A Posteriori Bounds for Linear Functional Outputs of Hyperbolic Partial Differential Equations

by

Hubert J.B. Vailong

Elève Diplômé de l'Ecole Polytechnique (Promotion X91)
Ingénieur Diplômé de l'Ecole Nationale Supérieure de l'Aéronautique et de l'Espace (1996)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Master of Science in Aeronautics and Astronautics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1997

Author . . . .
Department of Aeronautics and Astronautics
January 31, 1997

Certified by . . . .
Jaime Peraire
Professor, Fluid Dynamics Research Laboratory
Thesis Supervisor

Accepted by . . . .
Jaime Peraire
Chairman, Department Graduate Committee

Abstract

One of the major difficulties faced in the numerical resolution of the equations of physics is to decide on the right balance between computational cost and solutions accuracy, and to determine how solutions errors affect some given "outputs of interest".

This thesis presents a technique to generate upper and lower bounds for outputs of hyperbolic partial differential equations. The outputs of interest considered are linear functionals of the solutions of the equations. The method is based on the construction of an "augmented" Lagrangian, which includes a formulation of the output as a quadratic form to be minimized and the equilibrium equations as a constraint. The corresponding Lagrange multiplier, or adjoint μ, is determined by solving a problem involving the adjoint of the operator in the original equations. The bounds are then derived from the underlying unconstrained max-min problem. A predictor is also evaluated as the average value of the bounds. Because the resolution of the max-min problem implies the resolution of the original discrete equations, the adjoint on a fine grid is approximated by a hierarchical procedure that consists of the resolution of the problem on a coarser grid followed by an interpolation on the fine grid. The bounds derived from this approximation are then optimized by the choice of natural boundary conditions for the adjoint and by selecting he value of a stabilization parameter κ.

The Hierarchical Bounds Method is illustrated on three cases. The first one is the convection-diffusion equation, where the bounds obtained are very sharp. The second one is a purely convective problem, discretized using a Taylor-Galerkin approach. The third case is based on the Euler equations for a nozzle flow, which can be reduced to a single nonlinear scalar continuous equation. The resulting discrete nonlinear system of equations is obtained by a Taylor-Galerkin method and is solved by the Newton-Raphson method. The problem is then linearized about the computed solution to obtain a linear system similar to the previous cases and produce the bounds.

In a last chapter, the Domain Decomposition is introduced. The domain is decomposed into K subdomains and the problem is solved separately on each of them before continuity at the boundaries is imposed, allowing the computation of the bounds to be parallelized. Because the cost of sparse matrix inversion is of order $O(N^3)$, Domain Decomposition becomes very useful for two-dimensional problems, where the overall cost is divided by K^2.

Acknowledgements

At the end of this year spent at the Massachusetts Institute of Technology, and before starting with the hard stuff, I would really like to thank

- my thesis supervisor, Professor Jaime Z. Peraire, for his help, his constant encouragements and his advice ; it would be nice if he was as successful in raising his "little monster" as in educating his students ;
- Professor Anthony T. Patera, of the Department of Mechanical Engineering, for providing me with the topic of this research and for his always useful suggestions ;
- Professor Jin Au Kong, of the Department of Electrical Engineering and Computer Science, thanks to whom I am here today ; had it not been for his advice, I may not have been admitted to MIT in the first place ; thanks for the jokes too !
- Marius Paraschivoiu for his results : that was very helpful to check mine ! Thanks for the time spent with me and all the advice ;
- Tolulope Okusanya for his dotfiles (actually, I still wonder if taking *his* was that smart...), his help with computers (God knows I needed help !), his unmistakable laugh and his constant good mood ; thanks also for the funniest game of the year : the first person who pronounces his name correctly wins a FREE glass of water at the pub of his/her choice ;
- Angie Kelic for fixing Tolu's dotfiles on my account, rebooting endeavour from time to time and explaining to me the subtleties of UNIX, US Presidential Elections, and President Clinton's executive decisions ;
- Ed "The Man" Piekos for his ubiquitous presence (except when he was not there), for his help with LaTeX, with technical writing and with printers, and for wearing his

vegetarian T-shirt (I love the "Veni, Vidi, Vegie") ;

- Imran Haq for the FDRL T-shirts and the thesis proof-reading (and for his patience with my stupid English questions !) ;
- all the students of the Fluid Dynamics Research Laboratory and of the Space Systems Laboratory for welcoming me among them and forcing me to leave the lab from time to time : Dr. Jon Ahn, Ali Merchant, Vadim Khayms, Greg Giffin, Greg Yashko, Chris McLain, Josh Elliot, Karen Willcox, Ray Sedwick, Graeme Shaw (the chariot racing and "hounds hunt humans" event promoter), Mike Fife, Folusho Oyerokun and Bilal Mughal ;
- John Harper for the free Sprite (as Alanis Morissette might have sung in *Ironic*, "it's a free Sprite when you wanted to pay" !). Congrats for the Quals and good luck for the next four years, dude !
- all the people in the MIT European Club for making this year a great year.

Contents

List of Figures

Introduction

When solving an engineering problem, one has to evaluate some outputs of interest, or design variables that determine the performance of the design. These outputs are often functionals of fields that are in turn solutions of ordinary or partial differential equations. These functionals are often linear, or more generally convex, but they can be nonlinear as well. An important part of the process followed by the engineer therefore consists of modeling a given problem, i.e. of translating it into a mathematical model, which generally yields a set of partial differential equations such as the Euler, Navier-Stokes or Maxwell equations. The solutions of these equations are the fluid velocity and pressure fields or the electromagnetic field. From the engineering point of view, these solutions are not as critical as the outputs derived from them, like the lift of a wing, the drag of a body or the radar cross section of an aircraft. The fields are nonetheless worthier than just intermediary steps to these outputs of interest. Indeed, if the engineer finds that the outputs do not satisfy the design constraints, he can go back to the fields to find information on the reasons for the degradation of the performance (shocks, turbulent transition, scattering...).

Because the analytical solution to the equations of physical problems is not always (and one could almost say rarely) available, the fields have to be computed numerically. The resulting discretization of the equations leads to a necessary trade-off between cost and accuracy : on one hand, a very fine discretization step usually yields very accurate solutions, with a very high computational cost ; on the other hand, the computation of a solution on a coarse grid may be much cheaper, but it is always at the cost of accuracy.

Several approaches have been suggested to reduce the cost of computations while keeping an acceptable accuracy. The simplest one consists of using non-uniform meshes to discretize the equations, the grids being refined only at places where high resolutions are needed : close to walls, around the expected positions of shocks or where the solution is expected to vary very fast. This method is limited by two factors. First, the cost may remain high because of the regions where the mesh is refined. Second, the zones where solutions vary rapidly may not be known a priori. To put up with the latter, an iterative process can be used, involving adaptative mesh refinement. The key idea of adaptative techniques is to define a certain norm for the error, which is made up by adding up the contributions from each point, the mesh being refined in the regions where these contributions exceed a certain threshold. A difficulty encountered by this approach is that the process does not stop by itself : either it is stopped "manually", or a minimum size must be fixed for the elements, under which the refinement is suspended. These mesh refinement methods achieve cost reduction by performing accurate computations only where they are needed.

The technique adopted in this thesis does not aim at finding the "exact" solution, but rather seeks to estimate bounds for an output derived from the solution. The idea is to replace the direct solution of the equations on a fine grid by a *Hierarchical Bounds Method* (HBM) that gives a much cheaper estimation of the output of interest considered, assuming that the difference between the fine grid output and the exact output is negligible. Bounds for this fine grid output are actually computed, so that, if they fit into the design constraints, the "exact" solution need not be computed.

The first chapter of this thesis presents a general theory of the Hierarchical Bounds Method, yielding bounds for some given output. The following chapters show how this theory can be successfully applied to several typical problems, the convection-diffusion equation, the pure convection equation (linear problems) and an equation derived from the one-dimensional Euler equations for a nozzle flow (nonlinear problem). In each case the particularities and difficulties encountered are illustrated, as well as the necessary adjustments made to apply the HBM. In particular, the natural boundary conditions used for the adjoint are derived and the sharpness of the bounds is discussed. In the last chapter, the Domain Decomposition technique is presented as a tool, useful especially in two or more dimensions, that allows further computational cost reduction and parallelization of the HBM.

Chapter 1

General Theory for Bounds

1.1 Introduction

This first chapter introduces a general theory for the Hierarchical Bounds Method, yielding bounds for linear functionals of solutions (or *outputs*) of partial or ordinary differential equations. After a brief preliminary discussion about Sobolev spaces, the theory is developped in three steps. First, the continuous problem is discretized by a Galerkin finite element method. Second, the outputs of interest are cast as the stationary point values of a Lagrangian (saddle problem) and bounds for these outputs are derived. Third, a hierarchical procedure is applied to obtain these bounds in a more computationally efficient manner. The last part of this chapter presents a procedure for sharpening the bounds based on the optimization of a stabilization parameter introduced in the formulation. In this thesis only one-dimensional problems are considered.

1.2 Sobolev Spaces

In this section, Sobolev spaces are briefly introduced. For a more complete description, the reader may consult [1] or [2]. First, let Ω be an open subset of $I\!R^n$ $(n \geq 1)$. The space of infinitely continuously differentiable functions with a compact support on Ω is denoted $\mathcal{D}(\Omega)$. The space $\mathcal{D}'(\Omega)$ of distributions on Ω is defined as the dual space of $\mathcal{D}(\Omega)$, i.e. the space of linear forms that are "continuous" on $\mathcal{D}(\Omega)$. The duality between $\mathcal{D}'(\Omega)$ and $\mathcal{D}(\Omega)$ is denoted $< T, \phi >$ $\forall T \in \mathcal{D}'(\Omega)$ and $\forall \phi \in \mathcal{D}(\Omega)$.

Let $L^2(\Omega)$ be the space of square integrable functions on Ω with respect to Lebesgue's measure, i.e the set of functions such that

$$\int_\Omega |f|^2 \, dx < +\infty \tag{1.1}$$

A scalar product can be defined on $L^2(\Omega)$ by

$$(f,g)_{0,\Omega} = \int_\Omega f(x) g(x) \, dx \tag{1.2}$$

and the corresponding norm is :

$$||f||_{0,\Omega} = (f,f)_{0,\Omega}^{1/2} = \left(\int_\Omega f(x)^2 \, dx \right)^{1/2} \tag{1.3}$$

With this scalar product, $L^2(\Omega)$ is a Hilbert space. Distribution derivation is then defined as :

$$\text{If } T \in \mathcal{D}'(\Omega), \, \forall \, \phi \in \mathcal{D}(\Omega) \text{ and } \forall \, i \; (1 \leq i \leq n), \quad < \frac{\partial T}{\partial x_i}, \phi >= - < T, \frac{\partial \phi}{\partial x_i} > \tag{1.4}$$

The first order Sobolev space on Ω is then defined as

$$\mathcal{H}^1(\Omega) = \left\{ v \in L^2(\Omega) \;\middle|\; \frac{\partial v}{\partial x_i} \in L^2(\Omega) \, , 1 \leq i \leq n \right\} \tag{1.5}$$

A scalar product can be defined on $\mathcal{H}^1(\Omega)$ as

$$(u,v)_{1,\Omega} = \int_\Omega \left(\sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} + u \, v \right) \, dx \tag{1.6}$$

and the associated norm is :

$$||u||_{1,\Omega} = (u,u)_{1,\Omega}^{1/2} = \left\{ \int_\Omega \left[\sum_{i=1}^{n} \left(\frac{\partial u}{\partial x_i} \right)^2 + u^2 \right] \, dx \right\}^{1/2} \tag{1.7}$$

With this scalar product, $\mathcal{H}^1(\Omega)$ is a Hilbert space. The closure of $\mathcal{D}(\Omega)$ in $\mathcal{H}^1(\Omega)$, i.e. the set of all functions of $\mathcal{H}^1(\Omega)$ that are limits of converging sequences of functions of $\mathcal{D}(\Omega)$, is denoted $\mathcal{H}_0^1(\Omega)$. It can be shown that $\mathcal{H}_0^1(\Omega)$ is the set of functions of $\mathcal{H}^1(\Omega)$ that vanish on the boundary of Ω. For example, if $n = 1$ and $\Omega =]0,1[$, then $\mathcal{H}_0^1(\Omega)$ is the set of functions of $\mathcal{H}^1(\Omega)$ that vanish at $x = 0$ and $x = 1$.

The dual of $\mathcal{H}_0^1(\Omega)$ considered as a subset of $\mathcal{H}^1(\Omega)$ is denoted by $\mathcal{H}^{-1}(\Omega)$. That is, $\mathcal{H}^{-1}(\Omega)$ is the set of linear forms that are continuous on $\mathcal{H}_0^1(\Omega)$. One can show that the elements of $\mathcal{H}^{-1}(\Omega)$ are the finite sums of functions in $L^2(\Omega)$ and first order derivatives of functions in $L^2(\Omega)$.

1.3 Continuous Problem

The generality of the theory is kept if one assumes that the domain on which the differential equation is defined is $\mathcal{D} =]0,1[$. The corresponding Sobolev spaces are denoted $\mathcal{H}^1(\mathcal{D})$, $\mathcal{H}_0^1(\mathcal{D})$ and $\mathcal{H}^{-1}(\mathcal{D})$. The problem considered is a *second order linear* problem where the values of the solution u at 0 and 1 are imposed :

$$\begin{cases} f(x,u,u_x,u_{xx}) = g \\ u(0) = U_0 \qquad u(1) = U_1 \end{cases} \tag{1.8}$$

f is a linear function of its arguments and the forcing function g is assumed to be in $\mathcal{H}^{-1}(\mathcal{D})$. Let $\mathcal{H}_E^1(\mathcal{D})$ be the set of all functions $v(x)$ in $\mathcal{H}^1(\mathcal{D})$ that satisfy the boundary conditions in 0 and 1 :

$$\mathcal{H}_E^1(\mathcal{D}) = \Big\{ v \in \mathcal{H}^1(\mathcal{D}) \,|\, v(0) = U_0,\, v(1) = U_1 \Big\} \tag{1.9}$$

One looks for solutions u to (1.8) in $\mathcal{H}_E^1(\mathcal{D})$.

A weak formulation for the problem is obtained by multiplying the differential equation by $w \in \mathcal{H}_0^1(\mathcal{D})$ and integrating over $\mathcal{D}$, the second derivative terms being integrated by parts (see [3]). The final result can be written in the form :

$$\int_0^1 [w\, f_1(x,u,u_x) + w_x\, f_2(x,u,u_x)]\; dx = \int_0^1 w\, g\, dx \tag{1.10}$$

In a more abstract form, the problem (1.10) can be stated as finding the solution $u \in \mathcal{H}_E^1(\mathcal{D})$ such that :

$$a(u,w) = M(w) \qquad \forall\, w \in \mathcal{H}_0^1(\mathcal{D}) \tag{1.11}$$

We shall consider problems such that the bilinear form $a(u,w)$ is *coercive* :

$$\exists\, \alpha > 0 \quad \text{such that} \quad \forall\, u \in \mathcal{H}_E^1(\mathcal{D}), \qquad a(u,u) \geq \alpha\, ||u||_{1,\Omega}^2 \tag{1.12}$$

Lax-Milgram's Theorem then ensures that the problem (1.11) has a unique solution [3].

Although the theory to be presented can be easily generalized to nonlinear convex functionals, the outputs considered are *linear* functionals of the field u. These outputs are written as :

$$s^{(1)} = l^{(1)}(u),\ s^{(2)} = l^{(2)}(u), \dots \tag{1.13}$$

1.4 Discrete Equations

We consider a linear Galerkin finite element approximation [4] on a general mesh with a uniform grid-spacing δ.

Let n be the total number of nodes *interior* to the interval $[0,1]$. One has :

$$n = \frac{1}{\delta} - 1 \tag{1.14}$$

Let $x_j = j\,\delta$ be the coordinate of the jth node of the mesh $(0 \leq j \leq n+1)$. To each of these nodes is associated a piecewise linear function φ_j ("hat" function) equal to 1 at node x_j and to 0 at all other nodes. Let $X_E \subset \mathcal{H}^1_E(\mathcal{D})$ and $X_0 \subset \mathcal{H}^1_0(\mathcal{D})$ be the classical continuous-piecewise polynomial sets. They can be expressed as :

$$\begin{aligned} X_0 &= \operatorname{span}\{\varphi_1(x), \dots, \varphi_n(x)\} & (1.15)\\ X_E &= \{v_E(x) = U_0\,\varphi_0(x) + v(x) + U_1\,\varphi_{n+1}(x) \,|\, v(x) \in X_0\} & (1.16) \end{aligned}$$

where U_0 and U_1 are the boundary conditions of the problem.

The finite element method then consists of approximating the solution u by its decomposition on the basis functions $\{\varphi_0, \dots, \varphi_{n+1}\}$:

$$u(x) \approx U_0\,\varphi_0(x) + \sum_{j=1}^{n} u(x_j)\,\varphi_j(x) + U_1\,\varphi_{n+1}(x)\,, \tag{1.17}$$

The *Galerkin* finite element method is a particular case where the "test function" w is chosen to be $\varphi_j(x)\,(1 \leq j \leq n)$, which leads, after inserting (1.17) into (1.10) and evaluating the integrals, to a linear system of n equations with n unknowns that one can write as :

$$L\,u = f \tag{1.18}$$

where $u = [u_1, \ldots, u_n]^T$ now designates the set of unknowns (values of the solution u at the points *interior* to the domain $\mathcal{D}$). From now on, depending on the context, the notation u may denote either the solution to the continuous problem (function) or the solution to the discrete problem (vector). The context will prevent any confusion. L is an $n \times n$ matrix (not necessarily symmetric) and f is the *forcing term.* The latter can be either calculated exactly when the function g is simple enough :

$$f_j = \int_0^1 \varphi_j \, g \, dx + \alpha(0)\, \delta_{1,j} + \beta(1)\, \delta_{n,j} = \int_{x_{j-1}}^{x_{j+1}} \varphi_j \, g \, dx + \alpha(0)\, \delta_{1,j} + \beta(1)\, \delta_{n,j} \tag{1.19}$$

or computed numerically by projecting g onto the space spanned by the basis functions φ_i's and performing the exact integration of the products $\varphi_i \, \varphi_j$:

$$\begin{aligned} \int_0^1 \varphi_j \, g \, dx &\approx \sum_{i=0}^{n+1} g_i \int_0^1 \varphi_j \, \varphi_i \, dx \\ &= \frac{\delta}{6} \left(g_{j-1} + 4 g_j + g_{j+1} \right) \end{aligned} \tag{1.20}$$

In (1.19), $\delta_{1,j}$ and $\delta_{n,j}$ are the Kronecker symbols ; $\alpha(0)$ and $\beta(1)$ are coefficients that depend on the equation and contain the boundary conditions $u(0) = U_0$ and $u(1) = U_1$. $\alpha(0)$ and $\beta(1)$ appear only in the first and the last components of the vector f respectively.

A discrete linear output of the problem can now be expressed as a function of the solution vector u :

$$s = u^T \ell + c \tag{1.21}$$

with $\ell \in I\!R^n$ and $c \in I\!R$.

For the following analysis, we introduce the matrix A which is *twice* the symmetric part of the matrix L :

$$A = L + L^T \tag{1.22}$$

Because the bilinear form in (1.11) has been assumed coercive, the matrix A is positive definite.

1.5 Duality Approach to Bounds for the Outputs

Following [5], a quadratic "augmented" output functional is first constructed. The first step consists of pre-multiplying (1.18) by u^T, and post-multiplying the transpose of (1.18) by u, to obtain

$$u^T L\, u \;=\; u^T f \tag{1.23}$$

$$u^T L^T u \;=\; f^T u \tag{1.24}$$

Adding these two equations, dividing by 2 and noting that the right-hand sides are equal, one obtains, with (1.22) :

$$\frac{1}{2}\, u^T A\, u - u^T f = 0 \tag{1.25}$$

Let us now define the functional

$$\mathcal{S}^{\pm}(v) = \frac{\kappa}{2}\, v^T A\, v - \kappa\, v^T f \pm \left(v^T \ell + c\right) \qquad \forall\, v \in I\!R^n \text{ and } \forall\, \kappa \in I\!R^+ \tag{1.26}$$

From (1.21) and (1.25), the output and its opposite can be written :

$$\pm\, s = \mathcal{S}^{\pm}(u) \tag{1.27}$$

Because u is the unique solution of the system (1.18), it is the only element of the set $\{v \in I\!R^n \,|\, L\, v = f\}$, and the output can be rewritten as :

$$\pm\, s = \min_{\{v \in I\!R^n \,|\, L\, v = f\}} \mathcal{S}^{\pm}(v) \tag{1.28}$$

This trivial result transforms the original problem into a constrained minimization problem. Following [6], a Lagrange multiplier (or adjoint) μ can be used to build the constraint of the *primal* problem (1.28) into a Lagrangian :

$$\mathcal{L}^{\pm}(v, \mu) = \mathcal{S}^{\pm}(v) + \mu^T (Lv - f) \tag{1.29}$$

(1.29) can be interpreted as an *augmented* Lagrangian with respect to the output s, in which κ plays the role of a stabilization parameter.

The *dual* problem is obtained by eliminating v from the Lagrangian. To that end, the

Lagrangian is minimized with respect to v. The stationarity condition for this *unconstrained* minimization is

$$\kappa\, A\, v = \kappa\, f \mp \ell - L^T \mu \tag{1.30}$$

A being positive definite, (1.30) has a unique solution that can be inserted into (1.29) to obtain

$$\min_{v \in I\!R^n} \mathcal{L}^{\pm}(v, \mu) = -\frac{1}{2\kappa}\,(L^T\mu \pm \ell - \kappa\, f)^T A^{-1}\,(L^T\mu \pm \ell - \kappa\, f) \pm c - \mu^T f = -\mathcal{R}^{\pm}(\mu) \tag{1.31}$$

The dual problem can therefore be written :

$$\max_{\mu \in I\!R^n} -\mathcal{R}^{\pm}(\mu) = \max_{\mu \in I\!R^n} \left[-\frac{1}{2\kappa}\,(L^T\mu \pm \ell - \kappa\, f)^T A^{-1}\,(L^T\mu \pm \ell - \kappa\, f) \pm c - \mu^T f \right] \tag{1.32}$$

By definition,

$$\forall\,(v, \mu) \in I\!R^n \times I\!R^n, \quad -\mathcal{R}^{\pm}(\mu) \le \mathcal{L}^{\pm}(v, \mu) \tag{1.33}$$

Weak duality then follows from the equality of $\mathcal{L}^{\pm}(v, \mu)$ and $\mathcal{S}^{\pm}(v)$ when the constraint is satisfied :

$$\text{For all admissible } v \text{ and } \mu, \qquad -\mathcal{R}^{\pm}(\mu) \le \mathcal{S}^{\pm}(v) \tag{1.34}$$

A small trick can be used to extend this inequality to *all* vectors $(v, \mu) \in I\!R^n \times I\!R^n$: if the constraint is satisfied, then the value of $\mathcal{S}^{\pm}(v)$ is given by (1.26), whereas, if the constraint is not satisfied, the value of $\mathcal{S}^{\pm}(v)$ is set equal to $+\infty$. Thus, one has

$$\forall\,(v, \mu) \in I\!R^n \times I\!R^n, \qquad -\mathcal{R}^{\pm}(\mu) \le \mathcal{L}^{\pm}(v, \mu) \le \mathcal{S}^{\pm}(v) \tag{1.35}$$

It can be shown that the values of the constrained minimum of $\mathcal{S}^{\pm}(v)$ and of the maximum of $-\mathcal{R}^{\pm}(\mu)$ are equal :

$$\begin{aligned}
\mathcal{S}^{\pm}(v) + \mathcal{R}^{\pm}(\mu) &= \frac{\kappa}{2}\, v^T A\, v - \kappa\, v^T f \pm \left(v^T \ell + c\right) \\
&\quad + \frac{1}{2\kappa}\,(L^T\mu \pm \ell - \kappa\, f)^T A^{-1}\,(L^T\mu \pm \ell - \kappa\, f) \mp c + \mu^T f \\
&= \frac{1}{2\kappa}\,(\kappa\, A\, v + L^T\mu \pm \ell - \kappa\, f)^T A^{-1}\,(\kappa\, A\, v + L^T\mu \pm \ell - \kappa\, f) \\
&\quad - \frac{1}{\kappa}(\kappa\, A\, v)^T A^{-1}(L^T\mu \pm \ell - \kappa\, f) - \kappa v^T L\, v \pm v^T \ell + \mu^T L\, v \\
&= \frac{1}{2\kappa}\,(\kappa\, A\, v L^T\mu \pm \ell - \kappa\, f)^T A^{-1}\,(\kappa\, A\, v L^T\mu \pm \ell - \kappa\, f)
\end{aligned} \tag{1.36}$$

Equation (1.36) is obtained by making use of the fact that A is symmetric and that for all admissible v, the equality $L\,v = f$ is satisfied. The stationarity condition of the Lagrangian with respect to v being (1.30), the right-hand side in (1.36) vanishes, which proves the *Minimax Theorem* :

$$\min_{\{v \in \mathbb{R}^n \,|\, L\,v = f\}} \mathcal{S}^{\pm}(v) = \max_{\{\mu \in \mathbb{R}^n\}} -\mathcal{R}^{\pm}(\mu) \tag{1.37}$$

$\mathcal{S}^{\pm}(v)$ is by construction the maximum of the Lagrangian when μ varies and $-\mathcal{R}^{\pm}(\mu)$ is by definition the minimum of the Lagrangian when v varies, so the Minimax Theorem allows us to write :

$$\begin{aligned} \pm\, s &= \min_{\{v \in \mathbb{R}^n\}} \max_{\{\mu \in \mathbb{R}^n\}} \mathcal{L}^{\pm}(v, \mu) && (1.38) \\ &= \max_{\{\mu \in \mathbb{R}^n\}} \min_{\{v \in \mathbb{R}^n\}} \mathcal{L}^{\pm}(v, \mu) && (1.39) \end{aligned}$$

The solution $(u, \psi^{\pm})$ of this saddle problem, also called *saddlepoint*, is determined by the stationarity conditions derived from setting the derivatives of the Lagrangian with respect to v and to μ equal to 0 :

$$\begin{aligned} \kappa\, A\, u + L^T \psi^{\pm} - \kappa\, f \pm \ell &= 0 && (1.40) \\ L\, u - f &= 0 && (1.41) \end{aligned}$$

Equation (1.41), which is equivalent to (1.18), shows that u does not depend on the sign chosen for the output and the Lagrangian, whereas, from (1.40), $\psi^{\pm}$ does.

