The generalized Graetz problem in finite domains

Jérôme Fehrenbach [*1], Frédéric de Gournay [†2], Charles Pierre [‡3] and Franck. Plouraboué [§4]

[1] Institut de Mathématiques de Toulouse, CNRS,
Université Paul Sabatier, Toulouse, France
[2] LMV, Université Versailles-Saint Quentin, France.
[3] Laboratoire de Mathématiques et de leurs Applications, , UMR CNRS 5142,
Université de Pau et des Pays de l'Adour, France.
[4] Institut de Mécanique des Fluides de Toulouse, UMR CNRS 5002
Université Paul Sabatier, Toulouse, France.

February, 2011

Keywords: convection-diffusion, variational formulation, Hilbert spaces

Subject classification: 76R99, 35A15, 65N25

Abstract

We consider the generalized Graetz problem associated with stationary convection-diffusion inside a domain having any regular three dimensional translationally invariant section and finite or semi-infinite extent. Our framework encompasses any previous "extended" and "conjugated" Graetz generalizations and provides theoretical bases for computing the orthogonal set of generalized two-dimensional Graetz modes. The theoretical framework both includes heterogeneous and possibly anisotropic diffusion tensor. In the case of semi-infinite domains, the existence of a bounded solution is shown from the analysis of a two-dimensional operator eigenvectors which form a basis of L^2. In the case of finite domains a similar basis can be exhibited and the mode's amplitudes can be obtained from the inversion of newly defined finite domain operator. Our analysis both includes the theoretical and practical issues associated with this finite domain operator inversion as well as its interpretation as a multi-reflection image method. Error estimates are provided when numerically truncating the spectrum to a finite number of modes. Numerical examples are validated for reference configurations and provided in non-trivial cases. Our methodology shows how to map the solution of stationary convection-diffusion problems in finite three dimensional domains into a two-dimensional operator spectrum, which leads to a drastic reduction in computational cost.

[*]jerome.fehrenbach@math.univ-toulouse.fr
[†]gournay@math.uvsq.fr
[‡]charles.pierre@univ-pau.fr
[§]plourab@imft.fr

Introduction

The Graetz problem was first settled as the stationary convection-dominated transport problem inside an axi-symmetrical Poiseuille flow in a semi-infinite cylinder [7]. It is the cornerstone of many practical applications. The associated orthogonal Graetz modes are interesting to consider since their projections into the imposed entrance boundary conditions provide a nice set of longitudinally exponentially decaying solution whichever the applied lateral boundary conditions, or the considered velocity field (see for exemple [13]). Since many important convective heat transfer problems share similar properties, the computation of a similar orthogonal basis has been attractive in many studies in a context where intensive computer simulations were difficult [20, 3]. Nevertheless the generalization of this concept to simple situations is not straightforward. When, for example, for the problem is no longer convection-dominated and longitudinal diffusion is considered, a situation refered to as the "extended" Graetz configuration (see for example [12, 6, 21, 10]), it is not simple to find a set of orthogonal modes. The same difficulty arises when coupling the convection-diffusion arising into the Poiseuille flow to pure diffusion into a surrounding cylinder, a configuration generally denoted "conjugated" Graetz configuration [2, 11, 4].

It is as late as 1980 than Papoutsakis *et al.* [15, 14], realized that a matrix operator acting upon a two-component temperature/longitudinal gradient vector (for the Graetz axi-symmetrical configuration) could provide a symmetric operator to the "extended" Graetz problem. The mathematical properties of this operator were nevertheless not deeply analyzed in [15, 14]; neither the compacity of the resolvent, the spectrum structure and location, the involved functional spaces, nor the numerical convergence were studied. One has to admit that, even limited in scope, this important contribution remained poorly cited and recognized until the late nineties, when it was realized that a similar approach could be adapted to any concentric axi-symmetrical configurations [16, 17, 9, 8], adding nevertheless a larger number of unkowns. Recently a detailed mathematical study of a generalized version of the Graetz problem, referred to as *generalized Graetz problem* here, for general non-axisymmetrical geometries, for any bounded velocity profile and including heterogeneous diffusivity, was presented in [18] and applied to infinite (at both ends) cylinder configurations. This mathematical study has brought to the fore the direct relevance of a new reformulation of the problem into a mixed form: adding to the original scalar temperature unknown a vectorial auxiliary unknown. This reformulation involves an operator, referred to as the *Graetz operator*, acting both on the scalar and vectorial unknowns. The Graetz operator was showed to be self adjoint, with compact resolvant in a proper functional setting. Its spectrum was proved to be composed of a double infinite discrete set of eigenvalues: a positive set (downstream modes) and a negative one (upstream modes).

The aim of the present contribution is to provide the mathematical analysis and numerical methods for solving the generalized Graetz problem in semi-

infinite and finite domains, as well as effective numerical methods to estimate the Graetz modes in the non-axisymmetrical case. These results are interesting since finite domains represent the most relevant configurations for applications such as, for example, convective heat pipes, the size of which is obviously finite.

Let us now describe more precisely the context of this study. This contribution addresses convection-diffusion/thermal transfer in a generalized cylindrical geometry $\Omega \times I$, where $\Omega \subset \mathbb{R}^2$ is a connected open domain and $I \subset \mathbb{R}$ is an interval, possibly unbounded at one or both of its ends. The fluid velocity inside the tube is denoted by $\mathbf{v}(\xi, z)$, whereas its temperature is denoted by $T(\xi, z)$ for $\xi = (x, y) \in \Omega$ and $z \in I$.

The fluid velocity $\mathbf{v}$ is assumed to be directed along the z direction and constant in the z variable, that is $\mathbf{v}(\xi, z) = v(\xi)\mathbf{e}_z$, where $\mathbf{e}_z$ is the unit vector in the z direction. Moreover, the velocity profile is assumed to be bounded, i.e $v \in \mathrm{L}^\infty(\Omega)$.

The conductivity matrix is supposed to be symmetric bounded, coercive and anisotropic in the ξ direction only, i.e. it is of the form

$$\begin{pmatrix} \sigma(\xi) & 0 \\ 0 & c(\xi) \end{pmatrix},$$

and there exists a constant $C > 1$ such that

$$C|\eta|^2 \geq \eta^T \sigma(\xi)\eta \geq C^{-1}|\eta|^2 \text{ and } C \geq c(\xi) \geq C^{-1}, \qquad \forall \xi \in \Omega, \eta \in \mathbb{R}^2. \quad (1)$$

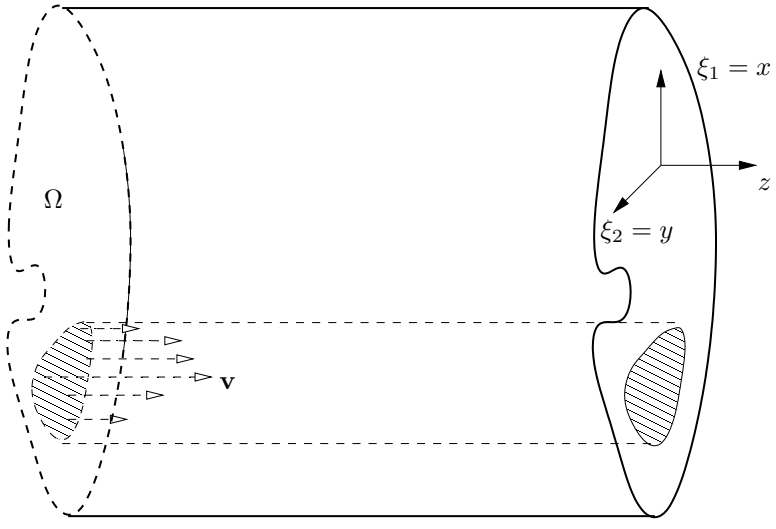

Figure 1: The geometry of the generalized Graetz problem

In this setting (see Figure 1), the steady convection-diffusion equation, refered to as the *generalized Graetz problem*, reads:

$$c(\xi)\partial_{zz}T + \mathrm{div}_\xi(\sigma(\xi)\nabla_\xi T) - \mathrm{Pe}v(\xi)\partial_z T = 0, \quad (2)$$

where Pe is the so-called Peclet number. In the sequel, the subscript ξ will be omitted and we will simply write: $\Delta = \Delta_\xi$, $\nabla = \nabla_\xi$, div $=$ div$_\xi$ for the Laplacian, gradient and divergence operators in the section Ω.

This problem is reduced to a system of two first order equations by introducing an additional vectorial unknown $\mathbf{p}$. Let $h = \mathrm{Pe}vc^{-1}$, we define the Graetz operator $\mathcal{A}$ by

$$\mathcal{A}\begin{pmatrix} T \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} hT - c^{-1}\mathrm{div}(\mathbf{p}) \\ \sigma\nabla T \end{pmatrix}, \tag{3}$$

in other words

$$\mathcal{A} = \begin{pmatrix} h & -c^{-1}\mathrm{div} \\ \sigma\nabla & 0 \end{pmatrix}. \tag{4}$$

The generalized Graetz problem defined in Equation (2) is then equivalent to the first-order system

$$\partial_z \psi(z) = \mathcal{A}\psi(z) \quad \text{with } \psi = \begin{pmatrix} \partial_z T \\ \sigma\nabla T \end{pmatrix}.$$

In [18] spectral properties of the operator $\mathcal{A}$ are established in order to derive exact solutions of the generalized Graetz problem on infinite geometries of the type $\Omega\times\mathbb{R}$ (unbounded ducts at both ends) involving a jump in the boundary conditions on $\partial\Omega$. It is proved that the spectrum consists of the eigenvalue 0 and two countable sequences of eigenvalues, one positive (downtream) and one negative (upstream) going both to infinity. Numerical approximations of this exact solution are given for axisymmetrical geometries.

However, on a semi-infinite duct $\Omega\times[0,+\infty)$, the projection of the entrance condition on the eigenmodes may provide non-zero coefficients associated to downstream modes. These coefficients yield a $T(z)$ that is unbounded as z goes to $+\infty$. The objective of the present work is then to provide a mathematical and numerical framework to solve the generalized Graetz problem on a semi-infinite duct that is adapted to any geometry of Ω. As a consequence of the forthcoming analysis, it is proved that the temperature components (T_n) of the upstream (resp. downstream) eigenmodes form a basis of $\mathrm{L}^2(\Omega)$. This analysis also provides a framework suitable to solve the problem on ducts of finite length. Error estimates for the operators induced on finite dimensional spaces associated to N upstream (or downstream) eigenmodes are provided. Finally a numerical implementation is proposed using a parametrization of the orthogonal of $ker\mathcal{A}$. Numerical examples provide a showcase of the power of the method.

The generalized Graetz problem is described in detail in Section 1, results obtained in [18] are recalled, and our main result (Theorem 1) is stated. In Section 2 we propose an equivalent formulation of this Theorem in the setting of finite sequences. In Section 3, our main result is proved in Proposition 2. Proposition 4 studies how the solution can be approximated when only the first modes of the operator $\mathcal{A}$ are known. These estimates are crucial in numerical

studies since only a part of the whole spectrum is computed. In Section 4 we solve different problems in semi-infinite and finite cylinders, and we show how the inequalities proved in Proposition 4 allow to obtain *a priori* inequalities on numerical approximations. After detailing the algorithm we use, Section 5 presents some of the numerical results we obtained.

1 Setting the problem

1.1 Spectral analysis

We recall the definition of the Sobolev spaces $\mathrm{L}^2(\Omega)$ and $\mathrm{H}^1(\Omega)$ on a smooth domain Ω. For that purpose, define the scalar products of functions:

$$(u,v)_0 = \int_\Omega u\bar{v}\,\mathrm{d}x \quad \text{and} \quad (u,v)_{\mathrm{H}^1(\Omega)} = \int_\Omega u\bar{v}\,\mathrm{d}x \; + \; \int_\Omega \nabla u \nabla \bar{v}\,\mathrm{d}x.$$

Then $\mathrm{L}^2(\Omega)$ (resp. $\mathrm{H}^1(\Omega)$) is defined as the subspace of measurable functions on Ω such that their $\mathrm{L}^2(\Omega)$ (resp $\mathrm{H}^1(\Omega)$)) norm induced by the corresponding scalar product is bounded. We also recall that the Sobolev space $\mathrm{H}^1_0(\Omega)$ is defined as the closure of the space of smooth functions with compact support for the $\mathrm{H}^1(\Omega)$ norm and that it can be identified with the subspace of functions of $\mathrm{H}^1(\Omega)$ that are equal to zero on $\partial\Omega$. In what follows the space $\mathrm{H}^1_0(\Omega)$ is endowed with the scalar product

$$(u,v)_1 = \int_\Omega \sigma \nabla u \nabla \bar{v}\,\mathrm{d}x$$

that defines a norm equivalent to the usual norm, thanks to the coercivity conditions (1) and the Poincaré inequality.

