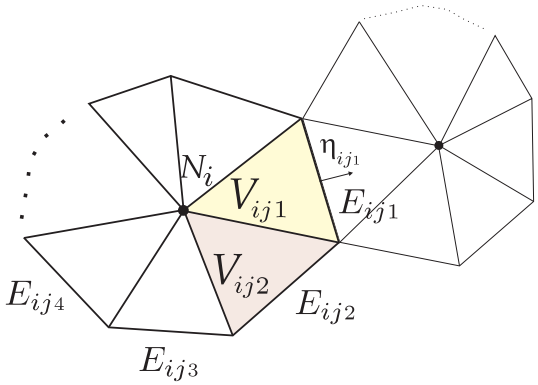


Figure 1: Finite volume discretization.

Given a family of paths Ψ, a *Roe linearization* of system (2) is a function

$$\mathcal{A}_\Psi : \Omega \times \Omega \times S^1 \to \mathcal{M}_N(\mathbb{R})$$

satisfying the following properties for each $W_L, W_R \in \Omega$ and $\boldsymbol{\eta} \in S^1$:

1. $\mathcal{A}_\Psi(W_L, W_R, \boldsymbol{\eta})$ has N distinct real eigenvalues

$$\lambda_1(W_L, W_R, \boldsymbol{\eta}) < \lambda_2(W_L, W_R, \boldsymbol{\eta}) < \cdots < \lambda_N(W_L, W_R, \boldsymbol{\eta}).$$

2. $\mathcal{A}_\Psi(W, W, \boldsymbol{\eta}) = \mathcal{A}(W, \boldsymbol{\eta})$.

3. $\mathcal{A}_\Psi(W_L, W_R, \boldsymbol{\eta}) \cdot (W_R - W_L) =$

$$\int_0^1 \mathcal{A}(\Psi(s; W_L, W_R, \boldsymbol{\eta}), \boldsymbol{\eta}) \frac{\partial \Psi}{\partial s}(s; W_L, W_R, \boldsymbol{\eta})\, ds. \tag{5}$$

We denote by $\Lambda_\Psi(W_L, W_R, \boldsymbol{\eta})$ the diagonal matrix whose coefficients are the eigenvalues $\lambda_j(W_L, W_R, \boldsymbol{\eta})$ and let $\mathcal{K}_\Psi(W_L, W_R, \boldsymbol{\eta})$ be the associated eigenvectors matrix. Let us define the positive and negative parts of $\mathcal{A}_\Psi(W_L, W_R, \boldsymbol{\eta})$ as

$$\mathcal{A}_\Psi^\pm(W_L, W_R, \boldsymbol{\eta}) = \mathcal{K}_\Psi(W_L, W_R, \boldsymbol{\eta}) \cdot \Lambda_\Psi^\pm(W_L, W_R, \boldsymbol{\eta}) \cdot \mathcal{K}_\Psi(W_L, W_R, \boldsymbol{\eta})^{-1},$$

where $\Lambda_\Psi^+(W_L, W_R, \boldsymbol{\eta})$ (respectively, $\Lambda_\Psi^-(W_L, W_R, \boldsymbol{\eta})$) is the diagonal matrix whose coefficients are the positive (respectively, negative) parts of the eigenvalues $\lambda_j(W_L, W_R, \boldsymbol{\eta})$.

In the particular case in which $\mathcal{A}_k(W)$, $k = 1, 2$, is the Jacobian matrix of a smooth flux function $F_k(W)$, property (5) does not depend on the family of paths and reduces to the usual Roe property:

$$\mathcal{A}_\Psi(W_L, W_R, \boldsymbol{\eta}) \cdot (W_R - W_L) = F_{\boldsymbol{\eta}}(W_R) - F_{\boldsymbol{\eta}}(W_L) \tag{6}$$

for any $\boldsymbol{\eta} \in S^1$.

The general expression of a Roe scheme in upwind form for solving (2) is given by ([3]):

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{|V_i|} \sum_{j \in \mathcal{N}_i} |E_{ij}| \mathcal{A}_{ij}^- \cdot (W_j^n - W_i^n) \tag{7}$$

where

$$\mathcal{A}_{ij}^- = \mathcal{A}_\Psi^-(W_i^n, W_j^n, \boldsymbol{\eta}_{ij}).$$

Additionally, a CFL condition must be imposed to ensure stability:

$$\Delta t \cdot \max \left\{ \frac{|\lambda_{ij,k}|}{d_{ij}};\ i = 1, \ldots, NV,\ j \in \mathcal{N}_i,\ \ k = 1, \ldots, N \right\} = \delta, \tag{8}$$

with $0 < \delta \leq 1$.

As in the case of systems of conservation laws, when sonic rarefaction waves appear it is necessary to modify the numerical scheme in order to obtain entropy-satisfying solutions. For instance, the Harten-Hyman entropy fix technique ([7]) can be easily adapted here.

Some general results concerning the consistency and well-balanced properties of Roe schemes have been studied in [3].

3.2 High-order extension

In this section we describe a high-order extension of scheme (7). Let us consider first a *reconstruction operator*, i.e., an operator that associates to a given family $\{W_i\}_{i=1}^{NV}$ of cell values two families of functions defined at the edges:

$$\gamma \in E_{ij} \mapsto W_{ij}^{\pm}(\gamma)$$

in such a way that, whenever

$$W_i = \frac{1}{|V_i|} \int_{V_i} W(\mathbf{x})\, d\mathbf{x} \tag{9}$$

for some smooth function W, then

$$W_{ij}^{\pm}(\gamma) = W(\gamma) + \mathcal{O}(\Delta^p), \quad \gamma \in E_{ij}.$$

It will be assumed that the reconstructions are calculated in the following way: given the family $\{W_i\}_{i=1}^{NV}$ of cell values, an approximation function is constructed at every cell V_i, based on the values W_j at some stencil of neighboring cells to V_i:

$$P_i(\mathbf{x}) = P_i\left(\mathbf{x};\ \{W_j\}_{j\in\mathcal{B}_i}\right)$$

for some set of indexes $\mathcal{B}_i$. These approximation functions are usually constructed by means of interpolation or approximation methods. The reconstructions at $\gamma \in E_{ij}$ are defined as

$$W_{ij}^-(\gamma) = \lim_{\mathbf{x}\to\gamma} P_i(\mathbf{x}), \quad W_{ij}^+(\gamma) = \lim_{\mathbf{x}\to\gamma} P_j(\mathbf{x}). \tag{10}$$

Clearly, for any $\gamma \in E_{ij}$ the following equalities are satisfied:

$$W_{ij}^-(\gamma) = W_{ji}^+(\gamma), \quad W_{ij}^+(\gamma) = W_{ji}^-(\gamma).$$

The reconstruction operator is assumed to satisfy the following properties:

(H1) It is conservative, i.e., the following equality holds for any cell V_i:

$$W_i = \frac{1}{|V_i|}\int_{V_i} P_i(\mathbf{x})d\mathbf{x}. \tag{11}$$

(H2) It is of order p, in the sense that

$$W(\gamma) - W_{ij}^{\pm}(\gamma) = \Delta^p g_{ij}(\gamma) + \mathcal{O}(\Delta^{p+1}), \quad \gamma \in E_{ij}$$

being g_{ij} a regular function.

(H3) It is of order q in the interior of the cells, i.e., if the operator is applied to a sequence $\{W_i\}$ satisfying (9) for some smooth function $W(\mathbf{x})$, then

$$P_i(\mathbf{x}) = W(\mathbf{x}) + \mathcal{O}(\Delta^q),\ \mathbf{x} \in \mathrm{int}(V_i). \tag{12}$$

(H4) The gradient of P_i provides an approximation of order m to the gradient of W:

$$\nabla P_i(\mathbf{x}) = \nabla W(\mathbf{x}) + \mathcal{O}(\Delta^m),\ \mathbf{x} \in \mathrm{int}(V_i). \tag{13}$$

The semidiscrete expression of the high-order extension of scheme (7), based on a given reconstruction operator, is the following (see [3] for more details):

$$\begin{aligned} W_i'(t) = -\frac{1}{|V_i|}\Bigg[&\sum_{j\in\mathcal{N}_i}\int_{E_{ij}} \mathcal{A}_{ij}^-(\gamma,t)\big(W_{ij}^+(\gamma,t) - W_{ij}^-(\gamma,t)\big)d\gamma \\ &+ \int_{V_i}\left(\mathcal{A}_1(P_i^t(\mathbf{x}))\frac{\partial P_i^t}{\partial x}(\mathbf{x}) + \mathcal{A}_2(P_i^t(\mathbf{x}))\frac{\partial P_i^t}{\partial y}(\mathbf{x})\right)d\mathbf{x}\Bigg] \end{aligned} \tag{14}$$

where P_i^t is the approximation function at time t:

$$P_i^t(\mathbf{x}) = P_i\left(\mathbf{x};\ \{W_j(t)\}_{j\in\mathcal{B}_i}\right),$$

$W_{ij}^{\pm}(\gamma,t)$ are given by

$$W_{ij}^{-}(\gamma,t) = \lim_{\mathbf{x}\to\gamma} P_i^t(\mathbf{x}), \quad W_{ij}^{+}(\gamma,t) = \lim_{\mathbf{x}\to\gamma} P_j^t(\mathbf{x}), \tag{15}$$

and

$$\mathcal{A}_{ij}(\gamma,t) = \mathcal{A}_\Psi\left(W_{ij}^{-}(\gamma,t), W_{ij}^{+}(\gamma,t), \boldsymbol{\eta}_{ij}\right).$$

The following result can be proved (see [3]):

Theorem 1 *Assume that $\mathcal{A}_1$ and $\mathcal{A}_2$ are of class $\mathcal{C}^2$ with bounded derivatives and $\mathcal{A}_\Psi$ is bounded. Suppose also that the reconstruction operator satisfies hypotheses (H1)–(H4). Then* (14) *is an approximation of order at least $\alpha = \min(p,q,m)$.*

Remark 2 *The conclusion of theorem 1 is rather pessimistic: the observed order in experiments is usually $\alpha = \min(p,q,m+1)$. See [3] for more details.*

In practice, the integral terms in (14) must be approximated numerically. A one-dimensional quadrature formula of order $\bar{r}$ is applied to calculate the line integrals:

$$\int_a^b f(s)ds = (b-a)\left(\sum_{l=1}^{n(\bar{r})} \omega_l f(x_l)\right) + \mathcal{O}(\Delta^{\bar{r}}), \tag{16}$$

while a two-dimensional quadrature formula of order $\bar{s}$ is used to compute the volume integrals:

$$\int_{V_i} f(\mathbf{x})\, d\mathbf{x} = |V_i| \sum_{l=1}^{n(\bar{s})} \alpha_l f(\mathbf{x}_l^i) + \mathcal{O}(|V_i|^{\bar{s}}). \tag{17}$$

To preserve the order of the numerical scheme, it is necessary to have $\bar{r} \geq \alpha$ and $\bar{s} \geq \alpha$.

Finally, the numerical scheme is written as follows:

$$\begin{aligned} W_i'(t) = -\frac{1}{|V_i|}\Bigg[\sum_{j\in\mathcal{N}_i} |E_{ij}| \sum_{l=1}^{n(\bar{r})} w_l \mathcal{A}_{ij,l}^{-}(t)\big(W_{ij,l}^{+}(t) - W_{ij,l}^{-}(t)\big) \\ + |V_i| \sum_{l=1}^{n(\bar{s})} \alpha_l \left(\mathcal{A}_1(P_i^t(\mathbf{x}_l^i)) \frac{\partial P_i^t}{\partial x}(\mathbf{x}_l^i) + \mathcal{A}_2(P_i^t(\mathbf{x}_l^i)) \frac{\partial P_i^t}{\partial y}(\mathbf{x}_l^i) \right) \Bigg] \end{aligned} \tag{18}$$

where

$$W_{ij,l}^{\pm}(t) = W_{ij}^{\pm}(a_{ij} + s_l(b_{ij} - a_{ij}), t)$$

and

$$\mathcal{A}_{ij,l}(t) = \mathcal{A}_\Psi(W_{ij,l}^{-}(t), W_{ij,l}^{+}(t), \boldsymbol{\eta}_{ij}),$$

being a_{ij} and b_{ij} the vertices of edge E_{ij}.

Remark 3 *A technique that avoids the explicit computation of* $\nabla P_i(\mathbf{x})$ *has been introduced in [11] in the one-dimensional case. The use of this technique, that is based on the trapezoidal rule and Romberg extrapolation, makes the expected order of accuracy to be* $\min(p, q)$*. The extension to two-dimensional problems is straightforward for structured meshes, while for unstructured meshes a Romberg extrapolation formula for triangles can be used (see [21]).*

For time stepping, high-order TVD Runge-Kutta methods like those described in [19] are applied. In particular, in this work we use a third-order reconstruction operator in space and a third-order TVD Runge-Kutta method to advance in time.

The reconstruction operator used here is the one proposed in [5]: it is a compact reconstruction operator of polynomial type, that is third-order accurate on each computational cell and it can be defined in general non-uniform quadrilateral meshes.

The well-balancedness properties of schemes (14) and (18) have been analyzed in [3].

In the particular case of the one-layer shallow water system, the numerical scheme (18) reads as follows (see [3] for more details):

$$U_i'(t) = -\frac{1}{|V_i|}\left[\sum_{j\in\mathcal{N}_i}|E_{ij}|\sum_{l=1}^{n(\bar r)} w_l \mathcal{D}^-(U^-_{ij,l}(\gamma,t), U^+_{ij,l}(\gamma,t), H^-_{ij,l}(\gamma), H^+_{ij,l}(\gamma), \boldsymbol{\eta}_{ij}) \right.$$
$$\left. - |V_i|\sum_{l=1}^{n(\bar s)}\alpha_l\left(S_1(P_i^t(\mathbf{x}_l^i))\frac{\partial P_i^H}{\partial x}(\mathbf{x}_l^i) + S_2(P_i^t(\mathbf{x}_l^i))\frac{\partial P_i^H}{\partial y}(\mathbf{x}_l^i)\right)\right], \quad (19)$$

where de following notation is used: P_i^t is the reconstruction function corresponding to the cell value $U_i(t)$, while P_i^H is the reconstruction function associated to the cell averages of the given bathymetry. $U_{ij}^{\pm}(\gamma,t)$ and $H_{ij}^{\pm}(\gamma)$ are given, respectively, by

$$U_{ij}^-(\gamma,t) = \lim_{\mathbf{x}\to\gamma} P_i^t(\mathbf{x}), \quad U_{ij}^+(\gamma,t) = \lim_{\mathbf{x}\to\gamma} P_j^t(\mathbf{x})$$

and

$$H_{ij}^-(\gamma) = \lim_{\mathbf{x}\to\gamma} P_i^H(\mathbf{x}), \quad H_{ij}^+(\gamma) = \lim_{\mathbf{x}\to\gamma} P_j^H(\mathbf{x}).$$

$U_{ij,l}^{\pm}$ (respectively $H_{ij,l}^{\pm}$) corresponds to $U_{ij}^{\pm}(\cdot)$ (respectively $H_{ij}^{\pm}(\cdot)$) evaluated at the quadrature points of the edge E_{ij}. Moreover,

$$\mathcal{D}^-(U_L, U_R, H_L, H_R, \boldsymbol{\eta}) = \boldsymbol{F}_{\boldsymbol{\eta}}(U_L) + P_{LR}^-\left(A_{LR}(U_R - U_L) - \boldsymbol{S}_{LR}(H_R - H_L)\right). \quad (20)$$

In the particular case of system (1),

$$A_{LR} = \begin{bmatrix} 0 & \eta_x & \eta_y \\ (-\bar u_x^2 + \bar c^2)\eta_x - \bar u_x\bar u_y\eta_y & 2\bar u_x\eta_x + \bar u_y\eta_y & \bar u_x\eta_y \\ -\bar u_x\bar u_y\eta_x + (-\bar u_y^2 + \bar c^2)\eta_y & \bar u_y\eta_x & \bar u_x\eta_x + 2\bar u_y\eta_y \end{bmatrix}$$

and

$$\boldsymbol{S}_{LR} = \left[0,\, \bar{c}^2\eta_x,\, \bar{c}^2\eta_y\right]^T$$

with

$$\bar{h} = \frac{h_L + h_R}{2}, \quad \bar{c} = \sqrt{g\bar{h}}, \quad \bar{u}_\alpha = \frac{\sqrt{h_L}u_{L,\alpha} + \sqrt{h_R}u_{R,\alpha}}{\sqrt{h_L} + \sqrt{h_R}}, \quad \alpha = x, y, \tag{21}$$

and the usual definitions of the velocity $u = q/h$. Finally,

$$P^{\pm}_{LR} = \frac{1}{2}\mathcal{K}_{LR}(I \pm \mathrm{sgn}(\Lambda_{LR}))\mathcal{K}^{-1}_{LR}, \tag{22}$$

where I is the identity matrix, Λ_{LR} is the diagonal matrix whose coefficients are the eigenvalues of A_{LR}, $\mathcal{K}_{LR}$ is a matrix whose columns are associated eigenvectors, and $\mathrm{sgn}(\Lambda_{LR})$ is the diagonal matrix whose coefficients are the signs of the eigenvalues of the matrix A_{LR}.

4 CUDA Implementation

In this section we briefly describe the potential data parallelism of the numerical scheme described in the previous section and its implementation in CUDA.

Initially, the finite volume mesh must be constructed from the input data with the appropriate setting of initial and boundary conditions. Then the time stepping is performed by applying a third-order Runge-Kutta TVD method, consisting on three steps. At each step, the spatial discretization (19) must be performed as follows:

1. **Reconstruction and volume integral computation:** First a reconstruction procedure at each cell and for each variable must be performed to define the functions $P_i(\mathbf{x})$. Next, the numerical approximation of the volume integral is computed using a third-order Gaussian quadrature formula

$$\Sigma_i = -|V_i| \sum_{l=1}^{n(\bar{s})} \alpha_l \left(S_1(P_i(\mathbf{x}^i_l)) \frac{\partial P^H_i}{\partial x}(\mathbf{x}^i_l) + S_2(P_i(\mathbf{x}^i_l)) \frac{\partial P^H_i}{\partial y}(\mathbf{x}^i_l) \right).$$

   The reconstructed values $U_{ij,l}$ at the quadrature points of each edge of the cell V_i are also computed. Again, a third-order Gaussian quadrature formula is used. Therefore, two values must be computed at each edge of V_i.

2. **Edge-based calculations:** The following computations must be performed at each edge E_{ij} common to cells V_i and V_j, using the reconstructed values $U^-_{ij,l}$ and $U^+_{ij,l}$ previously computed:

$$\Sigma^{\pm}_{ij} = |E_{ij}| \sum_{l=1}^{n(\bar{r})} w_l \mathcal{D}^{\pm}(U^-_{ij,l}(\gamma,t), U^+_{ij,l}(\gamma,t), H^-_{ij,l}(\gamma), H^+_{ij,l}(\gamma), \boldsymbol{\eta}_{ij}).$$

3. **Volume-based calculations:** At each cell V_i, the following computations must be performed:

   a) Computation of the local Δt_i for each volume.

   b) Computation of $U_i^{n+1,s}$: The $n+1, s$-th state of each volume must be approximated from the n-th and the $n+1, s-1$-th states using the data computed at the previous steps.

Several remarks can be made related to the description of the parallel algorithm. The computation steps required by the problem addressed here can be classified into two groups: computations associated to edges and computations associated to volumes. The scheme exhibits a high degree of data parallelism because the computation at each edge/volume is independent with respect to the computation performed at the rest of edges/volumes. Moreover, the scheme presents a high arithmetic intensity and the computation exhibits a high degree of locality. These remarks indicate that this problem is suitable for being implemented on GPUs using CUDA.

Concerning the implementation, each processing step previously described is assigned to a CUDA kernel. A kernel is a function executed on the GPU, which is executed forming a grid of thread blocks that run logically in parallel (see [13] for more details). Let us describe the implementation of a high order scheme on structured meshes using CUDA. Non-structured meshes need a more sophisticated data structure to be used on GPUs. More details can be found in [9] and [5].

- **Build the data structure:** For each volume, we store its state (h, q_x and q_y) and its depth H. We define an array of `float4` elements, where each element represents a volume and contains the former parameters. This array is stored as a 2D texture since texture memory is especially suited for each thread to access its closer environment in texture memory. The per-block shared memory, on the other hand, is more suitable when each thread needs to access many elements located in global memory, and each thread of a block loads a small part of these elements into shared memory. We first implemented a CUDA program using shared memory instead of a texture, where each thread of a block loaded the data of a volume into shared memory, but later we got better execution times by using a texture.

  The area of the volumes and the length of the vertical and horizontal edges are precalculated and passed to the CUDA kernels that need them.

  We can know at runtime if an edge or volume is frontier or not and the value of $\boldsymbol{\eta}_{ij}$ at an edge by checking the position of the thread in the grid.

- **Reconstruction and integral computation:** In this step, the reconstruction values $U_{ij,l}^{\pm}$, $l = 1, 2$, are computed and stored in four arrays located in global memory, each one being an array of `float3` elements. The size of each array is twice the number of volumes and

they are associated to the four edges of a cell (south, north, east and west). Moreover, the integral term Σ_i is also computed and stored in an accumulator placed in global memory. This accumulator is an array of `float4` elements and its size is the number of volumes. This accumulator is also used to store the contributions of the vertical edges. In this process, each thread represents a finite volume cell.

- **Process vertical and horizontal edges:** We divide the edge processing into vertical and horizontal edge processing. For vertical edges $\eta_{ij,y} = 0$, and for horizontal edges $\eta_{ij,x} = 0$. Therefore, all the operations where these terms take part can be avoided, thus increasing efficiency.

  Here, each thread represents a vertical or a horizontal edge, and computes the contribution to their adjacent volumes.

  The edges (i.e., threads) synchronize each other when contributing to a particular volume by means of two accumulators stored in global memory, each one being an array of `float4` elements. Note that one of them has been previously used to store the integral cell computation. The size of each accumulator is the number of volumes. Each element of the accumulators stores the edge contributions to the volume (a 3×1 vector, $\Sigma_{ij}^{\pm}$, and a `float` value storing $\|\Lambda_{ij}\|_\infty$). In the processing of vertical edges, each edge writes the contribution to its right-side volume in the first accumulator, and the contribution to its left-side volume in the second accumulator. Next, the processing of horizontal edges is performed in an analogous way, with the difference that the contribution is added to the accumulators.

- **Compute Δt_i for each volume:** In this step, each thread represents a volume and the local Δt_i of the volume V_i is computed using the CFL condition (8).

- **Get the minimum Δt:** This step finds the minimum of the local Δt_i of the volumes by applying a reduction algorithm on the GPU. The reduction algorithm applied is the kernel 7 (the most optimized one) of the reduction sample included in the CUDA Software Development Kit [13].

- **Compute $U_i^{n+1,s}$ for each volume:** In this step, each thread represents a volume and the state U_i of the volume V_i is updated. The final value is obtained by adding up the two 3×1 vectors stored in the positions corresponding to the volume V_i in both accumulators. Since a CUDA kernel cannot directly write into textures, the texture is initially updated by writing the results into a temporal array, which is then copied to the CUDA array bound to the texture.

5 Numerical experiments

Different implementations of the scheme have been performed: a sequential CPU code was written in C++ using double precision, a quadcore CPU code

Table 1: Structured meshes: Execution times in seconds for all meshes and programs (first order)

Volumes	CPU 1 core	CPU 4 cores	GTX 260 Cg	GTX 260 CUSP	GTX 260 CUDP	GTX 280 Cg	GTX 280 CUSP	GTX 280 CUDP
100×100	0.8	0.26	0.11	0.01	0.06	0.08	0.01	0.05
200×200	6.7	1.98	0.26	0.06	0.37	0.2	0.06	0.32
400×400	56.6	26.57	0.84	0.39	2.75	0.68	0.35	2.34
800×800	455.9	216.5	4.42	2.91	21.43	3.75	2.48	18.49
1600×1600	3639.9	1722.8	30.72	23.44	167.2	26.14	19.27	143.0
2000×2000	7135.7	3375.4	58.54	44.87	336.9	49.48	38.34	272.0

Table 2: Non-structured meshes: Execution times in seconds for all the meshes and programs (first order)

Volumes	CPU 1 core	CPU 4 cores	GTX 260 CUSP	GTX 260 CUDP	GTX 280 CUSP	GTX 280 CUDP
4016	0.26	0.10	0.012	0.043	0.011	0.041
16040	2.30	0.65	0.040	0.23	0.038	0.22
64052	21.19	7.38	0.24	1.66	0.21	1.56
256576	178.0	63.34	1.77	12.67	1.57	11.96
1001898	1442.0	518.4	13.46	97.91	12.14	92.22

using OPENMP ([15]), a GPU code implemented using CG and single precision, a GPU code implemented in CUDA using single precision (CUSP), and a GPU code implemented in CUDA using double precision (CUDP). The CPU was an Intel Xeon E5430 (2.66 GHz 12MB L2 Cache), while two different GPUs have been used: a NVIDIA GeForce GTX260 (192 stream processors with 869Mb) and a NVIDIA GeForce GTX280 (240 stream processors with 1Gb).

As test problem, we consider a circular dambreak problem in the $[-5,5] \times [-5,5]$ domain. The depth function is $H(x,y) = 1 - 0.4\, e^{-x^2-y^2}$ and the initial condition is:

$$W_i^0(x,y) = \begin{pmatrix} h^0(x,y) \\ 0 \\ 0 \end{pmatrix}, \text{ where } h^0(x,y) = \begin{cases} 1 + H(x,y) & \text{if } \sqrt{x^2+y^2} > 0.6 \\ 3 + H(x,y) & \text{otherwise} \end{cases}$$

The numerical schemes are run for different mesh sizes. Simulations are carried out in the time interval [0,1]. CFL parameter is $\delta = 0.9$ and wall boundary conditions ($\mathbf{q} \cdot \boldsymbol{\eta} = 0$) are considered.

Tables 1, 2 and 3 show the execution times in seconds for all the meshes and programs. As can be seen, the execution times seems to grow linearly with the number of volumes of the mesh, as expected. Figures 2, 3 and 4 show graphically the speedup obtained in all the implementations with respect to the monocore version.

Concerning the first order numerical scheme on structured meshes (see Table

Table 3: Structured meshes: Execution times in seconds for all the meshes and programs (third order)

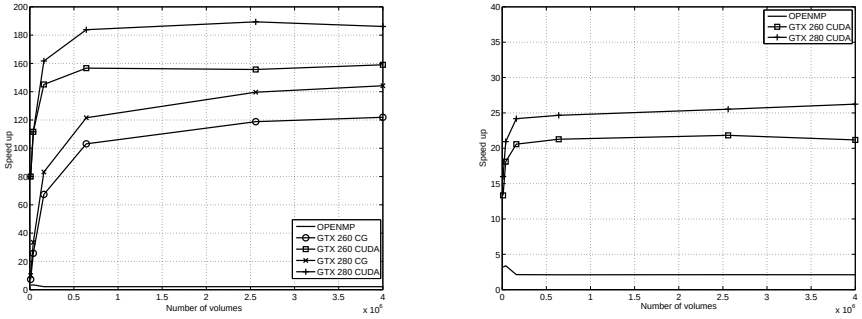
Volumes	CPU 1 core	GTX260 CUSP	GTX 280 CUSP
100×100	2.36	0.025	0.024
200×200	19.0	0.18	0.15
400×400	152.10	1.36	1.31
800×800	1218.32	10.89	9.47
1200×1200	4068.0	36.16	30.01

1 and Figure 2) we can see that the execution times of the single precision CUDA program (CUSP) outperform that of Cg in all cases with both graphics cards. Using a GTX 280, for big problems, CUSP achieves a speedup of two orders of magnitude with respect to the monocore version, reaching a performance gain of more than 180 (see Figure 2(a)). The double precision CUDA program (CUDP) has been about 7 times slower than CUSP for big problems in both graphics cards (see 2(b)), which seems logical considering that, in GT200 architecture, each multiprocessor has 8 single precision units and only one double precision unit. As expected, the OpenMP version only reaches a speedup less than four with respect to the monocore program in all meshes (see Figure 2).

The results for triangular non-structured meshes (see Table 2 and Figure 3) are similar to those obtained for structured meshes: in this case the achieved speedup for single precision is about 120, being the double precision CUDA program about 7 times slower than CUSP. Note that the performance of the CUDA codes has decreased. The reason for this decreasing is that the data structure needed to manage non-structured meshes in GPU is more complex than the one needed for structured meshes, where 2D *textures* can be used.

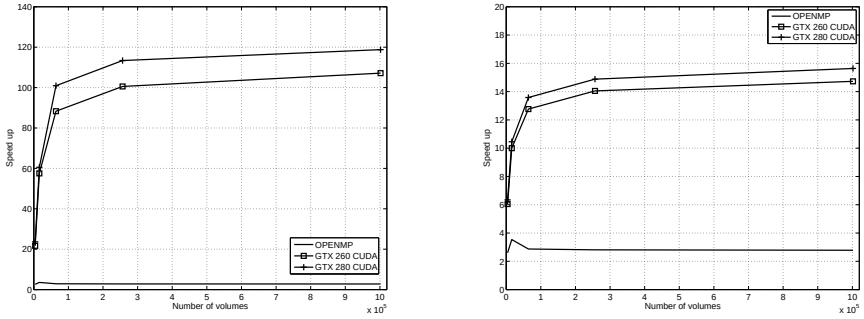
Similar results are obtained for the high order numerical scheme on structured meshes (see Table 3 and Figure 4), reaching a performance of about 140 with respect to the monocore version.

We also have compared the numerical solutions obtained in the monocore and the CUDA programs. The L1 norm of the difference between the solutions obtained in CPU and GPU at time $t = 1.0$ for all meshes was calculated. The order of magnitude of the L1 norm using CUSP vary between 10^{-4} and 10^{-6}, while that of obtained using CUDP vary between 10^{-12} and 10^{-14}, which reflects the different accuracy of the numerical solutions computed on the GPU using single and double precision.



(a) Single precision (b) Double precision

Figure 2: First order scheme: speedup on structured meshes. Single precision (left). Double precision (right).



(a) Single precision (b) double precision

Figure 3: First order scheme: speedup on non-structured meshes. Single precision (left). Double precision (right).

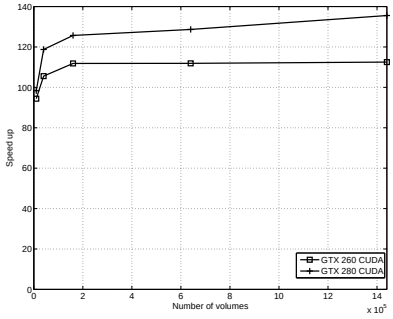


Figure 4: High order scheme: speedup on non-structured meshes (single precision).

6 Conclusions

Different implementations of a path conservative first and third order Roe type well-balanced finite volume scheme for the one-layer shallow-water system scheme have been performed. Optimization techniques to parallelize efficiently the numerical schemes on CUDA architecture have been considered. Simulations carried out on a GeForce GTX 280 card using single precision were found to be up to two orders of magnitude faster than a monocore version of the solver for big-size uniform problems, one order of magnitude faster than a quadcore implementation based on OpenMP, and also faster than a GPU version based on a graphics-specific language (Cg). The double precision version of the CUDA solver has been 7 times slower than the single precision version for big meshes. In any case, this factor of 7 will be dramatically reduced in the next generation of NVIDIA graphics cards (FERMI) where the number of double precision units will be increased. These simulations also show that the numerical solutions obtained with the solver are accurate enough for practical applications, obtaining better accuracy using double precision than using single precision. As further work, we propose to extend the strategy to enable efficient high order simulations on non-structured meshes and to extend the CUDA implementation to other models like two-layer shallow water systems.

References

[1] M.J. Castro, J.A. García, J.M. González and C. Parés (2006). A parallel 2d finite volume scheme for solving systems of balance laws with non-conservative products: application to shallow flows. Comp. Meth. Appl. Mech. Eng. 196, 2788-2815.

[2] M.J. Castro, J.A. García, J.M. González and C. Parés (2008). Solving shallow-water systems in 2D domains using finite volume methods and multimedia SSE instructions. J. Comput. App. Math., 221: 16-32.

[3] M.J. Castro, E.D. Fernández, A.M. Ferreiro, A. García, C. Parés (2009). High order extension of Roe schemes for two dimensional nonconservative hyperbolic systems. J. Sci. Comput. 39, 67–114.

[4] G. Dal Maso, P.G. LeFloch and F. Murat (1995). Definition and weak stability of nonconservative products. J. Math. Pures Appl. 74:483–548.

[5] J.M. Gallardo, S. Ortega, M. de la Asunción, J.M. Mantas (2010). Two-dimensional compact third-order polynomial reconstructions. Solving nonconservative hyperbolic systems using GPUs. Submitted to J. Sci. Comput.

[6] T.R. Hagen, J.M. Hjelmervik, K.A. Lie, J.R. Natvig, M. Ofstad (2005). Visual simulation of shallow-water waves. Sim. Modelling Pract. and Th. 13, 716–726.

[7] A. Harten, J.M. Hyman (1983). Self-adjusting grid methods for one-dimensional hyperbolic conservation laws. J. Comp. Phys. 50, 235–269.

[8] M. Lastra, J.M. Mantas, C. Ureña, M.J. Castro, J.A. García (2009). Simulation of shallow-water systems using graphics processing units. Math. Comp. Simul. 80, 598–618.

[9] M. de la Asunción, J.M. Mantas, M.J. Castro (2009). Simulation of one-layer shallow water systems on multicore and CUDA architectures. Accepted in J. Supercomputing.

[10] P.G.LeFloch. Shock waves for nonlinear hyperbolic systems in nonconservative form (1989). Institute for Math. and its Appl., Minneapolis, Preprint 593.

[11] S. Noelle, N. Pankratz, G. Puppo, J. Natvig (2006). Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. J. Comput. Phys. 213, 474–499.

[12] http://www.nvidia.com,

[13] NVIDIA. CUDA Zone. http://www.nvidia.com/object/cuda_home.html. Accessed November 2009.

[14] C. Parés (2006). Numerical methods for nonconservative hyperbolic systems: a theoretical framework. SIAM J. Num. Anal. 44, 300– 321.

[15] Chapman B, Jost G, van der Pas R (2007). Using OpenMP: Portable Shared Memory Parallel Programming, The MIT Press.

[16] J.D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krüger, A. E. Lefohn, T. Purcell (2005). A Survey of General-Purpose Computation on Graphics Hardware, Eurographics 2005 State of the Art Report.

[17] J.D. Owens, M. Houston, D.Luebke,S. Green, J.E. Stone,J.C Phillips (2008). GPU Computing. Proceedings of the IEEE, 96(5):879-899.

[18] M. Rumpf, R. Strzodka (2006). Graphics Processor Units: New Prospects for Parallel Computing, L. N. in Computational Science and Engineering 51, 89–121.

[19] C.-W. Shu, S. Osher (1998). Efficient implementation of essentially non-oscillatory shock capturing schemes. J. Comput. Phys. 77, 439–71.

[20] A.I.Volpert (1967). Spaces BV and quasilinear equations, Math. USSR Sbornik, 73 (1967): 255-302.

[21] G. Walz (1997). Romberg Type Cubature over Arbitrary Triangles. Mannheimer Mathem. Manuskripte Nr.225, Mannhein.

Sesiones especiales

INTEGRACIÓN GEOMÉTRICA

Sergio Blanes

David Martín de Diego

Juan Ignacio Montijano

RECENT ADVANCES IN ADAPTIVE FINITE ELEMENTS

Roland Becker

Stefano Micheletti

Bol. Soc. Esp. Mat. Apl.
n°50(2010), 47–61

SPLITTING METHODS WITH COMPLEX COEFFICIENTS

SERGIO BLANES*, FERNANDO CASAS†, ANDER MURUA‡

*Instituto de Matemática Multidisciplinar, Universidad Politécnica de Valencia, E-46022 Valencia, Spain.
†Institut de Matemàtiques i Aplicacions de Castelló and Departament de Matemàtiques, Universitat Jaume I, E-12071 Castellón, Spain.
‡Konputazio Zientziak eta A.A. saila, Informatika Fakultatea, EHU/UPV, Donostia/San Sebastián, Spain.

serblaza@imm.upv.es, Fernando.Casas@uji.es, Ander.Murua@ehu.es

Abstract

Splitting methods for the numerical integration of differential equations of order greater than two involve necessarily negative coefficients. This order barrier can be overcome by considering complex coefficients with positive real part. In this work we review the composition technique used to construct methods of this class, propose new sixth-order integrators and analyze their main features on a pair of numerical examples, in particular how the errors are propagated along the evolution.

1 Introduction

Splitting methods for the numerical integration of differential equations constitute an appropriate choice when the associated vector field can be decomposed into several pieces and each of them is explicitly integrable.

Given the initial value problem

$$x' = f(x), \qquad x_0 = x(0) \in \mathbb{R}^D \tag{1}$$

with $f : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ and solution $\varphi_t(x_0)$, let us suppose that f can be expressed as $f = \sum_{i=1}^m f^{[i]}$ for certain functions $f^{[i]} : \mathbb{R}^D \longrightarrow \mathbb{R}^D$, in such a way that the equations

$$x' = f^{[i]}(x), \qquad x_0 = x(0) \in \mathbb{R}^D, \qquad i = 1, \ldots, m \tag{2}$$

can be integrated exactly, with solutions $x(h) = \varphi_h^{[i]}(x_0)$ at $t = h$, the time step. Splitting methods intend to approximate the exact flow φ_h by a composition of flows $\varphi_h^{[i]}$. For instance,

$$\chi_h = \varphi_h^{[m]} \circ \cdots \circ \varphi_h^{[2]} \circ \varphi_h^{[1]}, \qquad \chi_h^* = \varphi_h^{[1]} \circ \varphi_h^{[2]} \circ \cdots \circ \varphi_h^{[m]} \tag{3}$$

both provide first-order approximations to the exact solution, since $\chi_h(x_0) = \varphi_h(x_0)+\mathcal{O}(h^2)$ and similarly for χ_h^* (which is called the adjoint of χ_h and verifies $\chi_h^* = \chi_{-h}^{-1}$).

It is possible to get higher order approximations by introducing more maps with additional real coefficients, $\varphi_{a_{ij}h}^{[i]}$, in (3). Perhaps the most popular splitting method is the second order symmetric composition

$$\mathcal{S}_h^{[2]} = \chi_{h/2} \circ \chi_{h/2}^* = \varphi_{h/2}^{[m]} \circ \cdots \circ \varphi_{h/2}^{[2]} \circ \varphi_h^{[1]} \circ \varphi_{h/2}^{[2]} \circ \cdots \circ \varphi_{h/2}^{[m]}. \tag{4}$$

When f in (1) is separable in two parts the above particularizes to

$$\chi_h = \varphi_h^{[2]} \circ \varphi_h^{[1]}, \qquad \chi_h^* = \varphi_h^{[1]} \circ \varphi_h^{[2]}, \qquad \mathcal{S}_h^{[2]} = \varphi_{h/2}^{[2]} \circ \varphi_h^{[1]} \circ \varphi_{h/2}^{[2]}, \tag{5}$$

and $\mathcal{S}_h^{[2]}$ is known as the Strang splitting [22], the leapfrog or the Störmer–Verlet method [26], depending on the context where it is used. More generally, one may choose the coefficients a_i, b_i to achieve order r with the composition

$$\psi_h = \varphi_{b_{s+1}h}^{[2]} \circ \varphi_{a_s h}^{[1]} \circ \varphi_{b_s h}^{[2]} \circ \cdots \circ \varphi_{b_2 h}^{[2]} \circ \varphi_{a_1 h}^{[1]} \circ \varphi_{b_1 h}^{[2]}. \tag{6}$$

It turns out that ψ_h can also be written in terms of χ_h and χ_h^*

$$\begin{aligned} \psi_h &= \left(\varphi_{\alpha_{2s}h}^{[2]} \circ \varphi_{\alpha_{2s}h}^{[1]}\right) \circ \cdots \circ \left(\varphi_{\alpha_2 h}^{[2]} \circ \varphi_{\alpha_2 h}^{[1]}\right) \circ \left(\varphi_{\alpha_1 h}^{[1]} \circ \varphi_{\alpha_1 h}^{[2]}\right) \\ &= \chi_{\alpha_{2s}h} \circ \chi_{\alpha_{2s-1}h}^* \circ \cdots \circ \chi_{\alpha_2 h} \circ \chi_{\alpha_1 h}^* \end{aligned} \tag{7}$$

as long as

$$a_j = \alpha_{2j-1} + \alpha_{2j}, \qquad b_{j+1} = \alpha_{2j} + \alpha_{2j+1}. \tag{8}$$

Equivalently,

$$\alpha_1 = b_1, \qquad \alpha_{2j+1} = b_1 + \sum_{k=1}^{j} (b_{k+1} - a_k), \qquad \alpha_{2j} = \sum_{k=1}^{j} (a_k - b_k), \tag{9}$$

with $\alpha_0 = \alpha_{2s+1} = 0$. This relation remains valid if $\sum_{i=1}^{s} a_i = \sum_{i=1}^{s+1} b_i$ [15]. A relevant consequence of this property is that, starting with the coefficients a_i, b_i of a given splitting method, we can get the coefficients α_i for the composition (7), which can be then applied in a more general setting with the maps χ_h and χ_h^* of (3). A particular case widely used in practice to achieve high order approximations consists in considering compositions using the Strang splitting (4) as basic method,

$$\psi_h = \mathcal{S}_{\alpha_s h}^{[2]} \circ \cdots \circ \mathcal{S}_{\alpha_2 h}^{[2]} \circ \mathcal{S}_{\alpha_1 h}^{[2]}. \tag{10}$$

Splitting methods are, in general, explicit, easy to implement and preserve structural properties of the exact solution, thus conferring to the numerical scheme a qualitative superiority with respect to other standard integrators, especially when long time intervals are considered (see [6] for a review).