The bounds immediately follow from (1.39), because

$$\forall \hat{\mu}^{\pm} \in \mathbb{R}^n \qquad \min_{\{v \in \mathbb{R}^n\}} \mathcal{L}^{\pm}(v, \hat{\mu}^{\pm}) \leq \pm s \tag{1.42}$$

which can also be written as :

$$\min_{\{v \in \mathbb{R}^n\}} \mathcal{L}^{+}\left(v, \hat{\mu}^{+}\right) \leq s \leq - \min_{\{v \in \mathbb{R}^n\}} \mathcal{L}^{-}\left(v, \hat{\mu}^{-}\right) \tag{1.43}$$

Thus, for any $\hat{\mu}^{\pm} \in \mathbb{R}^n$, solving (1.40) to obtain the vector $v \in \mathbb{R}^n$ that minimizes the left-hand side of (1.42) and plugging it into (1.29), we obtain upper and lower bounds for the output.

1.6 Hierarchical Procedure

Equation (1.43) can be used to give bounds for the discretized output of interest. From (1.38)–(1.39), the bounds are exactly equal to the output when the Lagrange multiplier $\hat{\mu}^{\pm}$ and the field v satisfy the stationarity conditions (1.40)–(1.41). Unfortunately, two difficulties complicate the choice of $\hat{\mu}^{\pm}$ in contradictory ways. First, the vector $\hat{\mu}^{\pm}$ chosen to compute the bounds must be as close to the "true" adjoint as possible, because the sharpness of the bounds is closely related to the quality of the approximation of the exact saddlepoint. Second, (1.41) shows that the computation of the adjoint requires the resolution of the original discrete system, hence makes the computation of the exact saddlepoint on a fine grid prohibitively expensive.

Because the saddlepoint cannot be computed cheaply on a fine grid, the Hierarchical Bounds Method consists of considering 2 different grids, i.e. 2 levels of discretization, one fine ("truth" mesh) and one coarse ("working" mesh), and solving for the saddlepoint only on the coarse grid. The adjoint is then interpolated to obtain an approximation of the exact saddlepoint on the fine grid and the bounds are obtained from (1.43) by solving only a symmetric problem. Further gains on the cost can be achieved, even in one dimension, by the use of the Domain Decomposition technique presented in the last chapter.

The general uniform grid-spacing δ can now be equal either to H (coarse grid) or to h (fine grid). The nodes of the coarse mesh are, from now on, assumed to be nodes of the fine mesh as well. The variables corresponding to the coarse grid (H-mesh, or "working" discretization) are denoted with an H subscript (e.g. $L_H, u_H, x_{Hj}, \ldots$), while the variables corresponding to the fine grid (h-mesh, or "truth" discretization) are denoted with an h subscript (e.g. $L_h, u_h, x_{hi}, \ldots$). The numbers of the points of the meshes that are interior to the domain $\mathcal{D}$ are N and n for the coarse and the fine grids respectively.

1.6.1 Computational Procedure

The saddlepoint on the coarse mesh $(u_H, \psi_H^{\pm})$ is first determined by solving the stationarity conditions $(\delta = H)$:

$$\kappa\, A_H\, u_H + L_H^T \psi_H^{\pm} - \kappa\, f_H \pm \ell_H = 0 \tag{1.44}$$

$$L_H\, u_H - f_H = 0 \tag{1.45}$$

(1.45) is solved for u_H (coarse grid solution), plugged into (1.44), which is then solved for $\psi_H^{\pm} \in I\!R^N$ by :

$$L_H^T \psi_H^{\pm} = -\left(\kappa\, A_H\, u_H - \kappa\, f_H \pm \ell_H\right) \tag{1.46}$$

Next, $\hat{\mu}_h^{\pm} \in I\!R^n$ is formed by interpolation of $\psi_H^{\pm}$ on the h-mesh ($\delta = h$). Let the boundary values for the adjoint in $x = 0$ and $x = 1$ be $\psi_0^{b\pm}$ and $\psi_1^{b\pm}$ respectively. One has $\forall\, i\ (1 \leq i \leq n)$:

$$\left(\hat{\mu}_h^{\pm}\right)_i = \psi_0^{b\pm}\, \varphi_{H,0}\left(x_{h,i}\right) + \sum_{j=1}^{N} \left(\psi_H^{\pm}\right)_j\, \varphi_{H,j}\left(x_{h,i}\right) + \psi_1^{b\pm}\, \varphi_{H,N+1}\left(x_{h,i}\right) \tag{1.47}$$

Equation (1.47) shows that the values of $\hat{\mu}_h^{\pm}$, especially at points close to $x = 0$ and $x = 1$, directly depend on the choice of the boundary conditions $\psi_0^{b\pm}$ and $\psi_1^{b\pm}$ for the adjoint $\psi_H^{\pm}$. Two considerations must be taken into account for this choice. First, the approximation $\hat{\mu}_h^{\pm}$ of the discrete adjoint has to be of good quality, otherwise accuracy is lost and the sharpness of the bounds is affected accordingly. Second, the interpolated adjoint has to be consistent with the underlying continuous problem. Because the choice for these *natural* boundary conditions depends essentially on the problem studied, they have to be determined for each particular case. Let us just mention, however, that the boundary conditions are derived from the continuous problem and a continuous equivalent of the discrete Lagrangian (1.29).

Finally, the bounds for the output s_h are computed on the fine grid as :

$$\left(s_h\right)_{LB}(H) = \min_{\{v \in I\!R^n\}} \mathcal{L}^+\left(v, \hat{\mu}_h^+\right) \tag{1.48}$$

$$\left(s_h\right)_{UB}(H) = -\min_{\{v \in I\!R^n\}} \mathcal{L}^-\left(v, \hat{\mu}_h^-\right) \tag{1.49}$$

The stationarity conditions for these two unconstrained minimization problems are simply (1.40) written on the fine grid with $\psi^{\pm} = \hat{\mu}_h^{\pm}$, i.e. :

$$\kappa\, A_h\, \hat{u}_h^{\pm} = -\left(L_h^T \hat{\mu}_h^{\pm} - \kappa\, f_h \pm \ell_h\right) \tag{1.50}$$

which is solved for $\hat{u}_h^{\pm}$. The computation of the bounds is then straightforward by plugging $\hat{u}_h^{\pm}$ and $\hat{\mu}_h^{\pm}$ into (1.29), (1.48) and (1.49) :

$$\left(s_h\right)_{LB}(H) = \mathcal{L}^+\left(\hat{u}_h^+, \hat{\mu}_h^+\right) \tag{1.51}$$

$$\left(s_h\right)_{UB}(H) = -\mathcal{L}^-\left(\hat{u}_h^-, \hat{\mu}_h^-\right) \tag{1.52}$$

Equation (1.50) shows that, although the first component of the exact saddlepoint $(u, \psi^{\pm})$ did not depend on the sign chosen for the output and the Lagrangian, $\hat{u}_h^{\pm}$ *does* depend on this sign, since it is determined from $\hat{\mu}_h^{\pm}$.

A simpler expression can now be derived for (1.51) and (1.52) : multiplying (1.50) on the left by $\hat{u}_h^{\pm T}$, one obtains

$$\kappa\, \hat{u}_h^{\pm T} A_h\, \hat{u}_h^{\pm} = -\left(\hat{u}_h^{\pm T} L_h^T \hat{\mu}_h^{\pm} - \kappa\, \hat{u}_h^{\pm T} f_h \pm \hat{u}_h^{\pm T} \ell_h\right) \tag{1.53}$$

Therefore,

$$\begin{aligned}
\mathcal{L}_h^{\pm}(\hat{u}_h^{\pm}, \hat{\mu}_h^{\pm}) &= \frac{\kappa}{2}\, \hat{u}_h^{\pm T} A_h\, \hat{u}_h^{\pm} - \kappa\, \hat{u}_h^{\pm T} f_h \pm \left(\hat{u}_h^{\pm T} \ell_h + c_h\right) + \hat{\mu}_h^{\pm T} (L_h\, \hat{u}_h^{\pm} - f_h) && (1.54)\\
&= -\frac{\kappa}{2}\, \hat{u}_h^{\pm T} A_h\, \hat{u}_h^{\pm} \pm c_h - \hat{\mu}_h^{\pm T} f_h && (1.55)
\end{aligned}$$

Equation (1.55) is obtained by adding (1.54) and (1.53). The bounds can thus be computed as :

$$\begin{aligned}
(s_h)_{LB}\,(H) &= -\frac{\kappa}{2}\, \hat{u}_h^{+T} A_h\, \hat{u}_h^{+} + c_h - \hat{\mu}_h^{+T} f_h && (1.56)\\
(s_h)_{UB}\,(H) &= \frac{\kappa}{2}\, \hat{u}_h^{-T} A_h\, \hat{u}_h^{-} + c_h + \hat{\mu}_h^{-T} f_h && (1.57)
\end{aligned}$$

1.6.2 Computational Cost

We now briefly discuss the advantages and drawbacks of the Hierarchical Bounds Method. The advantages of this procedure are not immediately obvious. As a matter of fact, the cost of the method resides essentially in the inversion of the matrix A_h to obtain $\hat{u}_h^{\pm}$: the inversion of L_h has been replaced by that of A_h, which is as costly, since both matrices have the same size and are tridiagonal in the one-dimensional case. Given that tridiagonal matrices can be easily and cheaply inverted, even when their sizes are large [7], the Hierarchical Bounds Method does not seem very advantageous.

However, the HBM presents two undeniable advantages. First, in two dimensions, the HBM represents a real improvement, because the matrices are not tridiagonal anymore, but *sparse*. Replacing the inversion of the matrix L_h, which is not necessarily symmetric, by the inversion of A_h, which *is* symmetric becomes critical. As a matter of fact, numerous very efficient algorithms are available to solve sparse symmetric problems (especially iterative processes, like the conjugate gradient [8, 9, 10]), whereas existing methods to invert non-symmetric matrices are usually neither systematic nor efficient. Second, the cost of the HBM can be improved, even in one dimension, by the use of the Domain Decomposition studied in more detail in the last chapter. The key idea here is to divide the domain $\mathcal{D}$ into several subdomains and solve in each subdomain Neumann or Neumann-Dirichlet *decoupled* problem with appropriate boundary fluxes. The cost is reduced because each subdomain contains many less points and the resolution of the problems on the subdomains can be easily parallelized.

1.7 Optimal Stabilization Parameter

The results of Sections 1.5 and 1.6 are valid for all positive values of the stabilization parameter κ. This additional parameter can now be used to optimize (i.e. sharpen) the bounds. To begin, the adjoint on the coarse grid is decomposed as :

$$\psi_H^{\pm} = \psi_H^{0\pm} + \kappa\, \psi_H^{1\pm} \tag{1.58}$$

where

$$\begin{aligned} \psi_H^{0\pm} &= \mp (L_H^T)^{-1} \ell_H && (1.59) \\ \psi_H^{1\pm} &= (L_H^T)^{-1} \left[f_H - A_H u_H \right] && (1.60) \end{aligned}$$

u_H being the solution of the system on the coarse grid (1.18). Similarly, the boundary values for this adjoint are decomposed as :

$$\begin{aligned} \psi_0^{b\pm} &= \psi_0^{0\,b\pm} + \kappa\, \psi_0^{1\,b\pm} && (1.61) \\ \psi_1^{b\pm} &= \psi_1^{0\,b\pm} + \kappa\, \psi_1^{1\,b\pm} && (1.62) \end{aligned}$$

where $\psi_0^{0\,b\pm}$, $\psi_0^{1\,b\pm}$, $\psi_1^{0\,b\pm}$ and $\psi_1^{1\,b\pm}$ are independent of κ.

The approximated adjoint on the fine grid can now be written as :

$$\hat{\mu}_h^{\pm} = \hat{\mu}_h^{0\pm} + \kappa\, \hat{\mu}_h^{1\pm} \tag{1.63}$$

where $\hat{\mu}^{0\pm}$ and $\hat{\mu}^{1\pm}$ are interpolations on the fine grid of $\psi_H^{0\pm}$ and $\psi_H^{1\pm}$ respectively :

$$\begin{aligned}
\left(\hat{\mu}_h^{0\pm}\right)_i &= \psi_0^{0\,b\pm}\varphi_{H0}\left(x_{hi}\right) + \sum_{j=1}^{N}\left(\psi_H^{0\pm}\right)_j \varphi_{Hj}\left(x_{hi}\right) + \psi_1^{0\,b\pm}\varphi_{H,N+1}\left(x_{hi}\right) && (1.64)\\
\left(\hat{\mu}_h^{1\pm}\right)_i &= \psi_0^{1\,b\pm}\varphi_{H0}\left(x_{hi}\right) + \sum_{j=1}^{N}\left(\psi_H^{1\pm}\right)_j \varphi_{Hj}\left(x_{hi}\right) + \psi_1^{1\,b\pm}\varphi_{H,N+1}\left(x_{hi}\right) && (1.65)
\end{aligned}$$

for all $i\,(1 \leq i \leq n)$. Following the notations of [5], the bounds are denoted as :

$$\begin{aligned}
\eta^+(\kappa) &\equiv (s_h)_{LB} && (1.66)\\
\eta^-(\kappa) &\equiv -(s_h)_{UB} && (1.67)
\end{aligned}$$

Two new vectors can be introduced :

$$\begin{aligned}
y_h^{\pm} &= L_h^T \hat{\mu}_h^{0\pm} \pm \ell_h && (1.68)\\
z_h^{\pm} &= L_h^T \hat{\mu}_h^{1\pm} - f_h && (1.69)
\end{aligned}$$

and the corresponding inner products are defined as :

$$\begin{aligned}
\alpha^{\pm} &= y_h^{\pm T} A_h^{-1}\, y_h^{\pm} && (1.70)\\
\beta^{\pm} &= z_h^{\pm T} A_h^{-1}\, z_h^{\pm} + 2\, f_h^T \hat{\mu}_h^{1\pm} && (1.71)
\end{aligned}$$

where A_h^{-1} is interpreted as the inverse of A_h.

Using (1.50), (1.55), (1.63) and (1.68)–(1.71), the bounds (1.66) and (1.67) can be written :

$$\begin{aligned}
\eta^{\pm}(\kappa) &= -\frac{\kappa}{2}\,\hat{u}_h^{\pm T} A_h\,\hat{u}_h^{\pm} \pm c_h - \hat{\mu}_h^{\pm T} f_h\\
&= -\frac{\kappa}{2}\left(A_h\,\hat{u}_h^{\pm}\right)^T A_h^{-1}\left(A_h\,\hat{u}_h^{\pm}\right) \pm c_h - \left(\hat{\mu}_h^{0\pm} + \kappa\,\hat{\mu}_h^{1\pm}\right)^T f_h\\
&= -\frac{1}{2\kappa}\left[L_h^T\left(\hat{\mu}_h^{0\pm} + \kappa\,\hat{\mu}_h^{1\pm}\right) - \kappa\, f_h \pm \ell_h\right]^T A_h^{-1}\left[L_h^T\left(\hat{\mu}_h^{0\pm} + \kappa\,\hat{\mu}_h^{1\pm}\right) - \kappa\, f_h \pm \ell_h\right]\\
&\qquad \pm c_h - \left(\hat{\mu}_h^{0\pm} + \kappa\,\hat{\mu}_h^{1\pm}\right)^T f_h\\
&= -\frac{1}{2\kappa}\left(y_h^{\pm} + \kappa\, z_h^{\pm}\right)^T A_h^{-1}\left(y_h^{\pm} + \kappa\, z_h^{\pm}\right) \pm c_h - \left(\hat{\mu}_h^{0\pm} + \kappa\,\hat{\mu}_h^{1\pm}\right)^T f_h\\
&= -\frac{1}{2\kappa}\,\alpha^{\pm} - \frac{\kappa}{2}\left(\beta^{\pm} - 2\, f_h^T \hat{\mu}_h^{1\pm}\right) - y_h^{\pm T} A_h^{-1}\, z_h^{\pm} \pm c_h - \left(\hat{\mu}_h^{0\pm} + \kappa\,\hat{\mu}_h^{1\pm}\right)^T f_h \qquad (1.72)
\end{aligned}$$

where the symmetry of A_h (therefore of A_h^{-1}) has been used. Finally, the bounds are expressed as functions of κ:

$$\eta^{\pm}(\kappa) = -\frac{1}{2\kappa}\,\alpha^{\pm} - \frac{\kappa}{2}\,\beta^{\pm} - y_h^{\pm T} A_h^{-1}\, z_h^{\pm} - f_h^T \hat{\mu}^{0\pm} \pm c_h \tag{1.73}$$

Taking the first derivative of $\eta^{\pm}(\kappa)$ with respect to κ :

$$\eta_\kappa^{\pm}(\kappa) = \frac{1}{2\kappa^2}\,\alpha^{\pm} - \frac{1}{2}\,\beta^{\pm} \tag{1.74}$$

The stationarity condition $\eta_\kappa^{\pm}(\kappa) = 0$ yields the value of the optimal stabilization parameter, $\kappa^{*\pm}$:

$$\kappa^{*\pm} = \sqrt{\frac{\alpha^{\pm}}{\beta^{\pm}}} \tag{1.75}$$

where the denominator is assumed positive for $\kappa^{*\pm}$ to be defined. This is not guaranteed, in particular for nonlinear problems.

Note that (1.59)–(1.62) and (1.64)–(1.65) show that $\hat{\mu}_h^{0-} = -\hat{\mu}_h^{0+}$ and $\hat{\mu}_h^{1-} = \hat{\mu}_h^{1+}$, so that $y_h^- = -y_h^+$ and $z_h^- = z_h^+$. Therefore, $\alpha^- = \alpha^+$ and $\beta^- = \beta^+$, which leads to

$$\kappa^{*+} = \kappa^{*-} = \kappa^* = \sqrt{\frac{\alpha^{\pm}}{\beta^{\pm}}} \tag{1.76}$$

A second consequence concerns the predictor (or predicted output) defined as the average value of the bounds :

$$s_{pre}(H) = \frac{1}{2}\left[(s_h)_{UB}(H) + (s_h)_{LB}(H)\right] \tag{1.77}$$

This predictor *does not* depend on the stabilization parameter κ. Indeed, we have :

$$\begin{aligned} s_{pre}(H) &= \frac{1}{2}\left[\eta^+(\kappa) - \eta^-(\kappa)\right] \\ &= \frac{1}{2}\Big[-\frac{1}{2\kappa}\,\alpha^+ - \frac{\kappa}{2}\,\beta^+ - y_h^{+T} A_h^{-1}\, z_h^+ - f_h^T \hat{\mu}^{0+} + c_h \\ &\qquad + \frac{1}{2\kappa}\,\alpha^- + \frac{\kappa}{2}\,\beta^- + y_h^{-T} A_h^{-1}\, z_h^- + f_h^T \hat{\mu}^{0-} + c_h\Big] \\ &= -y_h^{+T} A_h^{-1}\, z_h^+ - f_h^T \hat{\mu}^{0+} + c_h \end{aligned} \tag{1.78}$$

where κ does not appear anymore.

From a computational cost point of view, the optimal stabilization parameter needs to be computed only for one of the bounds, its value being the same for the other bound. The computation of κ^* requires the inversion of A_h to be performed twice, which was already the cost of the HBM without optimizing κ. Furthermore, there is no need for any other inversion of A_h, since the expression for the bounds is (1.73), the last inversion in this equation having already been performed to compute $\beta^{\pm}$.

Finally, taking the second derivative of $\eta^{\pm}(\kappa)$ with respect to κ, we obtain :

$$\eta^{\pm}_{\kappa\kappa}(\kappa) = -\frac{\alpha^{\pm}}{\kappa^3} \tag{1.79}$$

When $\kappa = \kappa^*$, this second derivative is negative, which means that the optimum obtained is indeed the maximum of $\eta^{\pm}(\kappa)$. Practically, (1.66) and (1.67) show that the maximum of the lower bound and the minimum of the upper bound are determined by this procedure.

The following chapters are devoted to the application of this theory to three different equations. Each of these problems have its particular features. The first one is the convection-diffusion equation, for which the general theory developed in [5] can be readily applied, without any modification. The second one is the purely convective case, where the straight finite element formulation needs to be modified to stabilize the solution and deal with the absence of the second boundary condition. The third one is a nonlinear equation derived from the steady Euler equations, where the problem needs to be linearized before it can be solved.

Chapter 2

The Convection-Diffusion Problem

In this chapter, the Hierarchical Bounds Method is applied to the one-dimensional convection-diffusion equation.

2.1 Continuous problem

The problem can be formulated in a differential form as :

$$-\nu\, u_{xx} + u_x = g \quad \forall x \in \mathcal{D} \tag{2.1}$$

$$u(0) = U_0, \quad u(1) = U_1 \tag{2.2}$$

where ν is a (small) positive constant, U_0 and U_1 are real numbers and $\mathcal{D} =]0,1[$. Because the problem is time-independent, (2.1) is an ordinary differential equation. The weak formulation of the problem can be written :

For $g \in \mathcal{H}^{-1}(\mathcal{D})$, find $u \in \mathcal{H}^1_E(\mathcal{D})$ such that

$$\int_0^1 (\nu\, w_x\, u_x + w\, u_x)\; dx = \int_0^1 w\, g\, dx \quad \forall w \in \mathcal{H}^1_0(\mathcal{D}) \tag{2.3}$$

The outputs of interest considered are the average value of the solution ($s^{(1)}$) and a pointwise value, i.e. the value of the solution u at a given point $\overline{x} < 1$ ($s^{(2)}$) :

$$s^{(1)} = l^{(1)}(u) \;=\; \int_0^1 u(x)\, dx \tag{2.4}$$

$$s^{(2)} = l^{(2)}(u) \;=\; u(\overline{x}) \tag{2.5}$$

2.2 Continuous Formulation

A continuous formulation for the problem is first introduced. This will prove useful to determine natural boundary conditions for the adjoint.

The first step consists of deriving the equivalent of the discrete Lagrangian (1.29) for the continuous case. Multiplying (2.1) by u and integrating between 0 and 1, one obtains :

$$\int_0^1 -\nu\, u\, u_{xx}\, dx + \int_0^1 u\, u_x\, dx - \int_0^1 u\, g\, dx = 0 \tag{2.6}$$

The first integral is integrated by parts and the second one can be integrated directly :

$$-\nu\, u_x(1)\, u(1) + \nu\, u_x(0)\, u(0) + \int_0^1 \nu\, (u_x)^2\, dx + \frac{u(1)^2 - u(0)^2}{2} - \int_0^1 u\, g\, dx = 0 \tag{2.7}$$

The continuous form of the augmented output (1.26) is then defined as :

$$\mathcal{S}^{\pm}(v) = \kappa \left[-\nu\, v_x(1)\, U_1 + \nu\, v_x(0)\, U_0 + \int_0^1 \nu\, (v_x)^2\, dx + \frac{U_1^2 - U_0^2}{2} - \int_0^1 v\, g\, dx \right] \pm l(v) \tag{2.8}$$

The trivial minimization follows :

$$\pm\, s = \min_{\left\{ v \in \mathcal{H}_E^1(\mathcal{D}) \,\middle|\, \int_0^1 (\nu\, w_x\, u_x + w\, u_x - w\, g)\, dx = 0 \;\; \forall w \in \mathcal{H}_0^1(\mathcal{D}) \right\}} \mathcal{S}^{\pm}(v) \tag{2.9}$$

and the corresponding Lagrangian becomes :

$$\mathcal{L}^{\pm}(v, \mu) = \mathcal{S}^{\pm}(v) + \int_0^1 \mu\, (-\nu\, v_{xx} + v_x - g)\, dx \tag{2.10}$$

Integrating the second derivative term by parts, one obtains :

$$\begin{aligned} \mathcal{L}^{\pm}(v, \mu) \;\; = \;\; & \kappa \left[-\nu\, v_x(1)\, U_1 + \nu\, v_x(0)\, U_0 + \int_0^1 \nu\, (v_x)^2\, dx + \frac{U_1^2 - U_0^2}{2} - \int_0^1 v\, g\, dx \right] \pm l(v) \\ & -\nu\, \mu(1)\, v_x(1) + \nu\, \mu(0)\, v_x(0) + \int_0^1 (\nu\, \mu_x\, v_x + \mu\, v_x - \mu\, g)\, dx \end{aligned} \tag{2.11}$$

Let us now consider the first output, namely the average value of the solution over the domain $\mathcal{D}$, i.e. $l^{(1)}(v) = \int_0^1 v\, dx$. The first variation of (2.11) with respect to variations $w(x) = v(x) - u(x)$ must be equal to 0 for u to be a stationary point :

$$\begin{aligned} & \kappa \left[-\nu\, w_x(1)\, U_1 + \nu\, w_x(0)\, U_0 + \int_0^1 2\nu\, w_x\, u_x\, dx - \int_0^1 w\, g\, dx \right] \pm \int_0^1 w\, dx \\ & \qquad -\nu\, \mu(1)\, w_x(1) + \nu\, \mu(0)\, w_x(0) + \int_0^1 (\nu\, \mu_x\, w_x + \mu\, w_x)\, dx \;\; = \;\; 0 \end{aligned} \tag{2.12}$$

The *natural* boundary conditions for the adjoint now appear simply by grouping the terms containing either $w_x(0)$ or $w_x(1)$, and setting them to 0, since the equation must be valid for any value of $w_x(0)$ and $w_x(1)$. One obtains :

$$\mu(0) = -\kappa U_0 \tag{2.13}$$

$$\mu(1) = -\kappa U_1 \tag{2.14}$$

As far as the second output is concerned, the reasoning is not modified and, since $0 < \overline{x} < 1$, the natural boundary conditions for the adjoint remain unchanged.

