The Spectral Element Method

Alberto Tibaldi

March 18, 2016

1 Introduction

Several physical and engineering models can be described as boundary value problems (BVPs) consisting of one or more differential equations to be solved in a domain defined by the boundary conditions (BCs). The boundary conditions, which should be specified at each extreme of the domain, can be classified into three categories: Dirichlet conditions (value of the solution), Neumann conditions (value of the derivative of the solution normal to the boundary edge, or flux), or Robin conditions (a combination of the solution and of its flux); when the value to be enforced is zero, conditions are called "homogeneous".

Among the techniques that can be used to solve problems based on ordinary or partial differential equations (ODEs or PDEs), spectral methods are very interesting candidates. Although spectral methods have been introduced in the mid-1940s, their first rigorous study was carried out by Gottlieb and Orszag in 1977 [1], who summarized the state of the art in their theory and application. These schemes derive from the method of weighted residuals, where a set of basis functions is used to approximate the solution of the ODE and a weak formulation is used to minimize the error in the expansion. The method accuracy can be improved by refining the approximation: the so-called h-refinement consists of decomposing the domain into several elements, whereas p-refinement techniques rely on the use of high order basis functions. Finite element methods (FEMs) are pure h-refinement methods, since the solution domain is divided into several small elements and low order local functions are used to expand the solution and to test the equations [2] - [4]; this allows to study arbitrarily-varying profiles of the physical parameters of the model (*e.g.,* refractive index in optical wave equations, or thermal conductivity in heat equation). On the other hand, in spectral methods both the expansion and test functions are chosen to be infinitely differentiable entire domain functions, leading to a purely p-refinement method. Even if these methods exhibit exponential convergence[1] when the physical parameters vary smoothly on the domain, if abrupt variations of these parameters occur (such as material discontinuities), the presence of Gibbs phenomena[2] strongly affect the solution accuracy.

[1] *i.e.,* exponential decay of the weighting coefficients of the basis functions by increasing the order of the method, so much faster than standard FEM where the basis functions are linear

[2] Gibbs phenomena consist of oscillations occurring when a discontinuous function or a function with non-continuous derivatives is approximated with smooth, *i.e.,* $C^{(\infty)}$, functions

A major leap towards the "mathematical Nirvana" of a method of arbitrarily high order capable of application to arbitrary physical parameters is represented by spectral element methods (SEMs). Spectral element methods are based on the decomposition of the domain of the BVP in a small number of sub-domains (or "patches"); then, a set of local basis functions is defined for each patch in a reference (or "parent") domain that can be transformed into the j-th patch through a linear mapping [5] - [7]. Then a unique, entire domain set of boundary adapted functions is synthesized starting from the local sets, through basis recombination procedures [8], [9].

The domain decomposition can be applied to each interface where an abrupt variation of the physical properties occurs. The resulting method can be interpreted as an application of the single-domain spectral method in each region where properties are varying smoothly, leading to exponential convergence and maintaining the flexibility of FEM. Additionally, different resolutions (*i.e.,* different orders of the basis functions) can be used in each patch, allowing efficient representations of the solution. Finally, just like in FEMs, this method can be efficiently coupled to other methods or to analytic solutions, leading to hybrid methods.

2 Method of weighted residuals

An ODE is a functional equation or, in other words, a relation between a linear operator $\mathcal{L}$ applied to a function $\psi(x)$, and another function $\varphi(x)$,

$$\mathcal{L}\psi(x) = \varphi(x), \qquad x \in \mathcal{D}. \tag{1}$$

Here $\mathcal{D}$ is the domain of the operator $\mathcal{L}$, defined by the boundary conditions of the problem. This operator can be seen as a mapping from a function space to another; in our case, $\mathcal{L}$ is a differential operator, so it maps a function space to another one, larger. As an example, if $\psi(x)$ is a function with one continuous derivative, $\varphi(x)$ is only continuous; since there are *more continuous functions than differentiable functions*, $\mathcal{L}$ maps the elements of a space to the ones of a larger space. Even if this functional analysis point of view is, at least for me, very fascinating, it is not so useful to solve our real problem. Indeed, in practical situations, $\varphi(x)$ is a known term, which is related to the "excitation" of the method, $\psi(x)$ is unknown, and our objective is its determination.

In most cases, the solution of an ODE cannot be performed by analytical derivations; moreover, even if in few peculiar situations the differential operator lends itself to simplifying manipulations, nothing can be done for arbitrary known terms $\varphi(x)$. People usually say something like "Yeah but we can solve the ODE with the computer!", and they are right, but the point is: a computer cannot solve a functional equation, since it cannot "understand" it; the functional equation should be transformed in something else. Computers are very good computators, rather than mathematicians: they do not understand problems! On the other hand, one of their main skills is the solution of linear systems, so, (1) should be transformed in a linear system, somehow.

This section describes a possible strategy for casting a functional equation into a linear system. A possible procedure is based on the application of the Rayleigh-Ritz theorem, which leads to a variational formulation of the ODE. The very same result can be obtained in a more intuitive way, by the method of weighted residuals [4]; this is based on a geometrical idea: the projections of the functional equation on the elements of a (finite) subspace are coupled linear algebraic equations, leading to a linear system. In this view, the first step is

$$\psi(x) = \sum_{n=1}^{N_{\text{fun}}} c_n f_n(x) + r(x), \tag{2}$$

which reads, *the solution is equal to a linear combination of known functions $f_n(x)$ weighted with unknown coefficients c_n, less than a residue $r(x)$*. So, by substituting in (1), it is obtained

$$\mathcal{L}\left[\sum_{n=1}^{N_{\text{fun}}} c_n f_n(x) + r(x)\right] = \sum_{n=1}^{N_{\text{fun}}} c_n \mathcal{L} f_n(x) + \mathcal{L} r(x) = \varphi(x).$$

