



**HAL**  
open science

## Introduction to spectral methods

Philippe Grandclement

► **To cite this version:**

Philippe Grandclement. Introduction to spectral methods. Stellar fluid dynamics and numerical simulations: from the sun to neutron stars, 2006, France. pp.153. hal-00091367

**HAL Id: hal-00091367**

**<https://hal.science/hal-00091367>**

Submitted on 6 Sep 2006

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Introduction to spectral methods

Philippe Grandclément

Laboratoire Univers et ses Théories, Observatoire de Paris,  
5 place J. Janssen, 92195 Meudon Cedex, France

This proceeding is intended to be a first introduction to spectral methods. It is written around some simple problems that are solved explicitly and in details and that aim at demonstrating the power of those methods. The mathematical foundation of the spectral approximation is first introduced, based on the Gauss quadratures. The two usual basis of Legendre and Chebyshev polynomials are then presented. The next section is devoted to one dimensional equation solvers using only one domain. Three different methods are described. Techniques using several domains are shown in the last section of this paper and their various merits discussed.

## I. INTRODUCTION

This proceeding constitutes a short introduction to *spectral methods*. The aim is not to present an exhaustive mathematical presentation of those methods. Numerous books can be consulted for this purpose (see the bibliography for a sample of them). The following material should instead be considered as a toolkit for implementing simple spectral methods solvers. Thus, a particular emphasize will be put on practical examples. For the sake of simplicity, we will restrict ourselves to one dimensional solvers.

Spectral methods are just one of the many ways to represent a function on a computer. The basic idea of all numerical techniques is to approximate any function by polynomials, those being the only functions than a computer can exactly calculate. So a function  $u$  will be approximate by  $\hat{u} = \sum_{n=0}^N \hat{u}_n \Phi_n$  where the  $\Phi_n$  are polynomials and called the *trial functions*. Depending on the choice of trial functions, one can generate various classes of numerical techniques. For example, the *finite difference* schemes are obtained by choosing local polynomials of low degree. In all the following, we will be interested in spectral methods where the  $\Phi_n$  are global polynomials, typically Legendre or Chebyshev ones. If spectral methods are basically more evolved than finite difference schemes, they have long prove their ability to tackle a wide variety of problems. In particular, they allow to reach very good accuracy with only moderate computational resources.

In the first part of this article, we present the basic properties of spectral methods. After a brief overview of the foundations of the spectral expansion, we show the power of the method on simple but non-trivial examples. In particular, we show that, for  $C^\infty$  functions, the error decays exponentially, as one increases the degree of the approximation. We then present some basic features of the two type of polynomials that are used in all the rest of this article, Legendre and Chebyshev polynomials.

The second section is devoted to the actual implementation of partial differential equation solvers. We restrict ourselves to one dimensional equations on a bounded domain (i.e.  $x \in [-1, 1]$ ). Three different types of solvers are presented: the *Tau-method*, the *collocation method* and the *Galerkin method*. They are explicitly implemented on a simple example.

The last part of this work is concerned with the implementation of *multi-domain* solvers. After having explained why this is often a valuable tool for the physicist, we present, once again, three different methods, all of them being tested an implemented on a simple, one dimensional equation, with a decomposition of space using two domains.

## II. FOUNDATIONS OF SPECTRAL METHODS

### A. Spectral expansion

#### 1. Orthogonal projection

Let us consider an interval  $\Lambda = [x_{\min}, x_{\max}]$ . In order to talk about basis, one needs to define a scalar product on  $\Lambda$ . If  $w$  is a positive function on  $\Lambda$ , one can define the scalar product of two functions  $f$  and  $g$ , with respect to the *measure*  $w$  as being

$$(f, g)_w = \int_{\Lambda} f(x) g(x) w(x) dx. \quad (1)$$

Using this scalar product, one can find a set of orthogonal polynomials  $p_n$ , each of them of degree  $n$ . The set composed of those polynomials, up to a given degree  $N$  is a basis of  $P_N$ .

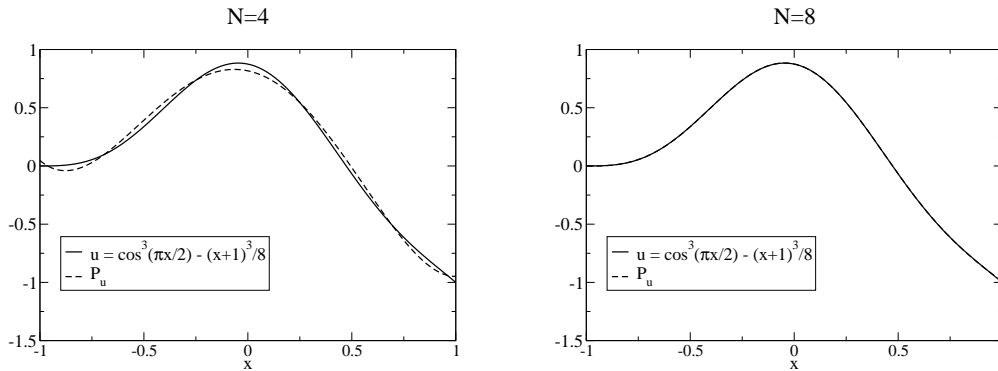


FIG. 1:  $u = \cos^3(\pi x/2) - (x+1)^3/8$  and its projection of Chebyshev polynomials of degree smaller than 4 (left panel) and 8 (right panel).

One can then hope to represent any function  $u$  on  $\Lambda$  by its projection on the polynomials  $p_n$ . Doing so, we define the projection of  $u$  simply by

$$P_N u = \sum_{n=0}^N \hat{u}_n p_n(x), \quad (2)$$

where the coefficients of the projections are given by  $\hat{u}_n = \frac{(u, p_n)}{(p_n, p_n)}$ . The difference between  $u$  and its projection is called the *truncation error* and one can show that it goes to zero when the order of the approximation increases :

$$\|u - P_N u\| \longrightarrow 0 \quad \text{when} \quad N \longrightarrow \infty. \quad (3)$$

This convergence is illustrated on Fig. 1, where a test function  $u$  is plotted, with its projection for  $N = 4$  and  $N = 8$ . Chebyshev polynomials are used. Let us note that, even if  $u$  is not a polynomial function, no discrepancy can be seen by eye, for  $N$  as small as 8.

This seems very appealing but for the fact that one needs to calculate the  $\hat{u}_n$  by computing integrals like  $\int_{\Lambda} u(x) p_n(x) w(x) dx$ . If one needs to evaluate  $u$  at a lot of points in order to do so (like when using standard Newton method), one would lose all the advantage of using spectral expansion.

## 2. Gauss quadratures

The solution to this problem is given by the Gauss quadratures. The theorem can be stated as follows.

*There exist  $N + 1$  positive reals  $w_n$  and  $N + 1$  reals  $x_n$  in  $\Lambda$  such that :*

$$\forall f \in \mathbb{P}_{2N+\delta}, \quad \int_{\Lambda} f(x) w(x) dx = \sum_{n=0}^N f(x_n) w_n. \quad (4)$$

The  $w_n$  are called the *weights* and the  $x_n$  the collocation points. The exact degree of applicability depends on the quadrature. The three usual choices are :

- Gauss :  $\delta = 1$
- Gauss-Radau :  $\delta = 0$  and  $x_0 = x_{\min}$ .
- Gauss-Lobatto :  $\delta = -1$  and  $x_0 = x_{\min}$  and  $x_N = x_{\max}$ .

Gauss quadrature is the best possible choice in terms of degree (it is not possible to find a quadrature with  $\delta > 1$ ). However, the other quadratures have the property that the boundaries of the interval coincides with collocation points, which can be useful, to enforce boundary conditions for example. In all the following, we will use the Gauss-Lobatto quadrature.

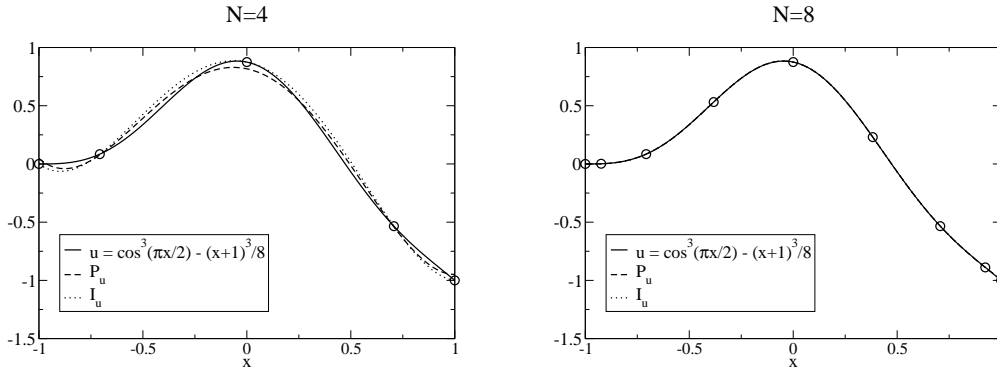


FIG. 2: Same as Fig. 1 with also the interpolant of  $u$ . The collocation points are denoted by the circles.