2.3 Discrete Equations

The problem (2.1)–(2.2) is discretized with a Galerkin finite element method. Denoting the size of the space discretization δ and the number of points interior to the domain n, the elements of the resulting $n \times n$ matrix of the system (1.18) are given by :

$$L_{ij} = \int_0^1 \left(\nu \frac{d\varphi_i}{dx} \frac{d\varphi_j}{dx} + \varphi_i \frac{d\varphi_j}{dx} \right) dx \qquad \forall (i,j)\ (1 \le i,j \le n) \tag{2.15}$$

where the basis functions φ_i $(1 \le i \le n)$ are the usual "hat" functions (piecewise linear functions). The matrix L then becomes :

$$L = \begin{pmatrix} \frac{2\nu}{\delta} & -\frac{\nu}{\delta} + \frac{1}{2} & 0 & \cdots & \cdots & 0 \\ -\frac{\nu}{\delta} - \frac{1}{2} & \frac{2\nu}{\delta} & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \frac{2\nu}{\delta} & -\frac{\nu}{\delta} + \frac{1}{2} \\ 0 & \cdots & \cdots & 0 & -\frac{\nu}{\delta} - \frac{1}{2} & \frac{2\nu}{\delta} \end{pmatrix} \tag{2.16}$$

The corresponding matrix $A = L + L^T$ can be obtained either from (2.16), or directly by discretizing the self-adjoint part of the differential operator in (2.3), i.e. :

$$A_{ij} = 2 \int_0^1 \left(\nu \frac{d\varphi_i}{dx} \frac{d\varphi_j}{dx} \right) dx \qquad \forall (i,j)\ (1 \le i,j \le n) \tag{2.17}$$

which leads to :

$$A = \begin{pmatrix} \frac{4\nu}{\delta} & -\frac{2\nu}{\delta} & 0 & \cdots & \cdots & 0 \\ -\frac{2\nu}{\delta} & \frac{4\nu}{\delta} & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \frac{4\nu}{\delta} & -\frac{2\nu}{\delta} \\ 0 & \cdots & \cdots & 0 & -\frac{2\nu}{\delta} & \frac{4\nu}{\delta} \end{pmatrix} \tag{2.18}$$

The forcing term f is defined in the general case, for $(1 \le i \le n)$, as :

$$f_i = \int_0^1 \left[\varphi_i \, g - \left(\nu \frac{d\varphi_i}{dx} \frac{d\varphi_{n+1}}{dx} - \varphi_i \frac{d\varphi_{n+1}}{dx} \right) U_1 - \left(\nu \frac{d\varphi_i}{dx} \frac{d\varphi_0}{dx} - \varphi_i \frac{d\varphi_0}{dx} \right) U_0 \right] dx \tag{2.19}$$

Assuming that g is discretized using the φ_i $(1 \le i \le n)$ basis, one obtains :

$$f_i = \frac{\delta}{6} \left(g_{i-1} + 4 g_i + g_{i+1} \right) + \left(\frac{\nu}{\delta} + \frac{1}{2} \right) U_0 \, \delta_{1,i} + \left(\frac{\nu}{\delta} - \frac{1}{2} \right) U_1 \, \delta_{n,i} \tag{2.20}$$

where $\delta_{1,i}$ and $\delta_{n,i}$ are the Kronecker symbols ; $\left(\frac{\nu}{\delta} + \frac{1}{2}\right) U_0$ and $\left(\frac{\nu}{\delta} - \frac{1}{2}\right) U_1$ are the terms $\alpha(0)$ and $\beta(1)$ seen in (1.19) respectively.

The discrete form of the outputs can be written as :

$$s = u^T \ell + c \tag{2.21}$$

where $\ell \in I\!R^n$ and $c \in I\!R$. For the two outputs considered in this chapter, one has $s^{(1)} = \int_0^1 u(x)\, dx \approx \int_0^1 \sum_{i=1}^{n} u_i \, \varphi_i \, dx = \sum_{i=1}^{n} u_i \, \delta + \frac{\delta}{2}$ and $s^{(2)} = u(\overline{x}) = u(x_j) = u_j$. Hence

$$\ell^{(1)} = \delta \, [1 \ldots 1]^T, \quad c^{(1)} = \frac{\delta}{2} \tag{2.22}$$

$$\ell^{(2)} = [0 \ldots 0 \, 1 \, 0 \ldots]^T, \quad c^{(2)} = 0 \tag{2.23}$$

In the case of the second output, the single nonzero component corresponds to $j = \overline{x}/\delta$ assumed integer.

2.4 Numerical Results

The general theory of the Hierarchical Bounds Method can be applied directly to the convection-diffusion equation. The numerical results obtained are presented for both ouputs with and without optimization of the stabilization parameter.

For these numerical simulations, the viscosity parameter ν has been taken equal to 0.1. The forcing term g in (2.1) is equal to 0 over the whole domain $\mathcal{D}$ and the boundary conditions are $U_0 = 0$ and $U_1 = 1$. The right-hand side of the discrete equations f therefore has all its components equal to 0 but the last, which is equal to $\left(\frac{\nu}{\delta} - \frac{1}{2}\right)$. The boundary conditions for the adjoint become :

$$\mu(0) = 0 \tag{2.24}$$

$$\mu(1) = -\kappa \tag{2.25}$$

The size of the fine grid cells is $h = 10^{-3}$, and the bounds presented have been computed for values of the coarse grid discretization H ranging between h and 0.1. More precisely, $H \in \{0.1\,;\, 0.05\,;\, 0.025\,;\, 0.02\,;\, 0.01\,;\, 5.10^{-3}\,;\, 4.10^{-3}\,;\, 2.10^{-3}\,;\, 10^{-3}\}$. Figure 2-1 presents the

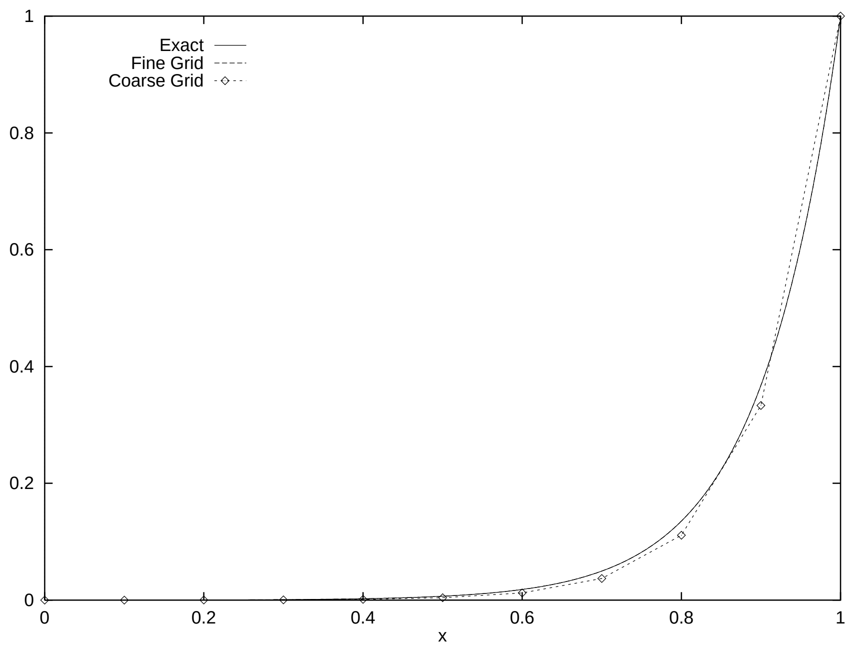


Figure 2-1: Solution of the convection-diffusion problem ($h = 10^{-3}$, $H = 0.1$)

solutions of the problem obtained on the fine and coarse ($H = 0.1$) grids. The exact solution of the problem (2.1)–(2.2), $u(x) = (e^{\nu x} - 1)\,/\,(e^{\nu} - 1)$ is given as a reference.

Good accuracy is observed, even for the coarsest grid. For the fine grid, the plot of the solution computed by the Galerkin finite element method cannot be distinguished from that of the analytical solution.

2.4.1 First Output : Average of the Solution

The results obtained for the first output (average of the solution over the domain $\mathcal{D}$) are now presented. The stabilization parameter κ is chosen equal to 1, which in this case happens to coincide with the optimal value, as described in Chapter 1. Figure 2-2 shows the adjoints $\psi_H^{\pm}$ for this output.

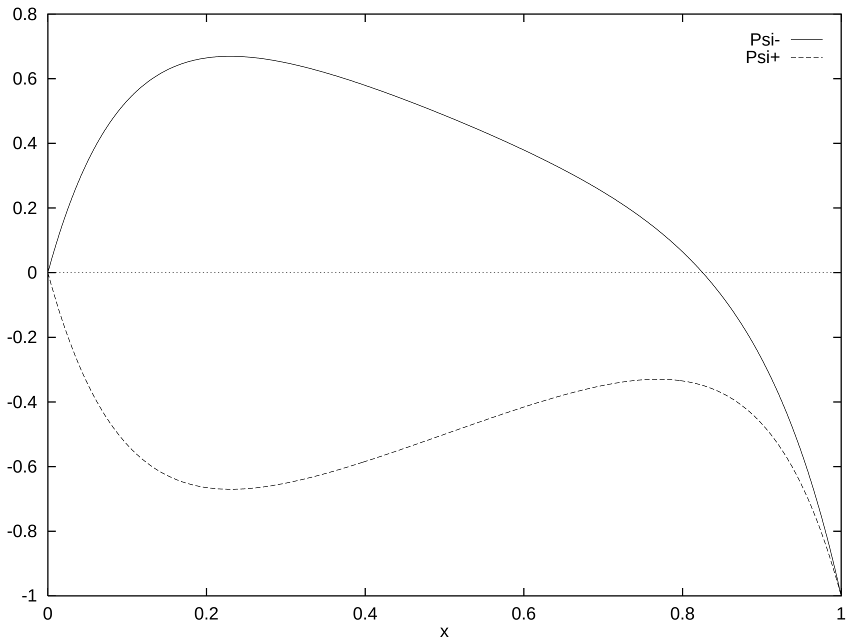


Figure 2-2: Adjoints $\psi_H^{\pm}$ $(H = h = 10^{-3})$

Figure 2-3 shows the bounds obtained. Three observations can be drawn from this graph. The first one is that the outputs computed on the coarse grids are very close to the output on the fine grid, since the curves seem to be on top of each other. The second one is that the bounds computed on the coarse grid estimate the "true" output within approximately 20%, even for large values of H (e.g. $H = 0.1$). The third one concerns the predictor. Like the coarse grid output, this predictor is so close to the h-mesh output that it becomes impossible to distinguish one curve from the other in the plot.

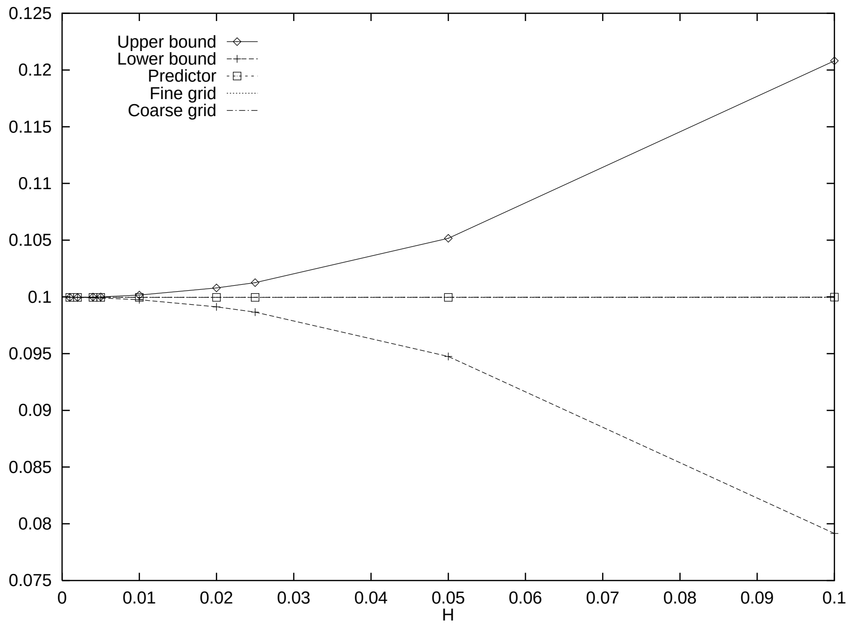


Figure 2-3: Bounds for the output : Average Value of the Solution

A closer look at the neighborhood of the fine grid output shows that the predictor is actually closer to the h-mesh output than the H-mesh output (see Figure 2-4). Therefore, the predictor estimates the "true" output better than the coarse grid.

To examine the convergence to the "true" output as a function of H, the rate of convergence r of the upper bound is defined as : $s_{UB}(H) - s_h = O\left(H^r\right)$. One has,

$$\log|s_{UB}(H) - s_h| = r\ \log H + \text{constant} \tag{2.26}$$

Therefore, the rate of convergence of the upper bound is given by the slope of a graph of $\log|s_{UB}(H) - s_h|$ as a function of $\log H$. The same reasoning holds for the lower bound, the predictor and the coarse grid output. Figure 2-5 presents plots showing the convergence rates of the upper and lower bounds as well as those of the predictor and the coarse grid output. The plot of the logarithm of the difference between the upper and the lower bounds is also represented. Some observations can be drawn from this graph. First, there is a large difference between the error bounds and the predictor error. Second, the errors corresponding to $H = h$ have of course not been shown, since by definition they are equal to 0. Third and final, the slope of the lines is equal to 2, which confirms the prediction of [5] according to which the convergence of the bounds is $O(H^2)$. This second order convergence

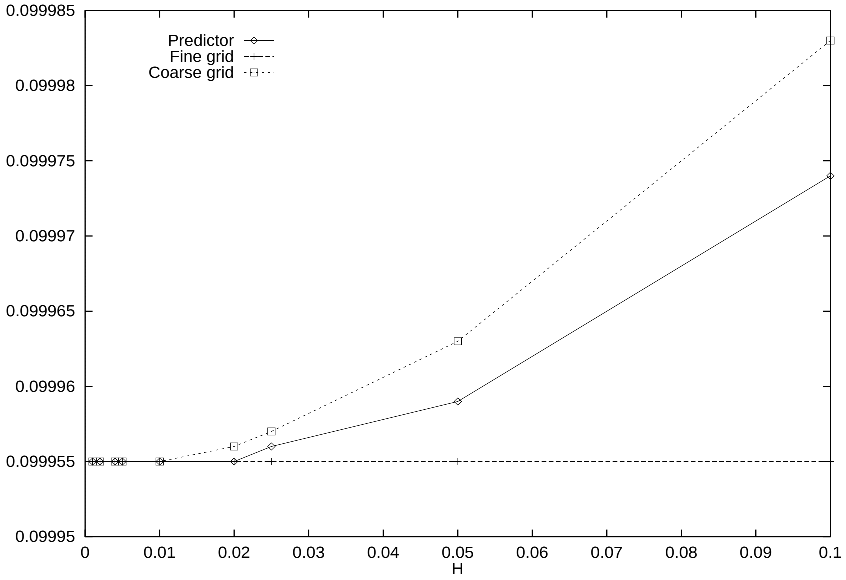


Figure 2-4: Predictor for the output : Average Value of the Solution

of the coarse grid output is characteristic of linear finite elements, at least as long as the problem is elliptic and the solution is sufficiently regular.

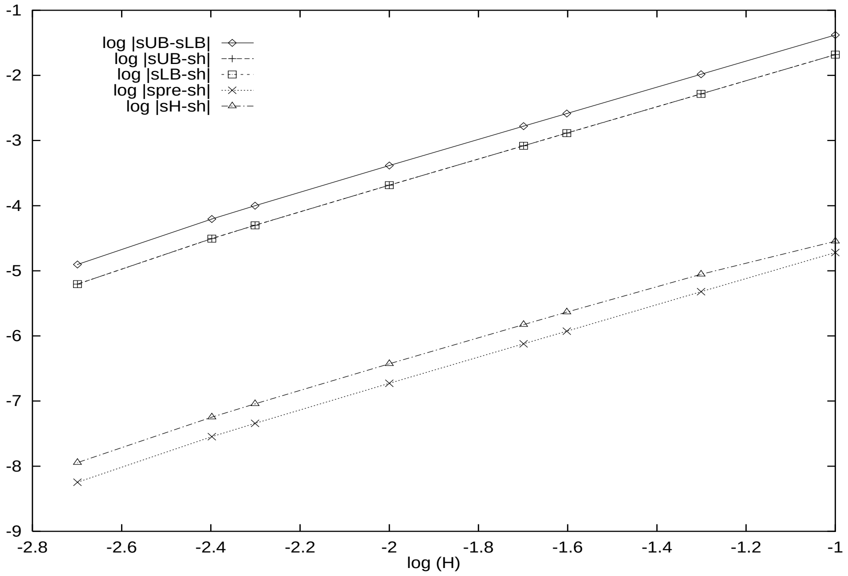
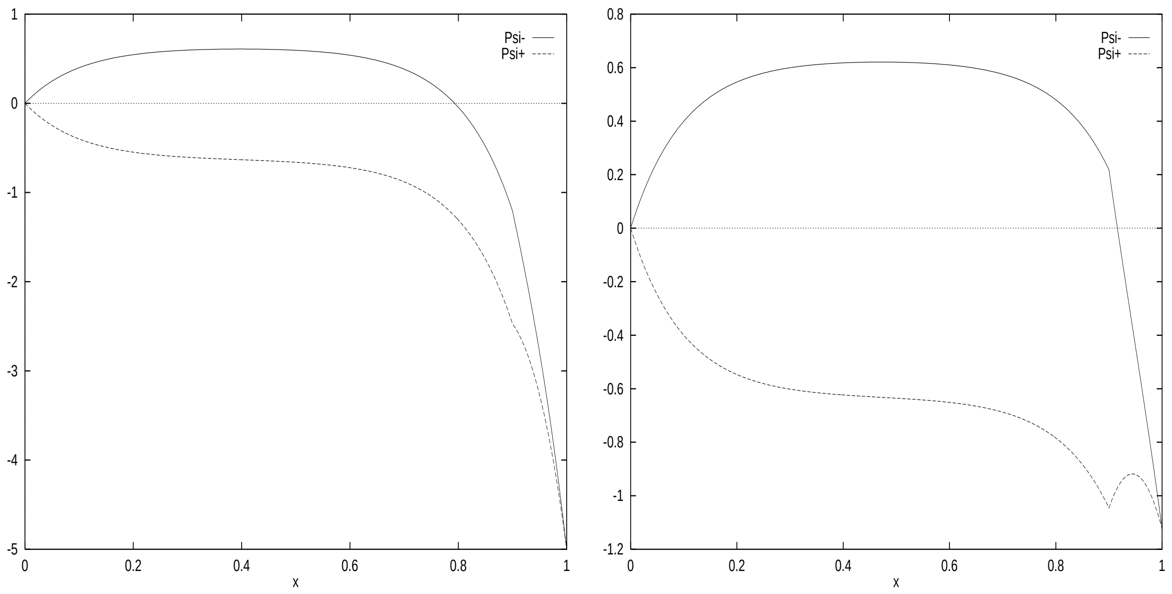


Figure 2-5: Convergence of the bounds for the output : Average Value of the Solution

2.4.2 Second Output : Pointwise value of the solution

In this subsection, we consider the value of the solution at point $\overline{x} = 0.9$, i.e. $s = u(0.9)$, as the output of interest. The discretizations are such that $x = 0.9$ corresponds to a grid point in all meshes. In this case, the optimal value of the stabilization parameter κ^* is not 1 anymore, but converges to a finite value (approximately 1.124) as H converges to h. Because this value is very close to 1, we present the results obtained for $\kappa = 5$ and $\kappa = \kappa^*$ to highlight the improvement of the bounds computed.

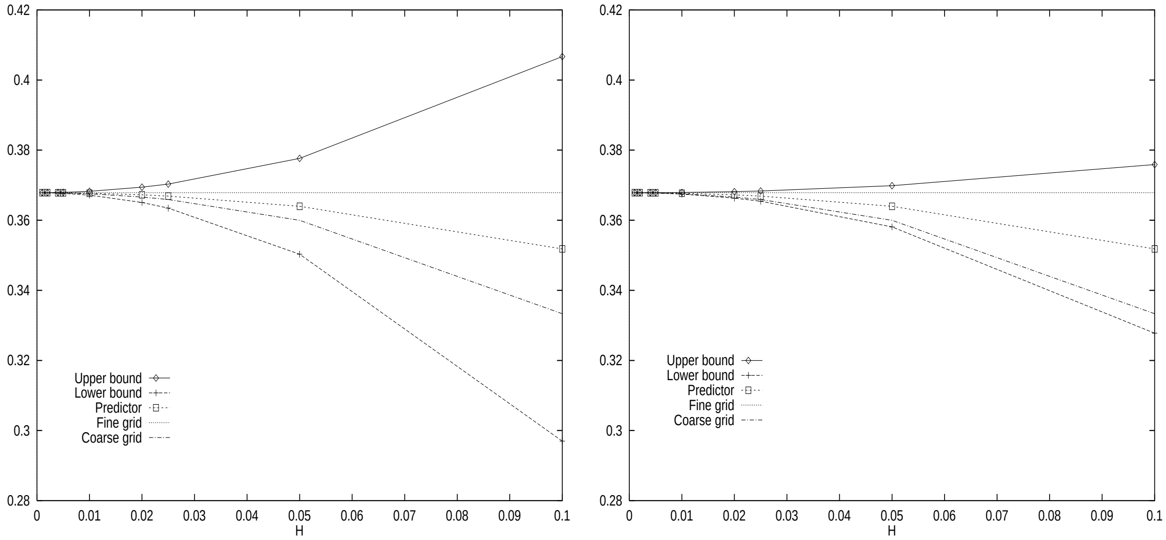
Figure 2-6 presents one characteristic feature of the chosen output : the first derivative of the adjoint shows a discontinuity, which occurs at the point $\overline{x} = 0.9$ where the solution is evaluated.



(a) $\kappa = 5$ (b) $\kappa = \kappa^*$

Figure 2-6: Adjoints for the output : Pointwise Value ($H = h = 10^{-3}$)

The improvement on the bounds between the case where $\kappa = 5$ and the one where κ takes its optimal value is significant, as shown on Figure 2-7. In particular, while the lower bound is below the output of the H-mesh in both cases, it is much closer to the fine grid output in the optimized case. Two other observations can be drawn from this figure. First, the bounds obtained on the coarsest grids give the "true" output within 20% for $\kappa = 5$, and within 10% for $\kappa = \kappa^*$, which is even better than for the average output. The bounds are thus very sharp. Second, although the coarse grid output is better than either of the



(a) $\kappa = 5$

(b) $\kappa = \kappa^*$

Figure 2-7: Bounds for the output : Pointwise Value

bounds, the predictor/estimator is significantly better than the coarse grid output. As expected, this predictor is also independant of κ.

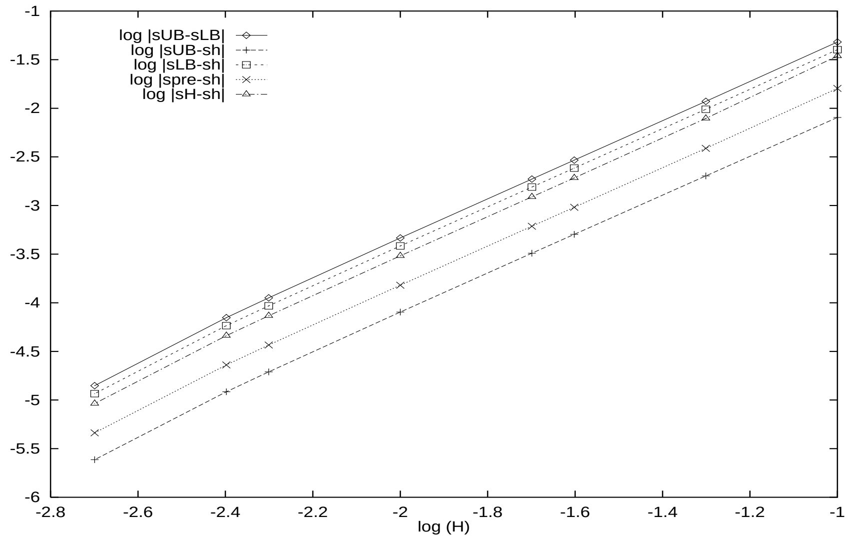


Figure 2-8: Convergence of the bounds for the output : Pointwise Value

The convergence of the bounds with the stabilization parameter κ optimized is given on Figure 2-8. As in the case of the average output, a second order convergence is numerically observed.