Here, linearity is used to switch $\mathcal{L}$ and the sum. Now, the problem has been transformed into the determination of coefficients, so of numbers, instead of functions. However, this is still a functional equation, which cannot be handled by a computer. Therefore, the second step consists in multiplying both sides times functions $g_m(x)$ and integrating on the solution domain $\mathcal{D} = [x_{\text{l}}, x_{\text{r}}]$ defined by the conditions in the left and right boundaries x_{l} and x_{r}; this is done for N_{fun} projections:

$$\sum_{n=1}^{N_{\text{fun}}} c_n \int_{x_{\text{l}}}^{x_{\text{r}}} \mathcal{L} f_n(x) g_m(x)\,\mathrm{d}x + \int_{x_{\text{l}}}^{x_{\text{r}}} \mathcal{L} r(x) g_m(x)\,\mathrm{d}x = \int_{x_{\text{l}}}^{x_{\text{r}}} \varphi(x) g_m(x)\,\mathrm{d}x, \qquad m = 1...N_{\text{fun}}.$$

The projection functions $g_m(x)$ are commonly referred to as "test functions", recalling the theory of distributions, where a distribution is characterized its functionals, defined as projections on smooth functions. These equations can be compactly re-written as

$$\sum_{n=1}^{N_{\text{fun}}} c_n L_{mn} + r_m = \varphi_m, \qquad m = 1...N_{\text{fun}},$$

with the following definitions:

$$\begin{aligned} L_{mn} &= \int_{x_{\text{l}}}^{x_{\text{r}}} \mathcal{L} f_n(x) g_m(x)\,\mathrm{d}x \\ r_m &= \int_{x_{\text{l}}}^{x_{\text{r}}} \mathcal{L} r(x) g_m(x)\,\mathrm{d}x \\ \varphi_m &= \int_{x_{\text{l}}}^{x_{\text{r}}} \varphi(x) g_m(x)\,\mathrm{d}x. \end{aligned}$$

Alternatively, this expression can be written in its vector form

$$\underline{\underline{L}}\,\underline{c} + \underline{r} = \underline{\varphi},$$

where this notation is built by exploiting the matrix row-column product. Now, the coefficients can be found as the solution of the linear system

$$\underline{\underline{L}}\,\underline{c} = \underline{\varphi}$$

obtained by omitting the $\underline{r}$ vector. Due to the projective nature of this simple approach, the solution coefficients $\underline{c}$ minimize the residual in the least squares sense.

2.1 Application example: Galërkin solution of a wave equation eigenproblem

In order to clarify conceptual notation, it is useful to apply it to a practical problem; since no explanation about the synthesis of the basis functions have been provided yet, the full solution of the problem cannot be obtained. The problem to be introduced is the determination of the eigenvalues of the wave equation

$$\frac{\mathrm{d}^2\psi(x)}{\mathrm{d}x^2} = -k^2\psi(x), \qquad x \in [0, \pi],$$

where Neumann boundary conditions are required for $x_\mathrm{l} = 0$, $x_\mathrm{r} = \pi$. Eigenvalue problems are homogeneous, so no forcing term appears: $\varphi(x) = 0$. It can be shown that the solutions of this problem are the integers

$$k = 0, 1, 2, ...$$

Indeed, since this is a problem defined within a bounded domain, the spectrum of this operator consists of a discrete infinity of eigenvalues (no continuum or residual spectra exist). The method of weighted residuals is now applied to this equation, starting from the expansion of the solution:

$$\sum_{n=1}^{N_\mathrm{fun}} c_n \frac{\mathrm{d}^2 f_n(x)}{\mathrm{d}x^2} + \frac{\mathrm{d}^2 r(x)}{\mathrm{d}x^2} = -k^2 \sum_{n=1}^{N_\mathrm{fun}} c_n f_n(x) - k^2 r(x).$$

In this example, as well in most cases, the Galërkin[3] version of the method of weighted residuals is applied; this consists of using test functions equal to the expansion functions. Then, by defining an equivalent term $\mathcal{L}r(x)$ including the residual terms (which will be eventually eliminated from the equation),

$$\sum_{n=1}^{N_\mathrm{fun}} c_n \int_0^\pi \frac{\mathrm{d}^2 f_n(x)}{\mathrm{d}x^2} f_m(x)\,\mathrm{d}x + \int_0^\pi \mathcal{L}r(x) f_m(x)\,\mathrm{d}x = -k^2 \sum_{n=1}^{N_\mathrm{fun}} c_n \int_0^\pi f_n(x) f_m(x)\,\mathrm{d}x.$$

[3]The method is commonly credited to the russian mathematician Boris Grigor'evic Galërkin, whose correct pronounce is "Galyorkin"; however, several contemporary researchers, such ass the swiss mathematician Walther Ritz, were working simultaneously on this subject.

The left-hand side integral can be integrated by parts, leading to

$$\int_0^\pi \frac{\mathrm{d}^2 f_n(x)}{\mathrm{d}x^2} f_m(x)\,\mathrm{d}x = \left.\frac{\mathrm{d}f_n(x)}{\mathrm{d}x} f_m(x)\right|_0^\pi - \int_0^\pi \frac{\mathrm{d}f_n(x)}{\mathrm{d}x}\frac{\mathrm{d}f_m(x)}{\mathrm{d}x}\,\mathrm{d}x.$$