### 3. Interpolation

If one applies the Gauss quadratures to approximate the coefficient of the expansion, one obtains

$$\tilde{u}_n = \frac{1}{\gamma_n} \sum_{j=0}^N u(x_j) p_n(x_j) w_j \quad \text{with} \quad \gamma_n = \sum_{j=0}^N p_n^2(x_j) w_j. \quad (5)$$

Let us precise that this is not exact in the sense that  $\hat{u}_n \neq \tilde{u}_n$ . However, the computation of  $\tilde{u}$  only requires to evaluate  $u$  at the  $N + 1$  collocation points. The *interpolant* of  $u$  is then defined as the following polynomial

$$I_N u = \sum_{n=0}^N \tilde{u}_n p_n(x). \quad (6)$$

The difference between  $I_N u$  and  $P_n u$  is called the *aliasing error*. The interpolant of  $u$  is the spectral approximate of  $u$  and one can show that it is the only polynomials of degree  $N$  that coincides with  $u$  at each collocation point :

$$[I_N u](x_i) = u(x_i) \quad \forall i \leq N. \quad (7)$$

Figure 2 shows the same function as Fig.1 but the interpolant is also plotted. One can see that, indeed  $I_N u$  coincides with  $u$  at the collocation points that are indicated by the circles. Once again, even with as few points as  $N = 8$ , no difference can be seen between the various functions.

Figure 3 shows the maximum difference between  $I_N u$  and  $u$  on  $\Lambda$ , as a function of the degree of the approximation  $N$ . We can observe the very general feature of spectral methods that the error decreases exponentially, until one reaches the machine accuracy (here  $10^{-14}$ , the computation being done in double precision). This very fast convergence explains why spectral methods are so efficient, especially compared to finite difference ones, where the error follows only a power-law in terms of  $N$ . We will later be more quantitative about the convergence properties of the spectral approximation.

Let us note that a function  $u$  can be described either by its value at each collocation point  $u(x_i)$  or by the coefficients of the interpolant  $\tilde{u}_i$ . If the values at collocation points are known one is working in the *configuration space* and in the *coefficient space* if  $u$  is given in terms of its coefficients. There is a bijection between the two descriptions and one simply goes from one space to another by using :

- $\tilde{u}_n = \frac{1}{\gamma_n} \sum_{j=0}^N u(x_j) p_n(x_j) w_j$  (configuration  $\rightarrow$  coefficient)
- $u(x_n) = \sum_{j=0}^N \tilde{u}_j p_j(x_n)$  (coefficient  $\rightarrow$  configuration)

Depending on the operation one has to perform, one choice of space is usually more suited than the other. For instance, let us assume that one wants to compute the derivative of  $u$ . This is easily done if  $u$  is known in the

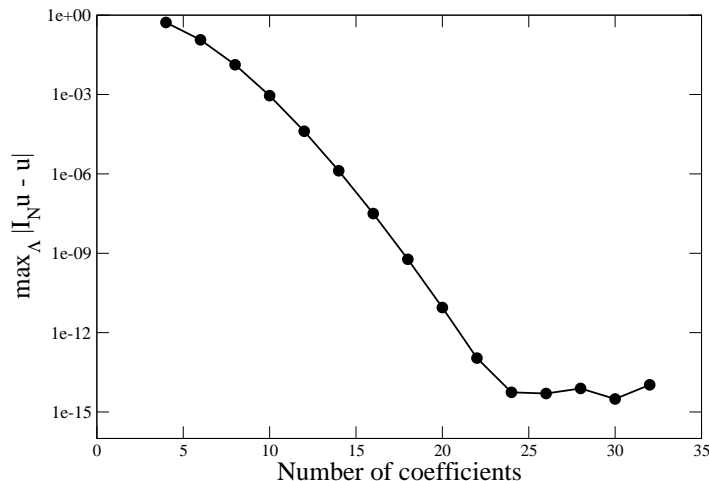


FIG. 3: Maximum difference between  $I_N u$  and  $u$  as a function of the degree of the approximation  $N$ .

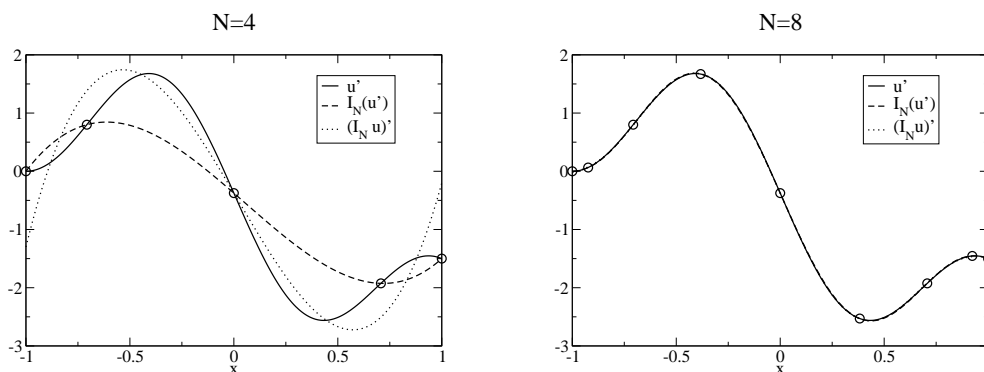


FIG. 4: First derivative of  $u = \cos^3(\pi x/2) - (x+1)^3/8$ , interpolant of the derivative and derivative of the interpolant, for  $N = 4$  and  $N = 8$ .

coefficient space. Indeed, one can simply approximate  $u'$  by the derivative of the interpolant :

$$u'(x) \approx [I_N u]' = \sum_{n=0}^N \tilde{u}_n p'_n(x). \quad (8)$$

Such an approximation only requires the knowledge of the coefficients of  $u$  and how the basis polynomials are derived. Let us note that the obtained polynomial, even if it is a good approximation of  $u'$ , is not the interpolant of  $u'$ . In other terms, the interpolation and the derivation are two operations that do not commute:  $(I_N u)' \neq I_N(u')$ . This is clearly illustrated on Fig. 4 where the derivative of  $u = \cos^3(\pi x/2) - (x+1)^3/8$  is plotted along with the functions  $(I_N u)'$  and  $I_N(u')$ . In particular, one can note that the functions used to represent  $u'$ , i.e.  $(I_N u)'$  does not coincide with  $u'$  at the collocation points. However, even with  $N = 8$  only, the three functions can not be distinguished by eye.

The maximum difference between  $u'$  and  $(I_N u)'$ , as a function of  $N$ , is shown on Fig. 5. Once, again, the convergence is exponential.

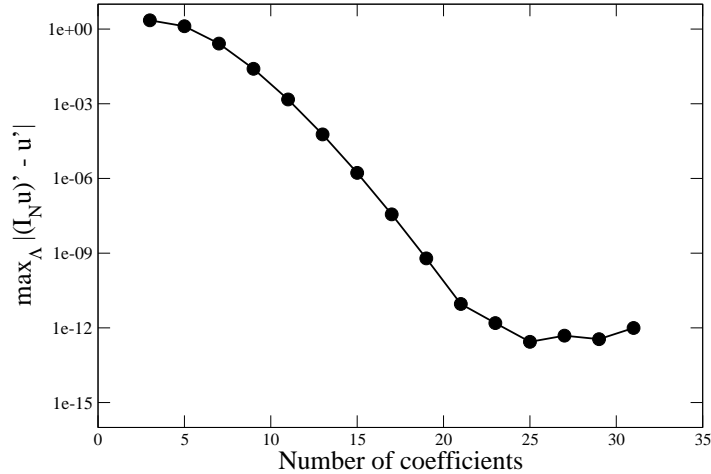


FIG. 5: Maximum difference between  $(I_N u)'$  and  $u'$  as a function of the degree of the approximation  $N$ .

## B. Usual families of polynomials

In this section, we briefly present some basic properties of Legendre and Chebyshev polynomials. Those two sets are the usual choice for non periodic problems.

### 1. Legendre polynomials

The Legendre polynomials, denoted by  $P_n$ , constitute a family of orthogonal polynomials on  $[-1, 1]$  with a measure  $w = 1$ . The fact that the measure is so simple is one of the main advantage of working with Legendre polynomials, especially from the analytical point of view.

The scalar product of two  $P_n$  is given by :

$$\int_{-1}^1 P_n P_m dx = \frac{2}{2n+1} \delta_{mn}. \quad (9)$$

The successive polynomials can be constructed by recurrence. Indeed given that  $P_0 = 1$  and  $P_1 = x$ , all the  $P_n$  can be obtained by using

$$(n+1) P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x). \quad (10)$$

It is then easy to see that the Legendre polynomials have the following simple properties : i)  $P_n$  has the same parity as  $n$ . ii)  $P_n$  is of degree  $n$ . iii)  $P_n(\pm 1) = (\pm 1)^n$ . iv)  $P_n$  has exactly  $n$  zero on  $[-1, 1]$ . The first polynomials are plotted on Fig. 6.