2.5 Conclusion

Three aspects of the Hierarchical Bounds Method have been highlighted in this chapter. First, the numerical results show that the bounds for the output computed on the fine grid are fairly sharp, and that the convergence is in $O(H^2)$ (second order). Second, although, in the case of the outputs considered so far, the coarse grid gives a better approximation than the bounds, the average value of the bounds (predictor) is still better than the coarse grid output in the sense that it is closer to the "true" output than the output computed on the H-mesh. Third, the improvement brought by the optimization of the stabilization parameter has been illustrated numerically.

Chapter 3

The Convection Problem

3.1 Introduction

This chapter considers the application of the Hierarchical Bounds Method to the linear pure convection equation. This equation is of interest for two reasons. First, the direct discretization of this equation with a Galerkin finite element method leads to a skew-symmetric matrix (hence with a zero symmetric part) and in certain cases to solutions with unphysical oscillations. Second, the convection equation is a first order ordinary differential equation that requires only one Dirichlet boundary condition.

In this chapter, we present a modification of the Galerkin procedure that allows for numerical solutions without oscillations to be computed. We then apply the theory developed in the first chapter, suitably adapted to this problem. We consider a problem in which the convection speed is from left to right and where a Dirichlet condition is applied at $x = 0$ while the solution at $x = 1$ is unknown.

The first step consists of finding a new formulation of the problem that eliminates the oscillations and allows for the computation of the solution at $x = 1$, yielding an "augmented" matrix L of dimensions $(n+1) \times (n+1)$. In the second step, a variation of the scheme leads to the *algebraic* computation of the boundary conditions for the adjoint $\psi_H^{\pm}$ at 0. This modification is necessary because, by definition, the adjoint is solution of the dual convection problem that involves a boundary condition at $x = 1$. This results in augmenting the dimensions of the matrix once again to $(n+2) \times (n+2)$.

3.2 Formulation of the problem

The purely convective problem can be written as :

$$\frac{du}{dx} = g(x) \tag{3.1}$$

$$u(0) = U_0 \tag{3.2}$$

where g is a function of the space variable x. For simple enough functions g, analytic solutions can be obtained to compare with the results of the numerical schemes.

In certain cases (in particular when g is not continuous), a direct Galerkin finite element method fails to give an acceptable solution (presence of unphysical oscillations). A possible way of avoiding this numerical difficulty consists of using a simple Taylor-Galerkin approach [11]. The basic idea of this approach is to introduce an artificial time-dependence :

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = g \tag{3.3}$$

and compute the solution to problem (3.1) as the steady-state solution of (3.3).

$$u_t = -u_x + g \tag{3.4}$$

$$u_{tt} = -u_{tx} \tag{3.5}$$

$$= u_{xx} - g_x \tag{3.6}$$

since g does not depend on the time variable t. $u(x, t + \Delta t)$ can be expanded in a Taylor series, and, keeping the first terms up to the second order, one gets :

$$u(x, t + \Delta t) = u(x, t) + \Delta t \, u_t(x, t) + \frac{\Delta t^2}{2} u_{tt}(x, t) + O(\Delta t^2) \tag{3.7}$$

$$= u(x, t) + \Delta t \, (-u_x + g) + \frac{\Delta t^2}{2} (u_{xx} - g_x) + O(\Delta t^2) \tag{3.8}$$

The solution of the original convection problem is then the steady-state solution of this differential equation. One thus writes :

$$u(x, t + \Delta t) = u(x, t) \tag{3.9}$$

The new scheme then consists of solving the equation

$$-\tau\, u_{xx} + u_x = g - \tau g_x \tag{3.10}$$

with $\tau = \Delta t/2$. Applying a CFL condition to the equation, one gets, for $\delta \in \{h, H\}$

$$\tau = \frac{\delta}{2} \tag{3.11}$$

A convection-diffusion problem can be recognized in (3.10), except that, in the Taylor-Galerkin formulation of the convection equation, the diffusion term goes to 0 as the size of the space discretization goes to 0. One should thus be able to apply a straight Galerkin finite element method to (3.10), just like in the case of the convection-diffusion problem.

Before performing the discretization, one difficulty must be solved concerning $u(1)$. Because the boundary condition on u at $x = 1$ is not specified in the original problem, the value of the solution at this point must be computed algebraically. $u(1) = u(x_{n+1})$ must therefore be considered as the $(n+1)$st unknown of the problem.

A weak form for (3.10) can now be written :

$$\int_0^1 (\tau\, w_x\, u_x + w\, u_x)\, dx = \int_0^1 (w\, g + \tau\, w_x\, g)\, dx \tag{3.12}$$

The output considered for this problem is the pointwise value of the solution u at point $x_{n+1} = 1$. This output offers a direct way of checking the convergence of the scheme to the exact value at 1 as the space discretization becomes small.

Using the Galerkin method with piecewise linear approximations leads to a matrix of the form :

$$L = \begin{pmatrix} \frac{2\tau}{\delta} & -\frac{\tau}{\delta}+\frac{1}{2} & 0 & \cdots & \cdots & 0 \\ -\frac{\tau}{\delta}-\frac{1}{2} & \frac{2\tau}{\delta} & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \frac{2\tau}{\delta} & -\frac{\tau}{\delta}+\frac{1}{2} \\ 0 & \cdots & \cdots & 0 & -\frac{\tau}{\delta}-\frac{1}{2} & \frac{\tau}{\delta}+\frac{1}{2} \end{pmatrix}$$

hence

$$L = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & \ddots & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix}$$

where L is of dimensions $(n+1) \times (n+1)$.

3.3 A New Formulation for the Adjoint

In order to be able to determine algebraically the values of the adjoints at the boundaries, the numerical algorithm is modified so that the Dirichlet boundary condition is imposed through the variational statement.

3.3.1 Additive term

Equation (3.12) can be written as :

$$\int_0^1 \left[w\,(u_x - g) + \tau\, w_x\,(u_x - g) \right] dx = 0$$

for all w and u equal to zero at $x = 0$. A modified scheme is obtained by allowing $u(0)$ to take any value and requiring :

$$\int_0^1 \left[w\,(u_x - g) + \tau\, w_x\,(u_x - g) \right] dx + w(0)\,u(0) = 0 \tag{3.13}$$

for all w and u (with no condition on u in 0). After integration by parts, one gets :

$$\int_0^1 \left[w\,(u_x - g) - \tau\, w\,(u_x - g)_x \right] dx + \left[\tau\, w\,(u_x - g) \right]_0^1 + w(0)\,u(0) = 0 \tag{3.14}$$

This is valid for all values of $w(0)$ and $w(1)$, so that the natural boundary conditions for the solution of this problem are :

$$u(0) - \tau\,(u_x - g)(0) = 0 \tag{3.15}$$

$$(u_x - g)(1) = 0 \tag{3.16}$$

The solution of the original convection problem (3.1)–(3.2) satisfies (3.15)–(3.16) as well as (3.14), but these boundary conditions are only used to check the validity of the solution given by the scheme. In the process, the initial single Dirichlet condition has been transformed into a pair of Neumann-Dirichlet conditions.

The absence of boundary conditions implies that, instead of looking for solutions of (3.13) in $\mathcal{H}^1_E(\mathcal{D})$ like in the general theory, we shall look for $u \in \mathcal{H}^1(\mathcal{D})$. The continuous-piecewise polynomial set X_E therefore needs to be redifined as :

$$X_E = \text{span}\,\{\varphi_0(x), \ldots, \varphi_{n+1}(x)\} \tag{3.17}$$

Moreover, because of the second term in the integral in (3.13), $g \in \mathcal{H}^{-1}(\mathcal{D})$ is not sufficient to ensure the existence of this integral. g must be such that the integral of the product $(w_x\, g)$ is defined, even for $w = \varphi_0$ or $w = \varphi_{n+1}$. Therefore, g is restricted to a subset of $\mathcal{H}^{-1}(\mathcal{D})$. The forcing functions chosen for the numerical tests do not raise any difficulty from that point of view.

The resulting system has dimensions $(n+2) \times (n+2)$. Three new terms appear in this matrix. First, a 1 appears in the upper left corner because of the new term $w(0)\, u(0)$ ($L_{00} = 1 + \frac{\tau}{\delta} - \frac{1}{2}$). The other two are linked to the addition of $u_0 = u(0)$ as a new unknown : the term previously containing the boundary condition is taken back to the left-hand side of the equation ($L_{10} = -\frac{\tau}{\delta} - \frac{1}{2}$) as well as the term expressing the dependance of u_1 on this boundary condition ($L_{01} = -\frac{\tau}{\delta} + \frac{1}{2}$).

The resulting matrix is :

$$L = \begin{pmatrix} 1+\dfrac{\tau}{\delta}-\dfrac{1}{2} & -\dfrac{\tau}{\delta}+\dfrac{1}{2} & 0 & \cdots & \cdots & 0 \\ -\dfrac{\tau}{\delta}-\dfrac{1}{2} & \dfrac{2\tau}{\delta} & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \dfrac{2\tau}{\delta} & -\dfrac{\tau}{\delta}+\dfrac{1}{2} \\ 0 & \cdots & \cdots & 0 & -\dfrac{\tau}{\delta}-\dfrac{1}{2} & \dfrac{\tau}{\delta}+\dfrac{1}{2} \end{pmatrix}$$

i.e.

$$L = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & \ddots & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix}$$

3.3.2 Natural Boundary Conditions for the Adjoint

Although they are computed algebraically, natural boundary conditions for the adjoint can be determined analytically from the continuous problem. The continuous Lagrangian corresponding to the new scheme can be written :

$$\begin{aligned} \mathcal{L}(v,\mu) &= \kappa \left\{ \left[\frac{1}{2}v^2\right]_0^1 - \int_0^1 v\,g\,dx + \tau \int_0^1 v_x^2\,dx - \tau \int_0^1 v_x\,g\,dx + v(0)^2 \right\} \pm l(v) \\ &\quad + \int_0^1 \left[\mu(v_x - g) + \tau\,\mu_x(v_x - g)\right]\,dx + \mu(0)\,v(0) \qquad (3.18) \\ &= \kappa \left\{ \left[\frac{1}{2}v^2\right]_0^1 - \int_0^1 v\,g\,dx + \tau \int_0^1 v_x^2\,dx - \tau \int_0^1 v_x\,g\,dx + v(0)^2 \right\} \pm l(v) \\ &\quad - \int_0^1 \left[v\,\mu_x + \mu\,g + \tau\,\left(v\,\mu_{xx} + \mu_x\,g\right)\right]\,dx + \mu(1)\,v(1) + \tau\,\left[\mu_x\,v\right]_0^1 \qquad (3.19) \end{aligned}$$

Taking the first variation of the Lagrangian with respect to variations $w(x) = v(x) - u(x)$ and considering the pointwise value of u in 1 as the output one obtains :

$$\begin{aligned} &\kappa \left\{ [w\,u]_0^1 - \int_0^1 \left[w\,g + \tau\,\left(-2w_x\,u_x + w_x\,g\right)\right]\,dx + 2w(0)\,u(0) \right\} \pm w(1) \\ &\qquad - \int_0^1 \left[w\,\left(\mu_x + \tau\,\mu_{xx}\right)\right]\,dx + \mu(1)\,w(1) + \tau\,\left[\mu_x\,w\right]_0^1 \;=\; 0 \qquad (3.20) \end{aligned}$$

The coefficients of $w(0)$ and $w(1)$ are set equal to zero, since all variations around 0 are allowed. This gives natural boundary conditions for the adjoint :

$$\mu(1) + \tau\,\mu_x(1) = -\left(\kappa\,u(1) \pm 1\right) \qquad (3.21)$$

$$\tau\,\mu_x(0) = \kappa\,u(0) \qquad (3.22)$$

Again, these conditions are not used directly in the numerical scheme, but are useful as a check of the numerical results.

3.4 Optimal Scaling

3.4.1 Optimal Value for κ

To compute the optimal value of κ in the case of the pointwise value output, we follow the procedure outlined in Chapter 1 and write :

$$\psi_H^{0\pm} = \mp (L_H^T)^{-1} \ell_H \tag{3.23}$$

$$\psi_H^{1\pm} = (L_H^T)^{-1} [f_H - A_H\, u_H] \tag{3.24}$$

Given the form of the matrix L_H (previously noted L for simplicity of notations), the inverse of its transpose is simple to find :

$$\left(L_H^T\right)^{-1} = \begin{pmatrix} 1 & \cdots & \cdots & 1 \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix} \tag{3.25}$$

With the output vector $\ell_H = [0 \cdots 0\, 1]^T$ (column vector of size $(N+2)$), the constant part of the adjoint for the coarse grid becomes :

$$\psi_H^{0\pm} = \mp \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{3.26}$$

This vector is then interpolated on the "truth" mesh. The values of the components of $\psi_H^{0\pm}$ being all equal, the components of the interpolated vector $\hat{\mu}_h^{0\pm}$ are also all equal :

$$\hat{\mu}_h^{0\pm} = \mp \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{3.27}$$

The interpolation process therefore does not change the form of the constant part of the adjoint, which means that, on the fine grid, one has :

$$\hat{\mu}_h^{0\pm} = \mp (L_h^T)^{-1} \ell_h \tag{3.28}$$

Equation (3.28) immediately shows that

$$p^{\pm} = L_h^T \hat{\mu}_h^{0\pm} \pm \ell_h = 0 \tag{3.29}$$

which also implies that

$$\alpha^{\pm} = p^{\pm T} A_h^{-1} p^{\pm} = 0 \tag{3.30}$$

The optimal value of κ given by (1.76) is thus equal to 0.

The natural boundary conditions for the adjoint given by (3.21) and (3.22) then become :

$$\begin{aligned} \hat{\mu}^{\pm}(1) + \tau \, \hat{\mu}_x^{\pm}(1) &= \mp 1 && (3.31) \\ \hat{\mu}_x^{\pm}(0) &= 0 && (3.32) \end{aligned}$$

Equation (3.28) and $\kappa^* = 0$ imply that $\left(\hat{\mu}_h^{\pm}\right)_i = \left(\hat{\mu}_h^{0\pm}\right)_i = \mp 1 \;\forall\, i$, so $\left(\hat{\mu}_h^{\pm}\right)_x = 0$ and thus, both equations (3.31) and (3.32) are satisfied.

3.4.2 Behaviour of the Bounds as κ goes to $\kappa^* = 0$

In this section, we investigate the limiting case where κ tends to zero. It is not clear that for $\kappa = 0$ the formulation presented applies as the Lagrangian is not strictly convex.

First, from (3.23) and (3.24), one has

$$\begin{aligned} \hat{\mu}_h^{0+} &= -\hat{\mu}_h^{0-} && (3.33) \\ \hat{\mu}_h^{1+} &= \hat{\mu}_h^{1-} && (3.34) \end{aligned}$$

all terms being independent of κ. $\hat{u}_h^{\pm}$ is then defined by

$$L_h^T \hat{\mu}_h^{\pm} = L_h^T \left(\hat{\mu}_h^{0+} + \kappa \, \hat{\mu}_h^{1+}\right) = \mp l_h - \kappa \left(A_h \, \hat{u}_h^{\pm} - f_h\right) \tag{3.35}$$

which, using (3.28) leads to

$$A_h \, \hat{u}_h^{\pm} = -L_h^T \hat{\mu}_h^{1\pm} + f_h \tag{3.36}$$

$\hat{u}_h^{\pm}$ is therefore independent of κ and one has

$$\hat{u}_h^+ = \hat{u}_h^- = \hat{u}_h \tag{3.37}$$

Furthermore, the bounds $s_{UB}(H)$ and $s_{LB}(H)$ can be computed by

$$\begin{aligned} s_{UB}(H) &= \frac{\kappa}{2}\,\hat{u}_h^T A_h\,\hat{u}_h + \hat{\mu}_h^{-T} f_h & (3.38)\\ s_{LB}(H) &= -\frac{\kappa}{2}\,\hat{u}_h^T A_h\,\hat{u}_h - \hat{\mu}_h^{+T} f_h & (3.39)\end{aligned}$$

Hence,

$$\begin{aligned} s_{UB}(H) &= \frac{\kappa}{2}\left[\hat{u}_h^T A_h\,\hat{u}_h + 2\hat{\mu}_h^{1-T} f_h\right] + \hat{\mu}_h^{0-T} f_h & (3.40)\\ s_{LB}(H) &= -\frac{\kappa}{2}\left[\hat{u}_h^T A_h\,\hat{u}_h + 2\hat{\mu}_h^{1+T} f_h\right] - \hat{\mu}_h^{0+T} f_h & (3.41)\end{aligned}$$

In both formulas, the factor in brackets and the second term of the right-hand side are independent of κ. This establishes the linear convergence of the bounds to a finite limit when κ converges to 0. The limits are :

$$\begin{aligned} \lim_{\kappa\to 0} s_{UB}(H) &= \hat{\mu}_h^{0-T} f_h & (3.42)\\ \lim_{\kappa\to 0} s_{LB}(H) &= -\hat{\mu}_h^{0+T} f_h & (3.43)\end{aligned}$$

Because of (3.28) and (3.33),

$$\begin{aligned} \hat{\mu}_h^{0-T} f_h &= \left[\left(L_h^T\right)^{-1} \ell_h\right]^T f_h & \\ &= l_h^T L_h^{-1} f_h & \\ &= \ell_h^T u_h & (3.44)\\ \hat{\mu}_h^{0+T} f_h &= -\ell_h^T u_h & (3.45)\end{aligned}$$

where u_h is the solution obtained on the fine ("truth") grid (by definition, $L_h u_h = f_h$). Since, by definition, $\ell_h^T u_h$ is the output s_h computed on the fine grid, one finally finds :

$$\lim_{\kappa\to 0} s_{UB}(H) = \lim_{\kappa\to 0} s_{LB}(H) = s_h \tag{3.46}$$

In other words, for any fixed value of H, when κ goes to 0, the upper bound and the lower bound computed from the coarse grid both converge to the output computed on the fine grid.

Equation (3.46) is valid as long as $\psi_H^{0\pm}$ is a linear function of x. Indeed, the interpolation $\hat{\mu}_h^{0\pm}$ has then the same form as $\psi_H^{0\pm}$, and (3.28) is satisfied. Otherwise (in the case of the average output for the convection-diffusion problem, for instance), one should rather write

$$L_h^T \hat{\mu}_h^{0\pm} = \mp \ell_h + \varepsilon_h \tag{3.47}$$

with all the terms being independent of κ. Thus,

$$A_h \, \hat{u}_h^{\pm} = -L_h^T \hat{\mu}_h^{1\pm} + f_h - \frac{\varepsilon_h}{\kappa} \tag{3.48}$$

The solution $\hat{u}_h^{\pm}$ can be decomposed as $\hat{u}_h^{\pm} = \hat{u}_h^{0\pm} + \hat{u}_h^{1\pm}/\kappa$ and it immediately appears from (3.38) and (3.39) that one has :

$$\lim_{\kappa \to 0} s_{UB}(H) = -\lim_{\kappa \to 0} s_{LB}(H) = +\infty \tag{3.49}$$

This explains why 0 is not an optimal value for the scaling factor in general.

3.5 Numerical results

Different functions have been considered for the forcing function g. The first one is a "step function" equal to 1 between 0.4 and 0.6, and to 0 everywhere else. Figure 3-1 shows the oscillations obtained when a Galerkin finite element method is directly applied to the convection equation. This figure demonstrates the need for another formulation to avoid an oscillatory numerical solution. The presence of oscillations only on one side of the domain is linked to the fact that, in one dimension, the finite element method is equivalent to central finite differences everywhere, but with an artificial Neumann boundary condition imposed in 1.

The other two functions tested for the forcing term are $g(x) = x$ and $g(x) = \cos x$. The numerical results are presented for $\kappa = 1$ and for the optimum κ.

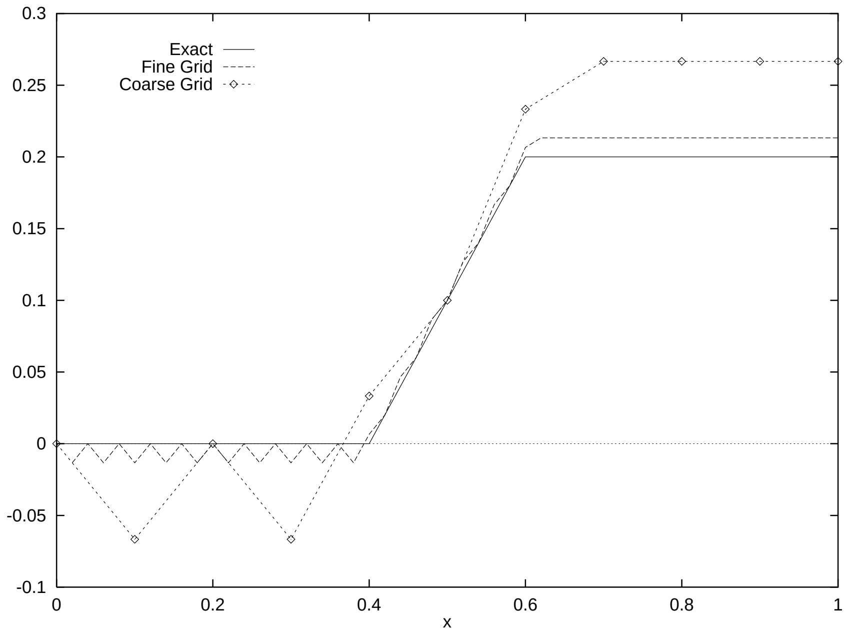
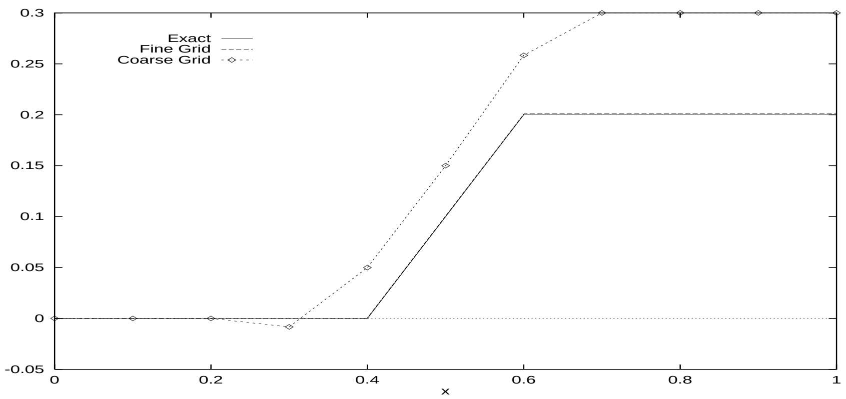


Figure 3-1: Finite Element Method directly applied to the convection equation

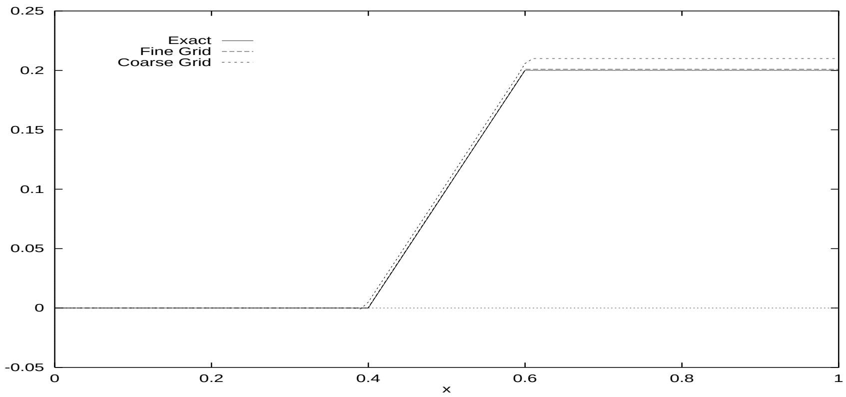
3.5.1 Results for a non-optimal stabilization parameter : $\kappa = 1$

The modifications proposed for the scheme lead to a numerical solution, which does not exhibit an oscillatory behaviour. The results obtained for different values of H are given in Figure 3-2.

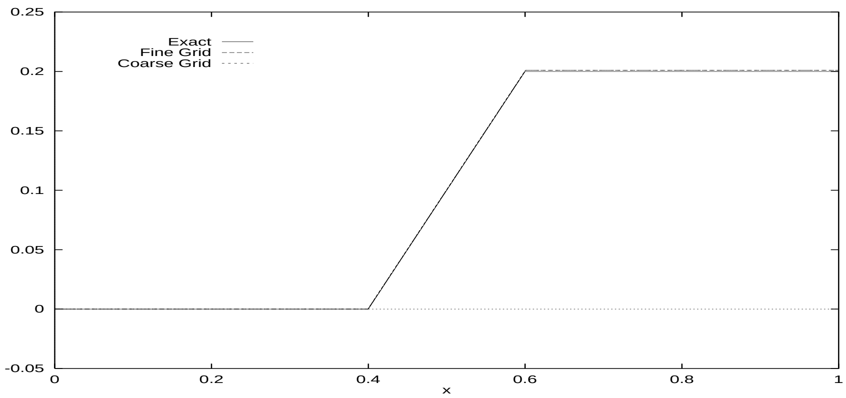
Although the error is rather large on the coarse mesh, it decays rapidly as the grid is refined. This error behaviour is characteristic of a non-continuous right-hand side g : when the function g is continuous, the solution computed by the modified scheme is very close to the exact solution, even for the coarsest grid ($H = 0.1$).



