ERROR ANALYSIS AND NUMERICAL EXPERIMENTS FOR PETROV-GALERKIN METHODS

B. W. SCOTNEY

NUMERICAL ANALYSIS REPORT 11/82

ABSTRACT

For diffusion-convection problems in which convection is dominant, Petrov-Galerkin methods are used to overcome the well-known inadequacies of the Galerkin method. The question arises as to how the test space should be chosen for a given trial space. We consider a factorisation of the differential operator which leads to the possibility of generating test spaces with a view to either producing a nodally accurate solution or a solution which is a best fit in a mixed norm. The Riesz Representation Theorem guarantees the existence of such spaces, and enables the relationship between the trial space - test space pairing and the optimality of the associated Petrov-Galerkin method to be investigated. In particular for the one dimensional diffusion-convection problem we establish explicit optimal error estimates for any conforming 'upwind' finite element method when a piecewise linear trial space is used, and our framework is exploited to analyse the commonly used test spaces.

For two-dimensional diffusion-convection problems a method aimed at producing a solution which is a best fit in a mixed norm is considered, and numerical experiments using this and some 'upwind' Petrov-Galerkin methods are presented.

1. INTRODUCTION

Suppose that Ω is a bounded open region in R^n with a polygonal boundary $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$, and $\partial\Omega_1 \cap \partial\Omega_2 = \phi$. Then if L is a linear second order differential operator, we will consider the problem

$$Lu = f \quad \text{in} \quad \Omega \quad , \tag{1.1}$$

$$u = g \quad \text{on} \quad \partial\Omega_1 \ , \quad \partial u/\partial n = 0 \quad \text{on} \quad \partial\Omega_2 ,$$

for a quantity u : f is a source term, and $\partial u/\partial n$ denotes differentiation in the direction of the outward normal on $\partial\Omega_2$.

1.1 Weak Formulation

Let $H^m(\Omega)$ denote the Sobolev space of functions with derivatives up to order m being square integrable over the region Ω. We suppose that $\partial\Omega$ is smooth enough for there to exist a unique, continuous, linear mapping $\gamma_0 : H^1(\Omega) \to L^2(\partial\Omega_1)$ such that for each $v \in H^1(\Omega)$, γ_0 is the restriction of v to $\partial\Omega_1$; we will suppose that g is such that there exists $G \in H^1(\Omega)$ such that $\gamma_0(G) = g$. Then $H^1_{E_0}(\Omega)$ is defined by

$$H^1_{E_0}(\Omega) = \{v \in H^1(\Omega) \mid \gamma_0 v = 0 \quad \text{on} \quad \partial\Omega_1\} , \tag{1.2a}$$

and

$$H^1_E(\Omega) = \{v \in H^1(\Omega) \mid v - G \in H^1_{E_0}(\Omega)\} . \tag{1.2b}$$

As a general reference see, for example, Ciarlet (1978).

Then the weak formulation of problem (1.1) is to find $u \in H^1_E(\Omega)$ such that

$$B(u,v) = \langle f,v \rangle \quad \forall \ v \in H^1_{E_0}(\Omega), \tag{1.3}$$

where the angular brackets $\langle w_1, w_2 \rangle$ denote the inner product $\int_\Omega w_1 w_2 d\Omega$, and $B(w_1,w_2)$ is the bilinear form resulting from integrating the second order terms of the inner product $\langle Lw_1, w_2 \rangle$ by parts, so that only derivatives

of w_1 and w_2 up to first order remain. In the case of vector arguments, $\langle \underline{w}_1, \underline{w}_2 \rangle$ denotes the integral $\int_\Omega \underline{w}_1 \cdot \underline{w}_2 \, d\Omega$.

1.2 Conforming Finite Element Method

We construct a finite dimensional subspace $S^h(\Omega)$ of $H^1(\Omega)$: $S^h(\Omega)$ is called the trial space, and is spanned by a set of basis functions $\{\phi_i, i = 1,\ldots,N\}$. Then the space $S_0^h(\Omega)$ is defined as

$$S_0^h(\Omega) = S^h(\Omega) \cap H_{E_0}^1(\Omega) . \tag{1.4}$$

We will consider only g for which there exists $G^h \in S^h(\Omega)$ such that $\gamma_0(G^h) = g$. The space $S_E^h(\Omega)$ may then be defined by

$$S_E^h(\Omega) = \{V = G^h + W \mid W \in S_0^h(\Omega)\} . \tag{1.5}$$

1.3 Galerkin Approximation

The Galerkin approximation to problem (1.3) is to find $U \in S_E^h(\Omega)$ such that

$$B(U,V) = \langle f, V \rangle \quad \forall \quad V \in S_0^h(\Omega). \tag{1.6}$$

In the case where L is a self-adjoint operator, we may write $L = T^*T$ where T^* is the formal adjoint of T. Then

$$B(v,w) = \langle Tv, Tw \rangle , \tag{1.7}$$

and the Galerkin approximation U is the optimal approximation from the trial space to the solution u of equation (1.3) in the norm $\| \cdot \|_T$ defined on $H_{E_0}^1(\Omega)$ by

$$\| \cdot \|_T^2 = \langle T\cdot, T\cdot \rangle . \tag{1.8}$$

The approximation problem now is purely one of selecting a trial space from which the solution to (1.3) can be adequately represented. As general references to the finite element approximation, see Ciarlet (1978), Strang & Fix (1973), and Zienkiewicz (1977).

2. NON-SELF-ADJOINT PROBLEMS

Conditions under which there exists a unique solution to problem (1.3) are given by the Lax-Milgram Theorem: see Ciarlet (1978). Existence and uniqueness of a solution to problem (1.6) are implied since $S_0^h(\Omega)$ is a subspace of $H^1_{E_0}(\Omega)$: see Aubin (1972) or Babuška and Aziz (1972).

Furthermore, if u is the unique solution to problem (1.3) and U the unique solution to the Galerkin equations (1.6), then

$$\| u-U \|_T \leq [1 + C_1/C_2] \inf_{V \in S_E^h(\Omega)} \| u-V \|_T, \tag{2.1}$$

where C_1 and C_2 are constants in the Lax-Milgram Theorem (see below), and the norm $\| \cdot \|_T$ is that given by (1.8), where $T^*T = \frac{1}{2}(L+L^*)$, and L^* is the formal adjoint of the operator L. As L moves away from self-adjointness, whilst the error estimate (2.1) remains of optimal order, the constant C_1/C_2 may become very large. The Galerkin formulation then becomes useless, and the problem is no longer purely one of choosing a trial space to adequately represent the solution to problem (1.3).

2.1 Petrov-Galerkin Methods

For non-self-adjoint problems Petrov-Galerkin methods have been put forward by many authors to overcome the inadequacies of the Galerkin formulation as exposed in the error estimate (2.1). These are generalisations of the Galerkin method in which a test space $T^h(\Omega) \subset H^1(\Omega)$ is employed. Setting $T_0^h(\Omega) = T^h(\Omega) \cap H^1_{E_0}(\Omega)$, the system (1.6) is replaced by the problem of finding $U \in S_E^h(\Omega)$ such that

$$B(U,V) = \langle f,V \rangle \qquad \forall\, V \in T_0^h(\Omega). \tag{2.2}$$

$T_0^h(\Omega)$ has the same dimension as the trial space $S_0^h(\Omega)$ and is spanned by a set of basis functions $\{\psi_i,\ i = 1,\ldots,N\}$. The problem of choosing a trial space $S_0^h(\Omega)$ to adequately represent the solution to problem (1.3) still remains, but it is supplemented, or rather overshadowed, by the problem of

choosing the test space $T^h(\Omega)$ to associate with a given $S^h(\Omega)$. The conditions under which there exists a unique solution to problem (2.2) are given by the Generalised Lax-Milgram Theorem (see Babuška & Aziz (1972)):-

The Generalised Lax-Milgram Theorem

Suppose that $B(\cdot,\cdot)$ is continuous on $H^1_{E_0}(\Omega) \times H^1_{E_0}(\Omega)$ and coercive on $S^h_0(\Omega) \times T^h_0(\Omega)$; i.e. suppose that $\|\cdot\|_{B_i}$ denotes a norm on $H^1_{E_0}(\Omega)$, and that there exist positive constants C_1 and C_2 such that

$$\text{(i)} \quad |B(v,w)| \le C_1 \|v\|_{B_i} \|w\|_{B_i} \qquad \forall\, v,w \in H^1_{E_0}(\Omega), \tag{2.3a}$$

$$\text{(ii)} \quad \inf_{V \in S^h_0(\Omega)} \sup_{W \in T^h_0(\Omega)} \frac{|B(V,W)|}{\|V\|_{B_i} \|W\|_{B_i}} \ge C_2, \tag{2.3b}$$

$$\text{(iii)} \quad \sup_{V \in S^h_0(\Omega)} |B(V,W)| > 0 \qquad \forall\, W \ne 0,\ W \in T^h_0(\Omega). \tag{2.3c}$$

Then there exists a unique solution U to (2.2), and the following error estimate holds:

$$\|u-U\|_{B_i} \le [1 + C_1/C_2] \inf_{V \in S^h_E(\Omega)} \|u-V\|_{B_i} . \tag{2.4a}$$

Morton (1981) has shown that the estimate (2.4a) can be improved to

$$\|u-U\|_{B_i} \le (C_1/C_2) \inf_{V \in S^h_E(\Omega)} \|u-V\|_{B_i} , \tag{2.4b}$$

and also the way in which for the Galerkin method the constant C_1 is related to the mesh Péclet number for diffusion-convection problems (see below).

In particular the tasks then are to construct $T^h_0(\Omega)$ in such a way as to guarantee existence of a solution to (2.2) for arbitrary mesh-size parameter h, and so that the constant C_1/C_2 in the estimate (2.4b) is small.