The values of the weights and collocation points can be written for the three usual quadratures :

- Legendre-Gauss :  $x_i$  are the nodes of  $P_{N+1}$  and  $w_i = \frac{2}{(1-x_i^2)[P'_{N+1}(x_i)]^2}$ .
- Legendre-Gauss-Radau :  $x_0 = -1$  and the  $x_i$  are the nodes of  $P_N + P_{N+1}$ . The weights are given by  $w_0 = \frac{2}{(N+1)^2}$  and  $w_i = \frac{1}{(N+1)^2}$ .
- Legendre-Gauss-Lobatto :  $x_0 = -1$ ,  $x_N = 1$  and  $x_i$  are the nodes of  $P'_N$ . The weights are  $w_i = \frac{2}{N(N+1)[P'_N(x_i)]^2}$ .

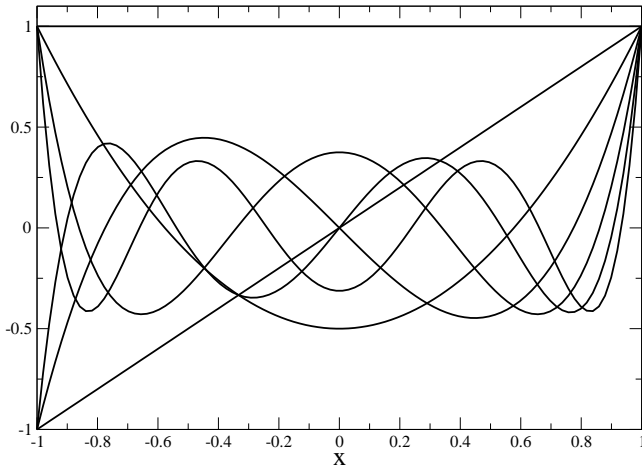


FIG. 6: The first Legendre polynomials, from  $P_0$  to  $P_6$ .

From the above formula, one can see that the position of the collocation points are not analytical and they have to be computed numerically, which is one of the main shortcoming of Legendre polynomials.

One can also derive the action of some linear operation in the coefficient space. Consider a function  $f$  given by its coefficients :  $f = \sum_{n=0}^N a_n P_n(x)$  and  $H$  be a linear operator acting on  $f$  so that  $Hf = \sum_{n=0}^N b_n P_n(x)$ . The relation between the  $a_n$  and  $b_n$  can be explicitly written in some cases. For example :

- If  $H$  is the multiplication by  $x$  then :

$$b_n = \frac{n}{2n-1} a_{n-1} + \frac{n+1}{2n+3} a_{n+1} \quad (n \geq 1). \quad (11)$$

- If  $H$  is the derivation :

$$b_n = (2n+1) \sum_{p=n+1, p+n \text{ odd}}^N a_p. \quad (12)$$

- If  $H$  is the second derivation :

$$b_n = (n+1/2) \sum_{p=n+2, p+n \text{ even}}^N [p(p+1) - n(n+1)] a_p. \quad (13)$$

Those kind of relations enable to represent the action of  $H$  as a matrix acting on the vector of the  $a_n$ , the product being the coefficients of  $Hf$ , i.e. the  $b_n$ .

## 2. Chebyshev polynomials

The Chebyshev polynomials  $T_n$  are an orthogonal set on  $[-1, 1]$  for the measure  $w = \frac{1}{\sqrt{1-x^2}}$ . More precisely one has

$$\int_{-1}^1 \frac{T_n T_m}{\sqrt{1-x^2}} dx = \frac{\pi}{2} (1 + \delta_{0n}) \delta_{mn}. \quad (14)$$

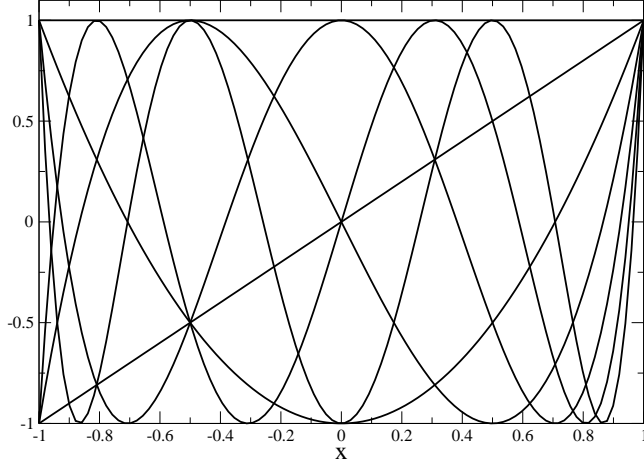


FIG. 7: The first Chebyshev polynomials, from  $T_0$  to  $T_6$ .

Chebyshev polynomials can be computed by knowing that  $T_0 = 1$ ,  $T_1 = x$  and by making use of the recurrence :

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \quad (15)$$

It follows that i)  $T_n$  has the same parity as  $n$ . ii)  $T_n$  is of degree  $n$ . iii)  $T_n(\pm 1) = (\pm 1)^n$ . iv)  $T_n$  has exactly  $n$  zero on  $[-1, 1]$ . The first polynomials are plotted on Fig. 7.

The weights and collocation points associated with Chebyshev polynomials can be computed :

- Chebyshev-Gauss :  $x_i = \cos \frac{(2i+1)\pi}{2N+2}$  and  $w_i = \frac{\pi}{N+1}$ .
- Chebyshev-Gauss-Radau :  $x_i = \cos \frac{2\pi i}{2N+1}$ . The weights are  $w_0 = \frac{\pi}{2N+1}$  and  $w_i = \frac{2\pi}{2N+1}$
- Chebyshev-Gauss-Lobatto :  $x_i = \cos \frac{\pi i}{N}$ . The weights are  $w_0 = w_N = \frac{\pi}{2N}$  and  $w_i = \frac{\pi}{N}$ .

Contrary to Legendre polynomials, the position of the collocation points are completely analytical which can somewhat simplify the computational task. In all the following, Chebyshev polynomials are used, except when otherwise stated.

Once again, one can find the relation between the coefficients  $a_n$  of a function  $f$  and the coefficients  $b_n$  of  $Hf$ , where  $H$  is a linear operator. For example :

- If  $H$  is the multiplication by  $x$  then :

$$b_n = \frac{1}{2} [(1 + \delta_{0n-1}) a_{n-1} + a_{n+1}] \quad (n \geq 1). \quad (16)$$

- If  $H$  is the derivation :

$$b_n = \frac{2}{(1 + \delta_{0n})} \sum_{p=n+1, p+n \text{ odd}}^N p a_p. \quad (17)$$

- If  $H$  is the second derivation :

$$b_n = \frac{1}{(1 + \delta_{0n})} \sum_{p=n+2, p+n \text{ even}}^N p(p^2 - n^2) a_p. \quad (18)$$



### C. Convergence properties

As already seen, for sufficiently regular functions, the difference between  $u$  and its interpolant  $I_N u$  goes to zero exponentially. This statement can be made more precise. Let us consider a  $C^m$  function  $u$ . Upper bounds on the difference between  $u$  and  $I_N u$  can be found for various norms and choice of polynomials.

- For Legendre :

$$\|I_N u - u\|_{L^2} \leq \frac{C}{N^{m-1/2}} \sum_{k=0}^m \|u^{(k)}\|_{L^2}. \quad (19)$$

- For Chebyshev :

$$\|I_N u - u\|_{L_w^2} \leq \frac{C}{N^m} \sum_{k=0}^m \|u^{(k)}\|_{L_w^2}. \quad (20)$$

$$\|I_N u - u\|_{\infty} \leq \frac{C}{N^{m-1/2}} \sum_{k=0}^m \|u^{(k)}\|_{L_w^2}. \quad (21)$$

Without going into too much details, one can note that the errors decay faster than some power of  $N$ . The upper bounds decay faster for more regular function (i.e. for higher  $m$ ). An interesting case is the one of a  $C^\infty$  function. In such a case, the error decays faster than any power of  $N$  and thus like an exponential. One talks of an *evanescent error* and this is the case previously observed.

A limit case can be provided by a discontinuous function. Consider a step function  $u$  such that  $u(x < 0) = 1$  and  $u(x > 0) = 0$ . If one tries to interpolate this function the convergence theorems previously exposed can not ensure convergence. This reflects on Fig. 8 where the step function is presented with interpolants for various number of points. As the number of collocation points increases, the maximum difference between  $u$  and  $I_N u$  stays constant (the amplitude of the oscillations can be seen to be constant). However, more and more oscillations are present, as the Chebyshev polynomials do their best to approximate the discontinuous step function. In a sense, the approximation is still better and better, the support of the difference being smaller and smaller. In other words, the integrated error :  $\int_{-1}^1 |I_N u - u| dx$  still goes to zero when  $N$  increases, as can be seen on Fig. 9. However, this convergence follows only a very slow-decaying power-law.

The simple example of the step function is a very general feature of the so-called *Gibbs phenomenon*, which occurs anytime one tries to interpolate a function that is not  $C^\infty$ . When this happens, the error is no longer evanescent and only converges as a power-law. In order to maintain evanescent errors, one should try to avoid any Gibbs phenomenon. In some cases, as will be seen later, this can be achieved by using a multi-domain decomposition of space.

