Rendiconti di Matematica, Serie VII
Volume 24, Roma (2004), 201-237

Optimal and approximate control of finite-difference approximation schemes for the 1D wave equation

ENRIQUE ZUAZUA

ABSTRACT: *We address the problem of control of numerical approximation schemes for the wave equation. More precisely, we analyze whether the controls of numerical approximation schemes converge to the control of the continuous wave equation as the mesh-size tends to zero.*
Recently, it has been shown that, in the context of exact control, i.e., when the control is required to drive the solution to a final target exactly, due to high frequency spurious numerical solutions, convergent numerical schemes may lead to unstable approximations of the control. In other words, the classical convergence property of numerical schemes does not guarantee a stable and convergent approximation of controls.
In this article we address the same problem in the context of optimal and approximate control in which the final requirement of achieving the target exactly is relaxed. We prove that, for those relaxed control problems, convergence (as the mesh-size tends to zero) holds. In particular, in the context of approximate control we show that, if the final condition is relaxed so that the final state is required to reach and ε-neighborhood of the final target with $\varepsilon > 0$, then the controls of numerical schemes (the so-called ε-controls) converge to the ε-controls of the wave equation. We also show that this result fails to be true in several space dimensions.
Although convergence is proved in the context of these relaxed control problems, the fact that instabilities occur at the level of exact control have to be considered as a serious warning in the sense that instabilities may ultimately arise if the control requirement is reinforced to exactly achieve the final target, i.e., as ε is taken smaller and smaller.

CONTENTS:

1 – Introduction

In recent years important progress has been made on problems of observation and control of wave phenomena. Much less is known about numerical approximation schemes.

The problems of observability and controllability can be stated as follows:

- *Observability.* Assuming that waves propagate according to a given wave equation and with suitable boundary conditions, can one guarantee that their whole energy can be estimated in terms of the energy concentrated on a given subregion of the domain (or its boundary) where propagation occurs in a given time interval?
- *Controllability.* Can solutions be driven to a given state at a given final time by means of a control acting on the system on that subregion?

It is well known that the two problems are equivalent provided one chooses an appropriate functional setting, which depends on the equation (see for, instance, [53], [83]).

But several different variants are meaningful and possible. In particular, at the level of the controllability problem, one can consider several degrees of

precision on the requirement of reaching the given target. For instance, one can require the control to drive the solution to the target exactly, this is the so-called *exact controllability* problem, or only in an approximate way, the so called *approximate controllability* one. One may also formulate the problem in the context of *optimal control*, minimizing a functional measuring the distance to the target in a suitable class of admissible controls.

Each of these control properties can be interpreted by duality as a suitable observability property. Obviously, stronger the control property under consideration is, stronger the corresponding observability property will be as well.

In this work we shall mainly focus on the issue of how these two properties behave under numerical approximation schemes for two particular control problems: *optimal* and *approximate control*. More precisely, we shall discuss the problem of whether, as the mesh-size tends to zero, the controls of numerical approximation schemes converge to the controls of the continuous wave equation.

This article is devoted to the wave equation as a simplified hyperbolic problem arising in many areas of Mechanics, Engineering and Technology. It is indeed, a model for describing the vibrations of structures, the propagation of acoustic or seismic waves, etc. Therefore, the control of the wave equation enters in a way or another in problems related with control mechanisms for structures, buildings in the presence of earthquakes, for noise reduction in cavities and vehicles, etc.

By now it is well known that, in the context of the exact controllability problem, the answer to the question is negative in the sense that exact controls of numerical approximation schemes may diverge as the mesh-size tends to zero. This is due to the classical numerical dispersion phenomena. Indeed, it is well known that the interaction of waves with a numerical mesh produces dispersion phenomena and spurious[(1)] high frequency oscillations [76], [74]. In particular, because of this nonphysical interaction of waves with the discrete medium, the velocity of propagation of numerical waves and, more precisely, the so called *group velocity*[(2)] may converge to zero when the wavelength of solutions is of the order of the size of the mesh and the latter tends to zero. As a consequence of this fact, the time needed to uniformly (with respect to the mesh size) observe (or control) the numerical waves *exactly* from the boundary or from a subset of the medium in which they propagate may tend to infinity

[(1)] The adjective spurious will be used to designate any component of the numerical solution that does not correspond to a solution of the underlying PDE. In the context of the wave equation, this happens at the high frequencies and, consequently, these spurious solutions weakly converge to zero as the mesh size tends to zero. Consequently, the existence of these spurious oscillations is compatible with the convergence (in the classical sense) of the numerical scheme, which does indeed hold for fixed initial data.

[(2)] At the numerical level it is important to distinguish the notions of phase and group velocity. Phase velocity refers to the velocity of propagation of individual monocromatic waves, while group velocity corresponds to the velocity of propagation of wave packets, that may significantly differ from the phase velocity when waves of similar frequencies are combined. See, for instance, [74].

as the mesh becomes finer. This is the reason for the unstable behavior of the control and observation properties of most numerical approximation schemes as the mesh-size tends to zero.

But that happens, as mentioned above, for the problems of exact observation and control. Exact observation means that the total energy of solutions is reconstructed from partial measurements uniformly, independently of the solution. Exact control means that one wishes to drive the solution exactly to a final target.

The main goal of this article is to show that when these requirements are relaxed, and one considers the problems of approximate and/or optimal control, then instabilities disappear and classical numerical schemes provide convergent approximations of controls.

In this paper we first briefly describe why numerical dispersion and spurious high frequency oscillations are an obstacle for the convergence of exact controls.

We then address the problems of approximate and optimal control. We prove, combining classical results on the convergence of numerical schemes and Γ-convergence arguments, that controls converge for the relaxed optimal and approximate control problems.

All we have said up to now concerning the wave equation can be applied with minor changes to several other models that are purely conservative like Schrödinger and beam equations (see the survey article [87] for a comparison between these models and their behavior in what concerns numerics and control).

However, many models from physics and mechanics have some damping mechanism built in. When the damping term is "mild" the qualitative properties are the same as those we discussed above. However, some other dissipative mechanisms may have much stronger effects. This is for instance the case for the thermal effects arising in the heat equation itself but also in some other more sophisticated systems, like the system of thermoelasticity. Roughly speaking one may say that the strong damping mechanisms help for the convergence of controls of numerical schemes. There is actually an extensive literature on optimal control of parabolic equations that confirms this fact [44], [68], [75], ... We also refer to E. Casas [9] for the analysis of finite-element approximations of elliptic optimal control problems and to [17] for an optimal shape design problem for the Laplace operator. But this has been done mainly in the context of optimal control and very little is known about the controllability issues that we address in this paper (we refer to [87] for a discussion of this topic and for a list of related open problems). For instance, as we shall see, in several space dimensions, the problem of analyzing the behavior of approximate controls for the heat equation is mainly open too.

Most of the analysis we shall present here has been also developed in the context of a more difficult problem, related to the behavior of the conservation/control properties in homogenization. There, the coefficients of the wave equation oscillate rapidly on a scale δ that tends to zero, so that the equation

homogenizes to a constant coefficient one. In that framework the interaction of high frequency waves with the microstructure produces localized waves at high frequency. These localized waves are an impediment for the uniform observation/control properties to hold. But, once more, this impediments do not arise if the control requirement is relaxed [12] and [48]. This was already observed in the context of homogenization and approximate control of the heat equation in [84]. The analogies between both problems (homogenization and numerical approximation) are clear: the mesh size h in numerical approximation schemes plays the role of the parameter δ in homogenization (see [85] and [14] for a discussion of the connection between these problems). Although the analysis of the numerical problem is much easier from a technical point of view, it was only developed after the problem of homogenization was understood. This is due in part to the fact that, from a control theoretical point of view, there was a conceptual difficulty to match the existing finite-dimensional and infinite-dimensional theories. This article may also be viewed as a further step in that direction showing that although the instabilities do arise at the level of exact control, optimal and approximate control problems ate often well-behaved.

This paper is mainly concerned with finite-difference numerical approximation schemes for $1D$ wave equations but the results and techniques extend easily to most common numerical approximation schemes, like finite-element methods, and also to fully discrete approximations. As we shall see, however, interesting open problems arise in several space dimensions where the questions under investigation exhibit new and not completely understood geometrical aspects. The rest of this paper is organized as follows.

In Section 2 we recall the basic ingredients of the finite-dimensional theory we will need along the paper. In particular we shall introduce the Kalman rank condition.

Section 3 is devoted to presenting and discussing the problems of observability and controllability for the constant coefficient wave equation. In Section 4 we discuss the finite-difference space semi-discretization of the $1D$ wave equation and recall the main results on the lack of controllability and observability. We also comment on some remedies and cures that have been introduced in the literature to avoid these instabilities.

In Section 5 and 6 we show that numerical approximation schemes are well behaved if the control requirement is relaxed to an approximate or optimal control problem, respectively. In Section 7 we briefly discuss the problem of stabilization. Finally, in Section 8 we formulate an interesting open problem related with the extension of the results in this paper to several space dimensions.

The interested reader is referred to the survey articles [81] and [83] for a more complete discussion of the state of the art in the controllability of partial differential equations and to [87] for what concerns numerical issues.

2 – Preliminaries on finite-dimensional systems

Most of this article is devoted to analyze the wave equation and its numerical approximations. Numerical approximation schemes and more precisely those that are semi-discrete (discrete in space and continuous in time) yield finite-dimensional systems of ODE's. There is by now an extensive literature on the control of finite-dimensional systems and the problem is completely understood for linear ones [50]. The problem of convergence of controls as the mesh-size in the numerical approximation tends to zero is very closely related to passing to the limit as the dimension of finite-dimensional systems tends to infinity. The later topic is widely open and this article may be considered as a contribution in this direction.

In this section we briefly summarize the most basic material on finite-dimensional systems that will be used along this article (we refer to [59] for more details).

Consider the finite-dimensional system of dimension N:

$$x' + Ax = Bv, \quad 0 \leq t \leq T; \quad x(0) = x_0, \tag{2.1}$$

where x is the N-dimensional state and v is the M-dimensional control, with $M \leq N$.

Here A is an $N \times N$ matrix with constant real coefficients and B is an $M \times N$ matrix. The matrix A determines the dynamics of the system and the matrix B models the way controls act on it.

Obviously, in practice, it would be desirable to control the N components of the system with a low number of controls. The best would be to do it by means of a scalar control, in which case $M = 1$. This is typically the situation when dealing with the boundary control of numerical approximation schemes of the $1D$ wave equation.

System (2.1) is said to be controllable in time T when every initial datum $x_0 \in \mathbb{R}^N$ can be driven to any final datum x_1 in $\mathbb{R}^N$ in time T and, more precisely, if for any x_0, $x_1 \in \mathbb{R}^N$ there exists $v \in L^2(0,T;\mathbb{R}^M)$ such that the solution of (2.1) satisfies

$$x(T) = x_1. \tag{2.2}$$

It turns out that for finite-dimensional systems there is a necessary and sufficient condition for controllability which is of purely algebraic nature. It is the so called *Kalman condition: System* (2.1) *is controllable in some time* $T > 0$ *iff*

$$\operatorname{rank}[B,\, AB, \ldots,\, A^{N-1}B] = N. \tag{2.3}$$

According to this, in particular, system (2.1) is controllable in some time T if and only if it is controllable for all time.

There is a direct proof of this result which uses the representation of solutions of (2.1) by means of the variation of constants formula. However, the methods we shall develop along this article rely more on the dual (but completely equivalent!) problem of observability of the adjoint system.

Consider the *adjoint system*

$$-\varphi' + A^*\varphi = 0, \quad 0 \le t \le T;\ \varphi(T) = \varphi_0. \tag{2.4}$$

It is not difficult to see that system (2.1) is controllable in time T if and only if the adjoint system (2.4) is *observable* in time T, i.e. if there exists a constant $C > 0$ such that, for all solution φ of (2.4),

$$| \varphi_0 |^2 \le C \int_0^T | B^*\varphi |^2 \, dt. \tag{2.5}$$

Before analyzing (2.5) in more detail let us see that this observability inequality does indeed imply controllability of the state equation.

Assume the observability inequality (2.5) holds and consider the following quadratic functional $J : \mathbb{R}^N \to \mathbb{R}$:

$$J(\varphi_0) = \frac{1}{2} \int_0^T | B^*\varphi(t) |^2 \, dt - < x_1, \varphi_0 > + < x_0, \varphi(0) > . \tag{2.6}$$

It is easy to see that, if $\tilde{\varphi}_0$ is a minimizer for J, then the control $v = B^*\tilde{\varphi}$, where $\tilde{\varphi}$ is the solution of the adjoint system (2.4) with that datum at time $t = T$, is such that the solution $x = x(t)$ of the state equation satisfies the control requirement (2.2). Indeed, it is sufficient to write down explicitly the fact that the differential of J at the minimizer vanishes.

Thus, the controllability problem is reduced to minimizing the functional J. Applying the Direct Method of the Calculus of Variations it can be shown that J achieves its minimum since the functional J is continuous and convex and it is also coercive according to the observability inequality (2.5). Indeed, note that when (2.5) holds the following variant holds as well, with possibly a different constant $C > 0$:

$$| \varphi_0 |^2 + | \varphi(0) |^2 \le C \int_0^T | B^*\varphi |^2 \, dt. \tag{2.7}$$

This gives a constructive way of building the controls, as a minimum of J.

The coercivity of J requires the Kalman condition (2.3) to be satisfied. The rank condition (2.3) turns out to be equivalent to the adjoint one

$$\operatorname{rank}[B^*,\ B^*A^*, \dots,\ B^*[A^*]^{N-1}] = N. \tag{2.8}$$

To see the equivalence between (2.7) and (2.8) let us note that, since we are in finite-dimension, using that all norms are equivalent[(3)], the observability inequality (2.7) is equivalent to a uniqueness property:

(2.9) (UP) Does the fact that $B^*\varphi \equiv 0$ for all $\leq t \leq T$ imply that $\varphi \equiv 0$?

And, as we shall see, this uniqueness property is precisely equivalent to the adjoint Kalman condition (2.8).

REMARK 2.1. Before proving this statement we note that $B^*\varphi$ is only an M-dimensional projection of the solution φ who has N components. Therefore, in order for this property (UP) to be true the operator B^* has to be chosen in a strategic way, depending of the state matrix A. The Kalman condition is the right test to check whether the choice of B^* (or B) is appropriate.

Let us finally prove that the uniqueness property (UP) holds when the adjoint rank condition (2.8) is fulfilled. In fact, taking into account that solutions φ are analytic in time, the fact that $B^*\varphi$ vanishes is equivalent to the fact that all the derivatives of $B^*\varphi$ of any order at time $t = T$ vanish. But the solution φ admits the representation $\varphi(t) = e^{A^*(t-T)}\varphi_0$ and therefore all the derivatives of $B^*\varphi$ at time $t = T$ vanish if and only if $B^*[A^*]^k\varphi_0 \equiv 0$ for all $k \geq 0$. According to the Cayley-Hamilton's theorem this is equivalent to the fact that $B^*[A^*]^k\varphi_0 \equiv 0$ for all $k = 0, \ldots, N-1$. Finally, the latter is equivalent to $\varphi_0 \equiv 0$ (i.e. $\varphi \equiv 0$) if and only if the adjoint Kalman rank condition (2.8) is fulfilled.

REMARK 2.2. It is important to note that in this finite-dimensional context, the time T of control plays no role. In particular, whether a system is controllable (or its adjoint observable) is independent of the time T of control.

REMARK 2.3. In the finite-dimensional context of this section we have only considered the problem of exact controllability. This is so since, in this case, approximate and exact controllability are equivalent properties. Approximate controllability refers to the situation in which the set of reachable states is dense in the space where solutions live. In this case, since we are in $\mathbb{R}^N$, this is equivalent to the fact that the set of reachable states in the whole $\mathbb{R}^N$ and this is precisely when exact controllability holds. The dual version of this equivalence property reads as follows: in finite-dimensions, the observability inequality (2.5) holds if and only if the uniqueness property (UP) is satisfied. None of these equivalences hold in general for infinite-dimensional dynamical systems.

The main task to be undertaken in order to pass to the limit in numerical approximations of control problems for wave equations as the mesh-size tends to zero is to explain why, even though at the finite-dimensional level the value

[(3)] This is the key point where finite and infinite dimensional systems behave so differently in what concerns controllability problems.

of the control time T is irrelevant, it may play a key role for the controllability/observability of the continuous PDE, as it is for instance the case in the context of the wave equation due to the finite speed of propagation.

3 – The constant coefficient wave equation

3.1 – Problem formulation: Observability

Let us consider the constant coefficient $1D$ wave equation:

$$\begin{cases} u_{tt} - u_{xx} = 0, & 0 < x < 1,\ 0 < t < T \\ u(0,t) = u(1,t) = 0, & 0 < t < T \\ u(x,0) = u^0(x),\ u_t(x,0) = u^1(x), & 0 < x < 1. \end{cases} \tag{3.1}$$

In (3.1) $u = u(x,t)$ describes the displacement of a vibrating string occupying the interval $(0,1)$.

The energy of solutions of (3.1) is conserved in time, i.e.

$$E(t) = \frac{1}{2}\int_0^1 \left[|u_x(x,t)|^2 + |u_t(x,t)|^2\right] dx = E(0),\ \ \forall 0 \le t \le T. \tag{3.2}$$

The problem of continuous boundary observability of (3.1) can be formulated, roughly, as follows: *to give sufficient conditions on the length of the time interval T such that there exists a constant $C(T) > 0$ so that the following inequality holds for all solutions of* (3.1):

$$E(0) \le C(T)\int_0^T |u_x(1,t)|^2\, dt. \tag{3.3}$$

This corresponds to the exact controllability property of the wave equation with control on $x = 1$ we shall discuss in the next subsection.

Inequality (3.3), when it holds, guarantees that the total energy of a solution can be "observed" or estimated from the energy concentrated or measured on the extreme $x = 1$ of the string during the time interval $(0,T)$ uniformly in the whole class of solutions of (3.1).

Here and in the sequel, the best constant $C(T)$ in inequality (3.3) will be referred to as the *observability constant*.

Of course, one can formulate a weakened version of this observability property which consists simply on the following uniqueness problem:

(3.4) If the solution u of (3.1) is such that $u_x(1,\, t) \equiv 0$ for $0 \le t \le T$, then $u \equiv 0$?

When this uniqueness property holds, we say that the system (3.1) is *weakly observable*. Of course, since we are now dealing with a PDE and therefore we are

necessarily in the context of an infinite dimensional dynamical system, unlike in the previous section, the fact that this uniquenes property holds does not automatically guarantee that the observability inequality (3.3) holds as well.

REMARK 3.1. This is just an example of a variety of similar observability problems. Among its possible variants, the following are worth mentioning: (a) one could observe the energy concentrated on the extreme $x = 0$ or in the two extremes $x = 0$ and 1 simultaneously; (b) the $L^2(0,T)$-norm of $u_x(1,t)$ could be replaced by some other norm, (c) one could also observe the energy concentrated in a subinterval (α, β) of the space interval $(0,1)$ occupied by the string, etc.

3.2 – Exact controllability

As we mentioned above, the observability problem above is equivalent to a boundary controllability one[(4)]. More precisely, the observability inequality (3.3) holds, if and only if, for any $\left(y^0, y^1\right) \in L^2(0,1) \times H^{-1}(0,1)$ there exists $v \in L^2(0,T)$ such that the solution of the controlled wave equation

$$\left\{\begin{array}{ll} y_{tt} - y_{xx} = 0, & 0 < x < 1,\ 0 < t < T \\ y(0,t) = 0;\ y(1,t) = v(t), & 0 < t < T \\ y(x,0) = y^0(x),\ y_t(x,0) = y^1(x), & 0 < x < 1 \end{array}\right. \tag{3.5}$$

satisfies

$$y(x,T) = y_t(x,T) = 0, \quad 0 < x < 1. \tag{3.6}$$

REMARK 3.2. Needless to say, in this control problem the goal is to drive solutions to equilibrium at the time $t = T$. Once the configuration is reached at time $t = T$, the solution remains at rest for all $t \geq T$, by taking null control for $t \geq T$, i.e. $v \equiv 0$ for $t \geq T$.

REMARK 3.3. It is convenient to note that (3.1) is not, strictly speaking, the adjoint of (3.5). The initial data for the adjoint system should be given at time $t = T$. But, in view of the time-irreversibility of the wave equations under consideration this is irrelevant. Obviously, one has to be more careful about this when dealing with time irreversible systems as the heat equation.

Let us check first that observability implies controllability since the proof is of a constructive nature and allows to build the control of minimal norm ($L^2(0,T)$-norm in the present situation) by minimizing a convex, continuous and coercive functional in a Hilbert space. In the present case, given $\left(y^0, y^1\right) \in$

[(4)] We refer to J. L. LIONS [53] for a systematic analysis of the equivalence between controllability and observability through the so called Hilbert Uniqueness Method (HUM).

$L^2(0,1) \times H^{-1}(0,1)$ the control $v \in L^2(0,T)$ of minimal norm for which (3.6) holds is of the form

$$v(t) = u_x^*(1,t), \tag{3.7}$$

where u^* is the solution of the adjoint system (3.1) corresponding to initial data $(u^{0,*}, u^{1,*}) \in H_0^1(0,1) \times L^2(0,1)$ minimizing the functional,

$$J((u^0,u^1)) = \frac{1}{2}\int_0^T |u_x(1,t)|^2 dt + \int_0^1 y^0 u^1 dx - < y^1, u^0 >_{H^{-1}\times H_0^1}, \tag{3.8}$$

in the space $H_0^1(0,1) \times L^2(0,1)$.

Note that J is convex. The continuity of J in $H_0^1(0,1) \times L^2(0,1)$ is guaranteed by the fact that the solutions of (3.1) satisfy the extra regularity property that $u_x(1,t) \in L^2(0,T)$ (a fact that holds also for the Dirichlet problem for the wave equation in several space dimensions, see [45], [53], [54]). More, precisely, for all $T > 0$ there exists a constant $C_*(T) > 0$ such that

$$\int_0^T \left[|u_x(0,t)|^2 + |u_x(1,t)|^2 \right] dt \leq C_*(T) E(0), \tag{3.9}$$

for all solution of (3.1).

Thus, in order to guarantee that the functional J achieves its minimum, it is sufficient to prove that it is coercive. This is guaranteed by the observability inequality (3.3).

Once coercivity is known to hold the Direct Method of the Calculus of Variations (DMCV) allows showing that the minimum of J over $H_0^1(0,1) \times L^2(0,1)$ is achieved. By the strict convexity of J the minimum is unique and we denote it, as above, by $(u^{0,*}, u^{1,*}) \in H_0^1(0,1) \times L^2(0,1)$, the corresponding solution of the adjoint system (3.1) being u^*.

The functional J is of class C^1. Consequently, the gradient of J at the minimizer vanishes and this is equivalent to

$$\int_0^1 y(T) w_t(T) dx - < y_t(T),\, w(T) >_{H^{-1}\times H_0^1} = 0, \tag{3.10}$$

for all $(w^0, w^1) \in H_0^1(0,1) \times L^2(0,1)$, w being the corresponding solution of (3.1). Obviously, this condition is equivalent to the exact controllability one ($y(T) \equiv y_t(T) \equiv 0$) since, whenever (w^0, w^1) covers the whole space $H_0^1(0,1) \times L^2(0,1)$, $(w(T), w_t(T))$ does it as well.

This argument shows that *continuous observability implies controllability.* The reverse is also true.

The main difference with respect to finite-dimensional systems is that the unique continuation property (3.4) does not imply the observability inequality to hold.

3.3 – Approximate controllability

Let us now discuss the control theoretical consequences of the weak observability or unique continuation property (3.4), a property that holds when $T \geq 2$ too. When this property holds the system is *approximately controllable* which means that, for all $\varepsilon > 0$ there is a control v_ε in $L^2(0,\, T)$ such that the solution satisfies

$$[\| y_\varepsilon(x,\, T) \|^2_{L^2(0,\,1)} + \| y_t(x,\, T) \|^2_{H^{-1}(0,\,1)}]^{1/2} \leq \varepsilon. \tag{3.11}$$

The control satisfying (3.11) can be built as above but this time the functional to be minimized has to be slightly perturbed[(5)]:

$$\begin{aligned} J_\varepsilon((u^0, u^1)) = \frac{1}{2}\int_0^T | u_x(1,t) |^2 \, dt + \varepsilon \| (u^0, u^1) \|_{H_0^1(0,1)\times L^2(0,1)} + \\ + \int_0^1 y^0 u^1 dx - \int_0^1 y^1 u_0 dx. \end{aligned} \tag{3.12}$$

In [21] it was proved, in the context of the approximate controllability of the heat equation, that adding the ε-term in the functional J_ε guarantess its coercivity as a direct consequence of the weak observability property, without requiring the observability inequality to hold.

The same is true in the present case: if weak observability holds then the functional J_ε satisfies the coercivity property

$$\underline{\lim}_{\|(u^0,\, u^1)\|_{H_0^1(0,\,1)\times L^2(0,\,1)}\to\infty} \frac{J_\varepsilon(u^0, u^1)}{\| (u^0, u^1) \|_{H_0^1(0,1)\times L^2(0,1)}} \geq \varepsilon. \tag{3.13}$$

Moreover the functional J_ε achieves its minimum at a single point $(u^{0,*}, u^{1,*})$ of $H_0^1(0,\, 1)\times L^2(0,\, 1)$. The control $v = u_x^*(1,\, t)$ is then such that (3.11) is satisfied.

Remark 3.4. In the present $1D$ case both the unique continuation and observability inequality hold if and only if $T \geq 2$. But, in several space dimensions, the observability inequality requires of further geometric constrainst. More precisely, it is required that the so-called Geometric Control Condition (GCC) is satisfied by the subset of the boundary where observation is being made (see [4]). Recall that, roughly speaking, GCC consists on requiring that all rays of Geometric Optics enter the control region in a time which is less than the control time.

[(5)] Here and in the sequel $-\int_0^1 y^1 u_0 dx$ denotes the duality pairing between $u^0 \in H_0^1(0,1)$ and $y^1 \in H^{-1}(0,1)$.

3.4 – Observability

The following holds:

PROPOSITION 3.1. *For any $T \geq 2$, system* (3.1) *is observable. In other words, for any $T \geq 2$ there exists $C(T) > 0$ such that* (3.3) *holds for any solution of* (3.1). *Conversely, if $T < 2$,* (3.1) *is not observable, or, equivalently,*

$$\sup_{u \text{ solution of (3.1)}} \left[\frac{E(0)}{\displaystyle\int_0^T |\, u_x(1,t)\,|^2 \, dt} \right] = \infty. \tag{3.14}$$

The proof of observability for $T \geq 2$ can be carried out in several ways. The simplest one uses the Fourier representation of solutions [87] but it is insufficient to deal with multidimensional problems. In several space dimensions one may use multipliers (KOMORNIK, [45]; LIONS, [53]), Carleman inequalities (ZHANG, [79]), and microlocal tools (BARDOS et al., [4]; BURQ and GÉRARD, [7]).

On the other hand, for $T < 2$ the observability inequality does not hold, due to the finite speed of propagation (= 1 in the model under consideration).

Summarizing, Proposition 3.1 states that, in one space dimension, a necessary and sufficient condition for the observability (both in its strong and weak version) to hold is that $T \geq 2$.

4 – 1*D* Finite-difference semi-discretizations

In this section we discuss the observability/controllability properties of a semi-discrete finite-difference approximation of the wave equation. This problem arises naturally in the numerical approximation of controls.

We describe the following results, of negative nature:

- The observability constant for the semi-discrete model tends to infinity for any T as the mesh-size h tends to zero.
- There are initial data for the wave equation for which the exact controls of the semi-discrete models diverge as $h \to 0$, This proves that one can not simply rely on the classical convergence (consistency + stability) analysis of the underlying numerical schemes to design stable algorithms for computing the controls.

We also briefly recall some of the basic cures that have been developed in the literature to avoid this high frequency numerical pathologies.

4.1 – Finite-difference approximations

Let us now formulate these problems and state the corresponding results in a more precise way.

Given $N \in \mathbf{N}$ we define $h = 1/(N+1) > 0$. We consider the mesh

$$x_0 = 0;\ x_j = jh,\ j = 1, \dots, N;\ x_{N+1} = 1, \tag{4.1}$$

which divides $[0,1]$ into $N+1$ subintervals $I_j = [x_j, x_{j+1}]$, $j = 0, \dots, N$.

Consider the following finite difference approximation of the wave equation (3.1):

$$\begin{cases} u_j'' - \dfrac{1}{h^2}\left[u_{j+1} + u_{j-1} - 2u_j\right] = 0, & 0 < t < T,\ j = 1, \dots, N \\ u_j(t) = 0, & j = 0,\ N+1,\ 0 < t < T \\ u_j(0) = u_j^0,\ u_j'(0) = u_j^1, & j = 1, \dots, N. \end{cases} \tag{4.2}$$

Observe that (4.2) is a coupled system of N linear differential equations of second order. The function $u_j(t)$ provides an approximation of $u(x_j, t)$ for all $j = 1, \dots, N$, u being the solution of the continuous wave equation (3.1). The conditions $u_0 = u_{N+1} = 0$ reproduce the homogeneous Dirichlet boundary conditions, and the second order differentiation with respect to x has been replaced by the three-point finite difference.

We shall use a vector notation to simplify the expressions. Then, system (4.2) reads as follows

$$\begin{cases} \vec{u}\,''(t) + A_h \vec{u}(t) = 0, & 0 < t < T \\ \vec{u}(0) = \vec{u}^0,\ \vec{u}\,'(0) = \vec{u}^1 \end{cases} \tag{4.3}$$

where the matrix A is given by:

$$A_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}, \tag{4.4}$$

and the column vector

$$\overrightarrow{u}(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_N(t) \end{pmatrix} \tag{4.5}$$

represents the whole set of unknowns of the system.

The solution $\overrightarrow{u}$ of (4.3) depends also on h but often this will not be made explicit in the notation.

The energy of the solutions of (4.2),

$$E_h(t) = \frac{h}{2}\sum_{j=0}^{N}\left[\mid u_j' \mid^2 + \left|\frac{u_{j+1}-u_j}{h}\right|^2\right], \tag{4.6}$$

is constant in time. It is a natural discretization of the continuous energy (3.2).

The problem of observability of system (4.2) can be formulated as follows: *to find* $T > 0$ *and* $C_h(T) > 0$ *such that*

$$E_h(0) \leq C_h(T)\int_0^T \left|\frac{u_N(t)}{h}\right|^2 dt \tag{4.7}$$

holds for all solutions of (4.2).

Observe that $\mid u_N/h \mid^2$ is a natural approximation of $\mid u_x(1,t) \mid^2$ for the solution of the continuous system (3.1). Indeed $u_x(1,t) \sim [u_{N+1}(t) - u_N(t)]/h$ and, taking into account that $u_{N+1} = 0$, it follows that $u_x(1,t) \sim -u_N(t)/h$.

System (4.2) is finite-dimensional. Therefore, if observability holds for some $T > 0$, then it holds for all $T > 0$ as we have seen in Section 2.

Inequality (4.7) does indeed hold for all $T > 0$ and $h > 0$. This can be seen analyzing the Kalman rank condition.

4.2 – Non uniform observability

But the observability constant $C_h(T)$ diverges as $h \to 0$. To see this let us consider the eigenvalue problem

$$-\left[w_{j+1} + w_{j-1} - 2w_j\right]/h^2 = \lambda w_j,\ j = 1, \dots, N; \quad w_0 = w_{N+1} = 0. \tag{4.8}$$

The spectrum can be computed explicitly in this case (Isaacson and Keller [4.2]), the eigenvalues and eigenvectors being

$$\lambda_k^h = \frac{4}{h^2}\sin^2\left(\frac{k\pi h}{2}\right) \tag{4.9}$$

and

$$\vec{w}_k^{\,h} = (w_{k,1}, \dots, w_{k,N})^T : w_{k,j} = \sin(k\pi j h),\ k, j = 1, \dots, N. \tag{4.10}$$

Obviously,

$$\lambda_k^h \to \lambda_k = k^2\pi^2, \text{ as } h \to 0 \tag{4.11}$$

for each $k \geq 1$, $\lambda_k = k^2\pi^2$ being the k-th eigenvalue of the continuous wave equation (3.1). On the other hand we see that the eigenvectors $\vec{w}_k^{\,h}$ of the discrete system (4.8) coincide with the restriction to the mesh-points of the eigenfunctions $w_k(x) = \sin(k\pi x)$ of the continuous wave equation (3.1).

The main negative result on the lack of uniform (as $h \to 0$) observability inequality is as follows [39], [40]:

THEOREM 4.1. *For any $T > 0$ it follows that, as $h \to 0$,*

$$\sup_{u \text{ solution of (4.2)}} \left[\frac{E_h(0)}{\displaystyle\int_0^T \mid u_N/h \mid^2 dt} \right] \to \infty. \tag{4.12}$$

This negative result is a consequence of the following identity

$$h \sum_{j=0}^{N} \left| \frac{w_{j+1} - w_j}{h} \right|^2 = \frac{2}{4 - \lambda h^2} \left| \frac{w_N}{h} \right|^2 \tag{4.13}$$

and the fact that

$$\lambda_N^h h^2 \to 4 \text{ as } h \to 0. \tag{4.14}$$

But, the fact that isolated eigenvectors are badly observed on the boundary is not the only obstacle for the boundary observability property to be uniform as the mesh-size tends to zero. Indeed, let us consider the following solution of the semi-discrete system (4.2), constituted by the last two eigenvectors:

$$\vec{u} = \frac{1}{\sqrt{\lambda_N}} \left[\exp(i\sqrt{\lambda_N}t)\vec{w}_N - \exp(i\sqrt{\lambda_{N-1}}t)\vec{w}_{N-1} \right]. \tag{4.15}$$

This solution is a wave packet obtained as superposition of two monochromatic semi-discrete waves corresponding to the last two eigenfrequencies of the system. The total energy of this solution is of the order 1 (because each of both components has been normalized in the energy norm and the eigenvectors are orthogonal one to each other). However, the trace of its discrete normal derivative tend to zero in $L^2(0,T)$ as $h \to 0$. This is due to two facts.

- First, the trace of the discrete normal derivative of each eigenvector is of order h compared to its total energy.
- Second and more important, the gap between $\sqrt{\lambda_N}$ and $\sqrt{\lambda_{N-1}}$ is of the order of h.

Thus, by Taylor expansion, the difference between the two time-dependent complex exponentials $\exp(i\sqrt{\lambda_N}t)$ and $\exp(i\sqrt{\lambda_{N-1}}t)$ is of the order Th.

This construction makes it possible to show that, whatever the time T is, the observability constant $C_h(T)$ in the semi-discrete system is at least of order $1/h$. In fact, this idea but combining an increasing number of high eigenfrequencies, can be used to show that the observability constant has to blow-up at infinite order. We refer to [58] for a precise analysis of the exponential blow-up of the observability constant.

The careful analysis of this negative example is extremely useful when designing possible remedies, i.e., to determine how one could modify the numerical scheme in order to reestablish the uniform observability inequality, since we have only found two obstacles and both happen at high frequencies. The first remedy is very natural: to cut off the high frequencies or, in other words, to ignore the high frequency components of the numerical solutions. This Fourier filtering method will be discussed later in some more detail. But let us first state the main consequences of the negative results above on the lack of uniform controllability.

4.3 – On the lack of uniform controllability

We have shown that the uniform observability property of the finite difference approximations (4.2) fails for any $T > 0$. In this subsection we explain the consequences of this result in the context of controllability.

The corresponding control system is:

$$\text{(4.16)} \qquad \begin{cases} y_j'' - \dfrac{1}{h^2}\left[y_{j+1} + y_{j-1} - 2y_j\right] = 0, & 0 < t < T,\ j = 1, \dots, N \\ y_0(0,t) = 0;\ y_{N+1}(1,t) = v(t), & 0 < t < T \\ y_j(0) = y_j^0,\ y_j'(0) = y_j^1, & j = 1, \dots, N, \end{cases}$$

and the question we consider is whether, for a given $T > 0$ and given initial data $(\vec{y}^{\,0}, \vec{y}^{\,1})$, there exists a control $v_h \in L^2(0,T)$ such that

$$\text{(4.17)} \qquad \vec{y}(T) = \vec{y}\,'(T) = 0.$$

System (4.2) being observable for all $h > 0$ and $T > 0$, system (4.16) is controllable for all $h > 0$ and $T > 0$, too.

However, this does not mean that the controls will be bounded as h tends to zero. In fact they diverge, even if $T \geq 2$. More precisely, we have the following main results:

- Taking into account that for all $h > 0$ the Kalman rank condition is satisfied, for all $T > 0$ and all $h > 0$ the semi-discrete system (4.16) is controllable. In other words, for all $T > 0$, $h > 0$ and initial data $(\vec{y}^{\,0}, \vec{y}^{\,1})$, there exists $v \in L^2(0,T)$ such that the solution $\vec{y}$ of (4.16) satisfies (4.17). Moreover, the

control v of minimal $L^2(0,T)$-norm can be built as in Section 3. It suffices to minimize the functional

$$J_h((\vec{u}^0,\vec{u}^1)) = \frac{1}{2}\int_0^T \left|\frac{u_N(t)}{h}\right|^2 dt + h\sum_{j=1}^N y_j^0 u_j^1 - h\sum_{j=1}^N y_j^1 u_j^0 \tag{4.18}$$

over the space of all initial data $(\vec{u}^0,\vec{u}^1)$ for the adjoint semi-discrete system (4.2).

Of course, in view of the observability inequality (4.7), this strictly convex and continuous functional is coercive and, consequently, has a unique minimizer.

Once we know that the minimum of J_h is achieved, the control is easy to compute. It suffices to take

$$v_h(t) = u_N^*(t)/h,\ 0<t<T, \tag{4.19}$$

as control to guarantee that (4.17) holds, where $\vec{u}^*$ is the solution of the semi-discrete adjoint system (4.2), corresponding to the initial data $(\vec{u}^{0,*},\vec{u}^{1,*})$ that minimize the functional J_h.

The control we obtain in this way is optimal in the sense that it is the one of minimal $L^2(0,T)$-norm. We can also get an upper bound on its size. Indeed, using the fact that $J_h \le 0$ at the minimum (which is a trivial fact since $J_h((0,0)) \le 0$), and the observability inequality (4.7), we deduce that

$$||v_h||_{L^2(0,T)} \le 4C_h(T)||(y^0,y^1)||_{*,h}, \tag{4.20}$$

where $||\cdot||_{*,h}$ denotes the norm

$$||(y^0,y^1)||_{*,h} = \sup_{(u_j^0,u_j^1)_{j=1,\dots,N}} \left[\left|h\sum_{j=1}^N y_j^0 u_j^1 - h\sum_{j=1}^N y_j^1 u_j^0\right| \Big/ E_h^{1/2}(u^0,u^1)\right]. \tag{4.21}$$

It is easy to see that this norm converges as $h\to 0$ to the norm in $L^2(0,1)\times H^{-1}(0,1)$. This norm can also be written in terms of the Fourier coefficients. It becomes a weighted euclidean norm whose weights are uniformly (with respect to h) equivalent to those of the continuous $L^2\times H^{-1}$-norm.

The estimate (4.20) is sharp and the constant $C_h(T)$ blows-up as h tends to zero. This has important consequences on the limit behavior of the control problem.

Indeed, according to Theorem 4.1, for all $T>0$ the constant $C_h(T)$ diverges as $h\to 0$. This shows, by the Banach-Steinhaus theorem, that there are initial data for the wave equation in $L^2(0,1)\times H^{-1}(0,1)$ such that the controls of the semi-discrete systems $v_h = v_h(t)$ diverge as $h\to 0$. There are

different ways of making this result precise. For instance, given initial data $(y^0, y^1) \in L^2(0,1) \times H^{-1}(0,1)$ for the continuous system, we can consider in the semi-discrete control system (4.16) the initial data that take the same Fourier coefficients as (y^0, y^1) for the indices $j = 1, \dots, N$. It then follows that, because of the divergence of the observability constant $C_h(T)$, there are necessarily some initial data $(y^0, y^1) \in L^2(0,1) \times H^{-1}(0,1)$ for the continuous system such that the corresponding controls v_h for the semi-discrete system diverge in $L^2(0,T)$ as $h \to 0$. Indeed, assume that for any initial data $(y^0, y^1) \in L^2(0,1) \times H^{-1}(0,1)$, the controls v_h remain uniformly bounded in $L^2(0,T)$ as $h \to 0$. Then, according to the uniform boundedness principle, we would deduce that the maps that associate the controls v_h to the initial data are also uniformly bounded. But this implies the uniform boundedness of the observability constant $C_h(T)$.

This lack of convergence is in fact easy to understand. As we have shown above, the semi-discrete system generates a lot of spurious high frequency oscillations. The control of the semi-discrete system has to take them into account. When doing this it gets further and further away from the true control of the continuous wave equation.

4.4 – Some remedies

Several remedies and cures have been proposed in the literature to avoid the unstabilities that high frequency numerical spurious solutions introduce both at the level of observation and control.

- **Fourier filtering**
  Filtering consists on considering subclasses of solutions of the adjoint system (4.2) constituted by the Fourier components corresponding to the eigenvalues $\lambda \leq \gamma h^{-2}$ with $0 < \gamma < 4$. This is equivalent to considering solutions whose only nontrivial components are those corresponding to the indices $0 < j < \delta h^{-1}$ with $0 < \delta < 1$. In these subclasses of solutions the observability inequality becomes uniform, i.e. the observability constant does not blow-up as h tends to zero. But for this to be true the time T of observability needs to be taken large enough and the value of the optimal observability time depends on the filtering parameters γ or δ. Note that these classes of solutions correspond to taking projections of the complete solutions by cutting off all frequencies with $\gamma h^{-2} < \lambda 4 h^{-2}$.
  More precisely, solutions of (4.2) can be developed in Fourier series as follows:

$$\vec{u} = \sum_{k=1}^{N} \left(a_k \cos\left(\sqrt{\lambda_k^h}\, t\right) + \frac{b_k}{\sqrt{\lambda_k^h}} \sin\left(\sqrt{\lambda_k^h}\, t\right) \right) \vec{w}_k^{\,h} \tag{4.22}$$

where a_k, b_k are the Fourier coefficients of the initial data, i.e.,

$$\vec{u}^0 = \sum_{k=1}^{N} a_k \vec{w}_k^h, \quad \vec{u}^1 = \sum_{k=1}^{N} b_k \vec{w}_k^h.$$

Given $0 < \delta < 1$, the classes of filtered solutions are of the form:

$$\mathcal{C}_\delta(h) = \left\{ \vec{u} \text{ sol. of (4.2) s.t. } \vec{u} = \sum_{k=1}^{[\delta/h]} \left(a_k \cos\left(\sqrt{\lambda_k^h} t\right) + \frac{b_k}{\sqrt{\lambda_k^h}} \sin\left(\sqrt{\lambda_k^h} t\right) \right) \vec{w}_k^h \right\}. \tag{4.23}$$

The Fourier filtering is natural since the numerical scheme, which converges in the classical sense, reproduces, at low frequencies, as $h \to 0$, the whole dynamics of the continuous wave equation. But, it also introduces a lot of high frequency spurious solutions. The scheme then becomes more accurate if we ignore the high frequency components and this makes the observability inequality uniform provided the time is taken to be large enough.