2.2 Diffusion-Convection Problems

We consider now the problem (1.1) in which the operator L takes the form

$$Lu \equiv -\underline{\nabla} \cdot (a\underline{\nabla}u - \underline{b}u) \quad \text{in} \quad \Omega \tag{2.5}$$

for a quantity u. Here, $a(\underline{x})$ is a scalar diffusion coefficient, $\underline{b}(\underline{x})$ is a vector convective velocity, and so $Lu = 0$ represents a conservation law for u. In physical problems of this type, u may be the concentration of a pollutant in a river, in which case the Péclet number, that is the ratio $|\underline{b}|\, L/a$ where L is a typical length in the domain, will typically be in the range $10^2 - 10^3$. Or in a cooling problem, u will represent the temperature of the coolant: for example, liquid sodium in a nuclear reactor, where the Péclet number will be of the order $2\text{-}3 \times 10^3$ (Wakil, 1962).

One can show (see Morton (1981)) that for L as in (2.5), there exists a unique solution to problem (1.3) provided that the following conditions hold:

$$\left.\begin{array}{ll} \text{(i)} & \underline{b}\cdot\underline{n} \geq 0 \quad \text{on} \quad \partial\Omega_2, \quad \text{where} \quad \underline{n} \quad \text{is the outward normal on} \quad \partial\Omega_2, \\ \text{(ii)} & f \in L_2(\Omega), \\ \text{(iii)} & 0 < a \in C^0(\bar{\Omega}), \quad \text{where} \quad \bar{\Omega} \quad \text{denotes the closure of} \quad \Omega, \\ \text{(iv)} & \underline{b} \in [H^1(\Omega)]^2 \quad \text{and} \quad \underline{\nabla}\cdot\underline{b} = 0. \end{array}\right\} \tag{2.6}$$

2.3 Factorisation and Riesz Representers

Under the conditions (2.6) the weak solution u to problem (1.3) may be written as $u = u_0 + G$ with $u_0 \in H^1_{E_0}(\Omega)$. Then u_0 satisfies

$$B(u_0, w) = F(w) \qquad \forall\, w \in H^1_{E_0}(\Omega), \tag{2.7}$$

where

$$B(w_1, w_2) \equiv \langle a\underline{\nabla}w_1, \underline{\nabla}w_2\rangle + \langle \underline{\nabla}\cdot(\underline{b}w_1), w_2\rangle \qquad \forall\, w_1, w_2 \in H^1_{E_0}(\Omega), \tag{2.8}$$

and

$$F(w) \equiv \langle f - \underline{\nabla}\cdot(\underline{b}G), w\rangle - \langle a\underline{\nabla}G, \underline{\nabla}w\rangle \qquad \forall\, w \in H^1_{E_0}(\Omega). \tag{2.9}$$

It is convenient to introduce the operators

$$T_1 \equiv a^{\frac{1}{2}}\,\underline{\nabla} \qquad \text{and} \qquad T_2 \equiv a^{\frac{1}{2}}\,\underline{\nabla} - \underline{b}/a^{\frac{1}{2}}, \tag{2.10}$$

(see Morton (1981)). We can then write the Petrov-Galerkin formulation (2.2)

as: find $U \in S_0^h(\Omega)$ such that

$$\langle T_2 U, T_1 V\rangle + \int_{\partial\Omega_2} \underline{b}\cdot\underline{n} UV d\Omega_2 = F(V) \qquad \forall\, V \in T_0^h(\Omega). \tag{2.11}$$

Let $H_1(\Omega)$ denote the Hilbert space $H^1_{E_0}(\Omega)$ equipped with the inner product

$$\langle v,w\rangle_1 \equiv \langle a^{\frac{1}{2}}\underline{\nabla}v,\ a^{\frac{1}{2}}\underline{\nabla}w\rangle = \langle T_1 v, T_1 w\rangle \tag{2.12}$$

and $H_2(\Omega)$ denote the Hilbert space $H^1_{E_0}(\Omega)$ equipped with the inner product

$$\langle v,w\rangle_2 \equiv \langle a\underline{\nabla}v, \underline{\nabla}w\rangle + \langle(\underline{b}\cdot\underline{b}/a)v,w\rangle \tag{2.13}$$

$$= \langle(a^{\frac{1}{2}}\underline{\nabla}-\underline{b}/a^{\frac{1}{2}})v,\ (a^{\frac{1}{2}}\underline{\nabla}-\underline{b}/a^{\frac{1}{2}})w\rangle + \int_{\partial\Omega_2} \underline{b}\cdot\underline{n} vw d\Omega_2$$

$$= \langle T_2 v, T_2 w\rangle + \int_{\partial\Omega_2} \underline{b}\cdot\underline{n} vw d\Omega_2 \ . \tag{2.14}$$

Consider the bounded linear functional

$$A_{w_2}(w_1) \equiv \langle T_2 w_1,\ T_1 w_2\rangle + \int_{\partial\Omega_2} \underline{b}\cdot\underline{n}\ w_1 w_2 d\Omega_2 \quad \forall\, w_1,w_2 \in H^1_{E_0}(\Omega). \tag{2.15}$$

Then, using the Riesz Representation Theorem and the continuity of $B(.,.)$ in $H_1(\Omega)$, there exists $R_1 : H_1(\Omega) \to H_1(\Omega)$ such that

$$\langle T_2 w_1,\ T_1 w_2\rangle + \int_{\partial\Omega_2} \underline{b}\cdot\underline{n}\ w_1 w_2 d\Omega_2 = \langle T_1 w_1,\ T_1(R_1 w_2)\rangle \quad \forall\, w_1,w_2 \in H_1(\Omega). \tag{2.16}$$

That is, $R_1 w_2$ is the Riesz representer of A_{w_2} in $H_1(\Omega)$. Similarly, there exists $R_2 : H_2(\Omega) \to H_2(\Omega)$ such that

$$\langle T_2 w_1,\ T_1 w_2\rangle + \int_{\partial\Omega_2} \underline{b}\cdot\underline{n} w_1 w_2 d\Omega_2 = \langle T_2 w_1,\ T_2(R_2 w_2)\rangle + \int_{\partial\Omega_2} \underline{b}\cdot\underline{n} w_1 (R_2 w_2) d\Omega_2 \quad \forall\, w_1,w_2 \in H_2($$

(2.17)

That is, $R_2 w_2$ is the Riesz representer of A_{w_2} in $H_2(\Omega)$. Hence we may consider generating an approximation $U \in S_0^h(\Omega)$ to the problem (2.11) using

a test space related to the trial space through the transformation

$$R_i\, T_0^h(\Omega) = S_0^h(\Omega), \qquad i = 1 \text{ or } 2. \tag{2.18}$$

We note that the existence of an inverse $R_i^{-1} : H_i(\Omega) \to H_i(\Omega)$, $i = 1,2$, is guaranteed by the coercivity of $B(\cdot,\cdot)$. (See Babuška and Aziz (1972)). Then using (2.17), problem (2.11) may be written as: find $U \in S_0^h(\Omega)$ such that

$$\langle T_2 U,\, T_2(R_2 V)\rangle + \int_{\partial\Omega_2} \underline{b}\cdot\underline{n}\, U(R_2 V)\, d\Omega_2 = F(V) \qquad \forall\, V \in T_0^h(\Omega). \tag{2.19}$$

Thus if the choice of test space in (2.18) with $i = 2$ could be achieved exactly, the approximation obtained from solving (2.19) is an optimal approximation from $S_0^h(\Omega)$ to u_0 in the norm $\| a^{\frac{1}{2}} \underline{\nabla}\cdot \|^2 + \| |\underline{b}|/a^{\frac{1}{2}} \cdot \|^2$ defined by T_2. Alternatively (2.16) may be used to write problem (2.11) as: find $U \in S_0^h(\Omega)$ such that

$$\langle T_1 U, T_1(R_1 V)\rangle = F(V) \qquad \forall\, V \in T_0^h(\Omega). \tag{2.20}$$

Hence if the choice of test space in (2.18) with $i = 1$ could be achieved exactly, the approximation $U \in S_0^h(\Omega)$ obtained by solving (2.20) is an optimal approximation from $S_0^h(\Omega)$ to u_0 in the norm $\| a^{\frac{1}{2}} \underline{\nabla}\cdot \|$ defined by T_1.

3. ERROR ANALYSIS FOR PETROV-GALERKIN SCHEMES

The following estimate, established by Morton (1981), relates the optimality of a Petrov-Galerkin method to the matching of a test space to a given trial space.

3.1 Basic Estimate

Suppose that $B(\cdot,\cdot) : H_m(\Omega) \times H_m(\Omega) \to R$ is a continuous and coercive bilinear form, where $H_m(\Omega)$ is $H^1_{E_0}(\Omega)$ equipped with an inner-product $\langle \cdot,\cdot \rangle_m$. Then by the Riesz Representation Theorem, for each $v \in H_m(\Omega)$ there exists a map $R_m : H_m(\Omega) \to H_m(\Omega)$ such that

$$B(v,w) = \langle v, R_m w \rangle_m \qquad \forall\, w \in H_m(\Omega). \tag{3.1}$$

Then if the constant Δ_m is defined by

$$\inf_{w \in T_0^h(\Omega)} \| V - R_m W \|_m \le \Delta_m \| V \|_m \qquad \forall\, V \in S_0^h(\Omega), \tag{3.2}$$

and U is the Petrov-Galerkin solution to problem (2.2), and u is the solution to (1.3), the following estimate holds:

$$\| u-U \|_m \le (1 - \Delta_m^2)^{-\frac{1}{2}} \inf_{V \in S_E^h(\Omega)} \| u-V \|_m \ . \tag{3.3}$$

3.2 Expressions for the Operators R_m in One Dimension

We will consider the one-dimensional Dirichlet problem:

$$-au'' + bu' = f \quad \text{on} \quad (0,1) \tag{3.4a}$$

$$u(0) = g_L, \ u(1) = g_R \ , \tag{3.4b}$$

where a and b are positive constants.