## III. DIFFERENTIAL EQUATION SOLVERS

### A. The weighted residual method

The type of problems with are concerned with in the section are ordinary differential equations, in one dimension only, on a bounded domain at the boundaries of which some conditions are enforced on the solution. Mathematically, one considers the following system :

$$Lu(x) = S(x) \quad \text{for } x \in U \quad (22)$$

$$Bu(y) = 0 \quad \text{for } y \in \partial U \quad (23)$$

where  $L$  and  $B$  are linear differential operators. A function  $u$  is then an admissible solution of this system, if and only if i) it satisfies Eq. (23) “exactly” (i.e. to machine accuracy) ii) it makes the residual  $R \equiv Lu - S$  small. In order to quantify what this “small” means, the weighted residual method relies on  $N + 1$  tests functions  $\xi_n$  and one asks that the scalar product of  $R$  with those functions is exactly zero :

$$(\xi_k, R) = 0, \quad \forall k \leq N. \quad (24)$$

Of course as  $N$  increases the obtained solution is closer and closer to the real one. Depending on the choice of spectral basis and of test functions, one can generate various different types of spectral solvers. In the following, three of the most used ones are presented and applied to a simple case.

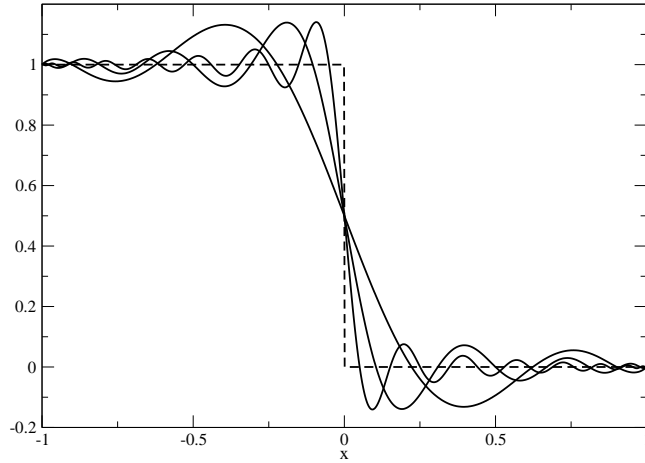


FIG. 8: Step function and some interpolants for  $N = 7$ ,  $N = 15$  and  $N = 31$ . As  $N$  increases the maximum difference stays constant and the number of oscillations increases.

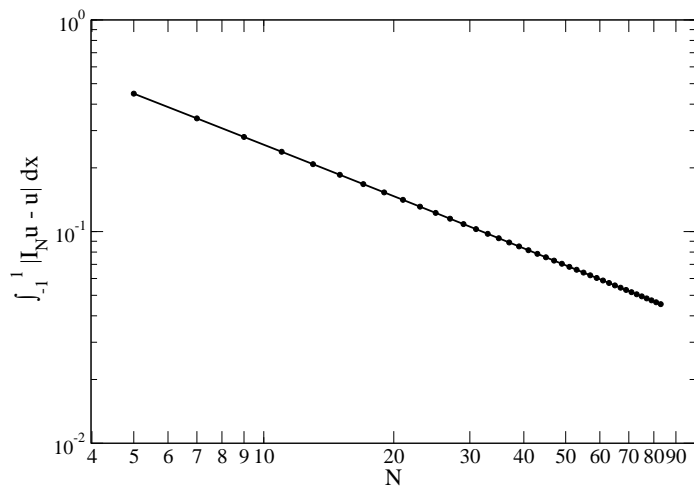


FIG. 9:  $\int_{-1}^1 |I_N u - u| dx$  as a function of  $N$ ,  $u$  being a step function. The convergence obeys a power-law which decays slower than  $N^{-1}$ .

### B. A test problem

We propose to solve the equation :

$$\frac{d^2 u}{dx^2} - 4 \frac{du}{dx} + 4u = \exp(x) + C, \quad (25)$$

with  $x \in [-1, 1]$  and  $C = -4e/(1 + e^2)$ . As boundary conditions, we simply ask that the solution is zero at the boundaries :

$$u(-1) = 0 \quad \text{and} \quad u(1) = 0. \quad (26)$$

Under those conditions, the solution is unique and analytical :

$$u_{\text{sol}}(x) = \exp(x) - \frac{\sinh(1)}{\sinh(2)} \exp(2x) + \frac{C}{4}. \quad (27)$$

Let us note that this solution is not a polynomial.

With our notations the linear operator  $L$  is  $\frac{d^2}{dx^2} - 4\frac{d}{dx} + 4\text{Id}$ . Using the elementary linear operations seen in Sec.II B, one can construct the matrix representation of  $L$ , which will come handy in the implementation of the various solvers.

Let us recall that if  $u = \sum_{i=0}^N \tilde{u}_i T_i(x)$  then  $Lu = \sum_{i=0}^N \sum_{j=0}^N L_{ij} \tilde{u}_j T_i(x)$ .

In this particular case, and for  $N = 4$ , one finds :

$$L_{ij} = \begin{pmatrix} 4 & -4 & 4 & -12 & 32 \\ 0 & 4 & -16 & 24 & -32 \\ 0 & 0 & 4 & -24 & 48 \\ 0 & 0 & 0 & 4 & -32 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix}. \quad (28)$$

### C. The Tau-method

In this method, the test functions  $\xi_n$  are chosen to be the same as the spectral functions of decomposition. Let us assume that one uses Chebyshev polynomials  $T_i$ . The residual equations (24) are then :

$$(T_n, Lu - S) = 0 \quad \forall n \leq N. \quad (29)$$

If one uses the definition of the matrix  $L_{ij}$ , this set of equations can be written as

$$\sum_{j=0}^N L_{nj} \tilde{u}_j = \tilde{s}_n \quad \forall n \leq N \quad (30)$$

where the  $\tilde{s}_n$  are the spectral coefficients of the source  $S$ .

However, due to the presence of homogeneous solutions of  $L$ , this set of  $N + 1$  equations is degenerate and one must impose the boundary conditions before solving it. In the Tau-method, the boundary conditions are enforced as additional equations. In our particular case they can be written in a very straightforward manner :

$$\begin{aligned} u(x = -1) = 0 &\implies \sum_{j=0}^N (-1)^j \tilde{u}_j = 0 \\ u(x = +1) = 0 &\implies \sum_{j=0}^N \tilde{u}_j = 0. \end{aligned} \quad (31)$$

One has then to relax the two *last* residual equations and to replace them by the two boundary conditions in order to get an invertible system for which the unknowns are the  $\tilde{u}_n$ . Relaxing the last two equations is not an issue. Indeed, if the functions are regular enough, the coefficients are rapidly decreasing and thus the obtained solution should converge nicely to the exact solution.

In our example, and for  $N = 4$  the system reads as :

$$\begin{pmatrix} 4 & -4 & 4 & -12 & 32 \\ 0 & 4 & -16 & 24 & -32 \\ 0 & 0 & 4 & -24 & 48 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \tilde{u}_0 \\ \tilde{u}_1 \\ \tilde{u}_2 \\ \tilde{u}_3 \\ \tilde{u}_4 \end{pmatrix} = \begin{pmatrix} -0.03 \\ 1.13 \\ 0.27 \\ 0 \\ 0 \end{pmatrix}. \quad (32)$$

Once inverted, the coefficients of the solution can be found  $\tilde{u}_0 \simeq 0.146$ ;  $\tilde{u}_1 \simeq 0.079$ ;  $\tilde{u}_2 \simeq -0.122$ ;  $\tilde{u}_3 \simeq -0.079$ ;  $\tilde{u}_4 \simeq -0.024$  (for  $N = 4$ ).

The numerical solution obtained is compared to the exact one on Fig. 10. If the difference is easily seen for  $N = 4$ , even with as few points as  $N = 8$ , there is not difference noticeable by eye between the two curves. One can also note, as expected when using the weighted residual method, that, for all  $N$ , the numerical solution "exactly" fulfills the boundary conditions.

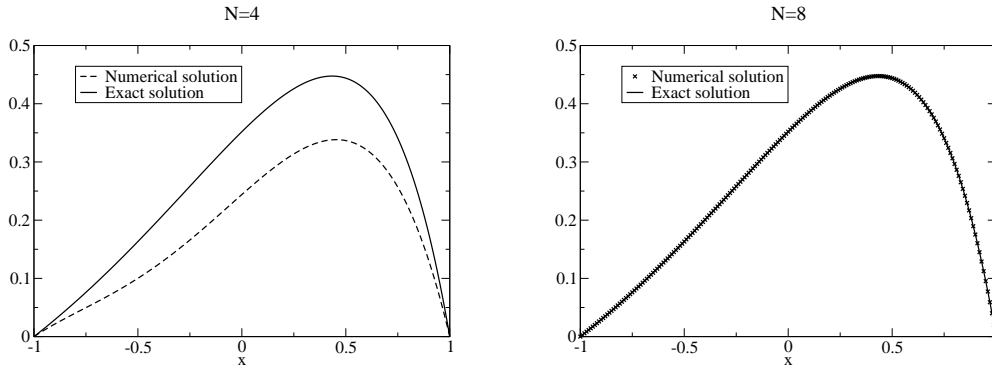


FIG. 10: Exact solution (solid line) and numerical one (dashed line and cross symbols), for the Tau-method, for  $N = 4$  and  $N = 8$ .