To prove the uniform (as $h \to 0$) observability result for filtered solutions of system (4.2), it is sufficient to combine a sharp analysis of the spectrum of the semi-discrete system under consideration and the classical *Ingham inequality* in the theory of nonharmonic Fourier series (see INGHAM [41] and YOUNG [77]). This analysis gives an explicit estimate of the optimal observability time in the class $\mathcal{C}_\delta(h)$: $T(\delta) = 2/\cos(\pi\delta/2)$. The minimal time $T(\delta)$ of uniform observability in this subclasses of filtered solutions is such that $T(\delta) \to 2$ as $\delta \to 0$ and $T(\delta) \to \infty$ as $\delta \to 1$, as one could expect. At the level of control, these results imply the uniform controllability of the projections of solutions of (4.16) over the subspace of the low eigenmodes that have not been cutted-off. One can then pass to the limit and prove the convergence towards the control of the continuous wave equation (3.5). This is so because, as h tends to zero, regardless of the value of the filtering parameter, one ends up recovering all the Fourier components of the state on the controlled projection.

We refer to [87] for a more details on the algorithm of control based on Fourier filtering.

In the context of the numerical computation of the boundary control for the wave equation the need of an appropriate filtering of the high frequencies was observed by R. GLOWINSKI [29]. This issue was further investigated numerically by M. ASCH and G. LEBEAU in [1]. There, finite difference schemes were used to test the Geometric Control Condition in various geometrical situations and to analyze the cost of the control as a function of time.

However, this method, which is natural from a theoretical point of view, can be hard to implement in numerical simulations. Indeed, solving the semi-discrete system provides the nodal values of the solution. One then needs

to compute its Fourier coefficients and, once this is done, to recalculate the nodal values of the filtered/truncated solution. Therefore, it is convenient to explore other ways of avoiding these high frequency pathologies that do not require going back and forth from the physical space to the frequency one. Several other possibilities have been introduced and analyzed in the literature. We mention them below.

- **Tychonoff regularization**
  GLOWINSKI et al. in [32] proposed a Tychonoff regularization technique that allows one to recover the uniform (with respect to the mesh size) coercivity of the functional that one needs to minimize to get the controls in the HUM approach. The method was tested to be efficient in numerical experiments. The convergence of the argument has been discussed in [87].

- **A two-grid algorithm**
  GLOWINSKI and LI in [31] introduced a two-grid algorithm that also makes it possible to compute efficiently the control of the continuous model. The method was further developed by GLOWINSKI in [29].
  The relevance and impact of using two grids can be easily understood in view of the analysis above of the $1D$ semi-discrete model. In view of the explicit expression of the eigenvalues of the semi-discrete system (4.9), all of them satisfy $\sqrt{\lambda} \leq 2/h$. We have also seen that the observability inequality becomes uniform when one considers solutions involving eigenvectors corresponding to eigenvalues $\sqrt{\lambda} \leq 2\gamma/h$, with $\gamma < 1$. Glowinski's 2-grid algorithm is based on the idea of using two grids: one with step size h and a coarser one of size $2h$. In the coarser mesh the eigenvalues obey the sharp bound $\lambda \leq 1/h^2$. Thus, the oscillations in the coarse mesh that correspond to the largest eigenvalues $\sqrt{\lambda} \sim 1/h$, in the finer mesh are associated to eigenvalues in the class of filtered solutions with parameter $\gamma = 1/2$. Then, this corresponds to a situation where the observability inequality is uniform for T large enough.
  The convergence of this method has recently been proved rigorously in [64] where the time of control was found to be $T > 4$, twice the control time for the continuous wave equation.

- **Mixed finite elements**
  An alternative approach consists in using mixed finite element methods rather than finite differences or standard finite elements, which require some filtering, Tychonoff regularization or multigrid techniques, as we have shown. First of all, it is important to underline that the analysis we have developed for the finite difference space semi-discretization of the $1D$ wave equation can be carried out with minor changes for finite element semi-discretizations as well. In particular, due to the high frequency spurious oscillations, uniform observability does not hold [40]. It is thus natural to consider mixed finite

element (m.f.e.) methods. This idea was introduced by BANKS et al. [2] in the context of boundary stabilization of the wave equation.
This method has been succesfully adapted in [11] for control purposes. It provides a good approximation of the wave equation and converges in classical terms. For this scheme the gap between the square roots of consecutive eigenvalues of its spectrum is uniformly bounded from below, and in fact tends to infinity for the highest frequencies as $h \to 0$. According to this and applying Ingham's inequality, the uniform observability property holds (see [11]).
The idea of correcting the dispersion diagram by modifying the numerical scheme has been previously explored in S. KRENK [46], for instance, where this was done by adding higher order terms in the approximation of the scheme. This approach has been also pursued by A. MUNCH [60] to enrich the class of schemes introduced in [11].

5 – Robustness of approximate controllability

In the previous sections we have shown that the exact controllability property behaves badly under most classical finite difference approximations. It is natural to analyze to what extent the high frequency spurious pathologies do affect other control problems and properties. In this section we focus on the problem of approximate controllability.

The approximate controllability problem is a relaxed version of the exact controllability one. The goal this time is to drive the solution of the controlled wave equation (3.5) not exactly to the equilibrium as in (3.6) but rather to an ε-state such that

$$\left[||y(T)||^2_{L^2(0,1)} + ||y_t(T)||^2_{H^{-1}(0,1)}\right]^{1/2} \leq \varepsilon. \tag{5.1}$$

When for all initial data (y^0, y^1) in $L^2(0,1) \times H^{-1}(0,1)$ and for all ε there is a control v such that (5.1) holds, we say that the system (3.5) is approximately controllable. Obviously, approximate controllability is a weaker notion than exact controllability and whenever the wave equation is exactly controllable, it is approximately controllable too.

As we have seen in Section 3.3, although exact controllability requires an observability inequality of the form of (3.3) to hold, for approximate controllability one only requires the uniqueness property (3.4).

This uniqueness property holds for $T \geq 2$ as well and can be easily proved using Fourier series or d'Alembert's formula. Its multidimensional version holds as well, as an immediate consequence of Holmgren's Uniqueness theorem (see [53]) for general wave equations with analytic coefficients and without geometric conditions, other than the time being large enough. In $1D$, because of the trivial

geometry, both the uniqueness property and observability inequality hold simultaneously for $T \geq 2$ but this is not longer true in several space dimensions.

Of course, the approximate controllability property by itself, as seen in Section 3.3, does not provide any information of what the cost of controlling to an ε-state as in (5.1) is, i.e. on what is the norm of the control v_ε needed to achieve the approximate control condition (5.1)[(6)]. But this issue will not be addressed here.

In what follows we fix some $\varepsilon > 0$. As mentioned above and seen in Section 3.3, once ε is fixed, we know that when $T \geq 2$, for all initial data (y^0, y^1) in $L^2(0,\, 1) \times H^{-1}(0,\, 1)$, there exists a control $v_\varepsilon \in L^2(0,\, T)$ such that (5.1) holds. Moreover, the control can be obtained minimizing a functional of the form (3.12).

The question we are interested in is the behavior of this property under numerical discretization.

Thus, let us consider the semi-discrete controlled version of the wave equation (4.16). We also fix the initial data in (4.16) "independently of h". This can be done in several ways:

a) When the data (y^0, y^1) of the continuous wave equation are smooth enough, for instance continuous, we may take the initial data for (4.16) as being the restriction of (y^0, y^1) to the mesh-points.
b) One may also take as initial for (4.16) the projection of the Fourier coefficients of (y^0, y^1) over the first N modes that can be represented on the discrete model.

Of course, (4.16) is also approximately controllable[(7)]. The question we address is as follows: *given initial data which are "independent of h", as above, with ε fixed, and given also the control time $T \geq 2$, is the control v_h of the semi-discrete system* (4.16) (*such that the discrete version of* (5.1) *holds*) *uniformly bounded as $h \to 0$?*

In the previous sections we have shown that the answer to this question in the context of exact controllability (which corresponds to taking $\varepsilon = 0$) is negative. However, we have also seen that relaxing the final requirement of reaching the target exactly may help. The following result shows that this is the case in the context of approximate control too.

THEOREM 5.1. *Assume that the initial data in* (4.16) *are essentially independent of h.*

[(6)] Roughly speaking, when exact controllability does not hold (for instance, in several space dimensions, when the GCC is not fulfilled), the cost of controlling blows up exponentially as ε tends to zero (see [66]). This type of result has been also proved in the context of the heat equation in [24]. But there the difficulty does not come from the geometry but rather from the regularizing effect of the heat equation.

[(7)] In fact, in finite dimensions, exact and approximate controllability are equivalent notions and, as we have seen, the Kalman condition is satisfied for system (4.16).

Assume that $T \geq 2$.
Then, for $\varepsilon > 0$ fixed, the controls v_h such that the solution of satisfies

$$\| (\vec{y_h}(T), \vec{y}'_h(T)) \|_{*,h} \leq \varepsilon, \tag{5.2}$$

are uniformly bounded in $L^2(0, T)$ as $h \to 0$.

Moreover, the controls v_h can be chosen such that they converge in $L^2(0, T)$ to a limit control v for which (5.1) *is realized for the continuous wave equation* (3.5).

This positive result on the uniformity of the approximate controllability property under numerical approximation when $\varepsilon > 0$ does not contradict the fact that the controls blow up for exact controllability (i.e. when $\varepsilon = 0$). These are in fact two complementary and compatible facts. For approximate controllability, one is allowed to concentrate an ε amount of energy on the solution at the final time $t = T$. For the semi-discrete problem this is done precisely in the high frequency components that are badly controllable as $h \to 0$, and this makes it possible to keep the control fulfilling (5.1) bounded as $h \to 0$.

The approximate control of the semi-discrete system can be obtained by minimizing the functional

$$J_h^*(\vec{u}^0, \vec{u}^1) = \frac{1}{2} \int_0^T \left| \frac{u_N(t)}{h} \right|^2 dt + \varepsilon ||(\vec{u}^0, \vec{u}^1)||_{\mathcal{H}^1 \times \ell^2} + h \sum_{j=1}^N y_j^0 u_j^1 - h \sum_{j=1}^N y_j^1 u_j^0 \tag{5.3}$$

over the space of all initial data $(\vec{u}^0, \vec{u}^1)$ for the adjoint semi-discrete system (4.2). In J_h^*, $|| \cdot ||_{\mathcal{H}^1 \times \ell^2}$ stands for the discrete energy norm, i.e. $|| \cdot || = \sqrt{2E_h}$. Note that there is an extra term $\varepsilon ||(\vec{u}^0, \vec{u}^1)||_{\mathcal{H}^1 \times \ell^2}$ in this new functional compared with the one we used to obtain the exact control (see (4.18)). On the other hand, the functional in (5.3) is a discrete version of the functional (3.12) one needs to minimize to get the approximate control for the continuous wave equation. In both cases, the controls one finds that way are those of minimal $L^2(0, T)$-norm.

Theorem 5.1 states the convergence of controls which are closely related to the minimizers of these functionals. Indeed, while the control v of the continuous wave equation (3.5) is defined as

$$v(t) = u_x^*(1, t), \tag{5.4}$$

u^* being the solution of the adjoint equation (3.1) with the initial data being the minimizer of the functional in (3.12), the control v_h of the semi-discrete equation (4.16) is defined as

$$v_h(t) = u_N^*(t)/h, \tag{5.5}$$

where u^* is the solution of the semi-discrete adjoint equation (4.2) with the minimizer of the functionals $J^*_{h,\,\varepsilon}$ in (5.3) as initial data.

Therefore, roughly speaking, Theorem 5.1 can be viewed as a Γ-convergence result [19] of the functional $J^*_{h,\,\varepsilon}$ towards J^*_ε.

Similar results have been proved in several different but related problems:

a) The approximate control of parabolic equations with rapidly oscillating coefficients and perforated domains in any space dimension (see [84] and [20], respectively) and the null control in $1D$ [55].
b) The exact controllability of the space semi-discretizations of the beam equation [51].

The key ingredient in the proof of Theorem 5.1 is the uniform (with respect to h) coercivity of the functionals $J^*_{h,\,\varepsilon}$. The following holds;

$$\underline{\lim}_{||(\vec{u}^{\,0},\vec{u}^{\,1})||_{\mathcal{H}^1\times\ell^2}\to\infty} \frac{J^*_h(\vec{u}^{\,0},\vec{u}^{\,1})}{||(\vec{u}^{\,0},\vec{u}^{\,1})||_{\mathcal{H}^1\times\ell^2}} \geq \varepsilon, \tag{5.6}$$

uniformly in h, provided $T \geq 2$.

Once the uniform observability property (5.6) holds, the minimizers are immediately uniformly bounded and the controls as well. Once this is done one can proceed in two steps:

a) First one shows that the weak limit (in $L^2(0,\,T)$) of the controls is a control for the limit system;
b) one then shows by Γ-convergence arguments that the limit control is precisely the one associated with the minimization of the limit functional J^*_ε;
c) finally one proves, using convexity and weak lower semicontinuity arguments, that $J^*_{h,\,\varepsilon}(\vec{u}^{\,*,0}_h, \vec{u}^{\,*,1}_h)$ tends to $J^*_\varepsilon(\vec{u}^{\,*,0}, \vec{u}^{\,*,1})$ as h tends to zero. This, together with the fact that the initial data to be controlled are essentially independent of h allows concluding that the $L^2(0,\,T)$-norms of the controls converge to the $L^2(0,\,T)$-norm of the limit controls. This guarantees that convergence holds in the strong topology.

We refer to [51] for the details of the proof in the closely related problem of the control of the beam equation.

Consequently, let us focus on the proof of the uniform coercivity property (5.6). At this level, the fact that $T \geq 2$ is essential. In order to show that the coercivity property above is uniform in $0 < h < 1$ we have to argue as in [84]. Mainly, we have to consider the case where $h \to 0$ and solutions of the adjoint semi-discrete system (4.2) converge to a solution of the continuous adjoint wave equation (3.1) such that $u_x(1,t) \equiv 0$ in $(0,T)$. Of course, if this happens with $T \geq 2$ we can immediately deduce that $u \equiv 0$ by the well known uniqueness property of the solutions of the wave equation discussed in Section 3.3. This suffices to conclude the uniform coercivity property.

This shows that the approximate controllability property is well-behaved under the semi-discrete finite-difference discretization of the wave equation. But the argument is in fact much more general and can be applied for other numerical approximation schemes. The two assumptions that are needed on the numerical scheme for this to hold are:

a) The scheme is convergent in the classical sense;
b) for all h the numerical scheme is controllable.

However, as we shall see, although these properties hold for most numerical schemes in $1D$, the second property may fail in several space dimensions unless some filtering is introduced or some extra geometric assumptions are imposed on the subset where the control is suppported.

6 – Robustness of optimal control

Finite horizon optimal control problems can also be viewed as relaxed versions of the exact controllability one.

Let us consider the following example in which the goal is to drive the solution of the wave equation (3.5) at time $t = T$ as closely as possible to the desired equilibrium state but penalizing the use of the control. In the continuous context the problem can be simply formulated as that of minimizing the functional

$$L^k(v) = \frac{k}{2}||(y(T), y_t(T))||^2_{L^2(0,1)\times H^{-1}(0,1)} + \frac{1}{2}||v||^2_{L^2(0,T)} \tag{6.1}$$

over $v \in L^2(0,T)$. This functional is continuous, convex and coercive in the Hilbert space $L^2(0,T)$. Thus it admits a unique minimizer that we denote by v_k. The corresponding optimal state is denoted by y_k. The penalization parameter establishes a balance between reaching the distance to the target and the use of the control. As k increases, the need of getting close to the target (the $(0,0)$ state) is emphasized and the penalization on the use of control is relaxed.

When exact (resp. approximate) controllability holds, i.e. when $T \geq 2$, it is not hard to see that the control one obtains by minimizing L^k converges, as $k \to \infty$, to an exact (resp. approximate) control for the wave equation (see [23]).

When the value of the parameter $k > 0$ is fixed, the optimal control v_k does not guarantee that the target ((0, 0) in this case) is achieved in an exact way. One can then measure the rate of convergence of the optimal solution $(y_k(T), y_{k,t}(T))$ towards $(0,0)$ as $k \to \infty$. When approximate controllability holds but exact controllability does not (a typical situation in several space dimensions when the GCC is not satisfied), the convergence of $(y_k(T), y_{k,t}(T))$ to $(0,0)$ in $L^2(0,1) \times H^{-1}(0,1)$ as $k \to \infty$ is very slow (roughly spaking, of logarithmic nature).

But here, once again, we fix any $k > 0$ and we discuss the behavior of the optimal control problem for the semi-discrete equation as $h \to 0$.

It is easy to write the semi-discrete version of the problem of minimizing the functional L^k. Indeed, it suffices to introduce the corresponding semi-discrete functional L_h^k replacing the $L^2 \times H^{-1}$-norm in the definition of L^k by the discrete norm introduced in (4.21). It is also easy to prove by the arguments we have developed in the context of approximate controllability, that, *as* $h \to 0$, *the control* v_h^k *that minimizes* L_h^k *in* $L^2(0,T)$ *converges to the minimizer of the functional* L^k *and the optimal solutions* y_h^k *of the semi-discrete system converge to the optimal solution* y^k *of the continuous wave equation in the appropriate topology*[(8)] *as* $h \to 0$ *too.*

In this case the proof of the uniform boundedness of the control is much easier since the uniform coercivity of the functionals L_h^k is obvious as soon as $k > 0$.

This shows that the optimal control problem is also well-behaved with respect to numerical approximation schemes, like the approximate control problem.

The reason for this is basically the same: in the optimal control problem the target is not required to be achieved exactly and, therefore, the pathological high frequency spurious numerical components are not required to be controlled.

In view of this discussion it becomes clear that the source of divergence in the limit process as $h \to 0$ in the exact controllability problem is the requirement of driving the high frequency components of the numerical solution exactly to zero. As we mentioned in the introduction, taking into account that optimal and approximate controllability problems are relaxed versions of the exact controllability one, even though they are theoretically well behaved under the numerical approximation process as our results above show, this negative result should be considered as a warning about the limit process as $h \to 0$ in general control problems.

7 – Stabilization

The problem of controllability has been addressed along this paper. The connections between the problems of controllability and stabilization are well known (see for instance [69], [78]) and similar developments could be carried out in the context of stabilization.

In the context of the wave equation, it is well known that the GCC suffices for stabilization and more precisely to guarantee the uniform exponential decay of solutions when a damping term, supported in the control region, is added

[(8)] Roughly, in $L^p([0,T]; L^2(0,1)) \cap W^{1,p}[0,T]; H^{-1}(0,1))$ for all $1 \leq p < \infty$, once the solution of the semi-discrete problem has been extended to the interior conveniently (as a piecewise linear and continuous function, for instance).

to the system. More precisely, when the subdomain ω of the domain Ω where the wave equation holds satisfies the GCC the solutions of the damped wave equation

$$y_{tt} - \Delta y + 1_\omega y_t = 0$$

with homogeneous Dirichlet boundary conditions are known to decay exponentially in the energy space. In other words, there exist constants $C > 0$ and $\gamma > 0$ such that

$$E(t) \leq Ce^{-\gamma t} E(0)$$

holds for every finite energy solution of the Dirichlet problem for this damped wave equation.

It is then natural to analyze whether the decay rate is uniform with respect to the mesh size for numerical discretizations. The answer is in general negative. Indeed, due to spurious high frequency oscillations, the decay rate fails to be uniform, for instance, for the classical finite difference semi-discrete approximation of the wave equation. This was established rigorously by F. Macià [56], [57] using Wigner measures. This negative result also has important consequences in many other issues related with control theory like infinite horizon control problems, Riccati equations for the optimal stabilizing feedback ([65]), etc.

We shall simply mention here that, even if the most natural semi-discretization schemes fail to be uniformly exponentially stable, the uniformity of the exponential decay rate can be reestablished if we add an internal viscous damping term to the equation (see [72], [73] and [61]).

In [72] we analyzed finite difference semi-discretizations of the damped wave equation

$$u_{tt} - u_{xx} + \chi_\omega u_t = 0, \tag{7.1}$$

where χ_ω denotes the characteristic function of the set ω where the damping term is effective. In particular we analyzed the following semi-discrete approximation in which an extra numerical viscous damping term is present:

$$\begin{cases} u_j'' - \dfrac{1}{h^2}\left[u_{j+1} + u_{j-1} - 2u_j\right] - \left[u_{j+1}' + u_{j-1}' - 2u_j'\right] - u_j'\chi_\omega = 0, \\ \qquad\qquad 0 < t < T,\ j = 1, \dots, N \\ u_j(t) = 0, \qquad\qquad 0 < t < T,\ j = 0,\ N+1 \\ u_j(0) = u_j^0, \quad u_j^1(0) = u_j^1, \qquad j = 1, \dots, N. \end{cases} \tag{7.2}$$

It was proved that this type of scheme preserves the uniform stabilization properties of the wave equation (7.1). To be more precise we recall that solutions of the $1D$ wave equation (7.1) in a bounded interval with Dirichlet boundary conditions decay exponentially uniformly as $t \to \infty$ when a damping term as above is added, ω being an open non-empty subinterval (see [80]). Using the numerical scheme above, this exponential decay property is kept with a uniform rate

as h tends to zero. The extra numerical damping that this scheme introduces adding the term $[u'_{j+1} + u'_{j-1} - 2u'_j]$ damps out the high frequency spurious oscillations that the classical finite difference discretization scheme introduces and that produce a lack of uniform exponential decay in the presence of damping.

The problem of whether this numerical scheme is uniformly observable or controllable as h tends to zero is an interesting open problem.

Note that the system above, in the absence of the damping term lozalized in ω, can be written in the vector form

$$\vec{u}'' + A_h\vec{u} + h^2A_h\vec{u}' = 0. \tag{7.3}$$

Here $\vec{u}$ stands, as usual, for the vector unknown $(u_1, \dots, u_N)^T$ and A_h for the tridiagonal matrix associated with the finite difference approximation of the Laplacian (4.4). In this form it is clear that the scheme above corresponds to a viscous approximation of the wave equation. Indeed, taking into account that A_h provides an approximation of $-\partial_x^2$, the presence of the extra multiplicative factor h^2 in the numerical damping term guarantees that it vanishes asymptotically as h tends to zero.

In [61] these results were extended to general domains in $2-d$. The subdomain ω was assumed to be a neighborhood of a subset of the boundary satisfying the classical multiplier condition, which constitutes a particular class of subdomains satisfying the GCC [80]. Then, adding a numerical viscosity term the uniform exponential decay was proved.

In the absence of geometric conditions on the subset ω, by only assuming that it is an open non-empty subset of Ω, using La Salle's invariance principle with the energy of the system as Lyapunov function, one can show that all solutions of the damped wave equation tend to zero as t goes to infinity without uniform exponential decay rate. This is true even in several space dimensions. This results turns out to be false at the semi-discrete level in the multi-dimensional case. Indeed, the property of decay relies on a unique continuation property similar to those we discussed in the context of approximate controllability. In the case of the continuous wave equation this property requires that whenever the solution u of the wave equation vanishes in $\omega \times (0, \infty)$, then it vanishes everywhere. This holds as a consequence of Holmgren's uniqueness theorem if $T>0$ is large enough. But it fails to be true for the semi-discrete equation without further restrictions on the subdomain ω as we shall see in open problem #2 below.

If one adds a numerical viscosity term, obviously, these difficulties dissapear and one recovers the decay of solutions of the semi-discrete system. But uniform (with respect to h) exponential decay rates can only be expected under geometric restrictions in ω as in [72] and [61]. Similar devepments have been carried out in [73] in the context of boundary damping in one-space dimension. Very likely similar results are true for boundary damping in several dimensions too. But a complete analysis of this issue using the techniques in [61] and [73] is still to be done.

8 – Open problems

1. *Moment problems techniques.* We have considered finite difference space semi-discretizations of the wave equation. We have addressed the problem of boundary observability and, more precisely, the problem of whether the observability estimates are uniform when the mesh size tends to zero.

We have proved that the uniform observability property does not hold for any time T. We have also described some possible remedies.

The main consequences concerning controllability have been mentioned. In particular, we have shown that exact controls of numerical approximation schemes may diverge.

By the contrary, we have proved that the problems of approximate and optimal control are well-behaved and that the convergence of the semi-discrete controls holds as the mesh-size h tends to zero.

It would be interesting to see if the moment problems techniques and the sharp estimates in [58] on biorthogonal families allow giving an alternative proof of these positive results with some explicit estimates on the size of the controls.

2. *Discrete unique-continuation.* As we mentioned above, the extension of Theorem 5.1 to the multi-dimensional case is no completely obvious. In fact, the results one gets change significantly.

Let us for instance discuss the simplest case of the constant coefficient wave equation in a square of $\mathbb{R}^2$. In [82] the instability of the controls was proved for finite difference semi-discrete approximations in the context of exact controllability. But, in view of Theorem 5.1, one could expect this not to be the case at the level of approximate controllability. But a new phenomena, producing new instabilities, arises in several space dimensions that we describe now.

In several space dimensions, for the continuous wave equation, approximate controllability holds from any open subset of the boundary if the control time is large enough (twice the diameter of the square domain is enough although a sharper estimate needs to take into account the geometry of the subset where the control is located). This means that the support of the control can be taken to be in any open subset of the domain or its boundary. But this fails to be true for the semi-discrete equation. Indeed, in $2-d$ the unique continuation or uniqueness property that is needed for the controllability of the semi-discrete approximation to hold is not satisfied automatically. In fact it is not even sufficient to assume that $h>0$ is small enough to guarantee that this uniqueness property is satisfied.

The following example due to O. Kavian [43] shows that, at the discrete level, new phenomena arise in what concerns the uniqueness problem. It concerns the eigenvalue problem for the 5-point finite difference scheme for the Laplacian in the square. A grid function taking alternating values ± 1 along a diagonal and vanishing everywhere else is an eigenvector with eigenvalue $\lambda = 4/h^2$. According to this example, even at the level of the elliptic equation, the domain ω where the solution vanishes has to be assumed to be large enough to guarantee the unique

continuation property. In [16] it was proved that when ω is a "neighborhood of one side of the boundary", then unique continuation holds for the discrete Dirichlet problem in any discrete domain. Here by a "neighborhood of one side of the boundary" we refer to the nodes of the mesh that are located immediately to one side of the boundary nodal points (left, right, top or bottom). Indeed, if one knows that the solution vanishes at the nodes immediately to one side of the boundary, taking into account that they vanish in the boundary too, the 5-point numerical scheme allows propagating the information and showing that the solution vanishes at all nodal points of the whole domain.

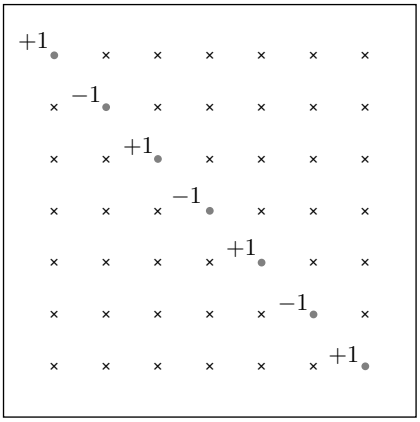

Fig. 1 The eigenvector for the 5-point finite difference scheme for the Laplacian in the square, with eigenvalue $\lambda = 4/h^2$, taking alternating values ± 1 along a diagonal and vanishing everywhere else in the domain.

Getting optimal geometric conditions on the set ω depending on the domain Ω where the equation holds, the discrete equation itself, the boundary conditions and, possibly, the frequency of oscillation of the solution for the unique continuation property to hold at the discrete level is an interesting and widely open subject of research.

One of the main tools for dealing with unique continuation properties of PDE are the so called *Carleman inqualities*. It would be interesting to develop the corresponding discrete theory.

Now, returning to the wave equation in the square domain and its semi-discrete approximations, we see that, in view of the explicit construction of the eigenvector above, one can build solutions of the semi-discrete system in separated variables that vanish everywhere in the domain except on the diagonal for all time. This example shows that the controllability property of the semi-discrete system fails for many open subsets of the boundary. Consequently,

the $1D$ result in Theorem 5.1 showing that, whenever the wave equation is approximately controllable, its semi-discrete approximations are controllable as well and the convergence of controls is false in several space dimensions without further geometric restrictions on the support of the controls.

The same pathology is an obstacle for the approximate controllability of the semi-discrete approximations of other models like, for instance, the heat or the Schrödinger equations. It is interesting to note that this obstacle of lack of unique continuation does not arise in the context of the problem of homogenization we mentioned in the introduction. Although, in principle, the later is more difficult to deal with from a technical point of view it turns out that the problem of approximate controllability is well-behaved in that context in several space dimensions for parabolic equations too [84].

It would be interesting to analyze if a filtering mechanism allows reestablishing the uniformity of the approximate controllability property without imposing additional geometric restrictions on the supports of the controls.

Concerning the problem of decay of solutions of wave equations in the presence of damping discussed in the previous section we emphasize that the counterexample above to unique continuation allows showing that, at the semi-discrete level, in contrast with what happens in the continuous case, the decay of solutions may fail without further restrictions on the geometry of the subdomain ω where the damping is effective.

Acknowledgements

The author acknowledges the invitation, warm hospitality and support.

REFERENCES

[1] M. Asch – G. Lebeau: *Geometrical aspects of exact boundary controllability of the wave equation. A numerical study*, ESAIM:COCV, **3** (1998), 163-212.

[2] H. T. Banks – K. Ito – C. Wang: *Exponentially stable approximations of weakly damped wave equations*, International Series in Numerical Analysis, **100** (1991), 1-33.

[3] C. Bardos – F. Bourquin – G. Lebeau: *Calcul de dérivées normales et méthode de Galerkin appliquée au problème de contrôlabilité exacte*, C. R. Acad. Sci. Paris Sér. I Math., **313** (11) (1991), 757-760.

[4] C. Bardos – G. Lebeau – J. Rauch: *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Cont. Optim., **30** (1992), 1024-1065.

[5] T. Z. Boulmezaoud – J. M. Urquiza: *On the eigenvalues of the spectral second order differentiation operator: Application to the boundary observability of the wave operator*, preprint (2002).

[6] N. Burq: *Contrôle de l'équation des ondes dans des ouverts peu réguliers*, Asymptotic Analysis, **14** (1997), 157-191.

[7] N. Burq – P. Gérard: *Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes*, C. R. Acad. Sci. Paris, **325** (1997), 749-752.

[8] N. Burq – G. Lebeau: *Mesures de défaut de compacité, application au système de Lamé*, Ann. Sci. École Norm. Sup., **34** (6) (2001), 817-870.

[9] E. Casas: *Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints*, A tribute to J. L. Lions, ESAIM Control Optim. Calc. Var., **8** (2002), 345-374.

[10] E. Casas – J. P. Raymond: *Error estimates for the numerical approximation of Dirichlet boundary control for semilinear elliptic equations*, preprint (2003).

[11] C. Castro – S. Micu: *Boundary controllability of a semi-discrete wave equation with mixed finite elements*, preprint (2004).

[12] C. Castro – E. Zuazua: *Contrôle de l'équation des ondes à densité rapidement oscillante à une dimension d'espace*, C. R. Acad. Sci. Paris, **324** (1997), 1237-1242.

[13] C. Castro – E. Zuazua: *Localization of waves in* $1D$ *highly heterogeneous media*, Archive Rational Mechanics and Analysis, **164** (1) (2002), 39-72.

[14] C. Castro – E. Zuazua: *Some topics on the controls and homogenization of parabolic Partial Differential Equations*, Homogenization 2001, Proceeding of the First HMS2000 International School and Conference on Homogenization, L. Carbone, R. De Arcangelis (eds.), GAKUTO Internat. Ser. Math. Sci. Appl., **18** (2003), Gakkotosho, Tokyo, Naples, 45-94.

[15] I. Charpentier – Y. Maday: *Identification numérique de contrôles distr ibués pour l'equation des ondes*, C. R. Acad. Sci. Paris, **322** (8) (1996), 779-784.

[16] D. Chenais – E. Zuazua: *Controllability of an elliptic equation and its finite difference approximation by the shape of the domain*, Numersiche Mathematik, **95** (2003), 63-99.

[17] D. Chenais – E. Zuazua: *Finite Element Approximation on Elliptic Optimal Design*, C. R. Acad. Sci. Paris, **338** (2004), 729-734.

[18] G. Cohen: *Higher-order numerical methods for transient wave equations*, Scientific Computation, Springer, 2001.

[19] G. Dal Maso: *An introduction to* Γ*-convergence*, Progress in Nonlinear Differential Equations and their Applications, **8**, Birkhäuser Boston, Inc., Boston, MA, 1993.

[20] P. Donato – A. Nabil: *Approximate controllability of linear parabolic equations in perforated domains*, ESAIM Control Optim. Calc. Var., **6** (2001), 21-38.

[21] C. Fabre – J. P. Puel – E. Zuazua: *Approximate controllability for the semilinear heat equation*, Proc. Roy. Soc. Edinburgh, **125A** (1995), 31-61.

[22] H. O. Fattorini: *Infinite Dimensional Optimization and Control Theory*, Encyclopedia of Mathematics and its Applications, **62**, Cambridge University Press, 1999.

[23] L. A. Fernández – E. Zuazua: *Approximate controllability of the semilinear heat equation involving gradient terms*, J. Opt. Theory Appls., **101** (2) (1999), 307-328.

[24] E. Fernández-Cara – E. Zuazua: *The cost of approximate controllability for heat equations: The linear case*, Advances Diff. Eqs., **5** (4–6) (2000), 465-514.

[25] E. Fernández-Cara – E. Zuazua: *Null and approximate controllability for weakly blowing-up semilinear heat equations*, Annales Inst. Henri Poincaré, Analyse non-linéaire, **17** (5) (2000), 583-616.

[26] A. V. Fursikov – O. Yu. Imanuvilov: *Controllability of evolution equations*, Lecture Notes Series # 34, Research Institute of Mathematics, Global Analysis Research Center, Seoul National University, 1996.

[27] P. Gérard: *Microlocal defect measures*, Comm. P.D.E., **16** (1991), 1761-1794.

[28] P. Gervasio – M. G. Naso: *Numerical approximation of controllability of trajectories for Euler-Bernouilli thermoelastic plates*, Math. Models Meth. Applied Sciences, **14** (5) (2004), 701-734.

[29] R. Glowinski: *Ensuring well-posedness by analogy; Stokes problem and boundary control of the wave equation*, J. Compt. Phys., **103** (2) (1992), 189-221.

[30] R. Glowinski – W. Kinton – M. F. Wheeler: *A mixed finite element formulation for the boundary controllability of the wave equation*, Int. J. Numer. Methods Engineering, **27** (1989), 623-635.

[31] R. Glowinski – C. H. Li: *On the numerical implementation of the Hilbert uniqueness method for the exact boundary controllability of the wave equation*, C. R. Acad. Sci. Paris Sr. I Math., **311** (2) (1990), 135-142.

[32] R. Glowinski – C. H. Li – J.-L. Lions: *A numerical approach to the exact boundary controllability of the wave equation (I). Dirichlet controls: Description of the numerical methods*, Japan J. Appl. Math., **7** (1990), 1-76.

[33] R. Glowinski – J.-L. Lions: *Exact and approximate controllability for distributed parameter systems*, Acta numerica, Cambridge Univ. Press, Cambridge, (1994), 269-378.

[34] P. Grisvard: *Contrôlabilité exacte des solutions de l'équation des ondes en présence de singularités*, J. Math. Pures Appl., **68** (2) (1989), 215-259.

[35] M. D. Gunzburger – L. S. Hou – L. Ju: *A numerical method for exact boundary controllability problems for the wave equation*, preprint (2002).

[36] A. Haraux – S. Jaffard: *Pointwise and spectral controllability for plate vibrations*, Revista Matemática Iberoamericana, **7** (1) (1991), 1-24.

[37] L. F. Ho: *Observabilité frontière de l'équation des ondes*, C. R. Acad. Sci. Paris, **302** (1986), 443-446.

[38] L. Hörmander: *The analysis of linear partial differential operators*, III & IV, Springer-Verlag, Berlin, 1985.

[39] J. A. Infante – E. Zuazua: *Boundary observability of the space-discretizations of the* $1D$ *wave equation*, C. R. Acad. Sci. Paris, **326** (1998), 713-718.

[40] J. A. Infante – E. Zuazua: *Boundary observability for the space-discretizations of the one-dimensional wave equation*, Mathematical Modelling and Numerical Analysis, **33** (1999), 407-438.

[41] A. E. Ingham: *Some trigonometric inequalities with applications to the theory of series*, Math. Zeits., **41** (1936), 367-379.

[42] E. Isaacson – H. B. Keller: *Analysis of Numerical Methods*, John Wiley & Sons, 1966.

[43] O. Kavian: *Private communication*, (2001).

[44] G. Knowles: *Finite element approximation of parabolic time optimal control problems*, SIAM J. Cont. Optim., **20** (1982), 414-427.

[45] V. Komornik: *Exact Controllability and Stabilization. The Multiplier Method*, Wiley, Chichester, Masson, Paris, 1994.

[46] S. Krenk: *Dispersion-corrected explicit integration of the wave equation*, Computer Methods in Applied Mechanics and Engineering, **191** (2001), 975-987.

[47] G. Lebeau: *Contrôle analytique I: Estimations a priori*, Duke Math. J., **68** (1) (1992), 1-30.

[48] G. Lebeau: *The wave equation with oscillating density: observability at low frequency*, ESAIM: COCV, **5** (2000), 219-258.

[49] G. Lebeau – E. Zuazua: *Null controllability of a system of linear thermoelasticity*, Archives for Rational Mechanics and Analysis, **141** (4) (1998), 297-329.

[50] E. B. Lee – L. Markus: *Foundations of Optimal Control Theory*, The SIAM Series in Applied Mathematics, John Wiley & Sons, 1967.

[51] L. León – E. Zuazua: *Boundary controllability of the finite difference space semi-discretizations of the beam equation*, ESAIM:COCV, A Tribute to Jacques-Louis Lions, Tome 2 (2002), 827-862.

[52] W. S. Levine: *Control Systems and Applications*, CRC Press, 2000.

[53] J. L. Lions: *Contrôlabilité Exacte, Stabilisation et Perturbations de Systèmes Distribués. Tome 1. Contrôlabilité Exacte*, Masson, Paris, RMA 8, 1988.

[54] J. L. Lions: *Exact controllability, stabilizability and perturbations for distributed systems*, SIAM Rev., **30** (1988), 1-68.

[55] A. López – E. Zuazua: *Uniform null controllability for the one dimensional heat equation with rapidly oscillating periodic density*, Annales IHP, analyse non linéaire, **19** (5) (2002), 543-580.

[56] F. Macià: *Propagación y control de vibraciones en medios discretos y continuos*, PhD Thesis, Universidad Complutense de Madrid, 2002.

[57] F. Macià: *Wigner measures in the discrete setting: high-frequency analysis of sampling & reconstruction operators*, preprint (2003).

[58] S. Micu: *Uniform boundary controllability of a semi-discrete* $1D$ *wave equation*, Numerische Mathematik, **91** (4) (2002), 723-768.

[59] S. Micu – E. Zuazua: *An Introduction to the Controllability of partial Differential Equations*, in: *Quelques questions de théorie du contrôle*, T. Sari, (eds.), Collection Travaux en Cours Hermann (2004), to appear.

[60] A. Munch: *Famille de schémas implicites uniformémemnt contrôlables pour l'équation des ondes* $1D$, peprint (2004).

[61] A. Munch – A. Pazoto: *Uniform stabilization and numerical analysis of a locally damped wave equation*, preprint (2004).

[62] M. Negreanu – E. Zuazua: *Uniform boundary controllability of a discrete* $1D$ *wave equation*, Systems and Control Letters, **48** (3-4) (2003), 261-280.

[63] M. Negreanu – E. Zuazua: *Discrete Ingham inequalities and applications*, C. R. Acad. Sci. Paris, **338** (2004), 281-286.

[64] M. Negreanu – E. Zuazua: *Convergence of a multigrid method for the controllability of a* $1D$ *wave equation*, C. R. Acad. Sci. Paris, **338** (4) (2004), 413-418.

[65] K. Ramdani – T. Takahashi – M. Tucsnack: *Uniformly exponentially stable approximations for a class of second order evolution equations*, Prépublication de l'Institute Elie Cartan de Nancy, 27/2003.

[66] L. Robbiano: *Fonction de coût et contrôle des solutions des équations hyperboliques*, Asymptotic Anal., **10** (2) (1995), 95-115.

[67] J. Rodellar et al.: *Advances in Structural Control*, CIMNE, Barcelona, 1999.

[68] T. Roubicek: *A stable approximation of a constrained optimal control problem for continuos casting*, Num. Funct. Anal. and Optim., **13** (1992), 487-494.

[69] D. L. Russell: *Controllability and stabilizability theory for linear partial differential equations. Recent progress and open questions*, SIAM Rev., **20** (1978), 639-739.

[70] D. L. Russell: *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., **52** (1973), 189-221.

[71] E. D. Sontag: *Mathematical Control Theory. Deterministic Finite-Dimensional Systems*, 2nd edition, Texts in Applied Mathematics, **6**, Springer-Verlag, New York, 1998.

[72] L. R. Tcheougoué Tebou – E. Zuazua: *Uniform exponential long time decay for the space semi-discretization of a locally damped wave equation via an artificial numerical viscosity*, Numerische Mathematik, **95** (2003), 563-598.

[73] L. R. Tcheougoué Tebou – E. Zuazua: *Uniform boundary stabilization of the finite difference space discretization of the* $1D$ *wave equation*, Advances in Computational Mathematics (2004), to appear.

[74] L. N. Trefethen: *Group velocity in finite difference schemes*, SIAM Rev., **24** (2) (1982), 113-136.

[75] F. Tröltz: *Semidiscrete Ritz-Galerkin approximation of nonlinear parabolic boundary control problems-strong convergence of optimal controls*, Appl. Math. Optim., **29** (1994), 309-329.

[76] R. Vichnevetsky – J. B. Bowles: *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*, SIAM Studies in Applied Mathematics, **5**, SIAM, Philadelphia, 1982.

[77] R. M. Young: *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

[78] J. Zabczyz: *Mathematical control theory: an introduction*, Systems & Control: Foundation & Applications, Birkhäuser Boston, Inc., Boston, MA, 1992.

[79] X. Zhang: *Explicit observability estimate for the wave equation with potential and its application*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., **456** (2000), 1101-1115.

[80] E. Zuazua: *Exponential decay for the semilinear wave equation with localized damping*, Communications in PDE, **15** (2) (1990), 205-235.

[81] E. Zuazua: *Some problems and results on the controllability of partial differential equations*, Progress in Mathematics, **169** (1998), Birkhäuser Verlag, 276-311.

[82] E. Zuazua: *Boundary observability for the finite-difference space semi-discretizations of the* $2D$ *wave equation in the square*, J. Math. Pures et Appliquées, **78** (1999), 523-563.

[83] E. Zuazua: *Controllability of Partial Differential Equations and its Semi-Discrete Approximations*, Discrete and Continuous Dynamical Systems, **8** (2) (2002), 469-513.

[84] E. Zuazua: *Approximate controllability for linear parabolic equations with rapidly oscillating coefficients*, Control and Cybernetics, **23** (4) (1994), 1-8.