(a) $H = 0.1$



(b) $H = 0.01$



(c) $H = 10^{-3}$

Figure 3-2: Numerical solution, g is a step function, $h = 10^{-3}$

The computed adjoint has to satisfy the natural boundary conditions (3.21) and (3.22). To check this property, the adjoint $\psi_H^{\pm}$ is plotted on Figure 3-3. The following values can

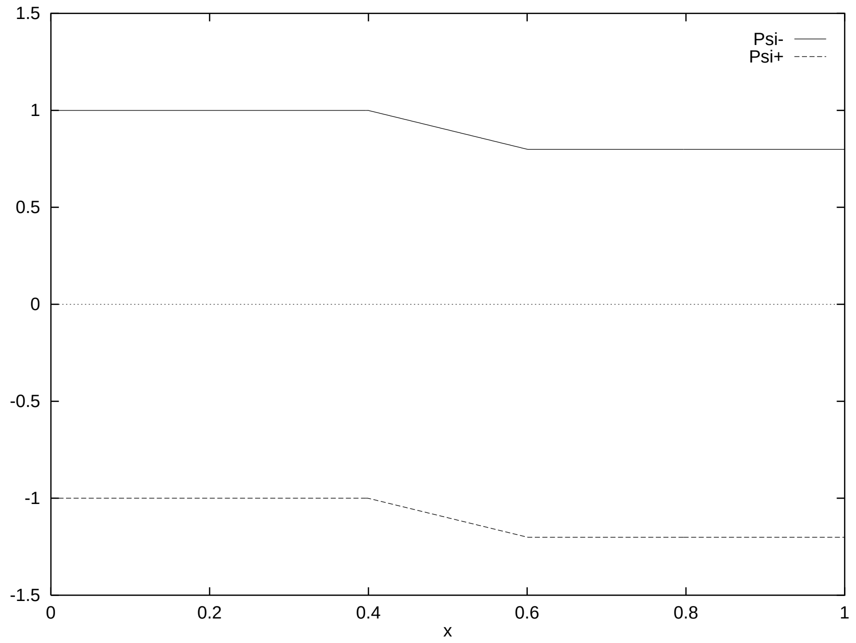


Figure 3-3: Adjoint for $\kappa = 1$, $H = h = 10^{-3}$

be determined from this figure :

$$
\begin{aligned}
\left(\psi_H^{\pm}\right)_x(0) &= 0 \\
\psi^{+}(1) &= -1.2 \\
\psi^{-}(1) &= 0.8 \\
\left(\psi_H^{\pm}\right)_x(1) &= 0
\end{aligned}
$$

so that with $u(0) = 0$ and $u(1) = 0.2$, the boundary conditions on the adjoint are immediately satisfied.

The bounds for the value of the fine grid solution at $x = 1$ are plotted on Figure 3-4. In this figure, $h = 10^{-3}$ and H varies from 0.1 down to 10^{-3}. Two conclusions can be drawn from this graph. First, the output computed on the coarse grid is closer to the "exact" output than both bounds, especially when the discretization step becomes very large. Second, the curves representing the bounds are symmetric with respect to the fine

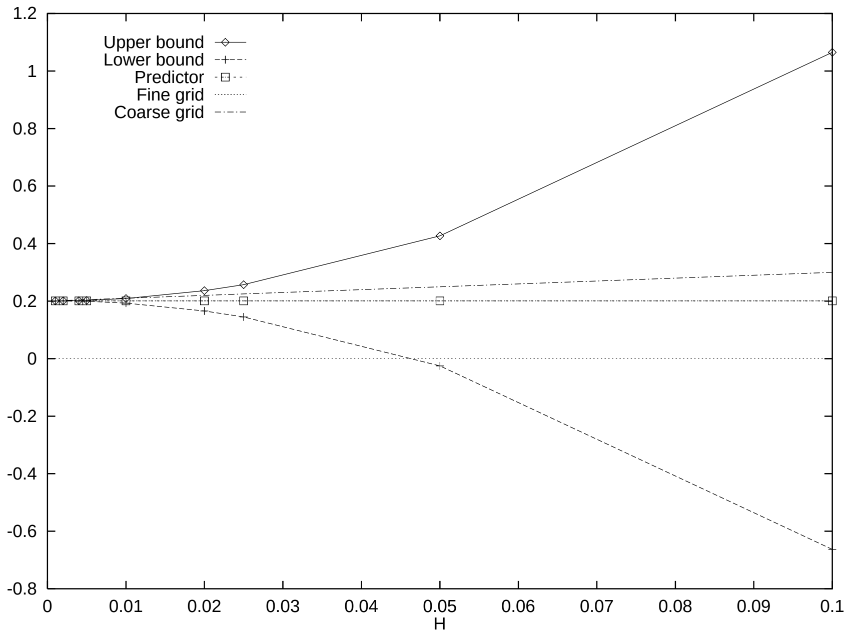


Figure 3-4: Bounds for $\kappa = 1$, $h = 10^{-3}$, $10^{-3} \leq H \leq 0.1$

grid output, so the average of the bounds, used as an estimation of the value of the output (predictor), matches exactly the output produced by the "truth" mesh. Again, we can verify that κ only affects the bounds and not the predicted value.

The convergence of the bounds is again given by a plot of the logarithm of the error as functions of the logarithm of the grid size. This graph is shown on Figure 3-5. Although, in the present case, the coarse grid output converges only linearly because of the discontinuity of the forcing term, the bounds still converge as $O(H^2)$. The error on the predictor/estimator has not been plotted because this error is equal to 0, the bounds being symmetric with respect to the fine grid output.

Four other functions have been tried for the forcing term g. The results obtained are given for two of them. First, we investigate the case of a linear forcing term : choose $g = x$ and solve $\{u_x = x \; \forall x \in \mathcal{D}, u(0) = 0\}$. The solutions obtained by the finite element method are given in Figure 3-6. One can immediately check that the nodal values of the solutions on both the coarse and the fine grids coincide with the exact solution ($u(x) = x^2/2$). This is to be expected, from the inspection of the resulting difference scheme.

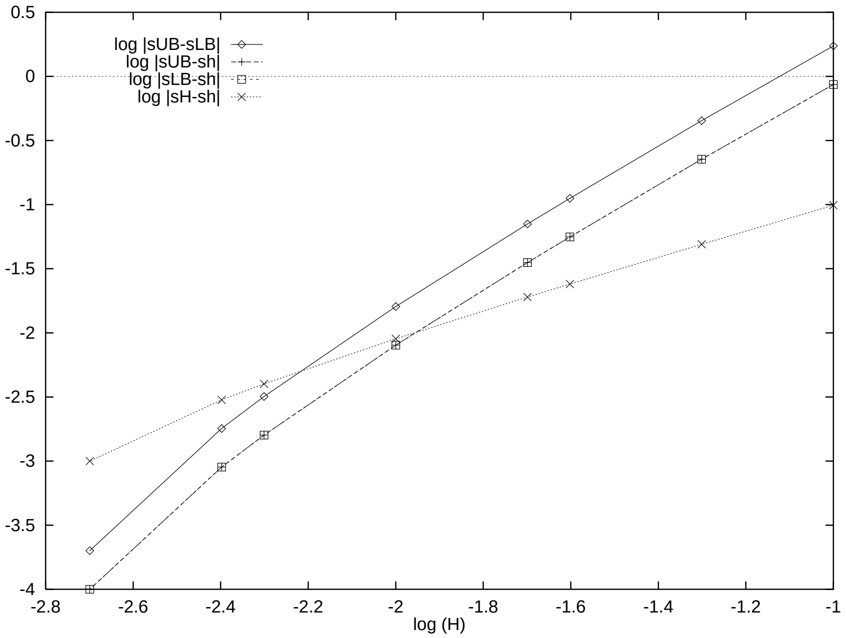


Figure 3-5: Convergence of the bounds for $u_x = g, u(0) = 0$ (step function), pointwise value output, $\kappa = 1$

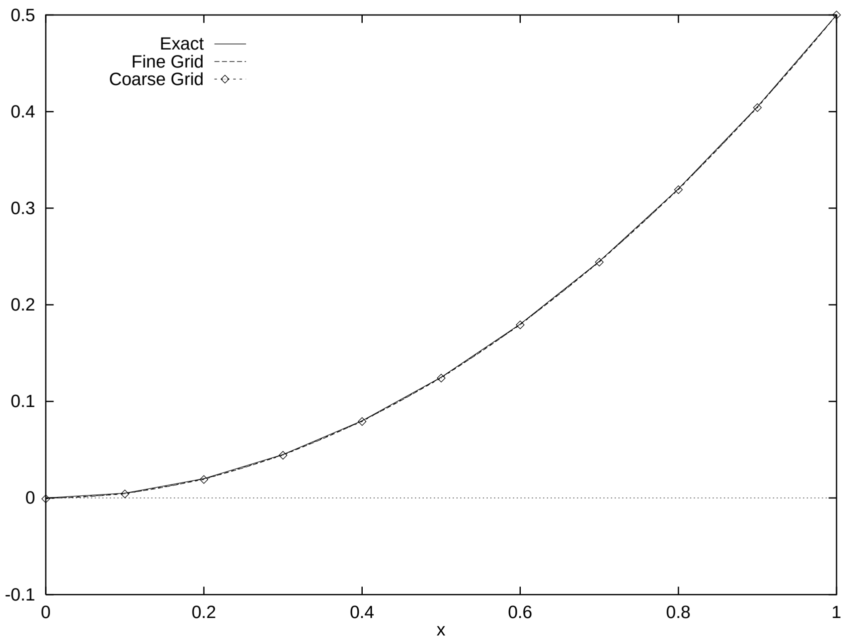


Figure 3-6: Solutions of $u_x = x, u(0) = 0$

The bounds computed are presented on Figure 3-7. Two observations can be drawn from this figure. First, the bounds appear to be very close to the fine grid output (even if they do not exactly match it) and the average of the bounds exactly matches the output computed on the h-mesh. The second observation concerns the sharpness of the bounds :

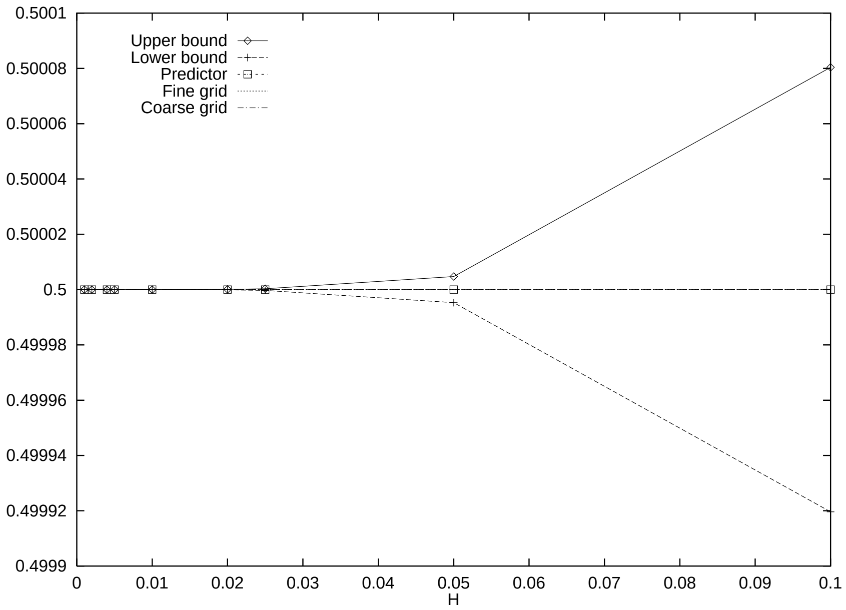


Figure 3-7: Bounds for $u_x = x, u(0) = 0$, pointwise value output, $\kappa = 1$

although the estimation obtained is not exact, the bounds remain within 0.016% of the exact solution.

The convergence of the bounds is now plotted on a log-log graph presented in Figure 3-8. This figure shows another feature of the bounds obtained in the purely convective case with a $\mathcal{C}^\infty$ forcing term : not only are the bounds very sharp, but they also converge very fast, a fourth order convergence is observed (i.e. $O\left(H^4\right)$). The errors on the coarse grid output and on the predictor are of the same order as machine precision. The bounds in this purely convective case with a linear forcing term are therefore sharper and converge faster than those obtained for the convection-diffusion model case.

Finally, we consider the case $g = \cos x$, i.e. the resolution of $u_x = \cos x, u(0) = 0$. The exact solution is $u(x) = \sin x$. The solutions given by the finite element method are shown in Figure 3-9. The forcing term being $\mathcal{C}^\infty$, the numerical solutions are much closer to the exact solution than when the forcing term is a step function. Also, the forcing term is not linear, so the coarse grid solution does not match the exact solution at the nodes anymore.

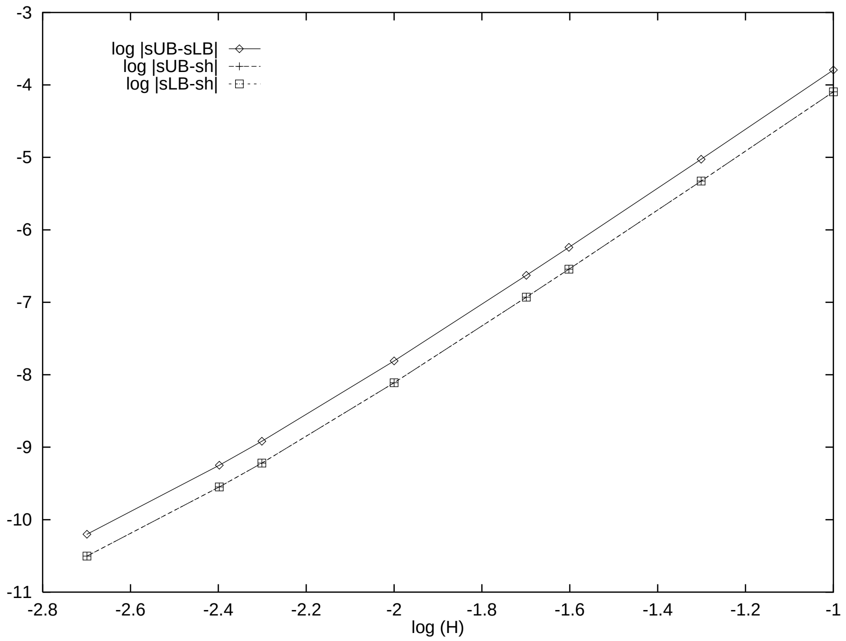


Figure 3-8: Convergence of the bounds for $u_x = x, u(0) = 0$, pointwise value output, $\kappa = 1$

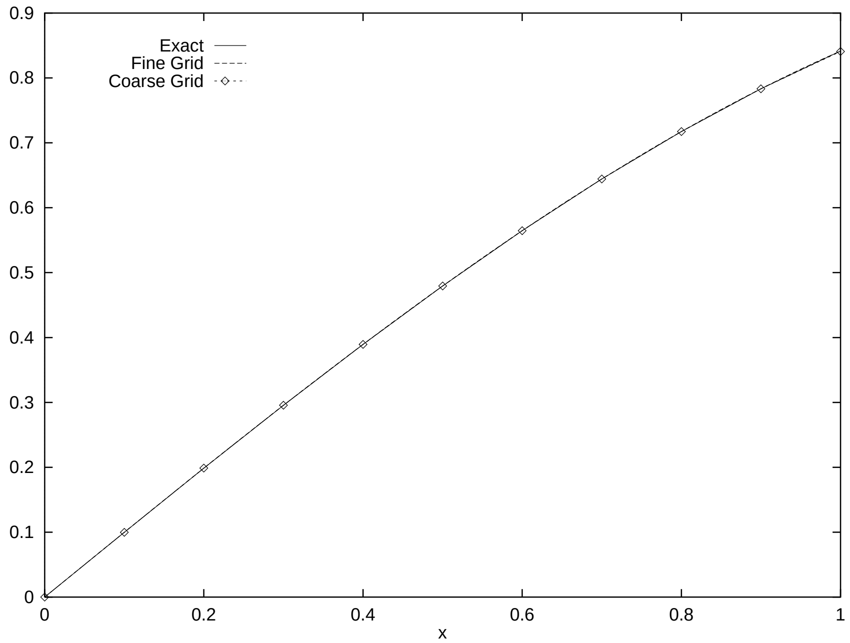


Figure 3-9: Solution of $u_x = \cos x, u(0) = 0$

The bounds obtained by the general method described in the previous sections are shown on Figure 3-10. Once again, the bounds are very sharp. In this case, they give even a better

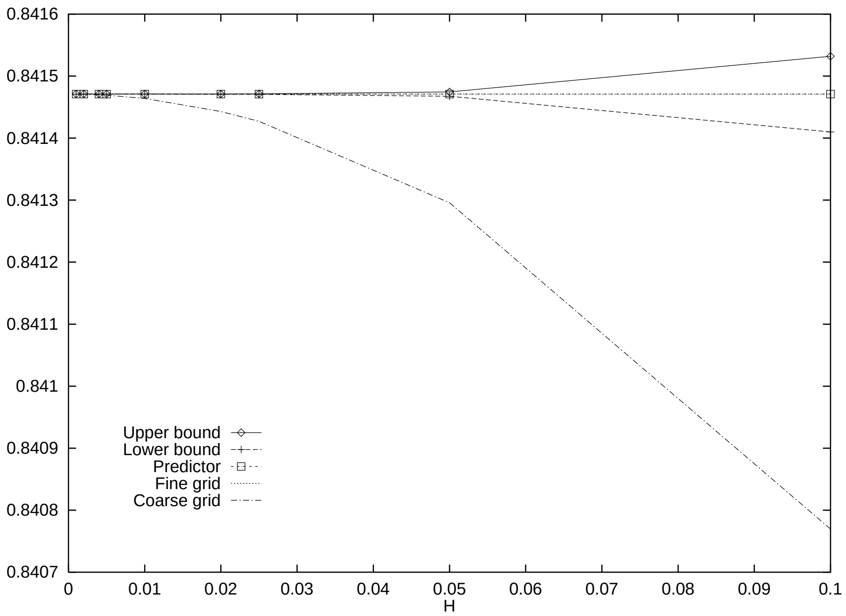


Figure 3-10: Bounds for $u_x = \cos x, u(0) = 0$, pointwise value output, $\kappa = 1$

estimation of the fine grid output than the coarse grid does. In other words, the bounds are closer to the "target" output than the coarse grid output. The convergence of the bounds is again $O\left(H^4\right)$, as shown on Figure 3-11. However, the convergence of the coarse grid output is still $O\left(H^2\right)$, which is usual for the linear finite element method.

In summary, one can say that in the case of the purely convective problem, when one chooses $\kappa = 1$ (which is not optimal), the method gives very good results, especially when the forcing term is continuous.

3.5.2 Optimization of the Stabilization Parameter : $\kappa^* = 0$

The results obtained for $\kappa = 1$ are already good, but the optimal value of the stabilizing factor is not 1 but 0, independantly of the forcing term g or of H. Let us see what happens numerically when $\kappa = 0$ is directly imposed for the case where g is the step function equal to 1 between 0.4 and 0.6 and to 0 everywhere else. The solutions remain unchanged, since κ comes into play only in the computation of the bounds. The bounds obtained are given on Figure 3-12. Apart from the coarse grid output that does not match the fine grid output, all the other curves are so close to each other that they cannot be distinguished on this plot.

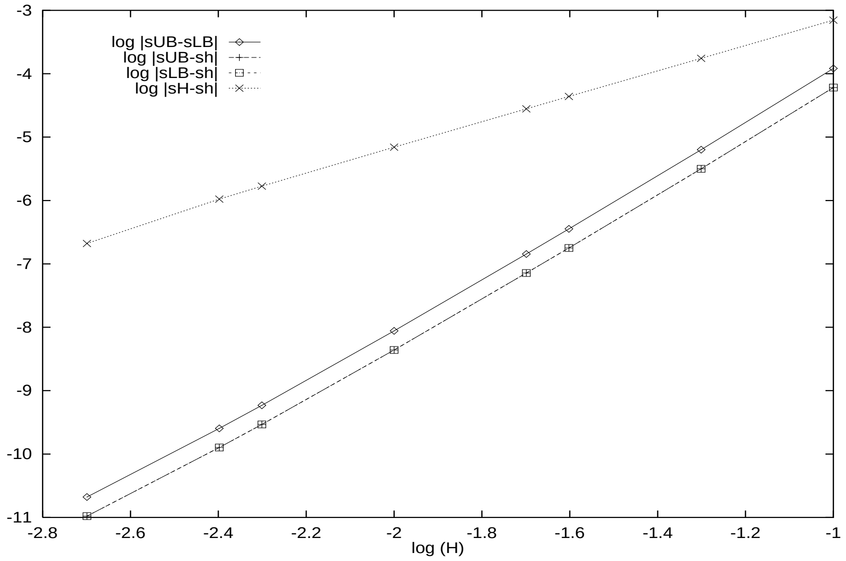


Figure 3-11: Convergence of the bounds for $u_x = \cos x, u(0) = 0$, pointwise value output, $\kappa = 1$

This is understandable : it has already been shown that they are theoretically supposed to match the “exact” output.

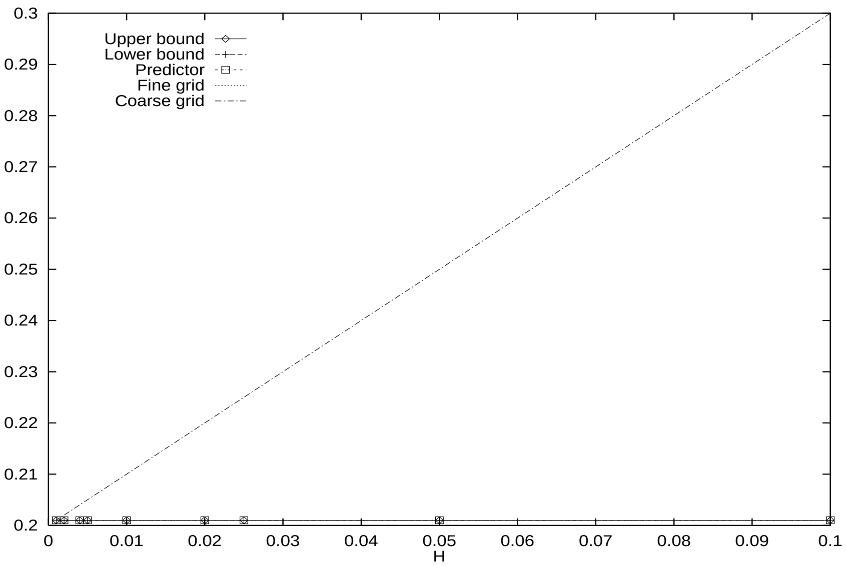


Figure 3-12: Bounds for $u_x = g, u(0) = 0$, $\kappa = \kappa^* = 0$, pointwise value output

Figure 3-13, which is the same graph, but with the coarse grid output removed, gives a better resolution, and it appears that, in perfect accordance with the theory, the bounds, as well as the predictor, match the fine grid output perfectly. This fine grid output is just

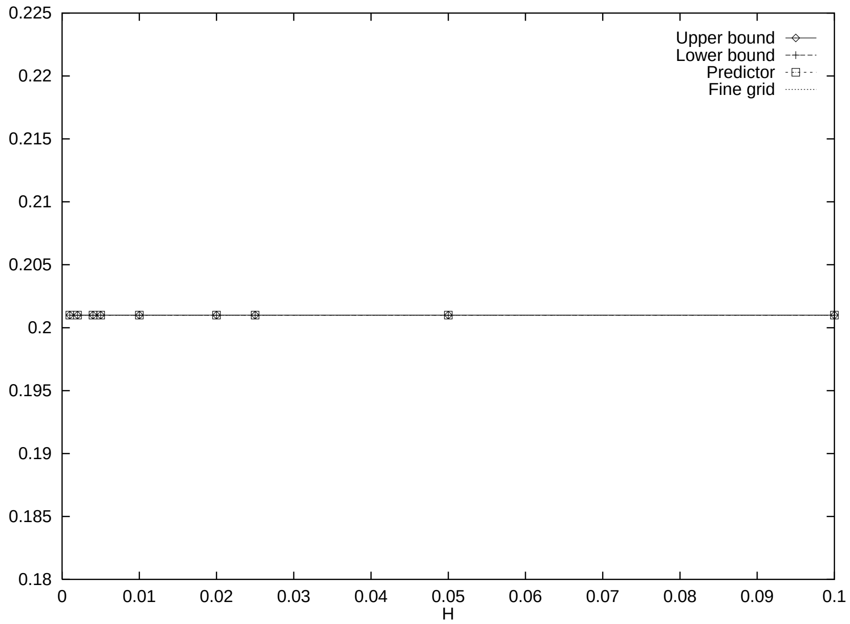


Figure 3-13: Bounds for $u_x = g, u(0) = 0$, $\kappa = \kappa^* = 0$, pointwise value output

above the exact value 0.2 because the point where the difference *is* negligible has not been reached with 1000 elements. Although this might seem surprisingly good, the plotting of the curves at higher resolutions (zooming around $y = 0.2$) gives exactly the same result. Furthermore, a look at the computed numerical values also reveals a perfect match between the fine grid output, the upper and lower bounds, and the predictor.