#### D. The collocation method

In this method, the test functions are no longer the functions of the spectral expansion, but are chosen to be zero but at one collocation point  $\xi_n = \delta(x - x_n)$ . The residual equations (24) are then equivalent to :

$$Lu(x_n) = S(x_n) \quad \forall n \leq N. \quad (33)$$

Given the expression of the coefficients of  $Lu$  in terms of the matrix  $L_{ij}$ , one can write those equations as a system for the unknowns  $\tilde{u}_n$  :

$$\sum_{i=0}^N \sum_{j=0}^N L_{ij} T_i(x_n) \tilde{u}_j = S(x_n) \quad \forall n \leq N. \quad (34)$$

As for the Tau-method, this system is degenerate and must be supplied with the boundary conditions. One again, they are enforced by two additional equations, which are the same than for the Tau-method (i.e. Eqs (31)). However, instead of relaxing the last two equations of (34), one has to relax the first and the last ones. Indeed, those are the equations that involved the extremal points, at which the boundary conditions have to be fulfilled.

For  $N = 4$ , the system for the collocation method is :

$$\begin{pmatrix} 1 & -1 & 1 & -1 & 1 \\ 4 & -6.83 & 15.3 & -26.1 & 28 \\ 4 & -4 & 0 & 12 & -12 \\ 4 & -1.17 & -7.31 & 2.14 & 28 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \tilde{u}_0 \\ \tilde{u}_1 \\ \tilde{u}_2 \\ \tilde{u}_3 \\ \tilde{u}_4 \end{pmatrix} = \begin{pmatrix} 0 \\ -0.80 \\ -0.30 \\ 0.73 \\ 0 \end{pmatrix} \quad (35)$$

and the coefficients of the solution are :  $\tilde{u}_0 \simeq 0.188$ ;  $\tilde{u}_1 \simeq 0.089$ ;  $\tilde{u}_2 \simeq -0.157$ ;  $\tilde{u}_3 \simeq -0.089$ ;  $\tilde{u}_4 \simeq -0.031$ .

Once again, the numerical solution and the exact one are shown on Fig. 11 and the same properties than for the Tau-method can be observed.

#### E. The Galerkin method

The basic idea of the Galerkin method is to expand the solution, not in terms of usual orthogonal polynomials, but on some linear combinations of polynomials that fulfill the boundary conditions. One then talks of *Galerkin basis*. The particular choice of basis is of course important and it is rather hard to give a general recipe. However, it is usually better if the Galerkin basis can be easily written in terms of the original basis.

For the boundary conditions of our example (26), it is easy to see that the following choice of Galerkin basis  $G_i$  is a valid one :

- $G_{2k}(x) = T_{2k+2}(x) - T_0(x)$

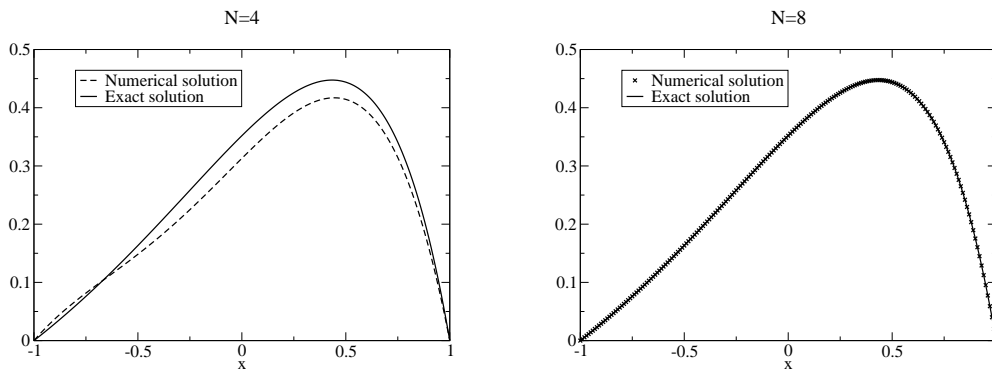


FIG. 11: Exact solution (solid line) and numerical one (dashed line and cross symbols), for the collocation method, for  $N = 4$  and  $N = 8$ .

- $G_{2k+1}(x) = T_{2k+3}(x) - T_1(x)$ .

Let us note that, in order to maintain a consistent degree of approximation, one has to consider only  $N - 1$  Galerkin polynomials (so that the Chebyshev polynomial of higher degree remains  $T_N$ ). This is evident on our example but is a general feature. The reader should also note that the  $G_i$ , in general, are not orthogonal polynomials.

It is convenient to define the *transformation matrix* as the matrix that relates the  $G_i$  to the standard polynomials, let us say the  $T_i$ . This transformation matrix  $M$  is not square but a  $(N + 1) \times (N - 1)$  matrix and is defined by

$$G_i = \sum_{j=0}^N M_{ji} T_j \quad \forall i \leq N - 2. \quad (36)$$

In our example  $M$  is easily constructed and for  $N = 4$  one gets

$$M_{ij} = \begin{pmatrix} -1 & 0 & -1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (37)$$

As already mentioned, the solution  $u$  is seek in terms of the Galerkin basis :  $u = \sum_{k=0}^{N-2} \tilde{u}_k^G G_k(x)$ . The transformation matrix enables us to get  $u$  in terms of  $T_i$ , and then to use the matrix  $L_{ij}$  to get the expression of

$$Lu(x) = \sum_{k=0}^{N-2} \tilde{u}_k^G \sum_{i=0}^N \sum_{j=0}^N M_{jk} L_{ij} T_i(x). \quad (38)$$

In order to use the weighted residual method, one has to precise our choice of test functions. In a standard Galerkin method, the  $\xi_n$  are chosen to be the Galerkin coefficients so that the residual equation reads  $(G_n, R) \quad \forall n \leq N - 2$ .

$(Lu, G_n)$  can be computed by using Eq. (38) and by expressing  $G_n$  in terms of Chebyshev polynomials, using, once again, the transformation matrix.

The situation for the source term is a bit different. Indeed, the source  $S$  has no reason at all to verify the same boundary conditions as the solution  $u$ . So, one has to expand  $S$  on the standard basis :  $S = \sum_{i=0}^N \tilde{s}_i T_i$ . The scalar product  $(S, G_n)$  can then be obtained by using  $M$  to go from the  $G_i$  basis to the  $T_i$  one.

Putting all the pieces together, the Galerkin system then reads :

$$\sum_{k=0}^{N-2} \tilde{u}_k^G \sum_{i=0}^N \sum_{j=0}^N M_{in} M_{jk} L_{ij} (T_i | T_i) = \sum_{i=0}^N M_{in} \tilde{s}_i (T_i | T_i), \quad \forall n \leq N - 2. \quad (39)$$

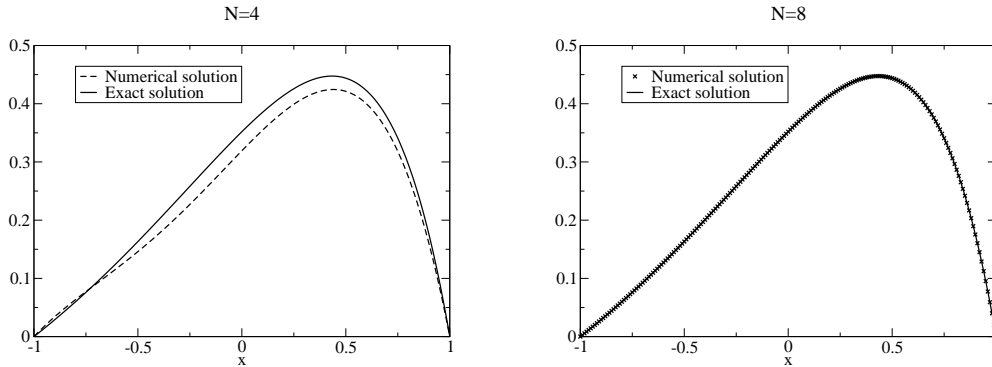


FIG. 12: Exact solution (solid line) and numerical one (dashed line and cross symbols), for the Galerkin method, for  $N = 4$  and  $N = 8$ .