This step is largely used in the solution of second-order equations, for at least two reasons. The superficial one, is that it allows to avoid the synthesis of the second derivatives of the basis functions, which may be an annoying task. The important point in applying integration by parts is the possibility to enforce Neumann conditions as **natural boundary conditions**. Natural boundary conditions are, usually, constraints on flux quantities, so conditions on the derivatives of the solution; such conditions can be enforced in the solution without restricting the set of expansion/test functions[4], but simply by introducing them in the functional to be minimized, *i.e.,* the residual. This can be done by defining an equivalent projected residual vector element as

$$r_m^{\mathrm{eq}} = \sum_{n=1}^{N_{\mathrm{fun}}} c_n \left.\frac{\mathrm{d}f_n(x)}{\mathrm{d}x} f_m(x)\right|_0^\pi + \int_0^\pi \mathcal{L}r(x) f_m(x)\,\mathrm{d}x.$$

Then, it is possible to complete the formulation of the problem by defining the following matrix elements:

$$K_{mn} = \int_0^\pi \frac{\mathrm{d}f_n(x)}{\mathrm{d}x}\frac{\mathrm{d}f_m(x)}{\mathrm{d}x}\,\mathrm{d}x \tag{3}$$

$$M_{mn} = \int_0^\pi f_n(x) f_m(x)\,\mathrm{d}x. \tag{4}$$

The matrices $\underline{\underline{K}}$ and $\underline{\underline{M}}$ are referred to as "stiffness' and "mass" matrices, according to the inherited civil engineering FEM notation; the matrix $\underline{\underline{M}}$ is also called Gram matrix. With these definitions, the linear system can be written as

$$-\underline{\underline{K}}\,\underline{c} = -k^2 \underline{\underline{M}}\,\underline{c}.$$

Here, the residual have not been introduced; by this way, the solution will minimize it in the least squares sense. The Neumann boundary condition, which is part of the functional to be minimized, will be satisfied progressively better by increasing the order of the method. Once that this method is understood, it is directly possible to write this expression, to save time. However, it is remarked that the natural Neumann boundary condition formulation requires the "integration by parts" step, otherwise, boundary conditions on derivatives should be enforced explicitly on the basis functions. If functions are not orthonormal, the mass matrix is not the identity, and this is a generalized linear eigenvalue problem. Instead than solving this, it is possible to write

$$\underline{\underline{M}}^{-1}\underline{\underline{K}}\,\underline{c} = k^2\,\underline{c}.$$

[4]This point will be more clear when discussing the synthesis procedure

The solutions of this problem are the eigenvalues k^2 and the eigenvectors $\underline{c}$, which are also the coefficients for rebuilding the solution. This problem will be resumed later, when the expressions of the integrals will be available.

3 Spectral method

The solution of Sturm-Liouville problems defined in the parent domain is typically used to synthesize spectral methods basis functions. Indeed, the spectral approximation of the solution of a differential problem is usually regarded as a finite expansion of eigenfunctions of a Sturm-Liouville problem; in spectral methods, the most appealing problems are the ones such that the expansion of an infinitely smooth function in terms of their eigenfunctions guarantees spectral accuracy. In particular, spectral accuracy is ensured if the Sturm-Liouville problem is singular [5]. Among these issues, particular importance rests with those problems whose eigenfunctions are algebraic polynomials, because of the efficiency with they can be evaluated and differentiated numerically. In this section the basis functions for spectral methods on single domains are defined starting from the main classes of orthogonal polynomials, and the mapping procedure from their parent domain to the physical one is described.

3.1 Legendre polynomials

Legendre polynomials are the eigenfunctions of the Legendre differential equation

$$\frac{\mathrm{d}}{\mathrm{d}u}\left[(1-u^2)\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{P}_n(u)\right] + n(n+1)\mathrm{P}_n(u) = 0. \tag{5}$$

The peculiarity of such polynomials is their orthonormality with respect to the scalar product

$$\int_{-1}^{+1} \mathrm{P}_n(u)\,\mathrm{P}_m(u)\,\mathrm{d}u = \frac{2}{2n+1}\delta_{mn} = ||\mathrm{P}_n(u)||_2^2\,\delta_{mn}. \tag{6}$$

This is useful to compute the (in this case, diagonal) mass matrix defined in (4). Even if MATLAB provides two functions to generate those polynomials, it is much more efficient to synthesize them "manually" by means of recurrence relations. So, starting from the first two orders,

$$\begin{aligned}
\mathrm{P}_0(u) &= 1 \\
\mathrm{P}_1(u) &= u \\
\mathrm{P}_{n+1}(u) &= \frac{1}{n+1}\left[(2n+1)u\mathrm{P}_n(u) - n\mathrm{P}_{n-1}(u)\right].
\end{aligned}$$

The $u = 1$ limit value of these polynomials is

$$\mathrm{P}_n(1) = 1.$$

Moreover, it can be shown that

$$\mathrm{P}_n(-u) = (-1)^n \mathrm{P}_n(u).$$

The first derivative of Legendre polynomials can be computed as

$$\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{P}_n(u) = \frac{n}{u^2-1}\left[u\mathrm{P}_n(u) - \mathrm{P}_{n-1}(u)\right].$$

At $u = 1$, this expression has a removable singularity, which should be treated analytically, leading to

$$\left.\frac{\mathrm{dP}_n(u)}{\mathrm{d}u}\right|_{u=1} = \frac{n(n+1)}{2}.$$

Legendre functions of even (odd) orders are even (odd) functions, so their derivatives are odd (even) functions; therefore, it can be written

$$\left.\frac{\mathrm{dP}_n(u)}{\mathrm{d}u}\right|_{-u} = (-1)^{n-1}\left.\frac{\mathrm{dP}_n(u)}{\mathrm{d}u}\right|_{u}.$$

The second derivative, if necessary, can be computed from (5):