With the $\| \cdot \|_1$ norm as in (2.12), it follows that, for $B(\cdot,\cdot)$ as in (2.8) and R_1 defined by (3.1),

$$(R_1 v)(x) = v(x) + (b/a)\int_0^x (v(t) - \bar{v})\,dt, \tag{3.5}$$

where $\bar{v} = \int_0^1 v(t)\,dt$. (See Morton (1981)).

Similarly, with the $\| \cdot \|_2$ norm as in (2.13), the operator R_2 defined by (3.1) may be explicitly written as

$$(R_2 v)(x) = v(x) + (b/a)\left(\int_0^x v(t) e^{b(x-t)/a}\,dt - K \sinh(bx/a)\right), \tag{3.6}$$

where

$$K = \int_0^1 e^{b(1-t)/a}\, v(t)\,dt \ / \ (\sinh(b/a)). \tag{3.7}$$

We note that, whilst the expression for $(R_1 v)(x)$ in (3.5) is considerably simpler than that in (3.6) and (3.7) for $(R_2 v)(x)$, the inverse

$$(R_1 w)^{-1}(x) = w(x) - (b/a)\left(\int_0^x w(t) e^{b(t-x)/a}\,dt - \left(\frac{e^{bx/a}-1}{e^{b/a}-1}\right)\int_0^1 w(t) e^{b(t-x)/a}\,dt\right)$$

is correspondingly more complicated than the inverse

$$(R_2 w)^{-1}(x) = w(x) - (b/a) \int_0^x w(t)dt - \frac{1-e^{-bx/a}}{1-e^{-b/a}} \int_0^1 w(t)dt \quad .$$

3.3 The Calculation of Δ_m

For a given trial space and test space pair, the degree to which the Petrov-Galerkin solution fails to achieve optimality is described by the constant Δ_m in the estimate (3.3). From (3.2) the smallest constant is given by

$$\Delta_m = \sup_{V \in S_0^h(\Omega)} \inf_{W \in T_0^h(\Omega)} \frac{\| V-R_m W \|_m}{\| V \|_m} \quad . \tag{3.8}$$

It will be assumed that $S_0^h(\Omega)$ and $T_0^h(\Omega)$ are of dimension N, so that any element $V \in S_0^h(\Omega)$ may be written as

$$V = \sum_{i=1}^{N} V_i \phi_i \quad , \tag{3.9}$$

and any element $W \in T_0^h(\Omega)$ may be written as

$$W = \sum_{i=1}^{N} W_i \psi_i \quad . \tag{3.10}$$

The N-vector $\underline{V}$ will denote the vector with components V_j, $j = 1,\ldots,N$, and $\underline{W}$ the vector with components W_j, $j = 1,\ldots,N$, etc. If we consider calculating $\inf_{W \in T_0^h(\Omega)} \| V-R_m W \|_m$ for a fixed element $V \in S_0^h(\Omega)$, we may write

$$\| V-R_m W \|_m^2 = \| V \|_m^2 - \| R_m W \|_m^2 + 2 \langle R_m W, R_m W - V \rangle_m \quad \forall\, W \in T_0^h(\Omega). \tag{3.11}$$

Then if $W^* \in T_0^h(\Omega)$ is the element which minimises $\| V-R_m W \|_m$, we have

$$\langle R_m W^* - V, R_m W \rangle_m = 0 \qquad \forall\, W \in T_0^h(\Omega), \tag{3.12}$$

and hence

$$\| V-R_m W^* \|_m^2 = \| V \|_m^2 - \| R_m W^* \|_m^2. \tag{3.13}$$

If the interval (0,1) is discretised into elements, with a set of $N+1$ nodal positions $\{x_j,\ j = 0,\ldots,N+1\}$ and $x_0 = 0$, $x_{N+1} = 1$, we may associate with node j the trial and test basis functions ϕ_j and ψ_j respectively. We then introduce the three $N \times N$ matrices A, B and C whose (i,j) entries are given by

$$\left.\begin{aligned} A_{ij} &= \langle R_m \psi_i,\ R_m \psi_j \rangle_m \\ B_{ij} &= \langle R_m \psi_i,\ \phi_j \rangle_m \\ C_{ij} &= \langle \phi_i,\ \phi_j \rangle_m \end{aligned}\right\} \begin{array}{l} i = 1,\ldots,N \\ j = 1,\ldots,N \end{array} \qquad (3.14)$$

and

respectively. Equation (3.13) may then be written in the form

$$\inf_{W \in T_0^h(\Omega)} \| V - R_m W \|_m^2 = \underline{V}^T C \underline{V} - \underline{W^*}^T A \underline{W^*} \ . \qquad (3.15)$$

Combining (3.15) with (3.8) we have

$$\Delta_m^2 = \sup_{V \in S_0^h(\Omega)} \frac{(\underline{V}^T C \underline{V} - \underline{W^*}^T A \underline{W^*})}{\underline{V}^T C \underline{V}}$$

$$= \sup_{\underline{V}} \left(1 - \frac{\underline{W^*}^T A \underline{W^*}}{\underline{V}^T C \underline{V}} \right) \ . \qquad (3.16)$$

The defining equation (3.12) for W^* may be written in the form

$$A \underline{W^*} = B \underline{V} \ . \qquad (3.17)$$

Substituting this into (3.16) and using the fact that A is symmetric and positive definite, gives

$$\Delta_m^2 = \sup_{\underline{V}} \left(1 - \frac{\underline{V}^T B^T A^{-1} B \underline{V}}{\underline{V}^T C \underline{V}} \right) \ . \qquad (3.18)$$

Since C is symmetric and positive definite, we may make the transformation $\underline{X} = C^{\frac{1}{2}} \underline{V}$, and hence

$$\Delta_m^2 = \sup_{\underline{X}} \left(1 - \frac{\underline{X}^T Q^T Q \underline{X}}{\underline{X}^T \underline{X}} \right) \ , \qquad (3.19)$$

where $Q = A^{-\frac{1}{2}} B\, C^{-\frac{1}{2}}$.

We could then obtain Δ_m^2 by computing the largest eigenvalue of the matrix $I - Q^TQ$, but such a method has the disadvantage that the square roots of the symmetric matrices A and C have to be computed. Alternatively, since C is symmetric and positive definite we could compute the smallest eigenvalue λ_m satisfying the generalised eigenvalue problem $B^TA^{-1}B\underline{V} = \lambda_m C\underline{V}$. (See, for example, Wilkinson (1965)). In the numerical calculations described later in this section, the more direct approach of obtaining $1/\Delta_m^2$ by minimising the expression

$$\frac{\underline{V}^T C \underline{V}}{\underline{V}^T(C-B^TA^{-1}B)\underline{V}} \tag{3.20}$$

as a quotient of two quadratic expressions in the N variables $V_1,\ldots,V_N$ was used. In the Section 3.5, however, for the norm $\|\cdot\|_1$, this eigenvalue problem is solved analytically for a large class of Petrov-Galerkin methods. The adoption of such an approach not only allows the constant Δ_m to be easily computed, but exposes the structure underlying the problem in this norm.

3.4 Analysis in One Dimension

From the definition (3.8) it is clear that $0 \le \Delta_m^2 < 1$. Further, if $\Delta_m = 0$, from (3.3) we have for the Petrov-Galerkin approximation

$$\|u-U\|_m = \inf_{V \in S_E^h(\Omega)} \|u-V\|_m \quad :$$

the trial space-test space relationship (2.18) has been satisfied exactly, and the optimal test space in the sense of reducing the constant in the estimate (3.3) has been found. Where $\Delta_m \neq 0$, the aim is to calculate the degree to which optimality in the estimate (3.3) is lost through having a test space which does not satisfy (2.18) exactly. In this case,

$$0 < \Delta_m^2 < 1,$$

and hence from (3.18) we have

$$0 < \frac{\underline{V}^T (B^TA^{-1}B)\underline{V}}{\underline{V}^T C \underline{V}} < 1 \qquad \forall\, \underline{V} \neq \underline{0}\ . \tag{3.21}$$

In the case when $m = 1$, for a piecewise linear trial space we note that the matrix C as defined by (3.14) is a tridiagonal $N \times N$ matrix of the form

$$\frac{a}{h}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & 0 \\ & -1 & 2 & -1 & & \\ & & & \ddots & \ddots & \ddots \\ & & & & -1 & 2 & -1 \\ & 0 & & & & -1 & 2 \end{bmatrix} \tag{3.22}$$

In the case of $m = 2$, C is again tridiagonal, with diagonal entries

$$2a/h + 2b^2h/3a \tag{3.23a}$$

and upper and lower diagonal entries

$$-a/h + b^2h/6a. \tag{3.23b}$$

Since $\langle \cdot, \cdot \rangle_m$, $m = 1,2$, is a norm, C is positive definite, and so in both cases,

$$\underline{V}^T C \underline{V} > 0 \qquad \forall\, \underline{V} \neq \underline{0}. \tag{3.24}$$

Combining (3.24) with (3.21) gives

$$\underline{V}^T (B^T A^{-1} B) \underline{V} > 0 \qquad \forall\, \underline{V} \neq \underline{0}. \tag{3.25}$$

Since C is symmetric and positive definite, there exists a matrix $C^{\frac{1}{2}}$, with an inverse $C^{-\frac{1}{2}}$. Hence for any $\underline{V}$ there exists a vector $\underline{X}$ such that $\underline{V} = C^{-\frac{1}{2}}\underline{X}$, and Δ_m^2 can be written as in (3.19). Hence

$$\Delta_m^2 = 1 - \lambda_m , \tag{3.26}$$

where λ_m is the smallest generalised eigenvalue satisfying

$$B^T A^{-1} B \ \underline{Y}_m = \lambda_m C \ \underline{Y}_m . \tag{3.27}$$

The complicated nature of R_1^{-1} leads to difficulty if this form involving A^{-1} is used for the analysis in the case $m = 1$. Consequently, having established (3.25) we may set $\underline{Y}_m = C^{-1}B^T\underline{P}_m$ and premultiply (3.27) by the matrix $(B^TA^{-1})^{-1}$ to give

$$BC^{-1}\,B^T\,\underline{P}_m = \lambda_m\,A\,\underline{P}_m. \tag{3.28}$$

The analysis to equations (3.27) and (3.28) is applicable for any choice of conforming test space $T_0^h(\Omega)$ for either the case $m = 1$, or $m = 2$. The remainder of this section is concerned with using equation (3.28) to establish optimal error bounds of the type (3.3) in the case $m = 1$ for a general class of Petrov-Galerkin methods, namely 'upwind' finite elements.