This system is well-posed and can be inverted to get the unknowns, which are the coefficient of  $u$  on the Galerkin basis, i.e. the  $\tilde{u}_i^G$ . Once those are known, the transformation matrix can be used, one last time, to get the solution in terms of the standard spectral basis :

$$u(x) = \sum_{i=0}^N \left( \sum_{n=0}^{N-2} M_{in} \tilde{u}_n^G \right) T_i. \quad (40)$$

In the case of our test problem, for  $N = 4$  the system is :

$$\begin{pmatrix} 2\pi & -4\pi & -4\pi \\ 8\pi & -8\pi & 0 \\ 0 & 8\pi & -26\pi \end{pmatrix} \begin{pmatrix} \tilde{u}_0^G \\ \tilde{u}_1^G \\ \tilde{u}_2^G \end{pmatrix} = \begin{pmatrix} 0.521 \\ -1.705 \\ 0.103 \end{pmatrix}. \quad (41)$$

The Galerkin coefficients are  $\tilde{u}_0^G \simeq -0.160$ ;  $\tilde{u}_1^G \simeq -0.092$ ;  $\tilde{u}_2^G \simeq -0.029$  and the standard ones  $\tilde{u}_0 \simeq 0.189$ ;  $\tilde{u}_1 \simeq 0.092$ ;  $\tilde{u}_2 \simeq -0.160$ ;  $\tilde{u}_3 \simeq -0.092$ ;  $\tilde{u}_4 \simeq -0.029$ . The numerical solution is compared to the exact one on Fig. 12.

#### F. The methods are optimal

On Fig. 13, the maximum error between the numerical solution and the analytical one is shown, as a function of the number of coefficients  $N$ . The three methods exhibit similar features : an exponential decay before reaching the machine accuracy and being limited by round-off errors. All the functions involved in our test being  $\mathcal{C}^\infty$ , it is not surprising that the error is evanescent. Also plotted on Fig. 13 is the maximum difference between the exact solution and its interpolant. The behavior is very similar to the three other curves. Thus, it seems that all the three resolution methods do not introduce more error than the one that would be already done when interpolating the exact solution.

This notion can be made more precise and is called *optimality*. Let us assume that the exact solution of an equation,  $u_{\text{exact}}$ , is known. One can compare it to its interpolant  $I_N u_{\text{exact}}$  and to the obtained numerical solution  $u_{\text{num}}$ . The resolution method is then said to be optimal if and only if  $\max(|u_{\text{exact}} - I_N u_{\text{exact}}|)$  and  $\max(|u_{\text{exact}} - u_{\text{num}}|)$  have the same behavior when  $N \rightarrow \infty$ . Of course, this is usually a difficult quality to assess, given that the exact solution is rarely known. However, on our particular simple example, all the three methods presented are optimal.

### IV. MULTI-DOMAIN SOLVERS

#### A. Motivation

As previously seen, spectral methods lose some of their appeal when they have to deal with non  $\mathcal{C}^\infty$  functions. Indeed, due to the Gibbs phenomenon, the convergence of the numerical solution to the real one is drastically slowed down.

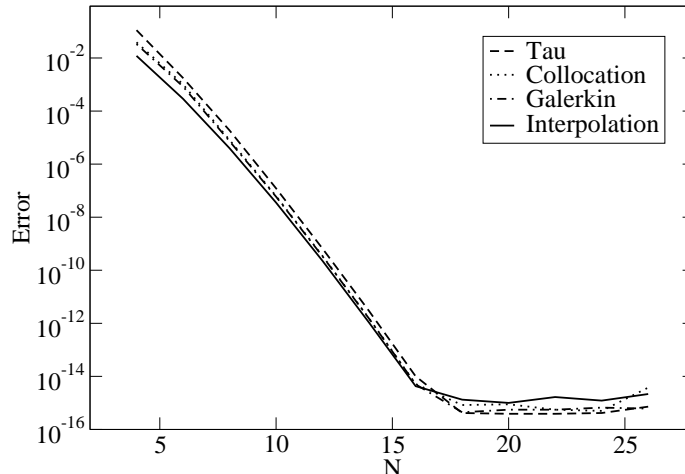


FIG. 13: Maximum relative difference between the exact solution and the numerical one as a function of the number of coefficients, for the Tau-method (dashed line), the collocation one (dotted one) and the Galerkin method (dashed-dotted line). Is also shown, the relative difference between the exact solution and its interpolant.

Nevertheless, they are a great deal of physical problems where such discontinuities occur. We can for example mention the shock front in fluid dynamics or the surface of a strange star where a jump of matter density occurs. From a mathematical point of view, it is also sometime useful to use different variables and functions in various regions of space, thus creating numerical discontinuities.

For all those reason, one has to introduce several numerical domains to cover the physical space of interest. If one can ensure that the various discontinuities lie at the boundaries, then, in each domain, one will deal only with  $C^\infty$  functions, thus recovering exponential convergence.

### B. Test problem

As in Sec. III, the various multi-domain solvers are presented on a simple test problem. The physical space is the interval  $-1 \leq x \leq 1$ . On this interval, one considers the following equation :

$$-\frac{d^2u}{dx^2} + 4u = S \quad (42)$$

with the boundary conditions  $u(-1) = 0$  and  $u(1) = 0$ . The source  $S$  is a step function :  $S(x < 0) = 1$  and  $S(x > 0) = 0$ . Under those conditions, the analytical solution is given by :

$$u(x < 0) = \frac{1}{4} - \left( \frac{e^2}{4} + B^- e^4 \right) \exp(2x) + B^- \exp(-2x) \quad (43)$$

$$u(x > 0) = B^+ \left( \exp(-2x) - \frac{1}{e^4} \exp(2x) \right), \quad (44)$$

where the constants are

- $B^- = -\frac{1}{8(1+e^2)} - \frac{e^2}{8(1+e^4)}$
- $B^+ = \frac{e^4}{8} \left( \frac{e^2}{(1+e^4)} - \frac{1}{(1+e^2)} \right)$ .

As for the one-domain case, this solution is not polynomial.

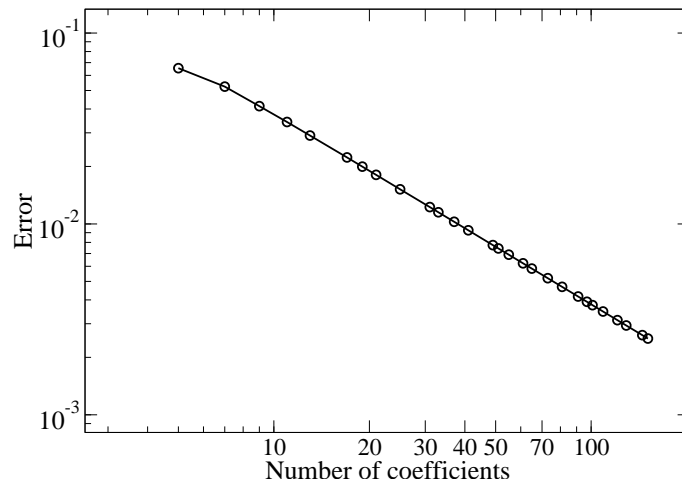


FIG. 14: Error done when using a Tau-method with a discontinuous source, as a function of the number of coefficients. The convergence is no longer exponential but follows a power-law in  $N^{-1}$ .

Of course, due to the discontinuity of the source, the solution is only  $\mathcal{C}^1$  at  $x = 0$ . This causes a great loss of accuracy for the one-domain solvers presented in Sec. III. For example, the error made when solving Eq. (42) with the standard Tau-method is presented on Fig. 14. As expected, even if one converges to the right solution, the decay of the error is no longer exponential but only follows a power-law (like  $N^{-1}$  in this particular case). For a practical point of view, the loss of accuracy is spectacular : with as many coefficients as  $N = 100$  one does not even reach a precision of  $10^{-3}$ . This is to be compared with the results of Sec.III, where machine precision could be reached with only 25 coefficients (see Fig. 13).

In order to cope with this problem, one can introduce two subdomains of the interval :

- Domain 1 :  $x \leq 0$  described by the variable  $x_1 = 2x + 1$  ( $-1 \leq x_1 \leq 1$ ).
- Domain 2 :  $x \geq 0$  described by  $x_2 = 2x - 1$  ( $-1 \leq x_2 \leq 1$ ).

In each subdomain, the functions are expanded on the basis polynomials, with respect to the auxiliary variable. For instance, in domain 1, the solution is given by :

$$u(x < 0) = \sum_{i=0}^N \tilde{u}_i^1 T_i(x_1(x)). \quad (45)$$

When implementing various operators, one has to remember that one is working with the auxiliary variables. For example  $\frac{d}{dx} = 2 \frac{d}{dx_1}$ .

### C. A multi-domain Tau-method

In each domain, one writes the usual residual equations for the Tau-method (see III C). For instance, in the domain 1, this gives rise to

$$(T_n, R) = 0 \implies \sum_{i=0}^N L_{ni} \tilde{u}_i^1 = \tilde{s}_n^1. \quad (46)$$

This is a set of  $N + 1$  equations among which one has to relax the last two, in order to impose various boundary and continuity conditions.  $L$  is the differential operator expressed with respect to  $x_1$ . The same is done in the second domain.