Examples of these structural features are symplecticity, volume preservation, time-symmetry and conservation of first integrals. In this sense, splitting methods constitute an important class of *geometric numerical integrators* [10, 12, 14, 16, 17, 20].

It has been shown that some of the coefficients in splitting schemes (6) are negative when the order $r \geq 3$ [9, 21, 24]. In other words, the methods always involve stepping backwards in time. An elementary proof of this feature can be worked out as follows. It is quite straightforward to check that one of the necessary condition for the composition (7) (respectively, (10)) to have order $r \geq 3$ is

$$\sum_{i=1}^{k} \alpha_i^3 = 0, \tag{11}$$

with $k = s$ (respect. $k = 2s$). Obviously, this sum vanishes only if at least one of the α_i is negative. In consequence, the flows $\varphi_h^{[j]},\ j = 1, \ldots, m-1$ in (4) and $\varphi_h^{[j]},\ j = 2, \ldots, m-1$ in (3) evolve with at least one negative fractional time step. On the other hand, by taking into account the link (8) among coefficients of (6) and (7), condition (11) with $k = 2s$ leads to

$$\sum_{j=1}^{s} (\alpha_{2j-1}^3 + \alpha_{2j}^3) = 0 \quad \Rightarrow \quad \exists\, k \,/\, \alpha_{2k-1}^3 + \alpha_{2k}^3 < 0 \quad \Rightarrow \quad a_k = \alpha_{2k-1} + \alpha_{2k} < 0.$$

In a similar way, using the same condition with $\alpha_0 = \alpha_{2s+1} = 0$, one has

$$\sum_{j=0}^{s} (\alpha_{2j}^3 + \alpha_{2j+1}^3) = 0 \quad \Rightarrow \quad \exists\, l \,/\, \alpha_{2l}^3 + \alpha_{2l+1}^3 < 0 \quad \Rightarrow \quad b_l = \alpha_{2l} + \alpha_{2l+1} < 0$$

and then at least one a_i as well as one b_i are negative. It must be stressed that condition (11) still persists when the processing technique is used, so that the same conclusion also follows in this case [4].

In summary, the presence of negative coefficients in splitting methods of order higher than two is unavoidable if one restricts oneself to *real* coefficients. Of course, this does not suppose any special impediment when the flow of the ODE evolves in a group (such as in the Hamiltonian case), but may be unacceptable when the differential equation is defined in a semigroup [16], as occurs, for instance, with the simple heat equation $u_t = \Delta u$ on the unit interval with homogeneous Dirichlet boundary conditions. Then the corresponding generated semigroup is well defined only for $t \geq 0$ [11].

More generally, consider the nonlinear heat equation

$$\frac{\partial}{\partial t} u(x,t) = \sum_{i=1}^{d} D_i (v_i(x) D_i u(x,t)) + F(x, u(x,t)) \tag{12}$$

with functions v_i real and positive, and $D_i \equiv \partial/\partial x_i$, on a certain domain $\Omega \in \mathbb{R}^d$. If a space discretization is carried out (either by finite differences or

by a pseudospectral method), a large system of ODEs results which has to be numerically integrated in time. To this end, we can split the resulting equation into linear and nonlinear parts, but schemes of the form (7) or (10) of at most order $r = 2$ can only be applied, since the resulting discrete Laplacian with negative fractional time steps is not well conditioned.

A closely related problem is the linear Schrödinger equation ($\hbar = 1$):

$$i\frac{\partial}{\partial t}\Psi(x,t) = \left(-\frac{1}{2m}\Delta + V(x)\right)\Psi(x,t). \tag{13}$$

A technique used in practice to obtain the eigenvalues and eigenfunctions for a given potential V consists in numerically integrating the equation (after spatial discretization) along pure imaginary times ($\tau = -it$). Equivalently, the equation to be analyzed is

$$\frac{\partial}{\partial \tau}\Psi = \left(\frac{1}{2m}\Delta - V(x)\right)\Psi, \tag{14}$$

which can be considered as a linear heat equation. The system evolves to the ground state whose norm decreases exponentially in proportion to the value of its energy (eigenvalue). By orthogonalization, one can make the system to evolve to any other eigenfunction [1, 13]. In any case, whereas there is no special difficulties with numerically integrating equation (13) using a splitting methods with negative fractional time steps, this is not the case for (12) and (14) due to the presence of the Laplacian.

It has been noticed, however, that higher order splitting methods with complex coefficients having positive real part do exist [3, 16, 23, 24, 25]. These schemes were reported mainly for theoretical purposes but received very little attention as practical numerical tools. Perhaps the main reason was that working with complex arithmetic makes the schemes more involved and, in many cases, also considerably more costly from a computational point of view (usually, four times more expensive).

It is only within recent years that a systematic search for new methods with complex coefficients has been carried out and the resulting schemes have been tested in different settings: Hamiltonian systems in celestial mechanics [8], the time-dependent Schrödinger equation in quantum mechanics [2, 19] and also in the more abstract setting of evolution equations with unbounded operators generating analytic semigroups [7, 11]. In this sense, we recall that the propagator $\exp(z\Delta)$ ($z \in \mathbb{C}$) associated with the Laplacian is well defined (in a reasonable distributional sense) if and only if $\mathrm{Re}(z) \geq 0$ [7]. More generally, it is possible to extend the semigroup related with parabolic PDEs into a sector in the right half plane of $\mathbb{C}$ [11].

The aim of this paper is to review some of the splitting methods with complex coefficients published in the literature, propose new sixth-order schemes in the class (10) and analyze them on a pair of simple numerical examples, to get a glance of the performance and main features of this kind of integrators and some of the difficulties involved.

2 Integrators with complex coefficients

Most of the existing splitting methods with complex coefficients have been constructed by applying the composition technique to the symmetric second-order leapfrog scheme $\mathcal{S}^{[2]}$. Thus, one gets a third-order method as

$$\mathcal{S}_h^{[3]} = \mathcal{S}_{\alpha h}^{[2]} \circ \mathcal{S}_{\beta h}^{[2]}, \tag{15}$$

where the coefficients have to satisfy (11) together with the consistency condition

$$\left.\begin{array}{rcl} \alpha+\beta & = & 1 \\ \alpha^3+\beta^3 & = & 0 \end{array}\right\} \Rightarrow \quad \alpha = \tfrac{1}{2} \mp i\tfrac{\sqrt{3}}{6}, \qquad \beta = \tfrac{1}{2} \pm i\tfrac{\sqrt{3}}{6}.$$

Due to its simplicity, this scheme has been rediscovered several times, either as the composition (15) [3, 23, 7] or by solving the order conditions required by (6) with $s = 2$ [8, 11].

A fourth-order integrator can be obtained with the symmetric composition

$$\mathcal{S}_h^{[4]} = \mathcal{S}_{\alpha h}^{[2]} \circ \mathcal{S}_{\beta h}^{[2]} \circ \mathcal{S}_{\alpha h}^{[2]}. \tag{16}$$

Although the necessary order conditions are the same, the time-symmetry of the composition rises the order by one (all the error terms at odd orders vanish identically):

$$\left.\begin{array}{rcl} 2\alpha+\beta & = & 1 \\ 2\alpha^3+\beta^3 & = & 0 \end{array}\right\} \Rightarrow \quad \alpha = \frac{1}{2 - 2^{1/3}\,\mathrm{e}^{2ik\pi/3}}, \qquad \beta = \frac{2^{1/3}\,\mathrm{e}^{2ik\pi/3}}{2 - 2^{1/3}\,\mathrm{e}^{2ik\pi/3}}$$

with $k = 0, 1, 2$. Notice that for $k = 1, 2$ it is true that $\mathrm{Re}(\alpha), \mathrm{Re}(\beta) > 0$.

Another fourth-order method can be obtained by symmetrizing the third-order scheme (15), i.e.,

$$\mathcal{S}_h^{[4]} = \mathcal{S}_{\alpha/2h}^{[2]} \circ \mathcal{S}_{\beta/2h}^{[2]} \circ \mathcal{S}_{\beta/2h}^{[2]} \circ \mathcal{S}_{\alpha/2h}^{[2]}. \tag{17}$$

Methods (15), (16) and (17) can be used to generate recursively higher order composition schemes as

$$\mathcal{S}_h^{[n+1]} = \mathcal{S}_{\alpha h}^{[n]} \circ \mathcal{S}_{\beta h}^{[n]}. \tag{18}$$

Here the coefficients have to verify the conditions $\alpha + \beta = 1$, $\alpha^{n+1} + \beta^{n+1} = 0$, whence

$$\alpha = \frac{1}{2} + i\,\frac{\sin(\frac{2l+1}{n+1}\pi)}{2 + 2\cos(\frac{2l+1}{n+1}\pi)} \quad \text{for} \quad \left\{ \begin{array}{ll} -\frac{n}{2} \le l \le \frac{n}{2} - 1 & \text{if } n \text{ is even,} \\ -\frac{n+1}{2} \le l \le \frac{n-1}{2} & \text{if } n \text{ is odd,} \end{array} \right.$$

and $\beta = 1 - \alpha$. The choice $l = 0$ gives the solutions with the smallest phase and allows one to build methods up to order six with coefficients having positive real part. This feature was stated in [25] and rediscovered in [11].

In a similar way, one may use recursively a symmetric three term composition, which allows to increase the order by two at each iteration:

$$\mathcal{S}_h^{[n+2]} = \mathcal{S}_{\alpha h}^{[n]} \circ \mathcal{S}_{\beta h}^{[n]} \circ \mathcal{S}_{\alpha h}^{[n]}, \tag{19}$$

with

$$2\alpha + \beta = 1, \qquad 2\alpha^{n+1} + \beta^{n+1} = 0.$$

The solutions providing coefficients with the smallest phase are

$$\alpha = \frac{\mathrm{e}^{i\pi/(n+1)}}{2^{1/(n+1)} - 2\,\mathrm{e}^{i\pi/(n+1)}}, \qquad \beta = 1 - 2\alpha,$$

and methods up to order eight with coefficients having positive real part are possible. Moreover, methods up to order fourteen of the more general form (10) with coefficients α_j with positive real part are attainable [7, 11]. An interesting (and open) question is to determine whether arbitrarily high orders can be attained or wether, as for the previous compositions, there is an order barrier for methods of the form (10) with $\mathrm{Re}(\alpha_j) > 0$. Observe that any method of the form (10) with coefficients having positive real part can be expressed in terms of the elementary flows $\varphi_{\lambda h}^{[j]}$ with $\mathrm{Re}(\lambda) > 0$ when $\mathcal{S}_h^{[2]}$ is taken as the leapfrog (5). This is also true, of course, in the more general case when f is split in m parts and $\mathcal{S}_h^{[2]}$ is taken as the symmetric second order basic method (4).

For instance, suppose that f in (1) is separable in two parts, so that $\mathcal{S}_h^{[2]}$ is given by (5). Then it is straightforward to check that the third order scheme (15) can be written as

$$\mathcal{S}_h^{[3]} = \varphi_{b_3 h}^{[2]} \circ \varphi_{a_2 h}^{[1]} \circ \varphi_{b_2 h}^{[2]} \circ \varphi_{a_1 h}^{[1]} \circ \varphi_{b_1 h}^{[2]} \tag{20}$$

with $a_1 = \frac{1}{2} + i\frac{\sqrt{3}}{6}$, $a_2 = a_1^*$, $b_1 = a_1/2$, $b_2 = 1/2$, $b_3 = b_1^*$. This particular symmetry of the coefficients results in a method whose leading error terms at order 4 are all strictly imaginary [8].

Another question of practical nature is the construction of methods of the form (10) with $\mathrm{Re}(\alpha_j) > 0$ involving the minimum number s of compositions for a prescribed order. For instance, the minimal number of compositions for achieving order 6 is $s = 7$. The corresponding order conditions can be written as [6, 10, 18]

$$\sum_{j=1}^{s} \alpha_j = 1, \quad \sum_{j=1}^{s} \alpha_j^k = 0, \quad k = 3, 5, \tag{21}$$

$$\sum_{j=1}^{s} \alpha_j^k c_j^\ell = 0, \quad (k, \ell) \in \{(3,1), (3,2), (3,3), (5,1)\}, \tag{22}$$

where for each $j = 1, \ldots, s$,

$$c_j = \frac{\alpha_j}{2} + \sum_{i=1}^{j-1} \alpha_i.$$

This system of algebraic equations has several solutions with $\mathrm{Re}(\alpha_j) > 0$. Among them, we have chosen the two sets of coefficients collected in Table 1. The first one corresponds to a symmetric method, $\alpha_{s+1-i} = \alpha_i$, as scheme (16), and was already found by Chambers [8]. The second method is apparently new, and possesses the special symmetry $\alpha_{s+1-i} = \alpha_i^*$, as scheme (15) (or (20) when expressed as (6)).

Table 1: Coefficients of two 7-stage sixth-order methods of type (10): $\mathrm{S}_7 6$ is a symmetric method and $\mathrm{S}_7^* 6$ is conjugate to a symmetric method (symmetric in the real part of the coefficients and skew-symmetric in the imaginary part).

$\mathrm{S}_7 6$
$\alpha_1 = 0.11690037554661284389 + 0.043428254616060341762\, i$
$\alpha_2 = 0.12955910128208826275 - 0.12398961218809259330\, i$
$\alpha_3 = 0.18653249281213381780 + 0.00310743071007267534\, i$
$\alpha_4 = 0.13401673670223327014 + 0.15490785372391915239\, i$
$\alpha_5 = \alpha_3, \quad \alpha_6 = \alpha_2, \quad \alpha_7 = \alpha_1$
$\mathrm{S}_7^* 6$
$\alpha_1 = 0.13374177891468362845 2 - 0.02883902837102555399 5\, i$
$\alpha_2 = 0.12134019583938803504 + 0.11585180844272788007\, i$
$\alpha_3 = 0.13489797942731665044 - 0.12906241362827633477\, i$
$\alpha_4 = 0.22004009163722337213$
$\alpha_5 = \alpha_3^*, \quad \alpha_6 = \alpha_2^*, \quad \alpha_7 = \alpha_1^*$

3 Numerical examples

3.1 Example 1: the harmonic oscillator

We consider the simple harmonic oscillator to illustrate some qualitative properties of the previous composition methods with complex coefficients. That is, we take the Hamiltonian function $H(q,p) = \frac{1}{2}(p^2 + q^2)$, with $q, p \in \mathbb{R}$. The corresponding equations of motion are linear and can be written as

$$x' \equiv \begin{pmatrix} q' \\ p' \end{pmatrix} = \Big[\underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{A} + \underbrace{\begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}}_{B} \Big] \begin{pmatrix} q \\ p \end{pmatrix} = (A+B)\, x, \tag{23}$$

so that the numerical solution at time $t = h$ furnished by method (6) is given by

$$x(h) = K(h) x_0 \equiv \mathrm{e}^{b_{s+1} hB}\, \mathrm{e}^{a_s hA}\, \mathrm{e}^{b_s hB} \cdots \mathrm{e}^{b_2 hB}\, \mathrm{e}^{a_1 hA}\, \mathrm{e}^{b_1 hB} x_0. \tag{24}$$

As is well known, for splitting methods with real coefficients the average error in energy remain constant for exponentially long times under suitable general

conditions on the Hamiltonian. For the particular case of the harmonic oscillator and with a sufficiently small time step, this is true for all times, and the average error in positions grows only linearly.

We propose here to check whether this also holds for methods with complex coefficients. To do that, we take as initial conditions $(q, p) = (1, 1)$ and integrate the system (23) for $t \in [0, 20000\pi]$ using a constant time step. We measure the error in position and energy of the output obtained by propagating the solution with the splitting method and then computing the real parts of the results $q_{out} = \mathrm{Re}(q)$, $p_{out} = \mathrm{Re}(p)$. Figure 1 shows the results obtained with the following methods: (i) $S_2 3$, the 2-stage third-order non-symmetric method (15), (ii) $S_3 4$, the 3-stage fourth-order symmetric method (16), (iii) $S_7^* 6$, the 7-stage sixth-order non-symmetric method, (iv) $S_7 6$, the 7-stage sixth-order symmetric method. The coefficients of these two 6th-order methods are collected in Table 1. The time step is chosen such that all methods require 27-28 evaluations per period. Notice the significant difference in the qualitative behavior of the numerical solution. Whereas the error grows exponentially for the symmetric methods $S_3 4$ and $S_7 6$, this is not the case for $S_2 3$ and $S_7^* 6$, which show a performance analogous to standard splitting methods with real coefficients: bounded energy error and linear growth of error in positions. Of course, such a behavior deserves a theoretical explanation, which we pursue next.

The matrix $K(h)$ in (24) is given explicitly by

$$K(h) = \begin{pmatrix} 1 & 0 \\ -b_{s+1}h & 1 \end{pmatrix} \begin{pmatrix} 1 & a_s h \\ 0 & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & a_1 h \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -b_1 h & 1 \end{pmatrix}.$$

In this way, one gets

$$K(h) = \begin{pmatrix} p(h) + d(h) & q(h) + e(h) \\ -q(h) + e(h) & p(h) - d(h) \end{pmatrix}$$

where $p(h)$, $d(h)$ (respectively, $q(h)$, $e(h)$) are even (resp. odd) polynomial functions having in general complex coefficients and $\det K(h) = p(h)^2 + q(h)^2 - d(h)^2 - e(h)^2 = 1$.

If the splitting method (6) is such that

$$a_{s-j+1} = a_j^*, \qquad b_{s-j+2} = b_j^* \tag{25}$$

(as happens, in particular, when it comes from a composition of the form (10) with $\alpha_{s-j+1} = \alpha_j^*$), then $K(h)^{-1} = K(-h)^*$. More specifically,

$$\begin{pmatrix} p(h) - d(h) & -q(h) - e(h) \\ q(h) - e(h) & p(h) + d(h) \end{pmatrix} = \begin{pmatrix} p(h)^* + d(h)^* & -q(h)^* - e(h)^* \\ q(h)^* - e(h)^* & p(h)^* - d(h)^* \end{pmatrix}.$$

This implies that $p(y)$, $q(h)$, and $e(h)$ are real polynomials, whereas the coefficients of $d(y)$ are purely imaginary. Notice that this is precisely the case of methods $S_2 3$ and $S_7^* 6$.

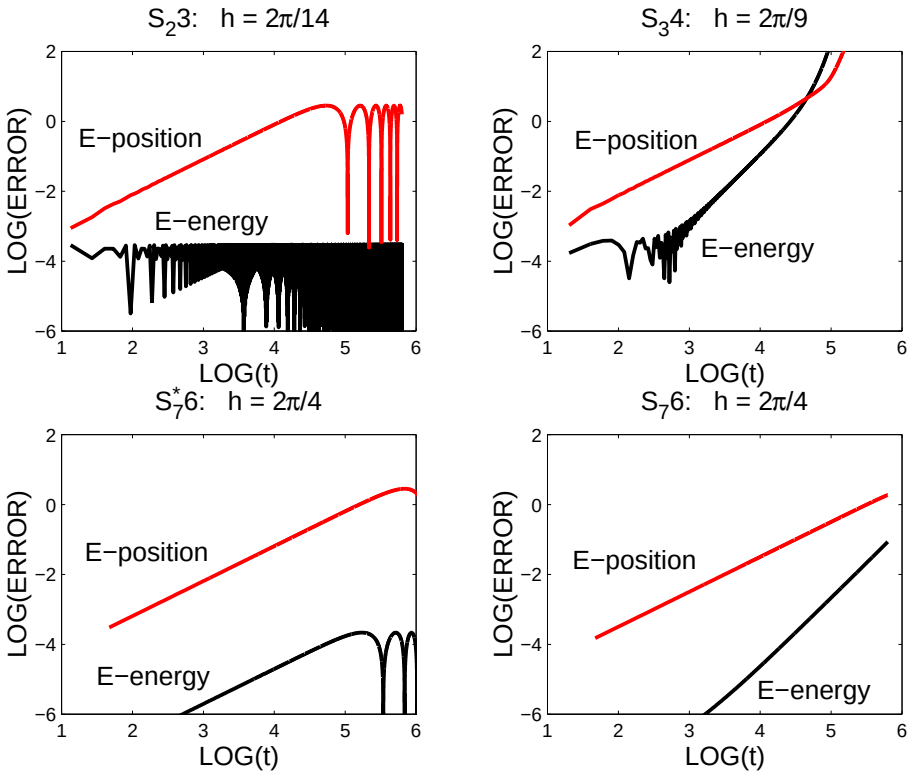


Figure 1: Error in position and energy (taking the real part from the output) obtained with the 4th- and 6th-order symmetric schemes $S_3 4$ and $S_7 6$, and non-symmetric methods $S_2 3$ and $S_7^* 6$. The time step is chosen such that all schemes require 27-28 evaluations per period.

If, on the other hand, the splitting method is symmetric, i.e., it is of the form (6) satisfying

$$a_{s-j+1} = a_j, \quad b_{s-j+2} = b_j$$

(as happens, in particular, when it comes from a composition of the form (10) with $\alpha_{s-j+1} = \alpha_j$), then $K(h)^{-1} = K(-h)$. This clearly implies that $d(h) \equiv 0$, but in general the polynomials $p(h)$, $q(h)$, and $e(h)$ have complex coefficients. For instance, methods (16) ($S_3 4$) and $S_7 6$ are such that $p(h)$ has non-real coefficients.

When a splitting method with matrix $K(h)$ is used to integrate the harmonic oscillator, it is essential that $p(h) \in \mathbb{R}$. Otherwise $K(h)^n$ grows exponentially with the number n of steps. As a matter of fact, the eigenvalues of $K(h)$ are

$\lambda_1 = \mathrm{e}^{i\phi(h)}$ and $\lambda_2 = \mathrm{e}^{-i\phi(h)}$, where

$$\phi(h) = \arccos(p(h)),$$

and thus $\max(|\lambda_1|, |\lambda_2|) > 1$ if $p(h) \notin \mathbb{R}$ (and also if $p(h) \in \mathbb{R}$ and $|p(h)| > 1$). That is precisely the situation with methods $S_3 4$ and $S_7 6$, and thus they are useless when integrating harmonic oscillators or systems that can be considered as close perturbations of harmonic oscillators with the partition (23).

From the previous comments, it is clear that instability will take place when integrating the harmonic oscillator unless $-1 \leq p(h) \leq 1$. In fact, the numerical solution can still be (weakly) unstable when $p(h)^2 = 1$ with $q(h)d(h)e(h) \neq 0$ [5]. Furthermore, it is shown in [5] that, for stable numerical solutions (that is, either $-1 < p(h) < 1$ or $p(h)^2 = 1$ with $q(h) = d(h) = e(h) = 0$), one has

$$K(h)^n = Q(h)^{-1} \begin{pmatrix} \cos(n\phi(h)) & \sin(n\phi(h)) \\ -\sin(n\phi(h)) & \cos(n\phi(h)) \end{pmatrix} Q(h),$$

with a suitable 2×2 matrix $Q(h)$ (typically close to the identity matrix). In consequence, the numerical solution $x_n = (q_n, p_n)$ is such that $\tilde{x}_n := Q(h)x_n$ corresponds to the exact solution at $t_n = nh$ of a harmonic oscillator with frequency $\tilde{\omega} = 1/h\phi(h) \approx 1$. This feature explains why schemes $S_2 3$ and $S_7^* 6$, when applied to the harmonic oscillator (23) with $h = \pi/7$ and $h = \pi/2$ respectively, exhibit a linear error growth in positions and a bounded error in energy, since for such methods, $p(h) = 1 - h^2/2 + \cdots$ is real and satisfies $p(h) \in (-1, 1)$ for the values of h considered in the numerical experiments.

3.2 Example 2: The Volterra–Lotka problem

Consider now the Volterra–Lotka problem

$$\dot{u} = u(v-2), \qquad \dot{v} = v(1-u). \tag{26}$$

This is a very simple nonlinear system which allows us to make a preliminary study about the behavior and performance of methods with complex coefficients in the transition process from a linear to a nonlinear problem. In a neighborhood of the steady state at $(u^*, v^*) = (1, 2)$ the system can be considered close to a harmonic oscillator. The nonlinear contributions are manifest as we move away from it. The system evolves along periodic trajectories around the equilibrium point in the region $0 < u, v$ determined by the first integral $I(u, v) = \ln(uv^2) - (u + v)$.

The vector field $f(u, v) = (u(v-2), v(1-u))$ can be separated in two solvable parts and this can be done in different ways. We consider the following split: $f_A = (u(v-2), 0)$ and $f_B = (0, v(1-u))$ (although the linear and nonlinear separation can also be considered).

We take as initial conditions $(u_0, v_0) = (2, 4)$, integrate up to $t = 20000 \times 2\pi$ and measure the relative error in the first integral, $|I - I_0|/|I_0|$. As in the previous example, we integrate using complex arithmetic and take the real parts of u and v only for representing the output. Figures 2-(a) and (b) show

the results obtained for time steps $h = \frac{4m\pi}{210}$ and four times smaller $h = \frac{m\pi}{210}$, with m the number of stages of each method. In this way, all methods require the same number of evaluations. Contrarily to the pure harmonic oscillator, we observe a secular error growth in the determination of the first integral for all methods which diminishes considerably when the time step is reduced. The observed behavior resembles what takes place with the so-called pseudo-symplectic methods (integrators of order n which preserve symplecticity up to order $p > n$), where the dominant errors behave as $\mathcal{E}_I = Ch^n + tDh^p$ for some constants C and D. If $p > n$ the secular part of the error does not manifest for relatively long times when the time step is reduced.

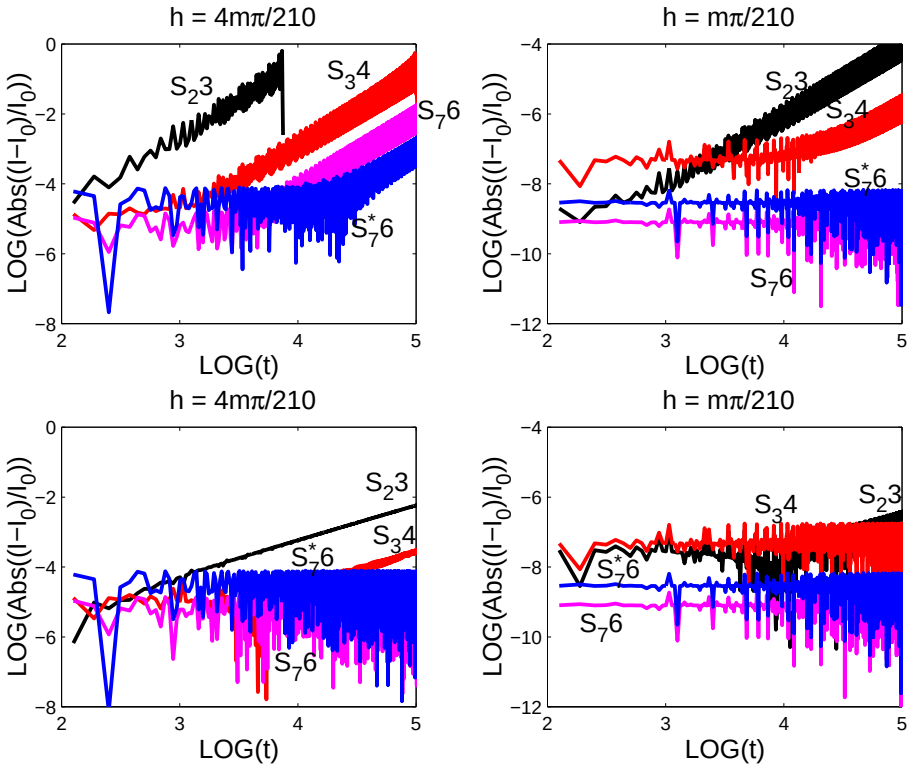


Figure 2: Relative error in the first integral $I = \ln(uv^2) - (u + v)$ for the Volterra–Lotka problem with initial conditions $(u_0, v_0) = (2, 4)$ for the time steps $h = \frac{4m\pi}{420}$ and $h = \frac{m\pi}{420}$, with m the number of stages of each method.

We have repeated the same experiment but, after each time step, we discard the imaginary part of u and v and initiate the next step only with their real part. In other words, we project each component on the real axis at the end of each integration step. The results obtained are shown in Figures 2-(c) and (d). Obviously, this way of proceeding does not preserve symplecticity any more but

the results obtained suggest that a significant improvement in accuracy can be achieved.

4 Conclusions and outlook

We have presented a short review of the splitting and composition technique to build methods of order greater than two with complex coefficients with positive real part. This procedure allows to overcome the order barrier where splitting methods of order greater than two involve necessarily negative coefficients in the real space. In general, splitting methods with complex coefficients are considerably more expensive than the corresponding methods with real coefficients (about four times more expensive), and this make them hardly competitive in practice. For this reason, one can think that the main application of the new methods could be on parabolic PDEs, where higher order methods with real coefficients (which necessarily have some negative coefficientes) can not be used. However, there is a number of problems which evolve in the complex space where using methods which complex coefficients does not necessarily mean increasing the cost of the algorithm. This can be the case, for instance, of the Schödinger equation (13).

As for the practical implementation of splitting methods with complex coefficients, in [8] it is claimed that one has to carry the numerical integration in complex variables, and (for problems with real solutions) one should take either the real part of the variables or their modulus only for the output. However, we have observed that removing the imaginary part at each step, i.e. projecting on the real space at each step, the error grow can be considerably diminished in some cases. In the numerical examples considered in previous section, the linear error grow in the first integrals originate from different sources depending on wether the projection onto the real domain is performed after each step or not. In the first case, the projection after each step destroys symplecticity but only at a higher order, and the schemes can be considered as pseudosymplectic. In the second case, the method is actually symplectic and can thus be (formally) considered as an exact solution of a Hamiltonian system in the complex domain, which have qualitatively different properties to trajectories in the real domain. We have also noticed that the higher order methods present a considerably reduced error grow. Then, it seems appropriate to look for efficient higher order methods with complex coefficients. In general, symmetric splitting methods are desirable. However, we have shown that for the harmonic oscillator symmetric methods (with non-real stability polynomial) present an exponential error grow, which is not the case for methods with the special symmetry (25). In a preliminary search of methods, we have presented a new sixth-order method with that special symmetry. This is an interesting subject to be further explored since many problems in different applications can be considered as perturbations to the harmonic oscillator.

Acknowledgements

This work has been supported by Ministerio de Ciencia e Innovación (Spain) under project MTM2007-61572 (co-financed by the ERDF of the European Union). SB also acknowledges financial support from Generalitat Valenciana through project GV/2009/032.

References

[1] J. Auer, E. Krotscheck, and S.A. Chin. A fourth-order real-space algorithm for solving local Schrödinger equations. *J. Chem Phys.*, 115:6841–6846, 2001.

[2] A.D. Bandrauk, E. Dehghanian, and H. Lu. Complex integration steps in decomposition of quantum exponential evolution operators. *Chem. Phys. Lett.*, 419:346–350, 2006.

[3] A.D. Bandrauk and Hai Shen, Improved exponential split operator method for solving the time-dependent Schrödinger equation. *Chem. Phys. Lett.*, 176:428–432, 1991.

[4] S. Blanes and F. Casas. On the necessity of negative coefficients for operator splitting schemes of order higher than two. *Appl. Numer. Math.*, 54:23–37, 2005.

[5] S. Blanes, F. Casas, and A. Murua. On the linear stability of splitting methods. *Found. Comp. Math.*, 8:357–393, 2008.

[6] S. Blanes, F. Casas, and A. Murua. Splitting and composition methods in the numerical integration of differential equations. *Bol. Soc. Esp. Math. Apl.*, 45:87–143, 2008.

[7] F. Castella, P. Chartier, S. Decombes, and G. Vilmart. Splitting methods with complex times for parabolic equations. *BIT*, 49:487–508, 2009.

[8] J.E. Chambers. Symplectic integrators with complex time steps. *Astron. J.*, 126:1119–1126, 2003.

[9] D. Goldman and T.J. Kaper. nth-order operator splitting schemes and nonreversible systems. *SIAM J. Numer. Anal.*, 33:349–367, 1996.

[10] E. Hairer, Ch. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer-Verlag, Second edition, 2006.

[11] E. Hansen and A. Ostermann. High order splitting methods for analytic semigroups exist. *BIT*, 49:527–542, 2009.

[12] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna. Lie-group methods. *Acta Numerica*, 9:215–365, 2000.

[13] L. Lehtovaara, J. Toivanen, and J. Eloranta. Solution of time-independent Schrödinger equation by the imaginary time propagation method. *J. Comput. Phys.*, 221:148–157, 2007.

[14] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics.* Cambridge University Press, 2004.

[15] R. I. McLachlan. On the numerical integration of ordinary differential equations by symmetric composition methods. *SIAM J. Numer. Anal.*, 16:151–168, 1995.

[16] R.I. McLachlan and R. Quispel. Splitting methods. *Acta Numerica*, 11:341–434, 2002.

[17] R.I. McLachlan and R. Quispel. Geometric integrators for ODEs. *J. Phys. A: Math. Gen.*, 39:5251–5285, 2006.

[18] A. Murua and J.M. Sanz-Serna. Order conditions for numerical integrators obtained by composing simpler integrators. *Philos. Trans. Royal Soc. London, ser. A*, 357:1079-1100, 1999.

[19] T. Prosen and I. Pizorn. High order non-unitary split-step decomposition of unitary operators. *J. Phys. A: Math. Gen.*, 39:5957–5964, 2006.

[20] J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian Problems.* AMMC 7. Chapman & Hall, 1994.

[21] Q. Sheng. Solving linear partial differential equations by exponential splitting. *IMA J. Numer. Anal.*, 9:199–212, 1989.

[22] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.

[23] M. Suzuki, Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations. *Phys. Lett. A*, 146:319-323, 1990.

[24] M. Suzuki. General theory of fractal path integrals with applications to many-body theories and statistical physics. *J. Math. Phys.*, 32:400–407, 1991.

[25] M. Suzuki. Hybrid exponential product formulas for unbounded operators with possible applications to Monte Carlo simulations. *Phys. Lett. A*, 201:425–428, 1995.

[26] L. Verlet. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules. *Phys. Rev.*, 159:98–103, 1967.

Bol. Soc. Esp. Mat. Apl.
nº50(2010), 61–82

SIMULATING NONHOLONOMIC DYNAMICS

M. KOBILAROV*, D. MARTíN DE DIEGO†, S. FERRARO‡

*California Institute of Technology, Control and Dynamical Systems
Pasadena, CA 91125, USA
†Instituto de Ciencias Matemáticas. CSIC-UAM-UC3M-UCM
Serrano 123, 28006 Madrid, Spain
‡Departamento de Matemática e Instituto de Matemática,
Universidad Nacional del Sur,
Av. Alem 1253, 8000 Bahía Blanca, Argentina

marin@cds.caltech.edu, david.martin@icmat.es, sferraro@uns.edu.ar

Abstract

This paper develops different discretization schemes for nonholonomic mechanical systems through a discrete geometric approach. The proposed methods are designed to account for the special geometric structure of the nonholonomic motion. Two different families of nonholonomic integrators are developed and examined numerically: the geometric nonholonomic integrator (GNI) and the reduced d'Alembert-Pontryagin integrator (RDP). As a result, the paper provides a general tool for engineering applications, i.e. for automatic derivation of numerically accurate and stable dynamics integration schemes applicable to a variety of robotic vehicle models.

Key words: *Geometric integrator, Nonholonomic mechanics, discrete variational calculus, reduction by symmetries*
AMS subject classifications: *58F15, 58F17, 53C35*

1 Introduction

Nonholonomic constraints have been the subject of deep analysis since the dawn of Analytical Mechanics. Hertz, in 1894, was the first to use the term "nonholonomic system", but we can even find older references in the work by Euler in 1734, who studied the dynamics of a rolling rigid body moving without slipping on a horizontal plane. Many authors have recently shown a new interest in that theory and also in its relation to the new developments in control theory, subriemannian geometry, robotics, etc (see, for instance, [24]). The main characteristic of this period is that Geometry was used in a systematic way (see L.D. Fadeev and A.M. Vershik [28] as an advanced and fundamental reference, and also, [1, 2, 9, 5, 18, 20] and references therein).

In the case of nonholonomic mechanics, these constraint functions are, roughly speaking, functions on the velocities that are not derivable from position constraints. Traditionally, the equations of motion for nonholonomic mechanics are derived from the Lagrange-d'Alembert principle which restricts the set of infinitesimal variations (or constrained forces) in terms of the constraint functions.

Recent works, such as [8, 10, 14, 23], have introduced numerical integrators for nonholonomic systems with very good energy behavior and properties such as the preservation of the discrete nonholonomic momentum map. In this paper, we will review and compare two new methods for nonholonomic mechanics, the Geometric Nonholonomic Integrator (GNI) [12] and the Reduced d'Alembert-Pontryagin Integrator (RDP) [17], examining their behavior in the numerical simulation of some of the most typical examples in nonholonomic mechanics: the Chaplygin sleigh and the snakeboard.

Finally, the developed algorithms are packaged as a general computational tool for automatic derivation of nonholonomic integrators given the system constraints and Lagrangian. It is available for download from
`http://www.cds.caltech.edu/~marin/index.php?n=nhi`

2 Introduction to Discrete Mechanics

Discrete variational integrators appear as a special kind of geometric integrators (see [13, 27]). These integrators have their roots in the optimal control literature in the 1960's and 1970's. In the sequel we will review the construction of this specific type of geometric integrators (see [22] for an excellent survey about this topic).

A discrete Lagrangian is a map $L_d \colon Q \times Q \to \mathbb{R}$, where Q is a finite-dimensional configuration manifold. For the construction of numerical integrators for a continuous Lagrangian system given by a Lagrangian $L : TQ \to \mathbb{R}$, the discrete Lagrangian may be considered as an approximation of the integral action

$$L_d(q_0, q_1) \approxeq \int_0^h L(q(t), \dot{q}(t))\, dt$$

where $q(t)$ is a solution of the Euler-Lagrange equations corresponding to L, that is,

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}(q(t), \dot{q}(t))\right) - \frac{\partial L}{\partial q}(q(t), \dot{q}(t)) = 0\,, \tag{1}$$

additionally satisfying $q(0) = q_0$ and $q(h) = q_1$, where h is the time step. Observe that this solution always exists if the Lagrangian is regular and h is small enough (see [25]).

Define the action sum $S_d : Q^{N+1} \to \mathbb{R}$ corresponding to the Lagrangian L_d by

$$S_d = \sum_{k=1}^{N} L_d(q_{k-1}, q_k),$$

where $q_k \in Q$ for $0 \leq k \leq N$. For any covector $\alpha \in T^*_{(x_1,x_2)}(Q \times Q)$, we have a decomposition $\alpha = \alpha_1 + \alpha_2$ where $\alpha_i \in T^*_{x_i}Q$. Therefore,

$$dL_d(q_0, q_1) = D_1 L_d(q_0, q_1) + D_2 L_d(q_0, q_1) \ .$$

The discrete variational principle states that the solutions of the discrete system determined by L_d must extremize the action sum given fixed points q_0 and q_N.

Extremizing S_d over q_k, $1 \leq k \leq N-1$, we obtain the following system of difference equations

$$D_1 L_d(q_k, q_{k+1}) + D_2 L_d(q_{k-1}, q_k) = 0 \ . \tag{2}$$

These equations are usually called the *discrete Euler-Lagrange equations*.