[85] E. Zuazua: *Observability of* $1D$ *waves in heterogeneous and semi-discrete media*, Advances in Structural Control, J. Rodellar *et al.* (eds.), CIMNE, Barcelona, 1999, 1-30.

[86] E. Zuazua: *Exact controllability for the semilinear wave equation in one space dimension*, Ann. IHP. Analyse non linéaire, **10** (1993), 109-129.

[87] E. Zuazua: *Propagation, observation, and control of waves approximated by finite difference methods*, (2004), http://www.uam.es/personal_pdi/ciencias/ezuazua/enrique_pub.html.

Indirizzo dell'Autore:

Enrique Zuazua – Departamento de Matemáticas – Universidad Autónoma – 28049 Madrid, Spain
E-mail: enrique.zuazua@uam.es.

Supported by grant BFM2002-03345 of the MCYT, Spain and the Networks "Homogenization and Multiple Scales" and "New materials, adaptive systems and their nonlinearities: modelling, control and numerical simulation (HPRN-CT-2002-00284)" of the EU. This article is originated at the "Lezioni Guido Castelnuovo" delivered by the author at the Dipartimento di Matematica, Università di Roma "La Sapienza" in November, 2003.

Rendiconti di Matematica, Serie VII
Volume 24, Roma (2004), 239-260

A method for global approximation of the solution of second order IVPs

F. COSTABILE – A. NAPOLI

ABSTRACT: *For the numerical solution of the second order initial value problem, a family of global methods is derived by finding Galerkin approximations on a given interval. For each $n \geq 1$ a method is defined that uses a particular inner product in the Galerkin equation. The methods are symmetric collocation on the zeros of Chebyshev polynomials of the second kind and are related to implicit Runge-Kutta-Nyström methods. Order, stability and error analysis are here studied. Numerical examples provide favorable comparisons with other existing methods.*

1 – Introduction

In this paper we will consider the initial value problem in ordinary differential equations

$$(1) \qquad \begin{cases} y''(x) = f(x, y(x)) & x \in [x_0, b] \\ y(x_0) = y_0 \\ y'(x_0) = y'_0 \,. \end{cases}$$

We suppose that $f(x, y(x))$ is a real function defined and continuous on the strip $S = [x_0, b] \times \mathbb{R}$ and a constant L exists so that the inequality

$$|f(x, y_1) - f(x, y_2)| \leq L |y_1 - y_2|$$

KEY WORDS AND PHRASES: *Initial value problem – Chebyshev series.*
A.M.S. CLASSIFICATION: 65L05 – 65L60

holds over the strip S. Under these hypotheses the problem (1) has a unique solution $y(x)$.

Problems of this kind arise in a variety of physical contexts such as molecular-dynamics calculations for liquid and gases, stellar mechanics, atomic and nuclear scattering problems.

We assume that (1) represents a single scalar equation, but nearly all of the numerical and theoretical considerations in this paper carry over systems of second order equations, where (1) could be treated in vector form.

For higher order differential equations, one may solve them numerically by first reducing them to systems of first order equations. However, for an equation of the form (1), it is simpler to attack it directly and it is well known that several advantages (substantial gain in efficiency, lower storage requirements, etc.) are realized when the equations are treated in their original second-order form.

We develop a family of direct methods which produce smooth, global approximations to $y(x)$ in the form of polynomial functions. The basic idea is to approximate $y''(x)$ on $[-1,1]$ by a linear combination of Chebyshev polynomials of second kind and then to require that it provides Galerkin approximation (Section 2).

In Section 3 we show that these methods (GCM) are also collocation methods; we study the global approximation error in Section 4; then we propose two algorithms to compute the numerical solution of (1) in the nodal points and present some numerical examples in comparison with Dormand-Prince methods.

In Section 6 we illustrate the corresponding implicit Runge-Kutta-Nyström form, we observe that these methods have even order and we compare them with other Runge-Kutta-Nyström methods.

In Section 7 the study of stability of the method for $n = 3$ shows that it compares quite favorably with other fourth-order methods.

Finally (Section 8), we show that GCM may be formulated as symmetric hybrid two-step methods.

2 – Some polynomial Galerkin-type methods

We may approximate y'' on $[-1,1]$ by an $(n-1)$-th degree polynomial

$$y_n''(x) = \sum_{k=1}^{n} c_{k-1} U_{k-1}(x) , \qquad x \in [-1,1] \tag{2}$$

where $U_k(x)$ is the k-th degree Chebyshev polynomial of second kind, which satisfies:

$$U_{k-1}(x) = \frac{\sin kt}{\sin t} \tag{3}$$

with $x = \cos t$.

The coefficients c_k, $k = 0, \ldots, n-1$, are determined by a polynomial Galerkin-type method: we require the residual function $y_n'' - f$ to be orthogonal to all polynomials U_{j-1}, that is what this relation holds:

$$(y_n'' - f(x, y_n), U_{j-1}) = 0$$

$j = 1, \ldots, n$, or equivalently

$$(y_n'', U_{j-1}) = (f(x, y_n), U_{j-1}) \tag{4}$$

$j = 1, \ldots, n$.
By defining the discrete inner product [10]

$$(u, v) = \sum_{i=1}^{n} u\left(\pi - \frac{\pi i}{n+1}\right) v\left(\pi - \frac{\pi i}{n+1}\right),$$

we have

$$\left(\sin^2 t\, U_{k-1}, U_{j-1}\right) = \sum_{i=1}^{n} \sin\frac{k\pi i}{n+1} \sin\frac{j\pi i}{n+1} = \begin{cases} 0 & j \neq k \\ \dfrac{n+1}{2} & j = k \end{cases}$$

and

$$\sin^2 t\,(y_n'', U_{j-1}) = \frac{n+1}{2} c_{j-1}\,. \tag{5}$$

By multiplying the right term of (4) by $\sin^2 t$, it becomes:

$$\left(\sin^2 t\, f(x, y_n), U_{j-1}\right) = \sum_{i=1}^{n} f(x_i, y_n(x_i)) \sin\frac{\pi i}{n+1} \sin\frac{j\pi(n+1-i)}{n+1} \tag{6}$$

with

$$x_i = \cos\left(\pi - \frac{\pi i}{n+1}\right) = -\cos\frac{\pi i}{n+1} \qquad i = 1, \ldots, n\,. \tag{7}$$

By equaling (5) and (6) we have:

$$c_{j-1} = \frac{2}{n+1} \sum_{i=1}^{n} \sin\frac{\pi i}{n+1} \sin\frac{j\pi(n+1-i)}{n+1} f(x_i, y_n(x_i))\,.$$

Using the identity

$$kU_{k-1}(x) = T_k'(x)\,, \qquad k \geq 1\,, \tag{8}$$

and integrating (2) between -1 and x we have

$$y'_n(x) = y'_0 + \sum_{i=1}^{n} {}^n\gamma_i(x)\, f(x_i, y_n(x_i)) \tag{9}$$

where

$${}^n\gamma_i(t) = \frac{2}{n+1}\sin\frac{\pi i}{n+1}\sum_{k=1}^{n}\frac{p_k(t)}{k}\sin\frac{\pi k(n+1-i)}{n+1}$$

with

$$p_k(x) = T_k(x) - (-1)^k$$

and $T_k(x)$ are the Chebyshev polynomials of first kind of degree k.

If $f(x, y(x))$ does not depend on $y(x)$, the (9)

$$\int_{-1}^{x} f(t)dt = \sum_{i=1}^{n} {}^n\gamma_i(x)\, f(x_i) \tag{10}$$

coincides with the modified Filippi Clenshaw-Curtis quadrature formula [10]. Hence, for $x \in [-1, 1]$, (10) is a positive quadrature procedure which converges for every $f \in C^0[-1, 1]$.

Integration of (9) gives

$$y_n(x) = y_0 + (x+1)\, y'_0 + \sum_{i=1}^{n} {}^n\beta_i(x)\, f(x_i, y_n(x_i)) \tag{11}$$

where

$${}^n\beta_i(x) = \frac{1}{n+1}\sin\frac{\pi i}{n+1}\left\{\sin\frac{\pi i}{n+1}(x+1)^2 + \right.$$
$$\left. + \sum_{k=2}^{n}\frac{1}{k}\sin\frac{k\pi(n+1-i)}{n+1}\left[\frac{T_{k+1}(x)}{k+1} - \frac{T_{k-1}(x)}{k-1} - 2\left(x + \frac{k^2}{k^2-1}\right)(-1)^k\right]\right\}.$$

Thus we obtain:

$$\begin{cases} y(x) \approx y_n(x) = y_0 + (x+1)y'_0 + \sum_{i=1}^{n} {}^n\beta_i(x) f(x_i, y_n(x_i)) \\ \\ y'(x) \approx y'_n(x) = y(x_0) = y'_0 + \sum_{i=1}^{n} {}^n\gamma_i(x) f(x_i, y_n(x_i)). \end{cases}$$

3 – Chebyshev-Galerkin methods as collocation methods

THEOREM 1. *Let us consider the initial value problem* (1) *with* $x_0 = -1$, $[x_0, b] = [-1, 1]$. *If* x_i, $i = 1, \ldots, n$ *are defined as* (7), *then the polynomial* (11) *of degree* $n+1$ *satisfies the relations*

$$\begin{aligned} & y_n(-1) = y_0 \\ & y_n'(-1) = y_0' \\ & y_n''(x_j) = f(x_j, y_n(x_j)), \quad j = 1, \ldots, n \end{aligned} \tag{12}$$

i.e. it is a collocation polynomial for (1) [11].

PROOF.

$$ {}^n\beta_i(x) = \int_{-1}^{x} {}^n\gamma_i(t)\, dt. $$

Hence, $\forall i, n \in \mathbb{N}$, we have

$$ {}^n\beta_i(-1) = 0 $$

and

$$ {}^n\beta_i'(x) = {}^n\gamma_i(x) \quad \Longrightarrow \quad {}^n\beta_i'(-1) = 0. $$

It follows that

$$ y_n(-1) = y_0\,; $$

$$ y_n'(-1) = y_0' + \sum_{i=1}^{n} {}^n\beta_i'(-1)\, y_n''(x_i) = y_0'. $$

Moreover, for the polynomial ${}^n\gamma_i(x)$, we get

$$ {}^n\gamma_i'(x) = \frac{2}{n+1} \sin\frac{\pi i}{n+1} \sum_{k=1}^{n} \frac{T_k'(x)}{k} \sin\frac{\pi k(n+1-i)}{n+1}. $$

Putting $x = \cos t$, we have

$$ {}^n\gamma_i'(\cos t) = \frac{2}{n+1} \sin\frac{\pi i}{n+1} \sum_{k=1}^{n} \frac{\sin kt}{\sin t} \sin\frac{k\pi(n+1-i)}{n+1} $$

and for $t = t_j = \arccos x_j$, from the orthogonality of the system of functions $\sin mt$ ([10]), it follows that

$$ {}^n\gamma_i'(x_j) = {}^n\gamma_i'(\cos t_j) = \delta_{ij}. $$

for $j = 1, \ldots, n$. Hence we have

$$ y_n''(x_j) = \sum_{i=1}^{n} {}^n\beta_i''(x_j)\, f(x_i, y_n(x_i)) = f(x_i, y_n(x_i)). $$

□

From Theorem 1 we give the following definition

DEFINITION 1. The polynomial (11) is a global approximation method for the solution of (1) in $[-1,1]$. It is a symmetric collocation method.

OBSERVATION. Now an alternative representation for the ${}^n\beta_i(x), i=1,\dots,n$ can be derived by observing from (12) that, since $y_n''(x)$ interpolates $f\left(x,y(x)\right)$ at x_i, using Lagrangian interpolation we have

$$w_n''(x)=\sum_{k=1}^{n} l_k(t) f\left(x_k, y_n\left(x_k\right)\right). \tag{13}$$

After two integrations and comparing with (11), we obtain

$${}^n\beta_i(x)=\int_{-1}^{x}\left(\int_{-1}^{s} l_i(t)dt\right)ds=\int_{-1}^{x}(x-t)l_i(t)dt \tag{14}$$

where $l_i(t)$ is a polynomial of Lagrange interpolation on the set of points $\{x_i\}$:

$$l_i(t)=\prod_{k=1\ k\neq i}^{n}\frac{t-x_k}{x_i-x_k}.$$

4 – Global error

For the global error

$$L_n\left(y,x\right)=y\left(x\right)-y_n\left(x\right).$$

the following theorem holds:

THEOREM 2. *For all fixed* $x\in[-1,1]$

$$y(x)-y_n(x)=\frac{1}{n!}\left[\int_{-1}^{x}(x-t)^{n+1}y^{(n+2)}(t)dt+\right.$$
$$\left.-n\sum_{k=1}^{n}{}^n\beta_k(x)\int_{x_k}^{x}(x_k-t)^{n+1}y^{(n+2)}(t)dt\right].$$

PROOF. We observe that for all fixed $x \in [-1,1]$

$$L_n(y,x) = \int_{-1}^{x} \left(\int_{-1}^{s} y''(l)dl \right) ds - \sum_{k=1}^{n} {}^n\beta_k(x)\, y''(x_k)$$

is a linear functional vanishing if $y(t)$ is a polynomial of degree less than or equal to $n+1$. In fact, if $y(t)$ is a polynomial, it can be written in the Lagrange form:

$$y_n''(x) = \sum_{k=1}^{n} l_k(t) y_n''(x_k),$$

and from (14)

$$L_n(y,x) = \int_{-1}^{x} \left(\int_{-1}^{s} y''(l)dl \right) ds - \sum_{k=1}^{n} \int_{-1}^{x} \left(\int_{-1}^{s} l_k(t)dt \right) ds\, y''(x_k) = 0.$$

Hence from Peano's Lemma [8],

$$y(x) - y_n(x) = \int_{-1}^{x} K(t,x)\, y^{(n+2)}(t)\, dt$$

where

$$K(t,x) = \frac{1}{(n-1)!} \left[\int_{-1}^{x} \left(\int_{-1}^{s} (l-t)_+^{n-1} dl \right) ds - \sum_{k=1}^{n} {}^n\beta_k(x)\, (x_k - t)_+^{n-1} \right].$$

The thesis follows after some calculations.

OBSERVATION. If $y^{(n+2)}(t)$ is continuous in $[-1,1]$, then there exist η_0, η_k, $k = 1, \ldots, n$ in $[-1,1]$ such that

$$y(x) - y_n(x) = \frac{1}{n!(n+2)} \Bigg[(x-1)^{n+2} y^{(n+2)}(\eta_0) + \\ + n \sum_{k=1}^{n} {}^n\beta_k(x)(x_k - x)^{n+2} y^{(n+2)}(\eta_k) \Bigg].$$

5 – Algorithms and numerical examples

In order to calculate the approximate solution of the initial value problem by (11) at $x \in [-1,1]$ we need the values $y_n(x_i)$, $i = 1,\dots,n$. For this aim we propose two algorithms:

A1. Solve the system

$$y_n(x_i) = y_0 + (x_i+1)y_0' + \sum_{k=1}^{n} {}^n\beta_k(x_i)\, f(x_k, y_n(x_k)) \quad i = 1,...,n.$$

by iterative methods, particularly a modified Newton-type method for general non linear case. For linear problems the computational cost is considerably lower.

A2. An alternative way to calculate $y_n(x_j)$ is the iterative algorithm

$$\begin{cases} G_{n,j}^0 = y_0 + (x_j+1)\, y_0' + (x_j+1)^2 f(-1, y_0)/2 \\ G_{n,j}^\nu = y_0 + (x_j+1)y_0' + \sum_{k=1}^{n} a_{kj} f\left(x_k, G_{n,k}^{\nu-1}\right) \quad \nu = 1,2,... \end{cases}$$

$j = 1,\dots,n$, $a_{kj} = {}^n\beta_k(x_j)$ and $G_{n,j}^\nu = G_{n,j}^\nu(x_j)$ where $G_{n,j}^0$ are the first three terms of Taylor approximation of $y_n(x_j)$ used to initialize the iterations.

We apply A1 to find numerical approximations of the solutions of some test problems. Similar results are obtained by algorithm A2.

Results are compared with the ones obtained by applying the Matlab ODE solver based on Dormand-Prince formula. We consider the following problems:

$$i) \quad \begin{cases} y'' = y + 2e^x \\ y(-1) = 0 \\ y'(-1) = \dfrac{1}{e} \end{cases}$$

with solution $y(x) = (x+1)e^x$

$$ii) \quad \begin{cases} y'' = -y + 2\cos x \\ y(-1) = -\sin(-1) \\ y'(-1) = \sin(-1) - \cos(-1) \end{cases}$$

with solution $y(x) = x\sin(x)$.

The figures (Fig. 1) and (Fig. 2) present the error function in the interval $[-1,1]$ in the case of Dormand-Prince approximation (dotted line) and in the case of approximation by GCM (solid line), algorithm A1.

In the first case (ode45) 85 function evaluations are needed for problem 1 and 67 for problem 2, while A1 requires 32 function evaluations if we use a modified

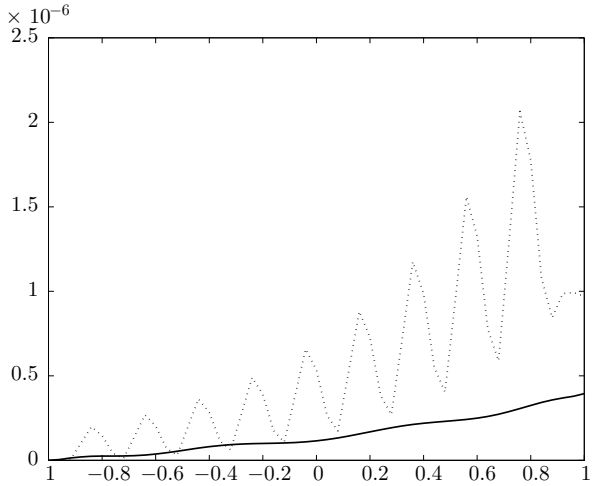

Fig. 1: Problem *i*.

Newton-type method and only 16 evaluations of functions of one variable if we use a direct method.

We can observe the smoothness of the error function in the case of approximation by A1.

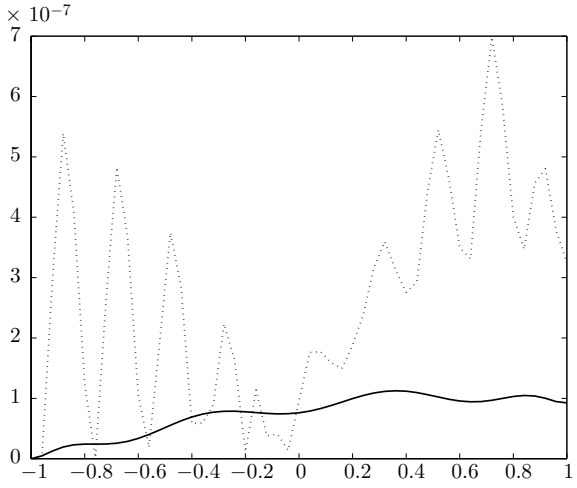

Fig. 2: Problem *ii*.

Moreover, with no additional cost, we have the approximation of the first derivative (Fig. 3 and 4).

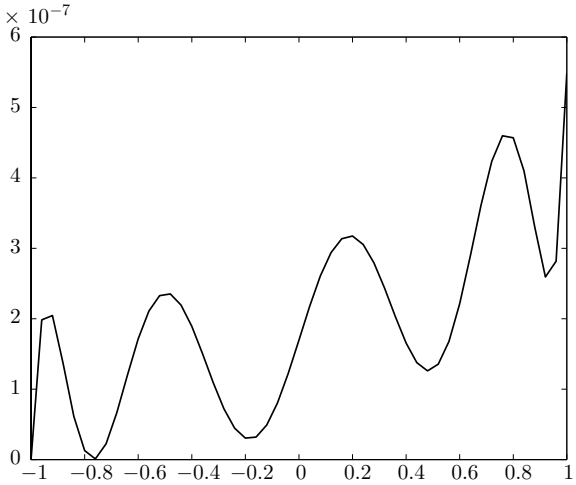

Fig. 3: Error function $|y'(x) - y'_n(x)|$ of problem i.

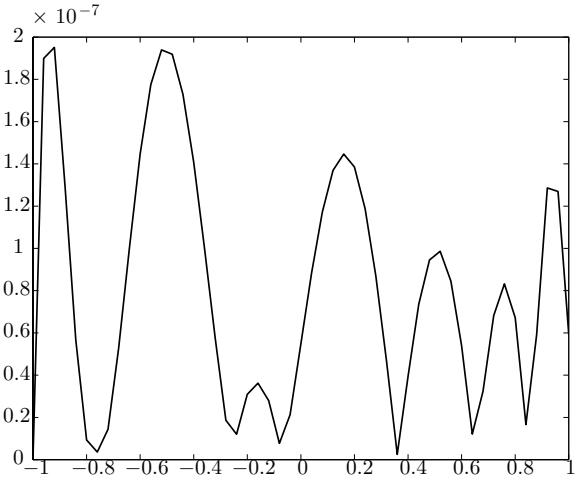

Fig. 4: Error function $|y'(x) - y'_n(x)|$ of problem ii.

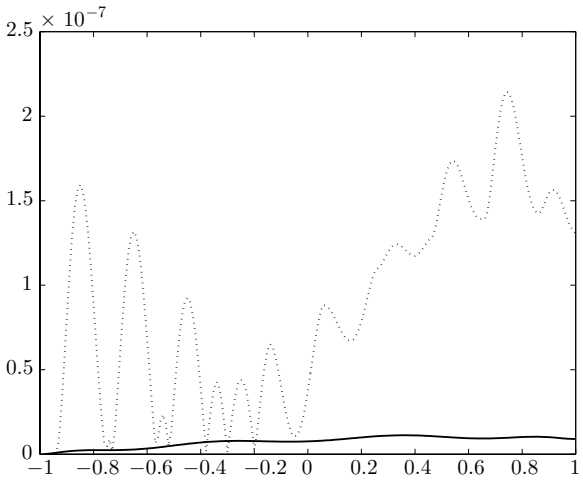

Fig. 5: Problem *iii*.

Now we consider the following non-linear problem:

$$iii) \quad \begin{cases} y'' = -\left(1 + 0.01y^2\right) y + 0.01\cos^3 x \\ y\,(-1) = \cos(-1) \\ y'(-1) = -\sin(-1). \end{cases}$$

The differential non-linear equation is a particular case of the undamped Duffing's equation, with a forcing term chosen so that the exact solution is $y\,(x) = \cos x$. Figure 5 show the error function in $[-1, 1]$ when we use ode45 (dotted line) and when we approximate by algorithm A1 (solid line).
The approximation by ode45 requires 67 function evaluations, algorithm A1 64.
Figure 6 presents the approximation of the first derivative using algorithm A1.

6 – Chebyshev-Galerkin methods as implicit Runge-Kutta-Nyström methods

Any one-step collocation method is equivalent to some implicit Runge-Kutta methods, where of course "equivalent" here means "matches the discrete values". Let $\chi : t_k = t_0 + kh$ be a uniform mesh with $t_0 = x_0$. On each subinterval we apply GCM (11), so that we have a collocation method on the points $t_{k+c_j} = t_k + c_j h$, $j = 1, \ldots, n$, with $c_j = \frac{1}{2}(x_j + 1)$, which are the images of the x_j under

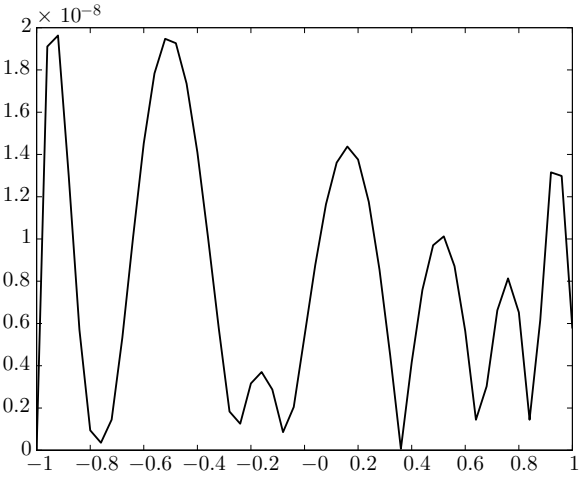

Fig. 6: Error function $|y'(x) - y'_n(x)|$ of problem *iii*.

a linear transform mapping $[-1, 1]$ onto $[0, 1]$:

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + h^2 \sum_{j=1}^{n} b_j f\left(t_{i+c_j}, y\left(t_{i+c_j}\right)\right)$$

where $b_j = \frac{1}{4}\beta_j(1)$. Putting

$$k_j = f\left(t_i + c_j h, y_i + c_j h y'_i + h^2 \sum_{m=1}^{n} b_{jm} k_m\right), \tag{15}$$

with $b_{jm} = \frac{1}{4}\beta_m(x_j)$,we have:

$$y_{i+1} = y_i + hy'_i + h^2 \sum_{j=1}^{n} b_j k_j \tag{16}$$

and from (9)

$$y'_{i+1} = y'_i + h \sum_{j=1}^{n} a_j k_j \tag{17}$$

with $a_j = \frac{1}{2}\gamma_j(1)$. Equations (15), (16) and (17) gives rise to an n-stage implicit Runge-Kutta-Nyström method (CRK) with

$$\sum_{j=1}^{n} b_j = \frac{1}{2} \qquad \sum_{j=1}^{n} a_j = 1.$$

Using Butcher's notation ([1]), the first three of these methods are presented in Tables 1,2 and 3.

Table 1.

$\frac{1}{2}$	$\frac{1}{4}$
	$\frac{1}{2}$
	1

Table 2.

$\frac{1}{4}$	$\frac{1}{24}$	$-\frac{1}{96}$
$\frac{3}{4}$	$\frac{9}{32}$	0
	$\frac{5}{12}$	$\frac{1}{12}$
	$\frac{1}{2}$	$\frac{1}{2}$

Table 3.

$\frac{2-\sqrt{2}}{4}$	$\frac{1}{64}$	$\frac{5-4\sqrt{2}}{96}$	$\frac{23-16\sqrt{2}}{192}$
$\frac{1}{2}$	$\frac{3+2\sqrt{2}}{48}$	0	$\frac{3-2\sqrt{2}}{48}$
$\frac{2+\sqrt{2}}{4}$	$\frac{23+16\sqrt{2}}{192}$	$\frac{5+4\sqrt{2}}{96}$	$\frac{1}{64}$
	$\frac{2+\sqrt{2}}{12}$	$\frac{1}{6}$	$\frac{2-\sqrt{2}}{12}$
	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Note that the c_j have the symmetry property

$$c_{n+1-j} = 1 - c_j, \qquad j = 1, \ldots, \left[\frac{n}{2}\right]. \tag{18}$$

A Runge-Kutta-Nyström method has order p [11] if for sufficiently smooth problems (1)

$$y(x_{i+1}) - y_{i+1} = O\left(h^{p+1}\right) \qquad y'(x_{i+1}) - y'_{i+1} = O\left(h^{p+1}\right).$$

Being the method (11) a collocation method, $\forall x \in [-1, 1]$ the following estimates hold [11]:

$$y(x) - y_n(x) = O\left(h^{n+2}\right), \qquad y'(x) - y'_n(x) = O\left(h^{n+1}\right).$$

So the method (11) has order at least n. We may prove that for odd n the order is $n + 1$. In fact, putting

$$M(t) = \prod_{i=1}^{n} (t - c_i),$$

if $n = 2k + 1$, we have

$$\int_0^1 M(t)dt = 0 \tag{19}$$

and this condition is equivalent to orthogonality to polynomials of degree $q = 0$, so the method has order [11] $p = n + q + 1 = n + 1$.

The (19) follows from the fact that $M(t) = (-1)^n M(-t)$, hence, if n is odd, $M(t)$ is an odd function.

Coleman and Booth in [5], starting from Panovsky-Richardson method [13], derived a class of Runge-Kutta-Nyström methods for differential equations of the form $y''(x) = f(x, y)$ which uses in each interval $[x_k, x_{k+1}]$ the set of $n + 1$ collocation points $\{x_k + c_j h, j = 0, \ldots, n\}$ where $c_j = \frac{1}{2}(x_j + 1)$ and x_j are the extrema of Chebyshev polynomials of first kind $T_k(x)$ of degree k.

In this context CRK method can be compared with other similar methods, among which the Coleman and Booth Runge-Kutta-Nyström one [5], which we indicate by CBRKN.

Thus we make a comparison between the forth order CBRKN, and the CRK method of order four derived in this section.

6.1 – Harmonic oscillator

Let's now solve the initial value problem

$$y'' = -y, \qquad y(0) = 1, \qquad y'(0) = 0 \tag{20}$$

using the forth order CBRKN and CRK methods. The results in Figures 7 and 8, produced by a MatLab code, show the absolute errors for the two methods (dotted line for the CBRKN) applied to problem (20) with steplengths respectively $h = 0.01$ and $h = 0.05$. Both methods have the same cost and are based upon the zeros of Chebyshev polynomials but of different degrees so their coefficients are different.

The maximum absolute errors on intervals $[0, x]$ with steplength $h = 0.01$ are displayed in Table 4.

Table 4.

x	CBRKN	CRK
1	$4.4 \cdot 10^{-12}$	$1.1 \cdot 10^{-12}$
2	$9.5 \cdot 10^{-12}$	$2.4 \cdot 10^{-12}$
5	$2.5 \cdot 10^{-11}$	$6.2 \cdot 10^{-12}$
10	$4.1 \cdot 10^{-11}$	$7.1 \cdot 10^{-12}$
20	$9.5 \cdot 10^{-11}$	$2.4 \cdot 10^{-11}$
50	$2.5 \cdot 10^{-10}$	$1.7 \cdot 10^{-11}$
100	$5.1 \cdot 10^{-10}$	$6.6 \cdot 10^{-11}$

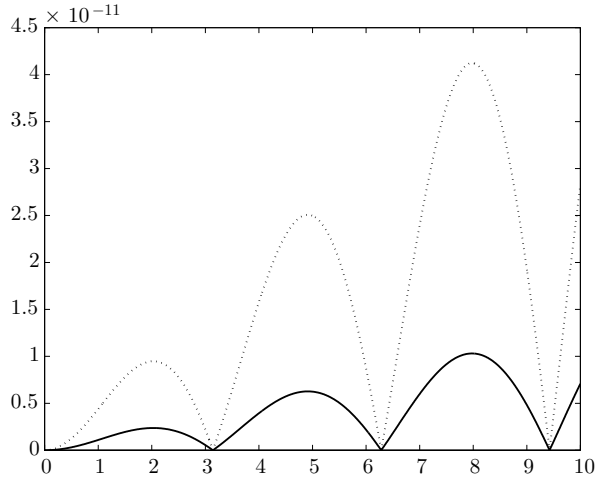

Fig. 7:

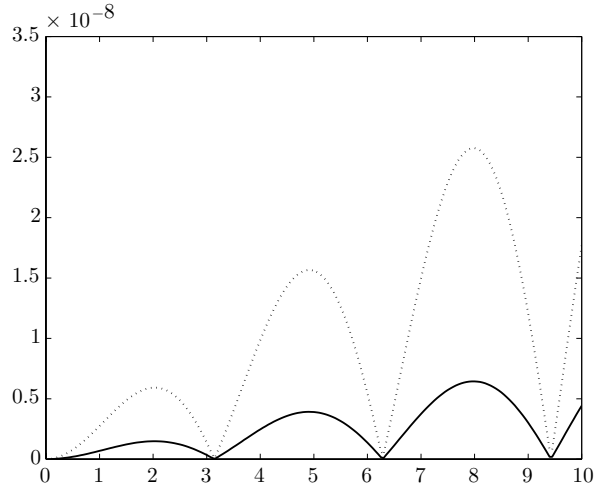

Fig. 8:

Table 5 illustrates the effects of different steplengths used over a given number of steps.

Table 5.

h	steps	CBRKN	CRK
0.1	500	$2.5 \cdot 10^{-6}$	$1.7 \cdot 10^{-7}$
0.1	1000	$5.1 \cdot 10^{-6}$	$6.6 \cdot 10^{-7}$
0.005	500	$5.9 \cdot 10^{-13}$	$1.2 \cdot 10^{-13}$
0.005	1000	$1.6 \cdot 10^{-12}$	$3.9 \cdot 10^{-13}$
0.002	500	$8.1 \cdot 10^{-15}$	$1.5 \cdot 10^{-15}$
0.002	1000	$1.5 \cdot 10^{-14}$	$3.5 \cdot 10^{-15}$
0.001	500	$1.9 \cdot 10^{-15}$	$2.0 \cdot 10^{-15}$
0.001	1000	$3.1 \cdot 10^{-15}$	$2.8 \cdot 10^{-15}$

Table 6 shows the maximum absolute errors on intervals $[0, x]$ for the same methods of order six applied to problem (20) with steplength $h = 0.01$:

Table 6.

x	CBRKN	CRK
1	$5.6 \cdot 10^{-16}$	$3.3 \cdot 10^{-16}$
2	$9.4 \cdot 10^{-16}$	$1.1 \cdot 10^{-16}$
5	$1.1 \cdot 10^{-15}$	$1.1 \cdot 10^{-16}$
10	$2.7 \cdot 10^{-15}$	$2.4 \cdot 10^{-15}$
20	$4.4 \cdot 10^{-15}$	$1.2 \cdot 10^{-15}$
50	$7.9 \cdot 10^{-15}$	$4.2 \cdot 10^{-15}$
70	$1.1 \cdot 10^{-14}$	$7.8 \cdot 10^{-16}$
100	$1.2 \cdot 10^{-14}$	$3.9 \cdot 10^{-14}$

Values of column 2 are the ones which appear in [5].
Figure 9 illustrates absolute errors when $h = 0.05$, in the case of order six.

6.2 – Two-body problem

A non-linear example frequently used to test numerical methods (see, e.g. [5]) is provided by the two-body problem:

$$\text{(21)} \qquad \begin{cases} y'' + \dfrac{y}{r^3} = 0, \quad y(0) = 1 - e, & y'(0) = 0 \\ z'' + \dfrac{z}{r^3} = 0, \quad z(0) = 0, & z'(0) = \sqrt{\dfrac{1+e}{1-e}} \end{cases}$$

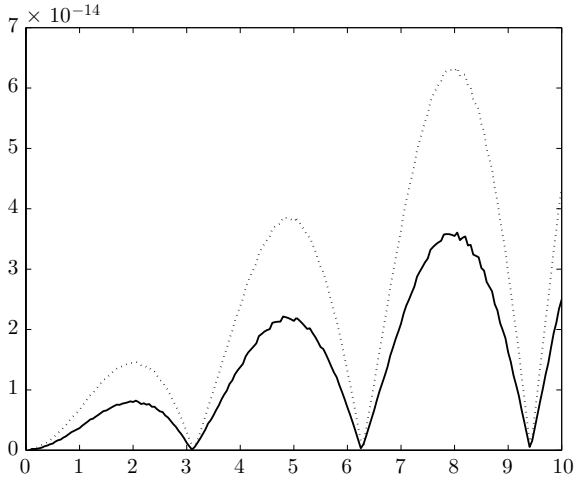

Fig. 9:

with $r^2 = y^2 + z^2$. The exact solution is

$$y = \cos E - e, \qquad z = \sqrt{1 - e^2}\sin E,$$

where e is the eccentricity of the orbit and E is implicitly defined as $x = E - e\sin E$.

In Table 7 we compare the maximum absolute errors on $[0, x]$ for the two fourth order methods, CRK and CBRKN, applied to (21) when $e = 0.1$ and steplength $h = 0.01$.

Table 7.

x	CBRKN	CRK
1	$2.9 \cdot 10^{-11}$	$7.4 \cdot 10^{-12}$
2	$4.0 \cdot 10^{-11}$	$9.9 \cdot 10^{-12}$
5	$1.2 \cdot 10^{-10}$	$2.9 \cdot 10^{-11}$
10	$2.8 \cdot 10^{-10}$	$7.1 \cdot 10^{-11}$
20	$8.4 \cdot 10^{-10}$	$2.1 \cdot 10^{-10}$
50	$2.1 \cdot 10^{-9}$	$5.4 \cdot 10^{-10}$
100	$4.2 \cdot 10^{-9}$	$1.0 \cdot 10^{-9}$

These results were produced by MatLab programs on a microcomputer and show that CRK method is favourably comparable to CBRKN one.

7 – Stability and periodicity

Now we investigate the numerical stability of method (11) for $n = 3$. Towards this aim we consider its equivalent implicit Runke-Kutta-Nyström form (CRK) and then we compare results with other forth-order methods.

We apply CRK to the test equation

$$y'' = -\alpha y$$

where α is a real number, and after some calculation we obtain:

$$(22)\qquad \begin{cases} y_{i+1} = \dfrac{-B}{A}\left(36864 + 17280h^2\alpha + 968h^4\alpha^2 + 11h^6\alpha^3\right) y_i + \\ \qquad\qquad -\dfrac{hB}{A}\left(36864 + 4992h^2\alpha + 136h^4\alpha^2 + h^6\alpha^3\right) y_i' \\ y_{i+1}' = \dfrac{-h\alpha B}{A}\left(36864 + 4992h^2\alpha + 120h^4\alpha^2\ \right) y_i + \\ \qquad\qquad -\dfrac{B}{A}\left(36864 + 17280h^2\alpha + 968h^4\alpha^2 + 11h^6\alpha^3\right) y_i' \end{cases}$$

where

$$\begin{aligned} A = {} & (17 + 12\sqrt{2})[4608 - 72h^2\alpha + (1 + 2\sqrt{2})h^4\alpha^2]\cdot \\ & \cdot\left(-36864 + 1152h^2\alpha - 8h^4\alpha^2 + h^6\alpha^3\right) \end{aligned}$$

$$B = 4608\left(17 + 12\sqrt{2}\right) - 72\left(17 + 12\sqrt{2}\right)h^2\alpha + \left(65 + 46\sqrt{2}\right)h^4\alpha^2$$

The equations (22) written in matrix notation are

$$(23)\qquad u_{i+1} = Mu_i$$

in which $u_i = [y_i, y_i']^T$, $M = (m_{ij})$,

$$\begin{aligned} m_{11} &= -\frac{B}{A}\left(36864 + 17280h^2\alpha + 968h^4\alpha^2 + 11h^6\alpha^3\right) \\ m_{12} &= -\frac{hB}{A}\left(36864 + 4992h^2\alpha + 136h^4\alpha^2 + h^6\alpha^3\right) \\ m_{21} &= -\frac{h\alpha B}{A}\left(36864 + 4992h^2\alpha + 120h^4\alpha^2\ \right) \\ m_{22} &= m_{11} \end{aligned}$$

We treat the cases $\alpha = k^2$ and $\alpha = -k^2$.

In the following we set $H = hk$ and denote the eigenvalues of the matrix M by $\mu_{1,2}$.

In the first case we get oscillating solutions, so it is important to have eigenvalues of M on or inside the unit circle.

In general the eigenvalues of the amplification matrix M are the roots of the characteristic equation

$$\lambda^2 - 2R\left(H^2\right)\lambda + P\left(H^2\right) = 0$$

where $R\left(H^2\right) = \frac{1}{2}$ trace (M) and $P\left(H^2\right) = \det(M)$ are rational functions of H^2; numerator and denominator of R are polynomials of degree $\leq n$ in H^2. It is known that for polynomial collocation $P\left(H^2\right) = 1$ when the collocation nodes are symmetric [12]. In this case, $R\left(H^2\right)$ is a rational approximation for $\cos H$, called stability function of the method.

Stability means that the numerical solutions remain bounded moving further away from the starting point.

Definition 2. A method is weakly stable in an interval $(0, r)$ if, for each H in $(0, r)$, $|\mu_1| = |\mu_2| = 1$.

Weak stability prevents the numerical solution u_i to spiral into the origin. Every symmetric collocation method is weakly stable in an interval of the form $(0, r)$ [12].

We have that the eigenvalues $\mu_{1,2}$ are complex when $0 \leq H^2 < 9.6$ and $|\mu| = 1 \; \forall H$ in $(0, 9.6)$.

The stability of method CRK compares quite favorably with other one-step fourth-order methods, for example, Runge-Kutta, Runge-Kutta-Nyström methods [9] and Chang-Gnepp method [2]. The stability range of the Runge-Kutta method is $0 \leq H^2 \leq 7.756$, of the Runge-Kutta-Nyström method is $0 \leq H^2 \leq 6.690$ and of the Chang-Gnepp method is $0 \leq H \leq 8.0722$.

Definition 3. An interval $\left(0, H_p^2\right)$ is said to be an interval of periodicity for a method (23) if, for all $H^2 \in \left(0, H_p^2\right)$, $\mu_{1,2}$ are distinct, complex and of modulii one.

If conditions of definition 3 are satisfied for all $H^2 > 0$, the method is P-stable, but one-step polynomial collocation does not provide any P-stable methods [4].

For method CRK, $n = 3$, the interval of periodicity is $(0, 9.6)$. The interval of stability of the fifth-order Nyström method in [3] is $(0, 8.46)$.

Let's now consider the case $\alpha = -k^2$. In the previous case we have oscillating solutions, here the solutions are exponential. We'll study the relative error of method CRK for the equation under discussion, in the case of small h, that is a large number of integration intervals, following the idea of Rutishauser [14].

The maximum eigenvalue of matrix M is

$$\mu = 1 + hk + \frac{1}{2}h^2k^2 + \frac{1}{6}h^3k^3 + \frac{1}{24}h^4k^4 + \frac{13}{1536}h^5k^5 + O\left(h^6k^6\right)$$

Thus the relative error is

$$F \approx \frac{hk - \ln \mu}{h} = \frac{\ln\left(e^{hk}\right) - \ln \mu}{h} = \frac{1}{h}\left(\frac{e^{hk} - \mu}{\mu}\right) \approx \frac{h^4 k^5}{1536}$$

for large x and small h. The relative error for the Runge-Kutta method is $F \approx \frac{h^4 k^5}{120}$, for the Runge-Kutta Nyström method is $F \approx \frac{h^4 k^5}{320}$ and for the method proposed by Chang and Gnepp ([3]) it is $F \approx \frac{h^4 k^5}{720}$.

8 – Chebyshev-Galerkin methods as hybrid symmetric two-step methods

Now we show that methods (11) may be formulated as symmetric two-step hybrid methods in which the position of the off-step points are determined by the x_i defined in (7). In [5] it was proved that a collocation method on the points $t_{k+c_i} = t_k + c_i h$, $i = 1, \ldots, n+1$ with $c_i = \frac{1}{2}(x_i + 1)$ is symmetric (that is the nodes are such that (18) holds). Then the approximations $y_i \approx y(t_i)$ and $z_i \approx y'(t_i)$ satisfy the equations

$$c_i h z_{k+1} = y_{k+1} - y_{k+c_{n+1-i}} + h^2 \sum_{j=1}^{n} b_{ij} f_{k+c_{n+1-j}} \tag{24}$$

for $i = 1, \ldots, n$, $k = 0, 1, \ldots$, where $f_{k+c_{n+1-j}} = f\left(t_k + c_{n+1-j} h, y_{k+c_{n+1-j}}\right)$. (Using the 18) and replacing $k+1$ by k in (24), we have:

$$c_i h z_k = y_k - y_{k-c_i} + h^2 \sum_{j=1}^{n} b_{ij} f_{k-c_j} \tag{25}$$

which, for $i = n+1$, may be written as

$$h z_k = y_k - y_{k-1} + h^2 \sum_{j=1}^{n} b_{n+1,j} f_{k-c_j} \tag{26}$$

and $b_{n+1,j} = \dfrac{1}{4}\beta_j(x_{n+1}) = b_j$.