We define $\mathcal{H} = \mathrm{L}^2(\Omega) \times (\mathrm{L}^2(\Omega))^2$ and for every $\psi_i \in \mathcal{H}$, we use the notation $\psi_i = (T_i, \mathbf{p}_i)$ throughout this paper. Once endowed with the scalar product

$$(\psi_1|\psi_2)_{\mathcal{H}} = \int_\Omega cT_1\bar{T}_2 + \sigma^{-1}\mathbf{p}_1\bar{\mathbf{p}}_2\,\mathrm{d}x,$$

the vector space $\mathcal{H}$ is an Hilbert space. Denote $\mathrm{H}_{\mathrm{div}}(\Omega)$ the space defined by

$$\mathrm{H}_{\mathrm{div}}(\Omega) = \{\mathbf{p} \in (\mathrm{L}^2(\Omega))^2 \text{ such that } \mathrm{div}(\mathbf{p}) \in \mathrm{L}^2(\Omega)\}$$

and define the unbounded operator $\mathcal{A} : D(\mathcal{A}) = \mathrm{H}^1_0(\Omega) \times \mathrm{H}_{\mathrm{div}}(\Omega) \to \mathcal{H}$ as

$$\mathcal{A} : \psi = (T, \mathbf{p}) \mapsto \mathcal{A}\psi = (hT - c^{-1}\mathrm{div}(\mathbf{p}), \sigma\nabla T) \in \mathcal{H}.$$

$\mathcal{A}$ is a self-adjoint operator with a compact resolvent and hence is diagonal on a Hilbertian basis of $\mathcal{H}$. It is shown in [18] that the spectrum of $\mathcal{A}$ is $Sp(\mathcal{A}) = \{0\} \cup \{\lambda_n; n \in \mathbb{Z}^*\}$, where the λ_n are eigenvalues of finite order that can be ordered as follows:

$$-\infty \leftarrow \lambda_{-n} \leq \ldots \lambda_{-1} \leq \lambda_0 = 0 \leq \lambda_1 \cdots \leq \lambda_n \to +\infty.$$

The kernel of $\mathcal{A}$ consists of vectors of the form $(0, \mathbf{p}) \in D(\mathcal{A})$ with $\mathrm{div}(\mathbf{p}) = 0$. It follows from Helmholtz decomposition that its orthogonal in $\mathcal{H}$, the range of $\mathcal{A}$, is given by

$$\mathcal{R}(\mathcal{A}) = \{(f, \sigma\nabla s) \text{ with } (f, s) \in \mathrm{L}^2(\Omega) \times \mathrm{H}^1_0(\Omega)\}.$$

Because $\mathcal{A}$ is symmetric, it is bijective from $D(\mathcal{A}) \cap \mathcal{R}(\mathcal{A})$ onto $\mathcal{R}(\mathcal{A})$.

Denote $(\psi_n)_{n\in\mathbb{Z}^*}$ an orthonormalized basis of $\mathcal{R}(\mathcal{A})$ composed of eigenvectors ψ_n of $\mathcal{A}$ associated respectively to the eigenvalues $\lambda_n \neq 0$, then each $\psi_n = (T_n, \mathbf{p}_n)$ verifies:

$$\left\{ \begin{array}{rcl} \lambda_n \mathbf{p}_n & = & \sigma\nabla T_n, \\ \lambda_n^2 T_n + c^{-1}\mathrm{div}(\sigma\nabla T_n) - h\lambda_n T_n & = & 0, \end{array} \right. \tag{5}$$

and for every $n, m \in \mathbb{Z}^*$:

$$\int_\Omega c T_n T_m + \frac{1}{\lambda_n \lambda_m} \sigma\nabla T_n \nabla T_m = \delta_{nm},$$

where δ_{nm} stands for the Kronecker's symbol.

The diagonalization of the operator $\mathcal{A}$ ensures that if $\psi_{|_{z=0}} \in \mathcal{R}(\mathcal{A})$ is given, there exists a unique $\psi(z) \in C^0(I, \mathcal{R}(\mathcal{A}))$ that verifies in the weak sense

$$\partial_z \psi(z) = \mathcal{A}\psi(z) \quad \psi(0) = \psi_{|_{z=0}},$$

where verifying the above differential equation in the weak sense is tantamount to verifying

$$\int_I (\psi(z)| - \partial_z X(z))_{\mathcal{H}} \,\mathrm{d}z = \int_I (\psi(z)|\mathcal{A}X(z))_{\mathcal{H}} \,\mathrm{d}z \quad \forall X \in C^1_c(I, \mathcal{D}(\mathcal{A}) \cap \mathcal{R}(\mathcal{A})).$$

Moreover this unique $\psi(z)$ verifies the equation

$$\psi(z) = \sum_{n\in\mathbb{Z}^*} (\psi(0)|\psi_n)_{\mathcal{H}} \psi_n e^{\lambda_n z}.$$

Coming back to the original setting, if $T_{|_{z=0}} \in \mathrm{H}^1_0(\Omega)$ and $\partial_z T_{|_{z=0}} \in \mathrm{L}^2(\Omega)$ are given, then there exists a unique $T(z, \xi) \in C^0(\mathbb{R}, \mathrm{H}^1_0(\Omega)) \cap C^1(\mathbb{R}, \mathrm{L}^2(\Omega))$ solution of (2) which is given by

$$\psi(z) = \sum_{n\in\mathbb{Z}^*} (\psi(0)|\psi_n)_{\mathcal{H}} \psi_n e^{\lambda_n z}, \quad \psi(z) = \begin{pmatrix} \partial_z T(z) \\ \sigma\nabla T(z) \end{pmatrix}, \quad \psi_{|_{z=0}} \in \mathcal{R}(\mathcal{A}). \tag{6}$$

As a remark, following [18], if the initial boundary conditions are slightly less regular, that is $T_{|_{z=0}} \in \mathrm{L}^2(\Omega)$ and $\partial_z T_{|_{z=0}} \in \mathrm{H}^{-1}(\Omega)$, then there is still a unique solution to (2) in $C^0(\mathbb{R}, \mathrm{L}^2(\Omega)) \cap C^1(\mathbb{R}, \mathrm{H}^{-1}(\Omega))$, given by

$$\widetilde{\psi}(z) = \sum_{n\in\mathbb{Z}^*} (\widetilde{\psi}(0)|\psi_n)_{\mathcal{H}} \psi_n e^{\lambda_n z}, \quad \text{with } \widetilde{\psi}(z) = \begin{pmatrix} T(z) \\ \sigma\nabla s(z) \end{pmatrix}, \tag{7}$$

where, for any z (and specially for $z = 0$), $s(z)$ is the unique solution in $\mathrm{H}^1_0(\Omega)$ of

$$\mathrm{div}(\sigma \nabla s) = chT - c\partial_z T.$$

We remark that the previous equation determines uniquely $s_{|z=0}$ and hence $\psi_{|z=0}$ from the knowledge of $T_{|z=0}$ and $\partial_z T_{|z=0}$. Of course, if the initial conditions are regular enough, then ψ and $\widetilde{\psi}$ are linked by $\psi = \partial_z \widetilde{\psi}$.

1.2 Main result

Following the previous discussion, if the problem is set on the semi infinite duct $\Omega \times \mathbb{R}^-$, the initial conditions $T_{|z=0}$ and $\partial_z T_{|z=0}$ determine uniquely $\psi_{|z=0}$ (or $\widetilde{\psi}_{|z=0}$) and hence any value of $\psi(z)$. But in general this set of conditions yields a $T(z)$ that may be unbounded as z goes to $-\infty$. A natural question to ask is then, given $T_{|z=0}$ (resp. $\partial_z T_{|z=0}$) in $\mathrm{L}^2(\Omega)$, is it possible to find $\partial_z T_{|z=0}$ (resp $T_{|z=0}$) such that $T(z)$ stays bounded for z going to infinity ?

We reformulate this question as: Given $f \in \mathrm{L}^2(\Omega)$, is it possible to find an $s \in \mathrm{H}^1_0(\Omega)$ (preferably unique) such that $\psi = (f, \sigma \nabla s)$ verifies:

$$(\psi | \psi_n)_{\mathcal{H}} = 0 \quad \text{for all } n < 0 \qquad ?$$

The answer to this question is given by the following Theorem, which is a consequence of Proposition 2

Theorem 1. Given $f \in \mathrm{L}^2(\Omega)$, there exists a unique sequence $\mathbf{u} = (u_i)_{i \in \mathbb{N}^*}$ such that

$$f = \sum_{i>0} u_i T_i.$$

In this case, setting $s \in \mathrm{H}^1_0(\Omega)$ as $s = \sum_{i>0} \lambda_i^{-1} u_i T_i$ ensures that the decomposition of $(f, \sigma \nabla s)$ on the eigenmodes of $\mathcal{A}$ only loads positive eigenvalues and hence goes to 0 as z goes to $-\infty$. Of course, changing z into $-z$ (or equivalently changing the sign of h) transforms the problem from a decomposition on the downstream modes to a decomposition on the upstream modes.

2 Decomposition on the upstream modes

2.1 Isomorphism with the space of sequences

The choice of an Hilbertian basis induces an isomorphism between $\mathcal{R}(\mathcal{A})$ and the space of square summable sequences. Denote the discrete $\mathrm{l}^2(\mathbb{Z}^*)$ and $\mathrm{h}^1(\mathbb{Z}^*)$ scalar product, defined for complex sequences $\mathbf{a} = (a_n)_{n \in \mathbb{Z}^*}$ and $\mathbf{b} = (b_n)_{n \in \mathbb{Z}^*}$ as

$$(\mathbf{a}|\mathbf{b})_{\mathrm{l}^2(\mathbb{Z}^*)} = \sum_{n \in \mathbb{Z}^*} a_n \bar{b}_n \text{ and } (\mathbf{a}|\mathbf{b})_{\mathrm{h}^1(\mathbb{Z}^*)} = \sum_{n \in \mathbb{Z}^*} \lambda_n^2 a_n \bar{b}_n$$

and define the $\mathrm{l}^2(\mathbb{Z}^*)$ (resp. $\mathrm{h}^1(\mathbb{Z}^*)$) Hilbert space as the subspace of complex sequences such that their $\mathrm{l}^2(\mathbb{Z}^*)$ (resp. $\mathrm{h}^1(\mathbb{Z}^*)$) norm is bounded.

The mapping

$$\begin{array}{rccc} \chi: & \mathrm{l}^2(\mathbb{Z}^*) & \to & \mathcal{R}(\mathcal{A}) \\ & \mathbf{a} & \mapsto & \displaystyle\sum_{i\in\mathbb{Z}^*} a_i\psi_i \end{array}$$

with adjoint $\chi^\star : \psi \mapsto ((\psi|\psi_n)_{\mathcal{H}})_{n\in\mathbb{Z}^*}$ is an isometry, i.e both $\chi\chi^\star$ and $\chi^\star\chi$ are the identities of their respective spaces. Moreover $\chi(\mathrm{h}^1(\mathbb{Z}^*)) = \mathcal{R}(\mathcal{A}) \cap D(\mathcal{A})$ and $\chi^\star(\mathcal{R}(\mathcal{A})\cap D(\mathcal{A})) = \mathrm{h}^1(\mathbb{Z}^*)$. Of course, this change of variable diagonalizes $\mathcal{A}$ in the sense that if D is the operator

$$\begin{array}{rccc} D: & \mathrm{h}^1(\mathbb{Z}^*) & \to & \mathrm{l}^2(\mathbb{Z}^*) \\ & \mathbf{a} & \mapsto & (\lambda_n a_n)_n \end{array}$$

then

$$\mathcal{A} = \chi D \chi^\star.$$

2.2 Reformulation of the problem in the setting of sequences

In order to reformulate our problem in a discrete setting, let us define the following operators

Definition 1. Define P_1 and P_2 as

$$\begin{array}{rccc} P_1: & \mathcal{R}(\mathcal{A}) & \longrightarrow & \mathrm{L}^2(\Omega) \\ & (f, \sigma\nabla s) & \longmapsto & c^{1/2} f \end{array} \qquad \begin{array}{rccc} P_2: & \mathcal{R}(\mathcal{A}) & \longrightarrow & \mathrm{H}^1_0(\Omega) \\ & (f, \sigma\nabla s) & \longmapsto & s \end{array}$$

with adjoints defined by

$$\begin{array}{rccc} P_1^\star: & \mathrm{L}^2(\Omega) & \longrightarrow & \mathcal{R}(\mathcal{A}) \\ & f & \longmapsto & (c^{-1/2} f, 0) \end{array} \qquad \begin{array}{rccc} P_2^\star: & \mathrm{H}^1_0(\Omega) & \longrightarrow & \mathcal{R}(\mathcal{A}) \\ & s & \longmapsto & (0, \sigma\nabla s). \end{array}$$

Then trivially $P_i P_i^\star = Id$, $P_i^\star P_i$ is a projection and $P_1^\star P_1 + P_2^\star P_2 = Id$. Moreover $P_i P_j^\star = 0$ if $i \neq j$.