3.5 'Upwind' Finite Elements

Consider the problem (3.4) on a uniform mesh of size h, with N interior nodes $\{x_j = jh,\ j = 1,\ldots,N\}$ and $x_0 = 0$, $x_{N+1} = 1$. Using the standard finite difference notation

$$\begin{aligned}\Delta_0 U_j &= (U_{j+1} - U_{j-1})/2\,,\\ \Delta_+ U_j &= U_{j+1} - U_j,\\ \Delta_- U_j &= U_j - U_{j-1},\end{aligned} \tag{3.29}$$

and

$$\delta^2 U_j = -U_{j-1} + 2U_j - U_{j+1}\,,$$

the left-hand side of (3.4) may be replaced by the difference operator

$$h^{-2}(a\,\delta^2 + bh(\alpha\Delta_- + (1-\alpha)\Delta_0))\,U_j \quad . \tag{3.30}$$

Choosing

$$\alpha = \coth(bh/2a) - (2a/bh) \tag{3.31}$$

gives the difference scheme first proposed by Allen & Southwell (1955); if f is constant this scheme will produce a nodally exact solution.

Numerous 'upwind' Petrov-Galerkin methods have been proposed with a view to reproducing the difference operator (3.30) on a uniform mesh for problem (3.4), and hence achieving nodal accuracy: see, for example, Il'in (1969),

Christie et al (1976), Hemker (1977), Heinrich et al (1977), Barrett (1977), Hughes (1978), Kellogg & Tsan (1978), Dixon, Harrison & Morgan (1979), Griffiths & Mitchell (1979), Hughes & Brooks (1979), Heinrich & Zienkiewicz (1979), Axelsson (1981), Kellogg (1980), Brooks (1981).

One may easily show that in the case of a piecewise linear trial space $S_E^h(\Omega)$, the best fit from this space to a function u in the norm $\|\cdot\|_1$ is nodally exact: the best fit $U^* \in S_E^h(\Omega)$ is defined by

$$\int_0^1 (u' - U^{*\prime})\phi_j' \, dx = 0, \qquad j = 1,\ldots,N \tag{3.32}$$

where

$$\phi_j(x) = \begin{cases} (x_{j+1}-x)/h_{j+1} , & x_j \le x \le x_{j+1} , \\ (x - x_{j-1})/h_j , & x_{j-1} \le x < x_j , \\ 0 , & \text{elsewhere} , \end{cases} \tag{3.33}$$

and $h_j = x_j - x_{j-1}$, $j = 1,\ldots,N+1$. Because ϕ_j' is constant on each element, equation (3.32) may be written as

$$\Delta_- (u(x_j) - U_j^*)/\Delta_- x_j = \Delta_+ (u(x_j) - U_j^*)/\Delta_+ x_j . \tag{3.34}$$

This implies that $(u(x_j) - U_j^*)/h_j - (u(x_{j-1}) - U_{j-1}^*)/h_j$ is a constant C, say, for $j = 1,\ldots,N+1$. Hence, summing from $j = 1$ to $j = N+1$ gives $C = 0$, and so

$$(u(x_j) - U_j^*)/h_j = (u(x_{j-1}) - U_{j-1}^*)/h_j, \quad j = 1,\ldots,N+1. \tag{3.35}$$

Since $u(x_0) = U_0^*$, then $u(x_j) = U_j^*$ for $j = 0,\ldots,N+1$.

It is thus particularly appropriate that the norm $\|\cdot\|_1$ should be used in the analysis of 'upwind' Petrov-Galerkin schemes.

In an 'upwind' finite element method, the test space $T_0^h(\Omega)$ is spanned by a set of basis functions $\{\psi_j, j = 1,\ldots,N\}$ of the form

$$\psi_j(x) = \phi_j(x) + \alpha_j(x), \qquad j = 1,\ldots,N, \tag{3.36}$$

where ϕ_j is as in (3.33) with $h_j = h$ for all j, and $\alpha_j(x)$ has the shifted form

$$\alpha_j(x) = \begin{cases} 0 & , \quad x < x_j - h \ , \\ \alpha((x - x_j + h)/h) & , \quad x_j - h \le x \le x_j \ , \\ -\alpha_j(x - h) & , \quad x_j < x \le x_j + h \ , \\ 0 & , \quad x_j + h < x \ . \end{cases} \tag{3.37}$$

For a conforming finite element method, $\alpha(t)$ may be any continuous function on the interval $[0,1]$, with the property that

$$\alpha(0) = \alpha(1) = 0.$$

The sign of $\int_0^1 \alpha(t)dt$ will depend on the sign of b in equation (3.4a). Typically, if b is positive, $\alpha(t)$ will be of the form

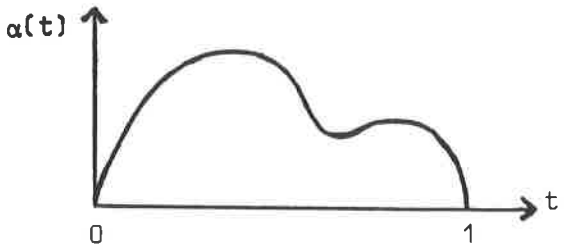


Fig. 3.1

Then $\alpha_j(x)$ takes the form

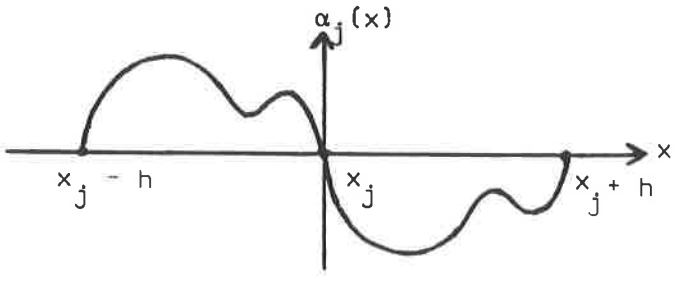


Fig. 3.2

on the interval $(x_j - h, x_j + h)$.

Note that despite the fact that many of the well-known 'upwind' methods use an $\alpha(t)$ which is even about $t = \frac{1}{2}$, such a restriction is not necessary in the following analysis.

Some results concerning the structure of the matrices A, B and C in (3.14) will now be established to enable the equation (3.28) to be fully exploited.

Lemma 3.1

Let I denote the N x N identity matrix, and E denote the N x N matrix in which each entry is unity. Then for the non-dimensionalised problem (3.4) in which $b \equiv 1$, and the uniform mesh size is h, the N x N matrix A in (3.14) has the form

$$A = (h/a)(I - hE + kC) \quad , \tag{3.38}$$

where

$$k = -a\int_0^1 \psi_j'\psi_{j\pm1}'dx - (1/a)\int_0^1 \psi_j\psi_{j\pm1}dx \quad . \tag{3.39}$$

Proof

For $j = 1,\ldots,N$, combining (3.5) with (3.14) gives a diagonal entry A_{jj} of the form

$$A_{jj} = (1/a)\int_0^1 (a\psi_j' + \psi_j - \bar{\psi}_j)(a\psi_j' + \psi_j - \bar{\psi}_j)dx \quad . \tag{3.40}$$

Noting that $\bar{\psi}_j = \int_0^1 \psi_j dx = h$, $\int_0^1 \psi_j' dx = 0$, and $\int_0^1 \psi_j\psi_j' dx = \tfrac{1}{2}\int_0^1 d/dx\,(\psi_j)^2 dx = 0$, (3.40) may be written as

$$A_{jj} = a\int_0^1 \psi_j'^2 dx + (1/a)\int_0^1 \psi_j^2\,dx - h^2/a \quad . \tag{3.41}$$

Also noting that $\int_0^1 (\psi_j + \psi_{j+1})^2 dx = 2\int_0^1 \psi_j^2 dx + 2\int_0^1 \psi_j\psi_{j+1}dx$,

and that $\psi_j + \psi_{j+1} = \phi_j + \phi_{j+1} = 1$ in (x_j, x_{j+1}),

so that $\int_0^1 (\psi_j + \psi_{j+1})^2 dx = \int_0^1 \psi_j^2 dx + h$, gives

$$\int_0^1 \psi_j^2 dx = h - 2\int_0^1 \psi_j\psi_{j+1}dx. \tag{3.42}$$

Further, since $\psi_j' + \psi_{j+1}' = 0$ in (x_j, x_{j+1}),

$$\int_0^1 \psi_j'^2 dx + 2\int_0^1 \psi_j'\psi_{j+1}' dx = \int_0^1 \psi_j'(\psi_j' + \psi_{j+1}' + \psi_{j-1}')$$

$$= 0,$$

giving

$$\int_0^1 \psi_j'^2 dx = -2\int_0^1 \psi_j'\psi_{j+1}' dx \quad . \tag{3.43}$$

Hence on substituting (3.42) and (3.43) into (3.41) we obtain

$$A_{jj} = h/a - 2a\int_0^1 \psi_j'\psi_{j+1}' dx - (2/a)\int_0^1 \psi_j\psi_{j+1} dx - h^2/a, \quad j = 1,\ldots,N. \tag{3.44}$$

For the upper diagonal entry A_{jj+1}, $j = 1,\ldots,N-1$, combining (3.5) with (3.14) shows that

$$A_{jj+1} = (1/a)\int_0^1 (a\psi_j' + \psi_j - h)(a\psi_{j+1}' + \psi_{j+1} - h)dx. \tag{3.45}$$

Equation (3.45) may be written as

$$A_{jj+1} = a\int_0^1 \psi_j'\psi_{j+1}' dx + \int_0^1 d/dx\,(\psi_j\psi_{j+1})dx + (1/a)\int_0^1 \psi_j\psi_{j+1} dx - h^2/a.$$

But $\int_0^1 d/dx\,(\psi_j\psi_{j+1})dx = 0$, giving

$$A_{jj+1} = a\int_0^1 \psi_j'\psi_{j+1}' dx + (1/a)\int_0^1 \psi_j\psi_{j+1} dx - h^2/a. \tag{3.46}$$

The lower diagonal entries A_{jj-1}, $j = 2,\ldots,N$ may be obtained through the symmetry of A.