Equations (15), (16), (17) may be put in the form:

$$\begin{cases} y_{k+1} = y_k + h z_k + h^2 \displaystyle\sum_{j=1}^{n} b_j f_{k+c_j} \\ y_{k+c_j} = y_k + h c_j z_k + h^2 \displaystyle\sum_{i=1}^{n} b_{ji} f_{k+c_i} \\ z_{k+1} = z_k + h \displaystyle\sum_{i=1}^{n} a_i f_{k+c_i} \end{cases} \tag{27}$$

so the method becomes:

$$(28)\qquad \begin{cases} y_{k+1} = 2y_k - y_{k-1} + h^2 \sum_{j=1}^{n} b_j \left(f_{k-c_j} + f_{k+c_j} \right) \\ y_{k+c_j} = 2y_k - y_{k-c_j} + h^2 \sum_{i=1}^{n} b_{ji} \left(f_{k-c_i} + f_{k+c_i} \right) \end{cases}$$

which is a symmetric, hybrid two-step method with $2n$ off-step points between t_{k-1} and t_{k+1} for each k.

If $x_i = \cos \dfrac{(n-i)\pi}{n}$, $i = 0, \ldots, n$, (28) coincides with Panovsky-Richardson implicit method [13]. In this case CRK method may be seen as an alternative formulation of Panovsky-Richardson method.

Equations (28) require starting values at x_0, x_1 and at any off-step points between x_0 and x_1. If these starting values provided by (28) are the approximations generated by CRK method on $[t_0, t_1]$, then, in exact arithmetic the two methods would yield identical results at all subsequent steps [13].

9 – Conclusions

This paper provides a family of numerical collocation methods for initial value problems of the form (1). For each positive integer n two polynomials, one of degree $n+1$, which approximates the exact solution of (1), and the other, of degree n, which approximates its first derivative, are given explicitly.

Numerical tests show that these methods perform as well as other existing methods in terms of stability, of magnitude of the absolute error, and of function evaluations.

REFERENCES

[1] J. C. Buthcher: *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods*, Wiley, Chichester 1987.

[2] S. Chang – P. C. Gnepp: *A direct method for the numerical solution of* $y'' = f(x, y)$, Calcolo, **21** (1984), 369-377.

[3] M. M. Chawla – S. R. Sharma: *Families of Fifth Order Nyström Methods for* $y'' = f(x, y)$ *and Intervals of Periodicity*, Computing, **26** (1981), 247-256.

[4] J. P. Coleman: *Rational approximations for the cosine function; P-acceptability and order*, Numerical Algorithms, **3** (1992), 143-158.

[5] J. P. Coleman – A. S. Booth: *The Chebyschev methods of Panovsky and Richardson as Runge-Kutta-Nyström methods*, J. Comput. Appl. Math., **61** (1995), 245-261.

[6] F. Costabile – A. Napoli: *A method for global approximation of the initial value problem*, Numerical Algorithms, **27** n.2 (2001), 119-130.

[7] F. Costabile – A. Napoli: *Stability of Chebyshev collocation methods*, Pubblicazione n.1 LAN (Unical), Rende (CS) 2002.

[8] P. Davis: *Interpolation and Approximation*, Blaisdell Publishing Company, New York 1975.

[9] J. T. Day: *A one-step method for the numerical integration of the differential equation* $y'' = f(x)y + g(x)$, Comput. J., **7** n.4 (1965), 314-317.

[10] H.Engels: *Numerical quadrature and cubature*, Acad. Press, London 1980.

[11] E.Hairer – S. P.Norsett – G.Wanner: *Solving Ordinary Differential Equations I*, Springer-Verlag, Berlin 1992.

[12] L. Kramarz: *Stability of collocation methods for the numerical solution of* $y'' = f(x, y)$, BIT, **20** (1980), 215-222.

[13] J. Panovsky – D. L. Richardson: *A family of implicit Chebyshev methods for the numerical integration of second-order differential equations*, J. Comput. Appl. Math., **23** n.1 (1988), 35-51.

[14] H. Rutishauser: *Bemerkungen zur numerischen Integration gewohnlicher Differentialgleichungen n-ter Ordnung*, Numer. Math., **2** (1960), 263-279.

Lavoro pervenuto alla redazione il 15 febbraio 2003
ed accettato per la pubblicazione il 4 dicembre 2003.
Bozze licenziate il 6 dicembre 2004

Indirizzo degli Autori:

F. Costabile, A. Napoli – Dipartimento di Matematica – Università della Calabria – 87036 Rende (Cs) Italy
E-mail: costabil@unical.it a.napoli@unical.it

Rendiconti di Matematica, Serie VII
Volume 24, Roma (2004), 261-279

Some applications of new spline spaces in computer aided geometric design

PAOLO COSTANTINI – CARLA MANNI

Abstract: *Aim of this paper is to describe how the so called Variable Degree Polynomial Spaces can be used for the construction of C^3 spatial curves, approximating or interpolating a given set of data. Their main advantages rely in the easy control on their shape, provided by the variable degrees, and in the low computational cost, comparable with that of standard quintic splines.*

1 – Introduction

Geometric continuous curves and surfaces based on polynomial or rational splines constitute the main tool of Computer Aided Geometric Design because of their simplicity and because of the easy and intuitive control on their shape provided by the so-called shape parameters. However, in some CAD/CAM applications, as, for instance, in the description of the motion of a milling machine, the physical meaning of the parameter is not negligible and a certain order of analytic continuity is often required; therefore new tools which encompass the new and the old requests would be highly desirable.

Aim of this paper is to describe the properties and some applications of new *quintic-like* spline spaces (called Variable Degree Polynomial Spaces, VDPS for short) which permit the construction of C^3 polynomial (or rational) curves and surfaces with the same simplicity, computational cost and ease of shape control as the classical quintics. Indeed, these spaces are isomorphic to the spaces of C^3

Key Words and Phrases: *Spline curves – Interpolation – Best approximation – Shape preservation – Tension property.*
A.M.S. Classification: 65D05 – 65D07 – 65D10 – 65D17

quintic splines and possess a control polygon (called *pseudo Bézier control net*) with all the usual geometric properties. Therefore, all the geometric construction that are used in CAGD can be repeated. Additionally, the degrees play the role of tension parameters, since their large values force the curve to have a piecewise linear appearance.

This paper is divided in five sections. In the next one the structure of VDPS will be briefly recalled and in Section 3 we will describe a simplification of the geometric construction for C^4 quintic splines, which is suitable for our purposes. Section 4 is devoted to show applications of VDPS in the interpolation and approximation of ordered spatial data. In the last section are reported some concluding remarks and open problems.

It is worthwhile to say that this paper has a structure very similar to [5]; in that paper an analogous geometric construction, also derived from C^4 quintic splines, is used to produce C^2 quintic splines with a third order *Frenet continuity* ($C^2 - FC^3$) – that is curvature and torsion continuous – splines. The advantages and disadvantages of [5] and of the present paper are, roughly speaking, symmetric and with an equivalent comprehensive effect: here, we have an higher continuity order (C^3 implies FC^3) at the price of the more complex space structure induced by the degrees; there, a lack in the continuity with the advantage of low degree splines, which so far constitute the standard *mathematical engines* of CAD/CAM environments.

2 – The spline space

In this section we want to briefly introduce the main properties of the C^3 *quintic-like* VDPS, referring for details to [4] and [7]. Let $\{u_0, u_1, \dots, u_m\}$ be an ordered knot sequence, let $h_i,\ i = 0, \dots, m-1$, be the knot spacing, and let

$$\mathbf{k} = \{k_i;\ i = 0, 1, \dots, m\},$$

with $k_i \geq 5$ be a given sequence of integers. For each interval $[u_i, u_{i+1}]$ we consider the six dimensional polynomial space:

$$VP_{k_i,k_{i+1}} := span\{(1-v), v, (1-v)^{k_i}, v(1-v)^{k_i-1}, v^{k_{i+1}-1}(1-v), v^{k_{i+1}}\}$$

with $v = (u-u_i)/h_i$, called *quintic-like variable degree polynomial space*. Denoting by $\mathbb{P}_n$ the space of algebraic polynomials of degree less than or equal to n, we remark that $VP_{k_i,k_{i+1}}$ is isomorphic to $\mathbb{P}_5$ and, in particular, $VP_{5,5} = \mathbb{P}_5$. Moreover, as it is shown in [4], $VP_{k_i,k_{i+1}}$ admits a *pseudo Bernstein-Bézier basis*

$$\{\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3, \mathcal{B}_4, \mathcal{B}_5\}$$

that is a basis with the usual properties – positivity, partition of unity – as the Bernstein-Bézier basis for $\mathbb{P}_5$. Therefore, for any $b \in VP_{k_i,k_{i+1}}$, we have

$$b = \sum_{j=0}^{5} b_{i,j}\mathcal{B}_{i,j}, \qquad \mathcal{B}_{i,j}(u) = \mathcal{B}_j((u-u_i)/h_i),$$

where the $b_{i,j}$ are called *pseudo Bézier ordinates* (the suffix i refers to the interval $[x_i, x_{i+1}]$) and play the same role as the usual control points for quintic polynomials. In particular, setting

$$\xi_{i,0} := u_i, \quad \xi_{i,1} := u_i + \frac{h_i}{k_i}, \quad \xi_{i,2} := u_i + 2\frac{h_i}{k_i},$$
$$\xi_{i,3} := u_{i+1} - 2\frac{h_i}{k_{i+1}}, \quad \xi_{i,4} := u_{i+1} - \frac{h_i}{k_{i+1}}, \quad \xi_{i,5} := u_{i+1}.$$

we have

$$u = \sum_{j=0}^{5} \xi_{i,j}\mathcal{B}_{i,j}(u).$$

For details we refer to [4].

Now, let us consider

$$VS_{\mathbf{k}} := \{s \in C^3[u_0, u_m] \text{ s.t. } s|_{[u_i,u_i+1]} \in VP_{k_i,k_{i+1}}\}$$

the space of quintic-like VDPS.

In [7] it is shown that $VS_{\mathbf{k}}$ admits a basis,

$$\{N_{2i}, N_{2i+1};\ i = -1, 0, \dots, m\}$$

defined, as usual, on an extended knot sequence

$$u_{-2} < u_{-1} < u_0 < u_1 < \cdots < u_m < u_{m+1} < u_{m+2},$$

having the classical properties of the B-spline basis.

The tension effect achieved with large degree values clearly appears from the plots of the B-spline basis functions shown in Figure 1; indeed, if the degrees tend simultaneously to infinity, the B-splines tend to the normalized piecewise linear B-splines.

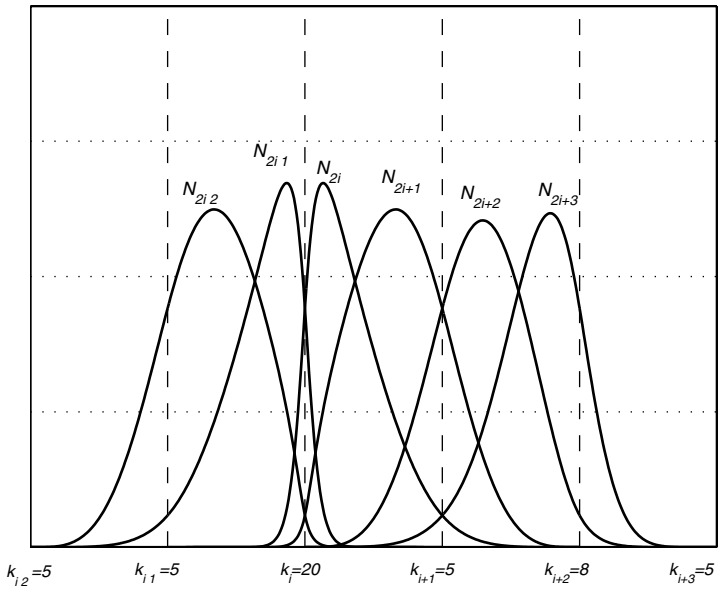

Fig. 1: Some B-spline basis functions.

3 – Geometric construction

We consider in this section variable degree spline curves in $\mathbb{R}^3$, that is

$$\mathbf{VS_k} = \{\mathbf{s} \text{ s.t. } \mathbf{s} : [u_0, u_m] \to \mathbb{R}^3, \text{ has components in } VS_\mathbf{k}\} .$$

For explaining our ideas, we start with standard C^4, quintic spline curves which can be seen as a particular case of VDPS with $k_0 = k_1 = \ldots = k_m = 5$. If we denote with $\{\tilde{N}_i,\ i = -2, -1, \ldots, m+2\}$ the sequence of normalized quintic B-splines of class C^4 and take a sequence of coefficients (often referred to as *de Boor control points*) $\{\mathbf{D}_i,\ \ i = -2, -1, \ldots, m+2\}$, a C^4 quintic spline curve can be expressed as

$$\mathbf{s} = \sum_{i=-2}^{m+2} \mathbf{D}_i \tilde{N}_i, \quad \mathbf{D}_i \in \mathbb{R}^3.$$

Obviously, $\mathbf{s}_i := \mathbf{s}|_{[u_i . u_{i+1}]}$ can be expressed in the Bernstein-Bézier form

$$\mathbf{s}_i = \sum_{n=0}^{5} \mathbf{b}_{i,n} \mathcal{B}_{i,n} \ ,$$

the coefficients $b_{i,j}$ are called *Bézier control points*.

One of the most attractive features of spline curves is their *geometric construction*, that is the possibility of constructing the Bézier control points $\mathbf{b}_{i,n}$ directly from the de Boor control points $\mathbf{D}_j$ via a *corner-cutting* process. This geometric construction can be schematically divided in two main steps (explained in Figures 2.a and 2.b). In the first one (see Figure 2.a) the polygonal legs $\mathbf{D}_i\mathbf{D}_{i+1}$, $i = -2, -1, \dots, m+1$ are divided in three segments proportional to $h_{i-2}+h_{i-1}$, h_i and $h_{i+1}+h_{i+2}$ and the additional points $\mathbf{F}_i^+$, $\mathbf{F}_{i+1}^-$ are inserted; then for $i = -1, 0, \dots, m+1$ the segment $\mathbf{F}_i^-\mathbf{F}_i^+$ is divided in three subsegments proportional to h_{i-2}, $h_{i-1}+h_i$, h_{i+1} and the points $\mathbf{p}_i$, $\mathbf{r}_i$ are placed on it. In the second step (see Figure 2.b) the segments $\mathbf{r}_i\mathbf{p}_{i+1}$, $i = -1, 0, \dots, m+1$ are subdivided with proportionality h_{i-1}, h_i, h_{i+1} and the Bézier control points $\mathbf{b}_{i,2}$, $\mathbf{b}_{i,3}$ are inserted; then, for $i = -1, 0, \dots, m+1$ the point $\mathbf{q}_i$ is inserted in $\mathbf{p}_i\mathbf{r}_i$ with proportionality h_{i-1} and h_i; finally, the same factors are used to insert the control points $\mathbf{b}_{i-1,4}$, $\mathbf{b}_{i,1}$ in the segments $\mathbf{b}_{i-1,3}\mathbf{q}_i$, $\mathbf{q}_i\mathbf{b}_{i,2}$ and $\mathbf{b}_{i-1,5} = \mathbf{b}_{i,0}$ in the segment $\mathbf{b}_{i-1,4}\mathbf{b}_{i,1}$ respectively. Note that this procedure (which is mathematically proved using the subdivision scheme given by the De Casteljau algorithm) automatically constructs C^4 curves. We refer to to [13], [8], [19] for the formal details.

Now let us consider curves in the more general space $\mathbf{VS_k}$. A spline curve $\mathbf{s} \in \mathbf{VS_k}$ can be expressed as

$$\mathbf{s} = \mathbf{r}_{-1}N_{-2} + \sum_{i=0}^{m}(\mathbf{p}_i N_{2i-1} + \mathbf{r}_i N_{2i}) + \mathbf{p}_{m+1}N_{2m+1} \ .$$

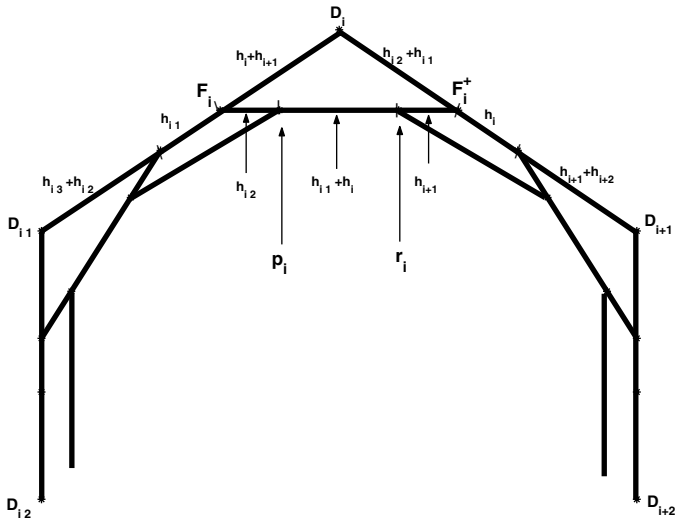

Fig. 2a: Geometric construction of C^4 quintic splines. First step.

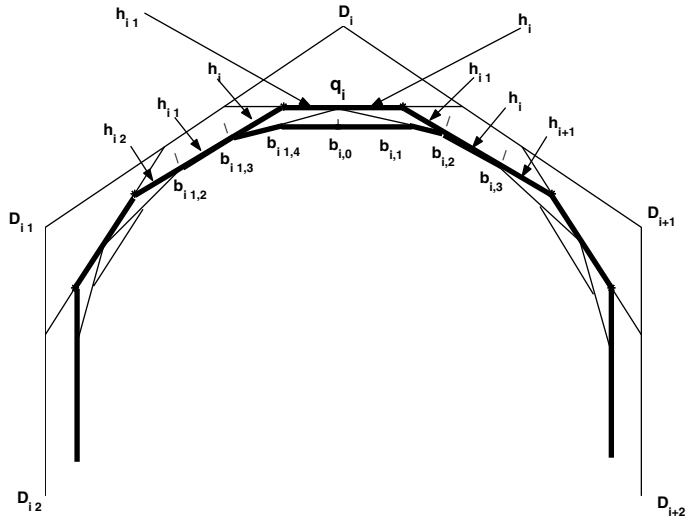

Fig. 2b: Geometric construction of C^4 quintic splines. Second step.

Obviously, $\mathbf{s}_i := \mathbf{s}|_{[u_i.u_{i+1}]}$ can be expressed in the Bernstein-Bézier form

$$\mathbf{s}_i = \sum_{n=0}^{5} \mathbf{b}_{i,n} \mathcal{B}_{i,n} \ ,$$

the coefficients $b_{i,j}$ are called *pseudo Bézier control points*.

In [4] it is shown that a geometric construction similar to that one illustrated in Figure 2.b holds also for the general case. More specifically it is possible to construct a C^3 curve belonging to $\mathbf{VS_k}$ starting from a control polygon connecting the control points $\mathbf{p}_i$, $\mathbf{r}_i$ as specified in Figure 3. The remarkable fact of this construction is that, for a large value of the degree k_i, both the points $\mathbf{p}_i$, $\mathbf{r}_i$ and the pseudo Bézier control points $\mathbf{b}_{i-1,3}, \ldots, \mathbf{b}_{i,2}$ are attracted by the central point $\mathbf{q}_i$. In other words, the degrees play the role of *tension parameters* and the shape of the curve can be easily modified to reach a piecewise linear appearance; in practice we have the same shape control as for the *geometric continuous* ([13]) splines with the advantage of maintaining the analytical continuity. It is worthwhile to recall that the computational cost does not depend on the degrees and is approximately the same as the quintic one. See [4] for details.

However, in this construction we have two control points associated to each knot and, instead of being an advantage, this flexibility implies the additional difficulty of choosing the slope of the segment $\mathbf{p}_i\mathbf{r}_i$.

The idea of this paper is very simple: to consider the points $\mathbf{r}_i$, $\mathbf{p}_i$ as obtained from the first corner-cutting step embedding the construction of C^3 VDPS of

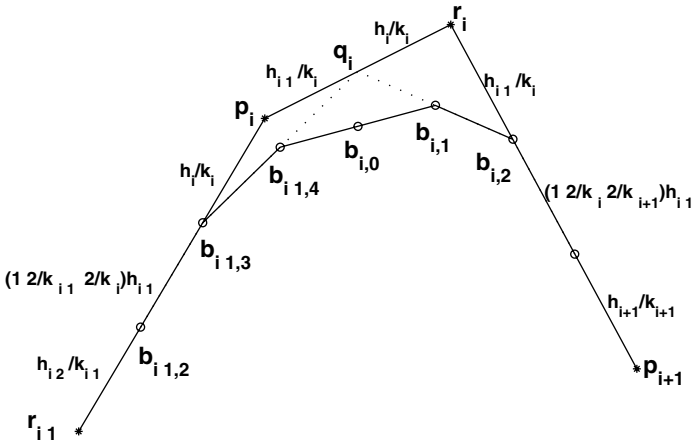

Fig. 3: Geometric construction of C^3 VDPS splines.

Figure 3 in the quintic C^4 scheme of Figures. 4. Thus, only *one* control point is associated to each knot. Of course in this way we are dealing with a subspace of $\mathbf{VS_k}$. The simplified geometric construction of the elements of this subspace is a consistent advantage both for their use in interpolation/approximation of spatial data and in free form design. We refer to [4], [7] for a comparison. More

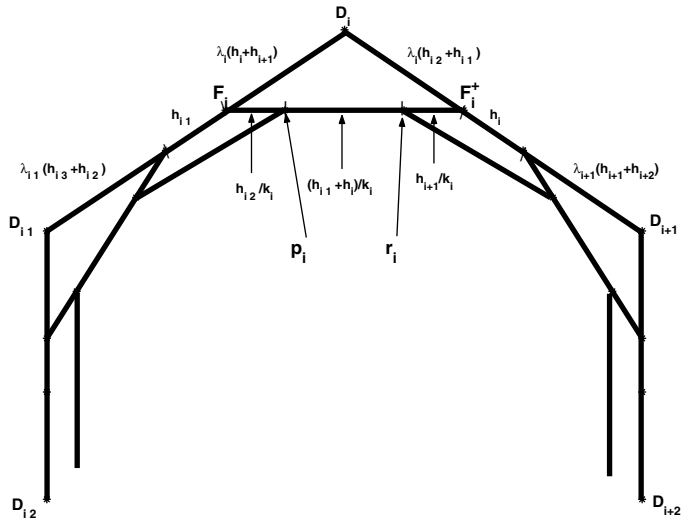

Fig. 4a: Geometric construction of C^3 VDPS. First step.

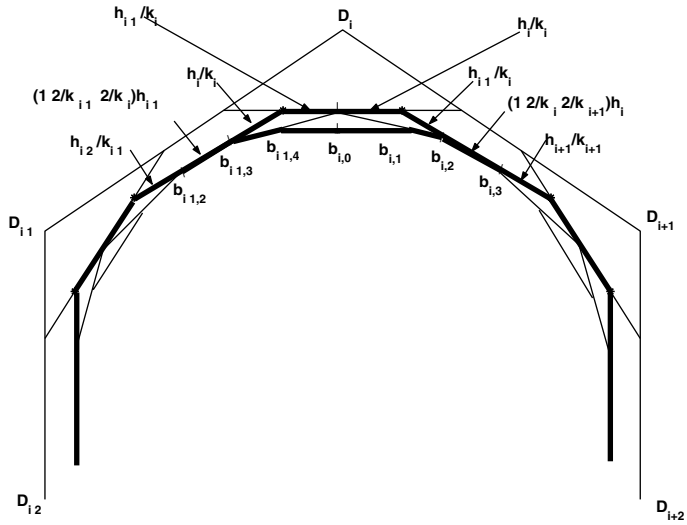

Fig. 4b: Geometric construction of C^3 VDPS. Second step.

specifically, the second step of the corner-cutting remains unchanged (see Figures. 3 and 4.b) while in the first step we introduce at each knot a new shape parameter, λ_i (see Figures. 2.a and 4.a). Obviously, we have again quintic C^4 splines for the choice $\lambda_i = 1$, $k_i = 5$, all i.

Since the polygonal legs $\mathbf{D}_i\mathbf{D}_{i+1}$ are divided in three segments proportional to $\lambda_i(h_{i-2} + h_{i-1})$, h_i and $\lambda_{i+1}(h_{i+1} + h_{i+2})$ it is clear that the points $\mathbf{F}_i^-$ and $\mathbf{F}_i^+$ are attracted by $\mathbf{D}_i$ for small values of λ_i. Therefore, the combined effect of small λ_i and large k_i produces a tension effect on the final curve. See Figure 5 for a graphical example, where we have chosen $\lambda_i = 1/k_i$ (obviously this is just one among the possible choices: λ_i and k_i can be chosen independently).

4 – Applications

The researches described in this paper have been mainly motivated by the necessity of constructing interpolating (for CAD/CAM applications) or approximating (for some reverse engineering applications) curves capable of maintaining the geometric characteristics (discrete curvature and discrete torsion [20]) of the data set.

4.1 – Interpolation of spatial data

We start with a brief description of the interpolation problem, referring to [1], [2], [3], [4], [10], [11], [12], [15], [16], [17] for related papers.

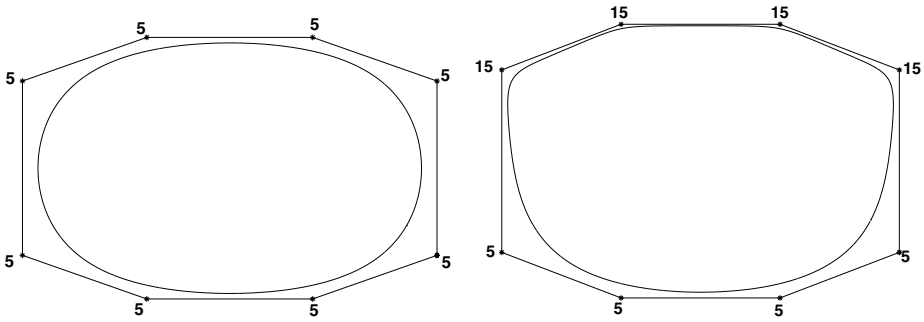

Fig. 5: Left: an example of C^4 quintic spline curve. Right: an example of C^3 VDPS curve. The numbers indicate the degree k_i associated to each de Boor control point

Let

$$\mathbf{I}_i \in \mathbb{R}^3, \qquad i = 0, \dots, m,$$

be the interpolation points with $\mathbf{I}_i \neq \mathbf{I}_{i+1}$. For a given matrix M let $|M| := \det(M)$. Define, for all admissible indices,

$$\mathbf{L}_i = \mathbf{I}_{i+1} - \mathbf{I}_i \ , \qquad i = 0, \dots, m-1,$$

$$\mathbf{B}_i := \begin{cases} \dfrac{\mathbf{L}_{i-1} \times \mathbf{L}_i}{\|\mathbf{L}_{i-1}\| \ \|\mathbf{L}_i\|}, & \text{if } \|\mathbf{L}_{i-1}\| \ \|\mathbf{L}_i\| > 0, \\ 0, & \text{elsewhere,} \end{cases} \qquad i = 1, \dots, m-1,$$

$$\Delta_i := \begin{cases} \dfrac{|\mathbf{L}_{i-1} \ \mathbf{L}_i \ \mathbf{L}_{i+1}|}{\|\mathbf{L}_{i-1} \times \mathbf{L}_i\| \ \|\mathbf{L}_i \times \mathbf{L}_{i+1}\|}, & \text{if } \|\mathbf{L}_{i-1} \times \mathbf{L}_i\| \|\mathbf{L}_i \times \mathbf{L}_{i+1}\| > 0, \\ 0, & \text{elsewhere,} \end{cases} \qquad i = 1, \dots, m-2,$$

The vectors $\mathbf{B}_i$ are the *discrete binormals* and the scalars Δ_i are the *discrete torsion* of the data ([20]).

Given a spline curve $\mathbf{s} = \mathbf{s}(u)$, we consider the corresponding *curvature vector* $\mathbf{K}(u)$ and the *torsion* $\tau(u)$:

$$\mathbf{K}(u) := \frac{\mathbf{s}'(u) \times \mathbf{s}''(u)}{\|\mathbf{s}'(u)\|^3}, \qquad \text{if } \ \mathbf{s}'(u) \neq \mathbf{0} \ ,$$

$$\tau(u) := \frac{|\mathbf{s}'(u) \ \mathbf{s}''(u) \ \mathbf{s}'''(u)|}{\|\mathbf{s}'(u) \times \mathbf{s}''(u)\|^2}, \qquad \text{if } \ \mathbf{s}'(u) \times \mathbf{s}''(u) \neq \mathbf{0},$$

and following the usual definitions (see, e.g., [1], [11], [7]), we formally define the *shape-constraints* (for a geometric interpretation the reader is referred, for example, to [1]). Let us denote with $\mathbf{I}$ the polygonal line connecting the data points $\{\mathbf{I}_0, \dots, \mathbf{I}_m\}$.

DEFINITION 1. Let $\mathbf{s}(u)$ be a spline curve defined for $u \in [u_0, u_m]$ and let ϵ_1, ϵ_2 two real, positive tolerances. We say that $\mathbf{s}(u)$ is $\mathbf{I}$-shape preserving if the following criteria are satisfied:

(i) *Weak collinearity criteria*
If $\|\mathbf{B}_i\| \le \epsilon_1$ and $\mathbf{L}_{i-1} \cdot \mathbf{L}_i > 0$ then $\left\| \frac{\mathbf{s}'(u)}{\|\mathbf{s}'(u)\|} \times \frac{\mathbf{L}_j}{\|\mathbf{L}_j\|} \right\| \le \epsilon_2$,
$j = i-1, i$, in each arbitrary but fixed closed subinterval of (u_{i-1}, u_{i+1}) where $\|\mathbf{s}'(u)\| \neq 0$.

(ii) *Convexity criteria*
(ii.1) If $\|\mathbf{B}_i\| \neq 0$, then $\mathbf{K}(u_i) \cdot \mathbf{B}_i > 0$.
(ii.2) If $\mathbf{B}_i \cdot \mathbf{B}_{i+1} > 0$, then $\mathbf{K}(u) \cdot \mathbf{B}_j > 0, j = i, i+1$, $u \in (u_i, u_{i+1})$.

(iii) *Weak coplanarity criteria*
If $|\Delta_i| \le \epsilon_1$ then $\left\| \frac{\mathbf{s}'(u) \times \mathbf{s}''(u)}{\|\mathbf{s}'(u) \times \mathbf{s}''(u)\|} \times \mathbf{B}_i \right\| \le \epsilon_2$, $u \in [u_i, u_{i+1}]$, if $\|\mathbf{K}(u)\| \neq 0$.

(iv) *Torsion criteria*
(iv.1) If $\Delta_{i-1}\Delta_i > 0$ then $\tau(u_i)\Delta_j > 0, j = i-1, i$.
(iv.2) If $\Delta_i \neq 0$ then $\tau(u)\Delta_i > 0$, n each arbitrary but fixed closed subinterval of (u_i, u_{i+1}).

Our goal is to construct an $\mathbf{I}$-shape preserving interpolating spline, that is

$$\mathbf{s} \in \mathbf{VS_k} \text{ such that } \mathbf{s}(u_i) = \mathbf{I}_i \ , \ i = 0, \dots, m$$

which satisfies conditions (i)-(iv) of Definition 1.

To uniquely solve the problem, four conditions must be added. Following [5], in the case of **closed curves**, we assume that $\mathbf{I}_0 = \mathbf{I}_m$ and we impose

$$\mathbf{D}_{-2} = \mathbf{D}_{m-2} \ , \ \mathbf{D}_{-1} = \mathbf{D}_{m-1} \ ; \mathbf{D}_{m+1} = \mathbf{D}_1 \ , \ \mathbf{D}_{m+2} = \mathbf{D}_2 \ .$$

The problem is more complex in the case of **open curves**; there are indeed many possible solutions which strictly depend on the applications and on the user desires. Very simple equations are

$$\mathbf{D}_{-2} = \mathbf{D}_{-1} = \mathbf{D}_0 \ ; \mathbf{D}_{m+2} = \mathbf{D}_{m+1} = \mathbf{D}_m \ ,$$

which imply that the shape of the curve at the end points is not influenced by extraneous conditions, but only depend on the first interpolation points, or

$$\begin{aligned}
\mathbf{D}_{-1} &= \mathbf{D}_0 + \alpha_0(\mathbf{D}_1 - \mathbf{D}_0) \ , \\
\mathbf{D}_{-2} &= \mathbf{D}_0 + \alpha_0(\mathbf{D}_1 - \mathbf{D}_0) + \beta_0(\mathbf{D}_2 - \mathbf{D}_1) \ ; \\
\mathbf{D}_{m+1} &= \mathbf{D}_m + \alpha_m(\mathbf{D}_m - \mathbf{D}_{m-1}) \ , \\
\mathbf{D}_{m+2} &= \mathbf{D}_m + \alpha_m(\mathbf{D}_m - \mathbf{D}_{m-1}) + \beta_m(\mathbf{D}_{m-1} - \mathbf{D}_{m-2}) \ ,
\end{aligned}$$

which produces natural-like end conditions, that is vanishing curvature and torsion at end points.

The theoretical aspects are grounded on the results of [4] and are completely similar to those of [5]; we limit therefore to sketch the main points, avoiding useless duplications.

Let us consider the augmented set of control points $\{\mathbf{D}_{-2}, \dots, \mathbf{D}_{m+2}\}$ (given by proper end conditions). The pseudo Bézier control points $\mathbf{b}_{0,0}, \mathbf{b}_{1,0}, \dots, \mathbf{b}_{m-1,0}, \mathbf{b}_{m,0} := \mathbf{b}_{m,5}$, play a particular role, since $\mathbf{b}_{i,0} = \mathbf{s}(u_i),\ i = 0, \dots, m$ and therefore the interpolation conditions can be rewritten as $\mathbf{b}_{i,0} = \mathbf{I}_i,\ i = 0, \dots, m$. Now, if we use the corner cutting process described in Figures 4 for their computations, we have linear equations of the form

$$\mathbf{b}_{i,0} = \sum_{j=i-2}^{i+2} \alpha_{i,j}\mathbf{D}_j \ .$$

The coefficients $\alpha_{i,j}$ can be computed using the Maple instructions reported in the appendix; the explicit expressions are extremely cumbersome and are not reported here for reasons of space. Let us denote with A the matrix obtained augmenting the collocation matrix $(\alpha_{i,j})_{0\le i\le m,-2\le j\le m+2}$ with the four rows given by one of the boundary conditions. We have the following result.

THEOREM 1. *For λ_i sufficiently small and k_i sufficiently large, the matrix A is strictly diagonally dominant.*

The proof can be obtained running the Maple program of the appendix. Obviously we have the following corollary.

COROLLARY 1. *For λ_i sufficiently small and k_i sufficiently large, there exists one and only one interpolating spline $\mathbf{s} \in \mathbf{VS_k}$*

We observe that, as in other interpolation problems with geometric continuous curves, we have not a formal proof for the existence of a solution for all λ_i, k_i. Additional results could be obtained, for specified boundary conditions, using a geometric analysis of the null space similar to those presented in [2] and [4]. We remark, however, that in the huge amount of numerical experiments singular matrices have never occurred. Therefore we safely conjecture that A is non-singular for any choice of λ_i, k_i, also supported by the consideration that the geometric structure of the corner cutting described in Figures. 4 – which produces the matrix elements – is not dependent on λ_i and k_i.

Using the same arguments of [5] we have the following asymptotic result.

THEOREM 2. *Let* $\lambda_i \to 0$ *and* $k_i \to \infty$. *Then* $\mathbf{b}_{\nu,\mu} \to \mathbf{D}_i$ *, for* $\nu = i-1, \mu = 3, 4, 5$ *and* $\nu = i, \mu = 0, 1, 2$.

The above theorem says that, for proper values of the shape parameters λ_i, k_i, the pseudo Bézier control net has the same shape of the *pseudo De Boor control net*, that is the polygonal line connecting the $\mathbf{D}_i$. This can be restated saying that both the pseudo Bézier and the pseudo De Boor control nets tend to the polygonal interpolating the data points. Taking a well-known result of [9] and repeating the same considerations of [4] we claim the following proposition.

PROPOSITION 1. *For* λ_i *sufficiently small and* k_i *sufficiently large the shape induced by the curvature and the torsion of* $\mathbf{s}$ *is the same as the shape induced by the discrete curvature and torsion of the pseudo Bézier control polygon.*

Summarizing we have the following result.

THEOREM 3. *It is possible to find sequences* $\lambda_0, \ldots, \lambda_m$ *and* $k_0, \ldots, k_m$ *such that the interpolating spline curve* $\mathbf{s}$ *is* $\mathbf{I}$*-shape preserving.*

Figure 6. (left) shows the plot of the C^4 quintic curve interpolating the so called *chair data*, [15]. For emphasizing the shape effect we have used the uniform, instead of the centripetal arc-length, parameterization. The choice $k_1 = k_{11} = 27$ ($\mathbf{I}_0$ is the highest point) reduces the unwanted inflections, as shown in Figure 6 right.

4.2 – Approximation of spatial data

Now let us turn to shape preserving approximation of spatial data. Despite its practical importance, this argument has received much less attention. For planar data the only papers seems to be [14] and [18] and, for the spatial case, [6], [7] and, partially, [5].

Let $\{(t_j, \mathbf{I}_j),\ j = 0, \ldots, N\}$, with $\mathbf{I}_j \in \mathbb{R}^3$ be a set of data points. The first problem we must solve is the definition of the *shape of the data*. Again, for reason

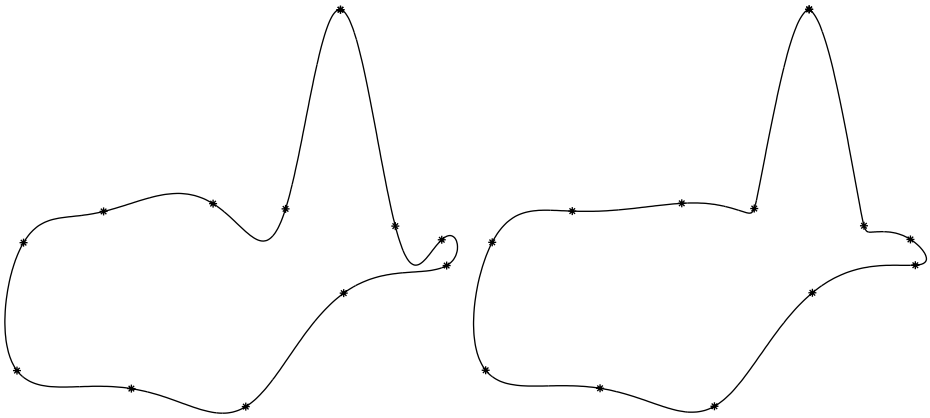

Fig. 6: Chair data. Left: C^4 quintic interpolating spline curve. Right: C^3 interpolating VDPS curve.

of space, we refer to [6] for details. The basic idea goes as follows. We extract from the data parameters a sequence of *significant* knots $\{u_0, u_1, \ldots, u_m\}$ with $u_0 = t_0$, $u_m = t_N$ and we define the space

$$\boldsymbol{\mathcal{L}} := \{\boldsymbol{\ell} \in C[u_0, u_m] \ \text{ s.t. } \boldsymbol{\ell}|_{[u_i, u_{i+1}]} \text{ has components in } \mathbb{P}_1\}.$$

We then take $\boldsymbol{\psi} \in \boldsymbol{\mathcal{L}}$, the best least squares approximation to data and we simply use the discrete curvature and torsion of $\boldsymbol{\psi}$ for defining the shape of the data; see Figure 7 for a planar example taken from [14].

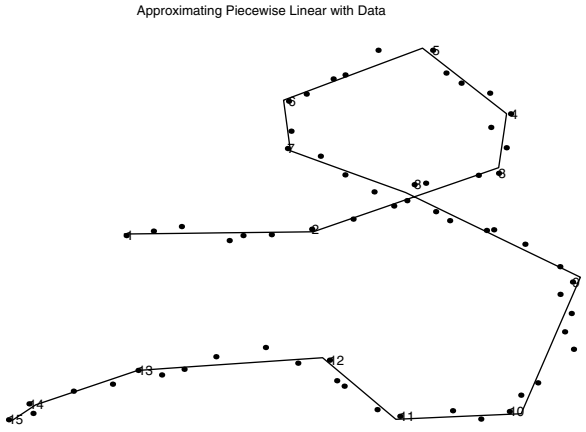

Fig. 7: The shape of a data set.

Our goal is to compute the spline curve of best approximation, that is $\mathbf{s}^*$ such that

$$|\mathbf{s}^* - \mathbf{I}| \leq |\mathbf{s} - \mathbf{I}|, \quad \forall\, \mathbf{s} \in \mathbf{VS_k} \ ,$$

where $|\mathbf{v}|^2 := \sum_{j=0}^{N} \|\mathbf{v}(t_j)\|^2$. Since we are mainly interested in CAD applications, our splines will be constrained to satisfy the boundary conditions of the previous section. However, following the same ideas of [6] and [7], we will not use constrained least square techniques but we simply perform an unconstrained minimization in the subspace

$$\mathbf{V\Sigma}_k := \{\mathbf{s} \in \mathbf{VS_k} \text{ such that } \mathbf{s} \text{ satisfies boundary conditions}\} \ ,$$

seeking for $\boldsymbol{\sigma}^* \in \mathbf{V\Sigma}_k$ such that

$$|\boldsymbol{\sigma}^* - \mathbf{I}| \leq |\boldsymbol{\sigma} - \mathbf{I}|, \quad \forall\, \boldsymbol{\sigma} \in \mathbf{V\Sigma}_k \ .$$

Again, the theoretical aspects are grounded on the results of [7] and are similar to those of [5].

Obviously, Proposition 1 and Theorems 1 and 2 still hold; since for $\lambda_i \to 0$ and $k_i \to \infty$, all i, the space $\mathbf{V\Sigma}_k$ approaches $\boldsymbol{\mathcal{L}}$, we state the following result.