As for the usual Tau method, the boundary conditions are enforced by two additional equations. In order to close the problem, one has to impose that the solution and its first derivative with respect to  $x$  are continuous at the boundary between the two domains :

- $u^1(x_1 = 1, x = 0) = u^2(x_2 = -1, x = 0)$
- $\frac{du^1}{dx}(x_1 = 1, x = 0) = \frac{du^2}{dx}(x_2 = -1, x = 0)$

We are left with a system of :  $2N - 2$  residual equations, 2 boundary conditions and 2 matching conditions, for a total of  $2N + 2$  equations. The unknowns are the coefficients of the solution in both domains which are in the right number  $2N + 2$ . For our test problem, when using  $N = 4$  in both domains, the system to be inverted is :

$$\begin{pmatrix} 4 & 0 & -16 & 0 & -128 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & -96 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & -192 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 0 & 1 & 4 & 9 & 16 & 0 & -1 & 4 & -9 & 16 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 4 & 0 & -16 & 0 & -128 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & -96 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & -192 \end{pmatrix} \begin{pmatrix} \tilde{u}_0^1 \\ \tilde{u}_1^1 \\ \tilde{u}_2^1 \\ \tilde{u}_3^1 \\ \tilde{u}_4^1 \\ \tilde{u}_0^2 \\ \tilde{u}_1^2 \\ \tilde{u}_2^2 \\ \tilde{u}_3^2 \\ \tilde{u}_4^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (47)$$

The solution is then given in both domains by

- $\tilde{u}_0^1 \simeq 0.082$ ;  $\tilde{u}_1^1 \simeq 0.044$ ;  $\tilde{u}_2^1 \simeq -0.036$ ;  $\tilde{u}_3^1 \simeq 0.0018$ ;  $\tilde{u}_4^1 \simeq -0.0007$
- $\tilde{u}_0^2 \simeq 0.038$ ;  $\tilde{u}_1^2 \simeq -0.044$ ;  $\tilde{u}_2^2 \simeq 0.008$ ;  $\tilde{u}_3^2 \simeq -0.0018$ ;  $\tilde{u}_4^2 \simeq 0.00017$

The convergence properties of this method will be presented in Sec. IV F.

#### D. Homogeneous solution method

This method is the closest to the analytical way of solving ordinary differential equations in various domains. The idea is just to compute some particular solution in both domains, to get the homogeneous solutions and to do the appropriate linear combination to fulfill the boundary conditions and continuity requirements.

Most of the time, the homogeneous solution are known analytically. Generically, their number is given by the order of the operator  $L$  of the equation. In our particular example,  $L$  is of second order and one gets 2 homogeneous solutions in each domain. One has to compute the coefficients of those solutions with respect to the auxiliary variables  $x_1$  and  $x_2$ . In our case, one finds :

- $h_1(x) = \exp(2x) = \exp(x_1 - 1) = \exp(x_2 + 1)$
- $h_2(x) = \exp(-2x) = \exp(-x_1 + 1) = \exp(-x_2 - 1)$ .

For instance, for  $h_1$  this gives raise to the following coefficients, in both domains :

- in the domain 1 : (0.47; 0.42; 0.1; 0.017; 0.0020)
- in the domain 2 : (3.4; 3.1; 0.74; 0.12; 0.015)

In order to get one particular solution in each domain, one can use whatever method already presented in the one-domain case (see Sec. III). The conditions enforced at the boundaries of each domain have no importance because they will be modified by linear combination with the homogeneous solutions. For instance, using a Tau-method in each domain, and demanding that the first two coefficients of the particular solutions are zero (equivalent to imposing two boundary conditions) gives the following system in the first domain (which reduces to a  $(N - 1) \times (N - 1)$  system) :

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 4 & 0 & -16 & 0 & -128 \\ 0 & 4 & 0 & -96 & 0 \\ 0 & 0 & 4 & 0 & -192 \end{pmatrix} \begin{pmatrix} \tilde{g}_0 \\ \tilde{g}_1 \\ \tilde{g}_2 \\ \tilde{g}_3 \\ \tilde{g}_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad (48)$$

which enables us to obtain the following particular solution, in domain 1  $\tilde{g}_0 = 0; \tilde{g}_1 = 0; \tilde{g}_2 \simeq -0.53; \tilde{g}_3 = 0; \tilde{g}_4 = -0.001$ .

So, in each domain, the solution is given by the sum of the particular solution  $g$  and the two homogeneous ones  $h_1$  and  $h_2$ , with unknowns coefficients  $\alpha^i$  and  $\beta^i$ ,

$$u(x) = g^i(x_i) + \alpha^i h_1^i(x_i) + \beta^i h_2^i(x_i) \quad (49)$$

where  $i$  denotes the domain.

In order to determine the appropriate values of the 4 coefficients  $\alpha^i$  and  $\beta^i$ , one has to impose :

- Boundary condition at  $x = -1$  i.e. at  $x_1 = -1$

$$g^1(-1) + \alpha^1 h_1^1(-1) + \beta^1 h_2^1(-1) = 0$$

- Boundary condition at  $x = 1$  i.e. at  $x_2 = 1$

$$g^2(1) + \alpha^2 h_1^2(1) + \beta^2 h_2^2(1) = 0$$

- Matching of the solution at  $x = 0$ , i.e. at  $x_1 = 1$  and  $x_2 = -1$

$$g^1(1) + \alpha^1 h_1^1(1) + \beta^1 h_2^1(1) = g^2(-1) + \alpha^2 h_1^2(-1) + \beta^2 h_2^2(-1)$$

- Matching of the first derivative at  $x = 0$ , i.e. at  $x_1 = 1$  and  $x_2 = -1$

$$g'^1(1) + \alpha^1 h_1'^1(1) + \beta^1 h_2'^1(1) = g'^2(-1) + \alpha^2 h_1'^2(-1) + \beta^2 h_2'^2(-1)$$

If the problem is well-posed, this system of 4 equations admits a unique solution for the 4 coefficients of the homogeneous solutions, thus allowing us to construct the solution of the equation in all space. For instance, in our particular case, for  $N = 4$ , the matching system is

$$\begin{pmatrix} 7.39 & 0.135 & 0 & 0 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 0 & 0 & 0.135 & 7.39 \end{pmatrix} \begin{pmatrix} \beta^1 \\ \alpha^1 \\ \beta^2 \\ \alpha^2 \end{pmatrix} = \begin{pmatrix} 0.055 \\ 0.055 \\ -0.23 \\ 0 \end{pmatrix} \quad (50)$$

which admits the solution :  $\beta^1 = 0.0048; \alpha^1 = 0.14; \beta^2 = 0.094; \alpha^2 = -0.0017$ . The coefficients of the numerical solution, in each domains are then :

- $\tilde{u}_0^1 \simeq 0.083; \tilde{u}_1^1 \simeq 0.044; \tilde{u}_2^1 \simeq -0.036; \tilde{u}_3^1 \simeq 0.0018; \tilde{u}_4^1 \simeq -0.0008$
- $\tilde{u}_0^2 \simeq 0.038; \tilde{u}_1^2 \simeq -0.044; \tilde{u}_2^2 \simeq 0.008; \tilde{u}_3^2 \simeq -0.0018; \tilde{u}_4^2 \simeq 0.00016$ .

Numerical accuracy of this method will be shown in Sec. IV F.

### E. Variational formulation

Contrary to all the methods previously presented, the variational one is only easily applicable when one is using Legendre polynomials. Indeed, as will be seen later, it requires that the measure is  $w(x) = 1$ .

The starting point is the usual residual equation, written here explicitly in its integral form

$$(\xi, R) = 0 \implies \int \xi (-u'' + 4u) dx = \int \xi S dx. \quad (51)$$

The term involving  $u''$  is then integrated by parts, which gives :

$$[-\xi u'] + \int \xi' u' dx + 4 \int \xi u dx = \int \xi S dx. \quad (52)$$

The method is easier to implement when, as for the collocation method, the test functions are :  $\xi_n = \delta(x - x_n)$ , being zero but at one collocation point.

In the following, it will also be convenient to define the derivative operator  $D$  in the configuration space. It relates the values of the function at each collocation point to the ones of its first derivative via :

$$g'(x_k) = \sum_{j=0}^N D_{kj} g(x_j). \quad (53)$$

The matrix associated to  $D$  is easily constructed. For example, for  $N = 4$  and Legendre polynomials, one gets :

$$D_{ij} = \begin{pmatrix} -10 & 13.5 & -5.33 & 2.82 & -1 \\ -2.48 & 0 & 3.49 & -1.53 & 0.52 \\ 0.75 & -2.67 & 0 & 2.67 & -0.75 \\ -0.52 & 1.53 & -3.49 & 0 & 2.48 \\ 1 & -2.82 & 5.33 & -13.5 & 10 \end{pmatrix} \quad (54)$$

Using this operator and the Gauss quadrature rule (4), one can compute the integrals involved in Eq. (52) :

$$\int \xi'_n u' dx = \sum_{i=0}^N \xi'_n(x_i) u'(x_i) w_i = \sum_{i=0}^N \sum_{j=0}^N D_{ij} D_{in} w_i u(x_j) \quad (55)$$

$$\int \xi_n u dx = \sum_{i=0}^N \xi_n(x_i) u(x_i) w_i = u(x_n) w_n \quad (56)$$

$$\int \xi_n s dx = \sum_{i=0}^N \xi_n(x_i) S(x_i) w_i = S(x_n) w_n. \quad (57)$$

For the points *strictly inside* each domain, the integrated term  $[-\xi u'] = 0$  so that the residual equations take the following form :

$$\sum_{i=0}^N \sum_{j=0}^N D_{ij} D_{in} w_i u(x_j) + 4u(x_n) w_n = S(x_n) w_n, \quad 0 < n < N \quad (58)$$

which is a set of  $2N - 2$  equations.