It is not clear that the theory applies directly with $\kappa = 0$, given that all the results were derived with the assumption that $\kappa \neq 0$. The question is thus to know what happens if one chooses $\kappa \neq 0$ and has this stabilization parameter tend to 0. Figure 3-14 shows that the bounds get closer and closer to the output obtained on the fine grid, which is a good indication of the consistency between the results and the theory. The upper and lower bounds have been plotted for $\kappa \in \{1\,;\, 0.5\,;\, 0.1\,;\, 10^{-2}\,;\, 10^{-3}\}$.

Figure 3-15 describes the evolution of the upper and lower bounds for H fixed equal to 0.1, κ varying from 1 down to 10^{-3}. This plot confirms that, for a fixed value of H, the upper and the lower bounds both converge to the "true" output when κ tends to 0, and that the convergence is linear, as predicted by the theory. This is why one can take $\kappa = 0$,

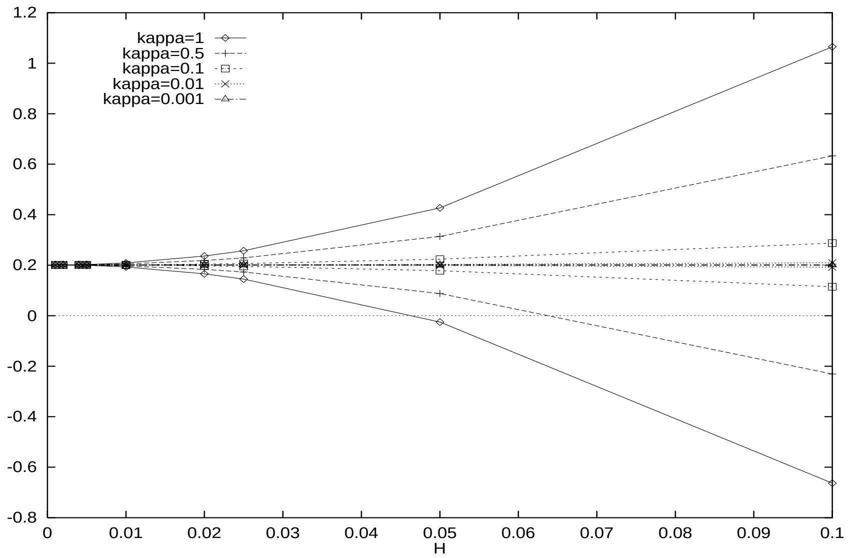


Figure 3-14: Bounds for $u_x = g, u(0) = 0$, κ converges to 0, pointwise value output

since the bounds present a regular behaviour near 0. In other words, when κ goes to 0, the bounds converge to the bounds obtained when $\kappa = 0$ is directly imposed in the equations.

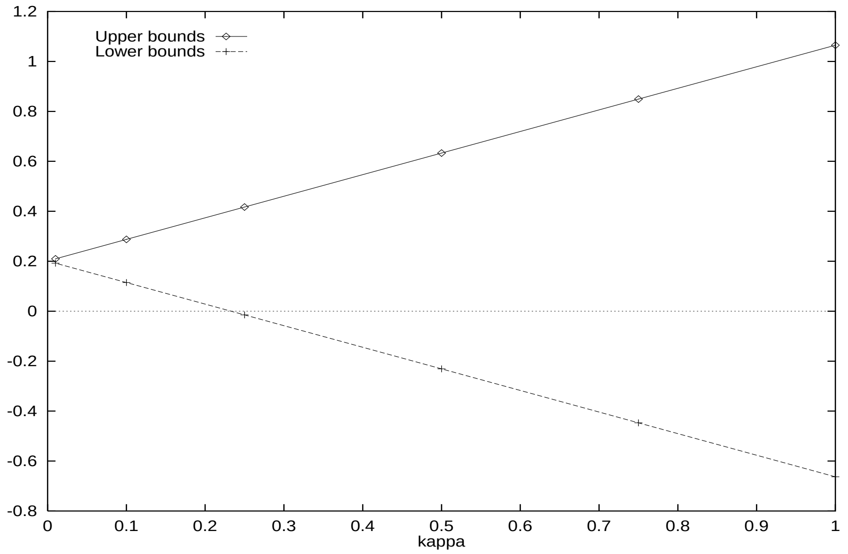


Figure 3-15: Bounds for $u_x = g, u(0) = 0$, κ converges to 0, $H = 0.1$ fixed, pointwise value output

3.6 Conclusion

Three important conclusions can be drawn from the results presented in this chapter. First, the Hierarchical Bounds Method can be adapted to first order differential equations, where only one boundary condition is available in the problem, and the numerical diffusion added to stabilize the numerical solution is sufficient to force the symmetric part of the discrete system matrix to be positive definite, so the Hierarchical Bounds Method can work.

Second, the boundary values of the adjoint are computed algebraically. Thanks to this property of the formulation, the natural boundary conditions need not be derived from the continuous problem, which is quite convenient, because the natural boundary conditions given by the continuous problem are not of the Dirichlet type, and are not very easy to implement.

Third, the bounds can be impressively sharp. In the case of the convection-diffusion equation, the coarsest grids give bounds within 20% of the "target" output (fine grid output). For the pure convection equation, the bounds are within less than 1% for the coarsest grids ($H = 0.1$) when the stabilization parameter is not optimal. What may be even more remarkable is that, even if the problem is solved only on a coarse grid, the optimization of the stabilization parameter produces bounds that match *exactly* the fine grid output, which is not the case for the convection-diffusion equation. This is of course particular to this problem, and not necessarily generalizable to other problems.

Chapter 4

Nonlinear Problem

In this chapter, the Hierarchical Bounds Method is extended to nonlinear problems. The model equation chosen is a scalar nonlinear equation derived from the one-dimensional steady Euler equations.

4.1 Governing Equations

The model problem considered is the steady flow of an inviscid fluid in a diverging nozzle. The problem can be modelled as one-dimensional, and it is thus governed by the 1D steady Euler equations :

$$\mathcal{F}_x + \mathcal{G} = 0 \quad 0 \leq x \leq 1 \tag{4.1}$$

where

$$\mathcal{F} = \begin{pmatrix} \rho\, u\, A \\ (\rho\, u^2 + p)\ A \\ (\rho\, E + p)\ u\, A \end{pmatrix}, \quad \mathcal{G} = \begin{pmatrix} 0 \\ -p\, A_x \\ 0 \end{pmatrix} \tag{4.2}$$

In (4.2), ρ is the density, u is the velocity, p is the pressure and A is the variable cross-section of the nozzle. Here, only the diverging part of the nozzle is modelled and A is assumed to be a given monotonically increasing and differentiable function. One also has

$$E = e + \frac{u^2}{2} \tag{4.3}$$

where e is the specific internal energy.

The pressure p is given by the equation of state of a perfect gas :

$$p = (\gamma - 1)\ \rho\, e \tag{4.4}$$

where $\gamma > 1$ is the gas constant. For air, $\gamma = 1.4$. Three types of flows can be observed physically : totally supersonic, totally subsonic and transonic. In the latter case, the flow is supersonic at the inlet and subsonic at the outlet.

In the case of a steady flow, the Euler equations (4.1)–(4.2) can be reduced to a single nonlinear equation (see [12]). To that end, the first and third equations are directly integrated to obtain :

$$\begin{aligned} \rho\, u\, A &= C && (4.5)\\ (\rho\, E + p)\ u\, A &= H && (4.6) \end{aligned}$$

With (4.3) and (4.4), (4.6) gives

$$\gamma\, e + \frac{u^2}{2} = H \tag{4.7}$$

The constants C (mass flow rate) and H (total enthalpy) can be evaluated either at the inflow boundary or at the outflow boundary, as they are conserved even through a shock. Using (4.5) and (4.7), ρ and e can be eliminated from the second component of (4.1). One writes :

$$\begin{aligned} \rho &= \frac{C}{u\, A} && (4.8)\\ e &= \frac{1}{\gamma}\left(H - \frac{u^2}{2}\right) && (4.9) \end{aligned}$$

Plugging this into

$$\left[\left(\rho\, u^2 + p\right)\ A\right]_x - p\, A_x = 0 \tag{4.10}$$

yields :

$$\left\{\frac{C}{u}\left[u^2 + \frac{\gamma - 1}{\gamma}\left(H - \frac{u^2}{2}\right)\right]\right\}_x - A_x\,\frac{C}{u\, A}\,\frac{\gamma - 1}{\gamma}\left(H - \frac{u^2}{2}\right) = 0 \tag{4.11}$$

Simplifying by C and multiplying by $\frac{2\gamma}{\gamma+1}$ leads to :

$$\left(u + 2\,\frac{\gamma - 1}{\gamma + 1}\,\frac{H}{u}\right)_x + \frac{A_x}{A}\,2\,\frac{\gamma - 1}{\gamma + 1}\left(\frac{u}{2} - \frac{H}{u}\right) = 0 \tag{4.12}$$

Denoting $\overline{\gamma} = \frac{\gamma-1}{\gamma+1}$ and $\overline{H} = 2\overline{\gamma}H$, one finally obtains :

$$f_x + g = 0 \tag{4.13}$$

where

$$f(u) = u + \frac{\overline{H}}{u} \tag{4.14}$$

$$g(u,x) = \frac{A_x}{A}\left(\overline{\gamma}u - \frac{\overline{H}}{u}\right) \tag{4.15}$$

Equation (4.13) is now a single scalar nonlinear equation, to which the HBM is applied.

Assuming that the direction of the flow is from left (inlet) to right (outlet), the solution of (4.13) satisfies the following properties :

- the flow is sonic at

$$u = u_* \equiv \sqrt{\overline{H}} \quad , \tag{4.16}$$

  subsonic for $u < u_*$ and supersonic for $u > u_*$;

- the shock jump from u_L (left of the shock) to u_R (right of the shock) satisfies the Rankine-Hugoniot condition

$$u_L \cdot u_R = \overline{H} \quad ; \tag{4.17}$$

- the entropy condition ensures that any shock is physically acceptable (i.e. no expansion shock) :

$$u_L > u_* > u_R \tag{4.18}$$

For a smooth flow, (4.13) being a first order ODE, only one boundary condition is required. Rewriting (4.13) as :

$$\frac{\partial f}{\partial u} u_x + g = 0, \tag{4.19}$$

Equation (4.19) has the form of a convection equation where the direction of the information flow is given by the sign of the coefficient of u_x : if this coefficient is positive, information is transported from the inlet to the outlet, if it is negative, information is transported from the outlet to the inlet. Equations (4.14) and (4.16) show that for completely supersonic flows, $f'(u) = \frac{\partial f}{\partial u}$ is strictly positive, whereas, for completely subsonic flows, $f'(u)$ is strictly negative. As a consequence, for a completely supersonic flow, the boundary condition on the velocity is to be imposed at the inlet. For a completely subsonic flow, the velocity should rather be imposed at the outlet. If a shock is present in the flow, then the inflow has to be supersonic and the outflow must be subsonic. For this reason, one boundary condition has to be imposed at each end of the nozzle.

For smooth solutions, the nonlinear equation (4.13) can now be integrated analytically : f_x is first expanded and one obtains :

$$\left(1 - \frac{\overline{H}}{u^2}\right) u_x + \frac{A_x}{Au} \overline{\gamma} \left(u^2 - 2H\right) = 0 \tag{4.20}$$

which leads to

$$A_x u \overline{\gamma} \left(2H - u^2\right) + A\, u_x \left(2\overline{\gamma}\, H - u^2\right) = 0 \tag{4.21}$$

and

$$(Au)_x \left(2H - u^2\right) + \frac{2}{\gamma - 1} A\, u^2 u_x = 0 \tag{4.22}$$

Let $r = \frac{1}{\gamma-1}$ and multiply (4.22) by $\left(2H - u^2\right)^{r-1}$, (4.22) becomes :

$$\frac{d}{dx}\left[Au\left(2H - u^2\right)^r\right] = 0 \tag{4.23}$$

and finally

$$Au\left(2H - u^2\right)^r = K \tag{4.24}$$

K is a constant determined from the boundary conditions. When there is a shock in the duct, the constants computed from the inflow and outflow boundary conditions are different and it can be shown that K, which is an entropy function, increases across a shock (see [12]). (4.24) actually defines a family of curves relating the area of the cross-section of the nozzle to the velocity of the flow. The shape of the curves is given in Figure 4-1.

Equation (4.24) and Figure 4-1 illustrate the discussion on the boundary conditions. When the solution is smooth, then a single boundary condition completely determines the flow in the nozzle. Indeed, this boundary condition defines the constant K, hence a particular curve of the family (4.24), and one remains on this curve throughout the duct. If a shock is present, then each of the two boundary conditions determines a constant K and a curve of the family. The shock then enables to “jump” from one curve to the other. Let

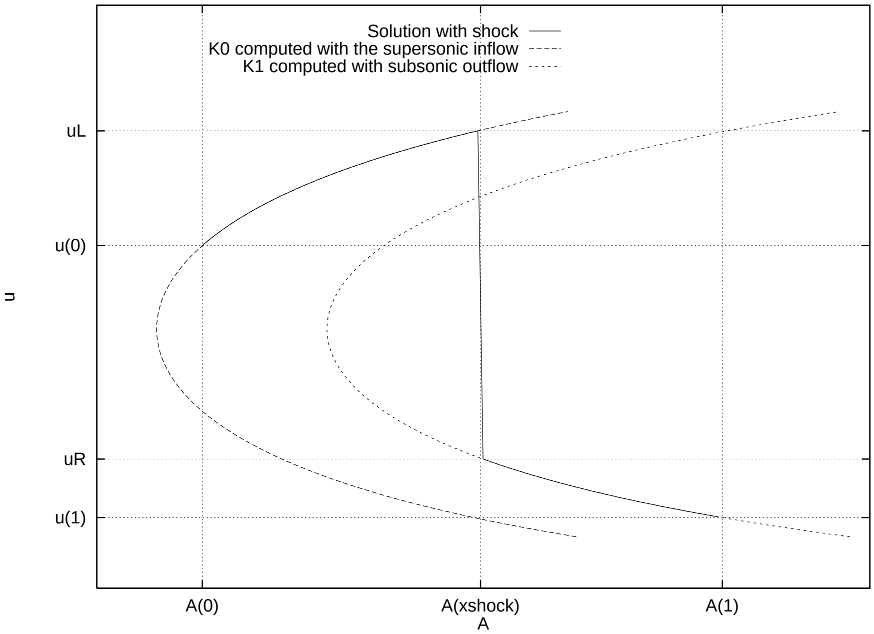


Figure 4-1: Curves for K_0 (supersonic), K_1 (subsonic) and Solution with shock

then K_0 and K_1 be the values of the constant computed based on the supersonic inflow and the subsonic outflow boundary conditions respectively. The function $z(u)$ being defined, as [12] :

$$z(u) = u\left(2H - u^2\right)^r \tag{4.25}$$

the position of the shock is given by :

$$\frac{1}{K_0}\, z(u_L) - \frac{1}{K_1}\, z\left(\frac{\overline{H}}{u_L}\right) = 0 \tag{4.26}$$

Equation (4.26) is derived by making use of (4.17) and of the fact that, at the shock, the area is the same on both sides. In the case where a shock is present, it can be shown that, since $K_0 < K_1$, (4.26) has a unique solution (see [12]).

4.2 Discrete Analysis Problem

The domain $\mathcal{D} =]0,1[$ is discretized with a uniform grid. The node coordinates are $x_i = i\,\delta\,(1 \leq i \leq n)$ and u_i is the (unknown) value of the solution at point x_i. Equation (4.13) looks very much like the pure convection equation, and better results can be expected if a Taylor-Galerkin approach is applied prior to the resolution of the equation. To that end, a non-physical time dependance is introduced :

$$u_t + f_x + g = 0 \tag{4.27}$$

from which

$$u_t = -(f_x + g) \tag{4.28}$$

leading to

$$u_{tt} = -f_{xt} - g_t \tag{4.29}$$

and the equation corresponding to (3.10) is then

$$f_x + g + \tau \ (f_{xt} + g_t) = 0 \tag{4.30}$$

Using Schwarz's Lemma, the order of the derivatives can be inverted and one obtains :

$$\begin{aligned} f_{xt} &= (f_t)_x \\ &= \left(\frac{\partial f}{\partial u} u_t\right)_x \\ &= \left[\frac{\partial f}{\partial u} \ (-f_x - g)\right]_x \end{aligned} \tag{4.31}$$

and similarly

$$g_x = \frac{\partial g}{\partial u} \ (-f_x - g) \tag{4.32}$$

Plugging (4.31)–(4.32) into (4.30), one obtains the following scheme :

$$f_x + g - \tau \left\{ \left[\frac{\partial f}{\partial u} \ (-f_x - g)\right]_x + \frac{\partial g}{\partial u} \ (-f_x - g) \right\} = 0 \tag{4.33}$$

Equation (4.33) can be "checked" intuitively by noting that, when $f = u$ and g is independant of u, like in the convection equation, (4.33) becomes exactly (3.10).

The corresponding weak form is then :

$$\int_0^1 w \, (f_x + g) \ dx + \int_0^1 \tau \, w_x \ \left[\frac{\partial f}{\partial u} \ (-f_x - g)\right] \ dx - \int_0^1 \tau \, w \, \frac{\partial g}{\partial u} \ (f_x + g) \ dx = 0 \tag{4.34}$$

4.2.1 Smooth Flow

Let us now assume a smooth solution with no shock. For simplicity, the flow is assumed supersonic in the rest of this section. The results obtained can be derived in a very similar way for totally subsonic flows. The solution is determined by the inflow boundary condition, and consequently, no outflow boundary condition can be imposed. Following the same reasoning as for the purely convective case, an additional term is introduced and the values of the solution u at all the points including the boundaries of the domain $\mathcal{D}$ are assumed to be unknowns. The scheme can now be written as

$$\int_0^1 \left(w + w_x\,\tau\,\frac{\partial f}{\partial u} - w\,\tau\,\frac{\partial g}{\partial u}\right)(f_x + g)\;dx + w(0)\;(u(0) - u_{in}) = 0 \tag{4.35}$$

where u_{in} is the boundary condition that defines the problem. By analogy with the linear case where τ was the term multiplying the w_x term, one now chooses

$$\tau = \frac{\delta}{2\max\left(\varepsilon, \left|\frac{\partial f}{\partial u}\right|\right)} \tag{4.36}$$

where ε is a fixed small number that prevents divisions by zero at sonic points (not exercised in this case).

If w is set equal to the piecewise linear basis function φ_i $(0 \le i \le n+1)$ and a one point integration is performed to approximate the resulting integrals, one obtains a *nonlinear* set of n equations with n unknowns, which can be written as :

$$W(u) = 0 \tag{4.37}$$

with

$$\begin{aligned} W^i(u) \;=\; & \left[\frac{\delta}{2} + \tau_i\,\frac{\partial f}{\partial u}\bigg|_{\overline{u}_i} - \frac{\delta}{2}\,\tau_i\,\frac{\partial f}{\partial u}\bigg|_{\overline{u}_i}\right]\left[\frac{\partial f}{\partial u}\bigg|_{\overline{u}_i}\,\frac{u_i - u_{i-1}}{\delta} + g\left(\overline{u}_i\right)\right] \\ & \left[\frac{\delta}{2} - \tau_{i+1}\,\frac{\partial f}{\partial u}\bigg|_{\overline{u}_{i+1}} - \frac{\delta}{2}\,\tau_{i+1}\,\frac{\partial f}{\partial u}\bigg|_{\overline{u}_{i+1}}\right]\left[\frac{\partial f}{\partial u}\bigg|_{\overline{u}_{i+1}}\,\frac{u_{i+1} - u_i}{\delta} + g\left(\overline{u}_{i+1}\right)\right] \end{aligned} \tag{4.38}$$

where $\overline{u}_i = \frac{u_i + u_{i-1}}{2}$ and $\tau_i = \tau(\overline{u}_i)$. In (4.38), one can recognize the contribution of element i (between x_{i-1} and x_i) on the first line and the contribution of element $(i+1)$ on the second line. The first contribution of course does not exist for $i = 0$ (in which case the additional term $u(0) - u_{in}$ appears), and the second does not appear for $i = n + 1$.

This nonlinear system can be solved by the Newton-Raphson method. This requires the computation of the *Jacobian matrix* of the nonlinear system, which will also be needed for the HBM. Practically, the Jacobian matrix is computed by assembling the matrices computed on each element i $(1 \leq i \leq n+1)$ as

$$\frac{\partial W_e}{\partial u} = \begin{pmatrix} \dfrac{\partial W_e^{i-1}}{\partial u_{i-1}} & \dfrac{\partial W_e^{i-1}}{\partial u_i} \\ \dfrac{\partial W_e^{i}}{\partial u_{i-1}} & \dfrac{\partial W_e^{i}}{\partial u_i} \end{pmatrix} \tag{4.39}$$

in (4.39), the index e indicates that the contribution of element e is considered.

The reasoning for subsonic flows is exactly the same, except that the boundary condition is imposed at $x = 1$ and the additive term is $w(1)\ (u(1) - u_{out})$ instead of $w(0)\ (u(0) - u_{in})$.

4.2.2 Normalization

In the rest of this chapter, the following normalization is used. The speed of sound at infinity is set equal to 1. This speed of sound is given by :

$$c_\infty^2 = \frac{\gamma\, p_\infty}{\rho_\infty} = \gamma\,(\gamma - 1)\, e_\infty \tag{4.40}$$

so that, with $c_\infty = 1$, one obtains

$$e_\infty = \frac{1}{\gamma\,(\gamma - 1)} \tag{4.41}$$

If one moreover assumes $\rho_\infty = 1$, one has

$$H_\infty = E_\infty + \frac{p_\infty}{\rho_\infty} = E_\infty + \frac{\gamma - 1}{\rho_\infty}\left(\rho_\infty\, E_\infty - \frac{1}{2}\rho_\infty\, u_\infty^2\right) = \gamma\, e_\infty + \frac{1}{2}\, u_\infty^2 \tag{4.42}$$

which gives the following consistent set of variables :

$$\rho_\infty = 1 \tag{4.43}$$

$$c_\infty = 1 \tag{4.44}$$

$$u_\infty = M_\infty \tag{4.45}$$

$$H_\infty = \frac{1}{\gamma - 1} + \frac{M_\infty^2}{2} \tag{4.46}$$

since by definition $u = c\,M$. The parameter $\overline{H}$ is then determined by

$$\overline{H} = 2\,\overline{\gamma}\,H_\infty = \frac{2}{\gamma+1} + \frac{\gamma-1}{\gamma+1}\,M_\infty^2 \tag{4.47}$$

4.3 Bounds for the Average Value of the Solution

The goal of the previous section was to provide a discrete solution to the problem (4.13). In this section, the Hierarchical Bounds Method is applied to the problem to find bounds for the average value of the solution u. Because the HBM can only be applied to linear problems, the equation must be first *linearized*. A heuristic approach is now used. The solution on the coarse grid is assumed sufficiently close to the solution on the fine grid, so that $W(u)$ can be linearized about the H-mesh solution. Let u_H and u_h be the solutions on the coarse and fine grids respectively, and let $W_H(u_H)$ and $W_h(u_h)$ be the corresponding systems of nonlinear equations. By definition, one has :

$$W_H(u_H) \;=\; 0 \tag{4.48}$$

$$W_h(u_h) \;=\; 0 \tag{4.49}$$

Let u_{iH} be the linear interpolation of u_H on the fine grid. If u_h is sufficiently close to u_H, one can write :

$$W_h(u_h) \approx W_h(u_{iH}) + \frac{\partial W_h}{\partial u}(u_{iH})\;(u_h - u_{iH}) \tag{4.50}$$

hence, from (4.49)

$$L_h\,u_h = f_h \tag{4.51}$$

where

$$L_h \;=\; \frac{\partial W_h}{\partial u}(u_{iH}) \tag{4.52}$$

$$f_h \;=\; L_h\,u_{iH} - W_h(u_{iH}) \tag{4.53}$$

One can verify that, in virtue of (4.48), the linearized equation on the coarse grid has the same form as (4.51).

The HBM is then applied to the linearized problem, with the possible extensions or modifications mentioned for the smooth case. Two main conceptual difficulties can be expected. First, because nothing ensures that the self-adjoint part of the linearized system matrix is positive definite, one cannot be sure that the HBM gives correct bounds for the output. Second, even when this matrix is positive definite, the bounds obtained correspond to the *linearized* problem, and nothing ensures that the fine grid output computed directly from the nonlinear system lies between these bounds.

4.4 Numerical Results

The simulations were performed using some of the numerical data given in [12]. More precisely, the diverging nozzle has a linearly growing area, with $A(0) = 1.05$ and $A(1) = 1.745$. The inflow boundary condition in the case of a completely supersonic flow is $u(0) = 1.299$, and the outflow boundary condition for the totally subsonic flow is $u(1) = 0.35$. The fine grid discretization step is $h = 10^{-3}$, and the coarse grid element size is refined from 0.1 to 10^{-3}.