Finally, for $i \neq j$, $j \pm 1$,

$$A_{ji} = (1/a)\int_0^1 (a\psi_j' + \psi_j - h)(a\psi_i' + \psi_i - h)dx. \tag{3.47}$$

Noting that the support of ψ_j nowhere coincides with that of ψ_i reduces (3.47) to

$$A_{ij} = -h^2/a \quad . \tag{3.48}$$

Hence combining (3.44), (3.46) and (3.48) produces the desired result. ■

Lemma 3.2

With I, E, h and b as in Lemma 3.1, and the N x N matrices B and C as in (3.14), the N x N matrix $BC^{-1}B^T$ has the form

$$BC^{-1}B^T = (h/a)(I - hE + \gamma C), \tag{3.50}$$

where

$$\gamma = \left(\frac{1}{ah}\right)\left[a + \int_{x_{j-1}}^{x_j} \alpha_j dx\right]^2 - \frac{h}{4a} \tag{3.51}$$

Proof

First we note that

$$\int_{x_j}^{x_{j+1}} (a\psi_j' + \psi_j)dx = \int_{x_j}^{x_{j+1}} (a\phi_j' + a\alpha_j' + \phi_j + \alpha_j)\, dx$$

$$= -a + \tfrac{1}{2}h + \int_{x_j}^{x_{j+1}} \alpha_j\, dx\,,$$

and

$$\int_{x_{j-1}}^{x_j} (a\psi_j' + \psi_j)dx = a + \tfrac{1}{2}h + \int_{x_{j-1}}^{x_j} \alpha_j\, dx.$$

Then the diagonal element B_{jj}, $j = 1,\ldots,N$ of the matrix B from (3.5) and (3.14) is given by

$$B_{jj} = \int_0^1 (a\psi_j' + \psi_j - h)\phi_j'\, dx \tag{3.52}$$

$$= (1/h)\int_{x_{j-1}}^{x_j} (a\psi_j' + \psi_j)dx - (1/h)\int_{x_j}^{x_{j+1}} (a\psi_j' + \psi_j)dx$$

$$= 2a/h + (1/h)\int_{x_{j-1}}^{x_j} \alpha_j dx - (1/h)\int_{x_j}^{x_{j+1}} \alpha_j dx,$$

from which we obtain

$$B_{jj} = 2a/h + (2/h) \int_{x_{j-1}}^{x_j} \alpha_j \, dx, \qquad j = 1,\ldots,N. \tag{3.53}$$

In a similar manner we obtain, for $j = 1,\ldots N-1$,

$$B_{jj+1} = -a/h + \tfrac{1}{2} - (1/h) \int_{x_{j-1}}^{x_j} \alpha_j \, dx, \tag{3.54}$$

and for $j = 2,\ldots,N$,

$$B_{jj-1} = -a/h - \tfrac{1}{2} - (1/h) \int_{x_{j-1}}^{x_j} \alpha_j \, dx. \tag{3.55}$$

These are clearly the only non-zero elements of B, and we see that

$$B_{jj} = 1 - 2\,B_{jj+1} \quad \text{and} \quad B_{jj-1} = B_{jj+1} - 1. \tag{3.56}$$

Thus B may be written as

$$B = -(ph/a)C + G \ , \tag{3.57}$$

where $p = B_{jj+1}$ and G is the $N \times N$ matrix with entries of 1 on the diagonal, -1 on the lower diagonal, and zero elsewhere, so that

$$G + G^T = (h/a)\, C. \tag{3.58}$$

Clearly

$$C^{-1}B^T = -(ph/a)I + C^{-1}G^T,$$

and hence

$$\begin{aligned} BC^{-1}B^T &= (ph/a)^2 C - (ph/a)(G + G^T) + GC^{-1}G^T \\ &= p(p-1)(h/a)^2 C + GC^{-1}G^T. \end{aligned} \tag{3.59}$$

It is clear that G^{-1} is the $N \times N$ triangular matrix with entries of 1 on the diagonal and below, and that $G^{T^{-1}}$ is the $N \times N$ triangular matrix with entries of 1 on the diagonal and above, so that

$$G^{T^{-1}} + G^{-1} = I + E \ . \tag{3.60}$$

Hence

$$\begin{aligned} G^{T^{-1}} C G^{-1} &= (a/h)\, G^{T^{-1}} (G + G^T) G^{-1} \\ &= (a/h)\,(I + E). \end{aligned} \tag{3.61}$$

But since $E^2 = NE$ and $h(1 + N) = 1$ we have

$$\begin{aligned}(I - hE)(I + E) &= I + (1 - h)E - hE^2 \\ &= I + (1 - h(1 + N))E \\ &= I. \end{aligned} \tag{3.62}$$

Thus

$$\begin{aligned} GC^{-1}G^T &= (h/a)(I + E)^{-1} \\ &= (h/a)(I - hE), \end{aligned} \tag{3.63}$$

enabling (3.59) to be written as

$$BC^{-1}B^T = (h/a)(I - hE + (p(p-1)h/a)C). \tag{3.64}$$

■

The forms (3.38) and (3.50) can now be used to write equation (3.28) in the following way:

$$(I - hE + \gamma C)\,\underline{P}_m = \lambda_m\,(I - hE + kC)\,\underline{P}_m. \tag{3.65}$$

Rearrangement of (3.65) gives

$$(1 - \lambda_m)(I - hE)\,\underline{P}_m = (\lambda_m k - \gamma)\,C\,\underline{P}_m\,, \tag{3.66}$$

and hence, noting that $0 < 1 - \lambda_m < 1$, and using (3.62) we have

$$((\lambda_m k - \gamma) / (1 - \lambda_m))(I + E)\,C\,\underline{P}_m = \underline{P}_m. \tag{3.67}$$

We will next show in Lemma 3.3 that the function $(\lambda\, k-\gamma)/(1-\gamma)$ is a monotonic function of λ, so that we can relate the smallest eigenvalue λ_m to the largest eigenvalue of the matrix $(I+E)C$.

Lemma 3.3

With k and γ defined as in (3.39) and (3.51) respectively, the function

$$f(\lambda) = (\lambda\, k-\gamma)/(1-\gamma) \tag{3.68}$$

is a monotonic increasing function of λ for any upwind test functions satisfying (3.36) and (3.37).

Proof

We first note that

$$\partial f/\partial\lambda = (k-\gamma)/(1-\lambda)^2 \ . \tag{3.69}$$

Combining (3,39), (3.50) and (3.64), we may write

$$a(k-\gamma) = -a^2 \int_{x_j}^{x_{j+1}} \psi_j'\psi_{j+1}'dx - \int_{x_j}^{x_{j+1}} \psi_j\psi_{j+1}dx - p(p-1)h, \tag{3.70}$$

where

$$p = \frac{1}{h}\int_{x_j}^{x_{j+1}} (a\psi_j' + \psi_j)dx \ . \tag{3.71}$$

Noting that on the interval (x_j,x_{j+1}),

$$\psi_{j+1} = 1 - \psi_j \qquad \text{and} \qquad \psi_{j+1}' = -\psi_j'$$

and using (3.71) enables (3.70) to be written as

$$a(k-\gamma) = a^2 \int_{x_j}^{x_{j+1}} \psi_j'^2\,dx - \int_{x_j}^{x_{j+1}} \psi_j(1-\psi_j)dx + \int_{x_j}^{x_{j+1}} (a\psi_j' + \psi_j)dx$$

$$- \frac{1}{h}\left[\int_{x_j}^{x_{j+1}} (a\psi_j' + \psi_j)dx\right]^2 \ . \tag{3.72}$$

Hence by using the Cauchy-Schwarz inequality we obtain

$$a(k-\gamma) \geq a^2 \int_{x_j}^{x_{j+1}} \psi_j'^2dx + \int_{x_j}^{x_{j+1}} \psi_j^2dx - a - \int_{x_j}^{x_{j+1}} (a\psi_j' + \psi_j)^2dx$$

$$= -a - 2a\int_{x_j}^{x_{j+1}} \psi_j\psi_j'\,dx \ . \tag{3.73}$$

We now note that since ϕ_j' is piecewise constant on (x_j,x_{j+1}) and $\alpha_j(x_j) = \alpha_j(x_{j+1}) = 0$, using (3.36) gives

$$\int_{x_j}^{x_{j+1}} \psi_j\alpha_j'dx = \int_{x_j}^{x_{j+1}} \phi_j\alpha_j'dx + \tfrac{1}{2}\int_{x_j}^{x_{j+1}} (\alpha_j^2)'dx = -\int_{x_j}^{x_{j+1}} \phi_j'\alpha_j dx = \frac{1}{h}\int_{x_j}^{x_{j+1}} \alpha_j dx$$

and

$$\int_{x_j}^{x_{j+1}} \psi_j\phi_j'dx = -\frac{1}{h}\int_{x_j}^{x_{j+1}} (\phi_j+\alpha_j)dx = -\tfrac{1}{2} - \frac{1}{h}\int_{x_j}^{x_{j+1}} \alpha_j dx \ ,$$

and hence

$$\int_{x_j}^{x_{j+1}} \psi_j \psi_j' dx = -\tfrac{1}{2} \quad .$$

Substituting into (3.73) gives the inequality

$$a(k-\gamma) \geq 0 \quad , \tag{3.74}$$

and hence from (3.69) we deduce that $f(\lambda)$ is a monotonic increasing function of λ. ■