THEOREM 4. *Let $\lambda_i \to 0$, $k_i \to \infty$ for $i = 0, 1, \ldots, m$. Then $\boldsymbol{\sigma}^* \to \boldsymbol{\psi}$.*

Note that the above theorem implies that also the pseudo Bézier and de Boor control polygons tend to $\boldsymbol{\psi}$; therefore we have the following result

COROLLARY 2. *It is possible to find sequences $\lambda_0, \ldots, \lambda_m$ and $k_0, \ldots, k_m$ such that the approximating spline curve $\boldsymbol{\sigma}^*$ is $\boldsymbol{\psi}$-shape preserving.*

The main drawback of the above result is that it is *global* in nature; if only some shape parameters tend to the limit values the space $\mathbf{V\Sigma}_k$ does not tend to $\boldsymbol{\mathcal{L}}$ and the asymptotic shape preserving properties vanish. The consequence is that all the segments of the curve are simultaneously stretched and the curve can assume an unpleasant appearance. In [7] is presented a solution which uses a *weighted* approximation, which can be here adopted. The basic idea is that we accept a compromise, obtaining the convergence at the price of a reduction in the approximation power. In order to force the spline to locally approach $\boldsymbol{\psi}$ when a local increase is applied, we work with an extension of the approximation problem. Let $w = \{w_0, \ldots, w_m\}$ be a sequence of positive weights. In the following we use the notation $\boldsymbol{\sigma} = \boldsymbol{\sigma}_{k,w}$ and use $\boldsymbol{\theta}_{k,w}$ and $\boldsymbol{\ell}_{k,w} = \boldsymbol{\ell}(\boldsymbol{\sigma}_{k,w})$ to denote, respectively, the piecewise linear curves interpolating the control points $\{\mathbf{D}_0, \ldots, \mathbf{D}_m\}$ and the spline at the knots $\{\boldsymbol{\sigma}(u_0), \ldots, \boldsymbol{\sigma}(u_m)\} = \{\mathbf{b}_{0,0}, \ldots, \mathbf{b}_{m,0}\}$. The basic idea is to push $\boldsymbol{\sigma}^*_{k,w}(u_i)$ towards $\boldsymbol{\psi}(u_i)$, by inserting $\boldsymbol{\psi}(u_0), \ldots, \boldsymbol{\psi}(u_m)$ as

weighted points in the approximation problem. The approximation points become:

$$\{(t_j, \mathbf{I}_j) : j = 0, \dots, N\} \cup \{(u_i, \boldsymbol{\psi}(u_i)) : i = 0, \dots, m\},$$

and we find $\boldsymbol{\sigma}^*_{k,w} \in \mathbf{V}\boldsymbol{\Sigma}_k$ which is the best weighted least squares approximation, that is which minimizes the following functional

$$\sum_{j=0}^{N} ||\boldsymbol{\sigma}_{k,w}(t_j) - \mathbf{I}_j||^2 + \sum_{i=0}^{m} w_i ||\boldsymbol{\sigma}_{k,w}(u_i) - \boldsymbol{\psi}(u_i)||^2, \qquad \text{s.t.} \quad \boldsymbol{\sigma}_{k,w} \in \mathbf{V}\boldsymbol{\Sigma}_k.$$

We can use the values of the weights w_i to control behavior of the curve at u_i; in particular if $w_i = 0, i = 0, \dots, m$ we have the old approximation problem, and

$$\lim_{w_i \to \infty} \boldsymbol{\sigma}^*_{k,w}(u_i) \to \boldsymbol{\psi}(u_i).$$

To be more precise, we use the weights for imposing that $\boldsymbol{\ell}_{k,w}$ has the same shape of $\boldsymbol{\psi}$ and the degrees for stretching the curve, that is for imposing that $\boldsymbol{\sigma}^*_{k,w}$ has the same shape of $\boldsymbol{\ell}_{k,w}$. Obviously, the larger are the weights, the more the approximation to the *true data* $\mathbf{I}_0, \dots, \mathbf{I}_N$ deteriorates, and we want to keep the weights as small as possible. We refer to [7] for details on the algorithm.

We limit ourselves to Figure 8 for a graphical comparison. In Figure 8 (left) are reported the data set (random perturbations of equally spaced points over an helix), $\boldsymbol{\psi}$ and $\boldsymbol{\sigma}^*_k$ obtained using the global scheme; in Figure 8 (right) have been depicted similar plots for the local scheme.

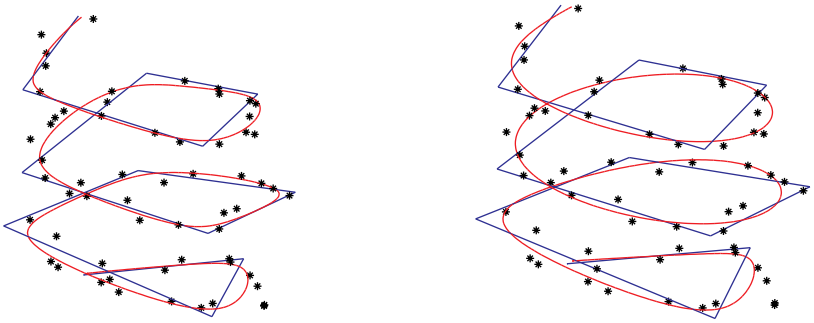

Fig. 8: Left: the global approximant. Right: the local approximant.

5 – Closure

We have presented a new class of C^3 functions which can be used for solving some important problems of CAGD. Their main advantage relies in the simple

geometric construction which, in turn, permits an easy description of the shape constraints and an easy choice of the optimal shape parameters.

It is worthwhile to remind that many CAD/CAM systems are based on standard NURBS with low degrees. The structure of Figures. 4 can be adapted to produce FC^3 (*Frenet-frame continuous* [13]) quintic curve; the corresponding results are reported in [5]. However, it is important to point out that, even if $C^2 - FC^3$ is a reasonable smoothness property both from the mechanical point of view (the motion of a point on the curve has a continuous acceleration) and from the geometric point of view (the tangent and curvature vectors and the torsion are continuous), the C^3 continuity is sometimes required. For instance, if the physical meaning of the parameter is time and the spline curve is used to control the motion of a robot, a smooth (C^1) acceleration will preserve the engines from harsh stresses.

We conclude the paper observing that the tensor-product extension seems straightforward; the non-obvious problems are to extract the information on the shape of the data, especially in the approximation case, and to assign to the data a parameterization suitable for our purposes. The corresponding researches are under study.

– APPENDIX – Maple instructions

Explicit expression of the points obtained in the corner cutting process described in Figures 4; check of the corresponding limits; explicit computation of the coefficients of the i-th row of the interpolation matrix M and check of its asymptotic diagonal dominance.

For notational simplicity we have set: `delta[i]:=1/k[i] .`

```
> Fp[i-2]:=((h[i-2]+lambda[i-1]*(h[i-1]+h[i]))*D[i-2]
> +lambda[i-2]*(h[i-4]+h[i-3])*D[i-1])/(lambda[i-2]*(h[i-4]+
> h[i-3])+h[i-2]+lambda[i-1]*(h[i-1]+h[i]));

> Fm[i-1]:=((lambda[i-2]*(h[i-4]+h[i-3])+h[i-2])*D[i-1]+
> lambda[i-1]*(h[i-1]+h[i])*D[i-2])/(lambda[i-2]*(h[i-4]+h[i-3])+
> h[i-2]+lambda[i-1]*(h[i-1]+h[i]));
> Fp[i-1]:=((h[i-1]+lambda[i]*(h[i]+h[i+1]))*D[i-1]+
> lambda[i-1]*(h[i-3]+h[i-2])*D[i])/(lambda[i-1]*(h[i-3]+h[i-2])+
> h[i-1]+lambda[i]*(h[i]+h[i+1]));

> Fm[i]:=((lambda[i-1]*(h[i-3]+h[i-2])+h[i-1])*D[i]+
> lambda[i]*(h[i]+h[i+1])*D[i-1])/(lambda[i-1]*(h[i-3]+h[i-2])+
> h[i-1]+lambda[i]*(h[i]+h[i+1]));
> Fp[i]:=((h[i]+lambda[i+1]*(h[i+1]+h[i+2]))*D[i]+
```

```
> lambda[i]*(h[i-2]+h[i-1])*D[i+1])/(lambda[i]*(h[i-2]+h[i-1])+
> h[i]+lambda[i+1]*(h[i+1]+h[i+2]));

> Fm[i+1]:=((lambda[i]*(h[i-2]+h[i-1])+h[i])*D[i+1]+
> lambda[i+1]*(h[i+1]+h[i+2])*D[i])/(lambda[i]*(h[i-2]+h[i-1])+
> h[i]+lambda[i+1]*(h[i+1]+h[i+2]));
> Fp[i+1]:=((h[i+1]+lambda[i+2]*(h[i+2]+h[i+3]))*D[i+1]+
> lambda[i+1]*(h[i-1]+h[i])*D[i+2])/(lambda[i+1]*(h[i-1]+h[i])+
> h[i+1]+lambda[i+2]*(h[i+2]+h[i+3]));

> Fm[i+2]:=((lambda[i+1]*(h[i-1]+h[i])+h[i+1])*D[i+2]+
> lambda[i+2]*(h[i+2]+h[i+3])*D[i+1])/(lambda[i+1]*(h[i-1]+h[i])+
> h[i+1]+lambda[i+2]*(h[i+2]+h[i+3]));

> p[i-1]:=collect(((h[i-2]+h[i-1]+h[i])*Fm[i-1]+h[i-3]*Fp[i-1])/
> (h[i-3]+h[i-2]+h[i-1]+h[i]),[D[i-2],D[i-1],D[i]]);
> r[i-1]:=collect((h[i-3]*Fm[i-1]+(h[i-2]+h[i-1]+h[i])*Fp[i-1])/
> (h[i-3]+h[i-2]+h[i-1]+h[i]),[D[i-2],D[i-1],D[i]]);

> p[i]:=collect(((h[i-1]+h[i]+h[i+1])*Fm[i]+h[i-2]*Fp[i])/
> (h[i-2]+h[i-1]+h[i]+h[i+1]),[D[i-1],D[i],D[i+1]]);
> limit(p[i],lambda[i]=0);
> r[i]:=collect((h[i-2]*Fm[i]+(h[i-1]+h[i]+h[i+1])*Fp[i])/
> (h[i-2]+h[i-1]+h[i]+h[i+1]),[D[i-1],D[i],D[i+1]]);
> limit(r[i],lambda[i]=0);

> p[i+1]:=collect(((h[i]+h[i+1]+h[i+2])*Fm[i+1]+h[i-1]*Fp[i+1])/
> (h[i-1]+h[i]+h[i+1]+h[i+2]),[D[i],D[i+1],D[i+2]]);
> r[i+1]:=collect((h[i-1]*Fm[i+1]+(h[i]+h[i+1]+h[i+2])*Fp[i+1])/
> (h[i-1]+h[i]+h[i+1]+h[i+2]),[D[i],D[i+1],D[i+2]]);

> b[i-1,3]:=collect((delta[i]*h[i]*r[i-1]+
> (delta[i-1]*h[i-2]+(1-2*delta[i-1]-2*delta[i])*h[i-1])*p[i])/
> (delta[i-1]*h[i-2]+(1-2*delta[i-1]-2*delta[i])*h[i-1]+
> delta[i]*h[i]),
> [D[i-2],D[i-1],D[i],D[i+1]]);

> simplify(limit(b[i-1,3],lambda[i]=0,delta[i]=0));

> b[i,2]:=collect((((1-2*delta[i]-2*delta[i+1])*h[i]+
> delta[i+1]*h[i+1])*r[i]+
> delta[i]*h[i-1]*p[i+1])/
> (delta[i]*h[i-1]+(1-2*delta[i]-2*delta[i+1])*h[i]+
```

```
> delta[i+1]*h[i+1]),
> [D[i-1],D[i],D[i+1],D[i+2]]);
> simplify(limit(b[i,2],lambda[i]=0,delta[i]=0));

> q[i]:=collect((h[i]*p[i]+h[i-1]*r[i])/(h[i-1]+h[i]),
> [D[i-2],D[i-1],D[i],D[i+1],D[i+2]]);
> simplify(limit(q[i],lambda[i]=0,delta[i]=0));

> b[i-1,4]:=collect((h[i]*b[i-1,3]+h[i-1]*q[i])/
> (h[i-1]+h[i]),[D[i-2],D[i-1],D[i],D[i+1],D[i+2]]);
> simplify(limit(b[i-1,4],lambda[i]=0,delta[i]=0));

> b[i,1]:=collect((h[i]*q[i]+h[i-1]*b[i,2])/
> (h[i-1]+h[i]),[D[i-2],D[i-1],D[i],D[i+1],D[i+2]]);
> simplify(limit(b[i,1],lambda[i]=0,delta[i]=0));

> b[i,0]:=collect((h[i]*b[i-1,4]+h[i-1]*b[i,1])/
> (h[i-1]+h[i]),[D[i-2],D[i-1],D[i],D[i+1],D[i+2]]);

> row:=coeffs(b[i,0],D[i-2],D[i-1],D[i],D[i+1],D[i+2]);
> limit(row[1],lambda[i]=0,delta[i]=0);
> limit(row[2],lambda[i]=0,delta[i]=0);
> simplify(limit(row[3],lambda[i]=0,delta[i]=0));
> limit(row[4],lambda[i]=0,delta[i]=0);
> limit(row[5],lambda[i]=0,delta[i]=0);
```

REFERENCES

[1] S. Asaturyan – P. Costantini – C. Manni: *Local shape- preserving interpolation by space curves*, IMA J. Numer. Anal., **21** (2001), 301-325.

[2] P. Costantini P.: *Curve and surface construction using variable degree polynomial splines*, Computer Aided Geometric Design, **17** (2000), 419-446.

[3] P. Costantini – T.N.T. Goodman – C. Manni: *Constructing C^3 shape preserving interpolating space curves*, Adv. Comput. Math., **14** (2001), 103-127.

[4] P. Costantini – C. Manni: *Shape–preserving C^3 interpolation: the curve case*, Adv. Comput. Math., **18** (2003), 41-63.

[5] P. Costantini – C. Manni: *Geometric construction of spline curves with tension properties*, Computer Aided Geometric Design, **20** (2003), 579-599.

[6] P. Costantini – F. Pelosi: *Shape–preserving approximation by space curves*, Num. Alg., **27** (2001), 219-316.

[7] P. Costantini – F. Pelosi: *Shape–preserving approximation of spatial data*, Adv. Comput. Math., **20** (2004), 25-51.

[8] M. Eck – D. Lasser: *B-spline-Bézier representation of geometric spline curves: quartics and quintics*, Computers Math. Applic., **23** (1992), 23-39.

[9] T.N.T. Goodman: *Total positivity and the shape of curves*, in *Total Positivity and its Applications*, M. Gasca and C.A. Micchelli (eds.), Kluwer, Dordrecht, 1996, pp. 157-186.

[10] T.N.T. Goodman – B.H. Ong: *Shape preserving interpolation by space curves*, Computer Aided Geomatric Design, **15** (1997), 1-17.

[11] T.N.T. Goodman – B.H. Ong: *Shape preserving interpolation by G^2 curves in three dimensions*, in *Curves and Surfaces with Applications in CAGD*, A. Le Méhauté, C. Rabut and L.L. Schumaker (eds.) Vanderbilt University Press, 1997, pp. 151-158.

[12] T.N.T. Goodman – B.H. Ong – M.L. Sampoli: *Automatic interpolation by fair, shape preserving, G^2 space curves*, Computer Aided Design, **30** (1998), 813-822.

[13] J. Hoschek – D. Lasser: *Fundamentals of Computer Aided Geometric Design*, A.K. Peters Ldt, Wellesley, Massachusetts, 1993.

[14] B. Jüttler: *Shape preserving least-squares approximation by polynomial parametric spline curves*, Computer Aided Geometric Design, **14** (1997), 731-747.

[15] P.D. Kaklis – M.T. Karavelas: *Shape preserving interpolation in* $\mathbb{R}^3$, IMA J. Numer. Anal., **17** (1997), 373-419.

[16] M.I. Karavelas – P.D. Kaklis: *Spatial shape preserving interpolation using ν–splines*, Numer. Alg., **23** (2000), 217-250.

[17] V.P. Kong – B.H. Ong: *Shape preserving interpolation using Frenet frame continuous curve of order 3*, preprint, 2001.

[18] R. Morandi – D. Scaramelli – A. Sestini: *A geometric approach for knot selection in convexity-preserving spline approximation*, in *Curve and Surface Design: Saint-Malo 1999*, Pierre-Jean Laurent, Paul Sablonniere and L.L. Schumaker (eds.), Vanderbilt University Press, Nashville, TN, 2000, pp. 287-296.

[19] P. Sablonnière: *Spline and Bézier polygons associated with a polynomial spline curve*, Computer Aided Design, **10** (1978), 257-261.

[20] R. Sauer R.: *Differenzengeometrie*, Springer Verlag, Berlin, 1970.

Lavoro pervenuto alla redazione il 15 febbraio 2003
ed accettato per la pubblicazione il 16 dicembre 2003.
Bozze licenziate il 6 dicembre 2004

Indirizzo degli Autori:

Paolo Costantini – Dipartimento di Scienze Matematiche ed Informatiche "R. Magari" – Pian dei Mantellini 44, 53100 Siena, Italy
email: costantini@unisi.it

Carla Manni – Dipartimento di Matematica - Università di Roma "Tor Vergata" – Via della Ricerca Scientifica, 00133 Roma, Italy
email: manni@mat.uniroma2.it

Work supported by Università di Siena, under P.A.R. 2001.

Rendiconti di Matematica, Serie VII
Volume 24, Roma (2004), 281-301

Application of cardinal radial basis interpolation operators to numerical solution of the Poisson equation

GIAMPIETRO ALLASIA – ALESSANDRA DE ROSSI

Abstract: *We consider the application of a new scattered data approximation scheme to numerically solving the Dirichlet problem for the Poisson equation. This collocation method, which is mesh-free and substantially independent on the space dimension, makes use of interpolation operators with cardinal radial basis and differs from the well-known discretization approach introduced by E. J. Kansa in 1990 and then extensively developed, based on Hardy's multiquadrics or others radial basis functions. In our method the discretization matrix, whose dimension equals the number of internal points in the domain, is symmetric and strictly diagonally dominant, so that the discrete problem is well-posed and also well-conditioned, since the matrix condition number is small. Numerical experiments show that the performance of our method is comparable in many cases with that of Kansa's method; moreover, the former works well even if the number of collocation points is large.*

1 – Introduction

In early the 1990s E. J. Kansa [13], [14] proposed a method to solve hyperbolic, parabolic and elliptic partial differential equations using Hardy's multiquadric radial basis functions (MQ-RBFs). Then several authors extended Kansa's idea applying MQ-RBFs and other radial basis functions (RBFs) to the numerical solution of various types of PDEs (see recent developments in [16]).

Key Words and Phrases: *Elliptic partial differential equations – Approximation scheme – Scattered data – Cardinal radial basis.*
A.M.S. Classification: 65N35 – 65N22 – 65D10 – 41A63

Two useful features of Kansa's method immediately appeared: it gives a truly mesh-free algorithm and its computational complexity does not increase consistently with spatial dimension. Mesh-free radial basis functions provide powerful discretization representations to simulate problems not merely in bidimensional domains, but in arbitrary n-dimensional domains with irregular boundaries, and they can be implemented wisely on massively parallel computers. There are also some well known disadvantages: MQ-RBFs (and in general RBFs) are globally supported and the discretization matrix is full and ill-conditioned. Several remedies have been proposed to circumvent the ill-conditioning, but this is yet a relevant and open problem in the resolution of PDEs by RBFs (see, e.g. [15]).

A seemingly new scheme for discretization of the Poisson equation and some other elliptic PDEs, which recalls Kansa's method but differs in some noteworthy aspects, has been proposed [4]. This collocation method, which is mesh-free and substantially independent on the space dimension, makes use of interpolation operators with cardinal radial basis (CRBIs). A suitable choice of the basis function type yields a discretization matrix, that is symmetric and strictly diagonally dominant, and has a dimension equal to the number of collocation points internal to the domain. So the discrete problem is well-posed and also well-conditioned, since the matrix condition number is small.

Numerical experiments show that the method considered gives solutions whose accuracy is in many cases comparable with that achieved by the MQ-RBF method. Furthermore, it appears clear that our method does not stop working well even if the set of collocation points is large.

2 – The CRBI Scheme

In order to investigate computational properties of our scattered data approximation scheme, we start outlining it and, for simplicity, focusing on the consideration of the Dirichlet problem for the Poisson equation.

Let $\Omega \subset \mathbf{R}^s$, $(s \geq 2)$, be an open, simply connected point set, bounded by a piecewise regular hypersurface Σ, and let $h(x) \in C(\Omega \cup \Sigma)$ and $k(x) \in C(\Sigma)$ be given real functions with $x = (x_1, x_2, \dots, x_s)$. The Dirichlet problem for the Poisson equation is that of finding a real function $u(x)$ in the space $\mathcal{U} = \{u \in C^2(\Omega) \cap C(\Sigma)\}$ such that

$$(1) \quad -\Delta u(x) \equiv -\sum_{k=1}^{s} \frac{\partial^2 u(x)}{\partial x_k^2} = h(x), \quad x \in \Omega, \quad \text{and} \quad u(x) = k(x), \quad x \in \Sigma.$$

To build up the discrete problem associated with (1), we consider a set of distinct points $S_N = \{\xi_i,\ i = 1, \dots, N\}$, in general arbitrarily distributed in the domain Ω, and a suitable family of cardinal basis functions $g_k \in C^2(\Omega) \cap C(\Sigma)$, $(k = 1, \dots, N)$, or rather $g_k \in C^2(\Omega \cup \Sigma)$, such that

$$(2) \qquad g_k(\xi_i) = \delta_{ki},$$

where δ_{ki} is the Kronecker delta. Each element F of the set $\mathcal{F}_N = \text{span}\{g_k, k = 1, \dots, N\}$, which is a linear space of dimension N and a subset of $\mathcal{U}$, is uniquely represented in the form

$$F(x) = \sum_{k=1}^{N} c_k g_k(x). \tag{3}$$

An approximate solution F of the problem (1) must satisfy the linear system

$$-\sum_{k=1}^{N} c_k \Delta g_k(\xi_i) = h(\xi_i), \quad \xi_i \in \Omega, \quad \text{and} \quad \sum_{k=1}^{N} c_k g_k(\xi_i) = k(\xi_i), \quad \xi_i \in \Sigma,$$

that is for (2)

$$-\sum_{k=1}^{N} c_k \Delta g_k(\xi_i) = h(\xi_i), \quad \xi_i \in \Omega, \quad \text{and} \quad c_i = k(\xi_i), \quad \xi_i \in \Sigma. \tag{4}$$

If we suppose for $k = 1, \dots, N$

$$w_k \in C^2(\Omega \cup \Sigma), \qquad w_k(x) = \begin{cases} 0, & \text{for } x = \xi_i,\ i \neq k, \\ > 0, & \text{for } x = \xi_k, \end{cases}$$

then we can set

$$g_k(x) = \frac{w_k(x)}{\sum_{j=1}^{N} w_j(x)}, \tag{5}$$

and these $g_k(x)$ can be interpreted as the basis functions in (3). Properties of the interpolation operator $F(x)$ are discussed in detail in [1], [2].

3 – The Basic Weight

The values of the second derivative of $F(x)$ at the nodes, which are needed in (4), depend from (5) on the choice of the weight $w_k(x)$. In a first approach, we choose a simple and classical weight setting for $k = 1, \dots, N$, and $\xi_i = (\xi_{i1}, \xi_{i2}, \dots, \xi_{is})$

$$w_k(x) = \prod_{\substack{i=1, \\ i \neq k}}^{N} d^2(x, \xi_i), \quad \text{with} \quad d^2(x, \xi_i) = \sum_{j=1}^{s} (x_j - \xi_{ij})^2. \tag{6}$$

Consequently, from (5) $g_k(x)$ takes the form

$$g_k(x) = \frac{w_k(x)}{\sum_{j=1}^{N} w_j(x)} = \frac{\prod_{\substack{i=1,\\ i\neq k}}^{N} d^2(x,\xi_i)}{\sum_{j=1}^{N}\prod_{\substack{i=1,\\ i\neq j}}^{N} d^2(x,\xi_i)}. \tag{7}$$

However, different choices are possible and worthy of careful consideration, in view of their numerical performance in applications.

As a result of rather tedious algebraic manipulations detailed in [4], we get for $p = 1, \ldots, s$

$$\frac{\partial^2 F(\xi_i)}{\partial x_p^2} = \frac{1}{w_i(\xi_i)}\left(\sum_{\substack{m=1,\\ m\neq i}}^{N} c_m \frac{\partial^2 w_m(\xi_i)}{\partial x_p^2} - c_i \sum_{\substack{k=1,\\ k\neq i}}^{N} \frac{\partial^2 w_k(\xi_i)}{\partial x_p^2} \right), \tag{8}$$

where

$$w_m(\xi_i) = \prod_{\substack{j=1,\\ j\neq m}}^{N} d^2(\xi_i,\xi_j), \quad \text{if} \quad m = i; \quad w_m(\xi_i) = 0, \quad \text{if} \quad m \neq i; \tag{9}$$

and

$$\begin{aligned} \frac{\partial^2 w_m(\xi_i)}{\partial x_p^2} &= \sum_{\substack{k=1,\\ k\neq m}}^{N} \left[2 \prod_{\substack{j=1,\\ j\neq m,k}}^{N} d^2(\xi_i,\xi_j) + 4(\xi_{ip} - \xi_{kp}) \sum_{\substack{h=1,\\ h\neq m,k}}^{N} (\xi_{ip} - \xi_{hp}) \times \right. \\ &\qquad \left. \times \prod_{\substack{j=1,\\ j\neq m,k,h}}^{N} d^2(\xi_i,\xi_j) \right], \text{ if } m = i; \\ \frac{\partial^2 w_m(\xi_i)}{\partial x_p^2} &= 2 \prod_{\substack{j=1,\\ j\neq m,i}}^{N} d^2(\xi_i,\xi_j), \text{ if } m \neq i. \end{aligned} \tag{10}$$

Note that

$$\frac{\partial^2 w_m(\xi_i)}{\partial x_p^2} = \frac{\partial^2 w_m(\xi_i)}{\partial x_1^2}, \qquad p = 1, \ldots, s.$$

From (4) and (8) we have that the condition to be satisfied by $F(x)$ at any node $\xi_i \in \Omega$ is given by

$$\frac{-1}{w_i(\xi_i)}\left\{ \sum_{\substack{m=1,\\ m\neq i}}^{N} c_m \; s \; \frac{\partial^2 w_m(\xi_i)}{\partial x_1^2} - c_i \sum_{\substack{k=1,\\ k\neq i}}^{N} s \; \frac{\partial^2 w_k(\xi_i)}{\partial x_1^2} \right\} = h(\xi_i),$$

whereas at any node $\xi_i \in \Sigma$ the Dirichlet condition yields

$$F(\xi_i) = c_i = k(\xi_i).$$

Hence, the system of linear equations $Ac = b$, obtained by discretization of the Dirichlet problem, is

$$\sum_{m=1}^{N} a_{im} c_m = b_i, \qquad i = 1, \dots, N, \tag{11}$$

where

(a) for $\xi_i \in \Omega$: $b_i = h(\xi_i)$ and

$$a_{im} = \begin{cases} -s \dfrac{1}{w_i(\xi_i)} \dfrac{\partial^2 w_m(\xi_i)}{\partial x_1^2}, & \text{if } m \neq i, \\ s \dfrac{1}{w_i(\xi_i)} \displaystyle\sum_{\substack{k=1, \\ k \neq i}}^{N} \dfrac{\partial^2 w_k(\xi_i)}{\partial x_1^2}, & \text{if } m = i; \end{cases} \tag{12}$$

(b) for $\xi_i \in \Sigma$: $b_i = k(\xi_i)$ and $a_{im} = \delta_{im}$.

If A is nonsingular, the solution of the $N \times N$ system $Ac = b$ is a vector c whose components $c_i, (i = 1, \dots, N)$, approximate the quantities appearing in the expression (3) of the interpolation operator F.

4 – Properties of Discretization Scheme

If N_1 nodes belong to Ω and N_2 to Σ, with $N = N_1 + N_2$, it is convenient to order the equations of the system (11) such that the first equations correspond to the first N_1 nodes. Thus, taking into account the point (b) above, the system can be rewritten as

$$\sum_{m=1}^{N_1} a_{im} c_m + \sum_{n=N_1+1}^{N} a_{in} c_n = h(\xi_i), \qquad \text{for} \quad i = 1, \dots, N_1,$$
$$c_j = k(\xi_j), \qquad \text{for} \quad j = N_1 + 1, \dots, N,$$

which is equivalent to

$$\sum_{m=1}^{N_1} a_{im} c_m = h(\xi_i) - \sum_{n=N_1+1}^{N} a_{in} k(\xi_n), \qquad \text{for} \quad i = 1, \dots, N_1.$$

This relation shows that the initial $N \times N$ system $Ac = b$ can be reduced to an $N_1 \times N_1$ system $\tilde{A}\tilde{c} = \tilde{b}$, namely

$$\sum_{m=1}^{N_1} \tilde{a}_{im}\tilde{c}_m = \tilde{b}_i, \qquad \text{for} \quad i = 1, \dots, N_1, \tag{13}$$

where

$$\begin{aligned} \tilde{a}_{im} &= a_{im}, \qquad \text{for} \quad i, m = 1, \dots, N_1, \\ \tilde{c}_m &= c_m, \qquad \text{for} \quad m = 1, \dots, N_1, \\ \tilde{b}_i &= h(\xi_i) - \sum_{n=N_1+1}^{N} a_{in}k(\xi_n), \qquad \text{for} \quad i = 1, \dots, N_1. \end{aligned} \tag{14}$$

We note that the dimension of the system depends only from the N_1 internal nodes. Nevertheless, the N_2 nodes on the boundary increase a little the computational effort to obtain the terms $\tilde{b}_i$ in (14).

For $i, m = 1, \dots, N_1$, $i \neq m$,

$$\tilde{a}_{im} = \frac{-s}{w_i(\xi_i)} \frac{\partial^2 w_m(\xi_i)}{\partial x_1^2} = \frac{-s}{\prod\limits_{\substack{j=1,\\ j\neq i}}^{N} d^2(\xi_i, \xi_j)} \; 2 \prod_{\substack{j=1,\\ j\neq m,i}}^{N} d^2(\xi_i, \xi_j) = \frac{-2s}{d^2(\xi_i, \xi_m)}, \tag{15}$$

which shows the symmetry of $\tilde{A}$.

The matrix $\tilde{A}$ is strictly diagonally dominant. In fact, the expression of the ith diagonal element $\tilde{a}_{ii}$ of $\tilde{A}$ can be rewritten by (9) and (10), in the form

$$\tilde{a}_{ii} = \frac{1}{\prod\limits_{\substack{j=1,\\ j\neq i}}^{N} d^2(\xi_i, \xi_j)} \sum_{\substack{k=1,\\ k\neq i}}^{N} 2s \prod_{\substack{j=1,\\ j\neq k,i}}^{N} d^2(\xi_i, \xi_j) = 2s \sum_{\substack{k=1,\\ k\neq i}}^{N} \frac{1}{d^2(\xi_i, \xi_k)}. \tag{16}$$

On the other hand, the sum of entries in the ith row, omitting the diagonal term, is

$$\sum_{k=1,k\neq i}^{N_1} \tilde{a}_{ik} = -\frac{\sum\limits_{k=1,k\neq i}^{N_1} 2s \prod\limits_{j=1,j\neq k,i}^{N} d^2(\xi_i, \xi_j)}{\prod\limits_{j=1,j\neq i}^{N} d^2(\xi_i, \xi_j)} = -2s \sum_{k=1,k\neq i}^{N_1} \frac{1}{d^2(\xi_i, \xi_k)}. \tag{17}$$

Comparing the quantities in (16) and (17), we have the inequalities

$$|\tilde{a}_{ii}| > \sum_{k=1,k\neq i}^{N_1} |\tilde{a}_{ik}| \qquad \text{for all} \quad i = 1, \dots, N_1,$$

because we sum N terms in (16) but only the first N_1 of them in (17).

Since a strictly diagonally dominant matrix is nonsingular, the system $\tilde{A}\tilde{c} = \tilde{b}$ has a unique solution.

5 – A More Localizing Weight

A crucial point in the considered scheme is the choice of the weight. Actually, the approximation accuracy obtained by using (6) is unsatisfactory, at least considering small domains, since the root mean square error (RMSE) and the maximum absolute error (MAE) are quite high in comparison with those arising from approximation with MQs. The reason is to be searched in the behavior of the basis function $g_k(x)$ in (7), which is not sufficiently localyzing. In fact, $g_k(x)$ can be rewritten as

$$g_k(x) = \begin{cases} \dfrac{1/d^2(x,\xi_k)}{\displaystyle\sum_{j=1}^{N} 1/d^2(x,\xi_j)}, & \text{if } x \neq \xi_k, \\ 1, & \text{if } x = \xi_k, \end{cases}$$

that shows the connection between the behaviors of $g_k(x)$ and

$$\phi(d^2(x,\xi_k)) = 1/d^2(x,\xi_k).$$

Now, the latter is too sensitive to the effects of any node ξ_k relatively far from the interpolation point x and, in particular, it happens when the considered distances are less than one. From a practical viewpoint, it is easier to consider the function $\phi(t) = 1/t^2$ instead of $\phi(d^2(x,\xi_k))$.

As a matter of fact, if the number N of nodes is large in comparison with the domain extension, it may be deemed expedient to strongly localize the weight so that more distant points do not work. A possible solution would consist in introducing in (3) localizing factors with reduced compact supports, but in such a way as to conserve the continuity of the second derivatives of the weights. Otherwise, one could consider beside (6) other weights which are strongly decaying as distance increases. Up to now, we have mainly explored functions of the second type.

A weight to be considered to improve the scheme is

$$\hat{w}_m(x) = \prod_{\substack{j=1,\\ j\neq m}}^{N} \{\exp[\alpha d^2(x,\xi_j)] - 1\}, \qquad \alpha \geq 1, \tag{18}$$

related to the function $\hat{\phi}(t) = 1/[\exp(\alpha t^2) - 1]$, whose localizing effect increases with α. The discretization scheme by using the weight (18) still enjoys all the properties discussed above. In particular, relations (12) (remembering (15) and (16)) become

$$a_{im} = \begin{cases} -\dfrac{1}{\hat{w}_i(\xi_i)}\, s\, \dfrac{\partial^2 \hat{w}_m(\xi_i)}{\partial x_1^2} = \dfrac{-2s\alpha}{d^2(\xi_i,\xi_m)}\,, & \text{if } m \neq i, \\ \dfrac{1}{\hat{w}_i(\xi_i)} \displaystyle\sum_{\substack{k=1,\\ k\neq i}}^{N} s\, \dfrac{\partial^2 \hat{w}_k(\xi_i)}{\partial x_1^2} = 2s\alpha \displaystyle\sum_{\substack{k=1,\\ k\neq i}}^{N} \dfrac{1}{d^2(\xi_i,\xi_k)}\,, & \text{if } m = i. \end{cases}$$

It must be noted that a drawback of using parameters in the weights is due to the requirement of determining their optimal values.

6 – Solving the Linear System

Since the system matrix $\tilde{A}$ is strictly diagonally dominant, Gaussian elimination algorithm can be applied to solve the system (13) without row or column interchanges, and the computations are stable with respect to the growth of roundoff errors [[12], pp. 181-182]. Since all the reduced matrices $\tilde{A}^{(k)}$ given by the algorithm are symmetric, the amount of work for the decomposition is approximately halved, that is, $O(n^3/6)$ multiplications/divisions and as many additions/subtractions are required.

Gaussian elimination ensures the factorization $\tilde{A} = LU$, where L is a lower triangular matrix with ones on the main diagonal and U is an upper triangular matrix. Other factorization methods can also be considered as the factorization LDL^T, where L is lower triangular with ones on its diagonal and D is the diagonal matrix with $a_{11}, a_{22}, \ldots, a_{nn}$ on its diagonal, and, since the matrix is positive definite, Choleski's factorization LL^T, where L has positive diagonal elements. They require computational efforts of the same order as Gaussian elimination, but the last is more easily handled in order to apply the iterative refinement [11], [9], [12].

The numerical stability allows to handle large systems and, in case of very large systems, Gaussian elimination can be efficiently performed in a parallel processing environment [10]. This feature is very important when the space dimension s is larger than two.

For the direct solution of linear systems we used subroutines given in LAPACK, because it represents a standard benchmark and enjoys an excellent documentation. Moreover, accompanying LAPACK is the set of lower-level operations called BLAS which has to be considered for implementation on both shared and local memory multiprocessors. Up to now, however, considering equations in two variables, we have done only serial computations.

7 – Smoothing Resulting Surfaces

After solving the system $\tilde{A}\tilde{c} = \tilde{b}$ and obtaining the values of all the coefficients c_m, $(m = 1, \dots N)$, in the expression of the approximation operator $F(x)$, one can proceed to obtain those values of $F(x)$ which are of interest or, rather, to give an approximate representation (possibly graphical) of the solution of the Dirichlet problem.

A computational problem arises from the very definition of $F(x)$ which is from (3) and (7)

$$F(x) = \sum_{k=1}^{N} c_k g_k(x) = \sum_{k=1}^{N} c_k \frac{\prod\limits_{\substack{i=1,\\ i\neq k}}^{N} d^2(x, \xi_i)}{\sum\limits_{j=1}^{n} \prod\limits_{\substack{i=1,\\ i\neq j}}^{N} d^2(x, \xi_j)}. \tag{19}$$

The question is if the product form (19) of the operator achieves more numerical stability than the equivalent barycentric form

$$F(x) = \begin{cases} \sum\limits_{k=1}^{N} c_k \dfrac{1/d^2(x, \xi_k)}{\sum\limits_{j=1}^{N} 1/d^2(x, \xi_j)}, & \text{if } x \neq \xi_k, \\ c_k, & \text{if } x = \xi_k, \end{cases} \tag{20}$$

or viceversa. Schneider and Werner [18] at first observe that the operator in one dimension may be considered for $p = 2$ a rational Hermite interpolation, i.e., the derivative is approximated (rather arbitrarily) by 0 at the data sites. Then they state that the barycentric form offers the advantage of a remarkable numerical stability: even in the presence of rounding errors, which may occur during the computation of c_k, the interpolation property is maintained.

As a matter of fact, in the practice of numerical calculation, the barycentric formula is definitely preferred (see more considerations in [1]). Obviously suitable tricks must be adopted to control the growth of rounding and truncation errors when x is very close to ξ_k.

In the graphical representation of $F(x)$ a drawback is given by the appearance of flat spots at the data points, since the partial derivatives of $F(x)$ vanish there (see [3]).

To avoid this generally undesirable property, we construct local approximants $Q_k(x)$ to $F(x)$ at ξ_k, obtained by means of the moving weighted least-squares method using weight functions with reduced compact support. These local approximations are to be used instead of the coefficients c_k in order to get a smoother surface. So the operator $F(x)$ is expressed as a convex combination of the local approximants

$$F(x) = \sum_{k=1}^{N} Q_k(x) \; \frac{w_k(x)}{\sum_{j=1}^{N} w_j(x)}.$$

Best performance is achieved by using for every node ξ_k a paraboloid $Q_k(x)$ which interpolates at the node.

8 – Boundary Effects

A common feature in all RBF approximations is how relatively inaccurate they are at boundaries. This accuracy degradation near boundaries in many cases severely limits the utility of methods based on RBFs. Actually, large boundary-induced errors of this type will contaminate less or more the solution everywhere across the domain [8]. In applying MQ-RBFs to the solution of PDEs the residual error is typically largest by one or two orders near the boundary compared to the residual error in the domain far from the boundary. An improvement has been proposed which consists in adding an additional set of nodes, lying inside or outside of the domain, and correspondly in adding an additional set of collocation equations obtained via collocation of the PDE on the boundary [7]. This PDE collocation on the boundary reduces dramatically the residual.

Our scheme allows to increase considerably the number of nodes collocated on the boundary, and this is done in a simple way with a very reduced computational cost. This feature could be particularly useful to control the difficulties possibly arising near the domain boundary.

9 – Numerical Results

In the following test examples we restrict ourselves to two dimensional Poisson and Laplace problems whose analytic solution are available. In all cases we use both MQ-RBF and CRBI approximations of the unknown function u. In particular CRBI method is applied both with the basic weight (6) and the localizing exponential weight (18).

EXAMPLE 1. We consider the Poisson problem studied in [6]

$$\begin{aligned} \Delta u(x,y) &= f(x,y), \qquad (x,y) \in \Omega, \\ u(x,y) &= g(x,y), \qquad (x,y) \in \Sigma, \end{aligned}$$

where $\Omega \cup \Sigma = [1,2] \times [1,2]$, the inhomogeneous term $f(x,y)$ is given by

$$\begin{aligned} f(x,y) = -\frac{751\pi^2}{144} \sin\frac{\pi x}{6} \sin\frac{7\pi x}{4} \sin\frac{3\pi y}{4} \sin\frac{5\pi y}{4} + \frac{7\pi^2}{12} \cos\frac{\pi x}{6} \cos\frac{7\pi x}{4} \times \\ \times \sin\frac{3\pi y}{4} \sin\frac{5\pi y}{4} + \frac{15\pi^2}{8} \sin\frac{\pi x}{6} \sin\frac{7\pi x}{4} \cos\frac{3\pi y}{4} \cos\frac{5\pi y}{4}, \end{aligned}$$

and

$$g(x,y) = \sin\frac{\pi x}{6} \sin\frac{7\pi x}{4} \sin\frac{3\pi y}{4} \sin\frac{5\pi y}{4}.$$

The exact solution $u(x,y)$ to this problem coincides with the boundary condition $g(x,y)$.

For this test we selected various uniform distributions of collocation points in the domain $[1,2] \times [1,2]$. Fig. 1 shows a uniform distribution on a grid of 81 collocation points. We solved the above problem using the MQ-RBF method with a shape parameter $c = 1$ and the CRBI method. The resulting algebraic systems were solved using Gauss elimination. Due to uncertainty of how to choose the values of the parameters for RBFs and CRBIs, we do not looked for

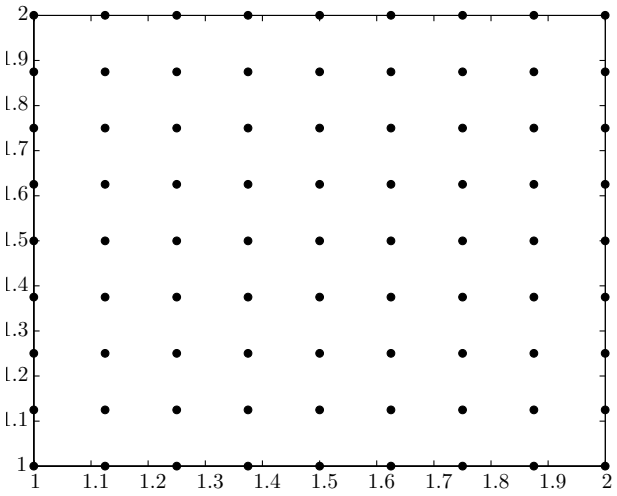

Fig. 1: Uniform distribution of 81 collocation points.

their optimal values. Simply, in this preliminary investigation, we set $c = 1$ as in [17] and α roughly proportional to the number of collocation points.

In Table 1 are listed the dimensions and the approximate condition numbers of discretization matrices. The subscripts K, C and Ce refer to Kansa's scheme and to CRBI based schemes with the two different weights, respectively.