4.4.1 Supersonic Flow

The completely supersonic flow is first investigated. The solutions obtained by the Newton-Raphson method applied to the modified scheme (4.35) are given on Figure 4-2. Even for coarse grids, the solution is very accurate, so one may expect the coarse grid output to be very close to the fine grid output. Contrary to the previous two equations studied is Chapters 2 and 3, the analytical solution is not readily available in the present case, at least not under the explicit form $u = u(x)$. The exact solution given as a reference was actually computed as $x = x(u)$ from (4.24) and the velocity appears to be a growing function of the space coordinate, as is expected from a supersonic flow in a diverging nozzle.

The adjoints associated to this problem are now plotted on Figure 4-3. The boundary values of the adjoints are computed algebraically, as in the purely convective case. One remark should be made at that point. The choice of the parameter τ is critical to the quality of the solution, as well as to the stability and the convergence of the scheme. Indeed, this parameter τ, which is "artificial" (it is introduced to improve the quality of the solution) can be interpreted from two different points of view. First, as the coefficient of the second derivative of the solution in the equation, τ can be seen as an (artificial) viscosity coefficient. Second, from (4.36), τ is also a time step that can be used for time marching.

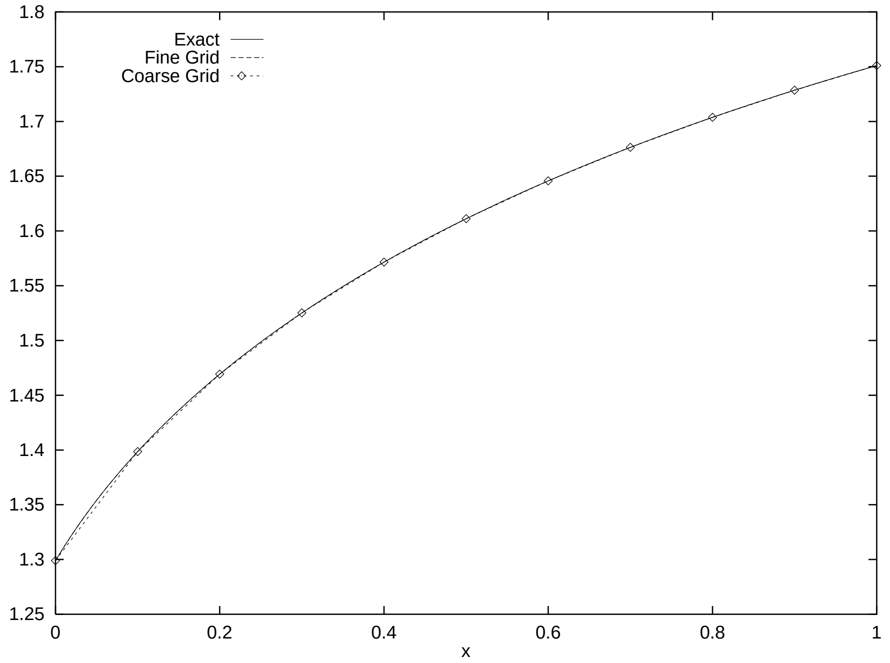


Figure 4-2: Modified Finite Elements, Completely Supersonic Flow, $h = 10^{-3}$, $H = 0.1$

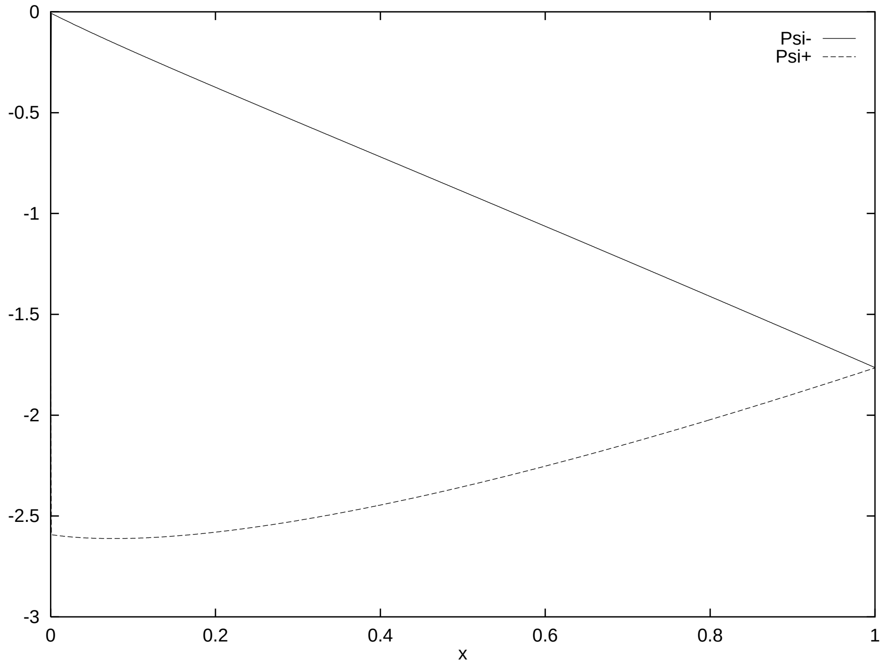


Figure 4-3: Adjoints $\psi_H^{\pm}$, Completely Supersonic Flow, $H = h = 10^{-3}$

Hence, a large CFL number multiplying the factor τ smooths the solution and prevents numerical oscillations by adding some artificial viscosity, but destabilizes the iterative Newton-Raphson procedure. On the contrary, a small CFL number stabilizes the numerical scheme, but allows parasite oscillations to appear in the solution. In the results given in this section, the CFL number is equal to 1, which seems numerically to be a good trade-off between stability and accuracy.

The bounds obtained for this problem are given on Figure 4-4. Four observations can

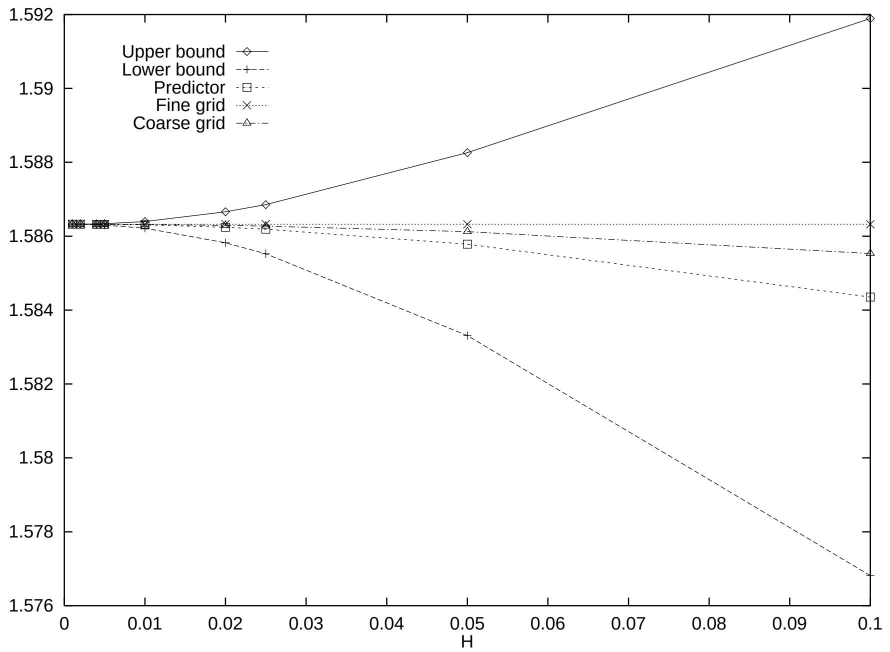


Figure 4-4: Bounds for the Output : Average of the Solution, Completely Supersonic Flow, $h = 10^{-3}$, $10^{-3} \leq H \leq 0.1$

be drawn from this graph. First, the fine grid output represented is the output computed directly on the fine grid, and not the linearized problem output. s_h is indeed between the bounds computed by using the linearized problem. The second observation concerns the sharpness of the bounds. Even though the lower and upper bounds are not as sharp as in the linear purely convective case, they estimate the fine grid output within less than 1% in the worst case, i.e. when the coarse grid is made of only 11 points (10 elements). The third observation concerns the coarse grid output. The main difference between the nonlinear case and the other previous cases is the closeness with which the the coarse grid output matches the fine grid. In the nonlinear case, the coarse grid is closer than either the bounds or the predictor. This of course does not mean that the method fails for the nonlinear case. Indeed, the present results show that one can compute strict bounds for the average of the velocity, which, for the engineer, is at least as important, if not more important than computing the output exactly. Finally, it has to be noted that in these simulations, the stabilization parameter κ has been optimized at each step. A characteristic feature of this problem is that the optimal value of κ becomes very large as H converges to h. Even if the optimal value of κ is not computed when $H = h$ because both the numerator and the denominator in (1.76) vanish, the evolution of κ for the last values of H indicates that the stabilization parameter goes to $+\infty$ as H converges to h.

Again, the convergence of the bounds is $O(H^2)$, as shown by Figure 4-5. This figure

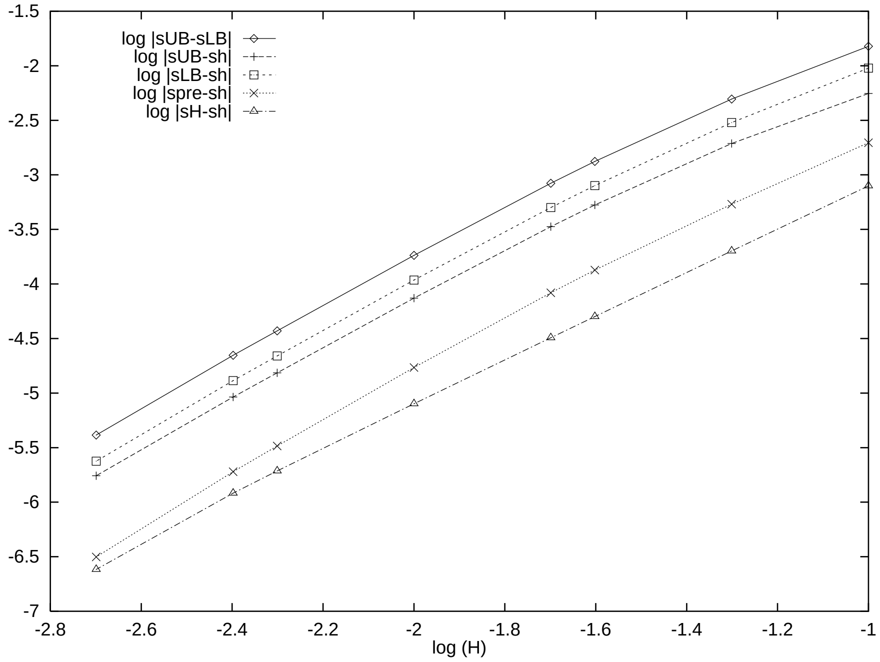


Figure 4-5: Convergence of the Bounds for the Output : Average of the Solution, Completely Supersonic Flow, $h = 10^{-3}$, $10^{-3} \leq H \leq 0.1$

shows that the coarse grid output and the predictor are not as close to the fine grid output as in the purely convective case (the gap between the lines is not as important), and that the convergence is much more similar to the convection-diffusion case.

4.4.2 Subsonic Flow

Similar numerical results can be obtained when the flow is subsonic, i.e. when a subsonic boundary condition is imposed on the outflow velocity. Not all subsonic boundary conditions are acceptable in this case. Indeed, when one "goes back" into the nozzle, the velocity grows, since in that direction, the nozzle is converging, so the sonic point may be reached before the inlet if the outflow velocity imposed is too large, and the flow cannot be solved. One can also understand this phenomenon by looking at Figure 4-1 : starting at $A(1)$ and $u(1)$ on the subsonic curve, the sonic point (extremum of the curve) is reached before $A(0)$ (the vertical line $A = A(0)$ is left of the extremum), so that no point of the curve corresponds to $A(0)$. This explains why the value $u(1) = 0.35$ has been chosen for the boundary condition. The solutions computed are given by Figure 4-6. Again, the accuracy of the solutions, even

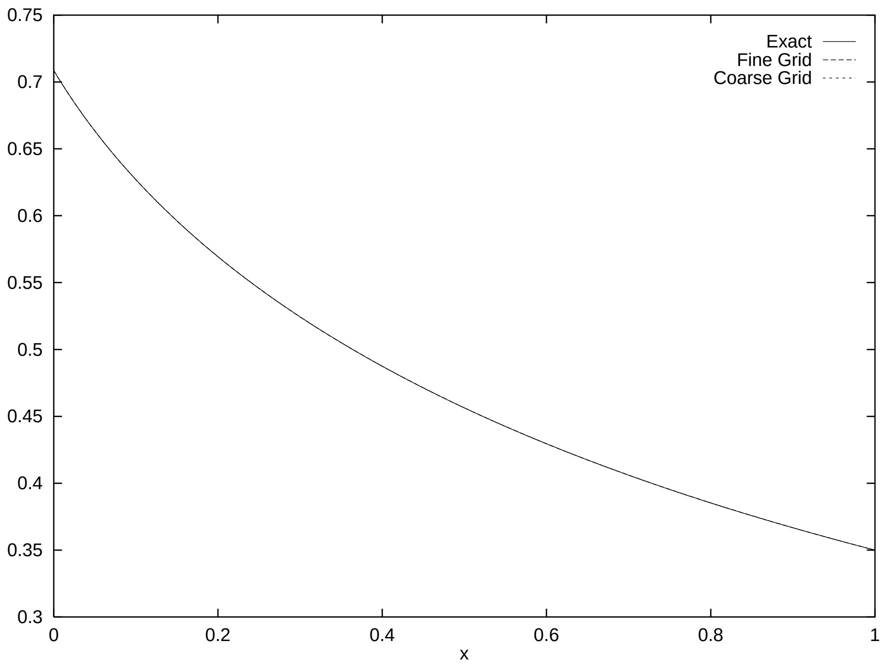


Figure 4-6: Modified Finite Elements, Completely Subsonic Flow, $h = 10^{-3}$, $H = 0.1$

for very coarse grids is remarkable and the coarse grid output is expected to be very close to the fine grid output.

The bounds obtained for the average output are given on Figure 4-7. Several points can be derived from this graph. First, the coarse grid output is very close to the fine grid output, as expected. Second, the bounds are clearly not as sharp as in the supersonic case. In particular, in spite of the optimization of the stabilization factor, the lower bound is

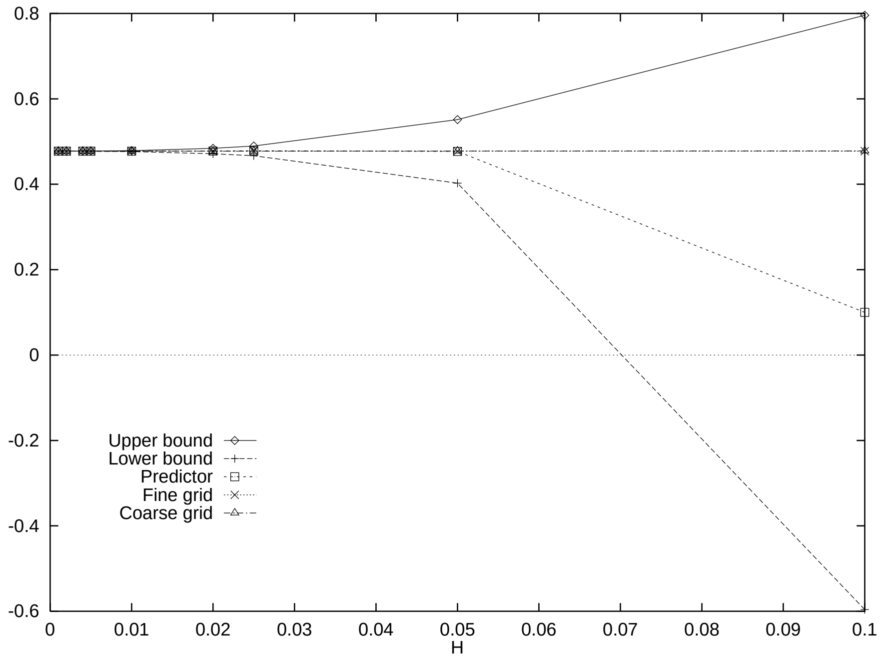


Figure 4-7: Bounds for the Output : Average of the Solution, Completely Subsonic Flow, $h = 10^{-3}$, $10^{-3} \leq H \leq 0.1$

very poor for the coarsest grid, which affects the estimator accordingly. Third, although the sharpness of the bounds is a little disappointing, their convergence is still fast and the fine grid output is indeed between the bounds. Fourth and finally, except for the coarsest mesh, the predictor estimates very well the fine grid output, since on this plot, the curve representing the predictor cannot be distinguished from the curve representing the fine grid output when $H \leq 0.05$. A closer look at the neighborhood of the fine grid output (Figure 4-8) shows that, even if the predictor remains close to the fine grid output (as shown by the vertical scale of the plot), its evolution is not as regular as in other cases, and the coarse grid output gives better results.

The convergence of the bounds is illustrated on Figure 4-9. The irregularities in the evolution of the predictor are more apparent on this plot, but one has to take into account that the general theory does not give any indication on the behaviour of this predictor. On the other hand, the convergence of the bounds is regular (but for the first points of the curves), and the slope of the lines is approximately 2.56 for the error bounds (the coarse grid output converges as $O(H^2)$).

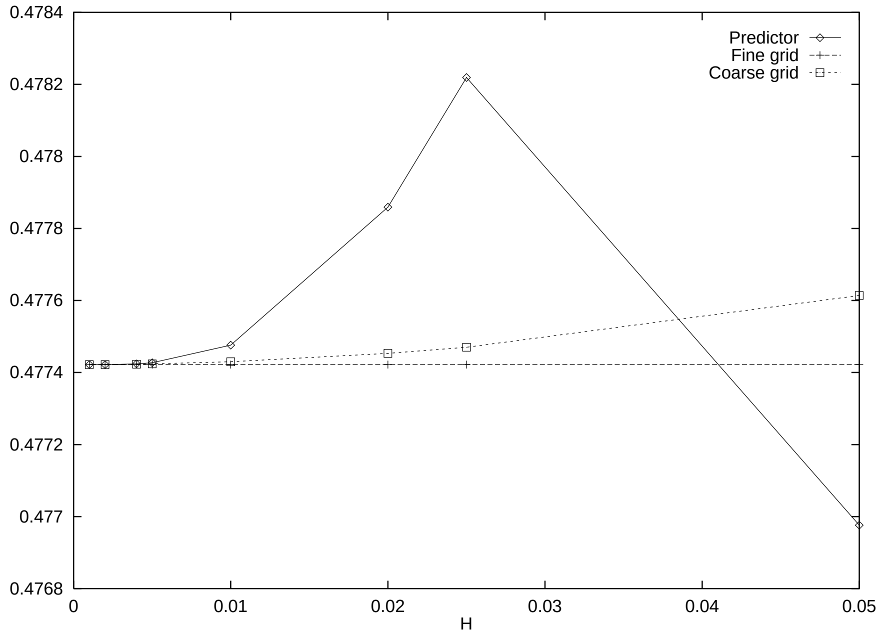


Figure 4-8: Predictor for the Output : Average of the Solution, Completely Subsonic Flow, $h = 10^{-3}$, $10^{-3} \leq H \leq 0.05$

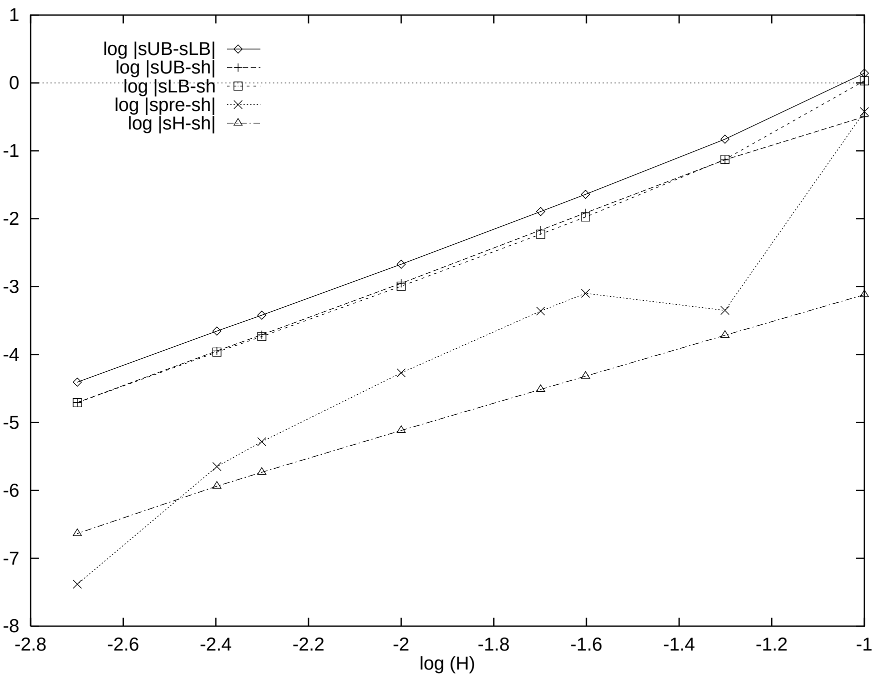


Figure 4-9: Convergence of the Bounds for the Output : Average of the Solution, Completely Subsonic Flow, $h = 10^{-3}$, $10^{-3} \leq H \leq 0.1$

4.5 Conclusion

The main conclusions of this chapter are twofold. First, the Hierarchical Bounds Method can be adapted to nonlinear cases, at least to a certain extent. Even though the results are not as good as those in the linear cases considered, the bounds provided by coarse grids remain fairly sharp. Second, the fine grid output computed directly from the nonlinear discrete system lies between the bounds in the cases considered. The problem is now that in the case where the flow is not smooth (presence of a shock), it seems that the HBM fails, mainly for two reasons. First, the self-adjoint part of the linearized system is not positive definite. Second, in certain cases, the optimal value of κ is not defined because the denominator $\beta^{\pm}$ of the fraction under the square root is negative.

Chapter 5

Domain Decomposition

5.1 Introduction

In the problems considered so far, the Hierarchical Bounds Method hardly brought any improvement with respect to the direct resolution of the equations on a fine grid. This is because, in one dimension, the matrix L_h of the discrete system is tridiagonal, inverting its symmetric part is as costly as the direct inversion of L_h.

However, two features of the HBM have been highlighted by these examples. First, the values of the fine grid output predicted by the bounds (average value of the upper and lower bounds) is usually closer to the exact solution than the coarse grid output. Second, in higher dimensions, the matrix of the discrete system is not tridiagonal anymore, but sparse, so the *symmetry* of A_h can be fully exploited. On one hand, a whole range of methods is available to invert symmetric matrices : direct inversion (Gauss's pivot, LU or QR Decompositions [7, 10]), iterative methods (conjugate gradient, GMRES [8], Krylov subspace methods in general [9]),... On the other hand, the existing methods to invert non-symmetric matrices like L_h are either not very efficient in general (Gauss's Pivot) or not systematic.

The *Domain Decomposition* is a technique that further reduces the cost by using the fact that the inversion of a general $n \times n$ matrix (even sparse) is a $O(n^3)$ process. The domain $\mathcal{D}$ is decomposed into K subdomains and the problem is solved on each subdomain before continuity is imposed at intersubdomain boundaries. For simplicity, and following [5], the Domain Decomposition is presented in this chapter for one-dimensional problems.

5.2 Domain Decomposition Formulation

5.2.1 Notations

The problem is defined on $\mathcal{D} =]0,1[$. The closure of $\mathcal{D}$ is noted $\overline{\mathcal{D}}$. A decomposition of $\mathcal{D}$ into K subdomains is now introduced. Let $\mathcal{K} = \{1, \ldots, K\} \subset I\!N$, one has :

$$\overline{\mathcal{D}} = \bigcup_{k \in \mathcal{K}} \overline{\mathcal{D}}^k \tag{5.1}$$

where $\overline{\mathcal{D}}^k$ is the closure of $\mathcal{D}^k = \left] a^k, a^{k+1} \right[$. The points $a^k \, (1 \leq k \leq K)$ are such that $a^1 = 0 < a^2 < \cdots < a^{K+1} = 1$.

The subdomain boundaries are assumed to be nodes of the grid, and the corresponding indices J^k are defined by :

$$\left\{ \begin{aligned} J^1 &= 1 \\ a^k &= x_{J^k} \qquad \text{for} \quad (2 \leq k \leq K) \\ J^{K+1} &= n \end{aligned} \right. \tag{5.2}$$

where n is the number of points of the mesh in $\mathcal{D}$ and $\{x_i\}_{1 \leq i \leq n}$ are the nodes of the grid.

Two sets of indices, local and global, are then defined on each subdomain. Let $\mathcal{M}^k = \left\{1, \ldots, M^k\right\}$ be the set of "local" indices on $\mathcal{D}^k$ for all $k \in \{1, \ldots, K\}$. The number of points in each subdomain $\overline{\mathcal{D}}^k$ interior to the domain $\mathcal{D}$ is thus

$$M^k = J^{k+1} - J^k + 1 \qquad \text{for} \quad k \in \mathcal{K} \tag{5.3}$$

In the enumeration of the points of the subdomains, the extremities are cited *twice*, once for each participating subdomain, so the number of degrees of freedom before imposing the continuity of the solution at the subdomain boundaries is $\tilde{n} = n + K - 1$.