Using Lemma 3.3 and equation (3.67) we see that if μ_{max} is the largest eigenvalue of the matrix $(I+E)C$, then the required λ_m satisfies

$$(1-\lambda_m)/(\lambda_m k-\gamma) = \mu_{max} \quad . \tag{3.75}$$

Equation (3.75) leads us to consider the eigensystem of the matrix $(I+E)C$, and the necessary results are given by Lemma 3.4:

Lemma 3.4

For positive integers p,

$$S_p = \left(\frac{4a}{h}\right) \sin^2 p\pi/N+1 \tag{3.76}$$

is a double eigenvalue of $(I+E)C$ corresponding to the two eigenvectors

$$\underline{V}_p = (\sin(2p\pi/(N+1)), \sin(4p\pi/(N+1)), \ldots, \sin(2Np\pi/(N+1))^T \tag{3.77}$$

and $$\underline{W}_p = (\sin^2(p\pi/(N+1)), \sin^2(2p\pi/(N+1)), \ldots, \sin^2(Np\pi/(N+1))^T.$$

If N is even, this accounts for the complete eigensystem of the matrix $(I+E)C$.

If N is odd, the one remaining eigenvalue is $4a/h$ and corresponds to the eigenvector

$$\underline{V}_N = (1, 0, 1, 0, \ldots, 0, 1, 0, 1)^T \quad . \tag{3.78}$$

Proof

The Lemma can be easily established by the use of trigonometric identities to show that for N even,

$$(I+E)C\,\underline{V}_p = S_p\,\underline{V}_p \quad \text{and} \quad (I+E)C\,\underline{W}_p = S_{p-p}\,\underline{W}_p, \quad p = 1,2,\ldots N/2,$$

and for N odd,

$$(I+E)C\,\underline{V}_p = S_{p-p}\,\underline{V}_p \quad \text{and} \quad (I+E)\,C\,\underline{W}_p = S_{p-p}\,\underline{W}_p, \quad p = 1,2,\ldots,\tfrac{1}{2}(N-1),$$

and
$$(I+E)C\,\underline{V}_N = (4a/h)\underline{V}_N. \quad \blacksquare$$

Hence, to summarise, the constant $(1-\Delta_m^2)^{-\frac{1}{2}}$ in the estimate (3.3) may be obtained through the use of (3.26) and (3.75) where γ and k are defined in (3.39) and (3.51) respectively, and μ_{max} is obtained through Lemma 3.4, to give

$$\lambda_m = (1 + \mu_{max}\gamma)/(1 + \mu_{max}k). \tag{3.79}$$

Hence when N is odd, λ_m is given by

$$\lambda_m = \frac{h + 4a\gamma}{h + 4ak}, \tag{3.80}$$

and when N is even, λ_m takes the explicit form

$$\lambda_m = \frac{h + 4a\gamma \sin^2 \frac{1}{2}N\pi/N+1}{h + 4ak \sin^2 \frac{1}{2}N\pi/N+1}. \tag{3.81}$$

Since for N even, $\mu_{max} = (4a/h)\sin^2 \frac{1}{2}N\pi/N+1$, and for N odd, $\mu_{max} = 4a/h$, we may write (3.79) as

$$\lambda_m = (h + 4a\mu\gamma)/(h + 4a\mu k)$$

in all cases, where $0 < \mu \leq 1$. We note from (3.50) and (3.64) that

$$\gamma = p(p-1)h/a,$$

and hence that

$$h + 4a\mu\gamma = h(1 + 4\mu p^2 - 4\mu p)$$

$$= h((1 - 2\mu p)^2 + 4p^2\mu(1-\mu))$$

$$\geq 0.$$

Further, from (3.39) we see that

$$h + 4a\mu k = h - 4\mu\left(a^2\int_{x_j}^{x_{j+1}}\psi_j'\psi_{j+1}'dx + \int_{x_j}^{x_{j+1}}\psi_j\psi_{j+1}dx\right)$$

$$= h + 4\mu\left(a^2\int_{x_j}^{x_{j+1}}\psi_j'^2dx - \int_{x_j}^{x_{j+1}}\psi_j dx + \int_{x_j}^{x_{j+1}}\psi_j^2dx\right)$$

$$= 4\mu\left(a^2\int_{x_j}^{x_{j+1}}\psi_j'^2dx + \int_{x_j}^{x_{j+1}}(\psi_j-\tfrac{1}{2})^2dx + h(1-\mu)\right)$$

$$> 0 \qquad \text{since } \psi_j \text{ is conforming.}$$

Hence, using the inequality (3.74) we have that for λ_m as in (3.79),

$$0 \leq \lambda_m \leq 1.$$

3.6 Test Space Examples

(a) Heinrich et al (1977)

The test functions are of the type described by (3.36) and (3.37), where $\alpha(t)$ is the quadratic form

$$\alpha(t) = -3\sigma t(t-1), \qquad 0 \leq t \leq 1, \tag{3.82}$$

and σ is a constant which determines the degree of upwinding. Heinrich et al make the choice of $\sigma_{AS} = \coth(bh/2a) - (2a/bh)$, so that for the constant coefficient problem (3.4) the Petrov-Galerkin scheme reproduces the Allen & Southwel difference operator (3.30) when a piecewise linear trial space is used.

The constants k and γ in equation (3.79) can be obtained from the definitions (3.39) and (3.51) respectively, and are given by

$$k = a(1 + 3\sigma^2)/h + (9\sigma^2 - 5)h/(30a), \tag{3.83}$$

and

$$\gamma = (h/a)\,((a/h)^2 + a\sigma/h + \sigma^2/4 - 1/4). \tag{3.84}$$

Then in the case where N is odd, μ_{max} takes the simple form $\mu_{max} = 4a/h$, and hence λ_m is given by

$$\lambda_m = \frac{h\,\sigma^2 + 4a\sigma + 4a^2/h}{(6h/5 + 12a^2/h)\sigma^2 + (h/3 + 4a^2/h)} \quad . \tag{3.85}$$

A similar expression may be derived when N is even, using the explicit form in (3.81).

In the limit of very high Péclet number, $a \to 0$, and

$$\lambda_m \to \frac{1 + (\mu_{max}h/4)(\sigma^2 - 1)}{1 + (\mu_{max}h/30)(9\sigma^2 - 5)} \quad . \tag{3.86}$$

When N is odd, (3.86) reduces to the particularly simple form

$$\lambda_m \to 15\sigma^2/(5 + 18\sigma^2) \quad \text{as} \quad a \to 0. \tag{3.87}$$

Choosing the parameter $\sigma = \sigma_{AS}$ as in Heinrich et al (1977) leads to $\sigma \to 1$ as $a \to 0$, and hence the smallest constant in the estimate (3.3) is

$$(1 - \Delta_m^2)^{-\frac{1}{2}} \to \sqrt{23/15} \quad \text{as} \quad a \to 0 \quad . \tag{3.88}$$

However, we may make use of the explicit form for λ_m in equation (3.85) to set $\partial\lambda_m/\partial\sigma = 0$, and hence show that the choice of σ which establishes the smallest possible error constant in (3.3) is

$$\sigma_{opt} = (5h/36a)(1 + 12(a/h)^2)/(1 + 10(a/h)^2), \tag{3.89}$$

and the constant is

$$(1 - \Delta_m^2)_{opt}^{-\frac{1}{2}} \to \sqrt{6/5} \quad \text{as} \quad a \to 0 \quad , \tag{3.90}$$

which is a considerable improvement on (3.88).

Values of the constants $(1 - \Delta_m^2)^{-\frac{1}{2}}$ and $(1 - \Delta_m^2)_{opt}^{-\frac{1}{2}}$ for a complete range of Péclet numbers for the problem (3.4) with $h = 0.1$ are given in Table 3.2. Figure 3.3 compares the parameters σ_{AS} and σ_{opt}.

(b) "Super" Hughes & Brooks

When a piecewise linear trial space is used, the Streamline Upwind Procedure introduced in Hughes & Brooks (1979) may be reformulated as a non-conforming Petrov-Galerkin method: see Hughes & Brooks (1981). The test functions are of the type described by (3.36) and (3.37) with

$$\alpha(t) = \tfrac{1}{2}b\,\sigma_{AS} \quad , \qquad 0 \le t \le 1. \tag{3.91}$$

Such test functions are clearly discontinuous, and hence lie outside the framework of the analysis in Section 3.5. However, we have modified the choice of $\alpha(t)$ in (3.91) slightly to construct conforming test functions in which $\alpha_j(x)$ takes the form in Fig. 3.4.