We shall also need the following technical definition

Definition 2. For $m < M$ in $\mathbb{Z}^*$, denote $l^2(\llbracket m, M \rrbracket)$ the subspace of $\mathrm{l}^2(\mathbb{Z}^*)$ of sequences $\mathbf{a}$ such that $a_n = 0$ if $n \notin \llbracket m, M \rrbracket$, and define the projection $\Pi_{m,M} : \mathrm{l}^2(\mathbb{Z}^*) \to l^2(\llbracket m, M \rrbracket)$ by

$$(\Pi_{m,M}\mathbf{u})_i = \begin{cases} u_i & \text{if } m \leq i \leq M \\ 0 & \text{if } i < m \text{ or } i > M. \end{cases}$$

For $m > 0$ the space $l^2(\llbracket m, \infty \llbracket)$ is the subspace of $\mathrm{l}^2(\mathbb{Z}^*)$ of sequences $\mathbf{a}$ such that $a_n = 0$ if $n < m$.

Proposition 1. Define the operator $K : \mathrm{l}^2(\mathbb{Z}^*) \longrightarrow \mathrm{l}^2(\mathbb{Z}^*)$ by

$$K = \chi^\star P_1^\star P_1 \chi.$$

Then $K = K^2$ (K is an orthogonal projection). Moreover proving Theorem 1 is equivalent to proving that

For every $\mathbf{a} \in \mathrm{l}^2(\mathbb{Z}^*)$ such that $K\mathbf{a} = \mathbf{a}$ there is a unique $\mathbf{u} \in l^2([\![1, \infty[\![)$ such that $K\mathbf{u} = \mathbf{a}$

Proof. The fact that $K^2 = K$ follows from the fact that $\chi\chi^\star = Id$ and $P_1 P_1^\star = Id$. By definition of P_1, χ, K for every $f \in \mathrm{L}^2(\Omega)$ and $\mathbf{u} = (u_i)$

$$f = \sum_i u_i T_i \Leftrightarrow f = c^{-1/2} P_1 \chi \mathbf{u} \Leftrightarrow \chi^\star P_1^\star \sqrt{c} f = K\mathbf{u},$$

where the last equivalence is proven using the definition of K for the direct implication and the property $(P_1\chi)(\chi^\star P_1^\star) = Id$ for the reciprocal implication. We now claim that

$$K\mathbf{a} = \mathbf{a} \Leftrightarrow \exists f \in \mathrm{L}^2(\Omega) \text{ such that } \mathbf{a} = \chi^\star P_1^\star \sqrt{c} f.$$

Once again, the reciprocal implication is proven by applying K on both sides of the identity and using $(P_1\chi)(\chi^\star P_1^\star) = Id$, whereas the direct implication is proven by setting $f = c^{-1/2}(P_1\chi)\mathbf{a}$ and using

$$\mathbf{a} = K\mathbf{a} = \chi^\star P_1^\star P_1 \chi \mathbf{a} = \chi^\star P_1^\star \sqrt{c} f.$$

□

In order to prove Theorem 1 using the equivalence from Proposition 1, we have to translate the eigenproblem equation in the setting of the space of sequences which is the purpose of the forthcoming theorem.

Theorem 2. For each $\mathbf{a} \in \mathrm{h}^1(\mathbb{Z}^*), \mathbf{b} \in \mathrm{l}^2(\mathbb{Z}^*)$, we have

$$KD^{-1}K = 0, \tag{8}$$
$$(Id - K)D(Id - K)\mathbf{a} = 0, \tag{9}$$

and

$$(KDK\mathbf{a}|\mathbf{b})_{l^2} = \int_\Omega h(P_1\chi\mathbf{a})(P_1\chi\mathbf{b})\,\mathrm{d}x. \tag{10}$$

Proof. By definition of $\mathcal{A}$, for any $(f, \sigma\nabla s) \in \mathcal{D}(\mathcal{A})$

$$\mathcal{A}\begin{pmatrix} f \\ \sigma\nabla s \end{pmatrix} = \begin{pmatrix} hf - c^{-1}\mathrm{div}(\sigma\nabla s) \\ \sigma\nabla f \end{pmatrix}.$$

This transforms into

$$\mathcal{A}P_1^\star(f) = P_1^\star(hf) + P_2^\star(c^{-1/2}f), \qquad \mathcal{A}P_2^\star(s) = P_1^\star(-c^{-1/2}\mathrm{div}(\sigma\nabla s)). \tag{11}$$

We prove (9) using $P_2 P_1^\star = 0$ and multiplying the second equation of (11) by P_2:

$$P_2 \mathcal{A} P_2^\star = 0 \Rightarrow P_2(\chi D \chi^\star) P_2^\star = 0 \Rightarrow \chi^\star P_2^\star (P_2 \chi D \chi^\star P_2^\star) P_2 \chi = 0.$$

This in turn implies that for any $\mathbf{a} \in \mathrm{h}^1(\mathbb{Z}^*)$, $(Id - K)D(Id - K)\mathbf{a} = 0$ since $Id - K = \chi^\star P_2^\star P_2 \chi$.

In order to prove (10), use $P_1 P_2^\star = 0$ and multiply the first equation of (11) by P_1. Then for each $f \in P_1(\mathcal{R}(\mathcal{A}) \cap D(\mathcal{A}))$

$$P_1(\chi D \chi^\star) P_1^\star(f) = P_1 \mathcal{A} P_1^\star(f) = hf.$$

If $\mathbf{a} \in \mathrm{h}^1(\mathbb{Z}^*)$ then $f = P_1 \chi \mathbf{a} \in P_1(\mathcal{R}(\mathcal{A}) \cap D(\mathcal{A}))$, the above equation applies and

$$\begin{aligned} & hP_1\chi\mathbf{a} = P_1\chi DK\mathbf{a} \\ \Rightarrow \quad & (hP_1\chi\mathbf{a}, P_1\chi\mathbf{b})_0 = (P_1\chi DK\mathbf{a}, P_1\chi\mathbf{b})_0 = (\chi^\star P_1^\star P_1 \chi DK\mathbf{a}, \mathbf{b})_{l^2} = (KDK\mathbf{a}, \mathbf{b})_{l^2} \end{aligned}$$

In order to prove (8), multiply the second equation of (11) by $P_1 \mathcal{A}^{-1}$ in order to get

$$0 = P_1 \mathcal{A}^{-1} P_1^\star(\mathrm{div}(c^{-1/2}\sigma\nabla s)) = P_1 \chi D^{-1} \chi^\star P_1^\star(c^{-1/2}\mathrm{div}(\sigma\nabla s)) \quad \forall s \in \mathrm{H}_0^1(\Omega).$$

For any $\mathbf{b} \in \mathrm{l}^2(\mathbb{Z}^*)$ define $f = P_1\chi\mathbf{b} \in \mathrm{L}^2(\Omega)$. There exists $s \in \mathrm{H}_0^1(\Omega)$ such that $\mathrm{div}(\sigma\nabla s) = c^{1/2} f$, and the above equation amounts to $KD^{-1}K\mathbf{b} = 0$. □

3 Properties of the sequential operators

3.1 The case $h = 0$

It is interesting to understand what happens in the purely diffusive case where $h = 0$. In this case, denote (S_n) the eigenvectors of the Laplacian associated to eigenvalues (μ_n^2) with $\mu_n > 0$:

$$-c^{-1}\mathrm{div}(\sigma\nabla S_n) = \mu_n^2 S_n \quad \text{ with } \int_\Omega c S_i S_j \,\mathrm{d}x = \delta_{ij} \text{ and } S_n \in \mathrm{H}_0^1(\Omega).$$

Then the eigenvectors of $\mathcal{A}$ are given exactly by

$$\psi_{\pm n} = \frac{1}{\sqrt{2}} \begin{pmatrix} S_n \\ \pm \mu_n^{-1} \sigma \nabla S_n \end{pmatrix} \text{ associated to the eigenvalues } \pm\,\mu_n,$$

and hence $T_n = T_{-n} = \frac{1}{\sqrt{2}} S_n$. In this case the restriction of K to the finite dimensional space $l^2([\![-N, N]\!])$ has the following simple form. Denote $\mathbf{e}_i = (\delta_{in})_{n\in\mathbb{Z}^*}$ the i^{th} vector of the canonical basis of the space of sequences. Then

$$\begin{aligned} (K\mathbf{e}_i|\mathbf{e}_j)_{\mathrm{l}^2(\mathbb{Z}^*)} = (P_1\chi\mathbf{e}_i|P_1\chi\mathbf{e}_j)_0 &= (P_1 \begin{pmatrix} T_i \\ \mu_i^{-1}\sigma\nabla T_i \end{pmatrix} | P_1 \begin{pmatrix} T_j \\ \mu_j^{-1}\sigma\nabla T_j \end{pmatrix})_0 \\ &= \int_\Omega c T_i T_j \,\mathrm{d}x. \end{aligned}$$

In the particular case $h = 0$,

$$\int_\Omega cT_iT_j \,\mathrm{d}x = \int_\Omega c\frac{1}{\sqrt{2}}S_{|i|}\frac{1}{\sqrt{2}}S_{|j|}\,\mathrm{d}x = \frac{1}{2}\delta_{|i|,|j|},$$

and we have

$$\Pi_{-N,N}K\Pi_{-N,N} = \frac{1}{2}\begin{pmatrix} Id & Id^\dagger \\ Id^\dagger & Id \end{pmatrix}, \quad Id^\dagger = \begin{pmatrix} 0 & \cdots & 1 \\ 0 & \diagup & 0 \\ 1 & \cdots & 0 \end{pmatrix}, \quad (Id^\dagger)_{i,j} = \delta_{i+j,N+1}.$$

In this setting, solving the problem of Proposition 1 is trivial. For any sequence $\mathbf{a} = (a_n)_n \in \mathrm{l}^2(\mathbb{Z}^*)$, $K\mathbf{a} = \mathbf{a}$ means that $a_{-n} = a_n$ and it is then sufficient to take $\mathbf{u} = (u_n)_n$ defined by:

$$\text{for } n < 0, \text{ take } u_n = 0 \text{ and for } n > 0, \text{ take } u_n = (a_n + a_{-n}) = 2a_n$$

This simple example is important to point out, since the case $h \neq 0$ is just a compact perturbation of the case $h = 0$. Indeed, coming back to equation (5), at order 0 when λ_n goes to infinity, we have:

$$\lambda_n^2 T_n + c^{-1}\mathrm{div}(\sigma\nabla T_n) = 0$$

and hence, when n goes to infinity, one expects $\lambda_{\pm n} \simeq \pm\mu_n$ and $T_{\pm n} \simeq \frac{1}{\sqrt{2}}S_n$, see Remark 1 for a precise statement of this assertion.

3.2 Existence and uniqueness of the solution

The next result is the main ingredient in the proof of Theorem 1.

Proposition 2. Suppose that $m \in \mathbb{N}^*$, $M > m$ possibly with $M = +\infty$ and denote $\pi = \Pi_{m,M}$.

For any $\mathbf{a} \in \mathrm{l}^2(\mathbb{Z}^*)$ there exists a unique $\mathbf{u} \in l^2([\![m, M]\!])$ solution of $\pi K\mathbf{u} = \pi\mathbf{a}$. Moreover this $\mathbf{u}$ satisfies

$$\|\mathbf{u}\|_{\mathrm{l}^2(\mathbb{Z}^*)} \leq (2 + \frac{\|h\|_{\mathrm{L}^\infty(\Omega)}}{\lambda_m})\|\pi\mathbf{a}\|_{\mathrm{l}^2(\mathbb{Z}^*)}. \tag{12}$$

Moreover, if $\mathbf{a} = K\mathbf{a}$, then $P_1\chi\mathbf{u}$ is the L^2 orthogonal projection of $P_1\chi\mathbf{a}$ on the space $Vect(c^{1/2}T_m, c^{1/2}T_{m+1}, \ldots, T_M)$.

Additionnaly, if $m = 1$ and $M = +\infty$ and $\mathbf{a} = K\mathbf{a}$, then we also have $K\mathbf{u} = \mathbf{a}$.

As an immediate corollary, the last assertion of this Proposition proves Theorem 1 via the equivalence pointed out in Proposition 1.