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}u}\left[(1-u^2)\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{P}_n(u)\right] + n(n+1)\mathrm{P}_n(u) = \\
\frac{\mathrm{dP}_n(u)}{\mathrm{d}u}\frac{\mathrm{d}}{\mathrm{d}u}(1-u^2) + (1-u^2)\frac{\mathrm{d}^2\mathrm{P}_n(u)}{\mathrm{d}u^2} + n(n+1)\mathrm{P}_n(u) = \\
= (1-u^2)\frac{\mathrm{d}^2\mathrm{P}_n(u)}{\mathrm{d}u^2} - 2u\frac{\mathrm{dP}_n(u)}{\mathrm{d}u} + n(n+1)\mathrm{P}_n(u) = 0.
\end{aligned}$$

So:

$$\frac{\mathrm{d}^2\mathrm{P}_n(u)}{\mathrm{d}u^2} = \frac{1}{1-u^2}\left[2u\frac{\mathrm{dP}_n(u)}{\mathrm{d}u} - n(n+1)\mathrm{P}_n(u)\right].$$

It is once again necessary to solve the removable singularity at $u = 1$; in this case, this is more complicated, therefore it has been done by means of Wolfram Mathematica, with the following script:

```
f[x_] = LegendreP[n,x]
Limit[f''[x], x -> 1]
```

the result of this operation is:

$$\left.\frac{\mathrm{d}^2\mathrm{P}_n(u)}{\mathrm{d}u^2}\right|_{u=1} = \frac{1}{8}(n^3 + 2n^2 - n - 2).$$

The second derivative of Legendre polynomials has the same parity of the Legendre polynomials, therefore

$$\left.\frac{\mathrm{d}^2\mathrm{P}_n(u)}{\mathrm{d}u^2}\right|_{-u} = (-1)^n \left.\frac{\mathrm{d}^2\mathrm{P}_n(u)}{\mathrm{d}u^2}\right|_{u}.$$

3.2 Chebyshev polynomials

An alternative basis for spectral methods is the Chebyshev polynomials one. Usually, Legendre polynomials are more indicated to be applied in spectral methods, since they are defined naturally as the orthogonal polynomials with the standard scalar product. Chebyshev polynomials are still orthogonal, but with respect to a different scalar product; however, it may be useful to define them.

Chebyshev polynomials (of first kind) are solutions of the Chebyshev differential equation

$$(1-u^2)\frac{\mathrm{d}^2\mathrm{T}_n(u)}{\mathrm{d}u^2} - u\frac{\mathrm{d}\mathrm{T}_n(u)}{\mathrm{d}u} + n^2\mathrm{T}_n(u) = 0. \tag{7}$$

One of the peculiarity of such polynomials is their synthesis easiness:

$$\mathrm{T}_n(u) = \cos(n\arccos(u)). \tag{8}$$

Moreover, their derivative can be written as

$$\frac{\mathrm{d}\mathrm{T}_n(u)}{\mathrm{d}u} = n\mathrm{U}_{n-1}(u), \tag{9}$$

where U_n denote the n-th order Chebyshev polynomial of second kind; these polynomials can be synthesized simply as well as the first kind ones:

$$\mathrm{U}_n(u) = \frac{\sin((n+1)\vartheta)}{\sin\vartheta}, \qquad \vartheta = \arccos u.$$

Unfortunately, as well as Legendre polynomials concern, this expression exhibits a removable singularity for $\vartheta = n\pi$; therefore, its limit should be computed analytically:

$$\lim_{\vartheta\to 0} \mathrm{U}_n(u(\vartheta)) = n+1.$$

indeed, if $\vartheta = 0$, $u = \cos\vartheta = 1$, so

$$\left.\frac{\mathrm{d}\mathrm{T}_n(u)}{\mathrm{d}u}\right|_{u=+1} = n^2.$$

Instead, if $u = -1$, $\vartheta = \pi$,

$$\lim_{u=-1} \mathrm{U}_n(u) = \lim_{\vartheta\to\pi} \frac{\sin((n+1)\vartheta)}{\sin\vartheta} = (-1)^n(n+1).$$

To sum up, by recalling (9),

$$\left.\frac{\mathrm{d}\mathrm{T}_n(u)}{\mathrm{d}u}\right|_{u=-1} = (-1)^{n-1} n^2.$$

The second derivative of first-kind Chebyshev polynomials can be computed from the differential equation (7) as

$$\frac{\mathrm{d}^2\mathrm{T}_n(u)}{\mathrm{d}u^2} = \frac{1}{1-u^2}\left[u\frac{\mathrm{d}\mathrm{T}_n(u)}{\mathrm{d}u} - n^2\mathrm{T}_n(u)\right].$$

Once again, there are removable singularities for $u = \pm 1$. The limit for $u = 1$ can be solved with Mathematica:

```
f[x_] = ChebyshevT[n,x]
Limit[f''[x], x -> 1]
```

This results in:

$$\left.\frac{\mathrm{d}^2\mathrm{T}_n(u)}{\mathrm{d}u^2}\right|_{u=+1} = \frac{1}{3}n^3\left(-1+n^2\right).$$

For the $u = -1$ limit, it can be shown that $\mathrm{T}_n(u)$ is even (odd) for even (odd) n; then, its second derivative has the same parity properties, and

$$\left.\frac{\mathrm{d}^2\mathrm{T}_n(u)}{\mathrm{d}u^2}\right|_{-u} = (-1)^n \left.\frac{\mathrm{d}^2\mathrm{T}_n(u)}{\mathrm{d}u^2}\right|_{u}.$$

3.3 Mapping to a generic bounded interval

Legendre or Chebyshev polynomials are usually defined in a "parent" domain, $u \in [-1, 1]$. The main reason behind this fact is the definition of these polynomials, which are chosen to be orthogonal (with the relevant weights) when integrated in $[-1, 1]$. Moreover, even if such polynomials can be evaluated out of this parent domain, they would explode, causing numerical problems.