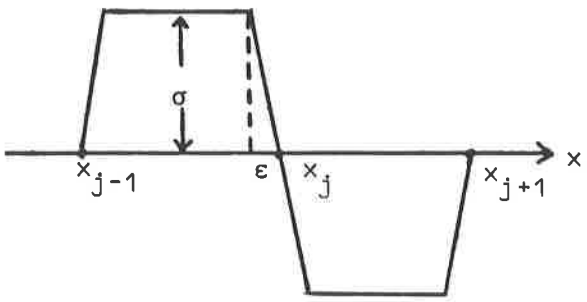


Fig. 3.4

By choosing

$$\bar{\sigma} = \sigma_N \equiv \tfrac{1}{2}\,h\,b\,\tilde{\sigma}_{AS}/(h-\varepsilon), \tag{3.92}$$

the test function is normalised to have the same area as the discontinuous test function using $\alpha(t)$ given by (3.91), and hence reproduces the same difference operator for problem (3.4) when a piecewise linear trial space is used. From the definitions (3.39) and (3.51) we may show that k and γ in equation (3.79) are given by

$$k = k_\varepsilon \equiv a/h + 2a\sigma^2/\varepsilon - h/6 - 4\sigma^2\varepsilon/3a + \sigma^2 h/a$$

and

$$\gamma = \gamma_\varepsilon \equiv a/h + 2\sigma(h-\varepsilon)/h + \sigma^2(h-\varepsilon)^2/ha - h/4a.$$

Hence in the case when N is odd, λ_m is given by

$$\lambda_m = \frac{4(a + \sigma(h-\varepsilon))^2}{(1/3+4\sigma^2)h^2 + 4a^2 + 8a^2\sigma^2 h/\varepsilon - 16\sigma^2\varepsilon h/3} . \qquad (3.93)$$

A similar expression may be derived using equation (3.81) when N is even.

If the form for σ in (3.92) is substituted into (3.93) we may consider λ_m as a function of ε. This relationship is displayed in Figure 3.5 for a range of mesh Péclet numbers. We may thus make use of (3.93) in selecting ε optimally in order to maximise the value of λ_m and hence minimise the smallest error constant in the estimate (3.3). The optimal choice of ε is shown in Table 3.1 for a range of mesh Péclet numbers when the element size $h = 0.1$:

Table 3.1

bh/a	ε_{opt}
2	3.1699×10^{-2}
5	2.4743×10^{-2}
10	1.5294×10^{-2}
20	8.2231×10^{-3}
50	3.4003×10^{-3}
100	1.7166×10^{-3}
500	3.4581×10^{-4}
5000	3.4635×10^{-5}
10000	1.7319×10^{-5}

From Table 3.1 and Figure 3.5 it is clear that for large Péclet numbers, ε_{opt} depends almost linearly on the mesh Péclet number, and that as $\beta = bh/a \to \infty$, $\varepsilon_{opt} \to 0$, and the discontinuous Hughes & Brooks test function is the limit towards which the optimal "Super" Hughes & Brooks test function tends as $\beta \to \infty$.

Furthermore, if the nonconforming Hughes & Brooks test functions are used to evaluate the entries in the matrices A, B and C in (3.14) with the

integrations being carried out element-by-element to avoid the inter-element contributions arising from the discontinuities in the test functions, the error constant thus obtained using equation (3.20) is

$$(1 - \Delta_{HB}^2)^{-\frac{1}{2}} \to 2/\sqrt{3} \qquad \text{as} \quad a \to 0, \tag{3.94}$$

which is the limit of the constant $(1 - \Delta_m^2)^{-\frac{1}{2}}$ for the conforming "Super" Hughes & Brooks test functions with $\varepsilon = \varepsilon_{opt}$.

Values of the constant $(1 - \Delta_m^2)^{-\frac{1}{2}}$ for a complete range of Peclet numbers for problem (3.4) with $h = 0.1$ are given in Table 3.2.

(c) Hemker (1977)

The test functions are of the type described by (3.36) and (3.37), with $\alpha(t)$ given by the exponential form

$$\alpha(t) = (e^{-ht/a} - e^{-h/a} - (1-t)(1 - e^{-h/a}))/(1 - e^{-h/a}). \tag{3.95}$$

From (3.95) we note that

$$\alpha + \partial\alpha/\partial x = -(1-t) + K \tag{3.96}$$

where K is the constant $a/h - e^{-h/a}/(1 - e^{-h/a})$, and hence that $\psi_j - a\alpha_j'$ is a constant K_j, say, on the interval (x_j, x_{j+1}). We may rearrange (3.72) to obtain

$$a(k\gamma) = a^2 \int_{x_j}^{x_{j+1}} \psi_j'^2 dx + \int_{x_j}^{x_{j+1}} \psi_j^2 dx + \int_{x_j}^{x_{j+1}} a\psi_j' dx - \frac{1}{h}\left[\int_{x_j}^{x_{j+1}} (a\psi_j' + \psi_j - a\phi_j')dx\right]^2$$

$$+ \frac{a^2}{h}\left[\int_{x_j}^{x_{j+1}} \phi_j' dx\right]^2 - \frac{2a^2}{h} \int_{x_j}^{x_{j+1}} \psi_j' dx \int_{x_j}^{x_{j+1}} \phi_j' dx - \frac{2a}{h} \int_{x_j}^{x_{j+1}} \psi_j dx \int_{x_j}^{x_{j+1}} \phi_j' dx$$

$$= a^2 \int_{x_j}^{x_{j+1}} \psi_j'^2 dx + \int_{x_j}^{x_{j+1}} \psi_j^2 dx + \int_{x_j}^{x_{j+1}} a\psi_j' dx - \frac{1}{h}\left[\int_{x_j}^{x_{j+1}} (a\psi_j' + \psi_j - a\phi_j')dx\right]^2$$

$$+ a^2 \int_{x_j}^{x_{j+1}} \phi_j'^2 dx - 2a^2 \int_{x_j}^{x_{j+1}} \psi_j' \phi_j' dx - 2a \int_{x_j}^{x_{j+1}} \psi_j \phi_j' dx \quad .$$

Having noted above that

$$\int_{x_j}^{x_{j+1}} a\psi_j' dx = \int_{x_j}^{x_{j+1}} a\phi_j' dx = -a = 2a \int_{x_j}^{x_{j+1}} \psi_j \psi_j' dx ,$$

we may therefore write

$$\begin{aligned} a(k-\gamma) &= \int_{x_j}^{x_{j+1}} (a\psi_j' + \psi_j - a\phi_j')^2 dx - \frac{1}{h}\left[\int_{x_j}^{x_{j+1}} (a\psi_j' + \psi_j - a\phi_j') dx\right]^2 \\ &= \int_{x_j}^{x_{j+1}} K_j^2 \, dx - \frac{1}{h}\left[\int_{x_j}^{x_{j+1}} K_j dx\right]^2 \\ &= 0 \end{aligned}$$

since K_j is constant on the interval (x_j, x_{j+1}).

Hence we see that $\gamma = k$, and so from (3.79) that $\lambda_m = 1$. The smallest error constant in the estimate (3.3) is thus unity for such an exponential test space, and the optimal trial space-test space pairing has been achieved for a piecewise linear trial space. The disadvantage of using a Petrov-Galerkin method employing this test space is that it involves the evaluation of inner products containing very steep exponential terms as in (3.95), which may be difficult to achieve accurately unless high order or special quadrature rules are us

(d) Galerkin

Finally we will use the framework of our analysis to show how poorly the Galerkin method may perform on problems of the type (3.4). With $\alpha(t) \equiv 0$, the constants k and γ in equation (3.79) are given by

$$k = h((a/h)^2 - 1/6a) \tag{3.97}$$

and

$$\gamma = h(a/h^2 - 1/4a), \tag{3.98}$$

and hence λ_m is

$$\lambda_m = (1 + \mu_{max} h((a/h)^2 - \tfrac{1}{4}))/(1 + \mu_{max}(a^2/h - h/6)). \tag{3.99}$$

When N is odd, (3.99) reduces to the particularly simple form

$$\lambda_m = 12/(12 + (h/a)^2), \qquad (3.100)$$

and hence

$$(1 - \Delta_m^2)^{-\frac{1}{2}} = \sqrt{h^2 + 12a^2} \,/\, 2\sqrt{3a} \;\to\; h/2\sqrt{3}\; a \quad \text{as} \quad a \to 0. \qquad (3.101)$$

From (3.3) and (3.101) we see that whilst the Galerkin approximation remains of optimal order, the smallest constant in the estimate (3.3) becomes unbounded as the mesh Péclet number increases.

Table 3.2

bh/a	$(1 - \Delta_m^2)^{-\frac{1}{2}}$			
	HHMZ	HHMZ*	SHB	GALERKIN
2	1.0060	1.0060	1.0178	1.1547
5	1.0468	1.0428	1.0597	1.7559
50	1.2022	1.0945	1.1406	14.468
500	1.2344	1.0954	1.1532	144.34
10^5	1.2383	1.0954	1.1546	28867.

HHMZ = Heinrich, Huyakorn, Mitchell & Zienkiewicz, with $\sigma = \sigma_{AS}$.

HHMZ* = Heinrich, Huyakorn, Mitchell & Zienkiewicz, with $\sigma = \sigma_{opt}$.

SHB = "Super" Hughes & Brooks, with $\sigma = \sigma_N$ and $\varepsilon = \varepsilon_{opt}$.

Figure 3.6(a) shows the different test functions for a mesh Péclet number bh/a = 5. Figure 3.6(b) is similar, with bh/a = 50.