The geometrical properties corresponding to this numerical method are obtained defining two discrete Legendre transformations associated to L_d by

$$\begin{array}{rccc} \mathbb{F}^- L_d : & Q \times Q & \longrightarrow & T^*Q \\ & (q_0, q_1) & \longmapsto & (q_0, -D_1 L_d(q_0, q_1)) \end{array}$$

$$\begin{array}{rccc} \mathbb{F}^+ L_d : & Q \times Q & \longrightarrow & T^*Q \\ & (q_0, q_1) & \longmapsto & (q_0, D_2 L_d(q_0, q_1)) \end{array}$$

and the 2-form $\omega_d = (\mathbb{F}^{\pm} L_d)^* \omega_Q$, where ω_Q is the canonical symplectic form on T^*Q. We will say that the discrete Lagrangian is regular if $\mathbb{F}^- L_d$ is a local diffeomorphism. We will have that:

$$\begin{array}{lcl} \mathbb{F}^- L_d \text{ is a local diffeomorphism} & \Leftrightarrow & \mathbb{F}^+ L_d \text{ is a local diffeomorphism} \\ & \Leftrightarrow & \omega_d \text{ is symplectic} \end{array}$$

Under this regularity condition, this implicit system of difference equations (2) defines a local discrete flow $\Upsilon : U \subset Q \times Q \longrightarrow Q \times Q$, by $\Upsilon(q_{k-1}, q_k) = (q_k, q_{k+1})$. The discrete algorithm determined by Υ preserves the symplectic form ω_d, i.e., $\Upsilon^* \omega_d = \omega_d$. Moreover, if the discrete Lagrangian is invariant under the diagonal action of a Lie group G, then the discrete momentum map $J_d : Q \times Q \to \mathfrak{g}^*$ defined by $\langle J_d(q_k, q_{k+1}), \xi \rangle = \langle D_2 L_d(q_k, q_{k+1}), \xi_Q(q_{k+1}) \rangle$ is preserved by the discrete flow. Here, ξ_Q denotes the fundamental vector field determined by $\xi \in \mathfrak{g}$:

$$\xi_Q(q) = \frac{d}{dt}\Big|_{t=0} (\exp(t\xi) \cdot q) \ .$$

Therefore, these integrators are **symplectic-momentum preserving integrators**.

In [21] we have obtained a geometric derivation of variational integrators that is also valid for reduced systems (on Lie algebras, quotient of tangent bundles by a Lie group action, etc.)

3 Description of the nonholonomic dynamics

The presence of nonholonomic (or holonomic) constraints gives rise to forces. Nonholonomic systems are described by the Lagrange-D'Alembert's principle which prescribes the constraint forces induced by the given nonholonomic constraints. In the following we will describe the equations of motion of a nonholonomic system in terms of Riemannian geometric tools (see [5]).

Let Q be an n-dimensional differentiable manifold, with local coordinates (q^i), $1 \leq i \leq n$. Consider a mechanical Lagrangian system $L : TQ \to \mathbb{R}$ defined by $L(v_q) = \frac{1}{2}\mathcal{G}(v_q, v_q) - V(q)$, $v_q \in T_qQ$ or, locally

$$L(q, \dot{q}) = \frac{1}{2} g_{ij}(q)\dot{q}^i\dot{q}^j - V(q) \; . \tag{3}$$

Here $\mathcal{G}$ is a Riemannian metric on Q (locally defined by the symmetric, positive definite matrix $(g_{ij}(q))_{1\leq i,j\leq n}$) and V represents a potential function. We know that the equations of motion for a Lagrangian system are (1) which, in the case of a mechanical Lagrangian system of the form (3), admits a nice expression in terms of standard Riemmanian geometric tools:

$$\nabla_{\dot{c}(t)}\dot{c}(t) = -\text{grad } V(c(t))$$

where ∇ is the Levi–Civita connection associated to $\mathcal{G}$ and, in coordinates, grad $V(c(t)) = g^{ij}\frac{\partial V}{\partial q^j}$ where (g^{ij}) is the inverse matrix of (g_{ij}).

Assume that the system is subjected to nonholonomic constraints, defined by a regular distribution $\mathcal{D}$ on Q, with rank $\mathcal{D} = n - m$. Locally the nonholonomic constraints are described by the vanishing of m independent functions

$$\phi^a = \mu_i^a(q)\dot{q}^i, \quad 1 \leq a \leq m \;\; \text{(the "constraint functions")}.$$

The Lagrange–d'Alembert principle states that the equations of motion for a nonholonomic system determined by the two data $(L, \mathcal{D})$ are:

$$\begin{aligned} \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^i}(q(t), \dot{q}(t))\right) - \frac{\partial L}{\partial q^i}(q(t), \dot{q}(t)) &= \lambda_a \mu_i^a(q(t)) \; , \\ \mu_i^a(q(t))\dot{q}^i(t) &= 0 \end{aligned} \tag{4}$$

where λ_a, $1 \leq a \leq m$ are Lagrange multipliers to be determined. Using the Levi-Civita connection we find an intrinsic equation for the nonholonomic equations:

$$\nabla_{\dot{c}(t)}\dot{c}(t) = -\text{grad } V(c(t)) + \bar{\lambda}(t), \quad \dot{c}(t) \in \mathcal{D}_{c(t)},$$

where $\bar{\lambda}$ is a section of $\mathcal{D}^\perp$ along c. Here $\mathcal{D}^\perp$ stands for the orthogonal complement of $\mathcal{D}$ with respect to the metric $\mathcal{G}$.

In coordinates, defining the n^3 functions Γ_{ij}^k (Christoffel symbols for ∇) by

$$\nabla_{\frac{\partial}{\partial q^i}}\frac{\partial}{\partial q^j} = \Gamma_{ij}^k \frac{\partial}{\partial q^k},$$

we may rewrite the nonholonomic equations of motion as

$$\ddot{q}^k(t) + \Gamma^k_{ij}(c(t))\dot{q}^i(t)\dot{q}^j(t) = -g^{ki}(c(t))\frac{\partial V}{\partial q^i} + \bar{\lambda}_a(t)g^{ki}(c(t))\mu^a_i(c(t)) \ ,$$
$$\mu^a_i(c(t))\dot{q}^i(t) = 0 \ .$$

4 Geometric Nonholonomic Integrator – GNI

Given a nonholonomic system $(L, \mathcal{D})$ where L is a Lagrangian system of mechanical type (3), using the metric $\mathcal{G}$, we may consider the complementary projectors

$$\mathcal{P}\colon TQ \to \mathcal{D} \hookrightarrow TQ$$
$$\mathcal{Q}\colon TQ \to \mathcal{D}^\perp \hookrightarrow TQ$$

and their duals considered as mappings from T^*Q to T^*Q.

The Geometric Nonholonomic integrator (GNI, in the sequel) for a nonholonomic system only needs to fix a discrete Lagrangian $L_d : Q \times Q \to \mathbb{R}$ to derive a numerical scheme, that is, it is not necessary to discretize the nonholonomic constraints for this type of integrator. The ***discrete nonholonomic equations*** proposed in [12] are

$$\mathcal{P}^*_{|q_k}(D_1L_d(q_k, q_{k+1})) + \mathcal{P}^*_{|q_k}(D_2L_d(q_{k-1}, q_k)) = 0 \tag{5a}$$
$$\mathcal{Q}^*_{|q_k}(D_1L_d(q_k, q_{k+1})) - \mathcal{Q}^*_{|q_k}(D_2L_d(q_{k-1}, q_k)) = 0. \tag{5b}$$

The first equation is the projection of the discrete Euler–Lagrange equations to the dual of the constraint distribution $\mathcal{D}$, while the second one can be interpreted as an elastic impact of the system against $\mathcal{D}$. This defines a unique discrete evolution operator if and only if the Lagrangian L_d is regular, in the sense of Section 2.

Define the pre- and post-momenta using the discrete Legendre transformations:

$$p^+_{k-1,k} = \mathbb{F}^+L_d(q_{k-1}, q_k) = (q_k, D_2L_d(q_{k-1}, q_k)) \in T^*_{q_k}Q$$
$$p^-_{k,k+1} = \mathbb{F}^-L_d(q_k, q_{k+1}) = (q_k, -D_1L_d(q_k, q_{k+1})) \in T^*_{q_k}Q.$$

In these terms, equation (5b) can be rewritten as

$$\mathcal{Q}^*_{|q_k}\left(\frac{p^-_{k,k+1} + p^+_{k-1,k}}{2}\right) = 0$$

which means that the average of post- and pre-momenta satisfies the nonholonomic constraints.

We can also rewrite the discrete nonholonomic equations as a jump of momenta:

$$p^-_{k,k+1} = (\mathcal{P}^* - \mathcal{Q}^*)\big|_{q_k}(p^+_{k-1,k}). \tag{6}$$

Reversibility. Note that the map $\mathcal{S} = \mathcal{P}^* - \mathcal{Q}^*$ is an involution, that is $\mathcal{S}^{-1} = \mathcal{S}$. Therefore, it acts equivalently in both directions, i.e. it creates a reversible and symmetric flow. Furthermore, it can be expressed as

$$\mathcal{S}(q) = U(q)DU^{-1}(q),$$

where D is a diagonal matrix with elements ± 1 corresponding to the eigenvalues of $\mathcal{S}$ while U is an invertible matrix with columns the eigenvectors of S. Thus, the update (6) can be written as

$$U^{-1}(q_k)p^-_{k,k+1} = DU^{-1}(q_k)p^+_{k-1,k} \tag{7}$$

based on which one can regard the momentum as either remaining unchanged (corresponding to $+1$ eigenvalues) or being reflected (corresponding to -1 eigenvalues) with respect to the basis defined by the mapping U^{-1}.

Preservation Properties. Suppose that Q is a manifold on which a Lie group G acts. Define for each $q \in Q$

$$\mathfrak{g}^q = \{\xi \in \mathfrak{g} \,|\, \xi_Q(q) \in \mathcal{D}_q\}\,, \tag{8}$$

where $\xi_Q(q)$ is the infinitesimal generator vector field corresponding to $\xi \in \mathfrak{g}$ at the point q. The bundle over Q whose fiber at q is $\mathfrak{g}^q$ is denoted by $\mathfrak{g}^{\mathcal{D}}$. Define the discrete nonholonomic momentum map $J_d^{\mathrm{nh}}\colon Q \times Q \to (\mathfrak{g}^{\mathcal{D}})^*$ as in [8] by

$$\begin{aligned} J_d^{\mathrm{nh}}(q_{k-1}, q_k)\colon \mathfrak{g}^{q_k} &\to \mathbb{R} \\ \xi &\mapsto \langle D_2 L_d(q_{k-1}, q_k), \xi_Q(q_k)\rangle\,. \end{aligned}$$

For any smooth section $\widetilde{\xi}$ of $\mathfrak{g}^{\mathcal{D}}$ we have a function $(J_d^{\mathrm{nh}})_{\widetilde{\xi}}\colon Q \times Q \to \mathbb{R}$, defined as $(J_d^{\mathrm{nh}})_{\widetilde{\xi}}(q_{k-1}, q_k) = J_d^{\mathrm{nh}}(q_{k-1}, q_k)\left(\widetilde{\xi}(q_k)\right)$.

If L_d is G-invariant and $\xi \in \mathfrak{g}$ is a horizontal symmetry (that is, $\xi_Q(q) \in \mathcal{D}_q$ for all $q \in Q$), then the GNI preserves $(J_d^{\mathrm{nh}})_\xi$ (see [12] for a proof).

In some cases of interest, it is possible to obtain an integrator preserving energy applying the following theorem (see [12]):

Theorem 1 *Let the configuration manifold be a Lie group with a bi-invariant Lagrangian and with an arbitrary distribution $\mathcal{D}$, and take a discrete Lagrangian that is left-invariant. Then the GNI* (5) *is energy-preserving.*

4.1 Nonholonomic version of the RATTLE and SHAKE methods

Consider a continuous nonholonomic system determined by the mechanical Lagrangian $L\colon \mathbb{R}^{2n} \to \mathbb{R}$:

$$L(q, \dot{q}) = \frac{1}{2}\dot{q}^T M \dot{q} - V(q)$$

(with M a constant, invertible matrix) and the constraints determined by $\mu(q)\dot{q} = 0$ where $\mu(q)$ is a $m \times n$ matrix with rank $\mu = m$.

Consider now the symmetric discretization

$$L_d(q_k, q_{k+1}) = \frac{1}{2}hL\left(q_k, \frac{q_{k+1}-q_k}{h}\right) + \frac{1}{2}hL\left(q_{k+1}, \frac{q_{k+1}-q_k}{h}\right)$$
$$= \frac{1}{2h}\left(q_{k+1}-q_k\right)^T M\left(q_{k+1}-q_k\right) - \frac{h}{2}\left(V(q_k)+V(q_{k+1})\right).$$

After some straightforward computations we obtain that equations (5a) and (5a) for the proposed nonholonomic discrete system are

$$q_{k+1} - 2q_k + q_{k-1} = -h^2 M^{-1}\left(V_q(q_k) + \mu^T(q_k)\lambda_k\right) \tag{9a}$$
$$0 = \mu(q_k)\left(\frac{q_{k+1}-q_{k-1}}{2h}\right), \tag{9b}$$

where λ_k are Lagrange multipliers. We recognize this set of equations as an obvious extension of the SHAKE method proposed by [26] to the case of nonholonomic constraints.

The momentum is approximated by $p_k = M(q_{k+1}-q_{k-1})/2h$. Denoting $p_{k+1/2} = M(q_{k+1}-q_k)/h$, equations (9a) and (9b) are now rewritten in the form

$$p_{k+1/2} = p_k - \frac{h}{2}\left(V_q(q_k) + \mu^T(q_k)\lambda_k\right),$$
$$q_{k+1} = q_k + hM^{-1}p_{k+1/2},$$
$$0 = \mu(q_k)M^{-1}p_k.$$

The definition of p_{k+1} requires the knowledge of q_{k+2} and, therefore, it is is natural to apply another step of the algorithm (9a) and (9b) to avoid this difficulty. Then, we obtain the new equations:

$$p_{k+1} = p_{k+1/2} - \frac{h}{2}\left(V_q(q_{k+1}) + \mu^T(q_{k+1})\lambda_{k+1}\right),$$
$$0 = \mu(q_{k+1})M^{-1}p_{k+1}.$$

The interesting result is that we obtain a natural extension of the RATTLE algorithm for holonomic systems to the case of nonholonomic systems. Unifying the equations above we obtain the following numerical scheme

$$p_{k+1/2} = p_k - \tfrac{h}{2}\left(V_q(q_k) + \mu^T(q_k)\lambda_k\right), \tag{10a}$$
$$q_{k+1} = q_k + hM^{-1}p_{k+1/2}, \tag{10b}$$
$$0 = \mu(q_k)M^{-1}p_k, \tag{10c}$$
$$p_{k+1} = p_{k+1/2} - \tfrac{h}{2}\left(V_q(q_{k+1}) + \mu^T(q_{k+1})\lambda_{k+1}\right), \tag{10d}$$
$$0 = \mu(q_{k+1})M^{-1}p_{k+1}. \tag{10e}$$

These equations allow us to take a triple (q_k, p_k, λ_k) satisfying the constraint equations (10c), compute $p_{k+1/2}$ using (10a) and then q_{k+1} using (10b). Then, equations (10d) and (10e) are used to compute the remaining components of the triple $(q_{k+1}, p_{k+1}, \lambda_{k+1})$. It is clear, applying Theorem 1 that, in the case $V = 0$, the numerical method is energy preserving.

Remark 1 From this Hamiltonian point of view, we have shown that the initial conditions for this numerical scheme are constrained in a natural way ((q_0, p_0) with $\mu(q_0)M^{-1}p_0 = 0$), that is, the initial conditions are exactly the same as those for the continuous system. Additionally, we select $\lambda_0 = 0$ (see [12]).

In [11], the following theorem is proven.

Theorem 2 *The nonholonomic RATTLE method is globally second-order convergent.*

4.2 Projected Version of the Nonholonomic RATTLE

The proposed nonholonomic RATTLE method can be expressed without the use Lagrangian multipliers by projecting the equations of motion onto the constraint distribution through the projection $\mathcal{P}$ defined in §4.

Assuming that the Lagrangian is regular and that matrix μ is full rank (i.e. rank m) (9) can be reformulated as

$$\mathcal{P}(q_k)^T M\left(q_{k+1} - 2q_k + q_{k-1}\right) = -h^2 \mathcal{P}(q_k)^T V_q(q_k) \tag{11a}$$

$$\mathcal{Q}(q_k)^T M\left(\frac{q_{k+1} - q_{k-1}}{2h}\right) = 0, \tag{11b}$$

where the $n \times n$ matrices $\mathcal{Q}$ and $\mathcal{P}$ represent both orthogonal projectors and have rank m and $(n - m)$, respectively, and are defined by

$$\mathcal{Q}(q) = M^{-1}\mu(q)^T\left(\mu(q)M^{-1}\mu(q)^T\right)^{-1}\mu(q), \tag{12a}$$

$$\mathcal{P}(q) = \mathrm{Id} - \mathcal{Q}(q), \tag{12b}$$

where Id is the identity matrix.

Eqs. (11) correspond to (5) for the case $Q = \mathbb{R}^n$ and furthermore can be put in the "momentum jump" form by adding (11a) and (11b) to get

$$q_{k+1} = q_k + \left(\mathrm{Id} - 2M^{-1}\mathcal{Q}(q_k)^T M\right)\left(q_k - q_{k-1}\right) - h^2 M^{-1}\mathcal{P}(q_k)^T V_q(q_k). \tag{13}$$

For a more realistic example, we can add control inputs $u \in U \subset \mathbb{R}^c$ acting in the basis defined by the $(n \times c)$ matrix $B(q)$ to obtain the following discrete equations:

$$q_{k+1} = q_k + \left(\mathrm{Id} - 2M^{-1}\mathcal{Q}(q_k)^T M\right)\left(q_k - q_{k-1}\right) + h^2 M^{-1}\mathcal{P}(q_k)^T f(q_k, u_k),$$

where the forces $f : Q \times U \to T^*Q$ are given by $f(q, u) = B(q)u - V_q(q)$.

In terms of momentum variables the integrator can be equivalently expressed as

$$p_{k+1/2} = \left(\mathrm{Id} - 2\mathcal{Q}(q_k)^T\right) p_{k-1/2} + h\mathcal{P}(q_k)^T f(q_k, u_k) \tag{14a}$$

$$q_{k+1} = q_k + hM^{-1}p_{k+1/2} \tag{14b}$$

providing an update scheme $(q_k, p_{k-1/2}) \Rightarrow (q_{k+1}, p_{k+1/2})$.

A remaining critical step in completing the algorithm is to establish the link between the discrete variables $(q_k, p_{k+1/2})$ for $k = 0, ..., N$ used in (14) and the continuous curve $(q(t), p(t))$. In that respect one can regard $p_k = (p_{k-1/2} + p_{k+1/2})/2$ as an approximation to the continuous momentum at time $t = kh$, i.e. $p_k \approx p(kh)$. The pair (q_k, p_k) satisfies the nonholonomic constraint by definition and is related, following from (14), to the "midpoint" momenta through

$$p_k = \mathcal{P}(q_k)^T p_{k+1/2} - \frac{h}{2}\mathcal{P}(q_k)^T f(q_k, u_k), \tag{15a}$$

$$p_k = \mathcal{P}(q_k)^T p_{k-1/2} + \frac{h}{2}\mathcal{P}(q_k)^T f(q_k, u_k). \tag{15b}$$

These expressions can be used to determine proper variables $(q_1, p_{1/2})$ to initialize the update (14) given continuous initial conditions $(q_0, p_0) \approx (q(0), p(0))$. Since there is a set of solutions $p_{1/2}$ satisfying (15) for a given p_0 the most natural choice is to pick $p_{1/2}$ satisfying the constraints at q_0. Therefore, the condition becomes

$$p_{1/2} = p_0 + \frac{h}{2}\mathcal{P}(q_0)^T f(q_0, u_0).$$

In summary, given initial conditions (q_0, p_0) satisfying the constraints, the dynamics is evolved forwards to reach the final state (q_N, p_N), also in the constraint submanifold, after N time steps through

$$\begin{aligned} p_{1/2} &= p_0 + \frac{h}{2}\mathcal{P}(q_0)^T f(q_0, u_0), \\ p_{k+1/2} &= \left(\mathrm{Id} - 2\mathcal{Q}(q_k)^T\right) p_{k-1/2} + h\mathcal{P}(q_k)^T f(q_k, u_k), \\ q_{k+1} &= q_k + hM^{-1} p_{k+1/2}, \\ p_N &= \mathcal{P}(q_N)^T p_{N-1/2} + \frac{h}{2}\mathcal{P}(q_N) f(q_N, u_N), \end{aligned} \tag{16}$$

for $k = 1, ..., N-1$.

5 Reduced d'Alembert-Pontryagin integrator-(RDP)

In this section we consider a class of mechanical systems which, in addition to nonholonomic constraints, also possess symmetries of motion arising from conservation laws. The interplay between the constraints and symmetries is linked to an intrinsic structure of the state space associated with important properties of the dynamics. Our goal in this section is to develop integrators that respect this structure and lead to more faithful numerical representation.

In §4 we introduced the action of a symmetry group G and its relation the evolution of the system momentum. Additional structure arises whenever the dynamics and constraints are G-invariant that permits the construction of *reduced* nonholonomic integrators [14, 17].

Following [1], define the subspaces $\mathcal{V}_q$ and $\mathcal{S}_q$ according to

$$\mathcal{V}_q = \{\xi_Q(q) \mid \xi \in \mathfrak{g}\}, \qquad \mathcal{S}_q = \mathcal{D}_q \cap \mathcal{V}_q.$$

Practically speaking, the *vertical* space $\mathcal{V}_q$ represents the space of tangent vectors parallel to symmetry directions while $\mathcal{S}_q$ is the space of symmetry directions that satisfy the constraints. Equivalently, $\mathcal{S}_q$ can be regarded as the space generated by elements in $\mathfrak{g}^q$, as defined in (8). The group G is chosen so that the Lagrangian L and distribution $\mathcal{D}$ are G-invariant. In addition, we make the standard assumption (see [1, 7]) that $T_qQ = \mathcal{D}_q + \mathcal{V}_q$, for each $q \in Q$.

Since our main interest is in a configuration space that is by construction of the form $Q = M \times G$ we will restrict any further derivations to the *trivial bundle* case. Using coordinates $(r, g) \in M \times G$ a basis for $\mathfrak{g}^q$ can be chosen as $\{e_b(r, g)\}$, for $b = 1, ..., \dim(\mathcal{S})$. Since $\mathcal{D}$ is G-invariant these elements can be expressed as $e_b(r, g) = \mathrm{Ad}_g\, e_b(r)$, where $\{e_b(r)\}$ is the body-fixed basis. We denote $\mathfrak{g}^r$ the space spanned by $\{e_b(r)\}$ at each $(r, e) \in Q$. Lastly, the system is subject to control force $f : [0, T] \to T^*M$ restricted to the shape space.

Nonholonomic Connection With these definitions we can define a principal connection $\mathcal{A} : TQ \to \mathfrak{g}$ with horizontal distribution that coincides with $\mathcal{H}_q$ at the point q, where $\mathcal{D}_q = \mathcal{S}_q \oplus \mathcal{H}_q$. This connection is called the *nonholonomic connection* and is constructed according to $\mathcal{A} = \mathcal{A}^{\mathrm{kin}} + \mathcal{A}^{\mathrm{sym}}$, where $\mathcal{A}^{\mathrm{kin}}$ is the kinematic connection enforcing the nonholonomic constraints and $\mathcal{A}^{\mathrm{sym}}$ is the mechanical connection corresponding to symmetries satisfying the constraints. These maps are defined according to

$$\begin{aligned} \mathcal{A}^{\mathrm{kin}}(q) \cdot \dot{q} &= 0, \\ \mathcal{A}^{\mathrm{sym}}(q) \cdot \dot{q} &= \mathrm{Ad}_g\, \Omega, \end{aligned} \tag{17}$$

where $\Omega \in \mathfrak{g}^r$ is called the *locked angular velocity*, i.e. the velocity resulting from instantaneously locking the joints described by the variables r. Intuitively, when the joints stop moving the system continues its motion uniformly along a curve (with tangent vectors in $\mathcal{S}$) with body-fixed velocity Ω and a corresponding spatial momentum that is conserved.

By definition the principal connection can be expressed as

$$\mathcal{A}(q) \cdot \dot{q} = \mathrm{Ad}_g(g^{-1}\dot{g} + \mathcal{A}(r)\dot{r}),$$

where $\mathcal{A}(r)$ is the local form and the two components in (17) can be added to obtain

$$g^{-1}\dot{g} + \mathcal{A}(r)\dot{r} = \Omega.$$

Numerical Formulation Since the Lagrangian is G-invariant, we can define the *reduced Lagrangian* $\ell : TM \times \mathfrak{g} \to \mathbb{R}$

$$\ell(r, \dot{r}, \xi) = L(r, \dot{r}, e, g^{-1}\dot{g}). \tag{18}$$

In [17] a nonholonomic integrator was derived using a discrete variational d'Alembert-Pontryagin principle based on the reduced Lagrangian ℓ, the connection $\mathcal{A}$ and a chosen trajectory discretization. In particular, a discrete trajectory with points $q_k = (r_k, g_k) \in M \times G$ and respective velocities $u_k \in TM$ and $\xi_k \in \mathfrak{g}$ was constructed so that

$$r_{k+1} - r_k = hu_k, \qquad \tau^{-1}(g_k^{-1} g_{k+1}) = h\xi_k,$$

where $\xi_k = \Omega_k - \mathcal{A}(r_{k+\alpha})u_k$, with $r_{k+\alpha} := (1-\alpha)r_k + \alpha r_{k+1}$ for a chosen $\alpha \in [0,1]$. The map $\tau : \mathfrak{g} \to G$ represents the *difference* between two configurations in the group by an element in its algebra and can be selected as:

- Exponential map $\exp : \mathfrak{g} \to G$, defined by $\exp(\xi) = \gamma(1)$, with $\gamma : \mathbb{R} \to G$ is the integral curve through the identity of the left invariant vector field associated with $\xi \in \mathfrak{g}$ (hence, with $\dot{\gamma}(0) = \xi$);
- Canonical coordinates of the second kind $\mathrm{ccsk} : \mathfrak{g} \to G$, $\mathrm{ccsk}(\xi) = \exp(\xi^1 e_1) \cdot \exp(\xi^2 e_2) \cdot \ldots \cdot \exp(\xi^n e_n)$, where $\{e_i\}$ is the Lie algebra basis.

A third choice for τ, valid only for certain *quadratic* matrix groups [6] (which include the rigid motion groups SO(3), SE(2), and SE(3)), is the *Cayley map* $\mathrm{cay} : \mathfrak{g} \to G$, $\mathrm{cay}(\xi) = (e - \xi/2)^{-1}(e + \xi/2)$. (See App. a for more details).

With these definitions in place the resulting reduced d'Alembert-Pontryagin (RDP) integrator can be stated [17]. For numerical convenience it is given in terms of vector-matrix notation, by treating the Lie algebra variables ξ and Ω as vectors of coordinates with respect to a chosen canonical basis (see App. b for an example).

The discrete flow satisfies the *reduced discrete dynamics*

$$\begin{bmatrix} \mathrm{Id} & [\mathcal{A}(r_k)]^T \\ 0 & [e_1(r_k), \ldots, e_c(r_k)]^T \end{bmatrix} \left(\begin{bmatrix} \partial_u \ell_k \\ (\mathrm{d}\tau^{-1}_{h\xi_k})^* \partial_\xi \ell_k \end{bmatrix} - \begin{bmatrix} \partial_u \ell_{k-1} \\ (\mathrm{d}\tau^{-1}_{-h\xi_{k-1}})^* \partial_\xi \ell_{k-1} \end{bmatrix} \right) = \begin{bmatrix} hf_k \\ 0 \end{bmatrix}, \quad (19)$$

where $\ell_k := \ell(r_{k+\alpha}, u_k, \xi_k)$ and $\xi_k = \Omega_k - \mathcal{A}(r_{k+\alpha})u_k$. The map $\mathrm{d}\tau_\xi : \mathfrak{g} \to \mathfrak{g}$ is the *right-trivialized tangent* of $\tau(\xi)$ defined by $\mathrm{D}\,\tau(\xi) \cdot \delta = TR_{\tau(\xi)}(\mathrm{d}\tau_\xi \cdot \delta)$ and $d\tau_\xi^{-1} : \mathfrak{g} \to \mathfrak{g}$ is its inverse (see App. a).

Equation (19) along with the *reconstruction equations*

$$g_{k+1} = g_k \tau(h\xi_k), \qquad r_{k+1} = r_k + hu_k, \quad (20)$$

constitute the complete RDP discrete evolution.

6 Examples

6.1 The Chaplygin Sleigh

The Chaplygin Sleigh [1] is a planar rigid body making a contact with the ground through a *skate* mounted at the central axis of the body at a distance a from

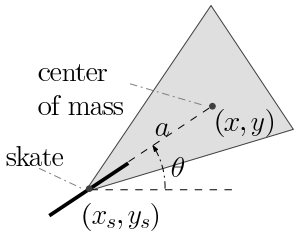


Figure 1: Chaplygin Sleigh model.

its center of mass (Fig. 6.1). The configuration space is the group $G = \mathrm{SE}(2)$ with coordinates $q = (\theta, x, y)$ describing the orientation and the position of the center of mass. The body has rotational inertia I and mass m and, therefore, its Lagrangian is defined by

$$L(q, \dot{q}) = \frac{1}{2} I \dot{\theta}^2 + \frac{1}{2} m(\dot{x}^2 + \dot{y}^2). \tag{21}$$

At the point of the skate contact $(x_s, y_s) = (x - a\cos\theta, y - a\sin\theta)$ the body must slide in the direction in which it is pointing. This condition is encoded by the nonholonomic constraint

$$a\dot{\theta} + \sin\theta\dot{x} - \cos\theta\dot{y} = 0.$$

A structure-preserving integrator was developed in [10] based on the discrete Lagrange-d'Alembert (DLA) principle with discrete momentum and measure preservation properties. Exploring this direction further, in this section we develop two alternative methods based on the GNI and RDP schemes.

GNI Integrator. From the mass matrix $M = \mathrm{diag}(I, m, m)$ and the constraint $\mu_1(q) = [a, \sin\theta, -\cos\theta]$, the projector $\mathcal{Q}$ can be computed using (12a) as

$$\mathcal{Q}(q) = \frac{1}{I + a^2 m} \begin{bmatrix} a^2 m & am\sin\theta & -am\cos\theta \\ aI\sin\theta & I\sin\theta^2 & -I\sin\theta\cos\theta \\ -aI\cos\theta & -I\sin\theta\cos\theta & I\cos^2\theta \end{bmatrix}. \tag{22}$$

Since the mass matrix is constant, the GNI integrator can be derived according to $v_{k+\frac{1}{2}} = (\mathrm{Id} - M^{-1}\mathcal{Q}(q_k)^T M) v_{k-\frac{1}{2}}$. In terms of the coordinates $v = (v^\theta, v^x, v^y)$, the discrete update becomes

$$\begin{aligned}
v^\theta_{k+\frac{1}{2}} &= \left(1 - \frac{2a^2 m}{I'}\right) v^\theta_{k-\frac{1}{2}} + \frac{am}{I'}\left(-2\sin\theta_k v^x_{k-\frac{1}{2}} + 2\cos\theta_k v^y_{k-\frac{1}{2}}\right), \\
v^x_{k+\frac{1}{2}} &= -\frac{2aI}{I'}\sin\theta_k v^\theta_{k-\frac{1}{2}} + \left(1 - \frac{2I}{I'}\sin^2\theta_k\right) v^x_{k-\frac{1}{2}} + \frac{2I}{I'}\sin\theta_k\cos\theta_k v^y_{k-\frac{1}{2}}, \\
v^y_{k+\frac{1}{2}} &= \frac{2aI}{I'}\cos\theta_k v^\theta_{k-\frac{1}{2}} + \frac{2I}{I'}\sin\theta_k\cos\theta_k v^x_{k-\frac{1}{2}} + \left(1 - \frac{2I}{I'}\cos^2\theta_k\right) v^y_{k-\frac{1}{2}},
\end{aligned}$$

where $I' = I + a^2 m$. It is straightforward to verify that the resulting update rule is energy-preserving, i.e. $\langle M v_{k-1/2}, v_{k-1/2} \rangle = \langle M v_{k+1/2}, v_{k+1/2} \rangle$. This property is inherent to the GNI construction as explained in [12].

RDP Integrator. The sleigh has no internal joints and therefore no shape space. Since the Lagrangian (21) is left-invariant to SE(2) group action, the reduced Lagrangian (18) can be expressed as

$$\ell(\xi) = L(e, g^{-1}\dot{g}),$$

where $\xi = (\omega, v, v^{\perp}) \in \mathfrak{g}$ describes the angular, forward, and sideways velocities with respect to the body frame fixed at the center of mass. The constrained symmetry space (8) of the sleigh can be identified as

$$\mathfrak{g}^q = \operatorname{span}\{e_1(g), e_2(g)\},$$

where $e_1 = (1, 0, a) \in \mathfrak{g}$ and $e_2 = (0, 1, 0) \in \mathfrak{g}$ form the constant basis in the body-fixed frame with $e_i(g) = \operatorname{Ad}_g e_i$, for $i = 1, 2$. The two components of the nonholonomic momentum $p_i = \langle \partial_\xi \ell, e_1 \rangle$ become

$$p_1 = (J + a^2 m)\omega, \qquad p_2 = mv,$$

corresponding to angular and forward momenta, respectively. The group trajectory can be reconstructed from the momentum according to

$$g^{-1}\dot{g} = \left(\frac{1}{I + a^2 m} p_1,\ \frac{1}{m} p_2,\ \frac{a}{I + a^2 m} p_1 \right).$$

The momentum components themselves evolve according to $\dot{p}_i = \langle \operatorname{ad}^*_\xi \partial_\xi \ell, e_i \rangle$ (see [2]), or equivalently

$$\dot{p}_1 = -\frac{a}{I + a^2 m} p_1 p_2, \qquad \dot{p}_2 = \frac{ma}{(I + a^2 m)^2} p_1^2.$$

Since the shape space consists of a single point, the discrete dynamics includes only the momentum equations (19) which become

$$\langle (\mathrm{d}\tau^{-1}_{h\xi_k})^* \partial_\xi \ell_k - (\mathrm{d}\tau^{-1}_{-h\xi_{k-1}})^* \partial_\xi \ell_{k-1}, e_i \rangle = 0,$$

for $i = 1, 2$. A simple form of these equations can be derived by choosing $\tau = \exp$ and truncating its tangent to first order, i.e. $\mathrm{d}\tau^{-1}_\xi \approx \mathrm{Id} - \frac{1}{2}\operatorname{ad}_\xi$. Using the notation $p_k = ((p_1)_k, (p_2)_k)$ the update becomes

$$\begin{aligned}
(p_1)_k - (p_1)_{k-1} &= -\frac{ha}{2(J + a^2 m)} \left[(p_1)_k (p_2)_k + (p_1)_{k-1} (p_2)_{k-1} \right], \\
(p_2)_k - (p_2)_{k-1} &= \frac{hma}{2(J + a^2 m)^2} \left[{(p_1)_k}^2 + {(p_1)_{k-1}}^2 \right].
\end{aligned}$$

These conditions are used to solve for the unknown next momentum p_k, e.g. through cubic equation root-finding. Note, that this particular choice

of approximation exactly matches a standard implicit central difference discretization of the continuous ODE. This is generally not case for systems with non-constant Lie algebra basis element e_i such as the snakeboard. Higher accuracy can be achieved through other choices of τ and better approximation of $\mathrm{d}\tau$. App. b details the cases $\tau = \exp$ and $\tau = \mathrm{cay}$ on SE(2).

The reconstruction equations are

$$g_{k+1} = g_k \exp(h\xi_k),$$

where

$$\xi_k = \left(\frac{1}{I + a^2 m}(p_1)_k,\ \frac{1}{m}(p_2)_k,\ \frac{a}{I + a^2 m}(p_1)_k \right).$$

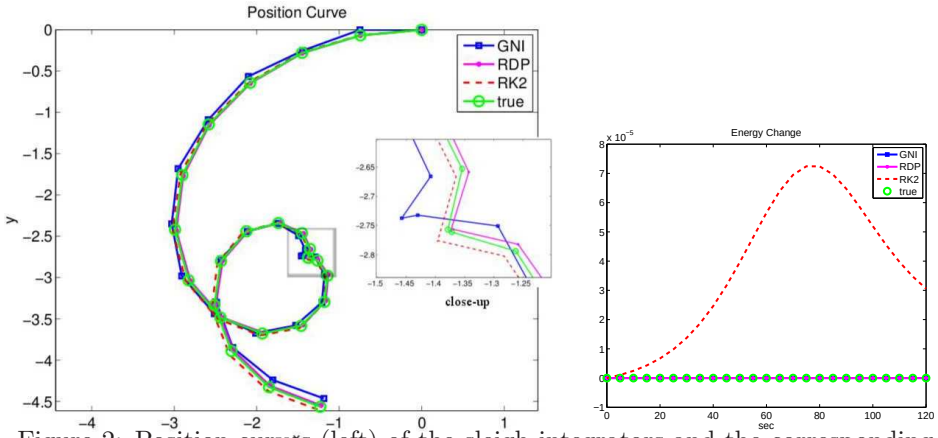


Figure 2: Position curves (left) of the sleigh integrators and the corresponding energy (right). The embedded close-up frame (left) zooms in on the cusp point of the "heart" shape.

Numerical Comparisons. The numerical behavior of the algorithms is now examined in terms of their ability to reproduce the true system trajectory and in terms of their energy preservation. Comparison to a standard Runge-Kutta second-order method is also included.

Note that the standard Chaplygin sleigh model (e.g. [1, 10, 12]) is studied in terms of the coordinates of the skate contact rather than the center off mass as in this work. For easier reference to such previous studies, we present the position curves below in terms of the skate coordinates (x_s, y_s). This representation enables the generation of the familiar "heart"-shaped curves (Fig. 6.1).

6.2 The Snakeboard

The snakeboard (Fig. 3) represents a type of system with an interesting interplay between constraints and symmetries. It has served as a classical example (e.g. [2, 7, 4]) of a system with non-trivial intersection of the constraint

distribution $\mathcal{D}$ and the vertical space $\mathcal{V}$. Our integrators capture the dynamics of such systems and their performance is examined in this section.

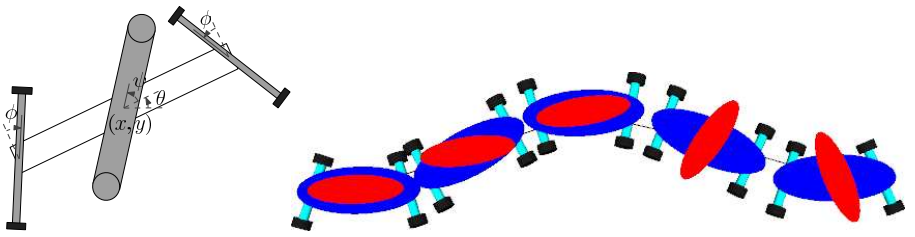


Figure 3: Snakeboard model(left) and a typical trajectory(right).

The shape space variables of the snakeboard are $r = (\psi, \phi) \in \mathrm{S}^1 \times \mathrm{S}^1$ denoting the rotor angle and the steering wheels angle, while its configuration is defined by (θ, x, y) denoting orientation and position of the board. This corresponds to a configuration space $Q = \mathrm{S}^1 \times \mathrm{S}^1 \times \mathrm{SE}(2)$ with shape space $M = \mathrm{S}^1 \times \mathrm{S}^1$ and group $G = \mathrm{SE}(2)$. Additional parameters are its mass m, distance l from its center to the wheels, and moments of inertia I and J of the board and the steering. The kinematic constraints of the snakeboard are:

$$\begin{aligned} -l\cos\phi d\theta - \sin(\theta+\phi)dx + \cos(\theta+\phi)dy = 0, \\ l\cos\phi d\theta - \sin(\theta-\phi)dx + \cos(\theta-\phi)dy = 0, \end{aligned} \tag{23}$$

enforcing the fact that the system must move in the direction in which the wheels are pointing and spinning. The constraint distribution is spanned by three covectors:

$$\mathcal{D}_q = \mathrm{span}\left\{\frac{\partial}{\partial\psi}, \frac{\partial}{\partial\phi}, c\frac{\partial}{\partial\theta} + a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y}\right\},$$

where $a = -2l\cos\theta\cos^2\phi$, $b = -2l\sin\theta\cos^2\phi$, $c = \sin 2\phi$. The group directions defining the vertical space are:

$$\mathcal{V}_q = \mathrm{span}\left\{\frac{\partial}{\partial\theta}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right\},$$

and therefore the constrained symmetry space becomes:

$$\mathcal{S}_q = \mathcal{V}_q \cap \mathcal{D}_q = \mathrm{span}\left\{c\frac{\partial}{\partial\theta} + a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y}\right\}. \tag{24}$$

Since $\mathcal{D}_q = \mathcal{S}_q \oplus \mathcal{H}_q$, we have $\mathcal{H}_q = \mathrm{span}\left\{\frac{\partial}{\partial\psi}, \frac{\partial}{\partial\phi}\right\}$. Finally, the Lagrangian of the system is $L(q,\dot{q}) = \frac{1}{2}\dot{q}^T M \dot{q}$ where

$$\mathrm{M} = \begin{bmatrix} I & 0 & I & 0 & 0 \\ 0 & 2J & 0 & 0 & 0 \\ I & 0 & ml^2 & 0 & 0 \\ 0 & 0 & 0 & m & 0 \\ 0 & 0 & 0 & 0 & m \end{bmatrix}.$$

The reduced Lagrangian can be expressed as $\ell(r,u,\xi) = (u,\xi)^T \mathrm{M}\ (u,\xi)$ by treating the velocity ξ as a vector in the standard $\mathfrak{se}(2)$ basis (defined in App. b).