Since generally the ODE should be solved in a physical interval $[x_\mathrm{l}, x_\mathrm{r}]$, a mapping between these two "worlds" should be defined:

$$[-1, 1] \longrightarrow [x_\mathrm{l}, x_\mathrm{r}].$$

By invoking our high school studies, this can be written as a simple straight line:

$$\frac{x - x_\mathrm{l}}{x_\mathrm{r} - x_\mathrm{l}} = \frac{u - (-1)}{1 - (-1)}.$$

let $L = x_\mathrm{r} - x_\mathrm{l}$; then,

$$x = \frac{L}{2}u + \frac{L}{2} + x_{\mathrm{l}}. \tag{10}$$

Similarly, the inverse mapping $u = u(x)$ can be obtained by inverting (10), so

$$u = \frac{2}{L}(x - x_{\mathrm{l}}) - 1. \tag{11}$$

3.4 Semi-unbounded intervals: Laguerre polynomials

In order to handle semi-infinite intervals, Laguerre polynomials should be used to build the basis functions:

$$f_n(u) = \mathcal{L}_n(u)\,\mathrm{e}^{-\frac{u}{2}}.$$

Since in this case the parent domain u is $u \in [0, +\infty)$, the only interesting mapping is

$$x = \alpha u + x_{\mathrm{l}}, \qquad \alpha = \pm 1.$$

The inverse mapping is

$$u = \frac{x - x_{\mathrm{l}}}{\alpha}.$$

Therefore, the basis functions in the spatial domain is

$$f_n(x) = \mathcal{L}_n(u(x))\mathrm{e}^{-\frac{1}{2}u(x)},$$

and

$$\frac{\mathrm{d}f_n(x)}{\mathrm{d}x} = \frac{\mathrm{d}f_n(u)}{\mathrm{d}u}\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{1}{\alpha}\left[\frac{\mathrm{d}\mathcal{L}_n(u)}{\mathrm{d}u} - \frac{1}{2}\mathcal{L}_n(u)\right]\mathrm{e}^{-\frac{u}{2}}.$$

3.4.1 Application example: determination of the effective refractive index

The eigenproblem of a planar dielectric waveguide is [12]:

$$\frac{\mathrm{d}^2u(x)}{\mathrm{d}x^2} + (k^2 - \beta^2)u(x) = 0,$$

where the problem is the determination of the z propagation constant β. This problem can be slightly modified as

$$\frac{\mathrm{d}^2u(x)}{\mathrm{d}x^2} + k_0^2(n^2(x) - n_{\mathrm{eff}}^2)u(x) = 0,$$

and then it is cast into the eigenproblem

$$\frac{\mathrm{d}^2u(x)}{\mathrm{d}x^2} + k_0^2n^2(x)u(x) = k_0^2n_{\mathrm{eff}}^2u(x),$$

with eigenfunction $u(x)$ and eigenvalue $k_0^2 n_{\text{eff}}^2$. This problem can be formulated with SEM, by using the indications of 2.1. Just an additional formulation, about integration: for the $z = -u$ case, $\mathrm{d}z = -\mathrm{d}u$, and

$$\int_{-\infty}^{0} f(z)\,\mathrm{d}z = -\int_{+\infty}^{0} f(u)\,\mathrm{d}u = +\int_{0}^{+\infty} f(u)\,\mathrm{d}u.$$

3.5 Definition of the basis functions and calculation of the integrals

The mapping described in the previous section can be used to complete the synthesis of the spectral method basis functions; focusing on Legendre polynomials:

$$f_n(x) = \mathrm{P}_n(u(x)). \tag{12}$$

Derivatives can be computed by applying the chain rule

$$\frac{\mathrm{d}f_n(x)}{\mathrm{d}x} = \frac{\mathrm{dP}_n(u)}{\mathrm{d}u}\frac{\mathrm{d}u}{\mathrm{d}x},$$

where, from (11),

$$\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{2}{L}.$$

Similar calculations can be performed for the second derivative:

$$\frac{\mathrm{d}^2 f_n(x)}{\mathrm{d}x^2} = \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{\mathrm{d}f_n(x)}{\mathrm{d}x}\right] = \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{\mathrm{d}f_n(u)}{\mathrm{d}u}\right] = \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{\mathrm{d}f_n(x)}{\mathrm{d}x}\right] = \left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)^2 \frac{\mathrm{d}^2 f_n}{\mathrm{d}u^2}.$$

The integrals can be cast in the parent domain by applying the substitution theorem; starting from

$$\int_{x_\mathrm{l}}^{x_\mathrm{r}} f(x)\,\mathrm{d}x,$$

by recalling (10), it is obtained

$$\mathrm{d}x = \frac{L}{2}\,\mathrm{d}u.$$

so

$$\int_{x_\mathrm{l}}^{x_\mathrm{r}} f(x)\,\mathrm{d}x = \frac{L}{2}\int_{-1}^{+1} f(u(x))\,\mathrm{d}u.$$

This formula is very useful because it is well suited for the Gauss-Legendre quadrature rule. Moreover, for Legendre polynomials, this can be also computed analytically, as described in the following.