TABLE 1: *Matrix dimensions* (*MD*) *and condition numbers* (*CN*) *for some uniform distributions.*

N	$\mathbf{6\times 6}$	$\mathbf{8\times 8}$	$\mathbf{10\times 10}$	$\mathbf{12\times 12}$
MD_K	36	64	100	144
CN_K	9.6026e07	1.8021e11	2.8824e14	5.7559e17
MD_C	16	36	64	100
CN_C	3.2483e00	4.6815e00	6.1136e00	7.5405e00
MD_{Ce}	16	36	64	100
CN_{Ce}	8.2780e00	1.5853e01	2.8020e01	4.2069e01

Table 1 shows the well-known disadvantage of ill-conditioning of the discretization matrices arising from the MQ-RBFs method and, on the contrary, how the matrices for the CRBI method have much smaller condition numbers. Moreover, as already pointed-out, the matrix dimension in CRBI approach is equal to the number of internal collocation points in the domain.

TABLE 2: *Root mean square errors* (*RMSE*) *and maximum absolute errors* (*MAE*) *for some uniform distributions.*

N	$\mathbf{6\times 6}$	$\mathbf{9\times 9}$	$\mathbf{11\times 11}$	$\mathbf{16\times 16}$	$\mathbf{21\times 21}$
$RMSE_K$	5.4416e−3	2.8910e−4	4.4902e−5	9.3432e−7	1.7773e−5
MAE_K	1.2663e−2	8.8215e−4	1.6874e−4	2.9001e−6	5.2733e−5
$RMSE_C$	8.5317e−2	1.1795e−1	1.3046e−1	1.5005e−1	1.6212e−1
MAE_C	3.3857e−1	4.5457e−1	4.7875e−1	5.1510e−1	5.4705e−1
$RMSE_{Ce}$	6.8051e−3	2.8701e−3	1.8761e−3	8.5708e−4	4.9346e−4
MAE_{Ce}	1.7341e−2	7.1686e−3	4.8149e−3	2.0704e−3	1.3317e−3

In order to test accuracy, we increased the number of collocation points considering several uniform grids in the domain. In Table 2 we list some of the results obtained. The root mean square errors and the absolute maximum errors, computed on the set of the collocation points, are better for the MQ-RBF method, but they begin to make worse as the number of collocation points increases starting from the 41×41 grid. For example, considering the uniform 51×51 grid (2601 collocation points) the RMSE goes down to 2.5025e−3 and the MAE to 5.7033e−3.

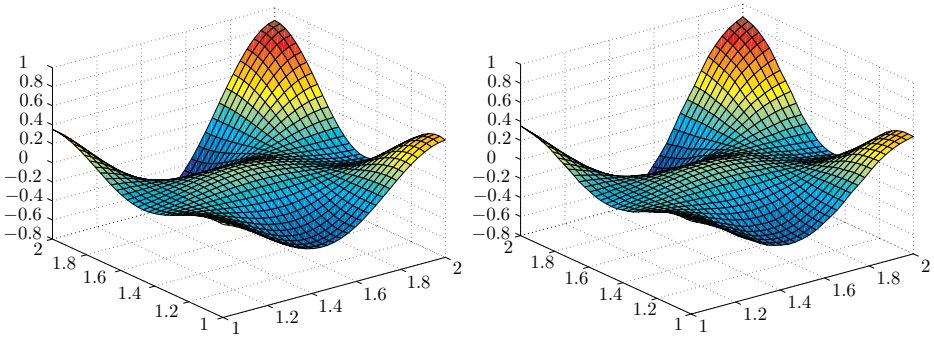

Fig. 2: The numerical solutions on the 33×33 grid obtained with MQ-RBF method and CRBI method with localizing weight using 81 collocation points.

Concerning the CRBI method, we observe that in this test example it is not accurate as MQ-RBF method, but it slowly improves when we use the exponential localizing weight. Instead, the accuracy of the numerical solution of CRBI with the basic weight is not satisfying. We must say that for this test problem is not necessary to consider a large set of collocation points (as show the plots in Fig. 2 and Fig. 3) but these remarks are important if we will consider a problem which require a large number of points to be numerically solved.

Fig. 2 shows the approximations obtained with the MQ-RBF method and the CRBI method with the localizing weight. We used 81 collocation points and the approximations are computed on a 33×33 grid. Considering the accuracy of the numerical solutions, the plot of the exact solution is not given. In Fig. 3 are

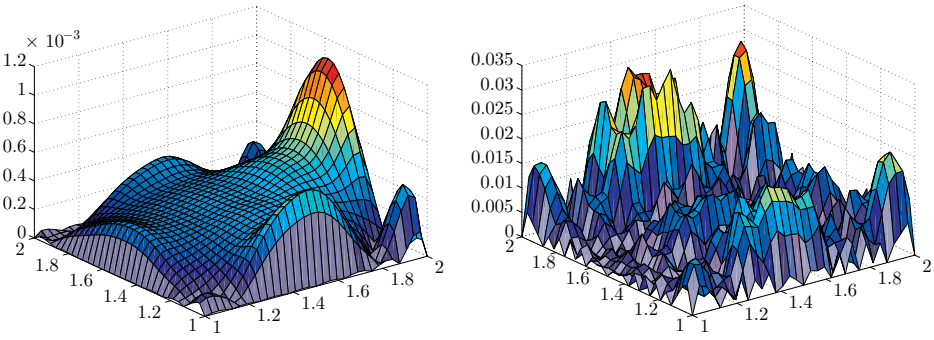

Fig. 3: The absolute errors computed on the 33×33 grid obtained with MQ-RBF method and CRBI method with localizing weight.

plotted the absolute errors in the two cases. Comparing the plots, we observe that the error is more uniformly distributed in the domain when CRBI method is used, while the error is more localized on the boundary for MQ-RBF method.

Example 2. Let us consider the Laplace equation which models the steady state temperature distribution in a thin plate [5]

$$\Delta u(x,y) = 0, \qquad (x,y) \in (0,1) \times (0,1)$$

with the Dirichlet boundary conditions

$$u(x,0) = 1, \quad u(x,1) = 0, \quad u(0,y) = 0, \quad u(1,y) = 0.$$

The analytic solution of this problem is given by

$$u(x,y) = \Big(\frac{4}{\pi}\Big) \sum_{i=0}^{\infty} \Big\{ \Big[\frac{1}{2i+1}\Big] \sin[(2i+1)\pi x] \times \\ \times \sinh[(1-y)(2i+1)\pi] \cdot \cosh[(2i+1)\pi] \Big\}.$$

For this test we selected various gridded and scattered data sets in the domain $[0,1] \times [0,1]$. Some of the data sets used are plotted in Fig. 4 and Fig. 5. The scattered data sets were generated randomly, selecting some of the points on the boundary.

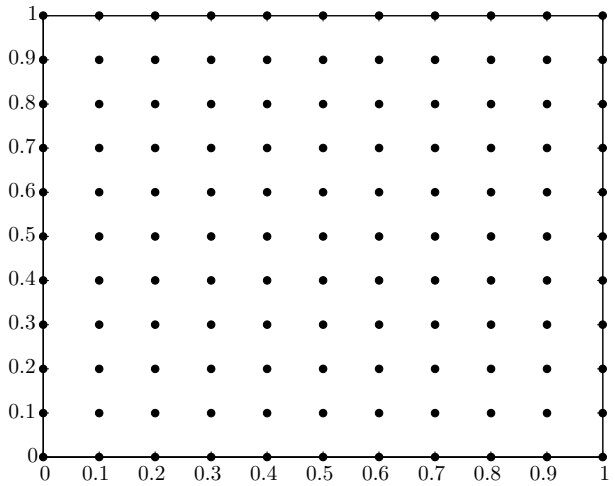

Fig. 4: Gridded data set of 121 collocation points.

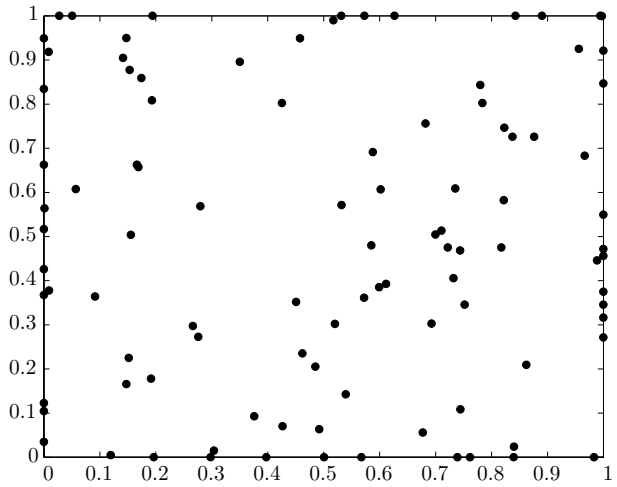

Fig. 5: Scattered data set of 100 collocation points.

In Table 3 we report the matrix dimensions and the approximated condition numbers for some uniform distributions. We remark again that the matrices arising from our approximation scheme are smaller and better conditioned in comparison with those from MQ-RBF scheme.

TABLE 3: *Matrix dimensions* (*MD*) *and condition numbers* (*CN*) *for some gridded data sets.*

N	$\mathbf{5 \times 5}$	$\mathbf{7 \times 7}$	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$
MD_K	25	49	81	121
CN_K	1.9153e06	4.0658e09	7.0456e12	7.4703e15
MD_C	9	25	49	81
CN_C	2.5342e00	3.9645e00	5.3981e00	6.8281e00
MD_{Ce}	9	25	49	81
CN_{Ce}	5.8284e00	1.3922e01	2.5233e01	3.9137e01

Tables 4 and 5 show the values of the root mean square and absolute maximum errors computed on the nodes, obtained with gridded data sets and scattered data sets, respectively. The number of boundary data points of the scattered data sets is indicated in brackets. Both tables show that in this example the CRBI method performs better than the MQ-RBF one. In particular, increasing the number of collocation points the errors increase for the latter, while on the contrary our scheme slowly improves.

TABLE 4: *Root mean square errors* (*RMSE*) *and absolute maximum errors* (*MAE*) *for some gridded data sets.*

N	$\mathbf{6\times 6}$	$\mathbf{8\times 8}$	$\mathbf{10\times 10}$	$\mathbf{12\times 12}$	$\mathbf{14\times 14}$
$\mathrm{RMSE_K}$	1.4017e−2	1.6773e−2	1.9210e−2	2.6067e−2	7.4839e−2
$\mathrm{MAE_K}$	3.3644e−2	4.7334e−2	5.8919e−2	9.4331e−2	3.5157e−1
$\mathrm{RMSE_C}$	6.5439e−2	8.3554e−2	9.8267e−2	1.0984e−1	1.1944e−1
$\mathrm{MAE_C}$	2.1083e−1	2.6841e−1	3.0426e−1	3.3161e−1	3.5143e−1
$\mathrm{RMSE_{Ce}}$	3.2226e−3	2.5500e−3	2.0659e−3	1.6933e−3	1.5749e−3
$\mathrm{MAE_{Ce}}$	1.0400e−2	7.5742e−3	6.6763e−3	6.4897e−3	5.2316e−3

TABLE 5: *Root mean square errors* (*RMSE*) *and absolute maximum errors* (*MAE*) *for some scattered data sets.*

$\mathbf{N(N_2)}$	**60**(**28**)	**100**(**36**)	**140**(**44**)	**240**(**60**)	**400**(**80**)
$\mathrm{RMSE_K}$	6.2720e−1	2.3253e−2	1.6883e+0	1.0566e−1	3.6538e−1
$\mathrm{MAE_K}$	2.0397e+0	1.3569e−1	1.8115e+1	7.7725e−1	9.8808e−1
$\mathrm{RMSE_C}$	6.3918e−2	8.8239e−2	1.1258e−1	1.2086e−1	1.2542e−1
$\mathrm{MAE_C}$	2.6953e−1	4.7216e−1	3.6631e−1	3.8096e−1	5.2064e−1
$\mathrm{RMSE_{Ce}}$	4.7033e−2	5.3307e−2	5.3559e−2	5.8081e−2	5.6451e−2
$\mathrm{MAE_{Ce}}$	1.7072e−1	3.8215e−1	1.5142e−1	4.6622e−1	3.7112e−1

Fig. 6 shows the plot of the exact solution. In Fig. 7 and Fig. 8 are plotted the numerical approximations obtained with the MQ-RBF method and the CRBI

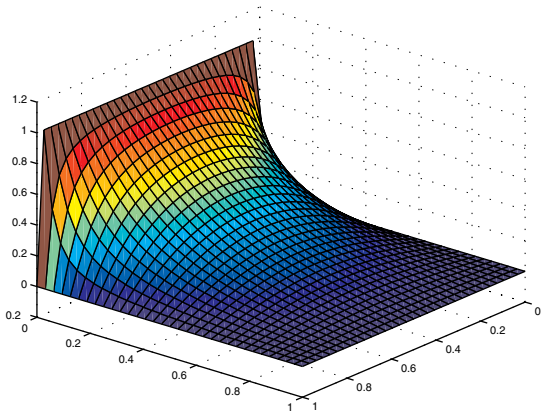

Fig. 6: Exact solution computed on the 26×26 grid.

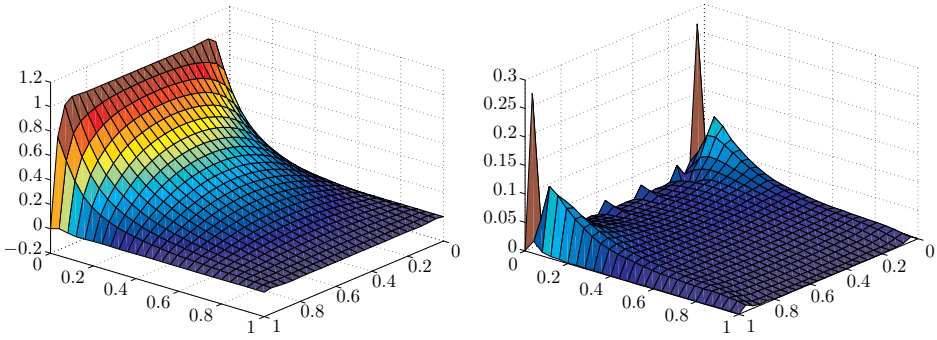

Fig. 7: Approximation with the MQ-RBF method using 121 collocation points and absolute error.

method with localizing weight using 121 collocation points and the related absolute errors computed on a 26×26 grid. We remark that, in comparison with the solution of the Poisson equation in Example 1, this solution has a behaviour difficult to be captured near the boundaries, where the values of the solution are prescribed equal to zero and one by the Dirichlet conditions. A comparison between the two plots points out that the CRBI method is more accurate near the boundaries where the solution is constrained to zero, while the MQ-RBF method is better approximating near the boundary $y = 0$, when the number of collocation points used is equal for the two scheme.

To improve the performance of the CRBI method we increase the number of collocation points. Fig. 9 shows the approximation obtained with 441 points and absolute errors. Extending data set of collocation points is justified by the properties of the discretization scheme and by the numerical stability.

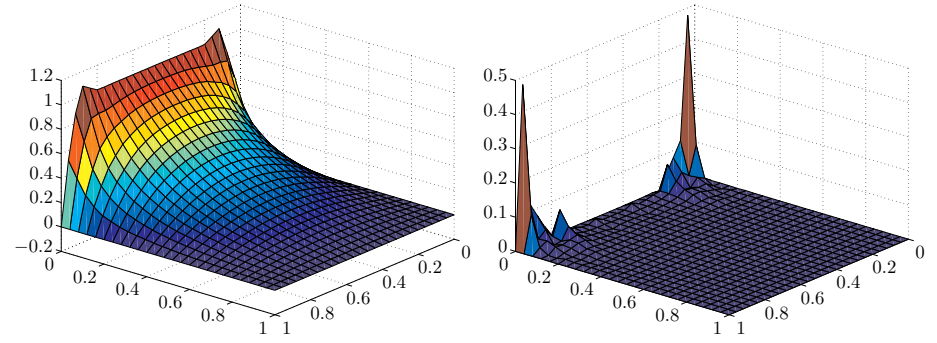

Fig. 8: Approximation with the CRBI method with localizing weight using 121 collocation points and absolute error.

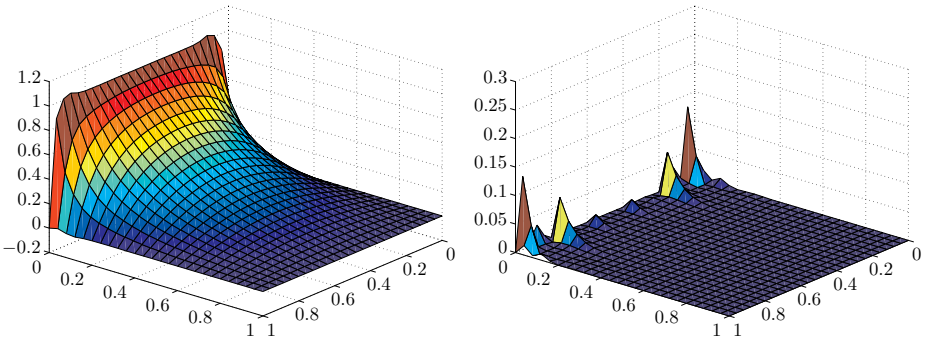

Fig. 9: Approximation with the CRBI method with the localizing weight using 441 collocation points and absolute error.

In order to test accuracy near the boundaries, we computed the local relative error $(u_t - u_n)/u_t$ where u_t and u_n are the analytical and numerical values, respectively, at the internal collocation points.

Fig. 10 and Fig. 11 show the relative error between the analytical and the numerical solutions when we used 81 uniformly distributed collocation points.

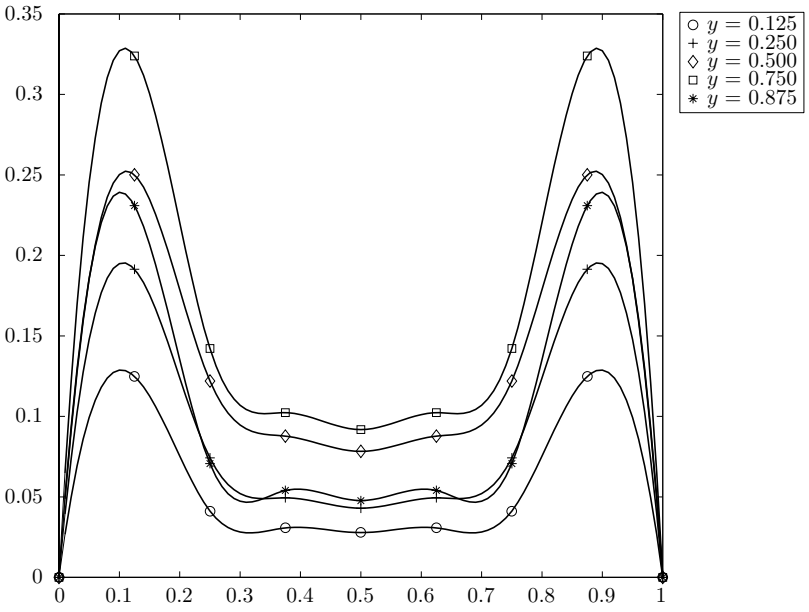

Fig. 10: Relative error obtained with the MQ-RBF method on the uniform distribution of 81 collocation points.

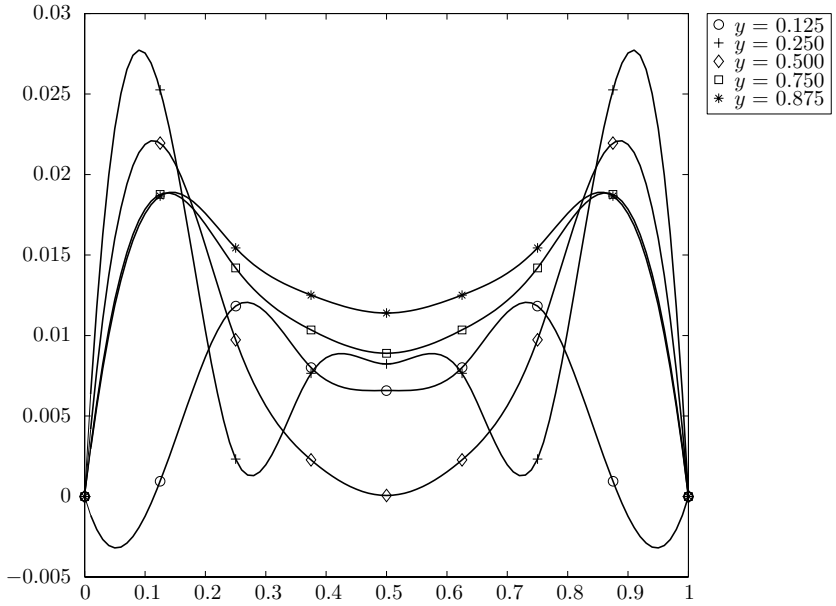

Fig. 11: Relative error obtained with the CRBI method with the localizing weight on the uniform distribution of 81 collocation points.

The numerical solutions were computed with MQ-RBF and the CRBI with localizing weight, respectively. In Fig. 10 we observe that the maximum relative errors are on the lines $y = 0.500$, $y = 0.750$, $y = 0.875$ and near the boundaries where the solution values were prescribed equal to zero; vice versa they decrease in the middle regions. Fig. 11 shows as the approximation with the CRBI method has a different behaviour. In fact the maximum relative error is more uniformly distributed in the domain. As already pointed-out in the literature on functional approximation of scattered data, the MQ function seems more appropriate to approximate rapidly varying functions, while CRBI method performs better when approximating slowly varying functions.

10 – Conclusions

Based on a theoretical establishment, a new method has been constructed to give a numerical solution to the Poisson equation. The method, which makes use of Cardinal Radial Basis Intepolation operators (CRBI), enjoys the following special features:

1. It is well-posed and numerically stable.
2. The discretization matrix is symmetric and strictly diagonally dominant (hence, positive definite).
3. Both error and sensitivity are reasonably small, that is, the method is not affected from the so-called uncertainty principle.
4. Being mesh-free and insensitive to dimension, it is particularly suitable for irregular domains and 3D problems.
5. Boundary effects can be controlled by increasing the number of collocation points on the boundary at the cost of a little computational effort.

Obviously, there are some factors affecting the accuracy of the proposed method:

1. First derivatives of the approximation operator at internal collocation points vanish so that flat spots appear on the rendered surface.
2. In general, when applied on a reduced number of collocation points (of the order of tens or hundreds), it does not work as well as MQ-RBF method.
3. The choice of the weight and the determination of optimal values of the possible parameters play important roles in the accuracy.
4. Numerically the method converges slowly and the rate of convergence has not yet been investigated theoretically.

Additional theoretical and numerical characteristics of the method as well as the application in solving other type of PDEs are currently under investigation by our research group.

REFERENCES

[1] G. Allasia: *A class of interpolating positive linear operators: theoretical and computational aspects*, in: *Approximation Theory, Wavelets and Applications*, S. P. Singh (ed.), Kluwer, Dordrecht, 1995, 1-36.

[2] G. Allasia: *Cardinal basis interpolation on multivariate scattered data*, Nonlinear Analysis Forum, (1) **6** (2001), 1-13.

[3] G. Allasia: *Simultaneous interpolation and approximation by a class of multivariate positive operators*, Numer. Algorithms, **34** (2003), 147-158.

[4] G. Allasia: *A scattered data approximation scheme for the multidimensional Poisson equation by cardinal radial basis interpolants*, Curve and surface fitting (Saint-Malo, 2002), Nashboro Press, Brentwood, TN, 2003, 11-20.

[5] H. S. Carslaw – J. C. Jaeger: *Conduction of heat in solids*, 2nd ed., Clarendon Press, Oxford, 1959.

[6] C. S. Chen – M. Ganesh – M. A. Golberg – A. H.-D. Cheng: *Multilevel compact radial functions based computational schemes for some elliptic problems*, Computers Math. Applic., **43** (2002), 359-378.

[7] A. I. Fedoseyev – M. J. Friedman – E. J. Kansa: *Improved multiquadric method for elliptic partial differential equations via PDE collocation on the boundary*, in: *Radial Basis Functions and Partial Differential Equations*, E. J. Kansa – Y. C. Hon (eds.), Computers Math. Applic., (3-5) **43** (2002), 439-455.

[8] B. Fornberg – T. A. Driscoll – G. Wright – R. Charles: *Observations on the behavior of radial basis function approximations near boundaries*, in: *Radial Basis Functions and Partial Differential Equations*, E. J. Kansa – Y. C. Hon (eds.), Computers Math. Applic., (3-5) **43** (2002), 473-490.

[9] G. H. Forsythe – C. B. Moler: *Computer solution of linear algebraic systems*, Prentice-Hall, Englewood Cliffs, 1967.

[10] T. L. Freeman – C. Phillips: *Parallel numerical algorithms*, Prentice Hall, U. K., 1992.

[11] A. George – J. W. Liu: *Computer solution of large sparse definite positive systems*, Prentice-Hall, Englewood Cliffs, 1981.

[12] N. Higham: *Accuracy and stability of numerical algorithms*, SIAM, Philadelphia, 1996.

[13] E. J. Kansa: *Multiquadrics-A scattered data approximation scheme with applications to computational fluid dynamics-I: Surface approximation and partial derivatives estimates*, Computers Math. Applic., (8/9) **19** (1990), 127-145.

[14] E. J. Kansa: *Multiquadrics-A scattered data approximation scheme with applications to computational fluid dynamics-II: Solution to parabolic, hyperbolic and elliptic partial differential equations*, Computers Math. Applic., (8/9) **19** (1990), 147-161.

[15] E. J. Kansa – Y. C. Hon: *Circumventing the ill-conditioning problem with multiquadric radial basis functions: Applications to elliptic partial differential equations*, Computers Math. Applic., **39** (2000), 123-137.

[16] E. J. Kansa – Y. C. Hon (eds.): *Radial Basis Functions and Partial Differential Equations*, Computers Math. Applic., (3-5) **43** (2002), 275-619.

[17] H. Power – V. Barraco: *A comparison analysis between unsymmetric and symmetric radial basis function collocation methods for the numerical solution of partial differential equations*, in: *Radial basis functions and partial differential equations*, E. J. Kansa – Y. C. Hon (eds.), Computers Math. Applic., (3) **43** (2002), 551-583.

[18] C. Schneider – W. Werner: *Hermite interpolation: the barycentric approach*, Computing, **46** (1991), 35-51.

Lavoro pervenuto alla redazione il 15 febbraio 2003
ed accettato per la pubblicazione il 20 dicembre 2003.
Bozze licenziate il 6 dicembre 2004

Indirizzo degli Autori:

Giampietro Allasia – Alessandra De Rossi – Department of Mathematics – University of Turin – Via Carlo Alberto 10 – 10123 Torino – Italy
E-mail: giampietro.allasia@unito.it alessandra.derossi@unito.it

Rendiconti di Matematica, Serie VII
Volume 24, Roma (2004), 303-320

A perturbative method for direct scattering problems

NADANIELA EGIDI – PIERLUIGI MAPONI

Abstract: *We present a numerical method to compute the solution of direct scattering problems, that is boundary-value problems for the Helmholtz equation in unbounded domains of the three dimensional real Euclidean space. Such problems arise, for example, from wave equation problems when the solution is assumed to be time-harmonic. We consider the T-matrix method for the solution of the direct scattering problems, which is a very classical numerical method for such a kind of problems. This method is based on the explicit construction of an operator T mapping the data of the problem to the solution of the problem. We propose a perturbative approach for the numerical approximation of the operator T. Finally we report the results of our numerical experience on a large number of test problems using the numerical method proposed here. This numerical experience shows very interesting results and it justifies further theoretical investigations.*

1 – Introduction

Let us begin with some basic definitions. Let $\mathbb{N}$, $\mathbb{R}$, $\mathbb{C}$ be the set of natural numbers, real numbers and complex numbers, respectively. Let $n \in \mathbb{N}$, we denote with $\mathbb{R}^n$, $\mathbb{C}^n$ the n-dimensional real Euclidean space and the n-dimensional complex Euclidean space, respectively. We denote with $(\cdot,\cdot)$ the Euclidean scalar product in $\mathbb{R}^n$, with $\|\cdot\|$ the corresponding Euclidean norm. Let $S^n = \{\underline{x} \in \mathbb{R}^{n+1} : \|\underline{x}\| = 1\}$. Let $\imath$ be the imaginary unit. Let $z \in \mathbb{C}$, we denote with $|z|$ the modulus of z and with $\mathrm{Re}(z)$, $\mathrm{Im}(z)$ the real and imaginary part of z respectively.

Key Words and Phrases: *Acoustic scattering – T-matrix method – Perturbative method.*
A.M.S. Classification: 65N35 – 65-04 – 35P25 – 35J05

Let $D \subset \mathbb{R}^3$ be a bounded simply connected open set with boundary ∂D and let $\overline{D}$ be its closure. We suppose that $\underline{0} \in D$. From the physical point of view we consider D as the position of an obstacle, or equivalently a scatterer, for the acoustic waves propagating in $\mathbb{R}^3 \setminus D$, in particular we suppose this scatterer contained in a homogeneous isotropic medium filling $\mathbb{R}^3 \setminus D$. Moreover for such a medium we suppose a constant pressure field P. Let $U^i(\underline{x}, t)$, $\underline{x} \in \mathbb{R}^3$, $t \in \mathbb{R}$ be an incident acoustic wave, where $\underline{x} \in \mathbb{R}^3$ denotes the space variables and $t \in \mathbb{R}$ denotes the time variable. Let $U^s(\underline{x}, t)$, $\underline{x} \in \mathbb{R}^3 \setminus D$, $t \in \mathbb{R}$ be the scattered acoustic wave generated by the interaction of U^i and the obstacle D. These waves can be considered as perturbations for the pressure field P; when such perturbations are small compared to P we have that U^i and U^s solve the wave equation, see [1], page 243 for details.

We suppose that U^i and U^s are time-harmonic, that is:

$$U^i(\underline{x}, t) = u^i(\underline{x}) e^{i\omega t}, \ \underline{x} \in \mathbb{R}^3, \ t \in \mathbb{R}, \tag{1}$$

$$U^s(\underline{x}, t) = u^s(\underline{x}) e^{i\omega t}, \ \underline{x} \in \mathbb{R}^3 \setminus D, \ t \in \mathbb{R}, \tag{2}$$

where u^i, u^s are suitable functions of the space variables and $\omega > 0$ is the time-frequency.

From the wave equation for U^i, U^s and from formulas (1), (2) we obtain the Helmholtz equation for u^i and u^s, that is

$$\Delta u^i(\underline{x}) + k^2 u^i(\underline{x}) = 0, \ \underline{x} \in \mathbb{R}^3, \tag{3}$$

$$\Delta u^s(\underline{x}) + k^2 u^s(\underline{x}) = 0, \ \underline{x} \in \mathbb{R}^3 \setminus \overline{D}, \tag{4}$$

where Δ is the Laplace operator with respect to the $\underline{x}$ variables, $k = \frac{\omega}{c} > 0$ is the wave number and $c > 0$ is the wave propagation velocity. We assume that D is an impenetrable acoustically soft obstacle, so that u^s satisfies the following boundary conditions:

$$u^s(\underline{x}) = -u^i(\underline{x}), \ \underline{x} \in \partial D, \tag{5}$$

see [2] page 67 for details. We note that impenetrable acoustically hard obstacles satisfy Neumann boundary condition, and obstacles having more complicated acoustic behaviour satisfy a boundary condition that can be given in terms of an acoustic surface impendance. Moreover we assume that the scattered acoustic wave u^s has the asymptotic behaviour of an outgoing spherical wave, so that u^s satisfies the Sommerfeld radiation condition, that is

$$\frac{\partial u^s}{\partial \hat{\underline{x}}}(\underline{x}) - \imath k u^s(\underline{x}) = o\left(\frac{1}{\|\underline{x}\|}\right), \ \|\underline{x}\| \to \infty, \tag{6}$$

where $\hat{\underline{x}} = \frac{\underline{x}}{\|\underline{x}\|} \in S^2$, $\|\underline{x}\| \neq 0$ and $o(\cdot)$ is the Landau symbol, see [3] page 189 for a more detailed discussion on the radiation condition.

Boundary-value problem (4)-(6) is uniquely solvable provided u^i in (5) is a continuous function and D is a class C^2 domain with connected complement, see [2] page 83, [4] pages 13 and 262 for details. Let us consider the following problem: from the knowledge of D, k and u^i compute the solution u^s of problem (4)-(6).

We consider the numerical approximation of such a problem. Many different methods for the solution of problem (4)-(6), or similar scattering problems, have been proposed in the scientific literature, see for example [5], [6], [7], for finite difference approaches or [8], [9], [10], [11] for finite element approaches. We note that these general purpose methods cannot be applied directly to problem (4)-(6) being this problem defined on an unbounded domain. A quite common technique to avoid this difficulty is to consider this problem in a domain $\mathcal{D} \setminus D$, where $\mathcal{D} \subset \mathbb{R}^3$ is a bounded open domain containing $\overline{D}$, and to substitute the Sommerfeld radiation condition with an auxiliary condition on the artificial boundary $\partial\mathcal{D}$. This condition is usually called transparent boundary condition or absorbing boundary condition, see [12] and the references therein. However some specialized numerical methods allow to deal with the unbounded domain of problem (4)-(6), see for example [2], [13], [14], [15], [16], [17], [18], [19], [20], [21] for integral equation approaches and [22], [23], [24], [25], [26], [27] for T-matrix approaches.

We study the T-matrix method which is a very classical method for the solution of scattering problems. This method consists in the construction of an operator $T = T(D, k)$, depending only on D and k, such that:

$$u^s = T\, u^i \tag{7}$$

for every continuous function $u^i : \partial D \to \mathbb{C}$. Usually functions u^i and u^s are expanded with respect to particular bases of functions defined in terms of the spherical harmonics, so that the operator T looks like a matrix with an infinite number of rows and an infinite number of columns. In practical situations we consider only a finite number of entries of T, whose computation foresees the solution of several linear systems where the entries of the coefficient matrix are obtained by the evaluation of several surface integrals on ∂D. We denote with $Q = Q(D, k)$ the matrix coefficient of this linear system. Usually Q is a dense matrix and, depending on D and k, it can be quite ill-conditioned, so that the solution of the corresponding linear system can produce a large error in the final solution. Moreover having in mind an efficient implementation of this method via parallel computations the step of the solution of such a linear system is an unpleasant step since it lowers considerably the parallel efficiency of the whole method.

To avoid a linear system solution in the T-matrix method we propose a perturbative method for the computation of the operator T, where the pertur-

bation is made with respect to the boundary ∂D of the obstacle D. As base point of this perturbation is considered the boundary ∂B of a generic obstacle B; in such a case for the construction of the operator T we have to solve several linear systems where the matrix coefficient is $Q(B,k)$. So that also in the perturbative method we really have to solve some linear systems, but now the matrix coefficient $Q(B,k)$ can be chosen in terms of B. We note that when the base point B is chosen as an axial-symmetric obstacle the matrix $Q(B,k)$, arising in the construction of the operator T, has a particular block-structure; when B is chosen as a sphere the matrix $Q(B,k)$ is a diagonal matrix, so that the solution of the corresponding linear system can be performed accurately, quickly and efficiently in a sequential computation as well as in a parallel computation. However in general we can compute, for example, the LU factorization of the matrix $Q(B,k)$ and we can use this factorization everytime the boundary ∂B of B is used as base point in the perturbative procedure.

Finally we report some of the results of our numerical experience obtained using the numerical method proposed here. We consider a large number of test problems, where we take into account axial-symmetric and non-axial-symmetric obstacles, convex and non-convex obstacles. In the numerical results convergence and stabilization features of the perturbative method proposed are outlined. This numerical experience shown very interesting results, so that we deserve further theoretical investigations to this introductory study.

The paper is organized as follows. In Section 2 we provide a brief introduction to the T-matrix method and we give some useful formulas for the development of the method proposed here. In Section 3 we present the perturbative method. In Section 4 we report some results of our numerical experience using the method presented in the previous section. In Section 5 we give some conclusions and the possible developments of the work.

2 – The T-matrix method

The construction of the operator T is usually given in terms of suitable bases of functions for the expansion of u^i, i.e. the datum of problem (4)-(6), and u^s, i.e. the unknown solution of problem (4)-(6). We denote with:

$$(8)\quad Y^{\sigma}_{l,m}(\underline{\hat{x}}(\theta,\phi)) = \gamma_{l,m}\begin{cases} P^m_l(\cos\theta)\cos(m\phi), & \sigma = 0,\ l = 0,1,\dots,\ m = 0,1,\dots,l, \\ P^m_l(\cos\theta)\sin(m\phi), & \sigma = 1,\ l = 1,2,\dots,\ m = 1,2,\dots,l, \end{cases}$$

the spherical harmonics, where $\underline{\hat{x}}(\theta,\phi) = (\sin\theta\cos\phi,\ \sin\theta\sin\phi,\ \cos\theta)^t \in S^2$, $\theta \in [0,\pi]$, $\phi \in [0,2\pi)$, and for $l = 0,1,\dots$, $m = 0,1,\dots,l$, P^m_l are the Legendre functions of order m and degree l and $\gamma_{l,m}$ are normalization coefficients, that is we have:

$$(9)\qquad \int_{S^2} (Y^{\sigma}_{l,m}(\underline{\hat{x}}))^2 ds(\underline{\hat{x}}) = 1,$$

where ds is the surface measure on S^2, see [28] page 331 for details. In the sequel we denote with ν the multi-index $(\sigma,\, l,\, m)$ and we denote with I the set of all possible values of ν given by formula (8), i.e. $I = \{\nu = (\sigma,\, l,\, m)\colon\ \sigma = 0, 1,\ l = \sigma, \sigma+1, \ldots,\ m = \sigma, \sigma+1, \ldots, l\}$. We note that the spherical harmonics verify an orthogonality property, that can be seen as a generalization of property (9), that is we have:

$$\int_{S^2} Y^{\sigma}_{l,m}(\hat{\underline{x}}) Y^{\sigma'}_{l',m'}(\hat{\underline{x}}) ds(\hat{\underline{x}}) = \delta_{\sigma,\sigma'}\delta_{l,l'}\delta_{m,m'}, \quad \nu,\ \nu' = (\sigma', l', m') \in I, \tag{10}$$

where δ denotes the kronecker delta.

In the construction of the operator T, introduced in (7), we use two bases of functions $\{\psi_\nu,\, \nu \in I\}$, $\{\mathrm{Re}\,\psi_\nu,\, \nu \in I\}$, which are defined as follows:

$$\psi_\nu(k\underline{x}) = h^{(1)}_l(k\,\|\underline{x}\|) Y^{\sigma}_{l,m}(\hat{\underline{x}}),\ \underline{x} \in \mathbb{R}^3 \setminus \{\underline{0}\},\ \nu \in I, \tag{11}$$

$$\mathrm{Re}\,\psi_\nu(k\underline{x}) = j_l(k\,\|\underline{x}\|) Y^{\sigma}_{l,m}(\hat{\underline{x}}),\ \underline{x} \in \mathbb{R}^3,\ \nu \in I, \tag{12}$$

where j_l denotes the spherical Bessel function of order l, $h^{(1)}_l$ denotes the spherical Hankel function of first kind and order l, see [28] page 435 for details. We note that for each $\nu \in I$ the complex-valued function ψ_ν is singular at the origin of the coordinate system, while the real-valued function $\mathrm{Re}\,\psi_\nu$ is regular at the origin of the coordinate system. Moreover from the separation of the Helmholtz operator in spherical coordinates it is easy to see that, for each $\nu \in I$, function $\mathrm{Re}\,\psi_\nu$ satisfies the Helmholtz equation in $\mathbb{R}^3$, function ψ_ν satisfies equation (4), being $\underline{0} \in D$, and it satisfies equation (6), for a detailed discussion see [29] page 1462.

Supposing that the functions u^i and u^s have the following expansion:

$$u^i(\underline{x}) = \sum_{\nu \in I} a_\nu\, \mathrm{Re}\,\psi_\nu(k\underline{x}),\ \underline{x} \in \mathbb{R}^3, \tag{13}$$

$$u^s(\underline{x}) = \sum_{\nu \in I} f_\nu \psi_\nu(k\underline{x}),\ \underline{x} \in \mathbb{R}^3 \setminus \overline{D}, \tag{14}$$

we obtain that the operator $T = T_{\nu;\nu'}(D,k)$, $\nu, \nu' \in I$, depending on the obstacle D and the wave number k, can be rewritten in a more practical way than formula (7), that is

$$f_\nu = \sum_{\nu' \in I} T_{\nu;\nu'}(D,k) a_{\nu'},\ \nu \in I. \tag{15}$$

We briefly recall the formulas useful for the computation of operator T. Let us define the following operator:

$$
\begin{aligned}
Q_{\nu;\nu'}(D,k) = &-\frac{\imath}{2}\delta_{\sigma,\sigma'}\delta_{l,l'}\delta_{m,m'} + \\
&+ \frac{k}{2}\int_{\partial D}\frac{\partial}{\partial\hat{\underline{n}}(\underline{x})}(\psi_\nu(k\underline{x})\,\mathrm{Re}\,\psi_{\nu'}(k\underline{x}))d\sigma(\underline{x}),\ \nu,\nu' \in I,
\end{aligned}
\tag{16}
$$

where $\hat{\underline{n}}(\underline{x})$ denotes the unit outward normal to ∂D at the point $\underline{x} \in \partial D$ and $d\sigma$ denotes the surface measure on ∂D. Let $\mathrm{Re}\,Q_{\nu;\nu'}(D,k) = \mathrm{Re}(Q_{\nu;\nu'}(D,k))$, $\nu,\ \nu' \in I$. The operator T is defined as the solution of the following equation:

$$
\sum_{\nu'\in I} T_{\nu;\nu'}(D,k)\,Q_{\nu';\nu''}(D,k) = -\,\mathrm{Re}\,Q_{\nu;\nu''}(D,k),\ \nu,\nu'' \in I.
\tag{17}
$$

Formula (16) and equation (17) are the results of simple but quite involved mathematical manipulations, which are mainly based on a representation formula for the solutions of the Helmholtz equation and on an expansion formula, with respect to the bases $\{\psi_\nu,\ \nu \in I\}$, $\{\mathrm{Re}\,\psi_\nu,\ \nu \in I\}$, of the free space Green's function of the Helmholtz operator with the Sommerfeld radiation condition at infinity, see [22] for a complete derivation of these formulas.

We note that in practical situations we consider only a finite number of elements for the operators Q and T previously defined. Given $L_{\max} \in \mathbb{N}$ we define the following finite set of multi-indices $I_{L_{\max}} = \{\nu = (\sigma,\ l,\ m)\colon\ \sigma = 0,1,\ l = \sigma,\sigma+1,\dots,L_{\max},\ m = \sigma,\sigma+1,\dots,l\}$ and in (16), (17) we consider $I_{L_{\max}}$ in place of I. So that, in particular, from (17) we have:

$$
\sum_{\nu'\in I_{L_{\max}}} T_{\nu;\nu'}(D,k)\,Q_{\nu';\nu''}(D,k) = -\,\mathrm{Re}\,Q_{\nu;\nu''}(D,k),\ \nu,\nu'' \in I_{L_{\max}}.
\tag{18}
$$

We abuse the notations Q and T for the matrices obtained from the corresponding operators. We note that the rows of matrix T can be computed as solutions of the linear system (18), where we have multiple right-hand sides, that is each row of matrix T corresponds to a different row of matrix $\mathrm{Re}\,Q$ through linear system (18).

We note that the T-matrix method is an interesting technique to solve problem (4)-(6), in fact matrix T depends only on D and k. Thus once matrix T is computed the solution of problem (4)-(6) can be easily obtained from formulas (14), (15) for every different incident acoustic wave u^i using the same matrix T.