There is only one direction along which snakeboard motions lead to momentum conservation: it is defined by the basis element

$$e_1(r) = 2l\cos^2\phi \begin{bmatrix} \frac{\tan\phi}{l} \\ -1 \\ 0 \end{bmatrix},$$

and, hence, there is only one momentum variable $p_1 = \langle \frac{\partial \ell}{\partial \xi}, e_1(r) \rangle$. Using this variable we can derive the connection according to [2] as

$$[\mathcal{A}] = \begin{bmatrix} \frac{I}{ml^2}\sin^2\phi & 0 \\ -\frac{I}{2ml}\sin 2\phi & 0 \\ 0 & 0 \end{bmatrix}, \text{and} \quad \Omega = \frac{p_1}{4ml^2\cos^2\phi} e_1(r).$$

GNI Integrator. The snakeboard constraints (23) can be expressed in terms of the one-forms

$$\mu_1(q) = (0,0,a,-c,0), \qquad \mu_2(q) = (0,0,b,0,-c).$$

The projector $\mathcal{Q}$ can then be computed from μ and the mass matrix M using (12a) to obtain

$$\mathcal{Q}(q) = \frac{1}{ml^2 - I\sin^2\phi} \begin{pmatrix} 0 & 0 & -m(a^2+b^2) & mac & mbc \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & m(a^2+b^2) & -mac & -mbc \\ 0 & 0 & -I'ac & mb^2+I'c^2 & -mab \\ 0 & 0 & -I'bc & -mab & ma^2+I'c^2 \end{pmatrix},$$

where $I' = ml^2 - I$ and $q = (\psi, \phi, \theta, x, y)$. Similarly to the Chaplygin sleigh §6.1, since the mass matrix is constant, the discrete dynamics is updated explicitly through $v_{k+\frac{1}{2}} = (\mathrm{Id} - M^{-1}\mathcal{Q}(q_k)^T M) v_{k-\frac{1}{2}}$.

RDP Integrator. The reduced discrete equations of motion will be derived by substituting the Lagrangian and the connection of the snakeboard into (19) and choosing the map $\tau = \exp$. Since, particularly for the snakeboard, $\mathfrak{s}$ is one dimensional and $\mathcal{A}(r)\cdot\delta$ is parallel to $e_1(r)$ for any $\delta \in T_r M$ the discrete dynamics simplifies (see [17, 16]) to

$$\langle p_k - p_{k-1}, e_1(r_k) \rangle = 0, \qquad \partial_u \ell_{k+\alpha} - \partial_u \ell_{k-1+\alpha} = 0,$$

where

$$p_k = (ml^2\xi_k^1 + Iu_k^\phi, m\xi_k^2, 0), \qquad \partial_u \ell_k = (I(u_k^\psi + \xi_k^1), 2Ju_k^\phi),$$

and the dynamics is derived by expressing $\xi_k = \Omega_k - \mathcal{A}(r_{k+\alpha})\cdot u_k$ in terms of $r_k = (\psi_k, \phi_k)$, $u_k = (u_k^\psi, u_k^\phi)$, and $(p_1)_k$. Note that the discrete dynamics is

linear in the unknowns u_k and $(p_1)_k$ and results in an efficient explicit integrator. The reconstruction equations are

$$g_{k+1} = g_k \exp(h\xi_k), \qquad r_{k+1} - r_k = hu_k.$$

Numerical Behavior. The studied snakeboard integrators are second-order methods. Their advantage over similar methods is shown through comparison to a typical second order Runge-Kutta method as well as to the actual true trajectory. Fig. 6.2 shows a trajectory with initial conditions $\psi(0) = \pi/2$, $\phi(0) = \pi/3$, $p_1(0) = -1$, $\dot{\psi}(0) = 2.5$, $\dot{\phi}(0) = -0.02$, $\theta(0) = 0$. Sinusoidal control inputs $u_\psi = \cos(20\pi t)$, $u_\phi = \sin(2\pi t)$ at the joints were used to create parallel parking maneuvers with cusp points. The CPU run-times of the compared methods are nearly identical and are not included in the plots.

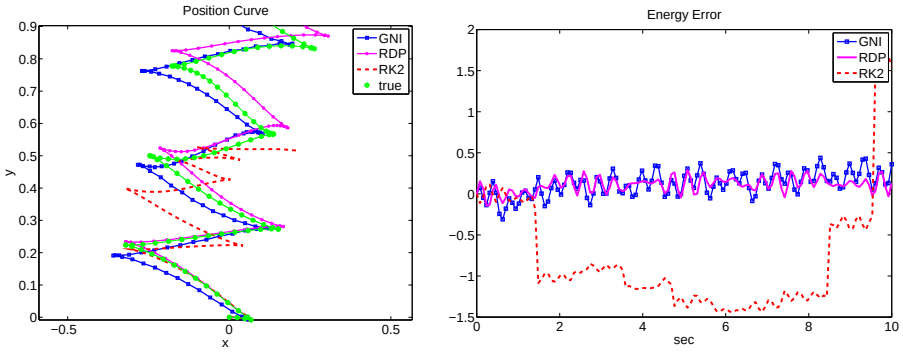


Figure 4: Snakeboard integrator numerics with $N = 128$ timesteps over the integration horizon $T = 10$ sec. At such coarse resolution RK2 method fails to follow the true trajectory while GNI and RDP have qualitatively correct behavior (position curve on left). One likely explanation lies in their better energy behavior (shown on right).

In special cases, for particular combinations of initial conditions and inertial parameters, the GNI integrator has shown non-physical oscillatory behavior. While this issue is most likely related to instabilities known to occur in projection-based methods, the exact cause remains to be determined in future work.

7 Conclusion

In this paper we have compared two geometric integrators for nonholonomic dynamics, the so-called GNI and RDP integrators. Both are constructed using differential geometric tools developed by geometric mechanics community through a careful study of nonholonomic dynamics during the last twenty years. This paper shows the importance of combining different research areas

(differential geometry, numerical analysis and mechanics) to produce methods with an extraordinary qualitative and quantitative behavior.

Such issues raise a number of future work directions. We therefore close with some open questions:

- Given one of the nonholonomic integrators (GNI or RDP), does there exist, in the sense of backward error analysis, a continuous nonholonomic system, such that the discrete evolution for the nonholonomic integrator is the flow of this nonholonomic system up to an appropriate order?
- Is it possible to use the nonholonomic Hamilton-Jacobi theory recently developed [15, 19] for the construction of these methods or new ones?

These questions will be part of the future work that we will develop in the next years.

Appendix

a Retraction map tangents

The two common choices for retraction maps are the exponential map $\tau = \exp$ and the Cayley map $\tau = \mathrm{cay}$. In this section we provide their right-trivialized tangents $\mathrm{d}\,\tau$ of these maps and their inverses $\mathrm{d}\,\tau^{-1}$ (see [3] for more details).

a.1 Exponential map

The right-trivialized derivative of the map exp and its inverse are defined as

$$\mathrm{dexp}_x\, y = \sum_{j=0}^{\infty} \frac{1}{(j+1)!}\, \mathrm{ad}_x^j\, y, \quad \mathrm{dexp}_x^{-1}\, y = \sum_{j=0}^{\infty} \frac{B_j}{j!}\, \mathrm{ad}_x^j\, y, \tag{25}$$

where B_j are the Bernoulli numbers. Typically, these expressions are truncated in order to achieve a desired order of accuracy. The first few Bernoulli numbers are $B_0 = 1$, $B_1 = -1/2$, $B_2 = 1/6$, $B_3 = 0$ (see [6, 13] for more details).

a.2 Cayley map

The derivative maps become (see [13] for derivation)

$$\mathrm{dcay}_x\, y = \left(\mathrm{I} - \frac{x}{2}\right)^{-1} y \left(\mathrm{I} + \frac{x}{2}\right)^{-1}, \quad \mathrm{dcay}_x^{-1}\, y = \left(\mathrm{I} - \frac{x}{2}\right) y \left(\mathrm{I} + \frac{x}{2}\right). \tag{26}$$

b Retraction Maps on SE(2)

The coordinates of SE(2) are (θ, x, y) with matrix representation $g \in \mathrm{SE}(2)$ given by:

$$g = \begin{bmatrix} \cos\theta & -\sin\theta & x \\ \sin\theta & \cos\theta & y \\ 0 & 0 & 1 \end{bmatrix}. \tag{27}$$

Using the isomorphic map $\widehat{\cdot}: \mathbb{R}^3 \to \mathfrak{se}(2)$ given by:

$$\widehat{v} = \begin{bmatrix} 0 & -v^1 & v^2 \\ v^1 & 0 & v^3 \\ 0 & 0 & 0 \end{bmatrix} \text{ for } v = \begin{pmatrix} v^1 \\ v^2 \\ v^3 \end{pmatrix} \in \mathbb{R}^3,$$

$\{\widehat{e}_1, \widehat{e}_2, \widehat{e}_3\}$ can be used as a basis for $\mathfrak{se}(2)$, where $\{e_1, e_2, e_3\}$ is the standard basis of $\mathbb{R}^3$.

The two maps $\tau : \mathfrak{se}(2) \to \mathrm{SE}(2)$ are given by

$$\exp(\widehat{v}) = \begin{cases} \begin{bmatrix} \cos v^1 & -\sin v^1 & \frac{v^2 \sin v^1 - v^3(1-\cos v^1)}{v^1} \\ \sin v^1 & \cos v^1 & \frac{v^2(1-\cos v^1)+v^3 \sin v^1}{v^1} \\ 0 & 0 & 1 \end{bmatrix} & \text{if } v^1 \neq 0 \\ \begin{bmatrix} 1 & 0 & v^2 \\ 0 & 1 & v^3 \\ 0 & 0 & 1 \end{bmatrix} & \text{if } v^1 = 0 \end{cases}$$

$$\mathrm{cay}(\widehat{v}) = \begin{bmatrix} \frac{1}{4+(v^1)^2} \begin{bmatrix} (v^1)^2 - 4 & -4v^1 & -2v^1v^3 + 4v^2 \\ 4v^1 & (v^1)^2 - 4 & 2v^1v^2 + 4v^3 \end{bmatrix} \\ 0 \qquad\qquad 0 \qquad\qquad 1 \end{bmatrix}$$

The maps $[\mathrm{d}\tau_\xi^{-1}]$ can be expressed as the 3×3 matrices:

$$[\mathrm{dexp}_{\widehat{v}}^{-1}] \approx \mathbf{I}_3 - \frac{1}{2}[\mathrm{ad}_v] + \frac{1}{12}[\mathrm{ad}_v]^2, \tag{28}$$

$$[\mathrm{dcay}_{\widehat{v}}^{-1}] = \mathbf{I}_3 - \frac{1}{2}[\mathrm{ad}_v] + \frac{1}{4}\begin{bmatrix} v^1 \cdot v & \mathbf{0}_{3\times 2} \end{bmatrix}, \tag{29}$$

where

$$[\mathrm{ad}_v] = \begin{bmatrix} 0 & 0 & 0 \\ v^3 & 0 & -v^1 \\ -v^2 & v^1 & 0 \end{bmatrix}.$$

References

[1] Anthony M. Bloch. *Nonholonomic mechanics and control*, volume 24 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, 2003.

[2] Anthony M. Bloch, P. S. Krishnaprasad, Jerrold E. Marsden, and Richard M. Murray. Nonholonomic mechanical systems with symmetry. *Arch. Rational Mech. Anal.*, 136(1):21–99, 1996.

[3] Nawaf Bou-Rabee and Jerrold E. Marsden. Hamilton-Pontryagin integrators on Lie groups. I. Introduction and structure-preserving properties. *Found. Comput. Math.*, 9(2):197–219, 2009.

[4] Francesco Bullo and Andrew D. Lewis. *Geometric control of mechanical systems*, volume 49 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2005. Modeling, analysis, and design for simple mechanical control systems.

[5] Frans Cantrijn, Jorge Cortés, Manuel de León, and David Martín de Diego. On the geometry of generalized Chaplygin systems. *Math. Proc. Cambridge Philos. Soc.*, 132(2):323–351, 2002.

[6] Elena Celledoni and Brynjulf Owren. Lie group methods for rigid body dynamics and time integration on manifolds. *Comput. Methods Appl. Mech. Engrg.*, 192(3-4):421–438, 2003.

[7] Hernán Cendra, Jerrold E. Marsden, and Tudor S. Ratiu. Geometric mechanics, Lagrangian reduction, and nonholonomic systems. In *Mathematics unlimited—2001 and beyond*, pages 221–273. Springer, Berlin, 2001.

[8] Jorge Cortés and Sonia Martínez. Non-holonomic integrators. *Nonlinearity*, 14(5):1365–1392, 2001.

[9] Jorge Cortés Monforte. *Geometric, control and numerical aspects of nonholonomic systems*, volume 1793 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2002.

[10] Yuri N. Fedorov and Dmitry V. Zenkov. Discrete nonholonomic LL systems on Lie groups. *Nonlinearity*, 18(5):2211–2241, 2005.

[11] S. Ferraro, D. Iglesias, and D. Martín de Diego. Numerical and geometric aspects of the nonholonomic shake and rattle methods. In *AIMS Proceedings*. (To appear).

[12] S. Ferraro, D. Iglesias, and D. Martín de Diego. Momentum and energy preserving integrators for nonholonomic dynamics. *Nonlinearity*, 21(8):1911–1928, 2008.

[13] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2002.

[14] David Iglesias, Juan C. Marrero, David Martín de Diego, and Eduardo Martínez. Discrete nonholonomic Lagrangian systems on Lie groupoids. *J. Nonlinear Sci.*, 18(3):221–276, 2008.

[15] David Iglesias-Ponte, Manuel de León, and David Martín de Diego. Towards a Hamilton-Jacobi theory for nonholonomic mechanical systems. *J. Phys. A*, 41(1):015205, 14, 2008.

[16] Marin Kobilarov, Keenan Crane, and Mathieu Desbrun. Lie group integrators for animation and control of vehicles. *ACM Transactions on Graphics*, 28(2), April 2009.

[17] Marin Kobilarov, Jerrold Marsden, and Gaurav Sukhatme. Geometric discretization of nonholonomic systems with symmetries. *Discrete Contin. Dyn. Syst. Ser. S*, 3(1):61–84, 2010.

[18] Jair Koiller. Reduction of some classical nonholonomic systems with symmetry. *Arch. Rational Mech. Anal.*, 118(2):113–148, 1992.

[19] Manuel de León, Juan C. Marrero, and David Martín de Diego. Linear almost Poisson structures and Hamilton-Jacobi theory. Applications to nonholonomic mechanics. 2008. Preprint, arXiv:0801.4358.

[20] Manuel de León and David Martín de Diego. On the geometry of nonholonomic Lagrangian systems. *J. Math. Phys.*, 37(7):3389–3414, 1996.

[21] Juan C. Marrero, David Martín de Diego, and Eduardo Martínez. Discrete Lagrangian and Hamiltonian mechanics on Lie groupoids. *Nonlinearity*, 19(6):1313–1348, 2006. Corrigendum: Nonlinearity **19** (2006), no. 12, 3003–3004.

[22] J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numer.*, 10:357–514, 2001.

[23] R. McLachlan and M. Perlmutter. Integrators for nonholonomic mechanical systems. *J. Nonlinear Sci.*, 16(4):283–328, 2006.

[24] Yuri I. Neĭmark and Nikolai A. Fufaev. *Dynamics of Nonholonomic Systems*. Translations of Mathematical Monographs, Vol. 33. American Mathematical Society, Providence, R.I., 1972.

[25] George W. Patrick. Lagrangian mechanics without ordinary differential equations. *Rep. Math. Phys.*, 57(3):437–443, 2006.

[26] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraint: molecular dynamics of n-alkanes. *J. Comput. Physics*, 23:327–341, 1977.

[27] J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian problems*, volume 7 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London, 1994.

[28] Anatoly M. Vershik and L. D. Faddeev. Differential geometry and Lagrangian mechanics with constraints. *Sov. Phys. Dokl.*, 17:34–36, 1972.

Bol. Soc. Esp. Mat. Apl.
n°50(2010), 83–99

HIGHLY STABLE RK TIME ADVANCING SCHEMES FOR COMPUTATIONAL AERO ACOUSTICS

M. CALVO, J.M. FRANCO, J.I. MONTIJANO, L. RÁNDEZ

IUMA-Departamento de Matemática Aplicada
Universidad de Zaragoza

{calvo, jmfranco, monti, randez}@unizar.es

Abstract

In this paper a brief survey of finite difference methods and time discretization schemes for the numerical simulation of problems in Computational Aero Acoustics (CAA), with special emphasis in the contributions of the authors in the last years to the subject, is presented. Due to the specific properties of these problems it is shown by means of some illustrative examples that standard schemes have some drawbacks and new numerical schemes have been derived taking into account not only the usual stability and accuracy requirements but also the dissipation and dispersion properties as well as low storage requirements. Some relevant contributions to the subject are presented comparing the relative merits by means of a Fourier analysis and numerical experiments.

1 Introduction.

In the field of Computational Aero Acoustics (CAA) many experiments require to simulate the propagation of sound waves far from its source with minimal dissipation and dispersion errors and these requirements cannot easily obtained with the standard spatial discretizations and time advancing integrators based on the well known Runge-Kutta (RK) methods that have been designed taking into account only the highest order and the best stability properties. Because of this, in the last 15 years a large number of publications have appeared proposing alternative discretizations that take into account not only the order and stability but also the dispersion and dissipation properties of the method.

Concerning the spatial discretization of the first derivative, it has been customary for a uniform spatial grid to take a $(2N+1)$-points stencil symmetric finite difference schemes of type

$$\partial_x u(x_j,t) \simeq \frac{1}{\Delta x} \sum_{l=-N}^{N} a_l\, u(x_{j+l},t), \qquad a_l + a_{-l} = 0,\ l = 0,\ldots,N, \quad (1)$$

This work was supported by project MTM2007-67530-C02-01

with maximum order $2N$ that are non dissipative. However, instead of choosing the available coefficients to attain the maximum order, several authors have determined these coefficients taking into account the dispersion properties for wave numbers $k\Delta x$ in some interval of type $[\alpha_l, \alpha_h]$ with $0 \leq \alpha_l < \alpha_h < \pi$. This idea was exploted by Tam and Webb [13] to derive the widely used Dispersion Relation Preserving (DRP) discretizations. A first scheme has a 7-point stencil, order 4, and it is able to resolve wavenumbers with an error $< 5 \times 10^{-3}$ for $k\Delta x \leq 1.16$. A second one with a 9-point stencil, has order 6 and resolve wavenumbers with an error smaller than 5×10^{-3} for $k\Delta x \leq 1.31$. More recently Bogey and Bailly [3], by minimizing the relative integral dispersion error between $\alpha_l = \pi/16$ and $\alpha_h = \pi/2$ have derived several optimized fourth order finite difference discretizations (1) for $N = 4, 5, 6$. In addition to the finite difference schemes (1), other implicit compact and ENO schemes with better dissipation and dispersion properties but with a higher computational cost have been considered also for the spatial discretizations.

Concerning the time advancing schemes, the classical fourth order four stage RK method has been the most popular in many applications because of their order and relatively large stability interval in the imaginary axis. However as we shall see later for CAA problems the step size of this method is not restricted by stability but by the dissipation and dispersion errors. This fact, together with the high dimensionality of the semidiscretizations, has led to a number of researchers to the derivation of special low-storage RK methods where the coefficients are selected taking into account the dissipation, dispersion, stability and order properties when applied to the linear wave test equation

$$\partial_t u(x,t) + c\partial_x u(x,t) = 0. \tag{2}$$

Among the contributions on this line of research we will mention the papers [2], [3], [4], [5], [8], [9], [10], [12].

The paper is organized as follows. In Section 2 a motivation for the development of DRP is given. Some schemes together with their main properties are presented with numerical examples that show the advantage of the new discretizations. In Section 3 we motivate the new time advancing RK schemes for CAA and the main contributions of the authors to this subject. Finally, in section 4 we present the optimization of spatial and temporal schemes taking into account the total dispersion and dissipation errors for $N = 5, 6$.

2 Centered finite difference schemes for the spatial discretization

For the spatial derivative we will consider centered finite difference schemes with a $(2N+1)$-point stencil of type

$$\partial_x u(x_j,t) \simeq \delta_x u_j \equiv \frac{1}{\Delta x} \sum_{l=-N}^{N} a_l\, u(x_{j+l},t) = \frac{1}{\Delta x} \sum_{l=1}^{N} a_l\, (u(x_{j+l},t) - u(x_{j-l},t)), \tag{3}$$

where Δx is the mesh spacing, $x_k = k\Delta x$, and to have no dissipation we will assume $a_j + a_{-j} = 0$, $j = 0, \ldots, N$. This assumption implies that the Taylor expansion of the right hand side of (3) around x_j only contains even powers of Δx and therefore the maximum order attainable by these schemes is $2N$.

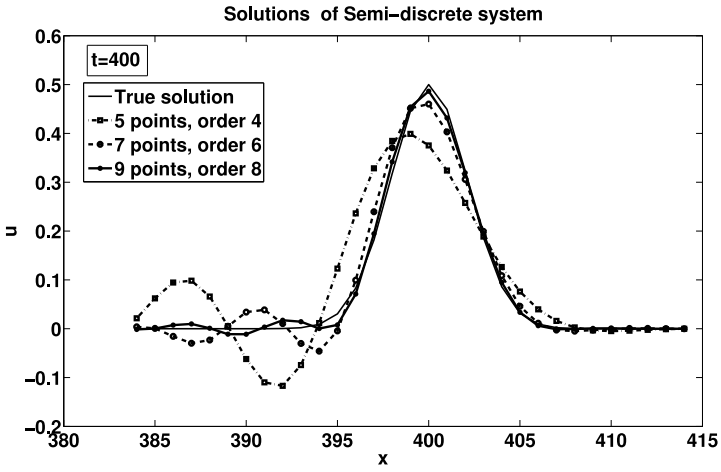


Figure 1: Numerical solution of the wave test equation semidiscretized with $(2N+1)$-point schemes for $N = 2, 3, 4$

We will start presenting the results of some simple numerical experiments. First of all in Fig. 1 we display the numerical solutions obtained by solving exactly the semidiscretization of the wave test equation (2) given by

$$\begin{cases} \partial_t u_j + c\, \delta_x u_j = 0, & t \in [0, 400] \\ u_j(0) = \Phi(x_j), \end{cases} \tag{4}$$

for the maximum order $(2N+1)$-point schemes (3) for $N = 2, 3, 4$, with $c = 1$, $\Phi(x) = \Phi_1(x) \equiv 0.5\ \exp(-x^2/9)$, $\Delta x = 1$, $x_j = j$, $j \in [-50, 450]$ at the time level $t = 400$ together with the exact solution of the wave equation. Observe that the exact solution is a Gaussian type wave that moves to the right with constant velocity $c = 1$ preserving its shape, so that $u(x,t) = \Phi_1(x - t)$.

As follows from Fig. 1 in the fourth order discretization the spatial dispersion errors introduces strong changes in the shape of the wave and these changes are smaller when the order is increased but in any case they are not negligible.

In Fig. 2 we present the solutions obtained by Sine-Gaussian initial condition given by $\Phi(x) = \Phi_2(x) \equiv \sin(\pi x/2)\ \exp\left(-x^2/9\right)$ for $\Delta x = 1$, $x \in [-50, 250]$ where the time interval is $t \in [0, 200]$. Now, since the problem is more difficult due to the spectral contents of $\Phi_2(x)$, we have taken the higher order spatial discretizations of orders 8 and 10 corresponding to 9 and 11 points respectively. Now even with the high order spatial discretizations the profile of the wave is destroyed.

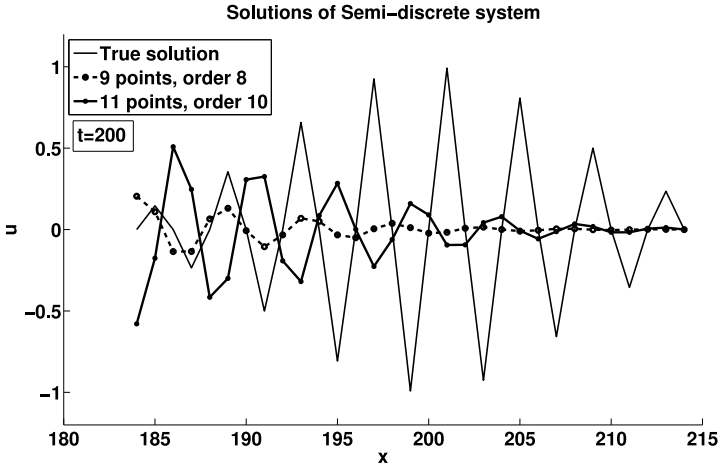


Figure 2: Numerical solution of the wave test equation, Sine-Gaussian initial condition, semidiscretized with $(2N+1)$-point schemes for $N = 4, 5$

To explain the behavior of the spatial difference schemes we apply the spatial Fourier transform given by

$$\widehat{u}(k,t) = \int_{-\infty}^{\infty} e^{-ikx}\, u(x,t)\, \mathrm{d}x, \quad \widehat{u}(k,t) = \Delta x \sum_{j=-\infty}^{j=+\infty} e^{-ikx_j}\, u(x_j,t)$$

to the original wave test equation and the semidiscretization (4) respectively obtaining

$$\partial_t \widehat{u}(k,t) + ick\, \widehat{u}(k,t) = 0, \qquad \partial_t \widehat{u}(k,t) + ick^*\, \widehat{u}(k,t) = 0$$

where

$$k^* = \frac{-i}{\Delta x} \sum_{l=-N}^{N} a_l\, e^{ilk\Delta x} = \frac{2}{\Delta x} \sum_{l=1}^{N} a_l\, \sin(lk\Delta x). \tag{5}$$

For a given N, the quantity $k^*\Delta x$ is usually called the effective (or numerical) wavenumber and as follows from (5) is a function of the exact wave number $k\Delta x$. Further for a discretization with order p, exact and effective wavenumbers agree up to order $(\Delta x)^{p+1}$ and $(k^*\Delta x)/(k\Delta x) \to 1$ as $\Delta x \to 0$. However when $k\Delta x \leq \pi$ separates from 0, the dispersion error may become very large. In Fig 3 we show the behavior of $(k^*\Delta x)$ as a function of $(k\Delta x)$ for $2N+1 = 5, 7, 9, 11$ when $k\Delta x \in [0, \pi]$. In all cases $(k^*\Delta x)$ have a unique maximum $k_{max}\Delta x$ in this interval and this implies that those waves with $k \geq k_{max}$ will be badly represented by the spatial discretization. In other words, for the $(2N+1)$-point stencil finite difference scheme of maximum order the spectral components of the initial condition $\Phi(x)$ with wavenumbers $k \geq k_{max}$ are not suitably represented by the discretization (3).

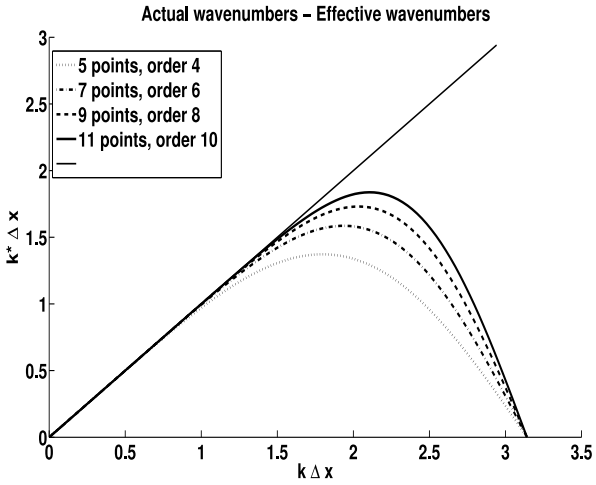


Figure 3: Numerical scaled wavenumbers versus actual scaled wavenumbers

In addition to this, it has been remarked by several authors [13] that the requirement $k \leq k_{max}$ is not enough to ensure a good dispersion behavior of the spatial discretization because even for $k \lesssim k_{max}$ the error in the numerical wavenumber can be large. Then, they introduced an additional condition

$$k \leq k_c = \max \{k \geq 0, |k - k^*| \leq \varepsilon\} \tag{6}$$

where ε is a small quantity that in some practical calculations has been taken as $\varepsilon = 10^{-3}$ because it leads to a reasonable dispersion error. In Table 1 we present the values of k_{max} and k_c for the high order $(2N+1)$-points schemes with $N = 2, 3, 4, 5$.

Table 1: Dispersion values, k_{max} and k_c, for $(2N+1)$-points schemes with $N = 2, 3, 4, 5$

$2N+1$	order	k_{max}	k_c
5	4	1.37	0.69
7	6	1.59	0.97
9	8	1.73	1.16
11	10	1.84	1.32

Taking into account Table 1 it is possible to explain the numerical results of the above examples. In the first example, since the spectral contents of $\Phi_1(x)$ is the right half of a Gaussian type function centered at $k = 0$, then the methods with 11 points are able to deal accurately most of the relevant spectral contents of $\Phi_1(x)$. Nevertheless in the second example, since the Fourier contents is a

Gaussian type function centered around $\pi/2$ all the considered methods cannot include the relevant spectral contents of $\Phi_2(x)$ and then it is not a suitable discretization for this function.

To cope with problems where the spectral contents of the solution is not close to the origin, several authors have proposed discretizations (3) with order smaller than the maximum order and using the available parameters to get larger values of k_{max} and k_c. This idea was used by Tam and Webb [13] to derive two 7- and 9-point stencil points discretizations called DRP (dispersion relation preserving methods) and more recently Bogey and Bailly [3] other 9-, 11- and 13-points discretizations. In Table 2 we collect the dispersion properties of these methods

Table 2: Dispersion values, k_{max} and k_c, for optimized difference schemes.

method	$2N+1$	order	k_{max}	k_c
Tam-Webb	7	4	1.73	1.45
Tam-Webb	9	6	1.77	1.28
Bogey-Bailly	9	4	1.1	1.5
Bogey-Bailly	11	4	1.98	1.66
Bogey-Bailly	13	4	2.14	1.92

Next the Figures 4 and 5 show the profiles of the semidiscrete and exact solutions at the final time in the above two examples for several optimized schemes. It must be noticed that in the second example, the 13-point discretization preserves quite accurately the shape of the wave.

3 Runge-Kutta time integration schemes

To illustrate the behavior of standard RK time integrators we start considering the application of the classical four-stage fourth order RK4 method with a fixed step size to the semidiscretization (4) with the initial condition $\Phi(x) = \Phi_1(x)$. Concerning the choice of the step size recall that when a scalar linear problem $y' = \lambda y$ is integrated by a one step method, the solution satisfies $y(t_n + \Delta t) = e^{\lambda \Delta t} y(t_n)$ whereas the numerical solution satisfies $y_{n+1} = R(\lambda \Delta t) y_n$, where $R(z)$ is the stability function (or amplification function) which approximates the exponential function. Observe that all eigenvalues of the semidiscretization (4) are pure imaginary simple $\pm iw$ and then to satisfy the stability requirement we should have $|R(iw\Delta t)| \leq 1$ for all $w \in [0, ck_{max}]$. On the other hand it is well known that the (imaginary) stability interval of the fourth order RK is $[-2\sqrt{2}, 2\sqrt{2}]$, thus if we take the DRP spatial discretization of Tam and Webb [13] with 9-points and order 6, according to Table 2, $1.77\Delta t \leq 2\sqrt{2}$ which implies $\Delta t \leq 1.597$. Figure 6 shows the profiles of the numerical and exact solution with $\Delta t = 1.597$ at the final time level. This example shows large phase and amplitude errors in the numerical solution. In addition, similar experiments with $\Delta t = 1$ show smaller but not negligible errors and to get

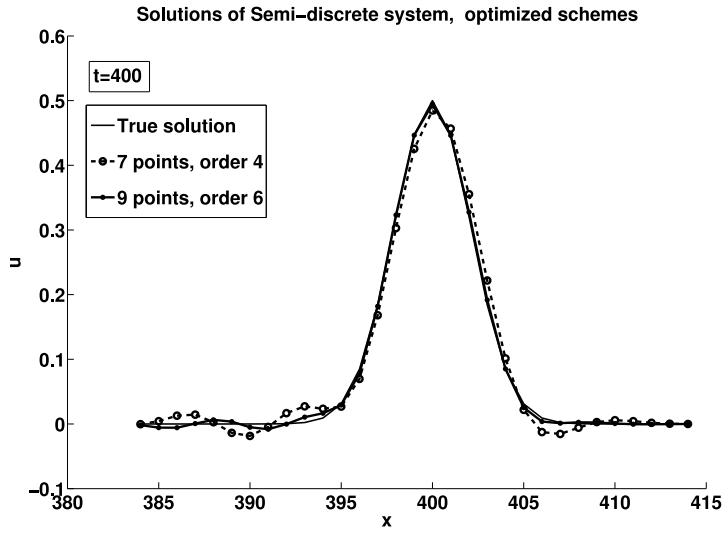


Figure 4: Numerical solution of the wave test equation, Gaussian initial condition, with optimized schemes with $N = 3, 4$

accurate solutions, time steps of $\Delta t \simeq 0.5$ are necessary. These experiments imply that in the standard RK4 stepsizes much smaller than the stability limit are necessary to preserve the dissipation and dispersion properties of this wave type solutions. This fact has led in the last decade to the construction of special RK time advancing schemes taking into account instead of the usual order-stability properties the following ones:

C1) Stability: $|R(iw\Delta t)| \leq 1$ for all $w \in [0, ck_{max}]$.

C2) Dissipation error: the quantity $|1 - R(iw\Delta t)|$ be small for all $w \in [0, ck_c]$.

C3) Dispersion error: the quantity $|\arg R(iw\Delta t) - w\Delta t|$ be small for all $w \in [0, ck_c]$.

C4) Maximum linear and non linear order. In the linear case defined by the maximum p such that $R(z) - e^z = \mathcal{O}(z^{p+1})$, and in the non linear case by Butcher's conditions.

C5) Low storage implementation.

The last requirement C5) has been introduced because in practical calculations (typically 2D and 3D-dimensional problems of CAA) the number of spatial grid points can be very high and consequently the dimensionality of the system. Therefore, as remarked by several authors (see e.g. [4], [7], [12], [14]), effective integrators for practical problems must use the minimum number of registers.

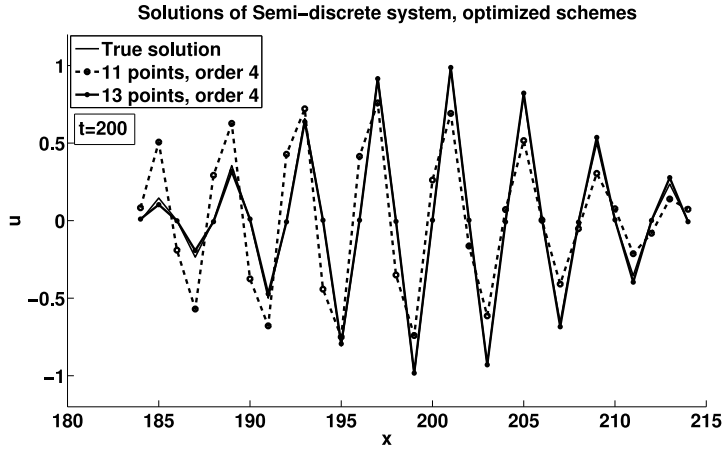


Figure 5: Numerical solution of the wave test equation, Sine-Gaussian initial condition, with optimized schemes with $N = 5, 6$

In general an explicit s-stage RK method for the numerical integration of the m-dimensional system $\partial_t U(t) = f(t, U(t))$, $U(t_n) = U_n \in \mathbf{R}^m$ is given by the equations

$$\begin{aligned} U_{n+1} &= U_n + \Delta t \sum_{j=1}^{s} b_j F_j, \\ F_j &= f\left(t_n + c_j \Delta t, U_n + \Delta t \sum_{k=1}^{j-1} a_{jk} F_k\right), j = 1 \ldots, s. \end{aligned} \tag{7}$$

where b_j, c_j, and a_{jk} are real constants that define the method. Clearly, advancing a step $U_n \to U_{n+1}$ requires, in general, the storage of s vectors of dimension m. However as observed by the authors in [4] if the $(s-1)s/2$ coefficients a_{jk} satisfy the conditions $a_{jk} = b_k$ for all $1 \le k \le j-2$, equations (7) reduce to

$$\begin{aligned} &V_0 = U_n, \quad F_0 = 0 \\ &V_j = V_{j-1} + b_j h F_{j-1} \\ &F_j = f(t_n + c_j \Delta t, V_j + \gamma_j \Delta t F_{j-1}), \quad \gamma_i = a_{j+1,j} - b_j, \qquad j = 1, \ldots, s \\ &U_{n+1} = V_s + b_s \Delta t F_s \end{aligned} \tag{8}$$

and therefore it is possible to implement the method by using only two m-vectors. An alternative low storage approach has been proposed by Williamson in [14].

Since the requirements C1)–C3) depend only on the amplification function $R(z) = \sum_{j=0}^{s} \gamma_j z^j, \gamma_0 = 1$, of the RK method (7) many authors have derived

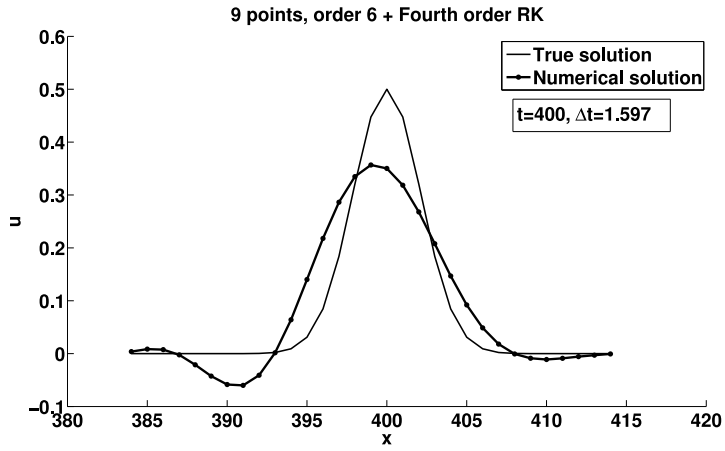


Figure 6: Numerical solution of the wave test equation, Gaussian initial condition, with DRP sixth order spatial discretization and classical fourth order RK for time discretization

"optimal" methods for several values of the number of steps s. Thus Hu et. al. [8] obtained optimal methods for $s = 4, 5, 6$. Also Bogey and Bailly [3] have derived optimal methods combined with spatial stencils with 9-, 10- and 11-points. In the same line the authors of the present paper derived optimal methods for $s = 5, 6$.

To measure the quality of a method defined by $R(z) = \sum_{j=0}^{s} \gamma_j z^j, \gamma_0 = 1$, it has been usual to compare the following quantities

$$\begin{aligned} S &= \max\{z > 0, |R(iz)| \leq 1\} \\ L_d &= \max\{z > 0, ||R(iz)| - 1| \leq 10^{-3}\} \\ L_\varphi &= \max\{z > 0, |\arg(R(iz)) - z| \leq 10^{-3}\} \end{aligned} \tag{9}$$

In Table 3 the values of these parameters corresponding to several methods are presented.

In Fig 7 we display the profiles of the exact and numerical solutions of example 1 for the RK method of Calvo et. al. with six stages and order four at the final time level. As can be seen, the shape of the wave is reproduced properly even with a time stepsize $\Delta t = 1.0$

To end this section let us note that in the frame of implicit RK there are methods such as Gauss methods that possess the best properties of stability and dissipation ($S = L_d = \infty$) although they have a non zero dispersion error that depends on s. For example, the fourth order Gauss method has $L_\varphi = 0.95$ In Figure 8 we display the profiles of the exact and numerical solutions of example 1 for the 2-stage Gauss method with order four at the final time level. Now, if the spatial discretization has good dissipation and dispersion properties, the time integrator reproduces quite accurately the shape of the solution with the

Table 3: Values of S, L_d and L_φ parameters for to several RK methods

method	order	order (lin.)	stages	S	L_d	L_φ
Classic	3	3	3	1.73	0.40	0.49
Classic	4	4	4	2.83	0.73	0.68
Hu et. al.	2	2	4	2.85	0.85	0.86
Hu et. al.	2	2	5	3.54	1.72	1.35
Hu et. al.	3	4	6	1.75	1.41	1.27
Calvo et. al.	3	3	5	3.48	0.91	1.09
Calvo et. al.	4	4	5	3.48	1.25	0.91
Calvo et. al.	4	4	6	3.82	2.00	1.14
Gauss	2	2	1	∞	∞	0.23
Gauss	4	4	2	∞	∞	0.95

step size $\Delta t = 1$. However the main drawback of Gauss methods for CAA problems is its implicitness that entails a very high computational cost.