The situation of the point lying at the boundary between the two domains is a bit tricky. Indeed the same physical point can be described either :

- as the last point of domain 1 :  $n = N$  and  $[-\xi u'] = -u'^1(x_1 = 1; x = 0)$  so that the residual equation is :

$$u'^1(x_1 = 1) = \sum_{i=0}^N \sum_{j=0}^N D_{ij} D_{iN} w_i u^1(x_j) + 4u^1(x_N) w_N - S^1(x_N) w_N. \quad (59)$$

- as the first point of domain 2 :  $n = 0$  and  $[-\xi u'] = u'^2(x_2 = -1; x = 0)$  and the residual equation is then

$$u'^2(x_2 = -1) = - \sum_{i=0}^N \sum_{j=0}^N D_{ij} D_{i0} w_i u^2(x_j) - 4u^2(x_0) w_0 + S^2(x_0) w_0. \quad (60)$$

This duality can be used to impose implicitly the continuity of the first derivative of  $u$  :

$$\begin{aligned} u'^1(x_1 = 1; x = 0) = u'^2(x_2 = -1; x = 0) &\implies \\ \sum_{i=0}^N \sum_{j=0}^N D_{ij} D_{iN} w_i u^1(x_j) + 4u^1(x_N) w_N + \sum_{i=0}^N \sum_{j=0}^N D_{ij} D_{i0} w_i u^2(x_j) + 4u^2(x_0) w_0 & \\ = S^1(x_N) w_N + S^2(x_0) w_0. & \end{aligned} \quad (61)$$

The other equations are i) the boundary condition on the left  $u^1(x_0) = 0$  ii) the boundary condition on the right  $u^2(x_N) = 0$  and iii) the matching of the solution at the boundary between the two domains  $u^1(x_N) = u^2(x_0)$ . This

is a total number of  $2N + 2$  equations for the  $2N + 2$  unknowns, which in this case, are the value of the solution at each collocation points, the  $u^i(x_n)$ .

For our example and  $N = 4$  in each domain, the system is :

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -16 & 28 & -12 & 2.5 & -0.7 & 0 & 0 & 0 & 0 \\ 2.1 & -12 & 22 & -12 & 2.1 & 0 & 0 & 0 & 0 \\ -0.7 & 2.5 & -12 & 28 & -16 & 0 & 0 & 0 & 0 \\ 0.2 & -0.7 & 2.1 & -16 & 29 & -16 & 2.1 & -0.7 & 0.2 \\ 0 & 0 & 0 & 0 & -16 & 28 & -12 & 2.5 & -0.7 \\ 0 & 0 & 0 & 0 & 2.1 & -12 & 22 & -12 & 2.1 \\ 0 & 0 & 0 & 0 & -0.7 & 2.5 & -12 & 28 & -16 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u^1(x_0) \\ u^1(x_1) \\ u^1(x_2) \\ u^1(x_3) \\ u^1(x_4) = u^2(x_0) \\ u^2(x_1) \\ u^2(x_2) \\ u^2(x_3) \\ u^2(x_4) \end{pmatrix} = \begin{pmatrix} 0 \\ 0.54 \\ 0.71 \\ 0.54 \\ 0.1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (62)$$

Once it is inverted, the solution is found in the configuration space :

- In domain 1 :  $u^1(x_n) = (0; 0.06; 0.12; 0.12; 0.092)$
- In domain 2 :  $u^2(x_n) = (0.092; 0.064; 0.030; 0.0090; 0)$ .

As for the other methods, the convergence properties are shown in the next section.

Before finishing with the variational method, it may be worthwhile to explain why Legendre polynomials were used. Suppose one wants to use Chebyshev polynomials. The measure is then  $w(x) = \frac{1}{\sqrt{1-x^2}}$ . When one integrates the term in  $u''$  by part one then gets

$$\int -u'' f w dx = [-u' f w] + \int u' f' w' dx \quad (63)$$

Because the measure is divergent at the boundaries, it is difficult, if not impossible, to isolate the term in  $u'$ . On the other hand this is precisely the term that is needed to impose the appropriate matching of the solution. There might be ways around this but this explains why the variational method has been presented with Legendre polynomials.

## F. Discussion

On Fig. 15, we show the maximum relative difference between the numerical solution and the analytical one. Contrary to the case of just one domain (see Fig. 14), exponential decay of the error is recovered, as expected, and machine accuracy is very rapidly reached. This illustrates the usefulness of a multi-domain decomposition.

From a numerical point of view, the method based on the homogeneous solutions is different from the two others. Indeed, when using it, the problem is treated domain by domain. The number of system to be solved is equal to the number of domains but each of those systems is roughly of size  $N^2$  only. This is to be compared to the Tau-method and the variational one where only one global system is inverted but which is of size  $(NN_D)^2$ . So, if the number of domain is important, one may consider the homogeneous solution method instead of the two others, which would give rise to huge systems.

On the other hand, the Tau-method and the variational one, do not require the knowledge of the homogeneous solutions which is a good point, especially when the differential operator  $L$  is complicated. The variational method may seem a little more involved, especially as it requires to work with Legendre polynomials but, from a mathematical point of view, it is the only one that has been prove to be optimal.

## V. CONCLUSION

Our introductory tour of spectral methods is now over. After having presented the mathematical foundations of this class of method, along with two families of standard polynomials, Legendre and Chebyshev ones, two simple test problems were treated by means of various methods. Three ways of solving an ordinary differential equation on a single domain were presented and tested : the Tau method, the collocation one and the Galerkin method. In the last part, a discontinuous source was treated by means of a multi-domain decomposition. Again, three methods were implemented and discussed : a multi-domain tau method, one based on the knowledge of homogeneous solutions and

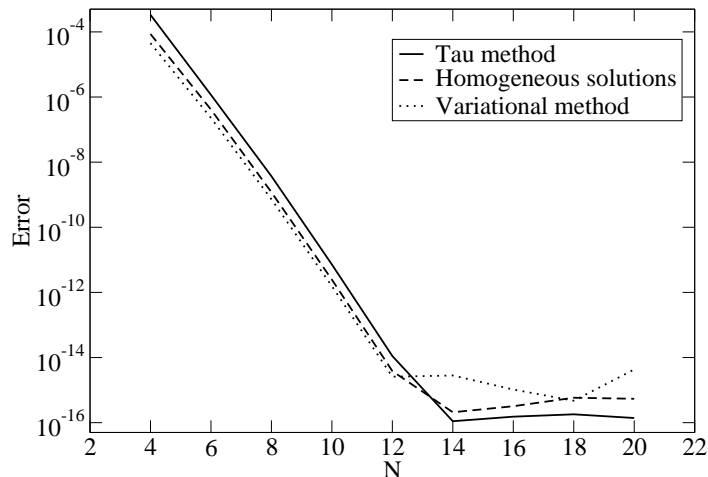


FIG. 15: Relative difference between the analytical solution and the numerical one, for the three multi-domain methods exposed. The error is evanescent.

a variational method using Legendre polynomials. This simple examples were intended to illustrate the usefulness of spectral methods to reach very good accuracy with moderate computational resources.

However, this work is far from covering the huge field of all spectral methods. Among the many aspects that were ignored one can mention the use of Fourier transform for periodic functions. We also have left out all issues concerning compactification of space, by means of auxiliary variables like  $1/r$ . Three dimensional problems were also left out as were non-linear problems. Finally, we did not present any problem involving time evolution. Should the reader feel the urge to learn more about all this, we recommend that he consults some of the books given in the bibliography.

- 
- [1] D. Gottlieb and S.A. Orszag, *Numerical analysis of spectral methods*, Society for Industrial and Applied Mathematics, Philadelphia (1977).
  - [2] C. Canuto, M.Y. Hussaini, A. Quarteroni and T.A. Zang, *Spectral methods in fluid dynamics*, Springer-Verlag, Berlin (1988).
  - [3] B. Mercier, *An introduction to the numerical analysis of spectral methods*, Spinger-Verlag, Berlin (1989).
  - [4] C. Bernardi and Y. Maday, *Approximations spectrales des problèmes aux limites elliptiques*, Springer-Verlag, Paris (1992).
  - [5] B. Fornberg, *A practical guide to pseudospectral methods*, Cambridge University Press, Cambridge (1998).
  - [6] J.P. Boyd, *Chebyshev and Fourier spectral methods - 2nd edition*, Dover, Mineola (2001).
  - [7] <http://www.lorene.obspm.fr>, *Numerical library for spectral methods in general relativity*.