Proof. We first suppose that $M < +\infty$, then $Im(\pi) = l^2([\![m, M]\!])$ is a finite dimensional subspace on which the endomorphism $\bar{K} = \pi K\pi$ is real symmetric, hence diagonalisable. It is sufficient to show that, on this space any eigenvalue

of $\bar{K}$ is greater than $C = (2 + \frac{\|h\|_{\mathrm{L}^\infty(\Omega)}}{\lambda_m})^{-1}$ in order to prove existence of $\mathbf{u}$, uniqueness and the bound in the l^2 norm.

Let ρ be an eigenvalue of $\bar{K}$ and $\mathbf{v}$ an associated normalized eigenvector: $\pi K \pi \mathbf{v} = \rho \mathbf{v}$, $(\mathbf{v}|\mathbf{v})_{l^2} = 1$ and $\pi \mathbf{v} = \mathbf{v}$. Since

$$\rho = (\pi K \pi \mathbf{v}|\mathbf{v})_{l^2} = (K\pi\mathbf{v}|\pi\mathbf{v})_{l^2} = (K\pi\mathbf{v}|K\pi\mathbf{v})_{l^2} = \|K\mathbf{v}\|^2_{l^2},$$

then $0 \leq \rho \leq 1$. In order to prove the lower bound on the l^2 norm, recall that since $\mathbf{v}$ is a finite sequence then (10) applies and

$$|(KDK\mathbf{v}|\mathbf{v})_{l^2}| = |\int_\Omega h(P_1\chi\mathbf{v})^2 \,\mathrm{d}x| \leq \|h\|_{\mathrm{L}^\infty(\Omega)}\|P_1\chi\mathbf{v}\|_0^2 = \|h\|_{\mathrm{L}^\infty(\Omega)}\|K\mathbf{v}\|^2_{\mathrm{l}^2(\mathbb{Z}^*)}.$$

using (9) $((Id-K)D(Id-K)\mathbf{v}|\mathbf{v})_{l^2} = 0$ and $\pi D = D\pi$, we have

$$(KDK\mathbf{v}|\mathbf{v})_{l^2} = (2\rho - 1)(D\mathbf{v}|\mathbf{v})_{l^2}$$

Since $|(D\mathbf{v}|\mathbf{v})_{l^2}| = |\sum_{n=m}^M \lambda_n v_n v_n| \geq \lambda_m (\mathbf{v}|\mathbf{v})_{l^2} \geq \lambda_m$, we have

$$|\lambda_m(2\rho - 1)| \leq \|h\|_{\mathrm{L}^\infty(\Omega)}\|K\mathbf{v}\|^2_{\mathrm{l}^2(\mathbb{Z}^*)} = \|h\|_{\mathrm{L}^\infty(\Omega)}\rho \tag{13}$$

Which in turns means that $\rho \geq C$.

Consider now the case $M = +\infty$ where any $\mathbf{a} \in l^2([\![1, +\infty[\![)$ is the strong l^2 limit of $\Pi_{m,p}\mathbf{a}$ as p goes to infinity. Passing to the limit, we recover

$$(\pi K \pi \mathbf{a}, \mathbf{a})_{\mathrm{l}^2(\mathbb{Z}^*)} \geq C\|\mathbf{a}\|^2.$$

The Lax-Milgram theorem applies and $\pi K \pi : l^2([\![m, +\infty[\![) \to l^2([\![m, +\infty[\![)$ is a bijection with a continuous inverse bounded by C in the operator norm.

We now turn our attention to the geometrical interpretation of $\mathbf{u}$. By definition, $c^{1/2}T_i = P_1\chi\mathbf{e}_i$, where $\mathbf{e}_i$ is the i^{th} canonical basis vector of $\mathrm{l}^2(\mathbb{Z}^*)$, hence, if $\mathbf{a} = K\mathbf{a}$, for all $i \in [\![m, M]\!]$

$$\begin{aligned}(P_1\chi\mathbf{a} - P_1\chi\mathbf{u}|c^{1/2}T_i) &= (P_1\chi\mathbf{a} - P_1\chi\mathbf{u}|P_1\chi\mathbf{e}_i) = (\chi^\star P_1^\star P_1 \chi(\mathbf{a}-\mathbf{u})|\mathbf{e}_i) = (K\mathbf{a} - K\mathbf{u}|\mathbf{e}_i)\\ &= (\mathbf{a} - K\mathbf{u}|\mathbf{e}_i) = (\mathbf{a} - K\pi\mathbf{u}|\pi\mathbf{e}_i) = (\pi\mathbf{a} - \pi K\pi\mathbf{u}|\mathbf{e}_i) = 0\end{aligned}$$

Hence $P_1\chi\mathbf{u} \in Vect(c^{1/2}T_i)_{i=m..M}$ is the L^2 orthogonal projection of $P_1\chi\mathbf{a}$ on $Vect(c^{1/2}T_i)_{i=m..M}$.

We finally prove that if $m = 1, M = +\infty$ and $K\mathbf{a} = \mathbf{a}$, then $K\mathbf{u} = \mathbf{a}$. Define $\mathbf{b} = K\mathbf{u} - \mathbf{a} = K(\mathbf{u} - \mathbf{a})$, then $K\mathbf{b} = \mathbf{b}$. Since we already have $\pi K\mathbf{u} = \pi\mathbf{a}$, then $\pi\mathbf{b} = 0$. Using (8): $KD^{-1}K = 0$, we have

$$0 = (KD^{-1}K\mathbf{b}|\mathbf{b}) = (D^{-1}K\mathbf{b}|K\mathbf{b}) = (D^{-1}\mathbf{b}|\mathbf{b}) = \sum_{i<0}\lambda_i|b_i|^2.$$

Since all the λ_i are strictly negative, then $b_i = 0$ for all $i < 0$ and since $\pi\mathbf{b} = 0$, we finally have $\mathbf{b} = 0$. □

Remark 1. The bound (12) is indeed sharp, since, in the case $h = 0$, we have $\mathbf{u} = 2\pi\mathbf{a}$. Indeed, in this case, the matrix $\Pi_{m,M}K\Pi_{m,M} = 1/2Id$. Moreover, when $\lambda_m > \|h\|_{\mathrm{L}^\infty(\Omega)}/2$, the bound (13) translates into:

$$2 - \frac{\|h\|_{\mathrm{L}^\infty(\Omega)}}{\lambda_m} \leq \rho^{-1} \leq 2 + \frac{\|h\|_{\mathrm{L}^\infty(\Omega)}}{\lambda_m}.$$

Hence, when m goes to $+\infty$ and $M > m$, every eigenvalue of the matrix $\Pi_{m,M}K\Pi_{m,M}$ goes to $\frac{1}{2}$. Anticipating on the results of Proposition 3 that asserts that every off-diagonal term K_{ij} of $\Pi_{m,M}K\Pi_{m,M}$ is bounded like $\|h\|_{\mathrm{L}^\infty(\Omega)}/(\lambda_i+\lambda_j)$, we can conclude that the matrix $\Pi_{m,M}K\Pi_{m,M}$ tends towards the matrix $\frac{1}{2}Id$ as m goes to $+\infty$. Hence, as expected, when m goes to infinity, the effect of h wears off and K behaves as if the compact perturbation h was inexistent.

3.3 Bounds for the approximation

The result of Proposition 2 states that the sought $\mathbf{u}$ solves the equation

$$\pi K \pi \mathbf{u} = \pi \mathbf{a}$$

with $\pi = \Pi_{1,\infty}$. But in practice, we can only compute this matrix for $\pi = \Pi_{1,N}$ with a finite N. Therefore, we wish to estimate the resulting error. For that purpose, we first prove that the off-diagonal terms of $\pi K \pi$ are small.

Proposition 3. For $i = 1, 2$, let $m_i, M_i \in \mathbb{N}^*$, and denote $\pi_i = \Pi_{m_i,M_i}$. We assume that $\pi_1\pi_2 = 0$, (or equivalently $[\![m_1, M_1]\!] \cap [\![m_2, M_2]\!] = \emptyset$). Then

$$\|\pi_1 K \pi_2 \mathbf{u}\|_{\mathrm{l}^2(\mathbb{Z}^*)} \leq \frac{\|h\|_{\mathrm{L}^\infty(\Omega)}}{\lambda_{m_1} + \lambda_{m_2}} \|\pi_2 \mathbf{u}\|_{\mathrm{l}^2(\mathbb{Z}^*)} \quad \forall \mathbf{u} \in \mathrm{l}^2(\mathbb{Z}^*).$$

Proof. Let ρ be the largest eigenvalue on $Im(\pi_2)$ of

$$\pi_2 K \pi_1 K \pi_2 \mathbf{v} = \rho \mathbf{v} \text{ with } \mathbf{v} = \pi_2 \mathbf{v} \in Im(\pi_2),$$

where $\mathbf{v}$ is a corresponding eigenvector such that $\|\mathbf{v}\|_{l^2} = 1$. We claim that it is sufficient to show that

$$0 \leq \rho \leq \left(\frac{\|h\|_{\mathrm{L}^\infty(\Omega)}}{\lambda_{m_1} + \lambda_{m_2}} \right)^2 \tag{14}$$

Indeed, the inequality to be proven in Proposition 3 is, for all $\mathbf{u} \in l^2\mathbb{Z}^*$:

$$(\pi_2 K \pi_1 K \pi_2 \mathbf{u} | \pi_2 \mathbf{u})_{l^2} = \|\pi_1 K \pi_2 \mathbf{u}\|^2_{\mathrm{l}^2(\mathbb{Z}^*)} \leq \left(\frac{\|h\|_{\mathrm{L}^\infty(\Omega)}}{\lambda_{m_1} + \lambda_{m_2}} \right)^2 \|\pi_2 \mathbf{u}\|^2_{\mathrm{l}^2(\mathbb{Z}^*)},$$

which is exactly tantamount to proving (14). First, ρ is positive since

$$\rho = (\pi_2 K \pi_1 K \pi_2 \mathbf{v}, \mathbf{v}) = (\pi_1 K \pi_2 \mathbf{v}, K \pi_2 \mathbf{v}) \geq 0.$$

In order to prove the upper bound on ρ, set $\mathbf{a} = \pi_1 K \pi_2 \mathbf{v}$ and $\mathbf{b} = \pi_2 \mathbf{v}$, then trivially $\pi_1 K \mathbf{b} = \mathbf{a}$ and the eigenvector equation reads $\pi_2 K \mathbf{a} = \rho \mathbf{b}$. Moreover, since D is a diagonal operator that commutes with π_1 and π_2, then

$$\mathbf{a} = \pi_1 \mathbf{a} \Rightarrow D\mathbf{a} = \pi_1 D\mathbf{a} \text{ and } \mathbf{b} = \pi_2 \mathbf{b} \Rightarrow D\mathbf{b} = \pi_2 D\mathbf{b}$$

and hence

$$\begin{aligned}(DK\mathbf{a}|\mathbf{b})_{l^2} + (KD\mathbf{a}|\mathbf{b})_{l^2} = (K\mathbf{a}|D\mathbf{b})_{l^2} + (D\mathbf{a}|K\mathbf{b})_{l^2} = (K\mathbf{a}|\pi_2 D\mathbf{b})_{l^2} + (\pi_1 D\mathbf{a}|K\mathbf{b})_{l^2} \\ = (\pi_2 K\mathbf{a}|D\mathbf{b})_{l^2} + (D\mathbf{a}|\pi_1 K\mathbf{b})_{l^2} = \rho(\mathbf{b}|D\mathbf{b})_{l^2} + (D\mathbf{a}|\mathbf{a})_{l^2}.\end{aligned}$$

Since $\pi_1 \pi_2 = 0$, then $(D\mathbf{a}|\mathbf{b})_{l^2} = 0$ and (9) turns into

$$(KDK\mathbf{a}|\mathbf{b})_{l^2} = (DK\mathbf{a}|\mathbf{b})_{l^2} + (KD\mathbf{a}|\mathbf{b})_{l^2}.$$

On the other hand, (10) reads

$$\begin{aligned}(KDK\mathbf{a}|\mathbf{b})_{l^2} &= \int_\Omega h(P_1\chi\mathbf{a})(P_1\chi\mathbf{b})\,\mathrm{d}x \leq \|h\|_{\mathrm{L}^\infty(\Omega)}\|(P_1\chi\mathbf{a})\|_0\|(P_1\chi\mathbf{b})\|_0 \\ &\leq \|h\|_{\mathrm{L}^\infty(\Omega)}\|\mathbf{a}\|_{\mathrm{l}^2(\mathbb{Z}^*)}\|\mathbf{b}\|_{\mathrm{l}^2(\mathbb{Z}^*)}.\end{aligned}$$

Collecting these three equations yields

$$\rho(\mathbf{b}|D\mathbf{b})_{l^2} + (D\mathbf{a}|\mathbf{a})_{l^2} \leq \|h\|_{\mathrm{L}^\infty(\Omega)}\|\mathbf{a}\|_{\mathrm{l}^2(\mathbb{Z}^*)}\|\mathbf{b}\|_{\mathrm{l}^2(\mathbb{Z}^*)}. \tag{15}$$

Since $\pi_2\mathbf{b} = \mathbf{b}$, then $(\mathbf{b}|D\mathbf{b})_{l^2} = \sum_{i=m_2}^{M_2} \lambda_i |b_i|^2 \geq \lambda_{m_2}\|\mathbf{b}\|_{l^2}^2$. Similarly $(\mathbf{a}|D\mathbf{a})_{l^2} \geq \lambda_{m_1}\|\mathbf{a}\|_{l^2}^2$. Moreover, using $\|\mathbf{b}\| = \|\pi_2\mathbf{v}\| = 1$ and

$$\|\mathbf{a}\|_{l^2}^2 = (\pi_1 K \pi_2 \mathbf{v}|\pi_1 K \pi_2 \mathbf{v})_{l^2} = (\pi_2 K \pi_1 K \pi_2 \mathbf{v}|\mathbf{v})_{l^2} = \rho,$$

Equation (15) turns into

$$\rho(\lambda_{m_1} + \lambda_{m_2}) \leq \|h\|_{\mathrm{L}^\infty(\Omega)}\sqrt{\rho},$$

which is exactly (14). □

The following proposition precisely states the error made when computing $\mathbf{u}$ with the limited information of the k first modes.