4 Coupled FD-RK schemes

In previous sections optimization of spatial and temporal schemes has been considered independently. Recently, Ramboer et al [11] considered the errors of both (spatial and temporal) discretizations in order to derive some six-stage RK time advancing schemes by selecting the free parameters of this RK to minimize the total dispersion and dissipation errors for two specific spatial discretizations: fourth-order Finite Volumes (FV) compact central and third-order FV upwind discretizations.

In this section we present a simultaneous analysis of both spatial and temporal discretization schemes. In order to do that, let us take into account that denoting by $\alpha = c\frac{\Delta t}{\Delta x}$ (CFL number), the total dispersion and dissipation errors are given, respectively by,

$$\phi_T(\alpha, z) = \alpha z + \arg(R(-i\alpha z^*)), \quad d_T(\alpha, z) = 1 - |R(-i\alpha z^*)|,$$

with $z^* = 2\sum_{j=1}^{N} a_j \sin(jz)$ and a_j the coefficients of the symmetric FD (1).

For the spatial part, we use symmetric fourth-order central finite differences (1) which are zero-dissipative. The approximation (1) is of order p if

$$\partial_x u(t, x_l) - \delta_x u(t, x_l) = \mathcal{O}(\Delta x^p), \quad p \geq 1.$$

For the time advancing scheme, a six-stage fourth-order RK method is used. It has the amplification function

$$R(\zeta) = 1 + \zeta + \frac{1}{2!}\zeta^2 + \frac{1}{3!}\zeta^3 + \frac{1}{4!}\zeta^4 + \beta_5\,\zeta^5 + \beta_6\,\zeta^6, \quad \beta_i \in \mathbb{R}.$$

Fourth-order FD-RK schemes with $2N + 1$ grid points and $N \geq 5$ were developed in [6], denoted by SFD_{2N+1}-RK_6, depending on the parameters

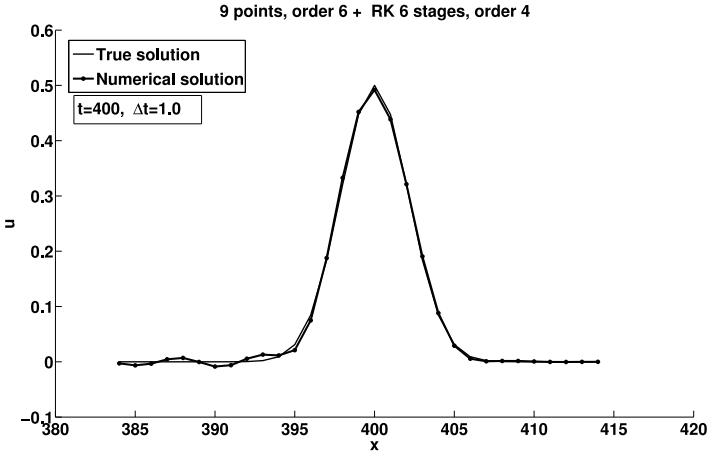


Figure 7: Numerical solution of the wave test equation, Gaussian initial condition, with DRP sixth order spatial discretization and Calvo et al. optimized six stages, fourth order RK for time discretization

Table 4: Coefficients of the FD-RK

$N = 5$ (11 points)	$N = 6$ (13 points)
$\beta_5 = 0.00785780000000$	$\beta_5 = 0.00785812800000$
$\beta_6 = 0.00094507900000$	$\beta_6 = 0.00094851200000$
$a_1 = 0.88131666666666$	$a_1 = 0.90280686066667$
$a_2 = -0.29651333333333$	$a_2 = -0.32759725333333$
$a_3 = 0.09657000000000$	$a_3 = 0.12294034000000$
$a_4 = -0.02315000000000$	$a_4 = -0.03812260000000$
$a_5 = 0.00292000000000$	$a_5 = 0.00835681000000$
	$a_6 = -0.00095450400000$

$a_3, \ldots, a_N, \beta_5$ and β_6. These free parameters were determined by minimizing the following error measure for $\alpha = 1$

$$\int_{\pi/16}^{\pi/2} \left[\left(\frac{\phi_T(z)}{\pi} \right)^2 + d_T^2(z) \right] \mathrm{d}z,$$

with the stability restriction

$$|R(iz)| < 1, \quad \forall z \in (0, \pi/2].$$

In Table 6, we give the coefficients of the optimized SFD_{2N+1}-RK_6 schemes for $N = 5, 6$.

In order to show the performance of the new schemes we will compare them with the FD-RK schemes that combine the spatial discretizations derived in

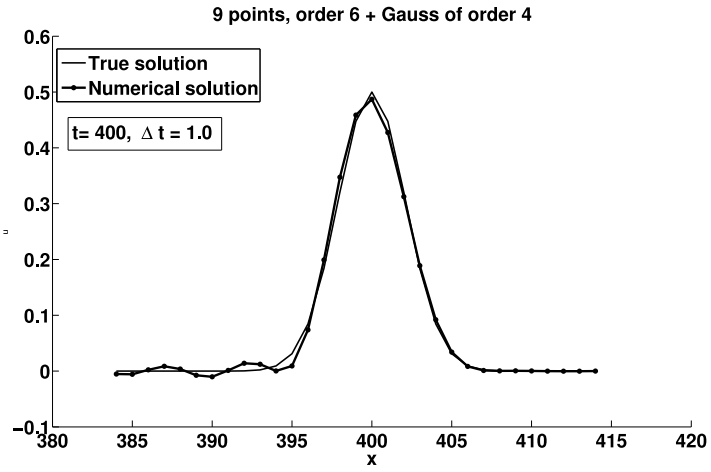


Figure 8: Numerical solution of the wave test equation, Gaussian initial condition, with DRP sixth order spatial discretization and fourth order Gauss RK for time discretization

Table 5: Dispersion and dissipation intervals

	Dispersion $\lvert\phi_T/\pi\rvert \leq 10^{-3}$	Dissipation $\lvert d_T\rvert \leq 10^{-4}$
$FDo11p$-RKB_6	[0,1.44]	[0,1.82]
$FDo13p$-RKB_6	[0,1.40]	[0,1.80]
SFD_{11}-RK_6	[0,1.63]	[0,1.74]
SFD_{13}-RK_6	[0,1.74]	[0,1.75]

Bogey et al [3] together with the fourth-order six-stage RK algorithm derived in Berland et al [1], referred as $FDo11p$-RKB_6 and $FDo13p$-RKB_6. In figures 9 and 10, we show the total dispersion and dissipation errors for $N = 6$.

In Table 5 we present for $N = 5, 6$ the intervals of dispersion and dissipation satisfying $|\phi_T/\pi| \leq 10^{-3}$ and $|d_T| \leq 10^{-4}$.

The next step is the construction of the low-storage RK method, according to (8), associated to the optimized stability function. RK methods with $s = 6$ stages and algebraic order 4, must satisfy

$$\mathbf{b}^T\mathbf{e} = 1, \quad \mathbf{b}^T\mathbf{c} = \tfrac{1}{2}, \quad \mathbf{b}^T\mathbf{c}^2 = 1/3, \quad \mathbf{b}^T\mathbf{A}^2\mathbf{c} = \tfrac{1}{24},$$
$$\mathbf{b}^T\mathbf{A}\mathbf{c} = \tfrac{1}{6}, \ \mathbf{b}^T\mathbf{c}^3 = \tfrac{1}{4}, \ \mathbf{b}^T(\mathbf{c}\cdot\mathbf{A}\mathbf{c}) = \tfrac{1}{8}, \ \mathbf{b}^T\mathbf{A}\mathbf{c}^2 = \tfrac{1}{12}.$$

Furthermore, the coefficients must satisfy the two additional conditions

$$\mathbf{b}^T\mathbf{A}^3\mathbf{c} = \beta_5, \quad \mathbf{b}^T\mathbf{A}^4\mathbf{c} = \beta_6,$$

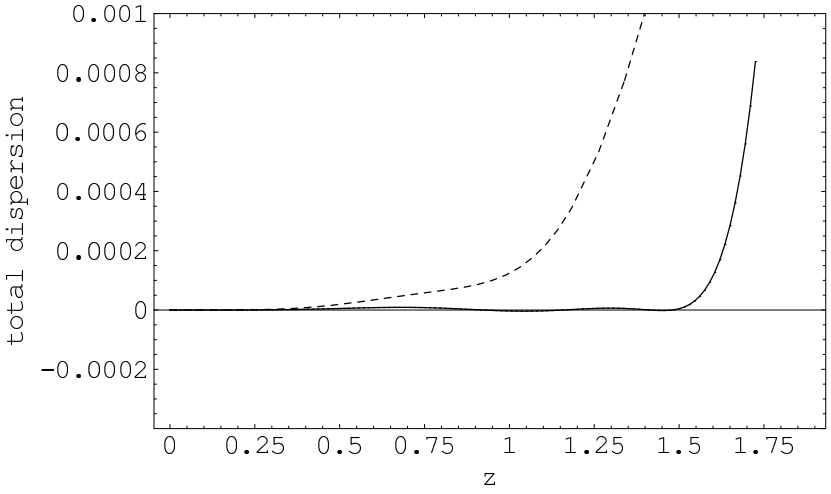


Figure 9: Total dispersion : — SFD_{13}-RK_6; - - - $FDo13p$-RKB_6.

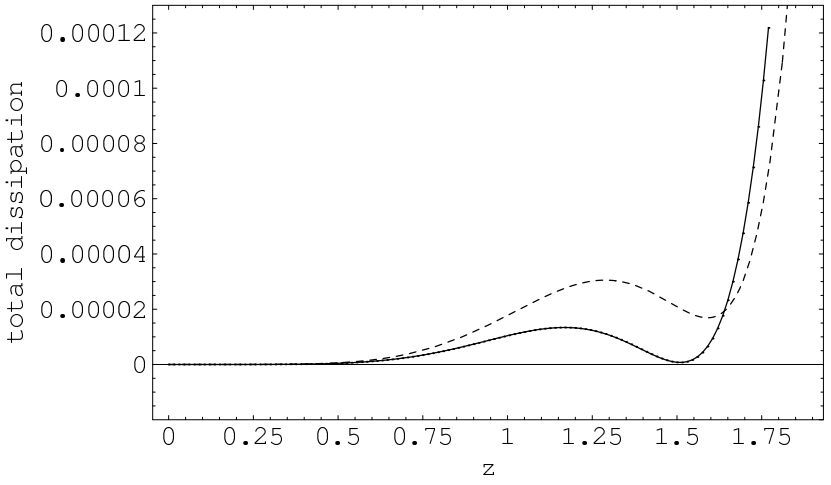


Figure 10: Total dissipation : — SFD_{13}-RK_6; - - - $FDo13p$-RKB_6.

Table 6: Coefficients of the low-storage RK methods

$c_1 = 0$	$b_1 = 0.10974285720869$	$\gamma_1 = 0.18025714279131$
$c_2 = 0.29000000000000$	$b_2 = 0.13448959704914$	$\gamma_2 = 0.13857764418235$
$c_3 = 0.38281009844018$	$b_3 = 0.38294944978031$	$\gamma_3 = 0.08426505141729$
$c_4 = 0.71144695545543$	$b_4 = -0.60216813067103$	$\gamma_4 = 0.66860432485845$
$c_5 = 0.69361809822556$	$b_5 = 0.49945631650501$	$\gamma_5 = 0.30908387736075$
$c_6 = 0.83355396723287$	$b_6 = 0.47552991012788$	SFD_{11}-RK_6
$c_1 = 0$	$b_1 = 0.11287033711698$	$\gamma_1 = 0.19025714279131$
$c_2 = 0.30000000000000$	$b_2 = 0.14141097168321$	$\gamma_2 = 0.13989004259688$
$c_3 = 0.38412249685471$	$b_3 = 0.36534072934351$	$\gamma_3 = 0.08964556109275$
$c_4 = 0.71682746513089$	$b_4 = -0.54871438354286$	$\gamma_4 = 0.66668799693474$
$c_5 = 0.69170177030185$	$b_5 = 0.47533937806862$	$\gamma_5 = 0.31691629028663$
$c_6 = 0.84138638015875$	$b_6 = 0.45375296733053$	SFD_{13}-RK_6

where β_5 and β_6 were obtained in the above optimization process, given in Table 4. We have a set of ten nonlinear equations for eleven coefficients b_i, γ_i with the constraints

- The weights satisfy $|b_i| \leq 2$, $i = 1, \ldots, 6$.
- The nodes satisfy $c_i \neq c_j$, $\forall i \neq j$ and $0 \leq c_i \leq 1$, $i = 1, \ldots, 6$.
- Minimize the $\|\cdot\|_2$-norm of the leading term of the local error.

In table 6 we display the coefficients of the RK method obtained in the above process.

Finally in Figures 11 and 12 we plot the errors integrating the problem (4) with initial condition $\Phi(x) = \sin\left(\frac{\pi x}{4}\right)\exp\left(-\log(2)x^2/9\right)$. This initial disturbance is propagated over $800\Delta x$ and we have used the CFL number $\alpha = 1$, with $\Delta x = \Delta t = 1$.

The numerical results show the advantage in using a global minimization procedure for generating a low dispersion and dissipation scheme. Comparisons with more "classical" schemes clearly illustrate this fact. On the other hand, more experience with non-linear problems in the field of aero-acoustics is needed.

References

[1] J. Berland, C. Bogey, O. Marsden and C. Bailly. *High-order, low dispersive and low dissipative explicit schemes for multiple-scale and boundary problems* J. of Comput. Phys. 224, 2, 637–662 (2007).

[2] V. Allampalli, R. Hixon, M. Nallasamy and S.D. Sawyer. *High-accuracy large-step explicit Runge-Kutta (HALE-RK) schemes for computational aeroacoustics* J. Comput. Phys. 228, 3837–3850 (2009).

[3] C. Bogey and C. Bailly. *A family of low dispersive and low dissipative explicit schemes for flow and noise computations* J. Comput. Phys. 194, 194–214 (2004).

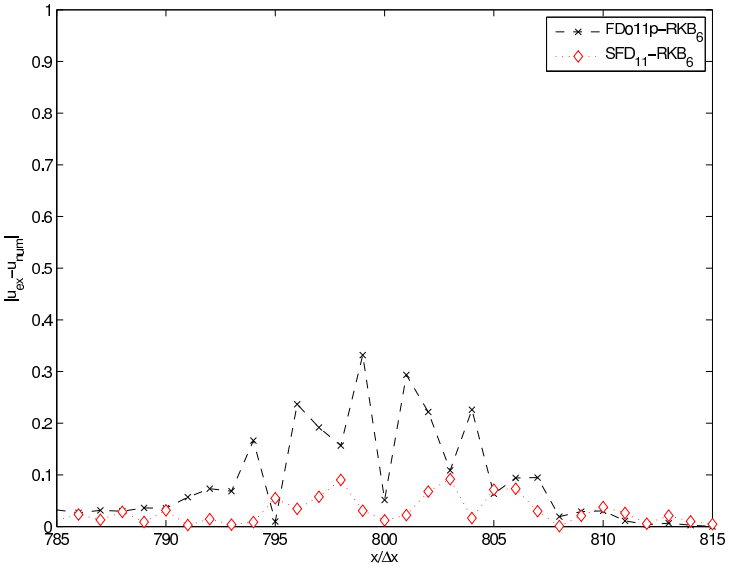


Figure 11: Errors for the schemes SFD_{11}-RK_6 and $FDo11p$-RKB_6. The numerical solution provided by SFD_{11}-RK_6 is better than $FDo11p$-RKB_6.

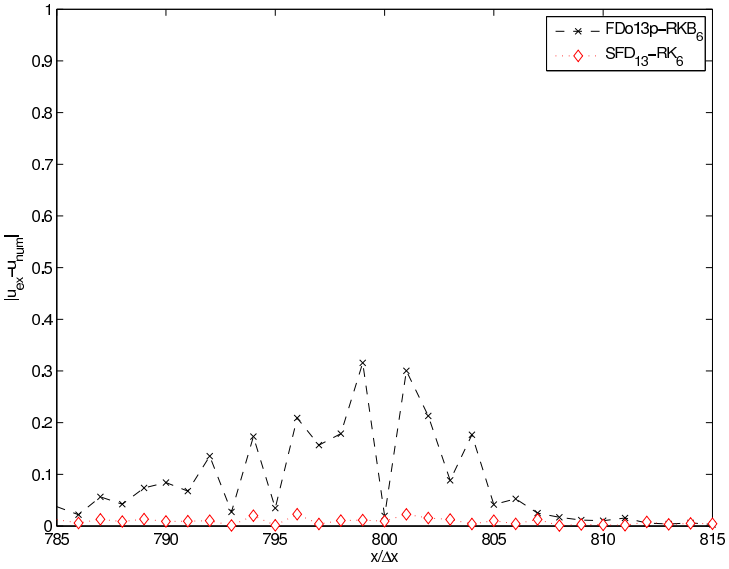


Figure 12: Errors for the schemes SFD_{13}-RK_6 and $FDo13p$-RKB_6. The numerical solution provided by SFD_{13}-RK_6 is better than $FDo13p$-RKB_6.

[4] M. Calvo, J.M. Franco, J.I. Montijano and L. Rández. *Minimum Storage Runge-Kutta Schemes for Computational Accoustics* Computers and Math. with Applications 45, pp. 535–545 (2003).

[5] M. Calvo, J.M. Franco and L. Rández. *A new minimum storage Runge-Kutta scheme for computational accoustics* J. Comput. Phys. 201, pp. 1–12 (2004).

[6] M. Calvo, J.M. Franco, J.I. Montijano and L. Rández. *Optimization of spatial and minimum storage RK schemes for computational acoustics* Numerical Analysis and Applied Mathematics, International Conference on Numerical Analysis and Applied Mathematics 2009. AIP Conference Proceedings, 1168 743–745 (2009).

[7] M. H. Carpenter and C. A. Kennedy . *A Fourth-Order 2N-Storage Runge-Kutta Scheme* NASA TR TM109112, June 1994.

[8] F. Q. Hu, M. Y. Hussaini and J. Manthey. *Low-dissipation and low dispersion Runge-Kutta schemes for computational accoustics* J. Comput. Phys. 124, 177–191 (1996).

[9] C. A. Kennedy M. H. Carpenter and R. M. Lewis. *Low-storage, explicit Runge-Kutta schemes for the compressible Navier-Stokes equations* NASA/CR-1999-209349, ICASE Report (99) (1999).

[10] J.L. Mead and R.A. Renaut. *Optimal Runge-Kutta Methods for First Order Pseudospectral Operators* J. Comput. Phys. 152, 404–419 (1999).

[11] J. Ramboer, T. Broeckhoven, S. Smirnov and C. Lacor. *Optimization of time integration schemes coupled to spatial discretization for use in CAA applications* J. Comput. Phys. 213, 777–802 (2006).

[12] D. Stanescu and W. G. Habashi. *2N-Storage Low dissipation and Dispersion Runge-Kutta schemes for Computational Accoustics* J. Comput. Phys. 143, 674–681 (1998).

[13] C.K.W. Tam and J.C. Webb. *Dispersion-relation-preserving finite difference schemes for computational acoustics* J. Comput. Phys. 107, 262 (1993).

[14] J.H. Williamson. *Low-storage Runge-Kutta schemes* J. Comput. Phys. 35, 48 (1980).

Bol. Soc. Esp. Mat. Apl.
n°50(2010), 99–114

ADAPTIVE NONCONFORMING FINITE ELEMENTS FOR THE STOKES EQUATIONS

ROLAND BECKER, SHIPENG MAO, DAVID TRUJILLO

Laboratoire de Matématiques Appliquées
Université de Pau et des Pays de l'Adour
INRIA Bordeaux Sud-Ouest, EPI Concha

roland.becker@univ-pau.fr, shipeng.mao@univ-pau.fr,
david.trujillo@univ-pau.fr

Abstract

We discuss some recent progress in the convergence analysis of adaptive finite element methods for the Stokes equations. First we present a result concerning the quasi-optimality of low-order non-conforming methods. Both the case of the Crouzeix-Raviart element on triangular meshes, and the Rannacher-Turek element on parallelogram elements are covered. Numerical experiments are conducted in order to appreciate the different variants of the algorithm.

Key words: *Adaptive finite element methods, nonconforming methods, quasi-optimality, Stokes equations.*
AMS subject classifications: *65N12, 65N15, 65N30, 65N50*

1 Introduction

We consider the Stokes equations with Dirichlet and Neumann-type boundary conditions in a bounded polygonal domain $\Omega \subset \mathbb{R}^2$:

$$\begin{cases} -\Delta u + \nabla p = f \quad \text{in } \Omega, \\ \operatorname{div} u = 0 \quad \text{in } \Omega, \\ u = g \quad \text{on } \Gamma_D, \quad \dfrac{\partial u}{\partial n} - pn = g \quad \text{on } \Gamma_N, \end{cases} \tag{1}$$

with given forces $f \in L^2(\Omega)^2$ and $g \in H^{1/2}(\Gamma_D)^2$ and $\Gamma_D, \Gamma_N \subset \partial\Omega$ such that $\partial\Omega = \Gamma_D \cup \Gamma_N$.

We will first consider the case of homogeneous Dirichlet boundary conditions, $\Gamma_N = \emptyset$ and $g = 0$.

The $L^2(\Omega)$, $L^2(\Omega)^d$, and $L^2(\Omega)^{d\times d}$-scalar product are denoted by $\langle\cdot,\cdot\rangle$. The corresponding norms are written as $\|\cdot\|$ and $Q = L^2_0(\Omega)$ is the space of square-integrable functions with mean zero. For a sub-domain $K \subset \Omega$ we use $\|\cdot\|_K$.

By means of the standard Sobolev space $V := H_0^1(\Omega)^d$, the considered weak formulation of (1) reads: Find $(u, p) \in V \times Q$ such that for all $(v, q) \in V \times Q$ there holds:

$$\langle \nabla u, \nabla v \rangle - \langle p, \operatorname{div} v \rangle + \langle \operatorname{div} u, q \rangle = \langle f, v \rangle. \tag{2}$$

The adaptive algorithm selects a sequence of meshes $\{h_k\}_{k \geq 0}$ in a family of admissible meshes $\mathcal{H}$ defined by the starting mesh h_0 and a local mesh refinement algorithm. For any $h \in \mathcal{H}$, let V_h and Q_h be the discrete velocity and pressure spaces. We denote by $\mathcal{K}_h$ the set of cells and set $N_h := \#\mathcal{K}_h$. With the piecewise gradient operator $\nabla_h : V_h \to L^2(\Omega)^{d \times d}$ defined by $(\nabla_h v_h)|_K := \nabla v_h|_K$ and the piecewise divergence operator $\operatorname{div}_h : V_h \to L^2(\Omega)$ defined by $(\operatorname{div}_h v_h)|_K := \operatorname{div} v_h|_K$ for all $K \in \mathcal{K}_h$, the discrete approximation of (2) reads: Find $(u_h, p_h) \in V_h \times Q_h$ such that for all $(v_h, q_h) \in V_h \times Q_h$ there holds:

$$\langle \nabla_h u_h, \nabla_h v_h \rangle - \langle p_h, \operatorname{div}_h v_h \rangle + \langle \operatorname{div}_h u_h, q_h \rangle = \langle f, v_h \rangle. \tag{3}$$

Let $\varepsilon_h^2 := \|\nabla_h(u - u_h)\|^2 + \|p - p_h\|^2$ be the accuracy on a given mesh h. With $\varepsilon_k := \varepsilon_{h_k}$ and $N_k := N_{h_k}$ $(k = 0, 1, \ldots)$, the quasi-optimality of the adaptive algorithm means that

$$\varepsilon_k \approx N_k^{-s}, \qquad N_k \approx \varepsilon_k^{-1/s}, \tag{4}$$

where $s > 0$ is the best possible exponent of error decrease bounded by the a priori error analysis of the interpolation error ($s = 1/d$ for a first-order method and a sufficiently regular solution). The first step to (4) is the proof of geometric convergence, that is, the existence of $0 < \rho < 1$ such that for $k = 0, 1, \ldots$

$$\varepsilon_{k+1} \leq \rho\, \varepsilon_k. \tag{5}$$

Notice that (4) is meaningless without $\lim_{k \to \infty} \varepsilon_k = 0$.

We have been recently able to prove (5) and (4) for the lowest-order nonconforming finite elements on triangular and parallelogram meshes, see[2]. Our analysis is a generalization of similar results concerning adaptive methods for elliptic problems. For conforming finite elements important progress has been achieved in recent years, including convergence proofs [11, 15] and complexity estimates [6, 17, 3]; see also [1] for similar results on mixed finite elements. The common main structure of proof of these results is as follows: convergence is based on a a global upper bound and a local lower bound ('discrete local efficiency'). The term 'global' refers to the error $e = u - u_h$, whereas the term 'local' refers to the difference $u_{h'} - u_h$ of the solutions on two consequent meshes h and h' generated by the adaptive algorithm. The important complexity estimates are based on two additional results: a global lower bound and a local upper bound.

The case of nonconforming finite elements leads to additional technical difficulties: in addition to the nonconformity of the discrete functions, the orthogonality of the error is lost and has to be replaced by an appropriate estimate. Such an estimate has been obtained for the Crouzeix-Raviart space in

[7], based on the Marini-relation with the Raviart-Thomas mixed finite element solution [14]. This estimate has been improved in [3], where in addition quasi-optimal complexity has been shown.

A posteriori error estimates for non-conforming finite elements for the Stokes equations are well-known, see for example [9], [12], and the references cited therein. The common structure of a posteriori error estimates for different low-order nonconforming methods on triangular and quadrilateral meshes has been worked out in [8]. In this paper, we amend these results, in the context of the Stokes equations, by local upper bounds and local estimates for the non-orthogonality, which are the main tools for complexity estimates.

The considered adaptive algorithm is based on a comparison of the two contributions of the estimator in each step of the algorithm. This idea has been introduced for conforming adaptive finite elements in [4]. It leads to a particularly simple marking strategy, avoiding refinement according to the smaller term. In previous work we have shown convergence and quasi-optimal complexity of this algorithm for the Poisson equation.

Convergence and quasi-optimality of a completely different adaptive method for the Stokes equations have been proven in [13]. The algorithm considered there is based on an infinite-dimensional Uzawa method discretized by conforming finite elements, which avoids the discrete saddle-point system. It does not yield locally divergence-free discrete velocity fields and seems to be less popular in engineering practice.

After presenting the a posteriori error estimator and adaptive algorithm in Section 2, we report on the quasi-optimality of the algorithm in Section 3. Extension to non-homogeneuous Dirichlet and Neumann-type boundary conditions is made in Section 4. Finally, Section 5 is devoted to some numerical experiments. For brevity we restrict ourselves to the case of parallelogram meshes in two space dimensions.

2 Adaptive algorithm

For the discretization of (2) we use the lowest-order non-conforming finite element spaces on a family of shape-regular locally refined meshes $\mathcal{H}$. In the following, we denote by C a generic constant. We say that such a constant is mesh-independent, if the estimate in which it appears holds for all $h \in \mathcal{H}$.

The family of possible meshes $\mathcal{H}$ is defined recursively by means of a local refinement algorithm $\mathcal{R}ef$ starting from a given conforming mesh h_0. The local mesh refinement algorithm $\mathcal{R}ef$ takes as input a coarse mesh $h \in \mathcal{H}$ and a subset $\mathcal{M} \subset \mathcal{K}_h$ of marked cells and produced a fine mesh $h' \in \mathcal{H}$. Possibly, additional cells have to be refined, either in order to guarantee conformity of a triangular mesh, or in order to satisfy the regularity condition (6) stated below. This leads to a set $\widetilde{\mathcal{M}} \subset \mathcal{K}_h$ of actually refined cells with $\widetilde{\mathcal{M}} \supset \mathcal{M}$. We say that h' is a refinement of h and write $(h', \widetilde{\mathcal{M}}) = \mathcal{R}ef(h, \mathcal{M})$.

For given $h \in \mathcal{H}$ we denote by $\mathcal{K}_h$ the set of cells and by $\mathcal{S}_h$ the set of interior sides defined to be the usual edges of a rectangle. The set of boundary sides is

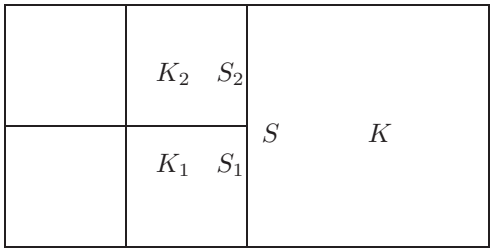


Figure 1: Refinement creating a hanging node and a hanging side S with subsides S_1 and S_2.

$\mathcal{S}_h^\partial$. The diameter and measure of $K \in \mathcal{K}_h$ (or $S \in \mathcal{S}_h$) are denoted by d_K (d_S) and $|K|$ ($|S|$), respectively. We denote by ω_K the set of neighboring cells of K.

In the case of a hanging node, see Figure 1, $\mathcal{S}_h$ contains a hanging edge S (the long edge in the figure) and subedges S_i with $S = \cup_i S_i$. We define the set of regular edges $\mathcal{S}_h^*$ by eliminating all hanging edges from $\mathcal{S}_h$. In addition we introduce the notation $\mathcal{S}_h^\perp$ for the set of hanging edges.

Let $\mathcal{N}_h$ be the set of regular nodes not lying on a hanging side and, for given $N \in \mathcal{N}_h$, let $\mathcal{K}(N) \subset \mathcal{K}_h$ be the set of cells such that $N \in \bar{K}$. In addition, we denote by $\mathrm{lev}(K)$ the refinement level of cell K. Then we impose the condition

$$\max_{K\in\mathcal{K}(N)} \mathrm{lev}(K) - \min_{K\in\mathcal{K}(N)} \mathrm{lev}(K) \le 1 \qquad \forall N \in \mathcal{N}_h, \tag{6}$$

which implies that only one hanging nodes is allowed per hanging side. In the case of conforming triangular meshes, the assumption (6) is meaningless.

We make the following hypothesis on the local mesh refinement algorithm.

1. The meshes consist either of triangles, tetrahedra or rectangles. All meshes are uniformly shape-regular, that is, there exists a mesh-independant constant C such that for all $h \in \mathcal{H}$ and $K \in \mathcal{K}_h$ there holds $d_K \le C\,|K|^{1/d}$.

2. There exist mesh-independent constants $0 < \kappa_1 < 1$ and $0 < \kappa_2 < 1$ such that for all $K \in \widetilde{\mathcal{M}}$ there holds:

$$|K'| \le \kappa_1\,|K| \quad \text{for all children } K' \text{ of } K \tag{7}$$

and

$$|S'| \le \kappa_2\,|S| \quad \text{for all children } S' \text{ of } S. \tag{8}$$

3. There exists a mesh-independant constant C_0 such that for any sequence $\{h_k\}_k$ with $(h_{k+1}, \widetilde{\mathcal{M}}_k) = \mathcal{R}ef(h_k, \mathcal{M}_k)$ and $n = 0, 1, \ldots$ there holds:

$$N_n \le N_0 + C_0 \sum_{k=0}^{n-1} \#\mathcal{M}_k. \tag{9}$$

The estimate (9) expresses the fact that the sum of the number of additionally refined cells $\#\widetilde{\mathcal{M}}_k - \#\mathcal{M}_k$ can be controlled. It is crucial for the estimate the complexity of the finite element meshes generated by the adaptive algorithm and is known to hold for the newest vertex bisection algorithm, see [6]. For quadrilateral meshes satisfying (6), we refer to [5].

For a given interior side $S \in \mathcal{S}_h^*$, let n_S be a chosen unit normal vector. Let $v_h \in L^2(\Omega)$ such that $v_h|_K \in C(\bar{K})$ for all $K \in \mathcal{K}_h$ and let $[v_h]_S$ be the jump defined as $[v_h]_S(x) := \lim_{\varepsilon \searrow 0} (v_h(x - \varepsilon n_S) - v_h(x + \varepsilon n_S))$ and $\{v_h\}_S$ be the mean defined as $\{v_h\}_S(x) := \frac{1}{2} \lim_{\varepsilon \searrow 0} (v_h(x - \varepsilon n_S) + v_h(x + \varepsilon n_S))$ for $x \in S$. For a boundary side, we set $n_S = n_{\partial\Omega}$ and $[v_h]_S(x) = v_h(x)$. The subscript S will be suppressed below, if this does not lead to confusion. The same notation is used for vector- and matrix-valued functions.

We denote by $\tilde{Q}^1(\mathbb{R}^2)$ the rotated bilinear space made out of $\{1, x, y, x^2 - y^2\}$. Let $h \in \mathcal{H}$. We define, generalizing the classical definition of [16] to the meshes with hanging nodes, the finite element spaces

$$V_h := \Big\{ v_h \in L^2(\Omega)^d : \; v_h|_K \in Q^1(\mathbb{R}^2)^2 \quad \text{for all } K \in \mathcal{K}_h \text{ and } \int_S [v_h]\, ds = 0 \quad \text{for all } S \in \mathcal{S}_h^* \cup \mathcal{S}_h^\partial \Big\}, \tag{10}$$

$$Q_h := \big\{ q_h \in L^2(\Omega) : \; q_h|_K \in P^0(\mathbb{R}^2) \quad \forall K \in \mathcal{K}_h \big\}. \tag{11}$$

For a hanging side $S \in \mathcal{S}_h^\perp$, the continuity requirement in (10) means that the degree of freedom associated with S are the mean of the ones associated to S_i.

The natural interpolation operator $\Pi_h : V \oplus V_h \to V_h$ is defined by

$$\frac{1}{|S|} \int_S \Pi_h v \, ds = \frac{1}{|S|} \int_S v \, ds \qquad \forall S \in \mathcal{S}_h^*, \, v \in V \tag{12}$$

and satisfies the projection property $\Pi_h v_h = v_h$ for all $v_h \in V_h$.

Next we extend the definition of canonical interpolation operator, for simplicity denoted the same. Let $h' \in \mathcal{H}$ be a refinement of $h \in \mathcal{H}$. We define $\Pi_h : V \oplus V_h \oplus V_{h'} \to V_h$ and $\Pi_{h'} : V \oplus V_{h'} \oplus V_h \to V_{h'}$ in the following way. For $S \in S_h^*$ there exist $S_i \in \mathcal{S}_{h'}^*$, $i = 1, \ldots, n$ such that $\bar{S} = \cup_{i=1} \bar{S}_i$; in case that S is not refined we set $n = 1$ and $S_1 = S$. Then, for given $v_{h'} \in V_{h'}$, we define $\Pi_h v_{h'} \in V_h$ by

$$\frac{1}{|S|} \int_S \Pi_h v_{h'} \, ds := \frac{1}{|S|} \sum_{i=1}^n \int_{S_i} v_{h'} \, ds. \tag{13}$$

In addition, for given $v_h \in V_h$, we define $\Pi_{h'} v_h \in V_{h'}$ by

$$\frac{1}{|S_i|} \int_{S_i} \Pi_h v_h \, ds := \frac{1}{|S_i|} \int_{S_i} \{v_h\} \, ds. \tag{14}$$

The interpolation operator Π_h has the following interpolation and stability property:

$$|K|^{-2/d} \|v_{h'} - \Pi_h v_{h'}\|_K^2 + \|\nabla \Pi_h v_{h'}\|_K^2 \le C \, \|\nabla v_{h'}\|_K^2. \tag{15}$$

The following technical results are stated without proofs, which can be found in [2].

Lemma 1 *The finite element spaces V_h and Q_h satisfy the following properties. For $v_h \in V_h$ and any $K \in \mathcal{K}_h$ we have $\Delta v_h|_K = 0$ and for any $S \in \mathcal{S}_h^*$ we have that $[\frac{\partial v_h}{\partial n_S}]$ is constant.*

Let $h' \in \mathcal{H}$ be a refinement of $h \in \mathcal{H}$. Then we have

$$\frac{1}{|S|}\int_S \Pi_h v_h \, ds = \frac{1}{|S|}\int_S \{v_h\} \, ds \qquad \forall S \in \mathcal{S}_{h'}^*,\ v_h \in V_h \tag{16}$$

and

$$\frac{1}{|S|}\int_S \Pi_h v_{h'} \, ds = \frac{1}{|S|}\int_S v_{h'} \, ds \qquad \forall S \in \mathcal{S}_h^*,\ v_{h'} \in V_{h'}. \tag{17}$$

Finally, there exists a mesh-independant constant $\gamma_{IS} > 0$ such that:

$$\sup_{v_{h'} \in V_h \setminus \{0\}} \frac{\langle \operatorname{div}_h v_{h'}, q_{h'} \rangle}{\|\nabla_{h'} v_{h'}\|} \geq \gamma_{IS} \|q_{h'}\| \quad \forall q_{h'} \in Q_{h'}. \tag{18}$$

Lemma 2 *Let in addition suppose that $(h', \widetilde{\mathcal{M}}) = \mathcal{R}ef(h, \mathcal{M})$. Then we have for $u \in V$*

$$\langle \nabla_{h'}(u - \Pi_{h'} u), \nabla_{h'} v_d \rangle = 0 \quad \forall v_d \in V_{h'} \oplus V_h. \tag{19}$$

The operator Π_h has the following properties: For arbitrary $u_{h'} \in V_{h'}$ there holds

$$\langle \nabla_{h'}(u_{h'} - \Pi_h u_h), \nabla_h v_h \rangle = 0 \quad \forall v_h \in V_h. \tag{20}$$

We use the a posteriori error estimator proposed in [9], consisting of a volume residual and an estimator for the nonconformity error defined on the edges. Let $K \in \mathcal{K}_h$ and $\mathcal{M} \subset \mathcal{K}_h$.

$$\eta_h(K) := |K|^{1/2} \|f\|_K, \quad \eta_h(\mathcal{M}) := \left(\sum_{K \in \mathcal{M}} \eta_h^2(K) \right)^{1/2}. \tag{21}$$

Let $S \in \mathcal{S}_h^*$. The estimator for the nonconformity involves the jump of the velocity vector and reads

$$J_h(S) := |S|^{-1/2} \|[u_h]\|_S, \quad J_h(\mathcal{M}) := \left(\sum_{K \in \mathcal{M}} \sum_{S \subset \partial K \setminus \partial\Omega} J_h^2(S) \right)^{1/2}. \tag{22}$$

In case the dependance of J_h on u_h is of importance, we write $J_h(u_h, S)$ and $J_h(u_h, \mathcal{M})$.

Remark 1 *The nonconformity estimator used in [9] reads*

$$\tilde{J}_h(S) := |S|^{1/2} \left\| [\frac{\partial u_h}{\partial t_S}] \right\|_S. \tag{23}$$

The equivalence of $\tilde{J}_h(S)$ and $J_h(S)$ follows from the weak continuity property of the nonconforming finite element space and an inverse estimate.

Note that no information on the pressure is involved in the estimator, and that the divergence of the discrete velocity field needs not to be measured, since it is zero on each cell.

Next we formulate the adaptive algorithm.

Adaptive Algorithm AFEM

---

(0) Choose parameters $0 < \theta, \sigma < 1$, $\gamma > 0$ and an initial mesh h_0, and set $n = 0$.

(1) Solve the discrete problem on mesh h_n with solution u_{h_n}.

(2) – If $J^2_{h_n}(\mathcal{K}_{h_n}) \leq \gamma\, \eta^2_{h_n}(\mathcal{K}_{h_n})$ then find a set $\mathcal{M}_n \subset \mathcal{K}_{h_n}$ with minimal cardinality such that

$$\eta^2_{h_n}(\mathcal{M}_n) \geq \theta\, \eta^2_{h_n}(\mathcal{K}_{h_n}). \tag{24}$$

– else find a set $\mathcal{M}_n \subset \mathcal{K}_{h_n}$ with minimal cardinality such that

$$J^2_{h_n}(\mathcal{M}_n) \geq \sigma\, J^2_{h_n}(\mathcal{K}_{h_n}). \tag{25}$$

(3) Adapt the mesh: $h_{n+1} := \mathcal{R}(h_n, \mathcal{M}_n)$.

(4) Set $n := n + 1$ and go to step (1).

For practical purposes, the algorithm has to be completed by a stopping criterion. Since we are interested in the analysis of the asymptotic behavior, we have skipped it here.