As a first "game problem" to introduce the integration procedure, let us consider the representation of an arbitrary function $f(u)$ with Legendre polynomials:

$$f(u) \simeq \sum_{n=0}^{N} c_n \mathrm{P}_n(u), \qquad u \in [-1, 1].$$

The approximation on the right-hand side minimizes the residual (omitted from the formulation for the sake of compactness) in the least squares sense, as usual. This can be performed in the spatial domain x as well, by using the aforementioned mapping, which is not introduced here to ease the calculations. As a first step, both members are projected on the p-th Legendre polynomial, for $p = 0...N$:

$$\int_{-1}^{+1} f(u)\,\mathrm{P}_p(u)\,\mathrm{d}u = \sum_{n=0}^{N} c_n \int_{-1}^{+1} \mathrm{P}_n(u)\,\mathrm{P}_p(u)\,\mathrm{d}u, \qquad p = 0...N.$$

Since $f(u)$ is an arbitrary function, the left-hand side integrals must be computed numerically; instead, for the right-hand side, from (6), it can be written

$$\int_{-1}^{+1} \mathrm{P}_n(u)\,\mathrm{P}_p(u)\,\mathrm{d}u = \delta_{pn}\frac{2}{2p+1}.$$

So, the equation reduces to (thanks to Kronecker's delta properties):

$$\int_{-1}^{+1} f(u)\,\mathrm{P}_p(u)\,\mathrm{d}u = c_p\frac{2}{2p+1},$$

then, by inverting,

$$c_p = \left(p + \frac{1}{2}\right)\int_{-1}^{+1} f(u)\,\mathrm{P}_p(u)\,\mathrm{d}u.$$

This can be written also in vector notation, as

$$\underline{c} = \underline{\underline{M}}^{-1}\underline{b}.$$

This problem helps us to introduce the calculation of

$$\int_{-1}^{+1} \frac{\mathrm{dP}_n(u)}{\mathrm{d}u}\mathrm{P}_p(u)\,\mathrm{d}u.$$

In other words: how can a derivative of $\mathrm{P}_n(u)$ be approximated with Legendre polynomials? This can be useful for spectral methods, to compute part of the system matrix. The objective is to write

$$\frac{\mathrm{dP}_n(u)}{\mathrm{d}u} = \sum_m c_m^{(n)}\mathrm{P}_m(u),$$

where $c_m^{(n)}$ is the coefficient to be multiplied to the m-th Legendre polynomial to obtain the derivative of the n-th Legendre polynomial. Then, by projecting, it is obtained

$$\int_{-1}^{+1} \frac{\mathrm{dP}_n(u)}{\mathrm{d}u} \mathrm{P}_p(u)\,\mathrm{d}u = c_p^{(n)} \frac{2}{2p+1},$$

or, more compactly,

$$\underline{d}^{(n)} = \underline{\underline{M}}\,\underline{c}^{(n)}.$$

This can be written in matrix form as

$$\underline{\underline{D}} = \underline{\underline{M}}\,\underline{\underline{C}},$$

by defining a matrix $\underline{\underline{D}}$ and a matrix $\underline{\underline{C}}$ with the column vectors $\{\underline{d}^{(n)}\}$, $\{\underline{c}^{(n)}\}$

$$\underline{\underline{D}} = \begin{bmatrix} \underline{d}^{(0)} & \underline{d}^{(1)} & \dots & \underline{d}^{(N)} \end{bmatrix} \qquad \underline{\underline{C}} = \begin{bmatrix} \underline{c}^{(0)} & \underline{c}^{(1)} & \dots & \underline{c}^{(N)} \end{bmatrix}.$$

The objective of this exercise is the determination of $\underline{\underline{D}}$, since $\underline{\underline{C}}$ can be found in textbooks, *e.g.,* [1, eq. (A.30)]. In order to obtain the second derivative projections,

$$\int_{-1}^{+1} \frac{\mathrm{d}^2\mathrm{P}_n(u)}{\mathrm{d}u^2} \mathrm{P}_p(u)\,\mathrm{d}u,$$

by using [1, eq. (A.31)].

The last integral that can be used in a spectral method formulation is

$$\int_{-1}^{+1} \frac{\mathrm{dP}_n(u)}{\mathrm{d}u} \frac{\mathrm{dP}_p(u)}{\mathrm{d}u}\,\mathrm{d}u.$$

This integral arises from an integration by parts, therefore the originating idea can be followed to obtain:

$$\int_{-1}^{+1} \frac{\mathrm{dP}_n(u)}{\mathrm{d}u} \frac{\mathrm{dP}_p(u)}{\mathrm{d}u}\,\mathrm{d}u = \underbrace{\frac{\mathrm{dP}_n(u)}{\mathrm{d}u} \mathrm{P}_p(u)\Big|_{-1}^{+1}}_{\text{analytic}} - \underbrace{\int_{-1}^{+1} \frac{\mathrm{d}^2\mathrm{P}_n(u)}{\mathrm{d}u^2} \mathrm{P}_p(u)\,\mathrm{d}u}_{[1,\text{(A.31)}]}.$$

3.5.1 Application example

Now, it is possible to complete the example reported in Section 2.1, since all the ingredients to build the system matrix have been provided. However, with this knowledge, it is not still possible to solve Dirichlet problems, since no information has been provided about the enforcement of Dirichlet boundary conditions; this will be the first point of the following section. As anticipated, if functions are not orthonormal, the mass matrix is not the identity, and this is a generalized linear eigenvalue problem, which should be solved with a specified routine, *e.g.,* in MATLAB,

```
[EigVc,EigVl]=eig(K,M);
```

3.5.2 Note on function approximation

If the function $f(u)$ to be approximated is a polynomial or at least it can be well approximated to a polynomial, so

$$f(u) = \sum_n a_n u^n,$$

it is possible to compute analytically the projection integrals, for Legendre functions, by computing

$$\int_{-1}^{+1} u^n \mathrm{P}_p(u)\,\mathrm{d}u,$$

where, for $n = 1$ the result is provided in [1, eq. (A.32)], and for $n = 2$ in [13]. By iterating the procedure described in the latter reference, further orders should be computed in a similar fashion.