Proposition 4. For any $\mathbf{a} \in \mathrm{l}^2(\mathbb{Z}^*)$, for any $k \in \mathbb{N}^*$, define $\pi = \Pi_{1,k}$. Define, by Proposition 2, $\hat{\mathbf{u}}f \in l^2(\llbracket 1, k \rrbracket)$ as the unique solution to $\pi K \widehat{\mathbf{u}} = \pi \mathbf{a}$.

Define $\mathbf{u} \in l^2(\llbracket 1, +\infty \rrbracket)$ the only solution to $\Pi_{1,\infty} K \mathbf{u} = \Pi_{1,\infty}\mathbf{a}$, i.e. $\mathbf{u} = \widehat{\mathbf{u}}$ when $k = +\infty$.

There exists a constant $C > 0$ independent of k and $\mathbf{a}$, there exists $k_0 \in \mathbb{N}^*$ such that for all $k \geq k_0$,

$$\|\mathbf{u} - \widehat{\mathbf{u}}\|_{\mathrm{l}^2(\mathbb{Z}^*)} \leq C\|(\Pi_{1,\infty} - \pi)(\mathbf{a} - K\widehat{\mathbf{u}})\|_{\mathrm{l}^2(\mathbb{Z}^*)},$$

$$\|\pi\mathbf{u} - \widehat{\mathbf{u}}\|_{\mathrm{l}^2(\mathbb{Z}^*)} \leq \frac{C}{\lambda_k}\|\mathbf{u} - \widehat{\mathbf{u}}\|_{\mathrm{l}^2(\mathbb{Z}^*)}.$$

Corollary 1. When $\mathbf{a} = \chi^\star P_1^\star f$, if f_{proj} is the L^2 orthogonal projection of f on the space $Vect(c^{1/2}T_1, \ldots, c^{1/2}T_n)$ then

$$\|\mathbf{u} - \widehat{\mathbf{u}}\|_{\mathrm{l}^2(\mathbb{Z}^*)} \leq C\|f - f_{proj}\|_0$$

Indeed when $\mathbf{a} = \chi^\star P_1^\star f$, then $K\mathbf{a} = \mathbf{a}$, $P_1\chi\mathbf{a} = f$ and thanks to Proposition 2 $P_1\chi\widehat{\mathbf{u}} = f_{proj}$. The corollary is then simply proven by

$$\|(\Pi_{1,+\infty} - \pi)(\mathbf{a} - K\widehat{\mathbf{u}})\|_{\mathrm{l}^2(\mathbb{Z}^*)} \leq \|(\mathbf{a} - K\widehat{\mathbf{u}})\|_{\mathrm{l}^2(\mathbb{Z}^*)} = \|K(\mathbf{a} - \widehat{\mathbf{u}})\|_{\mathrm{l}^2(\mathbb{Z}^*)} = \|P_1\chi\mathbf{a} - P_1\chi\widehat{\mathbf{u}}\|_{l^2(\Omega)}.$$

Proof. of Proposition 4. Define $\widetilde{\pi} = \Pi_{k+1,+\infty}$, $\mathbf{d} = \mathbf{u} - \widehat{\mathbf{u}}$, then the equations

$$\pi K\widehat{\mathbf{u}} = \pi\mathbf{a} \text{ and } (\widetilde{\pi} + \pi)K\mathbf{u} = (\pi + \widetilde{\pi})\mathbf{a}$$

yield the following system

$$\left\{ \begin{array}{rcl} (\pi K\pi)(\pi\mathbf{d}) + (\pi K\widetilde{\pi})(\widetilde{\pi}\mathbf{d}) & = & 0 \\ (\widetilde{\pi}K\pi)(\pi\mathbf{d}) + (\widetilde{\pi}K\widetilde{\pi})(\widetilde{\pi}\mathbf{d}) & = & \widetilde{\pi}\mathbf{a} - (\widetilde{\pi}K\pi)\widehat{\mathbf{u}} \end{array} \right.$$

Thanks to Proposition 2, the operators $\pi K\pi$ (resp. $\widetilde{\pi}K\widetilde{\pi}$) are invertible with an inverse bounded from above with a constant independent of k and then

$$\left\{ \begin{array}{rcl} \|\pi\mathbf{d}\|_{l^2} & \leq & C\|(\pi K\widetilde{\pi})\widetilde{\pi}\mathbf{d}\|_{l^2} \\ \|\widetilde{\pi}\mathbf{d}\|_{l^2} & \leq & C\left(\|\widetilde{\pi}(\mathbf{a} - K\pi\widehat{\mathbf{u}})\|_{l^2} + \|(\widetilde{\pi}K\pi)\pi\mathbf{d}\|_{l^2}\right) \end{array} \right.$$

Since $\widetilde{\pi}\pi = 0$, then Proposition 4 applies to $\pi K\widetilde{\pi}$ and $\widetilde{\pi}K\pi$ and

$$\|\pi\mathbf{d}\|_{l^2} \leq \frac{C}{\lambda_k}\|\widetilde{\pi}\mathbf{d}\|_{l^2} \text{ and } (1 - \frac{C}{\lambda_k^2})\|\widetilde{\pi}\mathbf{d}\|_{l^2} \leq C\|\widetilde{\pi}(\mathbf{a} - K\widehat{\mathbf{u}})\|_{l^2}$$

Letting k big enough so that $1 - \frac{C}{\lambda_k^2} > \frac{1}{2}$ and $\frac{1}{\lambda_k} < 1$ there exists another constant, also denoted by C such that

$$\|\mathbf{d}\|_{l^2} = \|\pi\mathbf{d}\|_{l^2} + \|\widetilde{\pi}\mathbf{d}\|_{l^2} \leq C\|\widetilde{\pi}(\mathbf{a} - K\widehat{\mathbf{u}})\|_{l^2} \text{ and } \|\pi\mathbf{d}\|_{l^2} \leq \frac{C}{\lambda_k}\|\mathbf{d}\|_{l^2}.$$

□

4 Solving semi-infinite and finite problems

4.1 The semi-infinite case with L^2 initial conditions

For a given $T_{\mathrm{ini}} \in \mathrm{L}^2(\Omega)$, we are interested in solving in the space $C^0(\mathbb{R}^-, \mathrm{L}^2(\Omega)) \cap C^1(\mathbb{R}^-, \mathrm{H}^{-1}(\Omega))$ the following equation:

$$\left\{ \begin{array}{rcl} c\partial_{zz}T - \mathrm{div}(\sigma\nabla T) - \mathrm{Pe}v\partial_z T & = & 0 \\ T_{|_{z=0}} = T_{\mathrm{ini}} & & \text{and } \lim_{z\to-\infty} T(z) = 0 \end{array} \right. \tag{16}$$

As developped in (7) in Section , T solves the differential equation (16), if and only if

$$\psi(z) = (T(z), \sigma\nabla s) \in C^0(\mathbb{R}^-, \mathcal{R}(\mathcal{A}))$$

verifies $\psi(z) = \sum_{n\in\mathbb{Z}^*} u_n e^{\lambda_n}\psi_n$ with some sequence $\mathbf{u} = (u_n)_{n\in\mathbb{Z}^*} \in \mathrm{l}^2(\mathbb{Z}^*)$ that verifies the boundary conditions in $z = 0$ and $z = -\infty$, that is:

$$T_{\mathrm{ini}} = \sum_{n\in\mathbb{Z}^*} u_n T_n \quad \text{and } u_n = 0 \quad \forall n < 0.$$

As stated in (6) in Section , a similar reduction can be performed if Neumann boundary conditions are enforced in $z = 0$, that is if

$$\partial_z T_{|_{z=0}} = F_{\mathrm{ini}}$$

is given instead of the value of $T_{|_{z=0}}$. In this case the problem would turn into

$$F_{\mathrm{ini}} = \sum_{n\in\mathbb{Z}^*} u_n T_n \quad \text{and } u_n = 0 \quad \forall n < 0 \text{ and } \psi = (\partial_z T, \sigma\nabla T).$$

Moreover, solving this equation for positive z instead of negative z can be done by changing z into $-z$, or equivalently by multiplying v by -1 which does not change the analysis.

Coming back to the original Dirichlet problem, setting $\mathbf{a} = \chi^\star P_1^\star \sqrt{c}T_{\mathrm{ini}} \in \mathrm{l}^2(\mathbb{Z}^*)$, we have $K\mathbf{a} = \mathbf{a}$ and $\mathbf{u}$ is given by Theorem 1 as the unique solution to

$$K\mathbf{u} = \mathbf{a} \text{ and } \mathbf{u} \in l^2([\![1, \infty]\!]).$$

Hence the existence and uniqueness of $T(z)$ in the considered space. In practice, one is able to compute only the k first eigenvectors. We wish to estimate the error made by an approximation of $T(z)$ if only the k first eigenmodes are considered. The following proposition sums up every property proved earlier.

Proposition 5. Suppose that $(\lambda_n, T_n)_{n=1..k}$, the k first positive eigenvalues/eigenvectors of $\mathcal{A}$ have been computed. Define $\widehat{\mathbf{a}} = (\int_\Omega cT_{\mathrm{ini}}T_n)_{n=1..k}$, set $\widehat{K} = (\int_\Omega T_i T_j)_{1\le i,j\le k}$ and find $\widehat{\mathbf{u}} = (\widehat{u}_n)_{n=1..k}$ the unique solution to

$$\widehat{K}\widehat{\mathbf{u}} = \widehat{\mathbf{a}}\ . \tag{17}$$

Define

$$\widehat{T}(z) = \sum_{n=1}^{k} c^{-1/2}\widehat{u}_n e^{\lambda_n z} T_n.$$

If $T(z)$ denotes the unique solution to problem (16), then for all $z \le 0$ we have

$$\|T(z) - \widehat{T}(z)\|_0 \le C\left(\frac{e^{\lambda_1 z}}{\lambda_k} + e^{\lambda_k z}\right)\|\sqrt{c}T_{\mathrm{ini}} - \sqrt{c}T_{\mathrm{proj}}\|_0,$$

where $\sqrt{c}T_{\mathrm{proj}}$ is the L^2-orthogonal projection of $\sqrt{c}T_{\mathrm{ini}}$ on the space spanned by $Vect(\sqrt{c}T_n)_{n=1..k}$.

We remark that since we are interested in the semi-cylinder defined by $z \leq 0$, the inequality gets better as z goes to $-\infty$ or as k grows.

Proof. Set $\pi = \Pi_{1,k}$, if $\mathbf{a} = \chi^\star P_1^\star \sqrt{c} T_{\text{ini}}$ then the solution of (16) is given by

$$(T(z), \sigma \nabla s(z)) = \sum_{n \in \mathbb{Z}^*} c^{-1/2} u_n e^{\lambda_n z} \psi_n,$$

where $\mathbf{u} = (\mathbf{u}_n)_{n \in \mathbb{Z}^*}$ is given by $K\mathbf{u} = \mathbf{a}$ and $\mathbf{u} \in l^2([\![1, +\infty[\![)$, see Proposition 2.