3 Quasi-optimality

We consider the error measure

$$\varepsilon^2_h := \sqrt{\|\nabla_h(u - u_h)\|^2 + \beta_1 \|p - p_h\|^2 + \beta_2\, \eta^2_h}, \tag{26}$$

with constants $\beta_1, \beta_2 > 0$ to be determined below. We show geometric convergence of the sequence $\{\varepsilon_k\}_{k\geq 1}$ for meshes generated by the adaptive algorithm. This implies that the $L^2(\Omega)$-error of pressure and the discrete H^1-error of velocities are bounded by a geometric series, as in the case of uniformly refined meshes under the regularity assumptions of standard a priori error analysis. In addition, we have the same result for the estimator. The following two theorems have been shown in [2].

Theorem 3 *Let $\{h_k\}_{k\geq 0}$ be a sequence of meshes generated by algorithm AFEM and let $\{(u_{h_k}, p_{h_k})\}_{k\geq 0}$ be the corresponding sequence of finite element solutions. Then there exist strictly positive β_1, β_2 and $0 < \rho < 1$ such that for all $k = 1, 2, \ldots$*

$$\varepsilon_{k+1} \leq \rho\, \varepsilon_k. \tag{27}$$

The convergence proof is based on three results: a global upper bound, a local lower bound, and an estimate for the non-orthogonality.

The following global upper bound has been established in [8] .

Lemma 4 *There exists a mesh-independant constant C_1 such that for $h \in \mathcal{H}$ and corresponding finite element solution $(u_h, p_h) \in V_h \times Q_h$*

$$\|\nabla_h(u - u_h)\|^2 + \|p - p_h\|^2 \leq C_1 \left(\eta_h^2(\mathcal{K}_h) + J_h^2(\mathcal{K}_h)\right). \tag{28}$$

For the local bound, we compare the discrete solutions belonging to two consequent meshes $h', h \in \mathcal{H}$.

Lemma 5 *Let $h \in \mathcal{H}$, $\mathcal{M} \subset \mathcal{K}_h$ and $(h', \widetilde{\mathcal{M}}) = \mathcal{R}ef(h, \mathcal{M})$. There exists a mesh-independant constant C_2 such that for the corresponding finite element solution $u_{h'} \in V_{h'}$ and $u_h \in V_h$*

$$J_h^2(\widetilde{\mathcal{M}}) \leq C_2 \|\nabla_h(u_{h'} - u_h)\|^2. \tag{29}$$

For a proof of a similar bound for the Poisson equation see Theorem 4.1 in [7]; Lemma 5 is a straightforward generalization.

The next Lemma provides an estimate for the decrease in η_h. The simple proof is omitted.

Lemma 6 *Let $h \in \mathcal{H}$, $\mathcal{M} \subset \mathcal{K}_h$ and $(h', \widetilde{\mathcal{M}}) = \mathcal{R}ef(h, \mathcal{M})$. Then there exist a mesh-independent constants $C_3 > 0$ such that*

$$\eta_{h'}^2(\mathcal{K}_{h'}) \leq \eta_h^2(\mathcal{K}_h) - C_3 \eta_h^2(\widetilde{\mathcal{M}}_h). \tag{30}$$

The following estimation of the non-orthogonality, which is at the heart of the convergence and complexity analysis has been proven in [2].

Lemma 7 *Let $' \in \mathcal{H}$, $\mathcal{M} \subset \mathcal{K}_{'}$ and $(h', \widetilde{\mathcal{M}}) = \mathcal{R}ef(', \mathcal{M})$. There exists a mesh-independant constant C_4, C_5 such that*

$$\langle \nabla_h(u - u_{h'}), \nabla_h(u_{h'} - u_h) \rangle \leq C_4\, \eta_h(\widetilde{\mathcal{M}}) \|\nabla_{h'}(u - u_{h'})\| \tag{31}$$

and

$$\langle p - p_h, p_h - p_H \rangle \leq C_5 \left(\eta_H(\widetilde{\mathcal{M}}) + \|\nabla_h(u_h - u_H)\|\right) \|p - p_h\|. \tag{32}$$

In order to express our assumptions on the regularity of the continuous solutions, we introduce some notation from nonlinear approximation theory, see [6, 10].

Let $\mathcal{H}_N$ be the set of all triangulations $h \in \mathcal{H}$ which satisfy $N_h \leq N$.

Next we define the approximation class

$$\mathcal{W}^s := \left\{(u,p,f) \in (H_0^1(\Omega)^d, L_0^2(\Omega), L^2(\Omega)^d) : \|(u,p,f)\|_{\mathcal{W}^s} < +\infty\right\} \tag{33}$$

where

$$\|(u,p,f)\|_{\mathcal{W}^s} := \sup_{N \geq N_0} N^s \inf_{h \in \mathcal{H}_N} \varepsilon_h.$$

We say that an adaptive finite element method is *quasi-optimal*, if, whenever $(u,p,f) \in \mathcal{W}^s$, it produces meshes $\{h_k\}_k$ such that $\{\varepsilon_k\}_k$ is geometrically convergent to zero and

$$\varepsilon_k \leq C\, N_k^{-s}. \tag{34}$$

Notice that the presented notion of quasi-optimality depends on the family $\mathcal{H}$ of admissible meshes.

The result on quasi-optimality of the adaptive algorithm is formulated next.

Theorem 8 *Under the condition that $0 < \theta < 1$ is small enough there exist $\beta_1 > 0$ and $\beta_2 > 0$ in the definition of the error (26) such that the algorithm AFEM is quasi-optimal.*

The proof given in [2] makes essential use of the estimation of non-orthogonality stated in Lemma 7. In addition, it requires a local upper bound and a global lower bound, which we establish first.

The global lower bound is a simple variant of its local counterpart, Lemma 5. The proof can be found in [8].

Lemma 9 *There exists a mesh-independent constant $C_6 > 0$ such that for the finite element solution (u_h, p_h) of (3), we have*

$$J_h^2(\mathcal{K}_h) \leq C_6 \left(\|\nabla_h(u-u_h)\|^2 + \|p-p_h\|^2\right). \tag{35}$$

The local upper bound is expressed next.

Lemma 10 *There exist a mesh-independent constant $C_7 > 0$ such that the following holds. Let $h' \in \langle$ be obtained as the local refinement of $h \in \mathcal{H}$ with a set of refined cells $\widetilde{\mathcal{M}} \subset \mathcal{K}_h$. Then the finite element solutions (u_h, p_h) and (u_H, p_H) on the two meshes verify:*

$$\frac{1}{2}\|\nabla_h(u_{h'} - u_h)\|_h^2 + \|p_{h'} - p_h\|^2 \leq C_7 \left(\eta_h^2(\widetilde{\mathcal{M}}) + J_h^2(\widetilde{\mathcal{M}})\right). \tag{36}$$

4 Extension to non-homogeneuous Dirichlet and Neumann-type boundary conditions

Let now $g \in H^{1/2}(\Gamma_D)^2$ be arbitrary and $\Gamma_N \subset \partial\Omega$ a non-degenerate boundary segment such that $|\Gamma_D| > 0$. We suppose that the finite element meshes match this partition of the boundary. Then there exists a divergence-free vector field $u_g \in H^1(\Omega)^2$ such that $\gamma_D(u) = g$ with the trace operator γ_D. Letting

from now on $V := \left\{v \in H^1(\Omega)^2 : \ \gamma_D(u) = 0\right\}$ and $Q = L^2(\Omega)$, the considered weak formulation of (1) reads: Find $(u, p) \in (u_g + V) \times Q$ such that for all $(v, q) \in V \times Q$ there holds:

$$\langle \nabla u, \nabla v\rangle - \langle p, \operatorname{div} v\rangle + \langle \operatorname{div} u, q\rangle = \langle f, v\rangle. \tag{37}$$

In order to extend our discretization, we replace the definition of (10). Let $\mathcal{S}_h^D$ be the set of edges on Γ_D. Then we set

$$\begin{aligned} V_h \quad := \quad \Big\{ & v_h \in L^2(\Omega)^d : \ v_h|_K \in Q^1(\mathbb{R}^2)^2 \quad \text{for all } K \in \mathcal{K}_h \text{ and} \\ & \int_S [v_h]\, ds = 0 \quad \text{for all } S \in \mathcal{S}_h^* \cup \mathcal{S}_h^D \Big\}. \end{aligned} \tag{38}$$

The definition of Q_h is unchanged beside the fact that the mean-zero is no longer imposed.

We now construct an approximation $u_{g,h}$ of u_g by imposing

$$\int_S u_{g,h}\, ds = \int_S u_g\, ds \quad \forall S \in \mathcal{S}_h. \tag{39}$$

The discrete approximation of (2) reads: Find $(u_h, p_h) \in (u_{g,h} + V) \times Q$ such that for all $(v_h, q_h) \in V \times Q$ there holds:

$$\langle \nabla_h u_h, \nabla v_h\rangle - \langle p_h, \operatorname{div}_h v_h\rangle + \langle \operatorname{div}_h u_h, q_h\rangle = \langle f, v_h\rangle. \tag{40}$$

In order to derive the error estimator for triangular meshes, we define the auxiliary problem: Find $(u^h, p^h) \in (u_g + V, Q)$ such that for all $(v, q) \in V \times Q$

$$\langle \nabla u^h, \nabla v\rangle - \langle p^h, \operatorname{div} v\rangle + \langle \operatorname{div} u^h, q\rangle = \langle \nabla_h u_h, \nabla v\rangle - \langle p_h, \operatorname{div} v\rangle + \langle \operatorname{div}_h u_h, q\rangle. \tag{41}$$

We now split the error as

$$\begin{aligned} & \|\nabla_h(u - u_h)\| + \|p - p_h\| \\ & \le \left(\|\nabla_h(u - u^h)\| + \|p - p^h\|\right) + \left(\|\nabla_h(u^h - u_h)\| + \|p^h - p_h\|\right) \\ & = I + II. \end{aligned} \tag{42}$$

We consider the first term. In order to bound the pressure let $v \in V$. Then, by (41), we find

$$\begin{aligned} \langle p - p^h, \operatorname{div} v\rangle \quad &= \quad \langle \nabla u, \nabla v\rangle - \langle f, v\rangle - \langle \nabla u^h, \nabla v\rangle + \langle \nabla_h u_h, \nabla v\rangle - \langle p_h, \operatorname{div} v\rangle \\ &= \quad \langle \nabla(u - u^h), \nabla v\rangle - \langle f, v\rangle + \langle \nabla_h u_h, \nabla_h \Pi_h v\rangle - \langle p_h, \operatorname{div}_h \Pi_h v\rangle \end{aligned}$$

which implies by the discrete Stokes equations

$$\langle p - p^h, \operatorname{div} v\rangle = \langle \nabla(u - u^h), \nabla v\rangle - \langle f, v - \Pi_h v\rangle. \tag{43}$$

Therefore we obtain from the continuous inf-sup condition

$$\gamma \, \|p - p^h\| \le \sup_{v \in V \setminus \{0\}} \frac{\langle p - p^h, \operatorname{div} v\rangle}{\|\nabla v\|} \le C \left(\|\nabla(u - u^h)\| + \eta_h\right). \tag{44}$$

Let us consider the velocity term. Since $w := u - u^h \in V$ we have by (43) with $w = v$

$$\begin{aligned}\|\nabla(u-u^h)\|^2 &= \langle\nabla(u-u^h),\nabla w\rangle\\ &= \langle f, w-\Pi_h w\rangle + \langle p-p^h, \operatorname{div} w\rangle\\ &= \langle f, w-\Pi_h w\rangle + \langle p-p^h, \operatorname{div}_h(u-u_h)\rangle\\ &\le C\,\eta_h\|\nabla(u-u^h)\|.\end{aligned}$$

Therefore, the volume part of the estimator does not need to be changed.

We next turn our attention to the second term. It measures the non-conformity and boundary data errors as can be seen as follows. With the continuous inf-sup condition and (41) with $q = 0$ we have

$$\begin{aligned}\gamma\,\|p^h - p_h\| &\le \sup_{v\in V\setminus\{0\}} \frac{\langle p^h - p_h, \operatorname{div} v\rangle}{\|\nabla v\|}\\ &= \sup_{v\in V\setminus\{0\}} \frac{\langle \nabla_h(u^h - u_h), \nabla v\rangle}{\|\nabla v\|}\\ &\le \|\nabla_h(u^h - u_h)\|.\end{aligned}$$

Let now $w \in u_g + V$ be arbitrary. We then have by (41) with $v = w - u^h$ and $q = 0$:

$$\begin{aligned}\|\nabla_h(u^h-u_h)\|^2 &= \langle\nabla_h(u^h-u_h),\nabla_h(u^h-u_h)\rangle\\ &= \langle\nabla_h(u^h-u_h),\nabla_h(w-u_h)\rangle - \langle p^h-p_h, \operatorname{div}(w-u^h)\rangle\\ &= \langle\nabla_h(u^h-u_h),\nabla_h(w-u_h)\rangle - \langle p^h-p_h, \operatorname{div}_h(w-u_h)\rangle\\ &\le \big(\|\nabla_h(u^h-u_h)\| + \|p^h-p_h\|\big)\,\|\nabla_h(w-u_h)\|\\ &\le C\,\|\nabla_h(u^h-u_h)\|\|\nabla_h(w-u_h)\|,\end{aligned}$$

where we have used (41) with $v = 0$ and $q = p_h - p^h$ in the third line. It follows from these estimates that

$$II \le C \inf_{w\in u_g+V} \|\nabla_h(u_h - w)\|. \tag{45}$$

We therefore have to modify the edge contributions as follows

$$J_h^2 := \sum_{S\in\mathcal{S}_h} |S|^{-1}\|[u_h]\|_S^2 + \sum_{S\subset\Gamma^D} |S|^{-1}\|g-g_h\|_S^2. \tag{46}$$

The proof of quasi-optimality in Section 3 has to changed in order to take into account the additional term in (46). The details are the subject of future work.

5 Numerical experiments

We consider an example of a crossing flow. The geometry with the flow configuration is shown in Figure 2. Singularities of the continuous solution is implied by the re-entrant corners.

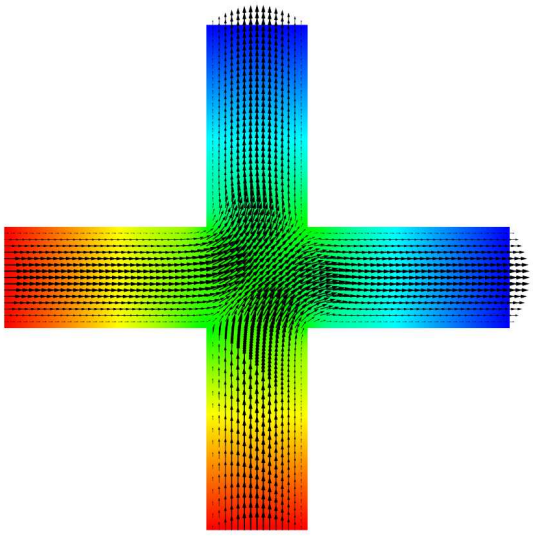


Figure 2: Domain and velocities for the crossing flow configuration.

In the following we apply the adaptive algorithm to this configuration. Typical meshes are shown in Figure 3.

A comparison of the decrease of the error estimator on the sequence of meshes in Figure 3 with uniform refinement can be seen in Figure 4.

References

[1] R. Becker and S. Mao, *An optimally convergent adaptive mixed finite element method*, Numer. Math., 111 (2008), pp. 35–54.

[2] R. Becker and S. Mao, *Quasi-optimality of adaptive non-conforming finite element methods for the stokes equations*. submitted, 2009.

[3] R. Becker, S. Mao, and Z.-C. Shi, *A convergent adaptive finite element method with optimal complexity*, Electron. Trans. Numer. Anal., 30 (2008), pp. 291–304.

[4] R. Becker, S. Mao, and Z.-C. Shi, *A convergent nonconforming adaptive finite element method with optimal complexity*. accepted for publication, 2009.

[5] R. Becker and D. Trujillo, *Convergence of an adaptive finite element method on quadrilateral meshes*, Research Report RR-6740, INRIA, 2008.

[6] P. Binev, W. Dahmen, and R. DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.

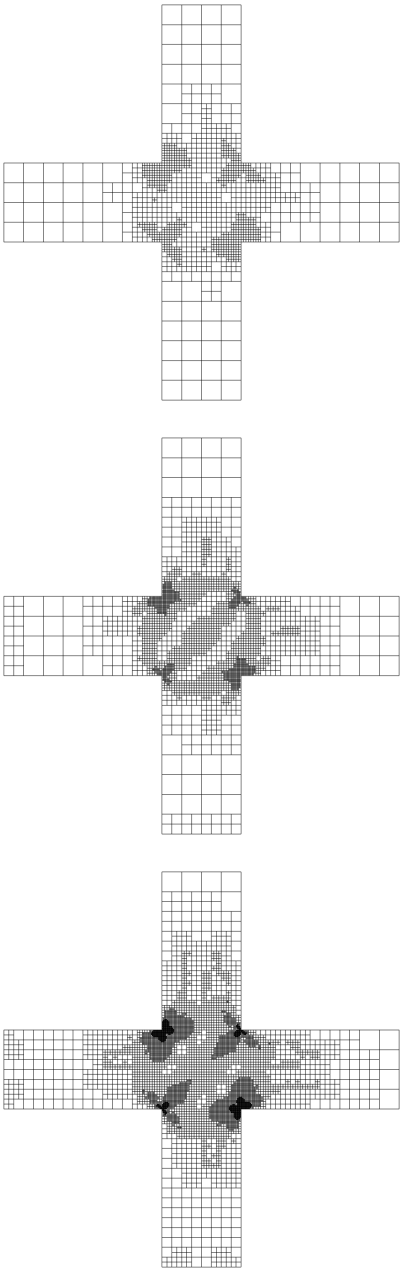


Figure 3: Sequence of locally refined meshed for the crossing flow configuration.

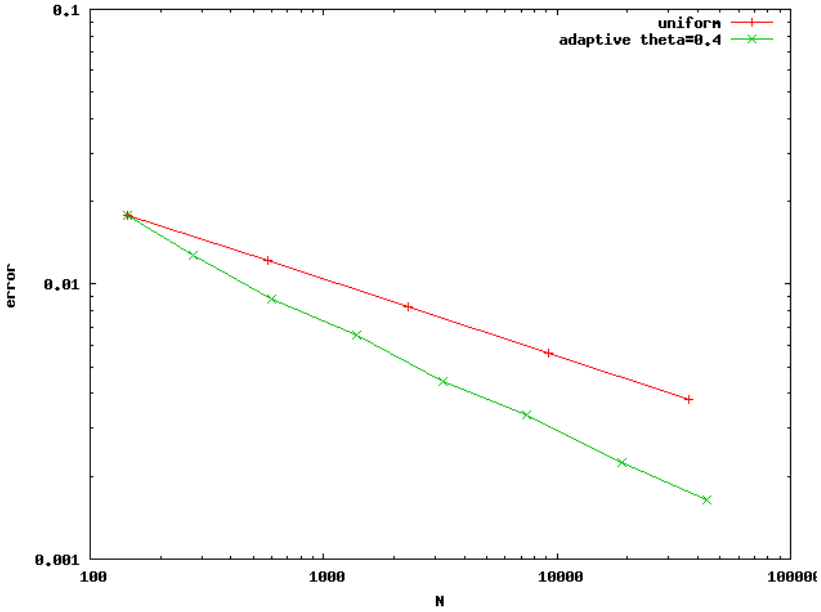


Figure 4: Comparison of adaptive and uniform refinement.

[7] C. Carstensen and R. Hoppe, *Convergence analysis of an adaptive nonconforming finite element method.*, Numer. Math., 103 (2006), pp. 251–266.

[8] C. Carstensen and J. Hu, *A unifying theory of a posteriori error control for nonconforming finite element methods*, Numer. Math., 107 (2007), pp. 473–502.

[9] E. Dari, R. Durán, and C. Padra, *Error estimators for nonconforming finite element approximations of the Stokes problem*, Math. Comp., 64 (1995), pp. 1017–1033.

[10] R. DeVore, *Nonlinear approximation.*, in Acta Numerica 1998, A. Iserles, ed., vol. 7, Cambridge University Press, 1998, pp. 51–150.

[11] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation.*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[12] W. Dörfler and M. Ainsworth, *Reliable a posteriori error control for nonconformal finite element approximation of Stokes flow*, Math. Comp., 74 (2005), pp. 1599–1619 (electronic).

[13] Y. Kondratyuk and R. Stevenson, *An optimal adaptive finite element method for the Stokes problem*, SIAM J. Numer. Anal., 46 (2008), pp. 747–775.

[14] L. D. Marini, *An inexpensive method for the evaluation of the solution of the lowest order raviart-thomas mixed method*, SIAM J Numer. Anal., 22 (1985), pp. 493–496.

[15] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM.*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[16] R. Rannacher and S. Turek, *Simple nonconforming quadrilateral stokes element*, Numerical Methods for Partial Differential Equations, (1992), pp. 97–111.

[17] R. Stevenson, *Optimality of a standard adaptive finite element method*, Found. Comput. Math., 7 (2007), pp. 245–269.

Bol. Soc. Esp. Mat. Apl.
n°50(2010), 115–138

A RECOVERY-BASED ERROR ESTIMATOR FOR ANISOTROPIC MESH ADAPTATION IN CFD

S. MICHELETTI†, S. PEROTTO†, P.E. FARRELL‡

‡ Applied Modelling and Computation Group,
Department of Earth Science and Engineering,
Imperial College London, SW7 2AZ, UK
† MOX – Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi", Politecnico di Milano
via Bonardi 9, I-20133 Milano, Italy

patrick.farrell06@imperial.ac.uk,
{stefano.micheletti,simona.perotto}@polimi.it

Abstract

We provide a unifying framework that generalizes the 2D and 3D settings proposed in [32] and [17], respectively. In these two works we propose a gradient recovery type a posteriori error estimator for finite element approximations on anisotropic meshes. The novelty is the inclusion of the geometrical features of the computational mesh (size, shape and orientation) in the estimator itself. Moreover, we preserve the good properties of recovery based error estimators, in particular their computational cheapness and ease of implementation. A metric-based optimization procedure, relying on the estimator, drives the anisotropic adaptation of the mesh. The focus of this work then moves to a goal-oriented framework. In particular, we extend the idea proposed in [32, 17] to the control of a goal functional. The preliminary results are promising, since it is shown numerically to yield quasi-optimal triangulations with respect to the error-vs-number of elements behaviour.

Key words: *anisotropic mesh adaptation; Zienkiewicz-Zhu a posteriori error estimator; computational fluid dynamics.*

AMS subject classifications: *65N15, 65N30, 65N50, 65K10*

1 Motivations

Numerical simulations in Computational Fluid Dynamics (CFD) usually demand a large computational effort due to multiple factors, such as the complexity of the phenomena involved, the unsteady character of the flows

Acknowledgement: UK Natural Environment Research Council grant NE/C52101X/1; Imperial College High Performance Computing Service; P.E. Farrell would like to thank AWE for their funding of his research through the Institute of Shock Physics

and the interest in 3D configurations. A natural objective is thus to contain the computational costs. A typical technique used to contain computational costs is to design ad-hoc meshes able to follow the dynamics of the phenomenon at hand, i.e., to adapted meshes. Most current numerical software include some adaptive capability to suitably modify the computational mesh. These usually deal with isotropic adapted meshes, namely, comprising regular elements clustered around the most critical areas for the phenomenon. Managing this kind of mesh is now rather straightforward, but a greater reduction of the computational effort can be achieved via anisotropic adapted meshes, i.e., meshes able to adapt the size as well as the shape and the orientation of the elements to the directional features of the flow (e.g., sharp fronts, shocks in compressible flows, steep boundary layers). Anisotropic mesh adaptation has proved to be a powerful strategy for improving the quality and the efficiency of flow simulations (see, e.g., [41, 21, 7] in aerospace applications, or [8] for multi-material flows in material processing applications or [36, 12, 34, 31] for modeling viscous flows around bodies).

Anisotropic adapted meshes can be obtained by employing either heuristic or theoretical approaches. In the first case one usually employs a numerical approximation of the Hessian of the solution, possibly coupled with an a priori error estimator (see, e.g., [36, 14, 40, 12, 2, 34, 15, 37, 24, 21]). Although the results are sometimes impressive, these techniques fail to link with a rigorous bound of the discretization error.
In the case of theoretical approaches, one moves from a posteriori error estimators, which can be developed both in a residual-based and in a goal-oriented framework ([3, 27, 19, 20, 41, 18]).
¿From a computational viewpoint, the theoretical approaches are in general more complicated, though they can provide optimal meshes in terms of convergence rate with respect to a quantity of interest.

In the engineering practice a very popular error estimator is the one proposed by Zienkiewicz-Zhu in 1987 for the linear elasticity problem ([43, 44, 45]). The broad diffusion of this technique is justified by its computational cheapness, ease of implementation and its good numerical performance in a huge variety of applications (see, e.g., [9, 28, 35]). Various efforts have been made to theoretically understand the amazingly good properties of the Zienkiewicz-Zhu error estimator; these have been confined to structured or mildly structured meshes ([26]). This estimator has been used mainly to drive isotropic mesh adaptation.

The computational cheapness and robustness of the Zienkiewicz-Zhu type estimator has prompted us to find a corresponding anisotropic counterpart ([32, 17]). In these works we consider piecewise linear finite elements and devise a simple recovery technique that is different from the standard one but more suited to incorporate anisotropic information. This recovery procedure leads to a Zienkiewicz-Zhu-like estimator which automatically includes the anisotropic features of the triangulation, i.e., size, stretching and orientation of each element.

In Sec.2-4 we provide a framework unifying both the 2D and the 3D settings proposed in [32] and [17], respectively, with a particular emphasis on

a challenging 3D application (Sec.4.2). In Sec.5, our focus moves to a goal-oriented framework. In this case quantities more general than the energy norm and meaningful from a physical viewpoint (pointwise stresses, fluxes, vorticity, drag or lift around bodies, etc.) can be controlled. In particular, we generalize the idea at the base of the estimators proposed in [32, 17] to the control of a goal functional. In the present setting, the choice of the functional is thoroughly general on $H_0^1(\Omega)$, whereas the differential problem coincides with the standard Poisson problem. Some conclusions are drawn in the last section.

2 Gradient recovery procedures

The approach proposed by Zienkiewicz-Zhu in [43, 44, 45] essentially consists of two steps: given an affine finite element function u_h, approximating the solution u to a certain partial differential problem, a recovery procedure for obtaining an improved approximation $P(\nabla u_h)$ of ∇u_h is first provided ([44]); then $P(\nabla u_h)$ is employed for devising an a posteriori error estimator for the H^1-seminorm of the discretization error $e_h = u - u_h$ ([45]). Standard notation is employed for the Lebesgue and Sobolev spaces ([30]).
In the sequel we aim at fitting these two steps in an anisotropic setting. This section focuses on the first step. To fix ideas, we consider the standard Poisson problem completed with homogeneous Dirichlet boundary conditions, i.e., find $u \in V \equiv H_0^1(\Omega)$, such that

$$\int_\Omega \nabla u \cdot \nabla v \, d\mathbf{x} = \int_\Omega f v \, d\mathbf{x} \quad \forall v \in V, \tag{1}$$

with Ω a polygonal/polyhedral domain in $\mathbb{R}^d$, for $d = 2, 3$, respectively, and $f \in L^2(\Omega)$.
Let $\mathcal{T}_h = \{K\}$ be a conforming partition of Ω consisting of triangles/tetrahedra and u_h be the Galerkin affine finite element approximation to (1).

Several approaches can be undertaken to pursue the gradient recovery step (see, e.g., [46, 44, 38, 1, 11, 33, 42, 6]). Essentially, they are all based on ad-hoc averagings/projections of the actual gradient ∇u_h over suitable element or nodal patches. In particular, we focus in this paper on the family of gradient recovery procedures proposed in [32, 17]: these provide a piecewise polynomial of degree r, $P^r(\nabla u_h)$, such that

$$P^r(\nabla u_h)|_{\Delta_K} \equiv P^r_{\Delta_K}(\nabla u_h) \in [\mathbb{P}_r(\Delta_K)]^d,$$

where $\Delta_K = \{T \in \mathcal{T}_h : T \cap K \neq \emptyset\}$ is the patch of elements associated with K and $\mathbb{P}_r(\Delta_K) = \text{span}\{x_1^{i_1} \ldots, x_d^{i_d} \, : \, i_1 + \ldots + i_d \leq r\}$ is the set of polynomials of (global) degree r defined on patch Δ_K. We seek $P^r_{\Delta_K}(\nabla u_h)$ such that

$$\int_{\Delta_K} (\nabla u_h - P^r_{\Delta_K}(\nabla u_h)) \cdot \mathbf{w} \, d\mathbf{x} = 0 \quad \forall \mathbf{w} \in [\mathbb{P}_r(\Delta_K)]^d. \tag{2}$$

The recovered gradient $P^r_{\Delta_K}(\nabla u_h)$ is strictly associated with K, and not to the elements comprising Δ_K: for any $T \in \Delta_K$, with $T \neq K$, $P^r_{\Delta_T}(\nabla u_h)$ is, in

general, different from $P^r_{\Delta_K}(\nabla u_h)$. In the particular case $r = 0$, we can write out the formula for the recovered gradient, given by

$$P^0_{\Delta_K}(\nabla u_h) = \frac{1}{|\Delta_K|} \sum_{T \in \Delta_K} |T| \, \nabla u_h|_T,$$

where $|\varpi|$ stands for the measure of a generic set $\varpi \subset \mathbb{R}^d$, namely we compute the area/volume-weighted average over the patch Δ_K of the gradients of the discrete solution. For $r > 0$, no simple explicit formula is available for $P^r_{\Delta_K}(\nabla u_h)$. Equation (2) leads to solve d least-squares problems of order $\binom{r+d}{d}$ for each patch Δ_K. The recovered gradient $P^r(\nabla u_h)$ is thus not conformal and the contributions due to adjacent elements overlap. Moreover, in contrast to the standard Zienkiewicz-Zhu procedure, the nodal re-interpolation step is no longer required.

3 Towards an anisotropic control of the mesh

To derive the estimator used in this work, we apply the recovery procedure described above in a convenient anisotropic setting. We choose the anisotropic framework introduced for the 2D case in [19] and extended to the 3D case in [17]. This step leads to a Zienkiewicz-Zhu-like error estimator, automatically including the anisotropic information (size, shape and orientation) of the mesh elements. The same is not so evident in the case of the standard Zienkiewicz-Zhu error estimator ([45]).

3.1 The anisotropic background

The general element $K \in \mathcal{T}_h$ is characterized geometrically by the properties of the affine map $T_K : \widehat{K} \to K$, where $\widehat{K}$ is the reference isotropic element, centred at the origin and inscribed in the unit d-sphere (see Figure 1, for the 2D case). In the 2D case, element $\widehat{K}$ coincides with the equilateral triangle centred at the origin, with coordinates $(-\sqrt{3}/2, -1/2)$, $(\sqrt{3}/2, -1/2)$, $(0, 1)$ and edge length $|\widehat{e}| = \sqrt{3}$. For the 3D case we pick the regular tetrahedron $\widehat{K}$ with coordinates $(-\sqrt{2/3}, -\sqrt{2}/3, -1/3)$, $(\sqrt{2/3}, -\sqrt{2}/3, -1/3)$, $(0, 2\sqrt{2}/3, -1/3)$, $(0, 0, 1)$ and edge length $|\widehat{e}| = 2\sqrt{2/3}$.
The affine map is defined as

$$\mathbf{x} = T_K(\widehat{\mathbf{x}}) = M_K \, \widehat{\mathbf{x}} + t_K,$$

where $M_K \in \mathbb{R}^{d \times d}$ is the Jacobian, $t_K \in \mathbb{R}^d$ is the shift vector, and $\mathbf{x} = (x_1, \ldots, x_d)^T$, $\widehat{\mathbf{x}} = (\widehat{x}_1, \ldots, \widehat{x}_d)^T \in \mathbb{R}^d$. The explicit expression of M_K and t_K is given by

$$M_K = \frac{1}{3} \begin{bmatrix} \sqrt{3}\,(x_1^2 - x_1^1) & 2\,x_1^3 - x_1^1 - x_1^2 \\ \sqrt{3}\,(x_2^2 - x_2^1) & 2\,x_2^3 - x_2^1 - x_2^2 \end{bmatrix}, \quad t_K = \frac{1}{3} \begin{bmatrix} x_1^1 + x_1^2 + x_1^3 \\ x_2^1 + x_2^2 + x_2^3 \end{bmatrix}$$

in the 2D case and by

$$M_K = \frac{1}{4}\begin{bmatrix} \sqrt{6}\,(x_1^2 - x_1^1) & \sqrt{2}\,(2\,x_1^3 - x_1^1 - x_1^2) & 3\,x_1^4 - x_1^1 - x_1^2 - x_1^3 \\ \sqrt{6}\,(x_2^2 - x_2^1) & \sqrt{2}\,(2\,x_2^3 - x_2^1 - x_2^2) & 3\,x_2^4 - x_2^1 - x_2^2 - x_2^3 \\ \sqrt{6}\,(x_3^2 - x_3^1) & \sqrt{2}\,(2\,x_3^3 - x_3^1 - x_3^2) & 3\,x_3^4 - x_3^1 - x_3^2 - x_3^3 \end{bmatrix},$$

$$t_K = \frac{1}{4}\begin{bmatrix} x_1^1 + x_1^2 + x_1^3 + x_1^4 \\ x_2^1 + x_2^2 + x_2^3 + x_2^4 \\ x_3^1 + x_3^2 + x_3^3 + x_3^4 \end{bmatrix}$$

in the 3D case, where $(x_1^i, \ldots, x_d^i)$, with $i = 1, \ldots, d+1$, are the coordinates of the general element K. Matrix M_K is factorized via the polar decomposition as

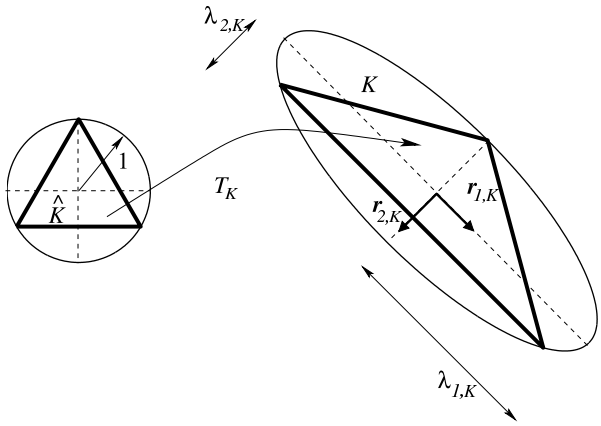


Figure 1: Geometric interpretation of the map T_K in the 2D case

$$M_K = B_K\, Z_K$$

where $B_K \in \mathbb{R}^{d\times d}$ is symmetric positive definite, and $Z_K \in \mathbb{R}^{d\times d}$ is orthogonal. Then matrix B_K is spectrally decomposed as

$$B_K = R_K^T\, \Lambda_K\, R_K,$$

where $R_K^T = [\mathbf{r}_{1,K}, \ldots, \mathbf{r}_{d,K}]$ is the eigenvector matrix and $\Lambda_K = \mathrm{diag}(\lambda_{1,K}, \ldots, \lambda_{d,K})$ is the matrix collecting the corresponding eigenvalues. The unit d-sphere circumscribing $\widehat{K}$ is changed, via T_K, into a d-ellipsoid circumscribing K. The unit vectors $\{\mathbf{r}_{i,K}\}_{i=1}^d$ identify the principal directions of the d-ellipsoid whereas each $\lambda_{i,K}$, with $i = 1, \ldots, d$, measures the length of the associated semi-axis (see Figure 1). Without loss of generality, we assume $\lambda_{1,K} \geq \ldots \geq \lambda_{d,K} > 0$, for any $K \in \mathcal{T}_h$. To characterize the shape of element K, we introduce the so-called stretching factors

$$s_{i,K} = \Big(\prod_{\substack{j=1 \\ j\neq i}}^{d} \lambda_{j,K}\Big)^{-2/d} \lambda_{i,K}^{2\,(d-1)/d} \quad \text{for} \quad i = 1, \ldots, d.$$

Notice that $s_{1,K} \geq s_{2,K} \geq \ldots \geq s_{d,K}$ and that

$$\prod_{i=1}^{d} s_{i,K} = 1. \tag{3}$$

These quantities measure the deformation of the element with respect to the isotropic case, where $s_{1,K} = s_{2,K} = \ldots = s_{d,K} = 1$.

We now state the anisotropic interpolation estimate that inspires the structure of the anisotropic estimator proposed in Sec.3.2. In particular we focus on functions $v \in H^1(\Omega)$ and we consequently adopt a Clément-like interpolant of degree 1, denoted by $I_h^1(v)$ [13, 39].

Proposition 1 *Let $v \in H^1(\Omega)$ and assume that, for any $K \in \mathcal{T}_h$,* $\mathrm{card}(\Delta_K) \leq \mathcal{D}$ *and* $\mathrm{diam}(\widehat{\Delta}_K) \leq \delta$, *where* $\mathrm{card}(\cdot)$ *stands for the cardinality,* $\mathrm{diam}(\cdot)$ *denotes the diameter, $\widehat{\Delta}_K = T_K^{-1}(\Delta_K)$ is the pullback of the patch Δ_K, $\mathcal{D} \in \mathbb{N}^*$ and $\delta \in \mathbb{R}^+$. Then there exists a constant $C = C(\mathcal{D}, \delta)$, such that*

$$\|v - I_h^1(v)\|_{L^2(K)} \leq C \Big(\sum_{i=1}^{d} \lambda_{i,K}^2 \, (\mathbf{r}_{i,K}^T G_{\Delta_K}(\nabla v)\, \mathbf{r}_{i,K}) \Big)^{1/2}, \tag{4}$$

$G_{\Delta_K}(\cdot) \in \mathbb{R}^{d \times d}$ being the symmetric semidefinite positive matrix whose general entry is given by

$$[G_{\Delta_K}(\mathbf{v})]_{l,m} = \sum_{T \in \Delta_K} \int_T v_l \, v_m \, d\mathbf{x} \quad \textit{with} \quad l, m = 1, \ldots, d, \tag{5}$$

for any vector-valued function $\mathbf{v} = (v_1, \ldots, v_d)^T \in [L^2(\Omega)]^d$.

Proof. We refer to [19] for the 2D case and to [17] for the 3D case. □

The hypotheses of Proposition 1 can be considered essentially as smoothness requirements on the mesh. They do not limit the anisotropy of every single element K; rather, they constrain the variation of $\{\mathbf{r}_{i,K}\}$ and $\{\lambda_{i,K}\}$ over patch Δ_K.

3.2 The recovery-based estimator

Proposition 1 prompts us to devise the desired anisotropic a posteriori error estimator. Let us choose $\mathbf{E}^r_{u,\Delta_K} = P^r_{\Delta_K}(\nabla u_h) - \nabla u_h|_{\Delta_K}$ as an approximation to the error $\nabla u - \nabla u_h$, over Δ_K. We define the anisotropic Zienkiewicz-Zhu-like local estimator for the H^1-seminorm of the discretization error as

$$\big[\eta^r_{K,\mathrm{A}}\big]^2 = \big(\prod_{i=1}^{d} \lambda_{i,K} \big)^{-2/d} \sum_{i=1}^{d} \lambda_{i,K}^2 \, \big(\mathbf{r}_{i,K}^T \, G_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K}) \mathbf{r}_{i,K} \big), \tag{6}$$

where matrix $G_{\Delta_K}(\cdot)$ is defined as in (5). Estimator (6) is heuristic even though some rationale can be provided. The sum in (6) is inspired by the interpolation

estimate (4) for $v = u - u_h$ and after replacing the partial derivatives of u with the corresponding components of $P^r_{\Delta_K}(\nabla u_h)$. Then product $(\prod_{i=1}^d \lambda_{i,K})^{-2/d}$ represents a scaling factor that guarantees the *consistency* of the anisotropic estimator with respect to the isotropic case: when $\lambda_{1,K} = \lambda_{2,K} = \ldots = \lambda_{d,K}$, (6) coincides with the isotropic Zienkiewicz-Zhu-like estimator ([32, 17])

$$\left[\eta^r_{K,\mathrm{I}}\right]^2 = \int_{\Delta_K} |\mathbf{E}^r_{u,\Delta_K}|^2 \, d\mathbf{x} \tag{7}$$

based on the patchwise recovered gradient (2). Moreover a *formal equivalence* between $\eta^r_{K,\mathrm{A}}$ and $|u - u_h|_{H^1(\Delta_K)}$ can be proved (see [32] for the 2D case). In the general case, for any $v \in H^1(\Omega)$, we have

$$s_{d,K}\,|v|^2_{H^1(\Delta_K)} \le \big(\prod_{i=1}^d \lambda_{i,K}\big)^{-2/d} \sum_{i=1}^d \lambda^2_{i,K}\,(\mathbf{r}^T_{i,K} G_{\Delta_K}(\nabla v)\,\mathbf{r}_{i,K}) \le s_{1,K}\,|v|^2_{H^1(\Delta_K)}.$$

Then the formal equivalence is somehow justified by replacing ∇v with $\mathbf{E}^r_{u,\Delta_K}$ in $G_{\Delta_K}(\nabla v)$, and v with $u - u_h$ in the seminorms.
The global error estimators associated with (6) and (7) are thus given by

$$\left[\eta^r_{\mathrm{A}}\right]^2 = \sum_{K\in\mathcal{T}_h} \left[\eta^r_{K,\mathrm{A}}\right]^2 \quad \text{and} \quad \left[\eta^r_{\mathrm{I}}\right]^2 = \sum_{K\in\mathcal{T}_h} \left[\eta^r_{K,\mathrm{I}}\right]^2, \tag{8}$$

respectively. Despite its heuristic derivation, estimator η^r_{A} satisfies a sort of patch test, at least in the 2D case, as proved in [32]. Moreover this estimator can be applied to more general problems, such as the elasticity or Navier-Stokes equations. In such a case one could replace, e.g., the gradient with the stress (rate) tensor ([44]). Alternately, the adaptation can be driven by the gradient of a scalar variable representative of the problem, like the pressure or the speed for the Navier-Stokes equations.
The estimator corresponding to $r = 0$ is extended to the 3D case in [17]. Here an adaptation driven by a scalar quantity (the speed for the Navier-Stokes equations and the density for a multimaterial application) is also assessed.