4 Spectral element method: synthesis of basis functions

In the previous section, basis functions were simply entire-domain polynomials, leading to a purely p-refined method. Now, the method will be refined by showing how a domain decomposition strategy can be introduced. In the first section, still useful for single-domain spectral methods, the procedure aimed at enforcing essential boundary conditions, such as Dirichlet or pseudo-periodicity conditions, will be described. Then, a similar technique will be adopted to enforce continuity and orthonormality conditions to the basis functions to be adopted in representing/testing the solution.

4.1 Treatment of essential boundary conditions

The example described in Section 2.1 was tailored in such a way to avoid the synthesis of boundary-adapted basis functions, relying on the fact that Neumann's boundary conditions can be included in the functional minimization (natural boundary condition). By contrast, essential boundary conditions, which involve the value of the solution instead of its flux, should be enforced by restricting the function space used to describe the solution and to test it. This idea can be actuated by means of a *basis recombination procedure*: starting from the original set of basis functions, they can be mixed, recombined, in such a way to obtain a new set of basis functions, which satisfy the desired boundary condition. The latter set is included (stricly or not) in the former; to make this clear, if we compare the space of continuous functions with the one of continuous functions that equal zero in a point, the latter is surely a subspace of the former: there are "more continuous functions" than "continuous functions that vanish in a well defined point" !

Now, after this "delicious" theoretical discussion, the engineer escapes from the Hyperuranion, and asks: "Ok, nice, how to do implement this basis recombination?". Well, let us start

from the set of non specialized functions $\{\phi_i^{(j)}(x)\}$ defined on each j-th patch (since different resolutions can be desired for each patch, these sets can be different); for example, in the case of Legendre polynomials[5],

$$\phi_i^{(j)} = \mathrm{P}_n(u(x)),$$

where $u(x)$ is the inverse of the mapping from the parent domain to the patch interval, which can be easily derived from the ideas presented in Section 3.3. Now, let $\{h_k^{(j)}(x)\}$ be the set of basis functions satisfying the essential boundary condition derived from the recombination of $\{\phi_i^{(j)}(x)\}$; let us consider, as clarifying example, a homogeneous Dirichlet boundary condition in $x = x_0$; then, these functions must satisfy

$$h_k^{(j)}(x_0) = 0, \qquad \forall k.$$

The basis recombination procedure allows to obtain the k-th function $h_k^{(j)}$ from a linear combination of all the basis functions of the non-specialized set, so

$$h_k^{(j)}(x) = \sum_i H_{ki}^{(j)} \phi_i^{(j)}(x).$$

The coefficient $H_{ki}^{(j)}$ is the weight of the linear combination for the i-th non-specialized basis function to obtain the k-th boundary-adapted function; this, for each j-th patch; the objective is the determination of the resulting matrix $\underline{\underline{H}}$. The condition to be enforced is

$$\sum_i H_{ki}^{(j)} \phi_i^{(j)}(x_0) = 0.$$

This relationship can be written in matrix form as

$$(\underline{\phi}^{(j)}(x_0))^{\mathrm{T}} \underline{\underline{H}}^{(j)} = 0,$$

where the sum is replaced by the matrix product, and this for each k-th boundary-adapted function. This is a homogeneous system, whose non-trivial solutions are the basis of the kernel of the vector $(\underline{\phi}^{(j)}(x_0))^{\mathrm{T}}$. To this aim, it is useful to compute its singular value decomposition (SVD)

$$(\underline{\phi}^{(j)}(x_0))^{\mathrm{T}} = \underline{\underline{U}}^{(\mathrm{h},j)} \underline{\underline{S}}^{(\mathrm{h},j)} \underline{\underline{V}}^{(\mathrm{h},j)\mathrm{H}}.$$

The SVD of a row vector with N elements provides $N - 1$ null singular values. So, since the columns of $\underline{\underline{V}}^{(\mathrm{h},j)}$ corresponding to null singular values (below a certain threshold) are a basis of the kernel of $(\underline{\phi}^{(j)}(x_0))^{\mathrm{T}}$, such columns are used to build the change of basis matrix $\underline{\underline{H}}$ from non-specialized to boundary-adapted basis functions.

Probably, the procedure described in these sections is not the most efficient. Indeed, basis recombination approaches not based on heavy numerical algorithms such as the SVD can be

[5] obviously, this method can be applied also with patches with different polynomial or even function types, for example to treat singular solution behaviors or unbounded intervals

found in the literature [14], [15]. Moreover, the SVD is quite "unstable", since small variations of the coefficients lead to huge variations of the $\underline{\underline{U}}$ or $\underline{\underline{V}}$ matrices. However, this instability does not affect the stability of the entire method; moreover, this method is important since it can be directly used in the extension of multi-domain spectral methods to 2-D geometries [8].

4.1.1 Exercise

At this point, the exercise proposed in Section 2.1 can be repeated with homogeneous Dirichlet conditions, instead that homogeneous Neumann conditions; the eigenvalues are the natural numbers $k = 1, 2, ...$, so with $k = 0$ excluded.

4.2 Enforcing orthonormality

Numerical methods work better with independent basis functions to avoid redundant representations of the unknown, which are reflected in an ill-conditioned mass matrix; in this view, orthonormal functions are surely the best choice. Then, this section explains how to synthesize, starting from the boundary-adapted basis functions $\{h_k^{(j)}(x)\}$ obtained from the previous section, a set of orthonormal functions $\{g_r^{(j)}(x)\}$. To this aim, it is necessary to apply the Gram-Schmidt algorithm on $\{g_r^{(j)}(x)\}$. Under a different point of view, the two sets of functions are once again related by a change of basis

$$g_r^{(j)}(x) = \sum_k G_{rk}^{(j)} h_k^{(j)}(x),$$

and our objective is to build the matrix $\underline{\underline{G}}^{(j)}$. Such matrix can be calculated once again by the SVD, starting from the mass matrix of the boundary-adapted functions, with element:

$$(\underline{\underline{M}}^{(\mathrm{h},j)})_{mn} = \int_{\mathcal{D}^{(j)}} h_m^{(j)}(x) h_n^{(j)}(x)\,\mathrm{d}x.$$