Extending by zero $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{a}}$ in $\mathrm{l}^2(\mathbb{Z}^*)$ then $\widehat{\mathbf{a}} = \pi\mathbf{a}$, $\widehat{K} = \pi K \pi$ and $\widehat{\mathbf{u}}$ verifies

$$\pi K \pi \widehat{\mathbf{u}} = \pi \mathbf{a} \text{ and } \widehat{\mathbf{u}} \in l^2([\![1, k]\!]).$$

Hence, $\widehat{\mathbf{u}}$ is unique and determined by Proposition 2. Moreover, Corollary 1 states that

$$\|\mathbf{u} - \widehat{\mathbf{u}}\|_{l^2} \leq C \|\sqrt{c} T_{\text{ini}} - \sqrt{c} T_{\text{proj}}\|_0.$$

$$\begin{aligned} \|T(z) - \widehat{T}(z)\|_0 &\leq C \sum_{n=1}^{k} |\mathbf{u}_n - \widehat{\mathbf{u}}_n|^2 e^{2\lambda_n z} + C \sum_{n>k} |\mathbf{u}_n - \widehat{\mathbf{u}}_n|^2 e^{2\lambda_n z} \\ &\leq C \|\pi\mathbf{u} - \widehat{\mathbf{u}}\|_{l^2}^2 e^{2\lambda_1 z} + C \|\mathbf{u} - \widehat{\mathbf{u}}\|_{l^2}^2 e^{2\lambda_k z}. \end{aligned}$$

The conclusion follows from Proposition 4 since $\|\pi\mathbf{u} - \widehat{\mathbf{u}}\|_{l^2} \leq \dfrac{C}{\lambda_k} \|\mathbf{u} - \widehat{\mathbf{u}}\|_{l^2}$. □

4.2 The finite case with Dirichlet condition on both ends

For given $L > 0$, $T_0, T_L \in \mathrm{L}^2(\Omega)$, we are interested in finding $T \in C^1([0, L], \mathrm{L}^2(\Omega)) \cap C^0([0, L], \mathrm{H}_0^1(\Omega))$, solution to the following equation

$$\left\{ \begin{array}{rll} c\partial_{zz} T + \operatorname{div}(\sigma \nabla T) - \mathrm{Pe} v \partial_z T &= 0 & \text{in } [0, L] \times \Omega \\ T_{|_{z=0}} = T_0 & \text{and} & T_{|_{z=L}} = T_L \end{array} \right. , \tag{18}$$

In this problem, two boundary conditions are imposed, one on each end of the finite cylinder. The mathematical proof of existence of solution is straightforward since this problem is the one of a three-dimensional Laplacian on $\Omega \times [0, L]$ with a transport term and Dirichlet boundary condition. We are looking here for an effective way to compute the solution of this problem by performing a reduction to a problem in two dimensions.

The first idea is to use upstream modes (negative eigenvalues) for the left-most boundary condition ($z = 0$), and to use downstream modes (positive eigenvalues) for the right-most boudary condition ($z = L$). Some corrections must be added in order to take into account the influence of each boundary on the other.

Proposition 6. Consider T_0 and T_L in $\mathrm{L}^2(\Omega)$. Then there exists a unique $(a_n)_{n\in\mathbb{Z}^*} \in \mathrm{l}^2(\mathbb{Z}^*)$ such that

$$\sum_{n<0} a_n T_n + \sum_{n>0} a_n e^{-L\lambda_n} T_n = T_0 \tag{19}$$

and

$$\sum_{n<0} a_n e^{L\lambda_n} T_n + \sum_{n>0} a_n T_n = T_L. \tag{20}$$

The solution of Problem (18) is then given by

$$T(z) = \sum_{n<0} a_n e^{\lambda_n z} T_n + \sum_{n>0} a_n e^{\lambda_n (z-L)} T_n \qquad \text{for } 0 \le z \le L.$$

Proof. For a given sequence $\mathbf{a} \in \mathrm{l}^2(\mathbb{Z}^*)$, denote $\mathbf{a}^+ = (a_n)_{n>0}$ and $\mathbf{a}^- = (a_n)_{n<0}$. We also introduce the operators

$$\begin{array}{rccc} U^\pm : & l^2(\mathbb{Z}^{*\pm}) & \longrightarrow & \mathrm{L}^2(\Omega) \\ & \mathbf{a}^\pm = (a_n)_n & \longmapsto & \displaystyle\sum_{\pm n>0} a_n T_n \end{array} \quad \text{and} \quad \begin{array}{rccc} C^\pm : & l^2(\mathbb{Z}^{*\pm}) & \longrightarrow & l^2(\mathbb{Z}^{*\pm}) \\ & \mathbf{a}^\pm = (a_n) & \longmapsto & (a_n e^{\mp L\lambda_n})_{\pm n>0}. \end{array}$$

Theorem 1 implies that U^+ and U^- are one-to-one. Then the two equations (19) and (20) read

$$\begin{pmatrix} U^- & U^+C^+ \\ U^-C^- & U^+ \end{pmatrix} \begin{pmatrix} \mathbf{a}^- \\ \mathbf{a}^+ \end{pmatrix} = \begin{pmatrix} T_0 \\ T_L \end{pmatrix}. \tag{21}$$

It remains to prove that the operator W from $l^2(\mathbb{Z}^{*-}) \times l^2(\mathbb{Z}^{*+})$ to $\mathrm{L}^2(\Omega)^2$ defined by

$$W = \begin{pmatrix} U^- & U^+C^+ \\ U^-C^- & U^+ \end{pmatrix} = \begin{pmatrix} Id & U^+C^+(U^+)^{-1} \\ U^-C^-(U^-)^{-1} & Id \end{pmatrix} \begin{pmatrix} U^- & 0 \\ 0 & U^+ \end{pmatrix}$$

is invertible. The endomorphism W_0 of $(\mathrm{L}^2(\Omega))^2$ defined by

$$W_0 = \begin{pmatrix} Id & M_+ \\ M_- & Id \end{pmatrix} \quad \text{with } M_\pm = U^\pm C^\pm (U^\pm)^{-1}$$

is invertible if and only if $Id - M_+M_-$ and $Id - M_-M_+$ are invertible which is the case since the operator $M_\pm$ is diagonal in the basis $(T_n)_{\pm n>0}$ with largest eigenvalue $e^{\mp L\lambda_{\pm 1}} < 1$. As a conclusion, the operator W is invertible, hence the equation (21) admits a unique solution $(\mathbf{a}^-, \mathbf{a}^+)$. □

Remark 2. A physical interpretation of the operator $M_\pm$ is the following. The operator M_+ acts on an element of $\mathrm{L}^2(\Omega)$ by decomposing this element on the downstream modes, and damps the modes with a damping factor corresponding to a length L. The operator M_- has the same interpretation except that upstream modes are concerned. These operators model the influence of one boundary condition on the other boundary of the cylinder.

Equation (21) can be rewritten

$$\begin{pmatrix} Id & M_+ \\ M_- & Id \end{pmatrix} \begin{pmatrix} U^+\mathbf{a}^+ \\ U^-\mathbf{a}^- \end{pmatrix} = \begin{pmatrix} T_0 \\ T_L \end{pmatrix}. \tag{22}$$

Such equation is of type

$$(Id + M_r)\mathbf{x} = \mathbf{y} \tag{23}$$

where $M_r = \begin{pmatrix} 0 & M_+ \\ M_- & 0 \end{pmatrix}$ is a reflection operator associated with the influence of the boundary conditions on the mode's amplitude. In our case the spectral radius of M_r is smaller than 1, and (23) can be solved using a power series:

$$\mathbf{x} = (Id + M_r)^{-1}\mathbf{y} = \mathbf{y} - M_r\mathbf{y} + M_r^2\mathbf{y} - M_r^3\mathbf{y} + ...$$

As stated above, this amounts to write that (in a first approximation) the solution is $\mathbf{x} \approx \mathbf{y}$: $\mathbf{x}$ is obtained by decomposing the boundary condition at $z = 0$ along the downstream modes, and the boundary condition at $z = L$ along the upstream modes. The next term in the power series reads $\mathbf{x} \approx \mathbf{y} - M_r\mathbf{y}$, this takes into account the corrective terms coming from the influence of each boundary condition on the other boundary of the cylinder. The higher order term $M_r^2\mathbf{y}$ takes into account the correction of the corrective terms and so on. In this sense our solution is a multi-reflection method, since each step provides an incremental reflection of the boundary influence. Nevertheless, as opposed to the image methods used for the computation of the Green functions in finite domains for which the convergence is algebraic, and thus rather poor, the successive terms in the sequence are exponentially decaying, providing an exponential convergence of our multi-reflection finite domain operator.

5 Numerical results

We present in this section more details on the implementation of the method, and illustrate the results in different configurations.

5.1 Implementation

The main obstacle to the numerical resolution of the eigenproblem

$$\mathcal{A}\psi = \lambda\psi \tag{24}$$

is the existence of the kernel of $\mathcal{A}$ which is infinite dimensional, since this prohibits applying effective numerical methods for the eigenvalues computation. The resolution can become effective when one restricts to a subspace of $\mathcal{R}(\mathcal{A})$. We have seen in section 1 that the space $\mathcal{R}(\mathcal{A})$ is given by

$$\mathcal{R}(\mathcal{A}) = \{(f, \sigma\nabla s) \text{ with } (f, s) \in \mathrm{L}^2(\Omega) \times \mathrm{H}_0^1(\Omega)\}.$$

We introduce the space $\mathcal{G}$ as

$$\mathcal{G} = \{(f, \sigma\nabla s) \text{ with } (f,s) \in \mathrm{H}^1_0(\Omega) \times \mathrm{H}^1_0(\Omega)\},$$

endowed with the norm

$$\|(f, \sigma\nabla s)\|_{\mathcal{G}} = \|f\|_{\mathrm{H}^1_0(\Omega)} + \|s\|_{\mathrm{H}^1_0(\Omega)}.$$

It is clear that $\mathcal{G}$ is a dense subset of $\mathcal{R}(\mathcal{A})$ for the $\mathcal{H}$ norm, that $D(A) \cap \mathcal{R}(\mathcal{A})$ is a dense subset of $\mathcal{G}$ for the $\mathcal{G}$ norm and that $\mathcal{G}$ belongs to the domain of $\mathcal{A}^{1/2}$ in the sense that

$$(\mathcal{A}\psi|\psi)_{\mathcal{H}} = \int_\Omega chT^2 + 2\sigma\nabla s \cdot \nabla T \leq C\|\psi\|^2_{\mathcal{G}} \quad \forall \psi = (T, \sigma\nabla s) \in D(A) \cap \mathcal{G}.$$

Solving the eigenproblem of finding $\psi_n \in \mathcal{D}(\mathcal{A}) \cap \mathcal{R}(\mathcal{A})$ such that for all $\psi \in \mathcal{R}(\mathcal{A})$

$$(\mathcal{A}\psi_n|\psi)_{\mathcal{H}} = \lambda_n(\psi_n, \psi)_{\mathcal{H}},$$

amounts to solving it for all $\psi \in \mathcal{G}$ (by density of $\mathcal{G}$ in $\mathcal{R}(\mathcal{A})$) and to seek $\psi_n \in \mathcal{G}$ if one defines, for all $\psi_i = (T_i, \sigma\nabla s_i) \in \mathcal{G}$

$$(\mathcal{A}\psi_1|\psi_2)_{\mathcal{H}} \;=\; \int_\Omega chT_1\overline{T_2} + \sigma\nabla s_1 \cdot \overline{\nabla T_2} + \sigma\overline{\nabla s_2} \cdot \nabla T_1. \tag{25}$$

We recall that the $\mathcal{H}$ scalar product reads for all $\psi_i = (T_i, \sigma\nabla s_i) \in \mathcal{G}$:

$$(\psi_1, \psi_2)_{\mathcal{H}} = \int_\Omega cT_1\overline{T_2} + \sigma\nabla s_1 \cdot \overline{\nabla s_2}. \tag{26}$$

If one approximates $\mathrm{H}^1_0(\Omega)$ by -say- P^1 finite element spaces, equation (26) allows to obtain the mass matrix M, and Equation (25) allows to assemble the stiffness matrix A of the eigenproblem

$$\text{Find } X, \lambda \text{ such that } AX = \lambda MX,$$

which is the discrete version of the eigenproblem (24), set on the orthogonal of the kernel of $\mathcal{A}$.

5.2 Solving the eigenproblem

The eigenproblem $\mathcal{A}\psi = \lambda\psi$, reduced to the generalized eigenvalue problem

$$AX = \lambda MX,$$

is solved using Lanczos method [5]. This algorithm provides the n eigenmodes whose associated eigenvalues are closest from zero (exepted 0 since we work in the orthogonal of the kernel). We denote by N' the number of eigenmodes associated to negative eigenvalues, and by N the number of eigenmodes associated to positive eigenvalues. Due to non-symmetry reasons (because of the

convective term) it is very likely that $N' \neq N$. One can of course restrict the number of eigenmodes to $\min(N', N)$ but this was not considered here.