4 Numerical assessment

We furnish the proposed mesh adaptive procedure driven by error estimator η^r_{A}. The effectiveness of both the estimator and the adaptive procedure are then assessed on a tough 3D test case.

4.1 A metric-based mesh generation procedure

We employ a *metric-based* adaptive procedure in a predictive fashion. Two opposite criteria are typically employed to construct the adapted grid: a) given a number of elements, one looks for the mesh minimizing the discretization error; b) given an accuracy of the numerical solution, one seeks the mesh with the least number of elements. We here focus on approach b). For this purpose we

build a mesh that is optimal with respect to a matching condition involving a suitable metric. This mesh is obtained via an iterative procedure that processes intermediate tentative meshes.

The concepts of a metric and a mesh are closely related. Essentially, a metric is a practical tool that allows one to identify a certain mesh. In the isotropic case, to define a mesh, it suffices to prescribe the size of every element, throughout the domain Ω. In the anisotropic case, the size as well as the shape and the orientation of each element have to be characterized. This may be accomplished by specifying, via a suitable metric tensor, the characteristic lengths of the element which vary according to position and direction.
More precisely, with a given mesh $\mathcal{T}_h$, we associate a metric, i.e., a symmetric positive-definite tensor field $\widetilde{M}_{\mathcal{T}_h} : \Omega \to \mathbb{R}^{d\times d}$ (see, e.g., [22]). We assume $\widetilde{M}_{\mathcal{T}_h}$ piecewise constant on $\mathcal{T}_h$, such that $\widetilde{M}_{\mathcal{T}_h}|_K = \widetilde{M}_K = B_K^{-2} = R_K^T \Lambda_K^{-2} R_K$, for any $K \in \mathcal{T}_h$, where matrices R_K and Λ_K are exactly defined as in Sec.3.1. With respect to this metric, any element K is equilateral, i.e., $\left(\mathbf{e}^T \widetilde{M}_K\, \mathbf{e}\right)^{1/2} = |\widehat{e}|$, with $\mathbf{e}$ the (arbitrarily oriented) vector identifying any edge e of K and $|\widehat{e}|$ the edge length of the reference element.
On the other hand, let now $\widetilde{M}$ be a given metric. We first diagonalize the tensor field $\widetilde{M}$, ideally for every $\mathbf{x} \in \Omega$, as $\widetilde{M} = \widetilde{R}^T \widetilde{\Lambda}^{-2} \widetilde{R}$, with $\widetilde{\Lambda} = \mathrm{diag}(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_d)$ and $\widetilde{R}^T = [\widetilde{\mathbf{r}}_1, \ldots, \widetilde{\mathbf{r}}_d]$ a positive diagonal and an orthogonal matrix, respectively. We then approximate quantities $\{\widetilde{\lambda}_i\}$, $\{\widetilde{\mathbf{r}}_i\}$ via piecewise constants over a tentative mesh $\mathcal{T}_h$, and denote these quantities by $\overline{\mathbf{r}}_{i,K} \in \mathbb{R}^d$, $\overline{\lambda}_{i,K} \in \mathbb{R}$, for any $K \in \mathcal{T}_h$ and with $i = 1, \ldots, d$. For example, this can be carried out by averaging the pointwise functions $\widetilde{\mathbf{r}}_i, \widetilde{\lambda}_i$ over K. The averaged quantities define a piecewise constant metric, say $\overline{M}_{\mathcal{T}_h}$.
Thus we state

Definition 1 *The mesh $\mathcal{T}_h$ matches $\widetilde{M}$ if, for any $K \in \mathcal{T}_h$, $\widetilde{M}_{\mathcal{T}_h}|_K = \overline{M}_{\mathcal{T}_h}|_K$.*

In the spirit of a predictive procedure the tensor field $\widetilde{M}$ represents the actual unknown. At each iteration of the adaptive process, say j, we deal with three quantities:

i) the actual mesh $\mathcal{T}_h^{(j)}$;

ii) the predicted metric $\widetilde{M}^{(j+1)}$ computed on $\mathcal{T}_h^{(j)}$ and piecewise constant;

iii) the updated mesh $\mathcal{T}_h^{(j+1)}$ matching $\widetilde{M}^{(j+1)}$.

In more detail, at each step j, first problem (1) is solved on $\mathcal{T}_h^{(j)}$; then $\widetilde{M}^{(j+1)}$ is built elementwise moving from estimator η_{A}^r and by solving suitable local optimization problems (one for each $K \in \mathcal{T}_h^{(j)}$). Then the new mesh $\mathcal{T}_h^{(j+1)}$ is built via $\widetilde{M}^{(j+1)}$ and the matching condition. For this purpose, for the 2D case we rely on the command `adaptmesh` available in the package `FreeFem++` [25]. On the other hand in the 3D case this task is accomplished by the

mesh optimisation procedure described in [34]. A quality function is defined for each element, measuring its conformity to the ideal element described by the tensor field (both size and shape). The quality of the mesh is defined to be the quality of the worst element within the mesh. Iterations of optimisation procedures such as edge collapsing, edge splitting, edge and face swapping and node movement are then applied and every operation is accepted if the quality of the mesh improves. This procedure is applied until the mesh quality function satisfies a user-specified threshold. For background reading on mesh optimisation procedures, see [22].

For practical reasons, in both the 2D and the 3D case, metric $\widetilde{M}^{(j+1)}$ is averaged nodewise before being passed to `FreeFem++` and to the mesh optimisation procedure respectively since they both take as input a piecewise-linear representation of the tensor field. This nodewise averaging can change the desired number of elements of the mesh encoded by the tensor field. Empirically, we found that rescaling the averaged piecewise-linear representation to match the expected number of elements of the piecewise-constant representation is important for the convergence of the adaptive procedure.

Concerning the local optimization procedure involved in point ii), it consists first in minimizing estimator $[\eta^r_{K,\mathrm{A}}]^2$ in (6) with respect to stretching and orientation, and then in computing the actual value of $\lambda_{1,K},\ldots,\lambda_{d,K}$ by an equidistribution criterion. For the purpose of minimization, we rewrite the local anisotropic estimator as

$$\begin{aligned}[\eta^r_{K,\mathrm{A}}]^2 &= \sum_{i=1}^{d} s_{i,K}\,(\mathbf{r}_{i,K}^T G_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K})\mathbf{r}_{i,K}) \\ &= \Big(\prod_{i=1}^{d}\lambda_{i,K}\Big)|\widehat{\Delta}_K|\sum_{i=1}^{d} s_{i,K}\,(\mathbf{r}_{i,K}^T \widehat{G}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K})\mathbf{r}_{i,K}), \qquad (9)\end{aligned}$$

where $\widehat{G}_{\Delta_K}(\cdot)$ is the scaled matrix $G_{\Delta_K}(\cdot)/|\Delta_K|$, and $\widehat{\Delta}_K$ is the patch defined as in Proposition 1. Thus it holds $|\Delta_K| = \Big(\prod_{i=1}^{d}\lambda_{i,K}\Big)\,|\widehat{\Delta}_K|$.

The idea behind expression (9) is that we have singled out the area/volume information (the term before the summation) from quantities that just depend on orientation and stretching. We can now state a result about the minimization of the terms involved in the summation, which collects, in a general format, the results obtained for the 2D and the 3D case, separately.

Proposition 2 *Let*

$$\mathcal{J}(\{s_{i,K},\mathbf{r}_{i,K}\}_{i=1,\ldots,d}) = \sum_{i=1}^{d} s_{i,K}\,(\mathbf{r}_{i,K}^T \widehat{G}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K})\mathbf{r}_{i,K}), \qquad (10)$$

and let $\{\mathbf{g}_i, g_i\}$, *for* $i = 1,\ldots,d$, *denote the eigen-pairs associated with* $\widehat{G}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K})$, *where it is understood that* $g_1 \geq \ldots \geq g_d > 0$ *as well as that the*

set $\{\mathbf{g}_i\}$ *defines an orthonormal frame. Under constraint* (3) *and that* $\{\mathbf{r}_{i,K}\}$ *are orthonormal,* $\mathcal{J}(\cdot)$ *is minimized when*

$$s_{i,K} = \Big(\prod_{i=1}^{d} g_i\Big)^{1/d} g_{d+1-i}^{-1}\,, \quad \mathbf{r}_{i,K} = \mathbf{g}_{\,d+1-i} \quad \text{for} \quad i = 1,\dots,d. \tag{11}$$

The corresponding optimal values for $\{\lambda_{i,K}\}$ *are thus given by*

$$\lambda_{i,K} = \left(\frac{\tau^2}{d\,\mathrm{card}(\mathcal{T}_h^{(j)})\,|\widehat{\Delta}_K|}\right)^{1/d} \left(\prod_{i=1}^{d} g_i\right)^{(d-2)/(2d^2)} g_{d+1-i}^{-1/2} \quad \text{for} \quad i = 1,\dots,d, \tag{12}$$

where τ *is the user-defined global tolerance on the* H^1*-seminorm of the discretization error.*

Proof. We refer to Proposition 2 in [32] for the 2D case and to Proposition 4.1 in [17] for the 3D case. □

Remark 1 *Computation of* $\{\lambda_{i,K}\}$ *requires a global criterion, i.e., the equidistribution criterion such that*

$$[\eta_{K,A}^r]^2 = \frac{\tau^2}{\mathrm{card}(\mathcal{T}_h^{(j)})}\,.$$

Notice that, due to the predictive fashion of the adaptive procedure, we are equidistributing the error with respect to the background mesh $\mathcal{T}_h^{(j)}$*.*

Remark 2 *The optimal* $\{\mathbf{r}_{i,K}\}$ *and* $\{s_{i,K}\}$ *in* (11) *equalize the* d *terms* $s_{i,K}\,(\mathbf{r}_{i,K}^T \widehat{G}_{\Delta_K}(\mathbf{E}_{u,\Delta_K}^r)\mathbf{r}_{i,K})$ *in* (10)*, i.e., we get*

$$s_{1,K}\,g_d = s_{2,K}\,g_{d-1} = \ldots = s_{d,K}\,g_1.$$

These equalities yield

$$\sum_{i=1}^{d} s_{i,K}\,(\mathbf{r}_{i,K}^T \widehat{G}_{\Delta_K}(\mathbf{E}_{u,\Delta_K}^r)\mathbf{r}_{i,K}) = d\Big(\prod_{i=1}^{d} g_i\Big)^{1/d},$$

that is, functional $\mathcal{J}(\cdot)$ *in* (10) *does not depend any longer on the stretching factors* $s_{i,K}$ *on the optimized mesh. Although we do not have a rigorous proof, we expect this property as well as the global equidistribution principle cited in Remark 1, to found the robustness of estimator* (8)*.*

Remark 3 *The hypothesis* $g_1 \geq \ldots \geq g_d > 0$ *in Proposition 2 can be relaxed by assuming* $g_1 \geq \ldots \geq g_d \geq 0$*. This is the case when* $\widehat{G}_{\Delta_K}(\mathbf{E}_{u,\Delta_K}^r)$ *is positive semidefinite. This degenerate case can be tackled by defining a minimum value for* $\{g_i\}$*, given by*

$$g_{min} = \frac{h_\Omega^{-d}\,\tau^2}{d\,\mathrm{card}(\mathcal{T}_h^{(j)})\,|\widehat{\Delta}_K|}\,,$$

where $h_\Omega \equiv \mathrm{diam}(\Omega)$. *The eigenvalues are thus selected such that* $g_i = \max(g_i, g_{min})$, *for* $i = 1, \ldots, d$. *If all the eigenvalues* $\{g_i\}$ *are degenerate, then, from* (12), $\lambda_{i,K}$ *is equal to* h_Ω, *for* $i = 1, \ldots, d$.

Proposition 2 allows us to identify elementwise the optimal metric $\widetilde{M}_K$ given by $\widetilde{M}_K = R_K^T \Lambda_K^{-2} R_K$, with $R_K^T = [\mathbf{r}_{1,K}, \ldots, \mathbf{r}_{d,K}]$ and $\Lambda_K = \mathrm{diag}(\lambda_{1,K}, \ldots, \lambda_{d,K})$, where $\{\mathbf{r}_{i,K}\}$ and $\{\lambda_{i,K}\}$ are defined in (11) and (12), respectively. The corresponding nodewise metric is thus given by

$$\widetilde{M}_N = \frac{1}{|\widehat{e}|^2 \, |\Delta_N|} \sum_{K \in \Delta_N} |K| \, \widetilde{M}_K,$$

where N is a generic node of the mesh $\mathcal{T}_h^{(j)}$ while Δ_N is the patch of the elements that share node N. The scaling factor $|\widehat{e}|^2$ shrinks the reference element to a unit edge one.

4.2 First test case: a 3D backward-facing step

The backward-facing step is a popular problem for investigating the simulation of the separation and reattachment of turbulent flows, as accurate experimental results for a wide range of flow regimes exist ([4]). For a review of the use of the backward-facing step case, in particular for comparative studies of different strategies for numerical simulation, see [10, §8.7.2].

A simulation of the three-dimensional backward-facing step at Reynolds number $Re = 10^3$ (using the definition of [4]) is performed on 4 processors. The geometry follows that of [29] and is shown in Figure 2. In this simulation, the length scales are given by $L_x = 30$, $L_i = 10$, $L_y = 4$, $h = 1$, and $L_z = 6$. The base of the domain is located at $z = 0$, the inflow plane is given by $\{x = -10\}$ with the step at $\{x = 0\}$, and the back of the domain in the spanwise direction is given by $\{y = 0\}$. An inflow boundary condition is imposed at the left hand boundary. An outflow boundary condition (homogenous Neumann on velocity and homogenous Dirichlet on pressure) is located at the right hand end of the geometry. A no slip boundary condition is imposed on the bottom boundary and no normal flow is imposed at the top and lateral boundaries.

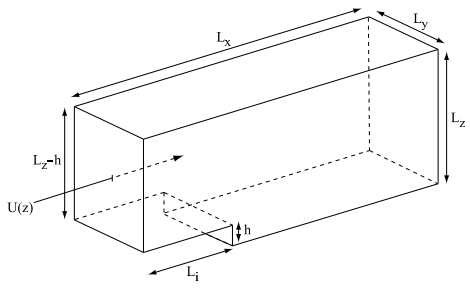


Figure 2: First test case: geometry of the computational domain

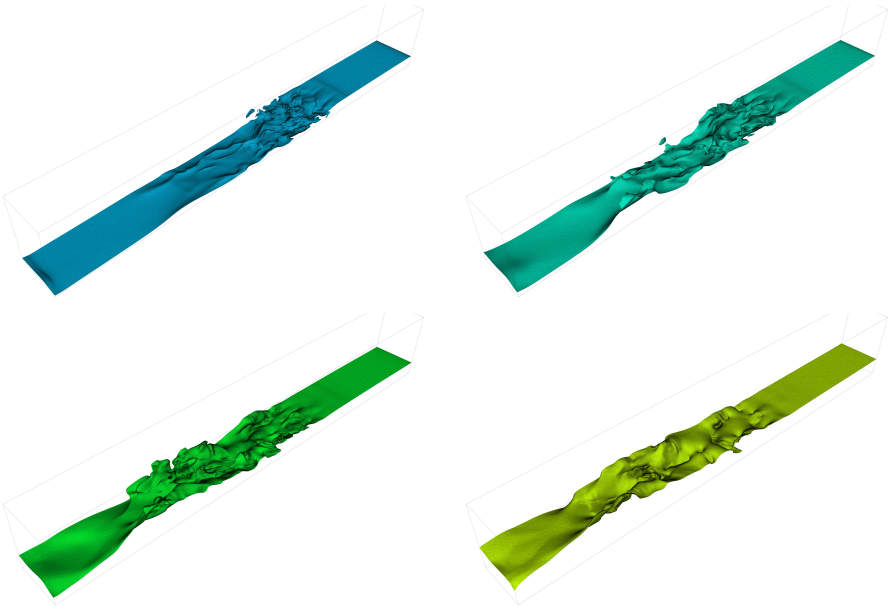


Figure 3: First test case: speed isosurfaces at 0.25 (top left), 0.5 (top right), 0.75 (bottom left) and 1 (bottom right)

The stabilised P1-P1 element pair is used to discretize velocity and pressure. The Crank-Nicolson time-stepping scheme is used, with the time-step adaptively chosen to keep the CFL number below 4. The mesh is adapted every 20 time-steps, and is limited to 5 million nodes. The anisotropic metric-based mesh generation procedure detailed in Sec.4.1 is now driven by the speed of the flow, while the global tolerance τ is set to 10. Visualisations of the simulation are shown in Figure 3, where the speed isosurfaces at 0.25, 0.5, 0.75 and 1 are displayed.

The number of nodes used in the simulation increases from an initial mesh of 250000 nodes and levels off at approximately 3.6 million nodes. As can be seen in Figure 4, the resolution is highly concentrated in regions of dynamical interest around the step and its wake. The adaptive strategy leads to the efficient computation of the flow features of interest (see Figure 5). The mesh is quite anisotropic before the step, but becomes isotropic in the wake.

5 An application to a 2D goal-oriented framework

When one is interested in controlling physically meaningful quantities, the main overhead consists of solving the dual problem corresponding to the selected goal functional (see, e.g., [16, 5, 23]). In this section we propose a naïve approach which extends the philosophy of estimator η_{A}^{r} to a goal-oriented framework. In the present setting, the choice of the functional is thoroughly general,

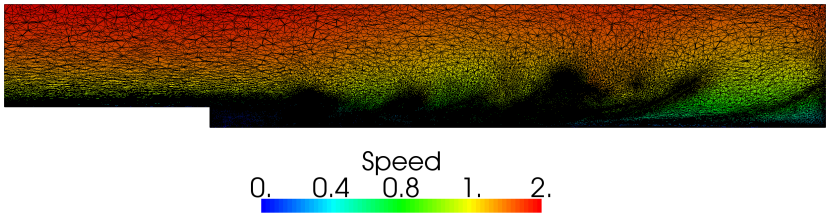


Figure 4: First test case: vertical half slice of the domain showing the final adapted mesh

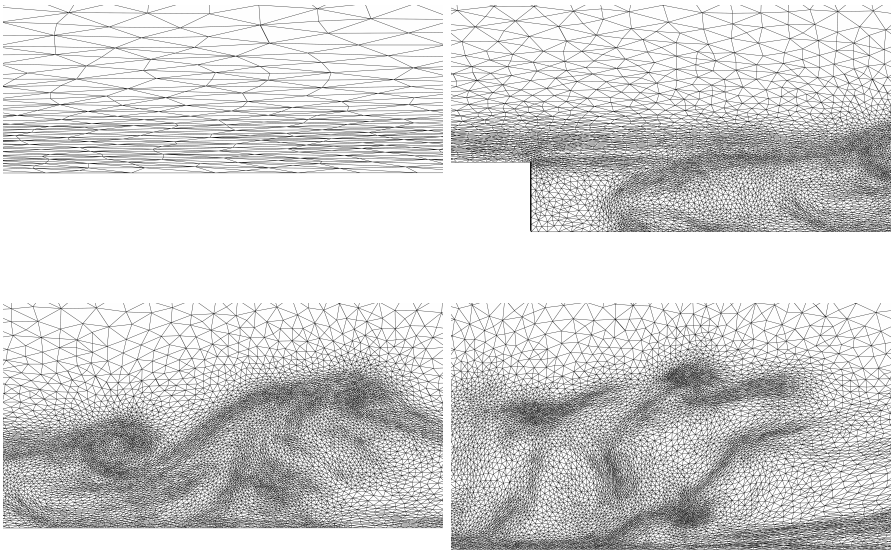


Figure 5: First test case: zooms of the final mesh on a vertical half slice of the domain

thanks to the Riesz representation theorem, whereas the differential problem is constrained to coincide with the standard Poisson problem. This choice is mandatory to exploit the Galerkin orthogonality property, i.e., to ensure the optimal convergence rate of the estimator.
Thus let $J(\cdot) : H_0^1(\Omega) \to \mathbb{R}$ be the linear and continuous functional we aim to control. Thanks to the Riesz representation theorem, functional J can be rewritten as

$$J(\varphi) = \int_\Omega \nabla\varphi \cdot \nabla g \, d\mathbf{x} \quad \forall \varphi \in H_0^1(\Omega), \tag{13}$$

with $g \in H_0^1(\Omega)$ the Riesz representant. Different choices of g lead to different quantities of interest. Choosing $\varphi = e_h$ in (13) and employing the Galerkin

orthogonality of problem (1), we have

$$J(e_h) = \int_\Omega \nabla e_h \cdot \big(\nabla g - \nabla v_h\big)\, d\mathbf{x} = \int_\Omega \big(\nabla u - \nabla u_h\big) \cdot \big(\nabla g - \nabla v_h\big)\, d\mathbf{x}, \quad (14)$$

where v_h belongs to the affine finite element space. In particular, we select v_h as the affine finite element Lagrange interpolant of g, denoted as Π_g.
Representation (14) prompts us to define the proposed local goal-oriented error estimator. Inspired by (7) we introduce

$$\eta_{K,\mathrm{I}}^{goal,r} = \frac{|K|}{|\Delta_K|} \int_{\Delta_K} \mathbf{E}^r_{u,\Delta_K} \cdot \mathbf{E}^r_{g,\Delta_K}\, d\mathbf{x}, \quad (15)$$

where $\mathbf{E}^r_{g,\Delta_K} = P^r_{\Delta_K}(\nabla \Pi_g) - \nabla \Pi_g|_{\Delta_K}$. By comparing $\eta_{K,\mathrm{I}}^{goal,r}$ with (14), we observe that ∇u and ∇g are replaced by $P^r_{\Delta_K}(\nabla u_h)$ and $P^r_{\Delta_K}(\nabla \Pi_g)$, respectively where $P^r_{\Delta_K}$ is defined as in (2). Moreover, the scaling factor $|K|/|\Delta_K|$ represents a variant to (7), justified by some recent numerical results, characterized by a sharper effectivity index. The idea is to lump somehow on K the information computed on patch Δ_K via a suitable rescaling.
Mimicking the approach in Sec.3.2, the anisotropic counterpart of (15) is thus obtained by projecting the recovered errors $\mathbf{E}^r_{u,\Delta_K}$ and $\mathbf{E}^r_{g,\Delta_K}$ along the anisotropic directions $\{\mathbf{r}_{i,K}\}$. This yields

$$\eta_{K,\mathrm{A}}^{goal,r} = \frac{|K|}{|\Delta_K|} \Big(\prod_{i=1}^{d} \lambda_{i,K}\Big)^{-2/d} \sum_{i=1}^{d} \lambda_{i,K}^2 \int_{\Delta_K} \big[\mathbf{E}^r_{u,\Delta_K} \cdot \mathbf{r}_{i,K}\big]\big[\mathbf{E}^r_{g,\Delta_K} \cdot \mathbf{r}_{i,K}\big]\, d\mathbf{x}. \quad (16)$$

To ease the numerical computation, we formulate (16) in terms of a suitable symmetric matrix $G^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K}, \mathbf{E}^r_{g,\Delta_K}) \in \mathbb{R}^{d\times d}$, given by

$$\big[G^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K}, \mathbf{E}^r_{g,\Delta_K})\big]_{lm} = \frac{|K|}{|\Delta_K|} \int_{\Delta_K} \mathrm{sym}\Big(\big[\mathbf{E}^r_{u,\Delta_K}\big]_l \big[\mathbf{E}^r_{g,\Delta_K}\big]_m\Big)\, d\mathbf{x},$$

with $l, m = 1, \ldots, d$, where operator $\mathrm{sym}(v_l w_m) = 0.5\ (v_l w_m + v_m w_l)$, for any vector $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, symmetrizes the matrix. Matrix $G^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K}, \mathbf{E}^r_{g,\Delta_K})$ allows us to replace (16) with

$$\widetilde{\eta}_{K,\mathrm{A}}^{goal,r} = \Big(\prod_{i=1}^{d} \lambda_{i,K}\Big)^{-2/d} \sum_{i=1}^{d} \lambda_{i,K}^2 \big(\mathbf{r}_{i,K}^T\, |G^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K}, \mathbf{E}^r_{g,\Delta_K})|\, \mathbf{r}_{i,K}\big), \quad (17)$$

where the modulus matrix is considered to bypass the cases when $G^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K}, \mathbf{E}^r_{g,\Delta_K})$ is indefinite. The global estimator $\widetilde{\eta}_{\mathrm{A}}^{goal,r}$ associated with (17) is then defined in a straightforward way by summing the local contributions.

A procedure analogous to the one exploited in Sec.4.1 to build up the optimal metric $\widetilde{M}^{(j+1)}$ can be applied also in this case. For this purpose we first scale

matrix $|G^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K}, \mathbf{E}^r_{g,\Delta_K})|$ and use the stretching factors $\{s_{i,K}\}$, to get

$$\begin{aligned}\widetilde{\eta}^{goal,r}_{K,\mathrm{A}} &= \Big(\prod_{i=1}^{d}\lambda_{i,K}\Big)^{-2/d}|K|\sum_{i=1}^{d}\lambda^2_{i,K}\big(\mathbf{r}^T_{i,K}\,|\widehat{G}^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K},\mathbf{E}^r_{g,\Delta_K})|\,\mathbf{r}_{i,K}\big)\\ &= |K|\sum_{i=1}^{d}s_{i,K}\big(\mathbf{r}^T_{i,K}\,|\widehat{G}^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K},\mathbf{E}^r_{g,\Delta_K})|\,\mathbf{r}_{i,K}\big)\\ &= \Big(\prod_{i=1}^{d}\lambda_{i,K}\Big)|\widehat{K}|\sum_{i=1}^{d}s_{i,K}\,\big(\mathbf{r}^T_{i,K}\,|\widehat{G}^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K},\mathbf{E}^r_{g,\Delta_K})|\,\mathbf{r}_{i,K}\big),\end{aligned}$$

where $\widehat{G}^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K},\mathbf{E}^r_{g,\Delta_K}) = G^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K},\mathbf{E}^r_{g,\Delta_K})/|K|$. Notice that the matrix is now scaled by the area of K instead of Δ_K.
Proposition 2 carries over the current goal-oriented case. The optimal recipes identifying metric $\widetilde{M}^{(j+1)}$ are now provided by

$$\begin{aligned}&s_{i,K} = \Big(\prod_{i=1}^{d}\tilde{g}_i\Big)^{1/d}\tilde{g}^{-1}_{d+1-i}\,,\quad \mathbf{r}_{i,K} = \tilde{\mathbf{g}}_{d+1-i}\quad \text{for}\quad i=1,\dots,d,\\ &\lambda_{i,K} = \left(\frac{\tau}{d\,\mathrm{card}(\mathcal{T}^{(j)}_h)\,|\widehat{K}|}\right)^{1/d}\left(\prod_{i=1}^{d}\tilde{g}_i\right)^{(d-2)/(2d^2)}\tilde{g}^{-1/2}_{d+1-i}\quad \text{for}\quad i=1,\dots,d,\end{aligned}\tag{18}$$

where $\{\tilde{\mathbf{g}}_i,\tilde{g}_i\}$, for $i = 1,\dots,d$, are the eigen-pairs associated with $|\widehat{G}^{goal}_{\Delta_K}(\mathbf{E}^r_{u,\Delta_K},\mathbf{E}^r_{g,\Delta_K})|$, with $\tilde{g}_1 \geq \dots \geq \tilde{g}_d > 0$ and $\{\tilde{\mathbf{g}}_i\}$ define an orthonormal frame. The degenerate case can be dealt with the same approach as in Remark 3.

In the next sections we assess the robustness of estimator $\widetilde{\eta}^{goal,r}_{\mathrm{A}}$ and of recipes (18) on two 2D problems.

5.1 Second test case

We solve problem (1) on $\Omega = (0,1)^2$, with f chosen such that $u(x_1,x_2) = \sin(\pi\,x_1)\,\sin(\pi\,x_2)$. Function g in (13) is the bubble function $g(x_1,x_2) = x_1\,x_2\,(x_1-1)\,(x_2-1)$, which identifies a functional of interest involving all the domain. We apply the adaptive procedure described above by enforcing a relative tolerance, τ_{rel}, such that $\tau = \tau_{\mathrm{rel}}\,|J(u)|$, where the exact value of the functional is $J(u) = 32/\pi^4$, while τ is the global tolerance in (18). As stopping criterion for the adaptive procedure we exploit the following double check:

$$\frac{\widetilde{\eta}^{goal,r}_{\mathrm{A}}}{|J(u_h)|} \leq 1.1\,\tau_{\mathrm{rel}}\quad \&\quad \frac{|\mathrm{card}(\mathcal{T}^{(j)}_h) - \mathrm{card}(\mathcal{T}^{(j+1)}_h)|}{\mathrm{card}(\mathcal{T}^{(j)}_h)} \leq 0.02.\tag{19}$$

The first check is an accuracy requirement which is relaxed by 10%; the second control ensures a sort of stagnation of the mesh cardinality before stopping the adaptive procedure.

Figure 6 and 7 gather the final adapted grids for the case $r = 0$ and $r = 1$,

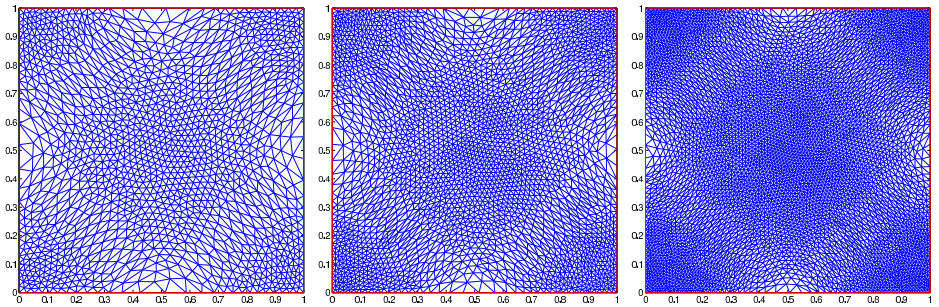


Figure 6: Second test case: final adapted meshes for $r = 0$, $\tau_{\rm rel} = 10^{-2}/2, 10^{-2}/4, 10^{-2}/8$ (from left to right)

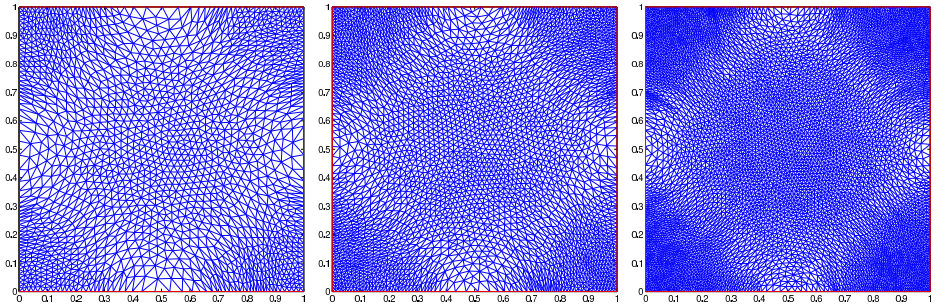


Figure 7: Second test case: final adapted meshes for $r = 1$, $\tau_{\rm rel} = 10^{-2}/16, 10^{-2}/32, 10^{-2}/64$ (from left to right)

respectively for different choices of the relative tolerance. The convergence is reached after a maximum of 5 iterations in all the cases. Although both u and g are very smooth functions, the adapted meshes exhibit a moderate stretching.

In Tables 1 and 2 a more quantitative analysis is provided. In particular the

Table 1: Second test case: convergence history for $r = 0$

$\tau_{\rm rel}$	$\mathrm{card}(\mathcal{T}_h)$	$\max s_{1,K}$	$\lvert J(u-u_h)\rvert/\lvert J(u)\rvert$	$\widetilde{\eta}_{\rm A}^{goal,0}/\lvert J(u_h)\rvert$	$\mathrm{E.I.}_{\rm A}^{goal,0}$
10^{-2}	1350	4.70	$4.643 \cdot 10^{-3}$	$9.630 \cdot 10^{-3}$	2.07
$10^{-2}/2$	2717	5.02	$2.496 \cdot 10^{-3}$	$4.764 \cdot 10^{-3}$	1.91
$10^{-2}/4$	5243	6.39	$1.380 \cdot 10^{-3}$	$2.595 \cdot 10^{-3}$	1.88
$10^{-2}/8$	10707	6.32	$6.463 \cdot 10^{-4}$	$1.297 \cdot 10^{-3}$	2.01

columns collect the values of: the relative tolerance $\tau_{\rm rel}$, the cardinality $\mathrm{card}(\mathcal{T}_h)$ of the mesh, the maximum value of the stretching factor $s_{1,K}$, the relative error $|J(u-u_h)|/|J(u)|$ on the goal functional, the relative estimator $\widetilde{\eta}_{\rm A}^{goal,r}/|J(u_h)|$,

Table 2: Second test case: convergence history for $r = 1$

τ_{rel}	$\text{card}(\mathcal{T}_h)$	$\max s_{1,K}$	$\lvert J(u-u_h)\rvert/\lvert J(u)\rvert$	$\widetilde{\eta}_{\text{A}}^{goal,1}/\lvert J(u_h)\rvert$	$\text{E.I.}_{\text{A}}^{goal,1}$
$10^{-2}/16$	3237	8.85	$2.155 \cdot 10^{-3}$	$6.030 \cdot 10^{-4}$	0.280
$10^{-2}/32$	6038	9.06	$1.180 \cdot 10^{-3}$	$3.154 \cdot 10^{-4}$	0.267
$10^{-2}/64$	12017	11.60	$6.368 \cdot 10^{-4}$	$1.530 \cdot 10^{-3}$	0.240
$10^{-2}/128$	23325	17.37	$3.478 \cdot 10^{-4}$	$7.693 \cdot 10^{-5}$	0.221

and the value of the effectivity index

$$\text{E.I.}_{\text{A}}^{goal,r} = \frac{\widetilde{\eta}_{\text{A}}^{goal,r}/\lvert J(u_h)\rvert}{\lvert J(u-u_h)\rvert/\lvert J(u)\rvert}\,.$$

All the quantities in the tables are referred to the last adapted mesh. Estimator $\widetilde{\eta}_{\text{A}}^{goal,0}$ overestimates the exact error just a little bit. On the other hand $\widetilde{\eta}_{\text{A}}^{goal,1}$ is underestimating even though the expected linear order of convergence is guaranteed as shown in Figure 8. Notice that the relative tolerances chosen for $r = 1$ are smaller with respect to the case $r = 0$ since the adapted grids yielded by $\widetilde{\eta}_{\text{A}}^{goal,1}$ are in general coarser, due to the underestimation. Moreover the choice $r = 1$ emphasizes the anisotropic features of the mesh (the maximum value for the stretching factor in Table 2 is about 18).

5.2 Third test case

We now consider a more complex functional $J(\varphi)$ aiming at controlling a localized quantity. In particular the Riesz representant is chosen as $g(x_1, x_2) = x_1\, x_2\, (x_1-1)\, (x_2-1)\, \big(\pi/2+\arctan(80\,(x_1-0.5))\big)$. The arctan function plays the role of a regularized characteristic function associated with the right-half of the

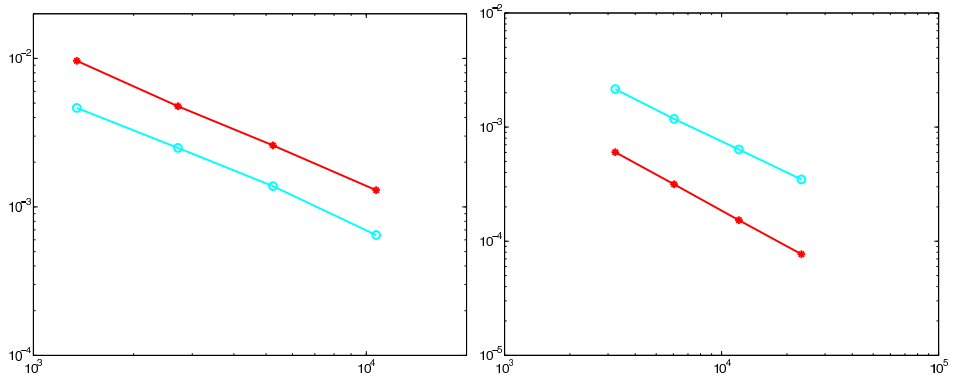


Figure 8: Second test case: error ('$\circ$'-marker) and estimator ('$*$'-marker) versus number of elements for $r = 0$ (left) and $r = 1$ (right)

Table 3: Third test case: convergence history for $r = 0$

$\tau_{\rm rel}$	$\mathrm{card}(\mathcal{T}_h)$	$\max s_{1,K}$	$\lvert J(u-u_h)\rvert/\lvert J(u)\rvert$	$\widetilde{\eta}_{\rm A}^{goal,0}/\lvert J(u_h)\rvert$	$\mathrm{E.I.}_{\rm A}^{goal,0}$
10^{-2}	2268	5.54	$2.507 \cdot 10^{-3}$	$9.714 \cdot 10^{-3}$	3.88
$10^{-2}/2$	4575	5.29	$1.231 \cdot 10^{-3}$	$5.233 \cdot 10^{-3}$	4.25
$10^{-2}/4$	9040	6.60	$6.046 \cdot 10^{-4}$	$2.602 \cdot 10^{-3}$	4.30
$10^{-2}/8$	17705	6.34	$3.128 \cdot 10^{-4}$	$1.362 \cdot 10^{-3}$	4.35

domain. The polynomial bubble $x_1\,x_2\,(x_1-1)\,(x_2-1)$ enforces the homogeneous Dirichlet boundary condition. The primal solution and the domain are the same as in the previous test case, while the exact value of the functional is now $J(u) = 0.51602455$.
The adaptive procedure is run, again picking different values for $\tau_{\rm rel}$ and $r = 0, 1$. Check (19) is adopted also in this case. For all the choices of the parameters

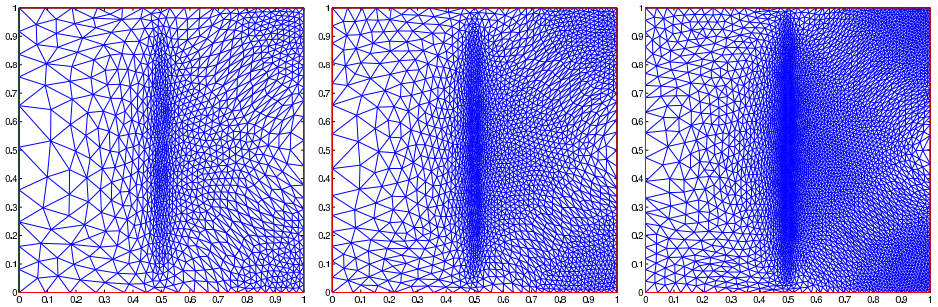


Figure 9: Third test case: final adapted meshes for $r = 0$, $\tau_{\rm rel} = 10^{-2}, 10^{-2}/2, 10^{-2}/4$ (from left to right)

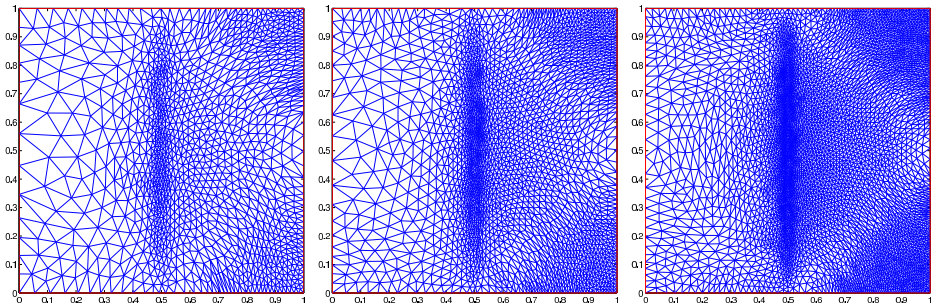


Figure 10: Third test case: final adapted meshes for $r = 1$, $\tau_{\rm rel} = 10^{-2}/4, 10^{-2}/8, 10^{-2}/16$ (from left to right)

the adaptive procedure converges within $6-7$ iterations. The final adapted

Table 4: Third test case: convergence history for $r = 1$

τ_{rel}	$\text{card}(\mathcal{T}_h)$	$\max s_{1,K}$	$\lvert J(u-u_h)\rvert/\lvert J(u)\rvert$	$\widetilde{\eta}_{\text{A}}^{goal,1}/\lvert J(u_h)\rvert$	$\text{E.I.}_{\text{A}}^{goal,1}$
$10^{-2}/4$	1264	7.30	$3.958 \cdot 10^{-3}$	$2.318 \cdot 10^{-3}$	0.586
$10^{-2}/8$	2388	9.42	$2.273 \cdot 10^{-3}$	$1.209 \cdot 10^{-3}$	0.533
$10^{-2}/16$	4701	7.54	$1.154 \cdot 10^{-3}$	$1.154 \cdot 10^{-3}$	0.534
$10^{-2}/32$	9287	9.78	$6.734 \cdot 10^{-4}$	$3.140 \cdot 10^{-4}$	0.467

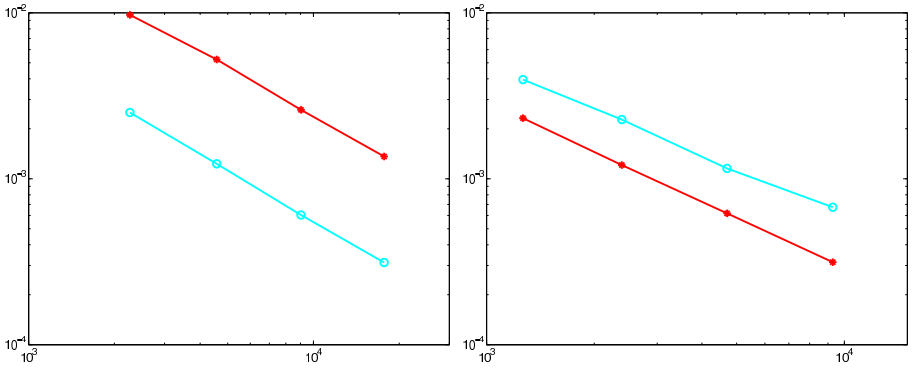


Figure 11: Third test case: error ('$\circ$'-marker) and estimator ('$*$'-marker) versus number of elements for $r = 0$ (left) and $r = 1$ (right)

grids corresponding to $\tau_{\text{rel}} = 10^{-2}, 10^{-2}/2, 10^{-2}/4$ for $r = 0$ are collected in Figure 9, while the ones associated with $\tau_{\text{rel}} = 10^{-2}/4, 10^{-2}/8, 10^{-2}/16$ and $r = 1$ are gathered in Figure 10. In all cases, the action of g is stressed by the presence of the vertical layer in the middle of the domain as well as by the coarse mesh on the left-hand side of the domain.
Tables 3 and 4 summarize the main quantities related to the adaptive procedure. As in the previous test case we observe the slightly overestimation of the estimator associated with $r = 0$, whereas for $r = 1$ we have a moderate underestimation. However, in both cases, the rate of convergence is quasi-optimal (see Figure 11). The maximum stretching factor reaches about the value 7 for $r = 0$ and 10 when $r = 1$.