Usually in problem (4)-(6) is considered an acoustic plane wave as the incident acoustic wave u^i, that is

$$
u^i(\underline{x}) = e^{\imath k(\underline{x},\underline{\alpha})},\ \underline{x} \in \partial D,
\tag{19}
$$

where $\underline{\alpha} \in S^2$ is the wave propagation direction. We note that function u^i given in (19) is a solution of equation (3) for every $\underline{\alpha} \in S^2$. When the choice (19) is made the expansion (13) can be given explicitely, that is we have:

$$e^{\imath k(\underline{x},\underline{\alpha})} = 4\pi \sum_{\nu \in I} i^l Y^{\sigma}_{l,m}(\underline{\alpha}) \operatorname{Re} \psi_\nu(k\underline{x}),\ \underline{x} \in \mathbb{R}^3,\ \underline{\alpha} \in S^2,\ k > 0, \tag{20}$$

see [29] page 1466 for details.

3 – The perturbative method

For the computation of matrix T we must perform two different steps: (i) computation of the entries of matrix Q using formula (16), (ii) solution of the linear system (18). Step (i) can be performed accurately and efficiently using parallel computations, in fact it consists in the approximation of several integrals that are independent one from the other. On the contrary step (ii) must be performed with special care since the ill-conditioning of the matrix Q can make the computation of matrix T not well accurate. We note that the condition number of Q depends on D, k and the value chosen for the truncation parameter $L_{\max}$. Moreover we note that step (ii) is not well suited for parallel computations, being the solution of a linear system with, in general, a dense matrix coefficient.

We propose a perturbative method to avoid the solution of the linear system (18). We note that similar perturbative techniques have been already used for the solution of Fredholm integral equations of the first kind that formulate problem (4)-(6), or similar problems. In such cases it has been noted that perturbative techniques take care of the ill-posedness of the corresponding problem, solving the difficulty of the problem at the various perturbative orders, see for example [13], [14], [15], [16], [17], [21].

We limit our discussion to star-like obstacles, that is we suppose there exists a function $r : S^2 \to \mathbb{R}$, such that:

$$\partial D = \{\underline{x} \in \mathbb{R}^3 : \underline{x} = r(\underline{\hat{x}})\underline{\hat{x}},\ \underline{\hat{x}} \in S^2\}, \tag{21}$$

so that from (16) we have that the entries of matrix Q can be rewritten as follows:

$$\begin{aligned} Q_{\nu;\nu'}(D,k) = &-\frac{\imath}{2}\delta_{\sigma,\sigma'}\delta_{l,l'}\delta_{m,m'} + \\ &+ \frac{1}{2}\int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta \Bigg(\rho^2 \frac{d(j_{l'}(\rho)h^{(1)}_l(\rho))}{d\rho} Y^{\sigma}_{l,m}(\underline{\hat{x}}(\theta,\phi)) Y^{\sigma'}_{l',m'}(\underline{\hat{x}}(\theta,\phi)) + \\ &- \frac{\partial\rho}{\partial\theta} j_{l'}(\rho) h^{(1)}_l(\rho) \frac{\partial}{\partial\theta} \left(Y^{\sigma}_{l,m}(\underline{\hat{x}}(\theta,\phi)) Y^{\sigma'}_{l',m'}(\underline{\hat{x}}(\theta,\phi)) \right) + \\ &- \frac{1}{\sin^2\theta} \frac{\partial\rho}{\partial\phi} j_{l'}(\rho) h^{(1)}_l(\rho) \frac{\partial}{\partial\phi} \left(Y^{\sigma}_{l,m}(\underline{\hat{x}}(\theta,\phi)) Y^{\sigma'}_{l',m'}(\underline{\hat{x}}(\theta,\phi)) \right) \Bigg), \quad \nu,\nu' \in I_{L_{\max}}, \end{aligned} \tag{22}$$

where $\rho(\underline{\hat{x}}) = kr(\underline{\hat{x}})$, $\underline{\hat{x}} \in S^2$.

In the perturbative approach we consider the obstacle D as a perturbation of a given obstacle B, where we suppose that there exists a function $r_B : S^2 \to \mathbb{R}$, such that:

$$\partial B = \{\underline{x} \in \mathbb{R}^3 : \underline{x} = r_B(\hat{\underline{x}})\hat{\underline{x}},\ \hat{\underline{x}} \in S^2\}. \tag{23}$$

Let $\epsilon \in \mathbb{R}$ with $0 \leq \epsilon \leq 1$, let

$$R(\hat{\underline{x}}, \epsilon) = r_B(\hat{\underline{x}}) + \epsilon H(\hat{\underline{x}}),\ \hat{\underline{x}} \in S^2, \tag{24}$$

where $H(\hat{\underline{x}}) = r(\hat{\underline{x}}) - r_B(\hat{\underline{x}})$, $\hat{\underline{x}} \in S^2$. Let D_ϵ be the star-like obstacle having boundary ∂D_ϵ parametrized by the function $R(\cdot, \epsilon)$. We note that $r_B(\hat{\underline{x}}) = R(\hat{\underline{x}}, 0)$, $r(\hat{\underline{x}}) = R(\hat{\underline{x}}, 1)$, $\hat{\underline{x}} \in S^2$, so that we have $B = D_0$ and $D = D_1$ and a similar relation for the matrices Q defined in (22), that is $Q(B, k) = Q(D_0, k)$, $Q(D, k) = Q(D_1, k)$.

Now, given $N \in \mathbb{N}$, we consider the approximation of $Q(D_\epsilon, k)$ given by a series in powers of ϵ truncated to the order N-th, that is:

$$Q(D_\epsilon, k) \approx Q^{(0)} + Q^{(1)}\epsilon + \cdots + \frac{1}{N!}Q^{(N)}\epsilon^N,\ 0 \leq \epsilon \leq 1, \tag{25}$$

where, for $n = 0, 1, \ldots, N$, $Q^{(n)}$ denotes the formal derivative of order n-th of $Q(D_\epsilon, k)$ with respect to ϵ and evaluated at $\epsilon = 0$. Moreover for the matrix $T(D_\epsilon, k)$ we suppose a similar approximation, that is

$$T(D_\epsilon, k) \approx T^{(0)} + T^{(1)}\epsilon + \cdots + \frac{1}{N!}T^{(N)}\epsilon^N,\ 0 \leq \epsilon \leq 1, \tag{26}$$

where $T^{(n)}$, $n = 0, 1, \ldots, N$ are suitable square matrices having the same order as of matrix T. So that substituting the approximations (25), (26) in equation (18) we obtain:

$$\begin{aligned} &\left(T^{(0)} + T^{(1)}\epsilon + \cdots + \frac{1}{N!}T^{(N)}\epsilon^N\right)\left(Q^{(0)} + Q^{(1)}\epsilon + \cdots + \frac{1}{N!}Q^{(N)}\epsilon^N\right) = \\ &\quad = -\operatorname{Re}\left(Q^{(0)} + Q^{(1)}\epsilon + \cdots + \frac{1}{N!}Q^{(N)}\epsilon^N\right), \end{aligned} \tag{27}$$

which is an equation for matrices $T^{(n)}$, $n = 0, 1, \ldots, N$. Solving this equation order by order with respect to the powers of ϵ, for matrices $T^{(n)}$, $n = 0, 1, \ldots, N$ we obtain the following expression:

$$\begin{aligned} T^{(0)} &= -\operatorname{Re}(Q^{(0)})(Q^{(0)})^{-1} \\ T^{(n)} &= -\left(\operatorname{Re}(Q^{(n)}) + \sum_{l=1}^{n}\binom{n}{l}T^{(n-l)}Q^{(l)}\right)(Q^{(0)})^{-1},\ n = 1, 2, \ldots, N, \end{aligned} \tag{28}$$

where $\binom{n}{l} = \frac{n!}{(n-l)!l!}$, $n, l \in \mathbb{N}$, $l \leq n$, is the binomial coefficient. Formula (28) gives an explicit expression for $T^{(n)}$, $n = 0, 1, \ldots, N$. More precisely, from the knowledge of $Q^{(0)}$ we can compute matrix $T^{(0)}$, then from the knowledge of $Q^{(0)}$, $Q^{(1)}$ and $T^{(0)}$ we can compute matrix $T^{(1)}$; we can compute the generic matrix $T^{(n)}$ from the knowledge of $Q^{(0)}$, $Q^{(1)}, \ldots, Q^{(n)}$ and $T^{(0)}$, $T^{(1)}, \ldots, T^{(n-1)}$ computed previously. The approximation of matrix $T(D, k)$ is obtained evaluating in $\epsilon = 1$ the truncated power series given in formula (26), where the matrices $T^{(n)}$, $n = 0, 1, \ldots, N$ are computed by formula (28) as explained.

Let us consider the computation of matrices $Q^{(n)}$, $n = 0, 1, \ldots, N$. The entries of these matrices are given by the derivatives of order n with respect to ϵ of the corresponding entries of matrix $Q(D_\epsilon, k)$ and these derivatives are evaluated at $\epsilon = 0$, that is

$$Q^{(0)}_{\nu;\nu'} = Q_{\nu;\nu'}(D_\epsilon, k)\Big|_{\epsilon=0}, \ \nu, \nu' \in I_{L_{\max}}, \tag{29}$$

$$Q^{(n)}_{\nu;\nu'} = \frac{d^n}{d\epsilon^n} Q_{\nu;\nu'}(D_\epsilon, k)\Big|_{\epsilon=0}, \ \nu, \nu' \in I_{L_{\max}}, \ n \geq 1. \tag{30}$$

When the differentiation operator with respect to ϵ can be exchanged with the integral operators appearing in the expression of $Q(D_\epsilon, k)$ and when also the limit as $\epsilon \to 0$ can be exchanged with these integral operators we obtain a more practical expression for the entries of matrices $Q^{(n)}$, $n = 0, 1, \ldots, N$, in fact we have:

$$Q^{(0)} = Q(B, k), \tag{31}$$

$$\begin{aligned} Q^{(n)}_{\nu;\nu'} = \frac{1}{2}\int_0^{2\pi} d\phi \int_0^{\pi} d\theta \sin\theta \frac{d^n}{d\epsilon^n}\Bigg(& \eta^2 \frac{d(j_{l'}(\eta)h^{(1)}_l(\eta))}{d\eta} Y^{\sigma}_{l,m}(\underline{\hat{x}}(\theta,\phi)) Y^{\sigma'}_{l',m'}(\underline{\hat{x}}(\theta,\phi)) + \\ & - \frac{\partial\eta}{\partial\theta} j_{l'}(\eta)h^{(1)}_l(\eta)\frac{\partial}{\partial\theta}\left(Y^{\sigma}_{l,m}(\underline{\hat{x}}(\theta,\phi)) Y^{\sigma'}_{l',m'}(\underline{\hat{x}}(\theta,\phi))\right) + \\ & - \frac{1}{\sin^2\theta}\frac{\partial\eta}{\partial\phi} j_{l'}(\eta)h^{(1)}_l(\eta)\frac{\partial}{\partial\phi}\left(Y^{\sigma}_{l,m}(\underline{\hat{x}}(\theta,\phi)) Y^{\sigma'}_{l',m'}(\underline{\hat{x}}(\theta,\phi))\right)\Bigg)\Bigg|_{\epsilon=0}, \\ & \nu, \nu' \in I_{L_{\max}}, \ n \geq 1, \end{aligned} \tag{32}$$

where $\eta = kR(\cdot, \epsilon)$, $0 \leq \epsilon \leq 1$, is the unique function in (32) that depends on ϵ. From formulas (31), (32) we can easily seen that the computation of the entries of matrices $Q^{(n)}$, $n = 0, 1, \ldots, N$ can be performed accurately and efficiently by a parallel computation being these entries defined as integrals of functions that are independent one from the other. But now also the computation of matrices $T^{(n)}$, $n = 0, 1, \ldots, N$ can be performed accurately, in fact from formula (28) we can

easily see that it consists in sums and products of matrices. However formula (28) foresees also the computation of $(Q(B,k))^{-1}$, but this matrix does not depend on the particular obstacle D. We note that the computation of $(Q(B,k))^{-1}$ can be quite easier than the computation of $(Q(D,k))^{-1}$; for example, when B is chosen as an axial-symmetric obstacle the matrix $Q(B,k)$ is a 2×2 block-diagonal matrix, so that $(Q(B,k))^{-1}$ can be given in terms of the inverses of its two diagonal blocks, see [22] for details. However the computation of $(Q(B,k))^{-1}$ can be performed only one time since $(Q(B,k))^{-1}$ can be stored and it can be used back for all the obstacles D that we decide to express in terms of B in the perturbative procedure. Moreover formula (28) is well suited for parallel computations, in fact for $n=0,1,\dots,N$ the computation of $T^{(n)}$ consists in $n+1$ matrix-matrix multiplications, where n of these multiplications are independent one from the other. Finally we note that the choice of B cannot be completely independent from D, in fact we expect that fast and accurate approximations of $T(D,k)$ can be obtained from formula (26) when B is close to D in a suitable normed space. This normed space, essential for an eventual investigation of the convergence properties of the approximation (26), is useless for the purpose of the present paper thus its definition is omitted.

We conclude describing the computational consequences of a particular choice for B, that seems quite interesting. In fact when B is chosen as a sphere of radius $r_S>0$, that is $r_B(\hat{\underline{x}})=r_S$, $\hat{\underline{x}}\in S^2$, the matrix $Q(B,k)$ becomes a diagonal matrix. More precisely, we have:

$$(33)\quad Q^{(0)}_{\nu;\nu'}=\left(-\frac{\imath}{2}+\frac{\rho_S^2}{2}\frac{d(j_l(\rho)h_l^{(1)}(\rho))}{d\rho}\bigg|_{\rho=\rho_S}\right)\delta_{\sigma,\sigma'}\delta_{l,l'}\delta_{m,m'},\ \nu,\nu'\in I_{L_{\max}},$$

$$\begin{aligned}(34)\quad Q^{(n)}_{\nu;\nu'}&=\frac{1}{2}\frac{d^n}{d\rho^n}\left(\rho^2\frac{d(j_{l'}(\rho)h_l^{(1)}(\rho))}{d\rho}\right)\bigg|_{\rho=\rho_S}\cdot\\ &\cdot\int_0^{2\pi}d\phi\int_0^{\pi}d\theta\sin\theta H^n(\hat{\underline{x}}(\theta,\phi))Y^{\sigma}_{l,m}(\hat{\underline{x}}(\theta,\phi))Y^{\sigma'}_{l',m'}(\hat{\underline{x}}(\theta,\phi))+\\ &-\frac{n}{2}\frac{d^{n-1}}{d\rho^{n-1}}\left(j_{l'}(\rho)h_l^{(1)}(\rho)\right)\bigg|_{\rho=\rho_S}\int_0^{2\pi}d\phi\int_0^{\pi}d\theta\sin\theta\cdot H^{n-1}(\hat{\underline{x}}(\theta,\phi))\cdot\\ &\cdot\left(\frac{\partial H(\hat{\underline{x}}(\theta,\phi))}{\partial\theta}\frac{\partial}{\partial\theta}\left(Y^{\sigma}_{l,m}(\hat{\underline{x}}(\theta,\phi))Y^{\sigma'}_{l',m'}(\hat{\underline{x}}(\theta,\phi))\right)+\right.\\ &\left.+\frac{1}{\sin^2\theta}\frac{\partial H(\hat{\underline{x}}(\theta,\phi))}{\partial\phi}\frac{\partial}{\partial\phi}\left(Y^{\sigma}_{l,m}(\hat{\underline{x}}(\theta,\phi))Y^{\sigma'}_{l',m'}(\hat{\underline{x}}(\theta,\phi))\right)\right),\\ &\qquad\qquad\nu,\nu'\in I_{L_{\max}},\ n\geq 1,\end{aligned}$$

where $\rho_S = kr_S$. We note that (33) follows from straightforward calculations using formulas (10), (22), (31) and formula (34) follows from formula (32). This is an interesting case since the fact that $Q^{(0)}$ is a diagonal matrix can be used effectively in formula (28) for the computation of $(Q^{(0)})^{-1}$, and besides the evident gain in the accuracy and in the computational cost of $(Q^{(0)})^{-1}$ we have that it improves the parallel efficiency of formula (28).

Finally, we note that the actual validation of the perturbative method proposed in this paper needs a rigorous convergence analysis of series

$$\sum_{n=0}^{\infty} \frac{T^{(n)}}{n!}\epsilon^n,\ 0 \le \epsilon \le 1, \tag{35}$$

generated by formulas (28)-(30). This theoretical analysis deserves to be considered with further investigations, so, at present we provide only some convincing numerical results for the experimental validation of the proposed method.

4 – Numerical results

We present some results extracted from our numerical experience using the perturbative method proposed in the previous section. The numerical results are relative to ten different obstacles and they show mainly the convergence and the stabilization features of the perturbative method. In particular, we consider star-like obstacles whose boundary is parametrized by the following functions:

$$\text{Oblate Ellipsoid}: \ r_1(\hat{\underline{x}}(\theta,\phi)) = \frac{1}{\sqrt{(\frac{2}{3}\sin\theta)^2 + \cos^2\theta}}, \tag{36}$$

$$\text{Prolate Ellipsoid}: \ r_2(\hat{\underline{x}}(\theta,\phi)) = \frac{1}{\sqrt{\sin^2\theta + (\frac{2}{3}\cos\theta)^2}}, \tag{37}$$

$$\text{Pseudo Apollo}: \quad r_3(\hat{\underline{x}}(\theta,\phi)) = \frac{3}{5}\sqrt{\frac{17}{4} + 2\cos(3\theta)}\,, \tag{38}$$

$$\text{Reverse Platelet}: \ r_4(\hat{\underline{x}}(\theta,\phi)) = 1 + \frac{1}{2}\sin^2\theta, \tag{39}$$

$$\text{Short Cylinder}: \quad r_5(\hat{\underline{x}}(\theta,\phi)) = \frac{1}{\sqrt[10]{(\frac{2}{3}\sin\theta)^{10} + \cos^{10}\theta}}, \tag{40}$$

$$\text{Long Cylinder}: \quad r_6(\hat{\underline{x}}(\theta,\phi)) = \frac{1}{\sqrt[10]{\sin^{10}\theta + (\frac{2}{3}\cos\theta)^{10}}}, \tag{41}$$

$$\text{Vogel's Nut}: \quad r_7(\hat{\underline{x}}(\theta,\phi)) = \frac{3}{2}\sqrt{1 - \frac{3}{4}\sin^2\theta}, \tag{42}$$

(43) Generic Ellipsoid : $$r_8(\hat{\underline{x}}(\theta,\phi)) = \frac{1}{\sqrt{\left(\left(\frac{3}{2}\sin\phi\right)^2 + \cos^2\phi\right)\sin^2\theta + \left(\frac{2}{3}\cos\theta\right)^2}},$$

(44) Corrugated Sphere : $$r_9(\hat{\underline{x}}(\theta,\phi)) = \left(1 + \frac{1}{20}\cos(4\theta) + \frac{1}{40}\cos(8\theta)\right) \cdot \left(1 + \frac{1}{20}\cos(4\phi) + \frac{1}{40}\cos(8\phi)\right),$$

(45) Cuboid : $$r_{10}(\hat{\underline{x}}(\theta,\phi)) = \frac{1}{\sqrt[10]{(\sin^{10}\phi + \cos^{10}\phi)\sin^{10}\theta + \cos^{10}\theta}},$$

where $\theta \in [0,\pi]$, $\phi \in [0,2\pi)$. We note that obstacles (36)-(42) are axial-symmetric obstacles, that is the corresponding parametrization of the boundary is a function independent from variable ϕ, obstacles (43)-(45) have not particular symmetry properties; all the obstacles are convex excepting (38), (42), (44) that are non-convex obstacles. In Figure 1 are shown the ten obstacles defined in (36)-(45). Finally, in problem (4)-(6) we always consider $k = 1$, and in equation (5) we choose function (19) with $\underline{\alpha} = \hat{\underline{x}}(\frac{\pi}{3}, \frac{\pi}{6})$.

The numerical results corresponding to obstacles (36)-(45) are reported in Table 1. For the computation of these results we have performed the sum in formula (26) using the arithmetic mean methods for the summation of divergent series. The simpler arithmetic mean method is the usual Cesàro means. This method can be generalized in several different ways obtaining, for example, the method of Hölder, the method of Cesàro, the method of Riesz; all these methods depend on a parameter usually called order of the method and they reduce to the usual Cesàro means when the order is equal to one, see [30] page 94 for a more detailed discussion. In particular for the results reported in Table 1 we have considered the Riesz method, that is given $\tau \in \mathbb{N}$ we define $\Sigma^{(N,\tau)}$ to be the sum of the matrices $T^{(n)}$, $n = 0, 1, \ldots, N$ according to the Riesz method of order τ, that is

(46) $$\Sigma^{(N,\tau)} = \sum_{n=0}^{N} \left(1 - \frac{n}{N+1}\right)^{\tau} \frac{T^{(n)}}{n!}.$$

The Riesz method is regular, that is, it does not modify the sum of convergent series. Thus, supposing that (35) is a convergent series we can compute $T(D,k)$ using either series (35) or $\Sigma^{(N,\tau)}$, as $N \to \infty$. In practice methods for the summation of divergent series are usually used for transforming slowly convergent into rapidly convergent series. From numerical results not reported in this paper series (35) seems to be convergent for all the considered obstacles, but the rate of convergence is quite dependent on the difficulty of the particular obstacle taken

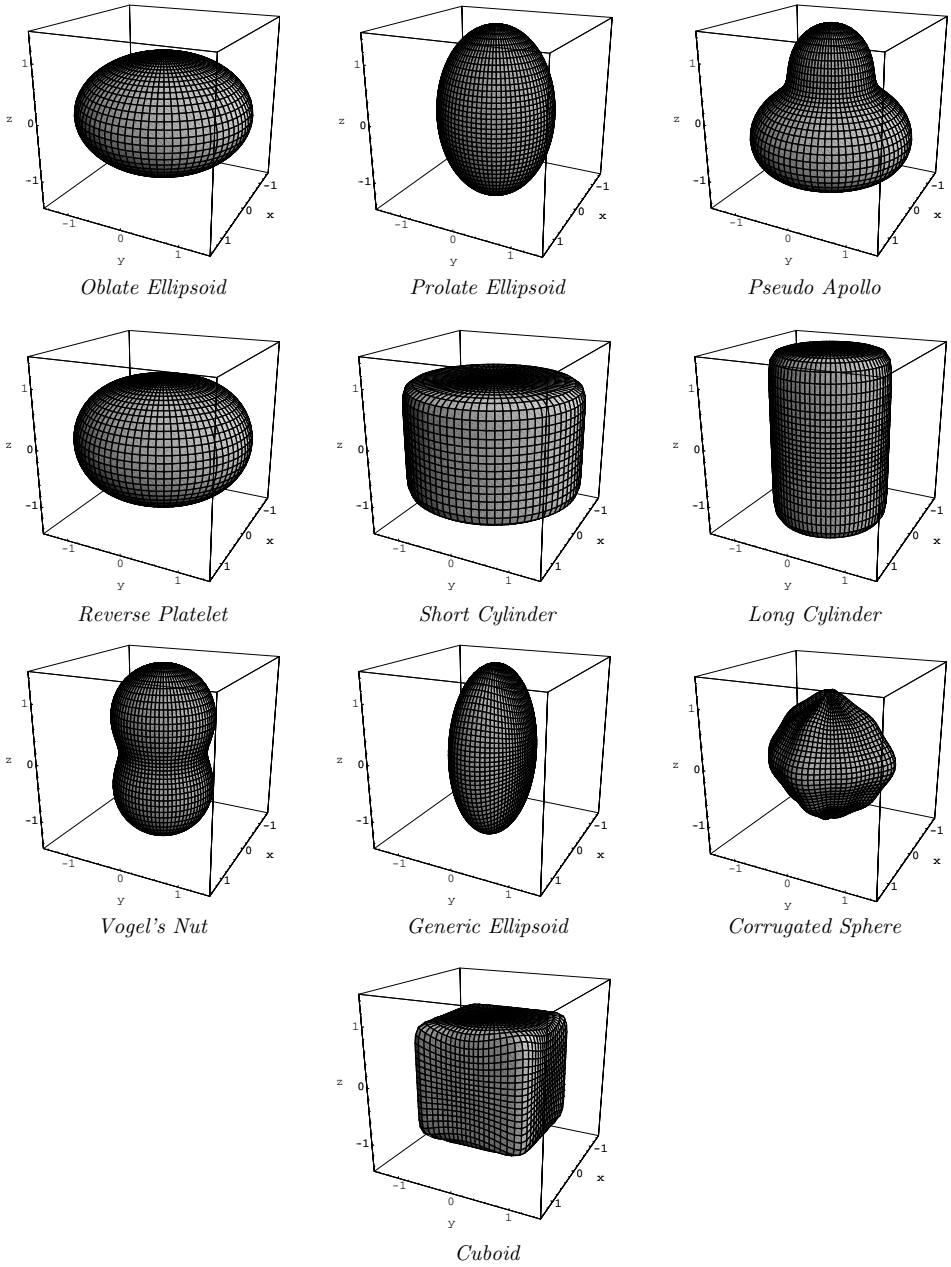

Fig. 1: The obstacles defined in (36)-(45).

into account. This unpleasant property of approximation (26) is attenuated by using the above mentioned methods for the summation of divergent series; in particular, we have that the method of Cesàro gives results similar to the ones obtained with the method of Riesz. The method of Hölder gives usually better results with respect to the method of Riesz when we consider hard obstacles, such as for example *Reverse Platelet*, but it gives much worse results when we consider easy obstacles, such as for example ellipsoids. So that given $N, \tau \in \mathbb{N}$ matrix $\Sigma^{(N,\tau)}$ is the computed approximation of the matrix $T(D,k)$; Table 1 shows the convergence properties of the sum $\Sigma^{(N,\tau)}$ to the matrix $T(D,k)$. We define the following performance index:

$$E_T^{(N,\tau)} = \frac{\left\|\Sigma^{(N,\tau)} - T(D,k)\right\|_\infty}{\|T(D,k)\|_\infty}. \tag{47}$$

where $\|\cdot\|_\infty$ denotes the operator matrix norm associated with the vector maximum norm. Moreover the approximation $\tilde{u}^{s,(N,\tau)}$ of the solution u^s of problem (4)-(6) is computed from formulas (14), (15) substituting $T(D,k)$ with $\Sigma^{(N,\tau)}$. Table 1 also shows a comparison between the approximation $\tilde{u}^s$ of the solution u^s of problem (4)-(6) obtained using the usual T-matrix method and the approximation $\tilde{u}^{s,(N,\tau)}$ obtained using the perturbative method. As a consequence of the discussion following formula (12) this comparison takes into account only the error in the approximation of condition (5), so that we consider the following two performance indices:

$$\begin{aligned} E_u = \frac{1}{92}\Bigg(&\left|\tilde{u}^s(\underline{\xi}_{0,0}) + u^i(\underline{\xi}_{0,0})\right| + \left|\tilde{u}^s(\underline{\xi}_{10,0}) + u^i(\underline{\xi}_{10,0})\right| + \\ &+ \sum_{i=1}^{9}\sum_{j=0}^{9}\left|\tilde{u}^s(\underline{\xi}_{i,j}) + u^i(\underline{\xi}_{i,j})\right|\Bigg), \end{aligned} \tag{48}$$

$$\begin{aligned} E_u^{(N,\tau)} = \frac{1}{92}\Bigg(&\left|\tilde{u}^{s,(N,\tau)}(\underline{\xi}_{0,0}) + u^i(\underline{\xi}_{0,0})\right| + \left|\tilde{u}^{s,(N,\tau)}(\underline{\xi}_{10,0}) + u^i(\underline{\xi}_{10,0})\right| + \\ &+ \sum_{i=1}^{9}\sum_{j=0}^{9}\left|\tilde{u}^{s,(N,\tau)}(\underline{\xi}_{i,j}) + u^i(\underline{\xi}_{i,j})\right|\Bigg), \end{aligned} \tag{49}$$

where $\underline{\xi}_{i,j} = r(\underline{\hat{x}}(\frac{\pi}{10}i, \frac{\pi}{5}j))\underline{\hat{x}}(\frac{\pi}{10}i, \frac{\pi}{5}j)$, $j, i = 0, 1, \ldots, 10$, and r is the parametrization of the boundary ∂D of the obstacle D under consideration. The indices E_u, $E_u^{(N,\tau)}$ can be seen as relative errors computed on a regular grid of ∂D; note that number 92, appearing in formulas (48), (49), represents the sum of the

absolute values of u^i, defined in (19), on such a grid. We note that the results shown in Table 1 are relative to the following choice of the parameters previously described: $L_{\max} = 6$, $\tau = 5, 10$, $N = 5, 10$. For a generic obstacle D the choice of the base point B in the perturbative procedure is given by the sphere having the nearest boundary ∂B to ∂D in the least-squares sense. Moreover the integrals appearing in formulas (22), (34) are approximated by a composite Gauss-Legendre formula and the solution of equation (18) is computed by the LU factorization of matrix Q with partial pivoting.

TABLE 1. The numerical results for the ten obstacles (36)-(45). For each obstacle the performance indices $E_T^{(N,\tau)}$, $E_u^{(N,\tau)}$, $N = 5, 10$, $\tau = 5, 10$ and E_u are reported.

		$E_T^{(N,\tau)}$		$E_u^{(N,\tau)}$		E_u
		$N = 5$	$N = 10$	$N = 5$	$N = 10$	
Oblate ellipsoid	$\tau = 5$	2.29(−2)	3.94(−2)	7.24(−2)	9.34(−2)	9.26(−2)
	$\tau = 10$	2.29(−2)	3.94(−2)	7.41(−2)	9.38(−2)	
Prolate ellipsoid	$\tau = 5$	2.88(−2)	4.94(−2)	5.33(−2)	7.90(−2)	4.31(−2)
	$\tau = 10$	2.88(−2)	4.94(−2)	5.32(−2)	7.90(−2)	
Pseudo apollo	$\tau = 5$	4.75(−2)	7.36(−2)	1.31(−1)	1.03(−1)	2.01(-1)
	$\tau = 10$	4.65(−2)	7.36(−2)	1.54(−1)	1.04(−1)	
Reverse platelet	$\tau = 5$	3.45(−1)	3.69(−1)	3.30	1.74	5.08
	$\tau = 10$	3.33(−1)	3.67(−1)	4.61	1.90	
Short cylinder	$\tau = 5$	2.90(−2)	5.16(−2)	2.59(−1)	2.13(−1)	3.24(−1)
	$\tau = 10$	2.93(−2)	5.17(−2)	2.66(−1)	2.18(−1)	
Long cylinder	$\tau = 5$	4.69(−2)	8.63(−2)	2.36(−1)	1.91(−1)	3.03(−1)
	$\tau = 10$	4.78(−2)	8.64(−2)	2.33(−1)	1.90(−1)	
Vogel's nut	$\tau = 5$	1.34(−1)	1.67(−1)	1.81(−1)	1.43(−1)	1.89(−1)
	$\tau = 10$	1.27(−1)	1.66(−1)	3.46(−1)	1.58(−1)	
Generic ellipsoid	$\tau = 5$	4.77(−2)	8.11(−2)	2.77(−1)	2.02(−1)	4.69(−1)
	$\tau = 10$	4.81(−2)	8.13(−2)	2.52(−1)	1.97(−1)	
Corrugated sphere	$\tau = 5$	7.86(−3)	1.18(−2)	5.17(−2)	5.35(−2)	5.45(−2)
	$\tau = 10$	7.85(−3)	1.18(−2)	5.17(−2)	5.35(−2)	
Cuboid	$\tau = 5$	3.01(−2)	4.53(−2)	1.30(−1)	9.11(−2)	1.88(−1)
	$\tau = 10$	2.93(−2)	4.51(−2)	1.31(−1)	9.14(−2)	

From Table 1 we can see very interesting results. In particular, we can note a quite rapid convergence, also due to the Riesz method, of the sum (46) to the matrix $T(D,k)$, in fact the indices $E_T^{(5,\tau)}$, $E_T^{(10,\tau)}$ are quite similar. Moreover, comparing $E_u^{(5,\tau)}$, $E_u^{(10,\tau)}$ we can see that high values of N need for obstacles having shape far from the sphere, such as for example *Long Cylinder*, *Vogel's Nut* and *Cuboid*, see Figure 1. We can also note that the use of a high value for parameter τ is usually useless and sometimes spoils the accuracy of the final approximation of $T(D,k)$. Moreover, comparing the indices E_u, $E_u^{(N,\tau)}$ reported in Table 1 it can be noted a quite general improvement of the solution obtained by the perturbative technique with respect to the one obtained by the usual T-matrix method. We can also note that the sensitivity of $E_u^{(N,\tau)}$ with respect to τ is larger than the one of $E_T^{(N,\tau)}$; furthermore it seems that the value of τ must be chosen according to the difficulty of the obstacle D, in fact for easy obstacles like ellipsoids we obtain better results for low values of τ, instead for the other obstacles we obtain better results for high values of τ.

5 – Conclusions

We consider the solution of direct scattering problems. These problems can be seen as boundary-value problems for the Helmholtz equation in unbounded domains. For the solution of these problems we study the so called T-matrix method, which is a very classical method for the solution of direct scattering problems. We propose a perturbative method based on the T-matrix method. From a large number of numerical experiments we have discussed the improvement in the accuracy of the T-matrix method due to the perturbative technique presented. In particular the numerical results shown in Section 4 are very interesting, so that we deserve further investigations of the method presented. The main question is, of course, the settlement of classes of obstacles for which the perturbative procedure proposed generates convergent approximations (see formula (26)) of the matrix $T(D,k)$. This investigation, unavoidable for a rigorous validation of the proposed method, can be integrated and completed with the study of the connection of formula (35) and the well known methods for the summation of divergent series. Another interesting question is also the development of versions of formulas (28), (31), (32), (33), (34) that are efficient for sequential computations and for parallel computations.

We conclude describing a possible very interesting application of the method proposed. The perturbative procedure presented here can deal in a natural way with the problem of scattering by random rough surface obstacles. This problem has been initially considered for the study of water waves on the ocean surface, but now it finds application in several different fields of engineering and natural sciences, such as for example detection of small defects in manufacturing processes or the study of the variations in height in natural ground surfaces, see [31], [32], [33] for a detailed discussion.

REFERENCES

[1] P.M. Morse – K.V. Ingard: *Theoretical Acoustics*, Mc Graw Hill, New York, 1968.

[2] D. Colton – R. Kress: *Integral Equation Methods in Scattering Theory*, John Wiley & Sons, New York, 1983.

[3] A. Sommerfeld: *Partial Differential Equations in Physics*, Academic Press, New York, 1964.

[4] P. Monk: *Finite Element Methods for Maxwell's Equations*, Oxford University Press, Oxford, 2003.

[5] I. Bar-On – Å. Edlund – U. Peskin: *Parallel solution of the multidimensional Helmholtz/Schroedinger equation using high order methods*, Proceedings of the Fourth International Conference on Spectral and High Order Methods, (Held in Herzliya, 1998), Eds J.S. Hesthaven, D. Gottlieb, E. Turkel, Applied Numerical Mathematics, **33** (2000), 95-104.

[6] K. Otto – E. Larsson: *Iterative solution of the Helmholtz equation by a second-order method*, SIAM Journal on Matrix Analysis and Applications, **21** (1999), 209-229.

[7] I. Singer – E. Turkel: *High-order finite difference methods for the Helmholtz equation*, Computer Methods in Applied Mechanics and Engineering, **163** (1998), 343-358.

[8] E. Giladi – J.B. Keller: *A hybrid numerical asymptotic method for scattering problems*, Journal of Computational Physics, **174** (2001), 226-247.

[9] P.E. Barbone – I. Harari: *Nearly H^1-optimal finite element methods*, Computer Methods in Applied Mechanics and Engineering, **190** (2001), 5679-5690.

[10] I. Babuška – F. Ihlenburg – E.T. Paik – S.A. Sauter: *A generalized finite element method for solving the Helmholtz equation in two dimensions with minimal pollution*, Computer Methods in Applied Mechanics and Engineering, **128** (1995), 325-359.

[11] A. Kirsch – P. Monk: *An analysis of the coupling of finite-element and Nyström methods in acoustic scattering*, IMA Journal of Numerical Analysis, **14** (1994), 523-544.

[12] Absorbing boundary conditions: *Papers from the IUTAM Symposium held in July 1997*, Ed. E. Turkel, Applied Numerical Mathematics, **27** (1998), 327-557.

[13] D.M. Milder: *An improved formalism for wave scattering from rough surface*, Journal of the Acoustical Society of America, **89** (1991), 529-541.

[14] D.M. Milder: *Role of the admittance operator in rough-surface scattering*, Journal of the Acoustical Society of America, **100** (1996), 759-768.

[15] R.A. Smith: *The operator expansion formalism for electromagnetic scattering from rough dielectric surfaces*, Radio Science, **31** (1996), 1377-1385.

[16] S. Piccolo – M.C. Recchioni – F. Zirilli: *The time harmonic electromagnetic field in a disturbed half space: an existence theorem and a computational method*, Journal of Mathematical Physics, **37** (1996), 2762-2786.

[17] L. Misici – G. Pacelli – F. Zirilli: *A new formalism for wave scattering from a bounded obstacle*, Journal of the Acoustical Society of America, **103** (1998), 106-113.

[18] J.C. Nédélec: *Acoustic and electromagnetic equations. Integral representations for harmonic problems*, Applied Mathematical Sciences, 144. Springer-Verlag, New York, 2001.

[19] G.F. Roach: *Boundary integral equation methods for elliptic boundary value problems*, Bulletin of the Institute of Mathematics and its Applications, **20** (1984), 82-88.

[20] T. Angell – R.E. Kleinman: *Modified Green's functions and the third boundary value problem for the Helmholtz equation*, Journal of Mathematical Analysis and Applications, **97** (1983), 81-94.

[21] P. Maponi – F. Zirilli: *The use of the operator expansion method to compute the generalized eigenfunctions of the Laplacian in a two-dimensional cavity*, in preparation.

[22] P.C. Waterman: *New formulation of acoustic scattering*, Journal of the Acoustical Society of America, **45** (1969), 1417-1429.

[23] A.G. Ramm: *Numerically efficient version of the T-matrix method*, Applicable Analysis, **80** (2001), 385-393.

[24] P.A. Martin: *Multiple scattering: an invitation*, Mathematical and numerical aspects of wave propagation (Mandelieu-La Napoule, 1995), SIAM, Philadelphia, 1995, 3-16.

[25] J.H. Lin – W.C. Chew: *BiCG-FFT T-Matrix method for solving for the scattering solution from inhomogeneous bodies*, IEEE Transactions on Microwave Theory and Techniques, **44** (1996), 1150-1155.

[26] *Acoustic, Electromagnetic and Elastic Wave Scattering: Focus on the T-matrix approach*: Eds. V.K. Varadan, V.V. Varadan, Pergamon Press, New York, 1980.

[27] Z. Wang – L. Hu – W. Ren: *Multiple scattering of acoustic waves by a half-space of distributed discrete scatterers with modified T matrix approach*, Waves in Random Media, **4** (1994), 369-375.

[28] M. Abramowitz – I.A Stegun: *Handbook of Mathematical Functions*, Dover Publications, New York, 1968.

[29] P.M. Morse – H.Feshbach: *Methods of theoretical physics*, Part II, Mc Graw Hill Book Company, New York, 1953.

[30] G.H. Hardy: *Divergent series*, Oxford University Press, London, 1967.

[31] J.A. Ogilvy: *Theory of Wave Scattering from Random Rough Surfaces*, Hilger, Bristol, 1991.

[32] G. Voronovich: *Wave Scattering from Rough Surfaces*, Springer, Berlin, 1999.

[33] K.F. Warnick – W.C. Chew: *Numerical simulation methods for rough surface scattering*, Wave Random Media, **11** (2001), 1-30.

Lavoro pervenuto alla redazione il 15 febbraio 2003
ed accettato per la pubblicazione il 8 marzo 2004.
Bozze licenziate il 6 dicembre 2004

Indirizzo degli Autori:

Nadaniela Egidi, Pierluigi Maponi – Dipartimento di Matematica e Informatica – Università di Camerino – 62032 Camerino, Italy
E-mail: nadaniela.egidi@unicam.it pierluigi.maponi@unicam.it

Rendiconti di Matematica, Serie VII
Volume 24, Roma (2004), 321-336

On the stability of the first eigenvalue of $A_p u + \lambda\, g(x) \mid u \mid^{p-2} u = 0$ with varying p

A. El KHALIL – P. LINDQVIST – A. TOUZANI

ABSTRACT: *The stability with respect to p of the nonlinear eigenvalue problem*

$$\sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left[\left(\sum_{m,k=1}^{N} a_{m,k}(x)\frac{\partial u}{\partial x_m}\frac{\partial u}{\partial x_k}\right)^{\frac{p-2}{2}} a_{i,j}(x)\frac{\partial u}{\partial x_j}\right] + \lambda g(x) \mid u \mid^{p-2} u = 0,$$

is studied.

1 – Introduction and notations

In this paper we study the continuity (stability) of the eigenvalue problem

$$\begin{cases} -A_p u = \lambda g(x) \mid u \mid^{p-2} u & \text{in } \Omega \\ u \in W_0^{1,p}(\Omega), \end{cases} \tag{1.1}$$

with respect to p which varies continuously in $(1,\infty)$. Here Ω is a bounded domain in $\mathbb{R}^N$ and $g \in L^\infty_{\mathrm{loc}}(\Omega) \cap L^r(\Omega)$ is an indefinite weight function. The exponent $r = r(N,p)$ satisfies the following conditions

$$\begin{cases} r \geq \frac{Np}{p-1} & \text{when } 1 < p \leq N \\ r = 1 & \text{when } p > N, \end{cases} \tag{1.2}$$

KEY WORDS AND PHRASES: *A_p-Laplacian – Indefinite weight – Stability – Nonlinear eigenvalue problem – Segment property.*
A.M.S. CLASSIFICATION: 35B35 – 35B32 – 35J70

and g can change its sign in Ω, we assume only that $\Omega^+ = \{x \in \Omega, g(x) > 0\}$ has positive measure. The so-called A_p-Laplacian operator is defined by

$$A_p u = \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i} \left[\left(\sum_{m,k=1}^{N} a_{m,k}(x) \frac{\partial u}{\partial x_m} \frac{\partial u}{\partial x_k} \right)^{\frac{p-2}{2}} a_{i,j}(x) \frac{\partial u}{\partial x_j} \right].$$

Where $A = (a_{i,j})_{i,j}$ is a matrix satisfying the conditions

$$(1.3) \quad \begin{cases} \text{(i)} \quad a_{i,j} \equiv a_{j,i} \in L^\infty(\Omega) \cap \mathcal{C}^1(\Omega) \\ \text{(ii)} \quad \mid \xi \mid_a^2 \equiv \sum_{i,j=1}^{N} a_{i,j}(x)\xi_i\xi_j \geq \mid \xi \mid^2 \qquad \text{when } x \in \Omega \text{ for all } \xi \in \mathbb{R}^N. \end{cases}$$

We will use the norm

$$\| v \|_{1,p} = \| \mid \nabla v \mid_a \|_p = \left(\int_\Omega \mid \nabla v \mid_a^p \, dx \right)^{\frac{1}{p}}.$$

We also define an inner product

$$\langle \xi, \zeta \rangle_a \equiv \sum_{i,j=1}^{N} a_{i,j}(x)\xi_i\zeta_j.$$

The A_p-Laplacian operator defined above was studied by Yu. G. Reshetnyak [13] and J. Mossino [11] and used in [8]. Many elliptic operators are particular cases of the A_p-Laplacian operator. For example, the p-Laplacian

$$\Delta_p u = \operatorname{div} (\mid \nabla u \mid^{p-2} \nabla u)$$

and the linear operator

$$A_2 u = \sum_{i,j=1}^{N} \frac{\partial}{\partial x_i} \left(a_{i,j}(x) \frac{\partial u}{\partial x_i} \right).$$

These operators,with $p \neq 2$, are used for non-Newtonian fluids (dilatant fluids have $p > 2$, pseudo-plastics have $1 < p < 2$), and appear in some reaction-diffusion problems as well as in nonlinear elasticity, and in glaciology ($p = \frac{3}{4}$).