Let $T_{\text{ini}} \in \mathrm{L}^2(\Omega)$. Consider $k \in \mathbb{Z}^*$. We denote T_{proj} the approximation of T_{ini} by the first k upstream modes if $k > 0$, and by the first $|k|$ downstream modes if $k < 0$. In other words, T_{proj} is the projection of T_{ini} on $Vect(T_1 \dots T_k)$ when $k > 0$ and $Vect(T_{-1} \dots T_{-k})$ when $k < 0$. Using the notations of Proposition 4, we recall that T_{proj} (for example in the case $k > 0$) is computed as

$$T_{\text{proj}} = \sum_{i=1}^{k} u_i T_i \text{ with } \pi K \pi \widehat{\mathbf{u}} = \mathbf{a} \text{ and } a_i = \int_\Omega T_{\text{ini}} T_i.$$

For a given value of k, the relative error is defined by

$$\frac{\|T_{\text{ini}} - T_{\text{proj}}\|_0}{\|T_{\text{ini}}\|_0}. \tag{27}$$

When N' upstream eigenmodes and N downstream eigenmodes are available, this allows to solve the problem in a cylinder of finite length. The computation of the eigenmodes allows to obtain an approximation of the operator W that appears at the left-hand side in Equation (21). The quantities $\mathbf{a}^+$ and $\mathbf{a}^-$ are then computed by solving Equation (21) in the least squares sense.

5.3 An axisymmetric case

We first consider an axisymmetric case. It allows a comparison with existing methods. Reference eigenvalues are computed using the "λ-analicity" method, as presented in [19] in a simpler case. This method provides an implicit analytical definition of the eigenvalues that makes possible their computation up to a given accuracy. The first eigenvalues were computed with this method with a precision of 10^{-10}, providing the reference eigenvalues, named 'analytical eigenvalues' in the sequel.

The domain Ω is the unit circle. The Peclet number is set to 10 and the velocity is supported in the disc B centered at the origin and of radius $r_0 = 1/2$. The velocity profile v is parabolic, culminating at the origin with the value 2, more precisely:

$$v(x, y) = 2(1 - \frac{x^2 + y^2}{r_0^2}) \text{ on } B.$$

The simulations were performed using Getfem [1] and Matlab. The problem was discretized using P1 finite elements, on different meshes containing respectively 164 points (`mesh 0`), 619 points (`mesh 1`), 2405 points (`mesh 2`) and 9481 points (`mesh 3`).

We computed the 50 eigenvalues that are closest to zero (multiplicity counted). These eigenvalues were compared with the analytical eigenvalues corresponding to axisymmetric eigenmodes. These results are presented in Figure 2. Note that the distribution of the eigenvalues is not symmetric with respect to 0, due to the convective term. In this case there are 30 downstream

modes, and 20 upstream modes. The relative error on the first upstream eigenvalue compared to the analytical eigenvalue, as a function of the mesh size is presented in Figure 3.

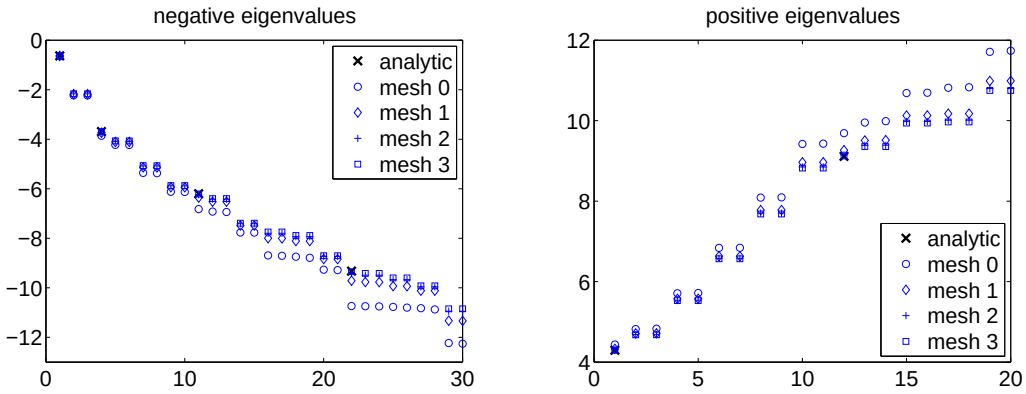

Figure 2: Left: the first eigenvalues for the downstream modes; right: the first eigenvalues for the upstream modes. The eigenvalues obtained for different discretizations are compared to the analytical eigenvalues (only for axisymmetric modes, indicated in black)

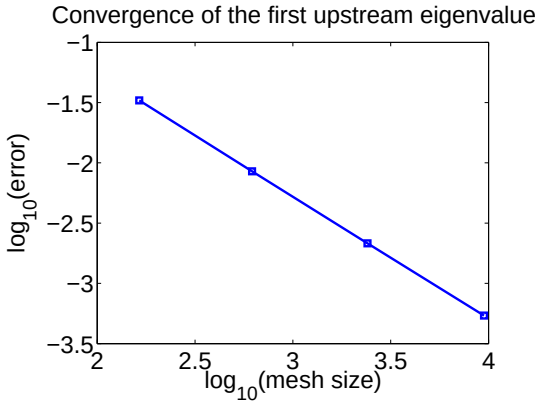

Figure 3: Numerical error for the first upstream eigenvalue, as a function of the mesh size (log scale).

As an illustration of Theorem 1, we decompose an element $T_{\text{ini}} \in \mathrm{H}^1_0(\Omega)$ along the downstream modes, and along the upstream modes. The field T_{ini} is $T_{\text{ini}}(x,y) = (1-x^2-y^2)(1+5x^3+xy)$. The total number of eigenvalues is 300. This computation uses the finest mesh `mesh 3`. We indicate in Figure 4 the relative error when the first k modes are taken into account, defined by Equation (27).

As another illustration of Theorem 1, we decompose another element $T_{\text{ini}} \in \mathrm{L}^2(\Omega)$ along the downstream modes, and along the upstream modes. The field T_{ini} is $T_{\text{ini}}(x,y) = 1$ and the convergence of the projections when an increasing number of modes taken into account is shown in Figure 5. Note that the

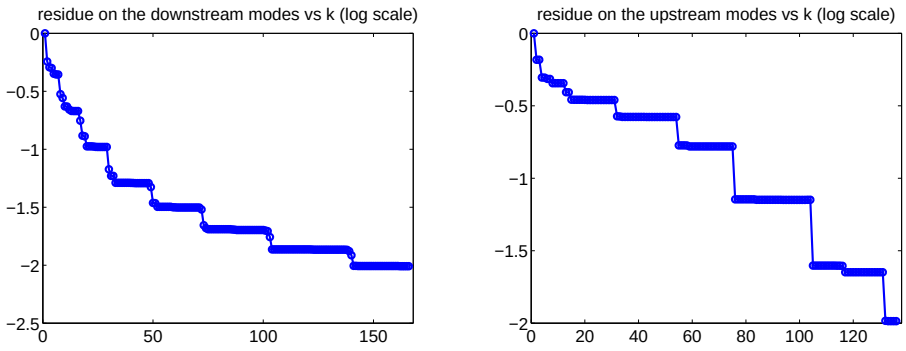

Figure 4: The log_{10} of the relative error of the projection of a field $T_{\text{ini}} \in \mathrm{H}_0^1(\Omega)$ on the first k eigenmodes plotted as a function of k for the downstream modes (left); the log_{10} of the relative error as a function of k for the upstream modes (right).

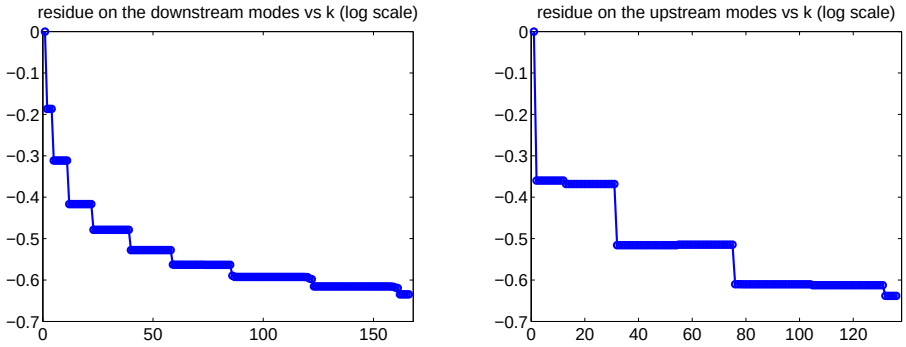

Figure 5: The log_{10} of the relative error of the projection of a field $T_{\text{ini}} \in \mathrm{L}^2(\Omega)$ on the first k eigenmodes plotted as a function of k for the downstream modes (left); the log_{10} of the relative error as a function of k for the upstream modes (right).

convergence is slower here than in the previous case (Figure 4), since in the previous case, the element T_{ini} belongs to $\mathrm{H}_0^1(\Omega)$ and in the present case to $\mathrm{L}^2(\Omega)$ only. We recall T_{ini} is projected on the space of eigenmodes which all belong to $\mathrm{H}_0^1(\Omega)$ and even if it is possible to approximate elements of $\mathrm{L}^2(\Omega)$ by elements of $\mathrm{H}_0^1(\Omega)$ in the L^2 norm, phenomenom of slow convergence (similar the well known Gibb's effect) will occur.

5.4 A non-axisymmetric case

In order to illustrate the capabilities of our approach, we present an illustration in a non-axisymmetric case.

The domain Ω is the unit circle. The Peclet number is set to 10 and the velocity is contained in the disc B centered at the point $(x_0, y_0) = (0.3, 0.2)$ and of radius $r_0 = 1/2$. The velocity profile v is parabolic in B culminating at

(x_0, y_0) with the value 2 (see Figure 6):

$$v(x,y) = 2(1 - \frac{(x-x_0)^2 + (y-y_0)^2}{r_0^2}) \text{ in } B$$

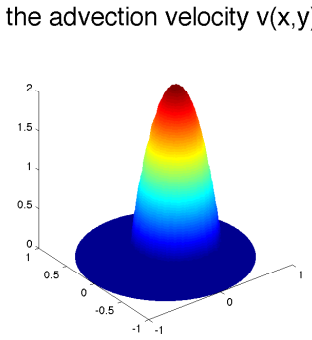

Figure 6: Velocity profile.

The problem was discretized on a mesh containing 9517 vertices. We computed the 50 eigenmodes that are closer to zero (multiplicity counted), see Figure 7. In this case there are 31 downstream modes, and 19 upstream modes.

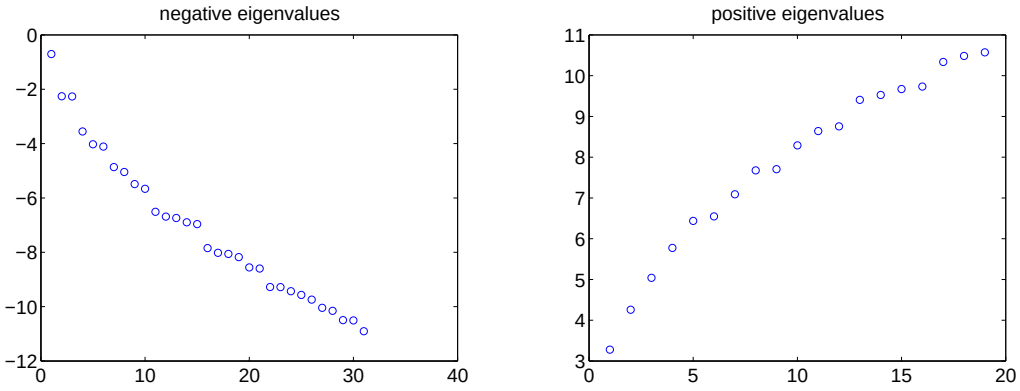

Figure 7: Left: the first eigenvalues for the downstream modes; right: the first eigenvalues for the upstream modes.

We present in Figures 8 and 9 the first downstream and upstream eigenmodes.

We document also the results of section 3 by showing the matrix $\Pi_{-N',N} K \Pi_{-N',N}$ for different values of the Peclet number, see Figure 10.