6 Conclusions

Despite its foundation on heuristic considerations, the proposed family of gradient recovery type a posteriori error estimators demonstrate very promising numerical results, as shown explicitly by the 3D numerical results in Sec. 4.2. The generalization of the approach proposed in [32, 17] to the goal-oriented framework also seems to be encouraging, moving from the quasi-optimal error-vs-number of elements behaviour exhibited in Sec. 5.1 and 5.2. Future work will

overcome the limitation of considering only the Poisson problem: in particular the aim will be to preserve the optimal convergence rate of the estimator, even in the presence of an asymmetric problem with possible stabilization terms.

References

[1] M.T. Ainsworth and J.T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley, New York, 2000.

[2] R.C. Almeida, R.A. Feijóo, A.C. Galeão, C. Padra, and R.S. Silva. Adaptive finite element computational fluid dynamics using an anisotropic error estimator. *Comput. Methods Appl. Mech. Engrg.*, 182(3-4):379–400, 2000.

[3] T. Apel. *Anisotropic Finite Elements: Local Estimates and Applications*. Advances in Numerical Mathematics. Teubner: Stuttgart, 1999.

[4] B.F. Armaly, F. Durst, J.C.F. Pereira, and B. Schönung. Experimental and theoretical investigation of backward-facing step flow. *J. Fluid Mech.*, 127:473–496, 1983.

[5] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001.

[6] C.L. Bottasso, G. Maisano, S. Micheletti, and S. Perotto. On some new recovery based a posteriori error estimators. *Comput. Methods Appl. Mech. Engrg.*, 195:4794–4815, 2006.

[7] Y. Bourgault, M. Picasso, F. Alauzet, and A. Loseille. On the use of anisotropic a posteriori error estimators for the adaptative solution of 3D inviscid compressible flows. *Int. J. Numer. Meth. Fluids*, 59(1):47–74, 2009.

[8] R. Boussetta, T. Coupez, and L. Fourment. Adaptive remeshing based on a posteriori error estimation for forging simulation. *Comput. Methods Appl. Mech. Engrg.*, 195(48-49):6626–6645, 2006.

[9] J.C. Bruch. Free surface seepage problems solved using: (a) the Zienkiewicz-Zhu error estimation procedure; and (b) a parallel computer. *Pitman Res. Notes Math. Ser.*, 282:98–104, 1993.

[10] A.S. Candy. *Subgrid scale modelling of transport processes*. PhD thesis, Imperial College London, 2008.

[11] C. Carstensen. All first-order averaging techniques for a posteriori finite element error control on unstructured grids are efficient and reliable. *Math. Comp.*, 73:1153–1165, 2004.

[12] M.J. Castro-Diaz, F. Hecht, B. Mohammadi, and O. Pironneau. Anisotropic unstructured mesh adaptation for flow simulations. *Int. J. Numer. Meth. Fluids*, 25(4):475–491, 1997.

[13] Ph. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, 2:77–84, 1975.

[14] E.F. D'Azevedo. Optimal triangular mesh generation by coordinate transformation. *SIAM J. Sci. Statist. Comput.*, 12(4):755–786, 1991.

[15] J. Dompierre, M.-G. Vallet, Y. Bourgault, M. Fortin, and W.G. Habashi. Anisotropic mesh adaptation: towards user-independent, mesh-independent and solver-independent CFD. III. Unstructured meshes. *Int. J. Numer. Meth. Fluids*, 39(8):675–702, 2002.

[16] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. Introduction to adaptive methods for differential equations. *Acta Numerica*, 4:105–158, 1995.

[17] P.E. Farrell, S. Micheletti, and S. Perotto. An anisotropic Zienkiewicz-Zhu a posteriori error estimator for 3D applications. Technical Report 25/2009, MOX, Dipartimento di Matematica "F. Brioschi", Politecnico di Milano, 2009.

[18] L. Formaggia, S. Micheletti, and S. Perotto. Anisotropic mesh adaptation in computational fluid dynamics: application to the advection–diffusion–reaction and the Stokes problems. *Appl. Numer. Math.*, 51(4):511–533, 2004.

[19] L. Formaggia and S. Perotto. New anisotropic a priori error estimates. *Numer. Math.*, 89(4):641–667, 2001.

[20] L. Formaggia and S. Perotto. Anisotropic error estimates for elliptic problems. *Numer. Math.*, 94(1):67–92, 2003.

[21] P.J. Frey and F. Alauzet. Anisotropic mesh adaptation for CFD computations. *Comput. Methods Appl. Mech. Engrg.*, 194:5068–5082, 2005.

[22] P.L. George and H. Borouchaki. *Delaunay Triangulation and Meshing: Application to Finite Elements*. Hermes: Paris, 1998.

[23] M.B. Giles and E. Suli. Adjoint methods for pdes: a posteriori error analysis and postprocessing by duality. *Acta Numerica*, pages 145–236, 2002.

[24] C. Gruau and T. Coupez. 3D tetrahedral, unstructured and anisotropic mesh generation with adaptation to natural and multidomain metric. *Comput. Methods Appl. Mech. Engrg.*, 194(48-49):4951–4976, 2005.

[25] F. Hecht. *Freefem++ Manual, Third Edition, Version 3.5*, 2009, `http://www.freefem.org/ff++/index.htm`.

[26] M. Krízek and P. Neittaanmäki. Superconvergence phenomenon in the finite element method arising from averaging gradients. *Numer. Math*, 45:105–116, 1984.

[27] G. Kunert. An a posteriori residual error estimator for the finite element method on anisotropic tetrahedral meshes. *Numer. Math.*, 86(3):471–490, 2000.

[28] K.L. Lawrence and R.V. Nambiar. The Zienkiewicz-Zhu error estimator for multiple material problems. *Commun. Appl. Numer. Methods*, 8:273–277, 1992.

[29] H. Le, P. Moin, and J. Kim. Direct numerical simulation of turbulent flow over a backward-facing step. *J. Fluid Mech.*, 330:349–374, 1997.

[30] J.L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications, vol. I.* Springer-Verlag, Berlin, 1972.

[31] S. Micheletti and S. Perotto. Output functional control for nonlinear equations driven by anisotropic mesh adaption: the Navier-Stokes equations. *SIAM J. Sci. Comput.*, 30(6):2817–2854, 2008.

[32] S. Micheletti and S. Perotto. Anisotropic adaptation via a Zienkiewicz-Zhu error estimator for 2D elliptic problems. *Proceedings of the eighth European Conference on Numerical Mathematics and Advanced Applications, ENUMATH 2009, June 29- July 3, Uppsala, Sweden*, 2009.

[33] A. Naga and Z. Zhang. A posteriori error estimates based on the polynomial preserving recovery. *SIAM J. Numer. Anal.*, 42:1780–1800, 2004.

[34] C.C. Pain, A.P. Umpleby, C.R.E. de Oliveira, and A.J.H. Goddard. Tetrahedral mesh optimisation and adaptivity for steady-state and transient finite element calculations. *Comput. Methods Appl. Mech. Engrg.*, 190(29–30):3771–3796, 2001.

[35] T.P. Pawlak, M.J. Wheeler, and S.M. Yunus. Application of the Zienkiewicz-Zhu error estimator for plate and shell analysis. *Int. J. Numer. Meth. Engng*, 29:1281–1298, 1990.

[36] J. Peraire, M. Vahdati, K. Morgan, and O.C Zienkiewicz. Adaptive remeshing for compressible flow computations. *J. Comput. Phys.*, 72(2):449–466, 1987.

[37] M.D. Piggott, C.C. Pain, G.J. Gorman, P.W. Power, and A.J.H. Goddard. h, r, and hr adaptivity with applications in numerical ocean modelling. *Ocean Model.*, 10(1-2):95–113, 2005.

[38] R. Rodríguez. Some remarks on Zienkiewicz-Zhu estimator. *Numer. Methods Partial Differential Equations*, 10:625–635, 1994.

[39] L.R. Scott and S. Zhang. Finite element interpolation of non-smooth functions satisfying boundary conditions. *Math. Comp.*, 54:483–493, 1990.

[40] R.B. Simpson. Anisotropic mesh transformations and optimal error control. *Appl. Numer. Math.*, 14:183–198, 1994.

[41] D.A. Venditti and D.L. Darmofal. Anisotropic grid adaptation for functional outputs: application to two-dimensional viscous flows. *J. Comput. Phys.*, 187:22–46, 2003.

[42] N. Yan and A. Zhou. Gradient recovery type a posteriori error estimation for finite element approximations on irregular meshes. *Comput. Methods Appl. Mech. Engrg.*, 190:4289–4299, 2001.

[43] O.C. Zienkiewicz and J.Z. Zhu. A simple error estimator and adaptive procedure for practical engineering analysis. *Int. J. Numer. Meth. Engng*, 24:337–357, 1987.

[44] O.C. Zienkiewicz and J.Z. Zhu. The superconvergent patch recovery and a posteriori error estimates. I: The recovery technique. *Int. J. Numer. Meth. Engng*, 33:1331–1364, 1992.

[45] O.C. Zienkiewicz and J.Z. Zhu. The superconvergent patch recovery and a posteriori error estimates. II: Error estimates and adaptivity. *Int. J. Numer. Meth. Engng*, 33:1365–1382, 1992.

[46] O.C. Zienkiewicz and J.Z. Zhu. The superconvergent patch recovery (SPR) and adaptive finite element refinement. *Comput. Methods Appl. Mech. Engrg.*, 101:207–224, 1992.

Tipo de evento:	Congreso
Nombre:	Congreso SIMAI–SEMA 2010
Lugar:	Università di Cagliari, ITALIA
Fecha:	6–10 septiembre 2010
Organiza:	Societá Italiana di Matematica Applicata e Industriale (SIMAI) y Sociedad Española de Matemática Aplicada (SeMA)
Información:	
E-mail:	Giorgio.Fotia@crs4.it
WWW:	http://openconference.simai.eu/index.php/sc/2010

Tipo de evento:	Workshop
Nombre:	Stochastic Partial Differential Equations: Approximation, Asymptotics and Computation
Lugar:	Isaac Newton Institute for Mathematical Sciences, Cambridge, UK
Fecha:	June 28 to July 2, 2010
Organiza:	Professor A. Debussche (ENS de Cachan) and Professor M. Hairer (Warwick) in association with the Newton Institute programme "Stochastic Partial Differential Equations"
Información:	
E-mail:	programmes@newton.ac.uk
WWW:	http://www.newton.ac.uk/programmes/SPD/spdw04.html

Tipo de evento:	Conference
Nombre:	The eleventh international conference on Integral Methods in Science and Engineering
Lugar:	University of Brighton, UK
Fecha:	12–15 July 2010
Organiza:	
Información:	
E-mail:	imse2010@brighton.ac.uk
WWW:	http://www.cmis.brighton.ac.uk/imse2010/

Tipo de evento:	Curso
Nombre:	Control of Partial Differential Equations
Lugar:	Cetraro (Cosenza), ITALIA
Fecha:	19–23 July 2010
Organiza:	FONDAZIONE CIME - Roberto CONTI. INTERNATIONAL MATHEMATICAL SUMMER CENTER 2010 COURSES
Información:	
E-mail:	cannarsa@mat.uniroma2.it
WWW:	http://php.math.unifi.it/users/cime/, http://www.ceremade.dauphine.fr/ glass/CIME/

Tipo de evento:	Curso
Nombre:	Quantum Many Body Systems
Lugar:	Cetraro (Cosenza), ITALIA
Fecha:	August 30 – September 4, 2010
Organiza:	FONDAZIONE CIME - Roberto CONTI. INTERNATIONAL MATHEMATICAL SUMMER CENTER 2010 COURSES
Información:	
E-mail:	giuliani@mat.uniroma3.it, mastropi@mat.uniroma2.it
WWW:	http://php.math.unifi.it/users/cime/, http://www.mat.uniroma3.it/users/giuliani/public_html/cime/index.html

Tipo de evento:	Conference
Nombre:	I ENJIM Encuentro Nacional de Jóvenes Investigadores en Matemáticas / I Spanish Young Researchers Meeting in Mathematics
Lugar:	Universidad de Sevilla
Fecha:	01–03 September 2010
Organiza:	María Anguiano Moreno, Manuel Ceballos González, Aurora Fernández León, Carlos Hugo Jiménez Gómez, Luis Felipe Rivero Garvía, Francisco Javier Suárez Gra
Información:	
E-mail:	enjim@us.es
WWW:	http://congreso.us.es/enjim/

Tipo de evento:	Escuela
Nombre:	XIV Escuela Hispano–Francesa sobre Simulación Numérica en Física e Ingeniería
Lugar:	Escuela Técnica Superior de Náutica y Máquinas, Campus de Riazor, 15011 A Coruña
Fecha:	6–10 septiembre 2010
Organiza:	María J. Esteban, CNRS y Université Paris–Dauphine; Carlos Vázquez Cendón, Universidad de A Coruña
Información:	
E-mail:	ehf2010@udc.es
WWW:	http://dm.udc.es/ehf2010/

Tipo de evento:	Workshop
Nombre:	Fluid–Kinetic Modelling in Biology, Physics and Engineering
Lugar:	Isaac Newton Institute for Mathematical Sciences, Cambridge, UK
Fecha:	6–10 September 2010
Organiza:	Professor J. A. Carrillo (Barcelona), Professor S. Jin (Wisconsin), Professor A. Juengel (Vienna) and Professor P. A. Markowich (Cambridge) in association with the Newton Institute programme Partial Differential Equations in Kinetic Theories
Información:	
E-mail:	programmes@newton.ac.uk
WWW:	http://www.newton.ac.uk/programmes/KIT/kitw01.html

Tipo de evento:	Curso
Nombre:	Topics in mathematical fluid-mechanics
Lugar:	Cetraro (Cosenza), ITALIA
Fecha:	6–11 September 2010
Organiza:	FONDAZIONE CIME - Roberto CONTI. INTERNATIONAL MATHEMATICAL SUMMER CENTER 2010 COURSES
Información:	
E-mail:	bveiga@dma.unipi.it, flandoli@dma.unipi.it
WWW:	http://php.math.unifi.it/users/cime/, http://php.math.unifi.it/users/cime/Courses/2010/course.php?codice=20104

Tipo de evento:	Conference
Nombre:	Highly Oscillatory Problems: From Theory to Applications
Lugar:	The Isaac Newton Institute, Cambridge, UK
Fecha:	12–17 September 2010
Organiza:	The European Science Foundation (ESF), in partnership with EMS and ERCOM/INI
Información:	
E-mail:	apiccolotto@esf.org
WWW:	http://www.esf.org/index.php?id=6532

Tipo de evento:	Conference
Nombre:	The Third International Congress on Mathematical Software
Lugar:	Kobe, JAPAN
Fecha:	13–17 September 2010
Organiza:	V. Dougalis, E. Gallopoulos, A. Hadjidimos, I. S. Kotsireas, D. Noutsos, Y. G. Saridakis, M. N. Vrahatis
Información:	
E-mail:	komei.fukuda@ifor.math.ethz.ch, noro@math.kobe-u.ac.jp
WWW:	http://www.math.kobe-u.ac.jp/icms2010/

Tipo de evento:	Conference
Nombre:	Conference in Numerical Analysis (NumAn 2010). Recent Approaches to Numerical Analysis: Theory, Methods and Applications
Lugar:	Chania, GREECE
Fecha:	15–18 September 2010
Organiza:	V. Dougalis, E. Gallopoulos, A. Hadjidimos, I. S. Kotsireas, D. Noutsos, Y. G. Saridakis, M. N. Vrahatis
Información:	
E-mail:	numan2010@science.tuc.gr
WWW:	http://numan2010.science.tuc.gr/

Tipo de evento:	Workshop
Nombre:	Numerical Methods for Continuous Optimization
Lugar:	Institute for Pure and Applied Mathematics (IPAM), UCLA, Los Angeles, California, USA
Fecha:	11–15 October 2010
Organiza:	Steven Wright, Chair (University of Wisconsin–Madison, Computer Science), Don Goldfarb (Columbia University, IEOR), Renato Monteiro (Georgia Institute of Technology, School of Industrial and Systems Engineering), Yurii Nesterov (Université Catholique de Louvain), Michael Overton (New York University), Kim Toh (National University of Singapore)
Información:	
E-mail:	opws2@ipam.ucla.edu
WWW:	http://www.ipam.ucla.edu/programs/opws2/

Tipo de evento:	Workshop
Nombre:	Numerical Solutions of Partial Differential Equations: Novel Discretization Techniques
Lugar:	University of Minnesota, Minneapolis, USA
Fecha:	1–5 November 2010
Organiza:	Susanne C. Brenner (Mathematics, Louisiana State University), Claudio Canuto (Matematica, Politecnico di Torino), Chi–Wang Shu (Applied Mathematics, Brown University)
Información:	
E-mail:	brenner@math.lsu.edu, ccanuto@polito.it, shu@dam.brown.edu
WWW:	http://www.ima.umn.edu/2010-2011/W11.1-5.10/

Tipo de evento:	Workshop
Nombre:	Applications of Optimization in Science and Engineering
Lugar:	Institute for Pure and Applied Mathematics (IPAM), UCLA, Los Angeles, California, USA
Fecha:	11–15 October 2010
Organiza:	Stephen Boyd (Stanford University, Engineering), Yonina Eldar (Technion - Israel Institute of Technology, Electrical Engineering), Tom Luo (University of Minnesota, Twin Cities), Bernhard Scholkopf (Max-Planck-Institute for Biological Cybernetics), Lieven Vandenberghe (University of California, Los Angeles (UCLA), EE)
Información:	
E-mail:	`opws5@ipam.ucla.edu`
WWW:	`http://www.ipam.ucla.edu/programs/opws5/`

Armesto Álvarez, José Antonio
Investigador. *Líneas de investigación:* Dinámica de fluidos (CFD) – Univ. de Cantabria – E. T. S. I. de Caminos, C. y Puertos – Instituto de Hidráulica Ambiental "IH Cantabria" – Avda. de Los Castros, s/n. 39005 Santander.
Tlf.: 942.201.810. *Fax:* 942.201.860.
e-mail: joseantonio.armesto@unican.es.

Becerro Molina, David
Estudiante. *Líneas de investigación:* Sistemas dinámicos complejos – Univ. de Barcelona – Fac. de Matemáticas – Dpto. de Matemática Aplicada y Análisis – Gran Vía de les Corts Catalanes, 585. 08007 Barcelona.
Tlf.: 934.021.597. *Fax:* 934.041.601.
e-mail: daxfiles@hotmail.com.

Bermúdez Edo, María Concepción
Prof. Titular de Escuela Universitaria. *Líneas de investigación:* Análisis numérico – Univ. Politécnica de Cartagena – E. T. S. de Ingeniería Naval y Oceánica – Dpto. de Matemática Aplicada y Estadística – Alfonso XIII, 52. 30203 Cartagena.
Tlf.: 968.325.583.
e-mail: concepcion.bermudez@upct.es.

Bohórquez Rodríguez de Medina, Patricio
Investigador. *Líneas de investigación:* Mecánica de fluidos computacional – Univ. de Jaen – Escuela Politécnica Superior – Dpto. de Ingeniería Mecánica y Minera – Campus Las Lagunillas. 23071 Jaén.
Tlf.: 953.212.872. *Fax:* 953.212.870.
e-mail: prmedina@ujaen.es.
http://blogs.ujaen.es/prmedina

Canadell Cano, Marta
Estudiante. – Univ. de Barcelona – Fac. de Matemáticas – Gran Vía, 585. 08007 Barcelona.

e-mail: marta.canadell.cano@gmail.com.

Carpio Huertas, Jaime
Prof. Titular de Universidad Interino. *Líneas de investigación:* Modelización y simulación numérica: MEF, técnicas adaptativas en espacio y tiempo, método DWR, esquemas semilagrangianos, convección-difusión-reacción, Navier-Stokes, combustión. Series temporales e identificación de parámetros en dinámica de estructuras – Univ. Politécnica de Madrid – E. T. S. I. Industriales – Dpto. de Ing. de Organización, Adm. Empresas y Estadística – C/ José Gutiérrez Abascal. 28006 Madrid.
Tlf.: 913.363.149.
e-mail: jaime.carpio@upm.es.
http://www.etsii.upm.es/ingor/estadistica/Jaime/jaime_carpio.htm

Crespo Cutillas, Francisco
Estudiante. *Líneas de investigación:* Equilibrios relativos en sistemas hamiltonianos, métodos aplicados a la mecánica celeste – Univ. de Murcia – Fac. de Informática – Dpto. de Matemática Aplicada – Campus de Espinardo. Murcia.

e-mail: francisco.crespo@um.es.

Díaz Cardell, Sara
Estudiante (Becario). *Líneas de investigación:* Códigos LDPC, complejidad de funciones booleanas, stream ciphers – Univ. de Alicante – Fac. de Ciencias – Dpto. de Estadística e Investigación Operativa – Campus de S. Vicente del Raspeig, Aptdo. Correos 99. 03080 Alicante.
Tlf.: 965.903.400. *Fax:* 965.903.902.
e-mail: s.diaz@ua.es.
http://www.eio.ua.es

Fernández Hernández, Marta
Estudiante. – Univ. Politécnica de Madrid – E. T. S. I. de Minas – Depto. de Matemática Aplicada y Métodos Informáticos – Alenza, 4. 28003 Madrid.

e-mail: marta@dmami.upm.es.

García García, Francisco
Prof. Asociado. *Líneas de investigación:* Funciones Bent – Univ. de Alicante – Fac. de Económicas – Dpto. de Fundamentos del Análisis Económico – Crtra. San Vicente del Raspeig, s/n. 03690 San Vicente del Raspeig (Alicante).
Tlf.: 965.903.400. *Fax:* 965.903.464.
e-mail: francisco.garcia@ua.es.

Gasso Matoses, María Teresa
Prof. Titular de Universidad. *Líneas de investigación:* Algebra matricial, topología – Univ. Politécnica de Valencia – Fac. de Informática – Dpto. de Matemática Aplicada – Camino de Vera, s/n. 46022 Valencia.
Tlf.: 963.879.496.
e-mail: mgasso@mat.upv.es.

Hidalgo López, Arturo
Prof. Titular de Universidad. *Líneas de investigación:* Simulación numérica, volúmenes finitos, problemas hiperbólicos, modelos climáticos, modelos geológicos – Univ. Politécnica de Madrid – E. T. S. I. de Minas – Dpto. de Matemática Aplicada y Métodos Informáticos – Ríos Rosas, 21. 28003 Madrid.
Tlf.: 913.363.233. *Fax:* 913.367.051.
e-mail: arturo.hidalgo@upm.es.

Malaver de la Fuente, Manuel
Profesor Asistente. *Líneas de investigación:* Ecuaciones diferenciales – Univ. Marítima del Caribe – Dirección de Gestión de Docentes – Dep. de Ciencias Aplicadas – Avda. del Ejército. 1162 Catia La Mar, Est. de Vargas (Venezuela).
Tlf.: 04.168.163.120 / 3.501.066.
e-mail: agujero_1@hotmail.com, mmf_umc@hotmail.com.

Martínez González, Alicia
Investigadora. *Líneas de investigación:* Modelos matemáticos y métodos multiescala para la señalización celular en tumores y terapias oncológicas – Univ. de Castilla-La Mancha – E. T. S. de Ingeniería Industrial – Dpto. de Matemáticas – Avda. Camilo José Cela, 3. 13071 Ciudad Real.

e-mail: alicia.martinez@uclm.es.
http://imaci.uclm.es

Pou Bueno, Marta
Estudiante. *Líneas de investigación:* Matemáticas financieras – Univ. de La Coruña – Fac. de Informática – Depto. de Matemáticas – Campus de Elviña, s/n. 15071 - A Coruña.
Tlf.: 981.167.000, Ext. 1301. *Fax:* 981.167.160.
e-mail: mpou@udc.es.
http://dm.udc.es

Rodríguez Rodríguez, Marcos
Estudiante (Becario). *Líneas de investigación:* Métodos numéricos para ODE's – Univ. de Zaragoza – Fac. de Ciencias – Dpto. de Matemática Aplicada – C/ Pedro Cerbuna, 1. 50009 Zaragoza.
Tlf.: 976.761.000. *Fax:* 976.761.140.
e-mail: marcos@unizar.es.
http://gme.unizar.es

Smith, Nadia A. S.
Estudiante (Becario). *Líneas de investigación:* Modelos matemáticos en ingeniería de alimentos – Univ. Complutense de Madrid – Fac. de Matemáticas – Dpto. de Matemática Aplicada – Plaza de las Ciencias, 3. 28040 Madrid.
Tlf.: 913.944.462. *Fax:* 913.944.613.
e-mail: nas.smith@mat.ucm.es.

Trillo Moya, Juan Carlos
Ayudante Doctor. *Líneas de investigación:* Wavelets, multirresolución no lineal y aplicaciones, interpolación, tratamiento digital de imágenes – Univ. Politécnica de Cartagena – E. T. S. de Ingeniería Industrial – Dpto. de Matemática Aplicada y Estadística – Edif. EUIT Civil y Naval. C/ Alfonso XIII. 30203 Cartagena.
Tlf.: 968.325.584. *Fax:* 968.325.694.
e-mail: jctrillo@upct.es.
http://www.dmae.upct.es/jcarlos

Direcciones útiles

Consejo Ejecutivo de SëMA

Presidente:

Carlos Vázquez Cendón. (carlosv@udc.es).
Dpto. de Matemáticas. Facultad de Informática. Univ. de A Coruña. Campus de Elviña, s/n. 15071 A Coruña. *Tel:* 981 16 7000-1335.

Vicepresidente:

Rosa María Donat Beneito. (Rosa.M.Donat@uv.es)
Dpto. de Matemática Aplicada. Fac. de Matemàtiques. Univ. de Valencia. Dr. Moliner, 50. 46100 Burjassot (Valencia) *Tel:* 963 544 727.

Secretario:

Carlos Castro Barbero. (ccastro@caminos.upm.es).
Dpto. de Matemática e Informática. E.T.S.I. Caminos, Canales y Puertos. Univ. Politécnica de Madrid. Av. Aranguren s/n. 28040 Madrid. *Tel:* 91 336 6664.

Vocales:

Sergio Amat Plata. (sergio.amat@upct.es)
Dpto. de Matemática Aplicada y Estadística. Univ. Politécnica de Cartagena. Paseo de Alfonso XIII, 52. 30203 Cartagena (Murcia). *Tel:* 968 325 694.

Rafael Bru García. (rbru@mat.upv.es)
Dpto. de Matemática Aplicada. E.T.S.I. Agrónomos. Univ. Politécnica de Valencia. Camí de Vera, s/n. 46022 Valencia. *Tel:* 963 879 669.

José Antonio Carrillo de la Plata. (carrillo@mat.uab.es)
Dpto. de Matemáticas. Univ. Autónoma de Barcelona. Edifici C. 08193 Bellaterra (Barcelona). *Tel:* 935 812 413.

Inmaculada Higueras Sanz. (higueras@unavarra.es).
Dpto de Matemática e Informática Univ. Pública de Navarra. Campus de Arrosadía, s/n. *Tel:* 948 169 526. 31006 Pamplona.

Carlos Parés Madroñal. (carlos_pares@uma.es).
Dpto. de Análisis Matemático. Fac. de Ciencias. Univ. de Málaga. Campus de Teatinos, s/n. 29080 Málaga. *Tel:* 952 132 017.

Pablo Pedregal Tercero. (Pablo.Pedregal@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. de Camilo José Cela, s/n. 13071 Ciudad Real. *Tel:* 926 295 436

Luis Vega González. (luis.vega@ehu.es).
Dpto. de Matemáticas. Fac. de Ciencias. Univ. del País Vasco. Aptdo. 644. 48080 Bilbao (Vizcaya). *Tel:* 944 647 700.

Tesorero:

Íñigo Arregui Álvarez. (arregui@udc.es).
Dpto. de Matemáticas. Fac. de Informática. Univ. de A Coruña. Campus de Elviña, s/n. 15071 A Coruña. *Tel:* 981 16 7000-1327.

Comité Científico del Boletín de SëMA

Daniel B. Szyld. (szyld@temple.dot.edu).
Department of Mathematics. College of Science and Technology. Temple University (038-16). 1805 N Broad Street. Philadelphia, PA 19122-6094, USA. *Tel:* +1 215 204 7288.

Luis Vega González. (mtpvegol@lg.ehu.es).
Dpto. de Matemáticas. Fac. de Ciencias. Univ. del País Vasco. Aptdo. 644. 48080 Bilbao (Vizcaya). *Tel:* 944 647 700.

Chi-Wang Shu. (shu@dam.brown.edu).
Division of Applied Mathematics Box F. 182 George Street Brown University Providence RI 02912 *Tel:* (401) 863-2549.

Enrique Zuazua Iriondo. (zuazua@bcamath.org).
Basque Center for Applied Mathematics Bizkaia Technology Park Building 208B 48170 - Zamudio (Vizcaya) *Tel:* 944 014 690.

Grupo Editor del Boletín de SëMA

Pablo Pedregal Tercero. (Pablo.Pedregal@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. Camilo José Cela, s/n. 13071 Ciudad Real. *Tel:* 926 295 300 ext. 3809

Enrique Fernández Cara. (cara@us.es).
Dpto. de Ecuaciones Diferenciales y An. Numérico. Fac. de Matemáticas. Univ. de Sevilla. Tarfia, s/n. 41012 Sevilla. *Tel:* 954 557 992.

Ernesto Aranda Ortega. (Ernesto.Aranda@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. Camilo José Cela, s/n. 13071 Ciudad Real. *Tel:* 926 295 300 ext. 3813

José Carlos Bellido Guerrero. (JoseCarlos.Bellido@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. Camilo José Cela, s/n. 13071 Ciudad Real. *Tel:* 926 295 300 ext. 3859

Alberto Donoso Bellón. (Alberto.Donoso@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. Camilo José Cela, s/n. 13071 Ciudad Real. *Tel:* 926 295 300 ext. 3859

Responsables de secciones del Boletín de SëMA

Artículos:

Enrique Fernández Cara. (cara@us.es).
Dpto. de Ecuaciones Diferenciales y An. Numérico. Fac. de Matemáticas. Univ. de Sevilla. Tarfia, s/n. 41012 Sevilla. *Tel:* 954 557 992.

Matemáticas e Industria:

Mikel Lezaun Iturralde. (mepleitm@lg.ehu.es).
Dpto. de Matemática Aplicada, Estadística e I. O. Fac. de Ciencias. Univ. del País Vasco. Aptdo. 644. 48080 Bilbao (Vizcaya). *Tel:* 944 647 700.

Educación Matemática:

Roberto Rodríguez del Río. (rr_delrio@mat.ucm.es).
Dpto. de Matemática Aplicada. Fac. de Químicas. Univ. Compl. de Madrid. Ciudad Universitaria. 28040 Madrid. *Tel:* 913 944 102.

Resúmenes de libros:

Fco. Javier Sayas González. (jsayas@posta.unizar.es).
Dpto. de Matemática Aplicada. Centro Politécnico Superior . Universidad de Zaragoza. C/María de Luna, 3. 50015 Zaragoza. *Tel:* 976 762 148.

Noticias de SëMA:

Carlos Castro Barbero. (ccastro@caminos.upm.es).
Dpto. de Matemática e Informática. E.T.S.I. Caminos, Canales y Puertos. Univ. Politécnica de Madrid. Av. Aranguren s/n. 28040 Madrid. *Tel:* 91 336 6664.

Anuncios:

Óscar López Pouso. (oscar.lopez@usc.es).
Dpto. de Matemática Aplicada. Fac. de Matemáticas. Univ. de Santiago de Compostela. Campus sur, s/n. 15782 Santiago de Compostela *Tel:* 981 563 100, ext. 13228.

Responsables de otras secciones de SëMA

Gestión de Socios:

Íñigo Arregui Álvarez. (arregui@udc.es).
Dpto. de Matemáticas. Fac. de Informática. Univ. de A Coruña. Campus de Elviña, s/n. 15071 A Coruña. *Tel:* 981 16 7000-1327.

Página web: www.sema.org.es/**:**

Carlos Castro Barbero. (ccastro@caminos.upm.es).
Dpto. de Matemática e Informática. E.T.S.I. Caminos, Canales y Puertos. Univ. Politécnica de Madrid. Av. Aranguren s/n. 28040 Madrid. *Tel:* 91 336 6664.

1. Los artículos publicados en este Boletín podrán ser escritos en español o inglés y deberán ser enviados por correo certificado a

   Prof. E. FERNÁNDEZ CARA
   Presidente del Comité Científico, Boletín S$\tilde{e}$MA
   Dpto. E.D.A.N., Facultad de Matemáticas
   Aptdo. 1160, 41080 SEVILLA

   También podrán ser enviados por correo electrónico a la dirección

   `boletin.sema@uclm.es`

   En ambos casos, el/los autor/es deberán enviar por correo certificado una carta a la dirección precedente mencionando explícitamente que el artículo es sometido a publicación e indicando el nombre y dirección del autor corresponsal. En esta carta, podrán sugerirse nombres de miembros del Comité Científico que, a juicio de los autores, sean especialmente adecuados para juzgar el trabajo.

   La decisión final sobre aceptación del trabajo será precedida de un procedimiento de revisión anónima.

2. Las contribuciones serán preferiblemente de una longitud inferior a 24 páginas y se deberán ajustar al formato indicado en los ficheros a tal efecto disponibles en la página web de la Sociedad (`http://www.sema.org.es/`).

3. El contenido de los artículos publicados corresponderá a un área de trabajo preferiblemente conectada a los objetivos propios de la Matemática Aplicada. En los trabajos podrá incluirse información sobre resultados conocidos y/o previamente publicados. Se anima especialmente a los autores a presentar sus propios resultados (y en su caso los de otros investigadores) con estilo y objetivos divulgativos.

Ficha de Inscripción Individual

Sociedad Española de Matemática Aplicada SëMA

Remitir a: Iñigo Arregui, Dpto de Matemáticas, Fac. de Informática,
Universidad de A Coruña. Campus de Elviña, s/n. 15071 A Coruña.
CIF: G-80581911

Datos Personales

- Apellidos: .....
- Nombre: .....
- Domicilio: .....
- C.P.: ..... Población: .....
- Teléfono: ..... DNI/CIF: .....
- Fecha de inscripción: .....

Datos Profesionales

- Departamento: .....
- Facultad o Escuela: .....
- Universidad o Institución: .....
- Domicilio: .....
- C.P.: ..... Población: .....
- Teléfono: ..... Fax: .....
- Correo electrónico: .....
- Página web: `http://` .....
- Categoría Profesional: .....
- Líneas de Investigación: .....

.....

Dirección para la correspondencia: ☐ Profesional ☐ Personal

Cuota anual para el año 2010

☐ Socio ordinario: 30€ ☐ Socio de reciprocidad con la RSME: 12€
☐ Socio estudiante: 15€

Datos bancarios

... de ..... de 201..

Muy Sres. Míos:

Ruego a Uds. que los recibos que emitan a mi cargo en concepto de cuotas de inscripción y posteriores cuotas anuales de SëMA (Sociedad Espa nola de Matemática Aplicada) sean pasados al cobro en la cuenta cuyos datos figuran a continuación

Entidad (4 dígitos)	Oficina (4 dígitos)	D.C. (2 dígitos)	Número de cuenta (10 dígitos)

- Entidad bancaria: .....
- Domicilio: .....
- C.P.: ..... Población: .....

Con esta fecha, doy instrucciones a dicha entidad bancaria para que obren en consecuencia.

Atentamente,

Fdo. .....

Para remitir a la entidad bancaria

... de ..... de 201..

Muy Sres. Míos:

Ruego a Uds. que los recibos que emitan a mi cargo en concepto de cuotas de inscripción y posteriores cuotas anuales de SëMA (Sociedad Espa nola de Matemática Aplicada) sean cargados a mi cuenta corriente/libreta ..... en esa Agencia Urbana y transferidas a

SEMA: 0128 - 0380 - 03 - 0100034244
Bankinter
C/ Hernán Cortés, 63
39003 Santander

Atentamente,

Fdo. .....

Ficha de Inscripción Institucional

Sociedad Española de Matemática Aplicada SëMA

Remitir a: Iñigo Arregui, Dpto de Matemáticas, Fac. de Informática,
Universidad de A Coruña. Campus de Elviña, s/n. 15071 A Coruña.
CIF: G-80581911

Datos de la Institución

- Departamento: .....
- Facultad o Escuela: .....
- Universidad o Institución: .....
- Domicilio: .....
- C.P.: ..... Población: .....
- Teléfono: ..... DNI/CIF: .....
- Correo electrónico: .....
- Página web: `http://` .....
- Fecha de inscripción: .....

Forma de pago

La cuota anual para el año 2009 como Socio Institucional es de 150€.
El pago se realiza mediante transferencia bancaria a

SEMA: 0128 - 0380 - 03 - 0100034244
Bankinter
C/ Hernán Cortés, 63
39003 Santander