The global indices of this "decoupled" enumeration are defined on each subdomain by the sets $\tilde{\mathcal{J}}^k = \left\{ \tilde{J}^k, \ldots, \tilde{J}^{k+1} - 1 \right\}$, where

$$\left\{ \begin{aligned} \tilde{J}^k &= J^k + k - 1 \qquad \text{for} \quad k \in \mathcal{K} \\ \tilde{J}^{K+1} &= \tilde{n} + 1 \end{aligned} \right. \tag{5.4}$$

5.2.2 Subdomain Operators

With the notations of the previous subsection, one defines two hybridization operators $Q \in I\!R^{\tilde{n}\times n}$ and $V \in I\!R^{(K-1)\times\tilde{n}}$ related to direct stiffness summation and intersubdomain continuity respectively. The matrix Q is given by

$$\forall k \in \mathcal{K} \quad \text{and} \quad p \in \mathcal{M}^k, \quad Q_{ij} = \begin{cases} 1 & \text{if} \quad i = \tilde{J}^k + p - 1 \quad \text{and} \quad j = J^k + p - 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

and the matrix V is defined as

$$\forall i \in \mathcal{K}, \quad V_{ij} = \begin{cases} -1 & \text{if} \quad j = \tilde{J}^{i+1} - 1 \\ 1 & \text{if} \quad j = \tilde{J}^{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

To be more explicit, let us consider a decomposition into 3 intervals of equal length, we have

$$Q = \begin{pmatrix}
1 & 0 & \cdots & 0 & & & & & & & \\
0 & 1 & \ddots & \vdots & & & & & & & \\
\vdots & \ddots & 1 & 0 & & & & & & & \\
0 & \cdots & 0 & 1 & & & & & & & \\
 & & & 1 & 0 & \cdots & 0 & & & & \\
 & & & 0 & 1 & \ddots & \vdots & & & & \\
 & & & \vdots & \ddots & 1 & 0 & & & & \\
 & & & 0 & \cdots & 0 & 1 & & & & \\
 & & & & & & 1 & 0 & \cdots & 0 \\
 & & & & & & 0 & 1 & \ddots & \vdots \\
 & & & & & & \vdots & \ddots & 1 & 0 \\
 & & & & & & 0 & \cdots & 0 & 1
\end{pmatrix} \tag{5.7}$$

and

$$V = \begin{pmatrix} 0 & \cdots & 0 & -1 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 & 0 & \cdots & 0 \end{pmatrix} \tag{5.8}$$

The columns of Q containing two nonzero entries correspond to the subdomain boundaries (columns J^k). The nonzero entries of V are in the columns $\tilde{J}^{k+1} - 1$ and $\tilde{J}^{k+1}$. The matrix V is defined only for $K > 2$.

Given any $v \in \mathbb{R}^n$ specifying the global nodal values of the solution, $Q \cdot v$ assigns these values to the local nodes on each subdomain. Conversely, given any $w \in \mathbb{R}^{\tilde{n}}$, $V \cdot w$ evaluates the jumps in w across the $K-1$ subdomain boundaries. From (5.7) and (5.8), the following equalities can easily be derived :

$$Q^T V^T = 0 \tag{5.9}$$

$$V\,V^T = 2\,I_{K-1} \tag{5.10}$$

where I_{K-1} is the $(K-1) \times (K-1)$ identity matrix.

The subdomain matrix operators L^k and A^k can now be defined along with the subdomain right-hand sides f^k by writing the weak form of the equation on each subdomain. For instance, for the convection-diffusion problem, one obtains :

$$\forall\, k \in \mathcal{K}, \qquad \forall\,(i,j) \in \left(\mathcal{M}^k\right)^2$$

$$\left(L^k\right)_{ij} = \int_{a^k}^{a^{k+1}} \left(\nu\, \frac{d\varphi_{i+J^k-1}}{dx}\, \frac{d\varphi_{j+J^k-1}}{dx} + \varphi_{i+J^k-1}\, \frac{d\varphi_{j+J^k-1}}{dx}\right) dx \tag{5.11}$$

$$\left(A^k\right)_{ij} = 2 \int_{a^k}^{a^{k+1}} \left(\nu\, \frac{d\varphi_{i+J^k-1}}{dx}\, \frac{d\varphi_{j+J^k-1}}{dx}\right) dx \tag{5.12}$$

$$\begin{aligned}\left(f^k\right)_i = \int_{a^k}^{a^{k+1}} \Big[\varphi_{i+J^k-1}\, g - \left(\nu\, \frac{d\varphi_{i+J^k-1}}{dx}\, \frac{d\varphi_{n+1}}{dx} - \varphi_{i+J^k-1}\, \frac{d\varphi_{n+1}}{dx}\right) U_1 \\ - \left(\nu \frac{d\varphi_{i+J^k-1}}{dx}\, \frac{d\varphi_0}{dx} - \varphi_{i+J^k-1}\, \frac{d\varphi_0}{dx} dx\right) U_0\Big]\, dx\end{aligned} \tag{5.13}$$

The output linear functionals also need to be decomposed. This decomposition is not unique. A judicious choice can however facilitate computations in certain cases as discussed in [5]. One possibility consists of writing :

$$\forall\, k \in \mathcal{K}\, \left(\ell^k\right)_i = \begin{cases} \ell_{i+J^k-1} & \text{if} \quad \begin{cases} i \in \left\{1, \ldots, M^k - 1\right\} \\ k = K \quad \text{and} \quad i = M^K \end{cases} \\[2ex] 0 & \text{if} \quad k \in \{1, \ldots, K-1\} \quad \text{and} \quad i = M^k \end{cases} \tag{5.14}$$

The block diagonal operators used to solve the problem numerically are now formed. These operators are denoted with an underlining bar to distinguish them from the original matrices on the global discretizations. These operators contain the unassembled, decoupled matrices (5.11)–(5.12). One thus defines $(\underline{L}, \underline{A}) \in \left(I\!R^{\tilde{n}\times\tilde{n}}\right)^2$ by :

$$\forall\, k \in \mathcal{K}$$

$$(\underline{L})_{ij} = \begin{cases} \left(L^k\right)_{i-\tilde{J}^k+1,\, j-\tilde{J}^k+1} & \text{if} \quad (i,j) \in \left(\tilde{J}^k\right)^2 \\ 0 & \text{otherwise} \end{cases} \tag{5.15}$$

$$(\underline{A})_{ij} = \begin{cases} \left(A^k\right)_{i-\tilde{J}^k+1,\, j-\tilde{J}^k+1} & \text{if} \quad (i,j) \in \left(\tilde{J}^k\right)^2 \\ 0 & \text{otherwise} \end{cases} \tag{5.16}$$

The vectors $\left(\underline{f}, \underline{\ell}\right) \in \left(I\!R^{\tilde{n}}\right)^2$ are also formed. They contain the "unassembled" and decoupled inhomogeneities (5.13) and output functionals (5.14) respectively :

$$\forall\, k \in \mathcal{K} \qquad \forall\, i \in \tilde{\mathcal{J}}^k$$

$$\underline{f}_i = \left(f^k\right)_{i+\tilde{J}^k-1} \tag{5.17}$$

$$\underline{\ell}_i = \left(\ell^k\right)_{i+\tilde{J}^k-1} \tag{5.18}$$

The operators (5.15)–(5.16) and the vectors (5.17)–(5.18) can be expressed simply using (5.7) and (5.8) as :

$$L = Q^T \underline{L}\, Q \tag{5.19}$$

$$A = Q^T \underline{A}\, Q \tag{5.20}$$

$$f = Q^T \underline{f} \tag{5.21}$$

$$\ell = Q^T \underline{\ell} \tag{5.22}$$

The consequences of the domain decomposition on the general theory provided in Chapter 1 are now discussed.

5.3 Duality Approach to Bounds for the Outputs

u is the unique solution of (1.18), so that, using (5.9) and (5.19) and letting

$$\underline{u} = Q\, u \tag{5.23}$$

$\underline{u}$ becomes the solution of the system :

$$
\begin{aligned}
Q^T \underline{L}\, \underline{w} &= f && (5.24)\\
V\, \underline{w} &= 0 && (5.25)
\end{aligned}
$$

Moreover, V is of rank $(K-1)$ and Q is of rank n, so (5.9) shows that the columns of Q span the nullspace of V. Adding to this the assumption that A is positive definite leads to the conclusion that this solution is also unique.

If (5.19), (5.20), (5.21) and (5.23) are now plugged into (1.25), one obtains :

$$\frac{\kappa}{2}\, \underline{u}^T \underline{A}\, \underline{u} - \kappa\, \underline{u}^T \underline{f} = 0 \tag{5.26}$$

An augmented output form is then defined for all $\underline{v} \in I\!R^{\tilde{n}}$:

$$\mathcal{S}^{\pm}(\underline{v}) = \frac{\kappa}{2}\, \underline{v}^T \underline{A}\, \underline{v} - \kappa\, \underline{v}^T \underline{f} \pm \left(\underline{v}^T \underline{\ell} + c \right) \tag{5.27}$$

and the output can be written, with (5.24) and (5.25), as

$$\pm\, s = \min_{\{\underline{v} \in I\!R^{\tilde{n}} \,|\, Q^T \underline{L}\, \underline{v} - f = 0,\, V\, \underline{v} = 0\}} \mathcal{S}^{\pm}(\underline{v}) \tag{5.28}$$

The Lagrangian corresponding to (1.29) introduces the additional adjoint ρ :

$$\mathcal{L}^{\pm}(\underline{v}, \mu, \rho) = \mathcal{S}^{\pm}(\underline{v}) + \mu^T (Q^T \underline{L}\, \underline{v} - f) + \rho^T V\, \underline{v} \tag{5.29}$$

and the duality result (1.38)–(1.39) can be written as

$$
\begin{aligned}
\pm\, s &= \min_{\{\underline{v} \in I\!R^{\tilde{n}}\}} \max_{\{\mu \in I\!R^n,\, \rho \in I\!R^{K-1}\}} \mathcal{L}^{\pm}(\underline{v}, \mu, \rho) && (5.30)\\
&= \max_{\{\mu \in I\!R^n,\, \rho \in I\!R^{K-1}\}} \min_{\{\underline{v} \in I\!R^{\tilde{n}}\}} \mathcal{L}^{\pm}(\underline{v}, \mu, \rho) && (5.31)
\end{aligned}
$$

The bounds are therefore given by :

$\forall\, \hat{\mu}^{\pm} \in I\!R^n$ and $\forall\, \hat{\rho}^{\pm} \in I\!R^{K-1}$

$$\min_{\{\underline{v} \in I\!R^{\tilde{n}}\}} \mathcal{L}^{+} \left(\underline{v}, \hat{\mu}^{+}, \hat{\rho}^{+}\right) \leq s \leq - \min_{\{\underline{v} \in I\!R^{\tilde{n}}\}} \mathcal{L}^{-} \left(\underline{v}, \hat{\mu}^{-}, \hat{\rho}^{-}\right) \tag{5.32}$$

and the saddlepoints $(u, \psi^{\pm}, \omega^{\pm})$ are solutions of the systems of equations :

$$\begin{aligned} \kappa\, \underline{A}\, \underline{u} + \underline{L}^T Q\, \psi^{\pm} + V^T \omega^{\pm} - \kappa\, \underline{f} \pm \underline{\ell} &= 0 && (5.33)\\ Q^T \underline{L}\, \underline{u} - f &= 0 && (5.34)\\ V\, \underline{u} &= 0 && (5.35) \end{aligned}$$

obtained as stationarity conditions for the Lagrangian (5.29) with respect to u, μ and ρ for (5.33), (5.34) and (5.35) respectively. $(\psi^{\pm}, \omega^{\pm})$ are the values of the adjoints that make the bounds (5.32) exact.

5.4 Hierarchical Procedure

Two levels of discretization are now introduced, one coarse ("working mesh") and one fine ("truth mesh"), and the domain decomposition is chosen such that all the intersubdomain boundaries are also nodes of both grids. $\left(\hat{\mu}_h^{\pm}, \hat{\rho}_h^{\pm}\right)$ are computed as linear interpolants of the best choice for the adjoints on the coarse grid, which are, from (5.33)–(5.35), $\left(\psi_H^{\pm}, \omega_H^{\pm}\right)$.

5.4.1 Computational Procedure

First, the solution on the coarse grid u_H and its "decomposed" version $\underline{u}_H = Q_H\, u_H$ are computed. The stationary conditions on the coarse grid are

$$\begin{aligned} \kappa\, \underline{A}_H\, \underline{u}_H + \underline{L}_H^T Q_H\, \psi_H^{\pm} + V_H^T \omega_H^{\pm} - \kappa\, \underline{f}_H \pm \underline{\ell}_H &= 0 && (5.36)\\ Q_H^T \underline{L}_H\, \underline{u}_H - f_H &= 0 && (5.37)\\ V_H\, \underline{u}_H &= 0 && (5.38) \end{aligned}$$

Pre-multiplying (5.36) by Q_H^T and using (5.9), (5.19)–(5.21) and (5.23), $\psi_H^{\pm}$ is obtained by

$$L_H^T \psi_H^{\pm} = -\left(\kappa\, A_H\, u_H - \kappa\, f_H \pm \ell_H\right) \qquad (5.39)$$

which is exactly (1.46), which does not include the domain decomposition. $\omega_H^{\pm}$ is then obtained by pre-multiplying (5.36) by V_H and using (5.10), which yields directly :

$$\omega_H^{\pm} = -\frac{1}{2} V_H\, \left(\kappa\, \underline{A}_H\, \underline{u}_H + \underline{L}_H^T Q_H\, \psi_H^{\pm} - \kappa\, \underline{f}_H \pm \underline{\ell}_H\right) \qquad (5.40)$$

The domain decomposition technique therefore does not add any complexity to the proceduresо far, at least from a numerical point of view. ψ_H can now be interpolated as in (1.47). Moreover, both $\omega_H^{\pm}$ and $\hat{\rho}_h^{\pm}$ have the same dimension $(K-1)$, so, given the results of the coarse grid computations, the best choice for $\hat{\rho}_h^{\pm}$ is

$$\hat{\rho}_h^{\pm} = \omega_H^{\pm} \tag{5.41}$$

The bounds are then computed as

$$\begin{aligned} (s_h)_{LB}(H) &= \min_{\{\underline{v} \in I\!R^{\tilde{n}_h}\}} \mathcal{L}_h^+ \left(\underline{v}, \hat{\mu}_h^+, \hat{\rho}_h^+ \right) && (5.42) \\ (s_h)_{UB}(H) &= -\min_{\{\underline{v} \in I\!R^{\tilde{n}_h}\}} \mathcal{L}_h^- \left(\underline{v}, \hat{\mu}_h^-, \hat{\rho}_h^- \right) && (5.43) \end{aligned}$$

From the stationarity condition (5.33) applied on the fine grid ($\delta = h$), if

$$\underline{\hat{u}}_h^{\pm} = \arg \min_{\{\underline{v} \in I\!R^{\tilde{n}}\}} \mathcal{L}_h^{\pm} \left(\underline{v}, \hat{\mu}_h^{\pm}, \hat{\rho}_h^{\pm} \right) \tag{5.44}$$

one must have

$$\kappa \, \underline{A}_h \, \underline{\hat{u}}_h^{\pm} = - \left(\underline{L}_h^T Q_h \, \hat{\mu}_h^{\pm} + V_h^T \hat{\rho}_h^{\pm} - \kappa \, \underline{f}_h \pm \underline{\ell}_h \right) \tag{5.45}$$

and the bounds become, as in (1.56) and (1.57) :

$$\begin{aligned} (s_h)_{LB}(H) &= \mathcal{L}_h^+ \left(\underline{\hat{u}}_h^+, \hat{\mu}_h^+, \hat{\rho}_h^+ \right) \\ &= -\frac{\kappa}{2} \, \underline{\hat{u}}_h^{+T} \underline{A}_h \, \underline{\hat{u}}_h^+ + c_h - \hat{\mu}_h^{+T} f_h && (5.46) \\ (s_h)_{UB}(H) &= -\mathcal{L}_h^- \left(\underline{\hat{u}}_h^-, \hat{\mu}_h^-, \hat{\rho}_h^- \right) \\ &= \frac{\kappa}{2} \, \underline{\hat{u}}_h^{-T} \underline{A}_h \, \underline{\hat{u}}_h^- + c_h + \hat{\mu}_h^{-T} f_h && (5.47) \end{aligned}$$

In view of (5.19), (5.20) and (5.23), these expressions for the bounds are actually exactly the same as in (1.56) and (1.57). Furthermore, using (5.9) and (5.21), (5.45) pre-multiplied by Q^T gives exactly (1.50).

5.4.2 Computational Cost

Given that the final results obtained are exactly the same as in Chapter 1, one might wonder whether the Domain Decomposition really improves the performance of the Hierarchical

Bounds Method. Furthermore, nothing guarantees that (5.45) has a solution. In fact, as shown in [5], there *is* always a solution and $\underline{\hat{u}}_h^{\pm}$ can be computed.

The main gain may not be obvious from this presentation because the procedure has been presented only in one dimension for simplicity. It appears nevertheless that the cost of the method once again comes from the inversion of $\underline{A}_h$. Only this time, (5.16) shows that this inversion is in fact decomposed in the resolution of K *decoupled* systems of size n/K. The Domain Decomposition is thus cost effective for two reasons. First, the resolution of subsystems can be parallelized. Second, in higher dimensions, the cost of the inversion of A_h (which is sparse instead of tridiagonal) is $O(n^3)$, whereas the cost of the inversion of $\underline{A}_h$ is the cost of K inversions of systems of sizes n/K, so the total cost is divided by K^2.

5.5 Optimal Stabilization Parameter

The determination of the optimal boundary conditions for the adjoint refers only to the continuous problem and not to the discretization, so that the natural boundary conditions for the adjoint are not modified by the Domain Decomposition. On the contrary, the determination of the optimal stabilization parameter κ^* has still to be discussed, since it depends on the domain decomposition.

The procedure is similar to the one presented in Section 1.7. The first equations (1.58)–(1.64) remain unchanged and are kept as such. But another adjoint must now be taken into account. $\omega_H^{\pm}$ is therefore also decomposed as :

$$\omega_H^{\pm} = \omega_H^{0\pm} + \kappa\, \omega_H^{1\pm} \tag{5.48}$$

where

$$\begin{aligned} \omega_H^{0\pm} &= -\frac{1}{2} V_H \left(\underline{L}_H^T Q_H \psi_H^{0\pm} \pm \underline{\ell}_H \right) && (5.49) \\ \omega_H^{1\pm} &= -\frac{1}{2} V_H \left(\underline{A}_H\, \underline{u}_H + \underline{L}_H^T Q_H \psi_H^{1\pm} - \underline{f}_H \right) && (5.50) \end{aligned}$$

and $\hat{\rho}_h^{\pm}$ can be written as

$$\hat{\rho}_h^{\pm} = \hat{\rho}_h^{0\pm} + \kappa\, \hat{\rho}_h^{1\pm} \tag{5.51}$$

where

$$\hat{\rho}_h^{0\pm} = \omega_H^{0\pm} \tag{5.52}$$
$$\hat{\rho}_h^{1\pm} = \omega_H^{1\pm} \tag{5.53}$$

because of (5.41). (1.68), (1.69), (1.70) and (1.71) are then replaced respectively by :

$$\underline{y}_h^{\pm} = \underline{L}_h^T Q_h\, \hat{\mu}_h^{0\pm} \pm \underline{\ell}_h + V_h^T \hat{\rho}_h^{0\pm} \tag{5.54}$$
$$\underline{z}_h^{\pm} = \underline{L}_h^T Q_h\, \hat{\mu}_h^{1\pm} - \underline{f}_h + V_h^T \hat{\rho}_h^{1\pm} \tag{5.55}$$
$$\alpha^{\pm} = \underline{y}_h^{\pm T} \underline{A}_h^{-1}\, \underline{y}_h^{\pm} \tag{5.56}$$
$$\beta^{\pm} = \underline{z}_h^{\pm T} \underline{A}_h^{-1}\, \underline{z}_h^{\pm} + 2\, f_h^T \hat{\mu}_h^{1\pm} \tag{5.57}$$

The values of $\alpha^{\pm}$ and $\beta^{\pm}$ are unchanged. Indeed, with (5.20), the pseudo-inverse of $\underline{A}_h$ is determined by :

$$\underline{A}_h^{-1} = Q\, A_h^{-1}\, Q^T \tag{5.58}$$

The notations for $\alpha^{\pm}$ and $\beta^{\pm}$ are therefore the same, with or without domain decomposition.

With the same notations as in Chapter 1, the bounds then become :

$$\eta^{\pm}(\kappa) = -\frac{1}{2\kappa}\,\alpha^{\pm} - \frac{\kappa}{2}\,\beta^{\pm} - \underline{y}_h^{\pm T} \underline{A}_h^{-1}\, \underline{z}_h^{\pm} - f_h^T \hat{\mu}^{0\pm} \pm c_h \tag{5.59}$$

which yields the same result as the theory without domain decomposition :

$$\kappa^* = \sqrt{\frac{\alpha^{\pm}}{\beta^{\pm}}} \tag{5.60}$$

On the contrary, the values of $\underline{y}^{\pm}$ and $\underline{z}^{\pm}$ are different from the values of $y^{\pm}$ and $z^{\pm}$ respectively, but because of (5.58), the value of the inner product $\underline{y}_h^{\pm T} \underline{A}_h^{-1}\, \underline{z}_h^{\pm}$ remains unchanged, so the values of the bounds are unchanged by domain decomposition.

The cost of the procedure is once again in the inversion of $\underline{A}$, and the reasoning of Section 1.7 still holds. The gain of the procedure using domain decomposition is then the same as for the resolution with any value of κ.

Conclusion

Three main conclusions can be drawn from the results obtained. First, the Hierarchical Bounds Method works and gives good results in the linear case, whether it is applied to second or first order differential equations. The bounds are sharp, even for coarse grids, and second order convergence has been observed in all cases. Second, the HBM is a *global* method that gives bounds for a fine grid output, but gives no information as regards the contribution of each element to the error. Its use in adaptative methods is thus not straightforward. Third, although the HBM also works in some cases for nonlinear problems, success is far from being guaranteed. In particular, the heuristic approach adopted in Chapter 4 does not automatically yield discrete systems that satisfy all the assumptions made in the general theory of the HBM (e.g. symmetric part positive definite).

Efforts for the future development of the HBM should therefore concentrate on two aspects of the problem. First, some further theoretical study should be conducted on the nonlinear case to come up with a formulation that either guarantees that the conditions of application of the HBM are gathered or at least that specifies the “functioning domain” of the method. Maybe the use of a simpler nonlinear equation (Burger’s equation for example) would be a good starting point. Second, the HBM needs to be implemented in two dimensions, because that is where the gains can be realized : inversion of symmetric sparse matrices instead of general matrices and Domain Decomposition. Some initial efforts in that direction are already available [13]. They seem very promising, and one can hope to see the HBM applied in the near future, for instance to validate low-order models like those developed in [14] without going to the full solution.

Bibliography

[1] J.-M. Bony : *Cours d'Analyse*. Ecole Polytechnique (1992)

[2] F.Treves : *Topological Vector Spaces, Distributions and Kernels*. Academic Press (1967)

[3] M. Raviart : *Résolution des Modèles aux Dérivées Partielles*. Ecole Polytechnique (1992)

[4] O. C. Zienkiewicz, K. Morgan : *Finite Elements and Approximation*. Wiley-Interscience (1983)

[5] M. Paraschivoiu, A. T. Patera : *A Hierarchical Duality Approach to Bounds for the Outputs of Partial Differential Equations*. Massachusetts Institute of Technology, Cambridge, MA (1996)

[6] G. Strang : *Introduction to Applied Mathematics*. Wellesley-Cambridge Press (1986)

[7] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery : *Numerical Recipes in C – The Art of Scientific Computing*, 2nd Edition. Cambridge University Press (1992)

[8] L.B. Wigton, N.J. Yu and D.P. Young : *GMRES Acceleration of Computational Fluid Dynamics Codes*. AIAA-85-1494 (1985)

[9] Y. Saad : *Krylov Subspace Techniques, Conjugate Gradients, Preconditioning and Sparse Matrix Solvers*, von Karman Institute for Fluid Dynamics Lecture Series (1994-05)

[10] B. Larrouturou, P.-L. Lions : *Méthodes Mathématiques pour les Sciences de l'Ingénieur : Optimisation et Analyse Numérique*. Ecole Polytechnique (1994)

[11] R. Codina : *Comparison of Some Finite Element Methods for Solving the Transient Convection-Diffusion Equation.* Universitat Politècnica de Catalunya

[12] P. D. Frank, G. R. Shubin : *A Comparison of Optimization-Based Approaches for a Model Computational Aerodynamics Design Problem*, Journal of Computational Physics *98*, 74-89 (1992)

[13] M. Paraschivoiu, J. Peraire, A. T. Patera : *A Posteriori Finite Element Bounds for Linear-Functional Outputs of Elliptinc Partial Differential Equations.* Symposium on Advances in Computational Mechanics, submitted to Comp. Meth. Appl. Meth. Engrg. (1997)

[14] K. Y. Tang, W. R. Graham, J. Peraire : *Active Flow Control Using a Reduced Order Model and Optimum Control.* 27th AIAA Fluid Dynamics Conference, AIAA 96-1946 (1996)

If riding in an airplane is "flying", then riding in a boat is "swimming".
To experience the element, you have to get out of the vehicle.

... And once you have tasted flight, you will walk the Earth, looking skyward,
For there you have been, and there you long to return.