5.5 A finite cylinder

The results of section 4.2 are documented here. The domain B, the Peclet number and the velocity profile v are the same as in section 5.4. We address

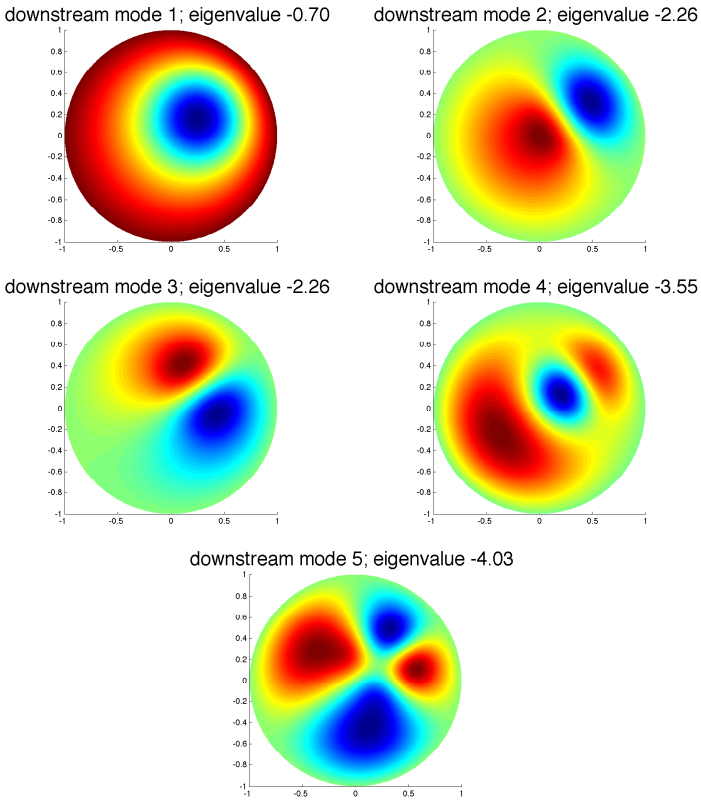

Figure 8: The first downstream eigenmodes.

the 3-dimensional problem in a cylinder of length L. Two boundary conditions are imposed on the extremities of this cylinder:

$$T_{|_{z=0}} = T_0 \qquad \text{and} \qquad T_{|_{z=L}} = T_L,$$

where

$$T_0(x,y) = \mathbf{1}_B(x,y) \qquad \textit{and} \qquad T_L(x,y) = 1 - x^2 - y^2.$$

This problem was discretized on a mesh comprising 9517 vertices. The 1000 eigenvalues closest to 0 are computed (527 downstream modes and 473 upstream modes). The matrix W defined in section 4.2 was assembled, the sequences $\mathbf{a}^+$ and $\mathbf{a}^-$ were computed, and the value of $T(z)$ at different sections, corresponding to different values of z are illustrated in Figures 11 and 12 for $L = 1$ and $L = 5$ respectively. Note that since the incoming condition T_0 is not in $\mathrm{H}_0^1(\Omega)$, the initial condition is poorly approximated (oscillations are visible). Note also that the downstream modes are damped slower than the upstream modes. The largest downstream eigenvalue is $\lambda_{-1} \approx -0.704$ which gives a characteristic length of $\ln(2)/|\lambda_{-1}| \approx 0.98$, while the smallest upstream eigenvalue is $\lambda_1 \approx 3.28$ which gives a characteristic length of $\ln(2)/\lambda_1 \approx 0.21$.

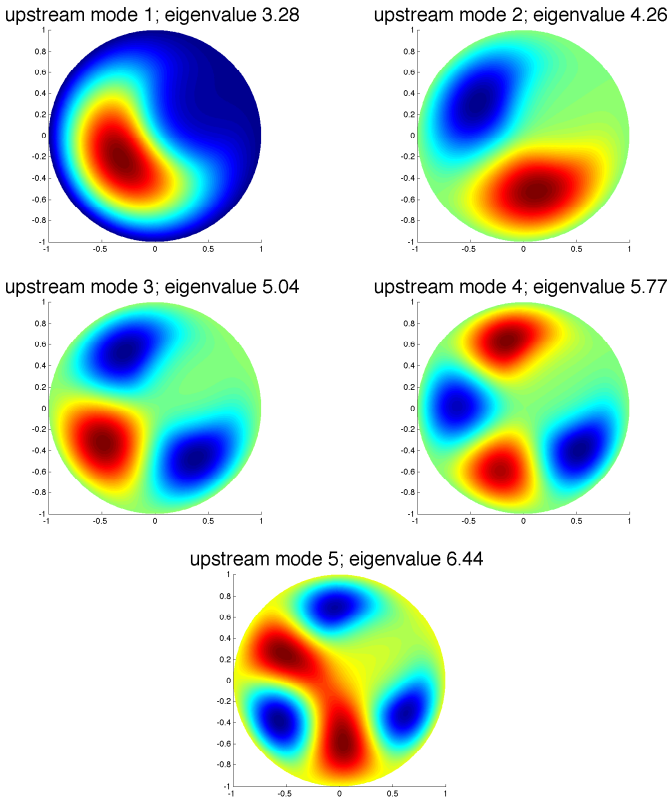

Figure 9: The first upstream eigenmodes.

Conclusion

It has been shown that the decomposition on the upstream (or downstream) modes is not only mathematically possible but also numerically feasible. Indeed, thanks to the bounds of Proposition 4, standard error analysis, as the one of Proposition 5, may be performed. Such analysis leads to effective algorithms that improve the state of the art on the generalized Graetz problem by many ways. First, non axisymmetrical geometries are allowed. Second, semi-infinite ducts and bounded ducts geometries are studied. Third, effective error analysis is available. We presented numerical examples that showcase the power of this method.

All these improvements pave the way to numerous applications, as for example, optimization of the velocity v in order to maximize (or minimize) heat transfer under constraints (for instance viscosity constraints if the velocity is the solution of a Stoke's problem). Nevertheless, some expected results still lack. For instance, the theory handles well L^2 bounds when L^2 initial data is given. But there isn't, as of today, any direct way to show H^1_0 bounds when H^1_0 initial data is given. An other improvement would be to understand if the information given by the eigenvectors with a positive eigenvalue is of any help when trying to decompose on the downstream modes. Indeed the algorithm we propose simply dumps this information in order to concentrate only on the

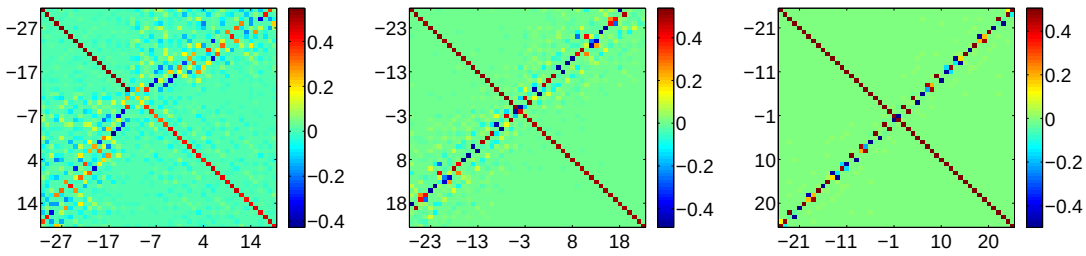

Figure 10: The matrix $\Pi_{-N',N} K \Pi_{-N',N}$. From left to right: Peclet = 10 (31 downstream and 19 upstream modes); Peclet = 1 (27 downstream and 23 upstream modes); Peclet = 0.1 (25 downstream and 25 upstream modes)

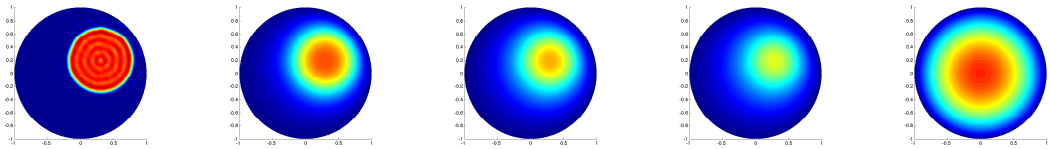

Figure 11: The finite cylinder with length $L = 1$. From left to right: the value of $T(z)$ for $z = 0, 0.25L, 0.5L, 0.75L, L$.

one given by the negative eigenvalues. It is also not clear how to proceed when Dirichlet and Neumann boundary conditions are mixed at the entrance and the exit. For instance, extending Graetz modes expansions for semi-infinite ducts when Ω is parted into two subsets Ω_D and Ω_N where respectively Dirichlet and Neumann boundary conditions are imposed is still an open question.

Such problems and extensions are currently under investigation.

References

[1] *http://download.gna.org/getfem/html/homepage/index.html*.

[2] V. A. Aleksashenko. Conjugate stationary problem of heat transfer with a moving fluid in a semi-infinite tube allowing for viscous dissipation. *J. Eng. Physics and Thermophysics*, 14(1):55–58, 1968.

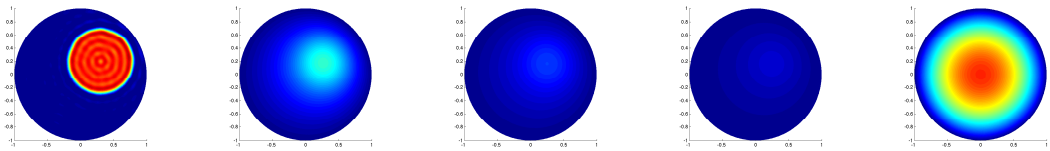

Figure 12: The finite cylinder with length $L = 5$. From left to right: the value of $T(z)$ for $z = 0, 0.25L, 0.5L, 0.75L, L$.

[3] R. F. Barron, X. Wang, R. O. Warrington, and T. Ameel. The extended Graetz problem with piecewise constant wall heat flux for pipe and channel flows. *International Communications in Heat and Mass Transfer*, 23(4):563–574, 1996.

[4] N. M. Belyaev, O. L. Kordyuk, and A. A. Ryadno. Conjugate problem of steady heat exchange in the laminar flow of an incompressible fluid in a flat channel. *J. Eng. Physics and Thermophysics*, 30(3):339–344, 1976.

[5] J. Demmel. Applied Numerical Linear Algebra. 1997.

[6] M. Ebadian and H. Zhang. An exact solution of extended Graetz problem with axial heat conduction. *Int. J. Heat Mass Transfer*, 82:1709 1717, 1989.

[7] L. Graetz. Uber die Wärmeleitungsfähigkeit von Flüssigkeiten. *Annalen der Physik*, 261,(7):337–357, 1885.

[8] C. Ho, H. Yeh, and W. Yang. Double-pass Flow Heat Transfer In A Circular Conduit By Inserting A Concentric Tube For Improved Performance. *Chem. Eng. Comm.*, 192(2):237–255, 2005.

[9] C.-D. Ho, H.-M. Yeh, and W.-Y. Yang. Improvement in performance on laminar counterflow concentric circular heat exchangers with external refluxes. *Int. J. Heat and Mass Transfer*, 45(17):3559–3569, 2002.

[10] J. Lahjomri, A. Oubarra, and A. Alemany. Heat transfer by laminar Hartmann flow in thermal entrance region with a step change in wall temperatures: the Graetz problem extended. *Int. J. Heat Mass Transfer*, 45(5):1127–1148, 2002.

[11] A. Luikov, V. Aleksashenko, and A. Aleksashenko. Analytical methods of solution of conjugated problems in convective heat transfer. *Int. J. Heat Mass Transfer*, 14:1047–1056, 1971.

[12] M. Michelsena and J. Villadsena. The Graetz problem with axial heat conduction. *Int. J. Heat Mass Transfer*, 17(11):1391–1402, 1974.

[13] R. Myonga, D. Lockerbyc, and J. Reesea. The effect of gaseous slip on microscale heat transfer: An extended Graetz problem. *International Communications in Heat and Mass Transfer*, 49(15-16):2502–2513, 2006.

[14] E. Papoutsakis, D. Ramkrishna, and H. C. Lim. The extended graetz problem with diriclet wall boundary conditions. *Appl. Sci. Res.*, 36:13–34, 1980.

[15] E. Papoutsakis, D. Ramkrishna, and H. C. Lim. The extended graetz problem with prescribed wall flux. *AIChE J.*, 26:779–787, 1980.

[16] E. Papoutsakis, D. Ramkrishna, and H.-C. Lim. Conjugated graetz problems. pt.1: general formalism and a class of solid-fluid problems. *Chemical Engineering Science*, 36(8):1381–1391, 1981.

[17] E. Papoutsakis, D. Ramkrishna, and H.-C. Lim. Conjugated Graetz problems. Pt.2: Fluid-Fluid problem. *Chemical Engineering Science*, 36(8):1393–1399, 1981.

[18] C. Pierre and F. Plouraboué. Numerical analysis of a new mixed-formulation for eigenvalue convection-diffusion problems. *SIAM J. Applied Maths*, 70(3):658–676, 2009.

[19] C. Pierre, F. Plouraboué, and M. Quintard. Convergence of the generalized volume averaging method on a convection-diffusion problem: a spectral perspective. *SIAM J. Appl. Math.*, 66(1):122–152, 2005.

[20] R. Shah and A. London. *Laminar Flow Forced Convection in Ducts*. Academic Press Inc., New York, 1978.

[21] B. Weigand, M. Kanzamarb, and H. Beerc. The extended Graetz problem with piecewise constant wall heat flux for pipe and channel flows. *Int. J. Heat Mass Transfer*, 44(20):3941–3952, 2001.