Under various conditions the simplicity of the first eigenvalue for the above case Δ_p were obtained by various authors. When $g \equiv 1$ the first eigenvalue for the A_p-Laplacian is simple as in the case of the ordinary p-Laplacian, see [3, 12, 14] for more general g. These results were extended to our problem in [15].

Recently, for $g \equiv 1$ and without any assumptions of regularity on the domain Ω, the simplicity of the first eigenvalue was proved in [9] for the p-Laplacian Δ_p. Its stability (continuity) with respect to p was studied in [10]. In some other cases, it was studied in [6].

The principal eigenvalue $\lambda_p(g)$ of the A_p-Laplacian with indefinite weight g is here defined as the least positive real number $\lambda > 0$ for which the problem (1.1) has a nontrivial solution.

We now describe some main results of this paper. We study the convergence of the first eigenfunctions in connection with the inequalities

$$\lim_{s \to p_-} \lambda_s(g) \leq \lambda_p(g) = \lim_{s \to p_+} \lambda_s(g),$$

proved in Theorem 3.2 and Corollary 3.1. In other words we explore the behavior of the principal eigenfunction $u_s \in W_0^{1,s}(\Omega)$ (required to be positive and $\int_\Omega g(x)|u_s|^s \, dx = 1$) to the equation

$$A_s u_s + \lambda_s(g)|u_s|^{s-2}u_s = 0,$$

as s varies continuously in $(1, \infty)$. This is why we are interested in the stability to the right.

In very irregular domains with $p \leq N$, the situation $\lim_{s \to p_-} \lambda_s(g) < \lambda_p(g)$ is possible. An example is given by [10] in the case $A_p = \Delta_p$ and $g \equiv 1$. This situation is as a consequence of a strange convergence phenomenon: The principal eigenfunctions u_s, $s < p$, converge to a positive solution of the first equation (1.1).

The limit function is in the Sobolev space $W^{1,p}(\Omega)$ and in every $W_0^{1,p-\epsilon}(\Omega)$, $\epsilon > 0$ small enough, but is not in the required $W_0^{1,p}(\Omega)$. If Ω satisfies the segment property then it follows from Theorem 2.1, that

$$W_0^{1,q}(\Omega) \cap W^{1,p}(\Omega) = W_0^{1,p}(\Omega), 1 < q < p.$$

In this case we show in Corollary 3.2 and Corollary 3.3 our main results related to the stability.

In Theorem 3.6 we show that the eigenfunctions and their gradients converge locally uniformly to a positive solution of the first equation problem (1.1), by the $\mathcal{C}^{1,\alpha}_{\text{loc}}$-regularity, see [4], and the L^∞-estimate established in the Appendix.

The paper is organized as follows: In Section 2, we establish some definitions and basic properties. In Section 3, we first give some general stability results with respect to p for the first positive eigenvalue of problem (1.1) and we restrict ourselves to bounded domain Ω having the segment property. This class of domains is fairly large. Then we prove the global stability using some results established in Section 2 and in Appendix. The segment property is needed here to guarantee the right boundary values of the limit function.

2 – Preliminary results

In defining the eigenvalues of the A_p-Laplacian operator with weight (in a given bounded domain $\Omega \subset \mathbb{R}^N$), we shall interpret Equation (1.1) in the weak sense.

Definition 2.1. We say that $\lambda \in \mathbb{R}$ is an eigenvalue, if there exists a function $u \in W_0^{1,p}(\Omega), u \not\equiv 0$, such that

$$\int_\Omega \mid \nabla u \mid_a^{p-2} \langle \nabla u, \nabla \varphi \rangle_a \, dx = \lambda \int_\Omega g(x) \mid u \mid^{p-2} u\varphi \, dx, \tag{2.1}$$

whenever $\varphi \in W_0^{1,p}(\Omega)$. The function u is called an eigenfunction.

2.1 – Basic properties

Under our conditions on $a_{i,j}$ and g, it is well-known that the problem (1.1) possesses at least a sequence of positive eigenvalues $\lambda_n, \lambda_n \nearrow^{+\infty}$, as $n \to +\infty$. These can obtained by the Ljusternick-Schnirelmann theory minimizing the energy functional,

$$\Phi(u) = \left(\frac{1}{p} |||\nabla u|_a||_p^p \right)^2 - \frac{1}{p} \int_\Omega g(x)|u|^p \, dx,$$

on $W_0^{1,p}(\Omega)$. See [2], see also [8] or [15].

Let now $\lambda_p(g)$ denote the first positive eigenvalue of (1.1). We recall that $\lambda_p(g)$ can be variationally characterized as

$$\begin{aligned} \lambda_p(g) &= \min \left\{ \int_\Omega \mid \nabla u \mid_a^p \, dx; \; u \in W_0^{1,p}(\Omega), \int_\Omega g(x) \mid u \mid_a^p \, dx = 1 \right\} = \\ &= \min \left\{ \frac{\int_\Omega \mid \nabla u \mid_a^p \, dx}{\int_\Omega g(x) \mid u \mid^p \, dx}; \; u \in W_0^{1,p}(\Omega), \int_\Omega g(x) \mid u \mid^p \, dx > 0 \right\}. \end{aligned} \tag{2.2}$$

Throughout this paper, the first eigenfunctions are those corresponding to $\lambda_p(g)$. The principal eigenfunction, denoted u_p, is the first eigenfunction normalized by $\int_\Omega g(x) \mid u_p \mid^p \, dx = 1$, and required to be positive. Hence

$$\lambda_p(g) \; = \; \int_\Omega \mid \nabla u_p \mid_a^p \, dx.$$

We end this paragraph by recalling some fundamental properties, found in [8], [15], which valid under our assumptions.

1) The first eigenfunctions are essentially unique in any bounded domain, i.e., they are merely constant multiples of each other.
2) The principal eigenfunction has no zeros in the domain the first eigenfunctions are only those not changing sign.
3) The solutions of problem (1.1) are known to be of class $C_{\rm loc}^{1,\alpha}(\Omega)$ for some $\alpha > 0$ depending on p and N, see [4].

2.2 – The segment property

We begin with defining a sharp class of domains for which the boundary is sufficiently regular to guarantee that

$$W^{1,p}(\Omega) \cap_{q<p} W_0^{1,q}(\Omega) = W_0^{1,p}(\Omega).$$

DEFINITION 2.2. An open subset Ω of $\mathbb{R}^N$ is said to have the segment property if, given any $x \in \partial\Omega$, there exist an open set G_x in $\mathbb{R}^N$ with $x \in G_x$ and y_x of $\mathbb{R}^N \setminus \{0\}$ such that, if $z \in \overline{\Omega} \cap G_x$ and $t \in]0,1[$, then $z + ty_x \in \Omega$.

This property allows us by a translation to push the support of a function u in Ω. The following result is essential here.

THEOREM 2.1. *Let Ω be a bounded domain in $\mathbb{R}^N$ having the segment property. If $u \in W^{1,p}(\Omega) \cap W_0^{1,q}(\Omega)$ for some $q \in]1,p[$, then $u \in W_0^{1,p}(\Omega)$.*

PROOF. The following technique is inspired by [1, Theorem 3.18]. The function

$$\tilde{u} = \begin{cases} u & \text{in } \Omega \\ 0 & \text{in } \mathbb{R}^N \setminus \Omega, \end{cases}$$

is in $W^{1,p}(\mathbb{R}^N)$. Indeed, we have $u \in W_0^{1,q}(\Omega)$, and so $\tilde{u} \in W^{1,q}(\mathbb{R}^N)$; moreover $\nabla\tilde{u} = \widetilde{\nabla u}$ weakly and a.e. on $\mathbb{R}^N$. On the other hand, $\tilde{u} \in L^p(\mathbb{R}^N)$ and $\widetilde{\nabla u} \in (L^p(\mathbb{R}^N))^N$, because $u \in W^{1,p}(\Omega)$. Finally, we conclude that $\tilde{u} \in W^{1,p}(\mathbb{R}^N))$. Let $K = \operatorname{supp} u =: \overline{\{x \in \Omega, u(x) \neq 0\}}^{\mathbb{R}^N}$, (closure in $\mathbb{R}^N$). Thus K is compact and $K \subset \overline{\Omega}$.

If $K \subset \Omega$, let j_ϵ be defined as in Section 2.17 of [1], thus the convolution $j_\epsilon * u \in C_0^\infty(\Omega)$, provided $0 < \epsilon < \operatorname{dist}(K, \partial\Omega)$, and $j_\epsilon * u \to u$ in $W^{1,p}(\Omega)$, as $\epsilon \to 0^+$. This shows that $u \in W_0^{1,p}(\Omega)$. We shall therefore suppose that $K \cap \partial\Omega \neq \emptyset$. From Definition 2.2, to each $x \in \partial\Omega$, there corresponds a neighborhood G_x and a vector $y_x \in \mathbb{R}^N \setminus \{0\}$. Put $F = K \cap (\overline{\Omega} \setminus \bigcup_{x\in\partial\Omega} G_x)$; then F is compact and $F \subset \Omega$. Thus there is an open set G_0 such that $F \subset G_0 \subset \Omega$, with $\overline{G}_0 \subset \Omega$. On the other hand, $K \cap \partial\Omega$ is compact in $\mathbb{R}^N$ and covered by the open sets G_x,$x \in \partial\Omega$. Therefore $K \cap \partial\Omega$ may be covered by finitely may of the G_x, say $G_1, G_2, ..., G_k$, and also the sets $G_0, G_1, ..., G_k$ form an open covering of K. By a similar argument as that in the proof of Theorem 3.18. of [1, p.55], we can construct open sets $G_0', G_1', ..., G_k'$ which form an open covering of K with $\overline{G_j'} \subset G_j$ for each j. Now, let $\Theta = \{\theta_j,\ 0 \leq j \leq k\}$ be a partition of unity subordinate to covering $\{G_j',\ 0 \leq j \leq k\}$ and put $u_j = \theta_j u, \forall j = 0, ..., k$. We have $u = \sum_{j=0}^N u_j$ and supp $u_j \subset G_j'$, for each $j = 0, ..., k$. Therefore, it suffices to show that each $u_j \in W_0^{1,p}(\Omega \cap G_j)$. Since $\overline{G_0'} \subset \Omega$, our discussion of the case $K \subset \Omega$ above shows that $u_0 \in W_0^{1,p}(\Omega)$. For $j \geq 1$, we have $u_j \in W^{1,p}(\Omega \cap G_j)$

and $\tilde{u}_j \in W^{1,p}(\mathbb{R}^N)$. Put $K_j = \operatorname{supp} u_j$ and let $u_{j,t} = \tilde{u}_j(x - ty_j)$, with $0 < t < \min\{1, |\, y_j \,|^{-1} \operatorname{dist}(G_j', G_j^c)\}$, y_j denoting the element associated with G_j as in Definition 2.2. Thus we have

$$\operatorname{supp} u_{j,t} \subset \Omega \cap G_j, \tag{2.3}$$

for each t satisfying $0 < t < \min\{1, |\, y_j \,|^{-1} \operatorname{dist}(G_j', G_j^c)\}$. Indeed, we have

$$\operatorname{supp} u_{j,t} = K_j + ty_j \subset G_j \cap \overline{\Omega} + ty_j \subset \Omega$$

by the segment property. On the other hand, let $x \in \operatorname{supp} u_{j,t}$. Then

$$\operatorname{dist}(x, G_j') \leq \operatorname{dist}(x, x - ty_j) + \operatorname{dist}(x - ty_j, K_j) + \operatorname{dist}(K_j, G_j') = \operatorname{dist}(x, x - ty_j).$$

We obtain

$$\operatorname{dist}(x, G_j') \leq \operatorname{dist}(x, x - ty_j) = |\, ty_j \,| \,.$$

Therefore, $\operatorname{dist}(x, G_j') < \operatorname{dist}(G_j', G_j^c)$ by the choice of t. Hence $x \in G_j$. This completes the proof of (2.3). We also have $u_{j,t} \in W^{1,p}(\mathbb{R}^N)$, because $\widetilde{u_j} \in W^{1,p}(\mathbb{R}^N)$; especially, we have $u_{j,t} \in W^{1,p}(\Omega \cap G_j)$ and from (2.3), we deduce that $u_{j,t} \in W_0^{1,p}(\Omega \cap G_j)$, for $t > 0$ sufficiently small. Translation is continuous in $L^p(\Omega \cap G_j)$ so $u_{j,t} \to u_j$ in $L^p(\Omega \cap G_j)$ and $\nabla u_{j,t} \to \nabla u_j$ in $(L^p(\Omega \cap G_j))^N$, as $t \to 0^+$ (note that $\nabla \widetilde{u_j} = \widetilde{\nabla u_j}$). Hence $u_{j,t} \to u_j$ in $W^{1,p}(\Omega \cap G_j)$. This together with the fact that $u_{j,t} \in W_0^{1,p}(\Omega \cap G_j)$, for $t > 0$ small enough, ends the proof. ∎

Remark 2.1. A bounded domain $\Omega \subset \mathbb{R}^N$ has the segment property if, and only if, it is in the class $\mathcal{C}$, cf. [5]. This means that locally the boundary has the continuous equation $x_N = f(x_1, x_2, ..., x_{N-1})$, after a notation of the coordinate axis.

3 – Stability of $s \longrightarrow \lambda_s(g)$

The first positive eigenvalue $\lambda_s(g)$ of the A_s-Laplacian with weight $g \in L^\infty_{\mathrm{loc}}(\Omega) \cap L^r(\Omega)$, where $r = r(N,p)$ satisfies (1.2), exists for each $s \in (1, \infty)$ which is near enough to p. Indeed, observe that (1.2) yields the following conditions: $r > \frac{N}{p}$ if $1 < p < N, r > N$ if $p = N$ and $r = 1$ if $p > N$, which imply the existence of $\lambda_s(g)$, (cf. [15]).

We will assume throughout this section that our conditions on g and $a_{i,j}$ are satisfied.

3.1 – Some inequalities

THEOREM 3.1. *The eigenvalues $\lambda_s(g)$ and λ_s satisfy*

$$p\lambda_p^{\frac{1}{p}}(g) \leq s\lambda_s^{\frac{1}{s}} \left(\frac{\lambda_s(g)}{\lambda_s}\right)^{\frac{1}{p}}, \tag{3.1}$$

when $1 < p < s$ and p, s are close enough.

PROOF. Let $\varphi = u_s^{\frac{s}{p}}$. Then $\varphi \in W_0^{1,p}(\Omega)$, because $s > p$. Moreover $\int_\Omega g|\varphi|^p dx = \int_\Omega g u_s^s dx = 1$, and $\nabla\varphi = \frac{s}{p} \mid u_s \mid^{\frac{s}{p}-1} \nabla u_s$. Observe that φ is admissible to compute $\lambda_s(g)$ in (2.2). Hence

$$\lambda_p^{\frac{1}{p}}(g) \leq \left(\int_\Omega \mid \nabla\varphi \mid_a^p dx\right)^{\frac{1}{p}} = \frac{s}{p}\left(\int_\Omega u_s^{s-p} \mid \nabla u_s \mid_a^p dx\right)^{\frac{1}{p}}.$$

From Hölder's inequality, we obtain the estimate

$$\lambda_p^{\frac{1}{p}}(g) \leq \frac{s}{p}\left(\int_\Omega u_s^s dx\right)^{\frac{1}{p}-\frac{1}{s}} \lambda_s^{\frac{1}{s}}(g).$$

On the other hand, we have

$$\lambda_s \int_\Omega u_s^s dx \leq \int_\Omega \mid \nabla u_s \mid_a^s dx = \lambda_s(g)$$

by (1.3 ii) and the minimizing property of λ_s. Hence

$$\lambda_p^{\frac{1}{p}}(g) \leq \frac{s}{p}\left(\frac{\lambda_s(g)}{\lambda_s}\right)^{\frac{1}{p}-\frac{1}{s}} \lambda_s^{\frac{1}{s}}(g) = \frac{s}{p}\lambda_s^{\frac{1}{s}}\left(\frac{\lambda_s(g)}{\lambda_s}\right)^{\frac{1}{p}}. \qquad \square$$

REMARK 3.1. If $g \in L^\infty(\Omega)$, the inequality (3.1) holds for each $1 < p < s < \infty$. Notice that the one-sided limits

$$\lim_{s\to p_-} \lambda_s(g) \quad \text{and} \quad \lim_{s\to p_+} \lambda_s(g)$$

exist.

COROLLARY 3.1. *We have*

$$\limsup_{s\to p_-} \lambda_s(g) \leq \lambda_p(g) \leq \liminf_{s\to p_+} \lambda_s(g).$$

PROOF. • When $s \to p_+$, we have $p < s < p+1$. Thus

$$s\lambda_s^{\frac{1}{s}} \leq (p+1)\lambda_{p+1}^{\frac{1}{p+1}}.$$

Hence the set $\{\lambda_s \mid p < s < p+1\}$ is bounded. Thus $\lambda_s^{\frac{p}{s}-1} \to 1$, as $s \to p_+$. Finally, from the inequality (3.1), we deduce that

$$\lambda_p(g) \leq \liminf_{s \to p_+} \lambda_s(g).$$

• For $s \to p_-$, with $1 < s < p$, we have from (3.1), the following inequalities

$$\left[\left(\frac{s}{p}\right)^{p-s} \lambda_s^{\frac{p}{s}-1}\right] \lambda_s(g) \leq \lambda_s(g)\lambda_s^{1-\frac{s}{p}} \leq \left(\frac{s}{p}\right)^s \lambda_p(g).$$

The first inequality is (3.1) for $g \equiv 1$. Hence

$$\left(\frac{s}{p}\right)^p \lambda_s^{\frac{p}{s}-1}\lambda_s(g) \leq \lambda_p(g).$$

Therefore

$$\limsup_{s \to p_-} \left[\left(\frac{s}{p}\right)^p \lambda_s^{\frac{p}{s}-1}\lambda_s(g)\right] \leq \lambda_p(g).$$

On the other hand, since $\lambda_s^{\frac{p}{s}-1} \to 1$, as $s \to p_-$, we obtain that

$$\limsup_{s \to p_-} \lambda_s(g) \leq \lambda_p(g). \qquad \square$$

REMARK 3.2. Observe that if $\lim_{s \to p} \lambda_s(g)$ exists, then this limit is necessarily equal to $\lambda_p(g)$. Therefore we will study the different cases $s \to p_+$ and $s \to p_-$.

3.2 – Stability to the right

THEOREM 3.2. *For an arbitrary bounded domain we have*

$$\lim_{s \to p_+} \lambda_s(g) = \lambda_p(g).$$

PROOF. Let $\varphi \in \mathcal{C}_0^\infty(\Omega)$ be such that

$$\int_\Omega g \mid \varphi \mid^p dx > 0, \tag{3.2}$$

and let $\epsilon > 0$ (small). Applying the Dominated Convergence Theorem, we find

$$\lim_{\epsilon \to 0_+} \int_\Omega g \mid \varphi \mid^{p+\epsilon} dx = \int_\Omega g \mid \varphi \mid^p dx > 0.$$

Hence, there is $\epsilon_0 > 0$ sufficiently small such that

$$\int_\Omega g \mid \varphi \mid^{p+\epsilon} dx > 0, \text{ when } 0 < \epsilon < \epsilon_0.$$

On the other hand, we have

$$\lambda_{p+\epsilon}(g) \leq \frac{\int_\Omega \mid \nabla \varphi \mid_a^{p+\epsilon} dx}{\int_\Omega g \mid \varphi \mid^{p+\epsilon} dx}.$$

It follows from the Dominated Convergence Theorem that

$$\limsup_{\epsilon \to 0_+} \lambda_{p+\epsilon}(g) \leq \frac{\int_\Omega \mid \nabla \varphi \mid_a^p dx}{\int_\Omega g \mid \varphi \mid^p dx}. \tag{3.3}$$

This, and the fact that φ is an arbitrary function satisfying (3.2), yield

$$\limsup_{\epsilon \to 0_+} \lambda_{p+\epsilon}(g) \leq \lambda_p(g).$$

Now the result follows from Corollary 3.1. □

THEOREM 3.3. *The principal eigenfunctions u_s associated with $\lambda_s(g)$ satisfy*

$$\lim_{s \to p_+} \int_\Omega \mid \nabla u_s - \nabla u_p \mid_a^p dx = 0. \tag{3.4}$$

PROOF. For $1 < p < s$ with s near p. Hölder's inequality implies that

$$\int_\Omega \mid \nabla u_s \mid_a^p \, dx \leq \mid \Omega \mid^{1-\frac{p}{s}} (\lambda_s(g))^{\frac{p}{s}} . \tag{3.5}$$

This shows that $\{u_s,\, s > p\}$ is a bounded set in $W_0^{1,p}(\Omega)$. Hence there is a sequence $s_1, s_2, ...$, converging to p_+ and there is a function $u \in W_0^{1,p}(\Omega)$ such that $u_{s_j} \rightharpoonup u$ (weakly) in $W_0^{1,p}(\Omega)$, as $j \to +\infty$. Using the Rellich-Kondrachov Compactness Theorem, (cf.[1, p.144]), we obtain that $u_{s_j} \to u$ in $L^{p+\frac{1}{N}}(\Omega)$, as $j \to +\infty$; in particular, $u_{s_j} \to u$ in $L^p(\Omega)$, as $j \to +\infty$. Passing to a subsequence if necessary, we can assume that $u_{s_j} \to u$ a.e. in Ω. We will prove that $u \equiv u_p$. The weak lower semicontinuity of the norm and (3.5) yield

$$\int_\Omega \mid \nabla u \mid_a^p \, dx \leq \lambda_p(g). \tag{3.6}$$

It suffices to have

$$\int_\Omega g u^p \, dx = 1.$$

Indeed, if we set $M_s = \max_\Omega u_s$, then from Lemma 4.1., we have $\max_{s\in[a,b]} M_s < M < \infty$. Here M is a constant not depending on s, and $[a, b]$ is any small interval containing p. Thus $0 < u_{s_j} \leq M$, and $0 \leq u \leq M$ a.e. on Ω. Hence

$$\mid g \mid \mid u_{s_j}^{s_j} - u^p \mid \leq \mid g \mid u_{s_j}^{s_j} + \mid g \mid u^p \leq \mid g \mid M^{s_j} + \mid g \mid u^p$$

a.e. on Ω. On the other hand, $M^{s_j} \leq M^{p+1} + 1$. Thus a.e. on Ω, we have

$$\mid g \mid \mid u_{s_j}^{s_j} - u^p \mid \leq \mid g \mid (M^{p+1} + 1 + M^p) \in L^1(\Omega).$$

The Dominated Convergence Theorem yields

$$\left| \int_\Omega g(u_{s_j}^{s_j} - u^p) dx \right| \leq \int_\Omega \mid g \mid \mid u_{s_j}^{s_j} - u^p \mid dx \to 0,$$

as $j \to +\infty$, since $g(u_{s_j}^{s_j} - u^p) \to 0$, a.e. in Ω, as $j \to +\infty$. From this it follows easily that $\int_\Omega g \mid u \mid^p dx = 1$. Finally, (3.6) and the variational characterization of $\lambda_p(g)$ yield

$$\int_\Omega \mid \nabla u \mid_a^p \, dx \, = \, \lambda_p(g).$$

By the uniqueness of the principal eigenfunction we have $u = u_p$. Thus the limit function u does not depend on the particular (sub)sequence $s_1, s_2, ...$ Therefore $u_s \to u_p$ at least in $L^p(\Omega)$, as $s \to p_+$.

The rest of the proof, i.e., the strong convergence (3.4) can be obtained from Clarkson's inequalities, (cf.[1]); but with the $||\,|\,|_a||_p$-norm in $W_0^{1,p}(\Omega)$. ∎

3.3 – Stability to the left

This case is more difficult, because if $u \in W_0^{1,p-\epsilon}(\Omega)$ then it is possible that $u \notin W_0^{1,p}(\Omega)$.

THEOREM 3.4. *Let Ω be an arbitrary bounded domain. If we suppose that*

$$\lim_{s\to p_-} \int_\Omega \mid \nabla u_s - \nabla u_p \mid_a^s \, dx = 0, \tag{3.7}$$

then we have

$$\lim_{s\to p_-} \lambda_s(g) = \lambda_p(g).$$

PROOF. (3.7) and the Hölder inequality imply

$$\lim_{s\to p_-} \int_\Omega \mid \nabla u_s - \nabla u_p \mid_a^{p-\epsilon} \, dx = 0,$$

for any $\epsilon > 0$ sufficiently small so that $0 < p - s < \epsilon < p - 1$. Therefore $\nabla u_s \to \nabla u_p$ in $(L^{p-\epsilon}(\Omega))^N$, as $s \to p_-$. For $\epsilon > 0$ small enough, the Hölder's inequality implies that

$$|| \mid \nabla u_s \mid_a ||_{p-\epsilon} \leq \mid \Omega \mid^{\frac{s+\epsilon-p}{s(p-\epsilon)}} || \mid \nabla u_s \mid_a ||_s \, .$$

Hence

$$|| \mid \nabla u_p \mid_a ||_{p-\epsilon} \leq \mid \Omega \mid^{\frac{\epsilon}{p(p-\epsilon)}} \liminf_{s\to p_-} \lambda_s^{\frac{1}{s}}(g). \tag{3.8}$$

Letting $\epsilon \to 0^+$, the Fatou lemma yields

$$\lambda_p^{\frac{1}{p}}(g) = || \mid \nabla u_p \mid_a ||_p \leq \liminf_{s\to p_-} \lambda_s^{\frac{1}{s}}(g).$$

This completes the proof, in view of Corollary 3.1. □

REMARK 3.3. The converse of the theorem is an open question in the case $p \leq N$.

However, we have the following partial result for any bounded domain and every p in $(1, +\infty)$.

THEOREM 3.5. *Under the same assumptions, suppose that $\lim_{s\to p_-} \lambda_s(g) = \lambda_p(g)$. Then each sequence of real numbers tending to p from below contains a subsequence $s_1, s_2, \ldots$ such that*

$$\lim_{j\to+\infty} \int_\Omega \mid \nabla u_{s_j} - \nabla u \mid_a^{s_j} \, dx \, = \, 0, \tag{3.9}$$

for some function $u \in W^{1,p}(\Omega) \cap W_0^{1,p-\epsilon}(\Omega)$, whenever $\epsilon > 0$, $\int_\Omega g \mid u \mid^p dx = 1$, $u \geq 0$ a.e. on Ω and $\int_\Omega \mid \nabla u \mid_a^p dx \leq \lambda_p(g)$. The function u may be depend on the sequence, but it is a weak solution to the equation

$$A_p u + \lambda_p(g)|u|^{p-2}u = 0.$$

PROOF. Let us fix $\epsilon_0 > 0$ small enough, so that $0 < p - \epsilon_0 < s < p$ and that $(p+\epsilon) < (p-\epsilon)^*$ for $0 < \epsilon < \epsilon_0$; where for $t \in (1, +\infty)$, $t^* = \frac{Nt}{N-t}$ if $1 < t < N$ and $t^* = +\infty$ if $t \geq N$. Using Hölder's inequality, we obtain

$$|| \mid \nabla u_s \mid_a ||_{p-\epsilon} \leq \mid \Omega \mid^{\frac{s+\epsilon-p}{s(p-\epsilon)}} \lambda_s^{\frac{1}{s}}(g),$$

when $0 < \epsilon < \epsilon_0$. From (3.1), we conclude that the norms $|| \mid \nabla u_s \mid_a ||_{p-\epsilon}, 0 < \epsilon < \epsilon_0$, are uniformly bounded, in view of the assumption $\lim\limits_{s \to p_-} \lambda_s(g) = \lambda_p(g)$. Thus we can find a function $u \in W_0^{1,p-\epsilon}(\Omega), 0 < \epsilon < \epsilon_0$; and find a sequence $s_1, s_2, ...$ converging to p_- such that $u_{s_j} \rightharpoonup u$ (weakly) in $W_0^{1,p-\epsilon}(\Omega)$, as $j \to +\infty$, for each $\epsilon \in (0, \epsilon_0)$ and hence $u_{s_j} \to u$ in $L^{p+\epsilon}(\Omega)$. Passing to a subsequence if necessary, we can assume that $u \geq 0$ a.e. on Ω. Clearly $u \in L^p(\Omega)$ and is independent of ϵ. On the other hand, the weak lower semicontinuity of the norm and the assumption $\lim\limits_{j \to +\infty} \lambda_{s_j}(g) = \lambda_p(g)$ imply that

$$|| \mid \nabla u \mid_a ||_{p-\epsilon} \leq \mid \Omega \mid^{\frac{\epsilon}{p(p-\epsilon)}} \lambda_p^{\frac{1}{p}}(g).$$

Then letting $\epsilon \to 0^+$, we obtain with Fatou's lemma that $\nabla u \in (L^p(\Omega))^N$ and

$$|| \mid \nabla u \mid_a ||_p \leq \lambda_p^{\frac{1}{p}}(g). \tag{3.10}$$

The normalization: $\int_\Omega g \mid u \mid^p dx = 1$, is preserved and

$$\lim_{j \to +\infty} \int_\Omega g \left(\frac{u_{s_j} + u}{2} \right)^{s_j} dx = 1, \tag{3.11}$$

for a subsequence if necessary. On the other hand, we have

$$\lambda_{s_j}(g) \leq \frac{\int_\Omega \mid \frac{\nabla u_{s_j} - \nabla u}{2} \mid_a^{s_j} dx}{\int_\Omega g (\frac{u_{s_j}+u}{2})^{s_j} dx}, \tag{3.12}$$

for j sufficiently large, because by (3.11) there is an index j_0 so large that

$$\int_\Omega g \left(\frac{u_{s_j}+u}{2} \right)^{s_j} dx > 0,$$

when $j \geq j_0$. Clarkson's inequality yields

$$\int_\Omega \left| \frac{\nabla u_{s_j} - \nabla u}{2} \right|_a^{s_j} dx \leq \frac{1}{2} \lambda_{s_j}(g) + \frac{1}{2} || \mid \nabla u \mid_a ||_{s_j}^{s_j} - \lambda_{s_j}(g) \int_\Omega g \left(\frac{u_{s_j} + u}{2} \right)^{s_j} dx,$$

if $s_j \geq 2$. Now (3.11) and the assumption $\lim_{s\to p_-} \lambda_s(g) = \lambda_p(g)$ imply

$$\limsup_{j\to+\infty} \int_\Omega \left| \frac{\nabla u_{s_j} - \nabla u}{2} \right|_a^{s_j} dx \leq \frac{1}{2} \, ||\, |\, \nabla u \,|_a||_p^p - \frac{1}{2}\lambda_p(g).$$

From this and (3.10), it follows easily that

$$\lim_{j\to+\infty} \int_\Omega \, |\, \nabla u_{s_j} - \nabla u \,|_a^{s_j} \, dx \, = \, 0,$$

for the case $p > 2$.

For the case $1 \leq p \leq 2$, we argue as follows. There is $j_1 \in \mathbb{N}$ such that $1 \leq s_j \leq 2$, for each $j \geq j_1$. Let $j_2 = max(j_1, j_0)$. Then Clarckson's inequality associated with s_j and (3.12) yield

$$\left\{ \int_\Omega |\frac{\nabla u_{s_j} - \nabla u}{2}|_a^{s_j} dx \right\}^{\frac{1}{s_j - 1}} + \left\{ \lambda_{s_j}(g) \int_\Omega g(\frac{u_{s_j} + u}{2})^{s_j} dx \right\} \leq$$
$$\leq \left\{ \frac{1}{2}\lambda_{s_j}(g) + \frac{1}{2} \int_\Omega |\nabla u|_a^{s_j} dx \right\}^{\frac{1}{s_j - 1}}.$$

On the other hand from Hölder's inequality and (3.10) we deduce that

$$\int_\Omega |\nabla u|_a^{s_j} dx \leq |\Omega|^{\frac{p-s_j}{p}} \lambda_p(g)^{\frac{s_j}{p}}.$$

Thus

$$\left\{ \int_\Omega |\frac{\nabla u_{s_j} - \nabla u}{2}|_a^{s_j} dx \right\}^{\frac{1}{s_j - 1}} \leq \left\{ \frac{1}{2}\lambda_{s_j}(g) + \frac{1}{2}|\Omega|^{\frac{p-s_j}{p}} \lambda_p(g)^{\frac{s_j}{p}} \right\}^{\frac{1}{s_j - 1}} +$$
$$- \left\{ \lambda_{s_j}(g) \int_\Omega g \left(\frac{u_{s_j} + u}{2} \right)^{s_j} dx \right\}^{\frac{1}{s_j - 1}}.$$

Now, (3.11) and the assumption $\lim_{s\to p_-} = \lambda_p(g)$ imply that

$$\left\{ \limsup_{j\to+\infty} \int_\Omega |\frac{\nabla u_{s_j} - \nabla u}{2}|_a^{s_j} dx \right\}^{\frac{1}{p-1}} \leq \left\{ \frac{1}{2}\lambda_p(g) + \frac{1}{2}\lambda_p(g) \right\}^{\frac{1}{p-1}} - \lambda_p(g)^{\frac{1}{p-1}} = 0.$$

Hence

$$\lim_{j\to+\infty} \int_\Omega |\nabla u_{s_j} - \nabla u|_a^{s_j} dx.$$

REMARK 3.4. (i) If the limit function $u \in W_0^{1,p}(\Omega)$, then $u \equiv u_p$ by the uniqueness of the principal eigenfunction and (3.10).

(ii) When $p \leq N$, in a very irregular domain the defect $\lim\limits_{s \to p_-} \lambda_s(g) < \lambda_p(g)$ is possible. See the counterpart in [10] for the case

$$\Delta_p u + \lambda |u|^{p-2} u = 0.$$

COROLLARY 3.2. *For any bounded domain Ω having the segment property, we have*

$$\lim_{s \to p_-} \lambda_s(g) = \lambda_p(g)$$

if and only if

$$\lim_{s \to p_-} \int_\Omega \mid \nabla u_s - \nabla u_p \mid_a^s \, dx \, = \, 0.$$

PROOF. Suppose that $\lim\limits_{s \to p_-} \lambda_s(g) = \lambda_p(g)$. From Theorem 3.5, the limit function u satisfies for $\epsilon > 0$ small enough $u \in W^{1,p}(\Omega) \cap W_0^{1,p-\epsilon}(\Omega)$ such that

$$u \geq 0 \ \text{ a.e. in } \ \Omega, \int_\Omega g \mid u \mid^p \, dx = 1 \ \text{ and } \ \int_\Omega \mid \nabla u \mid_a^p \, dx \leq \lambda_p(g).$$

Since Ω has the segment property, thus $u \in W_0^{1,p}(\Omega)$ by Theorem 2.1. Thus u is admissible in the definition of $\lambda_p(g)$. Consequently,

$$\lambda_p(g) = \int_\Omega \mid \nabla u \mid_a^p \, dx.$$

Hence $u \equiv u_p$ by the uniqueness of the principal eigenfunction. So by (3.9), we obtain

$$\lim_{s \to p_-} \int_\Omega \mid \nabla u_s - \nabla u_p \mid_a^s \, dx \, = \, 0,$$

since the limit function does not depend on the choice of the sequence. The converse is immediate, in view of Theorem 3.4. ∎

Using the $C^{1,\alpha}_{\text{loc}}$-regularity of the principal eigenfunctions u_s, s proved in [6] and the L^∞-estimate to be established in Lemma 4.1., we can state the following result generalizing [10].

THEOREM 3.6. *Assume that the conditions on g and $a_{i,j}$ are satisfied. Then each sequence converging to p_-, contains a subsequence $s_1, s_2, ...$ such that $u_{s_j} \to u$ and $\nabla u_{s_j} \to \nabla u$ locally uniformly, where u is some function in $C^1(\Omega)$. Moreover, u is a weak solution of the equation*

$$(\mathcal{E}) \qquad A_p u + \lambda g(x) \mid u \mid^{p-2} u = 0 \quad ,$$

where $\lambda = \lim\limits_{j \to +\infty} \lambda_{s_j}(g)$.

We know that only the first eigenfunctions are not changing signs. Thus if λ is an eigenvalue of $(\mathcal{E})$, then $\lambda = \lambda_p(g)$, and by normalization, we have $u \equiv u_p$. We have come to an important point: though the limit function u of $\{u_s\}$, as $s \to p_-$, is in $\in W^{1,p}(\Omega) \cap W_0^{1,p-\epsilon}(\Omega)$, for any $\epsilon > 0$ chosen sufficiently small, it is not always the right eigenfunction u_p, i.e., u is not necessary in $W_0^{1,p}(\Omega)$. Therefore u is not admissible in the definition of $\lambda_p(g)$. But, If Ω satisfies the segment property, then $u = u_p$, $\lambda = \lambda_p(g)$ and

$$\lim_{s \to p_-} \lambda_s(g) = \lambda_p(g).$$

So we have the following result.

COROLLARY 3.3. *For any bounded domain Ω having the segment property, we have*

$$\lim_{s \to p} \lambda_s(g) = \lambda_p(g).$$

4 – Appendix

The technique to uniformly bound u_p in an arbitrary domain is originally due to Ladyzhenskaya and Urlatseva, cf. [7].

LEMMA 4.1. *Let the assumptions on g and $a_{i,j}$ be fulfilled. Then for any bounded domain Ω, $\max_\Omega u_p$ is bounded uniformly in p, (u_p denotes the normalized principal eigenfunction).*

PROOF. If $p > N$, then from [5, Theorem 3.18., p.240] we have

$$|| u_p ||_\infty \leq C \mid \Omega \mid^{\frac{1}{N} - \frac{1}{p}} || \nabla u_p ||_p \leq C \mid \Omega \mid^{\frac{1}{N} - \frac{1}{p}} \lambda_p^{\frac{1}{p}}(g),$$

where $C = \frac{1}{N}[\frac{1-\frac{1}{p}}{\frac{1}{N}-1}]^{\frac{1}{p'}} \omega_N^{\frac{1}{N}}$, and ω_N is the volume of the unit ball in $\mathbb{R}^N$.

For $1 < p \leq N$, we keep track of various "constants" in Proposition 2.16. of [15]; we obtain the lower bound

$$\|u_p\|_{\infty,\Omega} \leq b \|\nabla u_p\|_{1,\Omega}\,;$$

where

$$b = (2^p \lambda_p(g) \|g\|_{r,\Omega})^{\frac{Nr}{pr-N}} (C^\theta + C^{\frac{\theta}{p}} \omega_N^{\frac{\theta(1-p)}{N}})^{1 - \frac{pr-N}{(p-1)Nr}},$$

$$\theta = \frac{Nr}{pr(1+N) - N(1+r)}$$

and $C = (\frac{(N-1)p}{2(N-p)\sqrt{N}})^p$ if $1 < p < N$, and $C = [\max\{N, \frac{r(N-1)}{r-1}\}]^{\frac{N}{r}}$ if $p = N$. This concludes the proof of the Lemma. ∎

Acknowledgement

The authors are indebted to Professor J.P. Gossez for a helpful discussion about the segment property.

REFERENCES

[1] R. Adams: *Sobolev Spaces*, Academic Press, New-York, 1975.

[2] H. Amann: *Ljusternik-Schnirelman theory and nonlinear eigenvalue problems*, Math. Ann., **199** (1972), 55-72.

[3] A. Anane: *Simplicité et isolation de la première valeur propre du p-Laplacien*, C. R. Acad. Sci. Paris, **305** (1987), 725-728.

[4] E. Di Benedetto: *$C^{1+\alpha}$-local regularity of weak solutions of degenerate elliptic equations*, Nonlinear Analysis T.M.A., **7** (1983), 827-859.

[5] D. E. Edmunds – W. D. Evans: *Spectral Theory and Differential Operators*, Clarendon Press-Oxford, 1990.

[6] A. El Khalil: *Sur le problème nonlinaire A_p-Laplacien: Stabilité-Bifurcation*, thèse de 3ième cycle, Faculté des siences Dhar-Mahraz Fés, 1996.

[7] O. Ladyzheskaya – N. Uralatseva: *Linear and Quasilinear Elliptic Equation*, Academic Press, New York, 1968.

[8] E. Lami Dozo – A. Touzani: *Autovalores con peso indefinito del A_p-Laplaciano*, Centro latinoamericano de Matematica e Informatica CLAMI, 1992.

[9] P. Lindqvist: *On the equation* $\mathrm{div}(|\nabla u|^{p-2}\nabla u) + \lambda |u|^{p-2} u = 0$, Proc. of Amer. Math. Soc., **109** (1990), 157-164.

[10] P. Lindqvist: *On Non-Linear Rayleigh Quotients*, Potential Analysis, **2** (1993), 199-218.

[11] J. Mossino: *Inégalités isopémetriques et applications en physique*, Paris, Hermann, 1984.

[12] M. Otani – T. Theshima: *On the first eigenvalue of some quasilinear elliptic equations*, Pro. Japon Acad., **64**, Ser. A (1988), pp. 8-10.

[13] Yu. G. Reshetnyak: *Set of singular points of solutions of certain nonlinear elliptic equations*, (in Russian), Sibirskij Mat. Z., **9** (1968), 354-368.

[14] S. Sakaguchi: *Concavity properties of solutions to some degenerate quasilinear elliptic Direchlet problem*, Annali della Scuola Normale Superiore de Pisa, Serie IV (Classe di Scienze), **14** (1987).

[15] A. Touzani: *Quelques résultats sur le A_p-Laplacien avec poids indefini*, thèse de Doctorat, U.L.B. (1991-1992).

Lavoro pervenuto alla redazione il 31 gennaio 2003
ed accettato per la pubblicazione il 23 marzo 2004.
Bozze licenziate il 6 dicembre 2004

Indirizzo degli Autori:

Abdelouahed El Khalil – Department of Mathematics and Industrial Genie – Polytechnic School of Montreal – P.O. 6079, Succ. Centre-Ville – Montréal (Quebec) H3C 3A7
E-mail: lkhlil@hotmail.com

Peter Lindqvist – Department of Mathematical Sciences – University of Science and Technology – N-7491 Trondheim, Norway
E-mail: lqvist@math.ntnu.no

Abdelfattah Touzani – Department of mathematics – Faculty of sciences Dhar-Mahraz – P.O. Box 1796 Atlas – Fez 30000, Morocco
E-mail: atouzani@iam.net.ma