Then, its SVD is computed, leading to

$$\underline{\underline{M}}^{(\mathrm{h},j)} = \underline{\underline{U}}^{(\mathrm{g},j)}\,\underline{\underline{S}}^{(\mathrm{g},j)}\,\underline{\underline{V}}^{(\mathrm{g},j)\mathrm{H}}.$$

The columns of $\underline{\underline{U}}^{(\mathrm{g},j)}$ corresponding to the non-zero singular values, *i.e.,* above a certain threshold, are an orthonormal basis of the range of $\underline{\underline{M}}^{(\mathrm{h},j)}$; such columns are used to build the change of basis matrix $\underline{\underline{G}}^{(j)}$.

4.3 Synthesis of entire-domain basis functions

The final step is the synthesis of entire-domain basis functions that satisfy continuity (or derivability) conditions at the junction points between two patches, starting from the orthonormal, boundary-adapted functions $\{g_r^{(j)}(x)\}$. Our basis recombination approach philosophy can be applied in this case as well, after defining a unique set of basis functions $\{g_\alpha(x)\}$ as the union of all the patch functions:

$$\{g_\alpha(x)\} = \bigcup_{j=1}^{\mathrm{N_p}} \left\{g_r^{(j)}(x)\right\}.$$

By this way, the multi-index α spans both the local and patch indexes r and j. Then, the entire-domain, orthonormal, boundary adapted functions $\{f_n(x)\}$ can be obtained as

$$f_n(x) = \sum_\alpha C_{n\alpha} g_\alpha(x).$$

If these functions should be just continuous in the point x_c between two patches i and j,

$$\sum_{\alpha(i)} C_{n\alpha(i)} g_{\alpha(i)}(x_\mathrm{c}) - \sum_{\alpha(j)} C_{n\alpha(j)} g_{\alpha(j)}(x_\mathrm{c}) = 0$$

should be satisfied. Here, $\alpha(i)$ and $\alpha(j)$ denote the sub-indexes in the patches i and j. Then, a row vector $\underline{d}$ with length equal to the sum of all functions belonging to the set $\{g_\alpha(x)\}$ is built, and it is non-zero only for the indices $\alpha(i)$ and $\alpha(j)$; finally, this vector is filled, in the relevant positions, with $(g_{\alpha(i)}(x_\mathrm{c}))^\mathrm{T}$, and $-g_{\alpha(j)}(x_\mathrm{c})$. Finally, a matrix $\underline{\underline{D}}$ is built with the rows obtained from each condition. Finally,

$$\underline{\underline{D}} = \underline{\underline{U}}^{(\mathrm{c})} \underline{\underline{S}}^{(\mathrm{c})} \underline{\underline{V}}^{(\mathrm{c})\mathrm{H}},$$

and, just like for essential boundary conditions, the change of basis matrix $\underline{\underline{C}}$ is built with the columns of $\underline{\underline{V}}^{(\mathrm{c})}$ corresponding to the null singular values.

References

[1] D. Gottlieb and S. A. Orszag, "Numerical analysis of spectral methods: theory and applications," *Society for Industrial and Applied Mathematics,* Philadelphia, Pennsylvania, 1977.

[2] P. P. Silvester and R. L. Ferrari, "Finite elements for electrical engineers," Third Edition, Cambridge University Press, Cambridge, 1996.

[3] M. Koshiba, "Optical waveguide theory by the finite element method," KTK Scientific Publishers, Tokyo, 1992.

[4] R. Vichnevetsky, "Computer methods for partial differential equations," vol. 1, Prentice-Hall Series in Computational Mathematics, New Jersey, 1981.

[5] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, "Spectral methods in fluid dynamics," *Springer-Verlag,* Berlin, Germany, 1988.

[6] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, "Spectral methods: fundamentals in single domains," *Springer-Verlag,* Berlin, Germany, 2006.

[7] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, "Spectral methods: evolution to complex geometries and applications to fluid dynamics," *Springer-Verlag,* Berlin, Germany, 2007.

[8] A. Tibaldi, "A mortar element method for the analysis of electromagnetic passive devices," Ph. D. Dissertation, Jan. 2014.

[9] O. A. Peverini, G. Addamo, G. Virone, R. Tascone, and R. Orta, "A spectral-element method for the analysis of 2-D waveguide devices with sharp edges and irregular shapes," *IEEE Trans. Microw. Theory Techn.,* vol. 59, no. 7, pp. 1685-1695, July 2011.

[10] M. Abramowitz and I. A. Stegun, "Handbook of mathematical functions," National Bureau of Standards, Applied Mathematics Series, Tenth printing, Dec. 1972.

[11] W. Magnus, F. Oberhettinger, and R. P. Soni, "Formulas and theorems for the special functions of mathematical physics," *Grundlehren der mathematischen Wissenschaften*, vol. 52, Berlin, 1966.

[12] R. Orta, "Modes of planar waveguides," *Notes from the "Passive Optical Components" course*, 2011.

[13] A. Tibaldi, "Analysis and design of high-performance horn antennas," M. Sc. Dissertation, Nov. 2011.

[14] R. H. Morf, "Exponentially convergent and numerically efficient solution of Maxwell's equations for lamellar gratings," *J. Opt. Soc. Am. A,* vol. 12, no. 5, pp. 1043-1056, May 1995.

[15] S. Bastonero, O. A. Peverini, R. Orta, and R. Tascone, "Anisotropic surface relief diffraction gratings under arbitrary plane wave incidence," *Opt. Quant. Electron.*, vol. 32, no. 6-8, pp. 1013-1025, 2000.