THE PARAMETER σ IN THE HEINRICH, HUYAKORN, MITCHELL & ZIENKIEWICZ UPWIND SCHEME

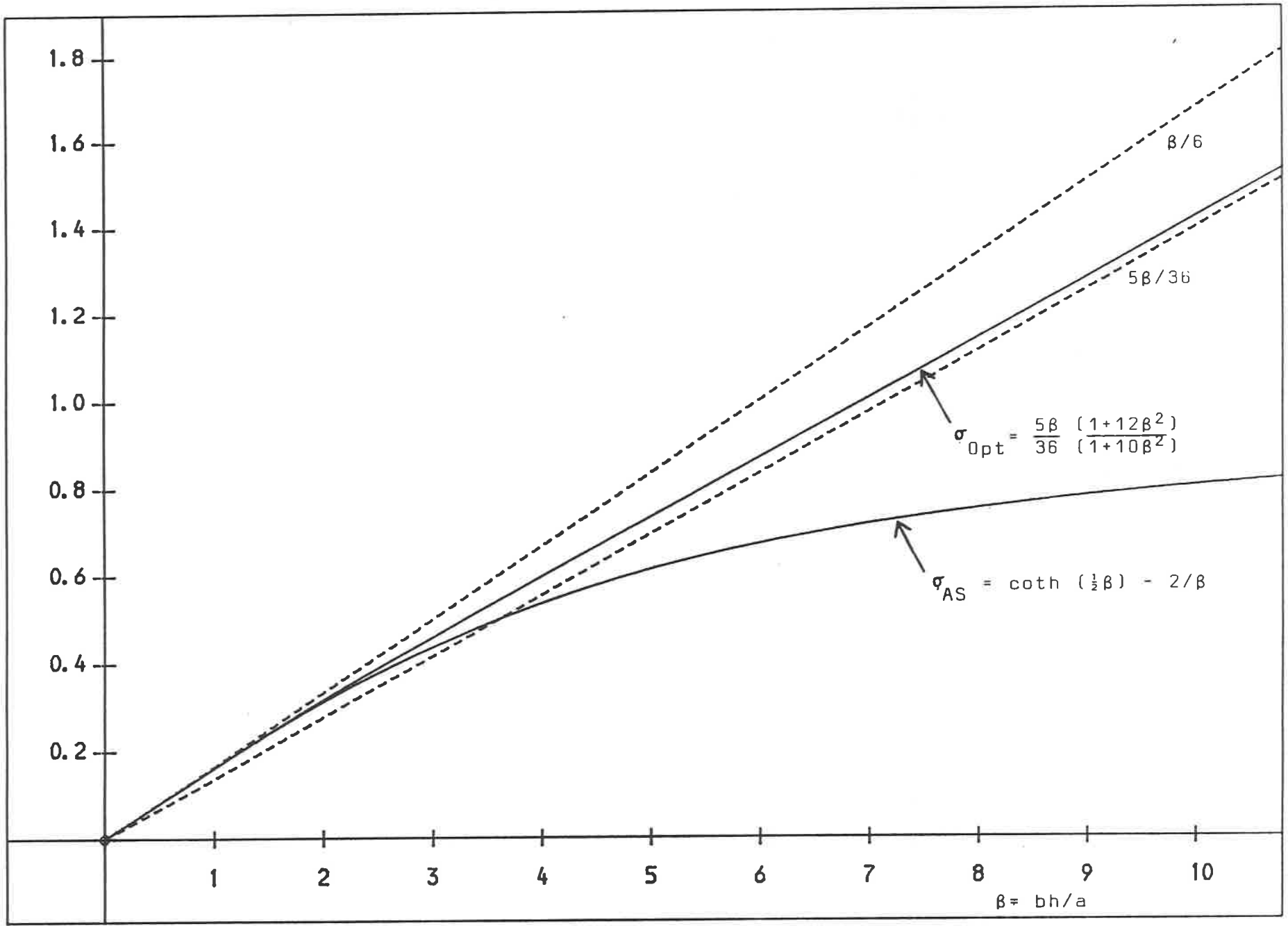


Figure 3.3

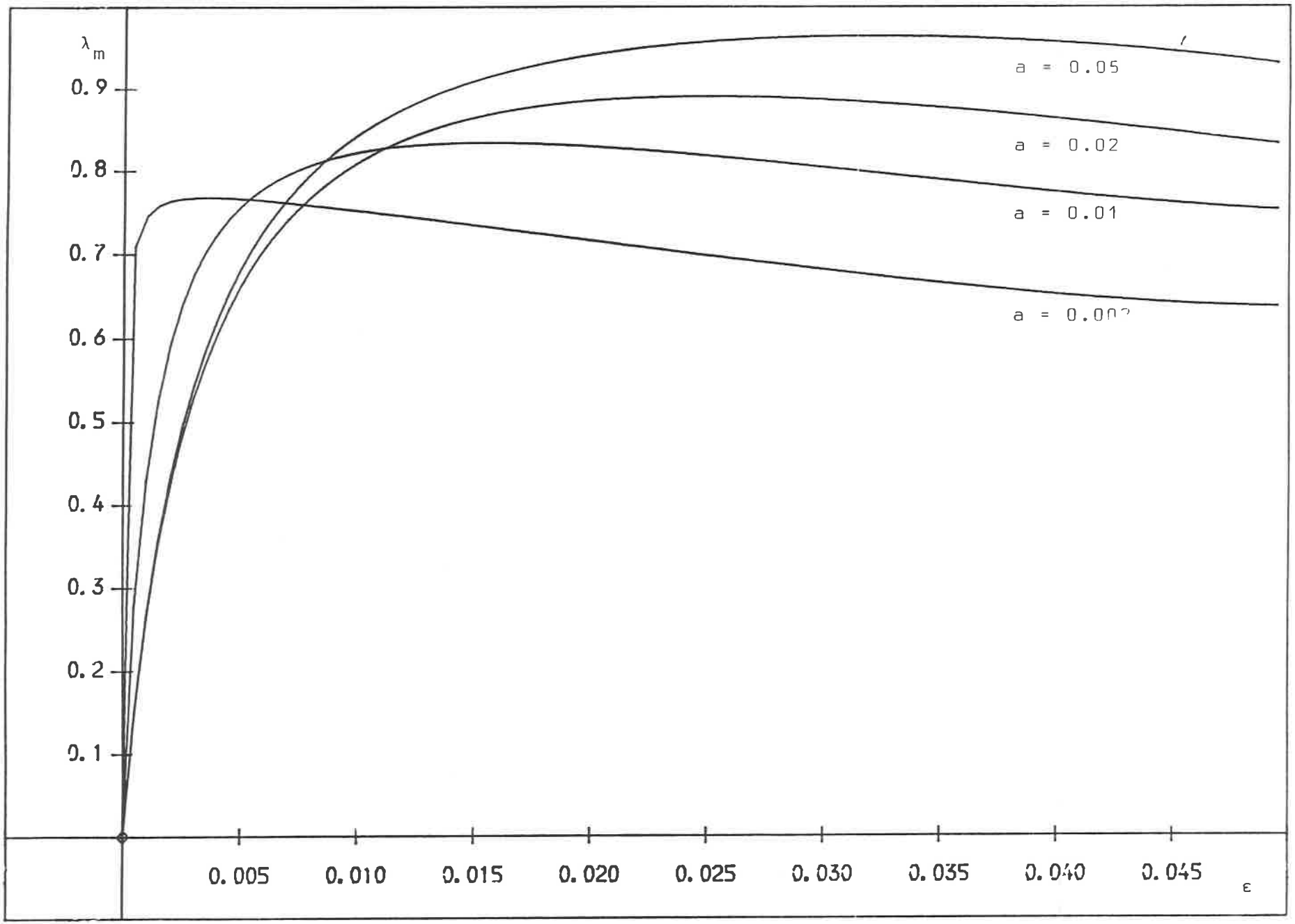

VARIATION OF λ_m WITH ε IN THE "SUPER" HUGHES & BROOKS SCHEME
b = 1, h = 0.1
λ_m
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
a = 0.05
a = 0.02
a = 0.01
a = 0.00?
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
0.045
ε

Figure 3.5

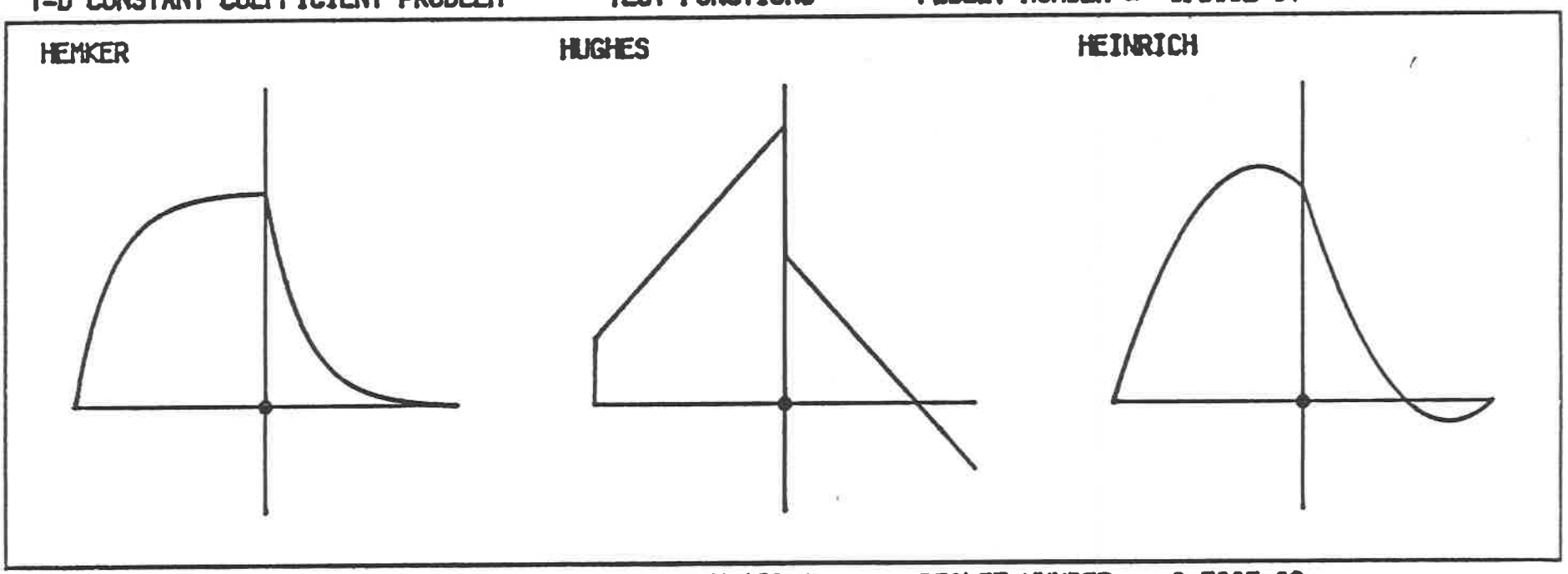


Figure 3.6a

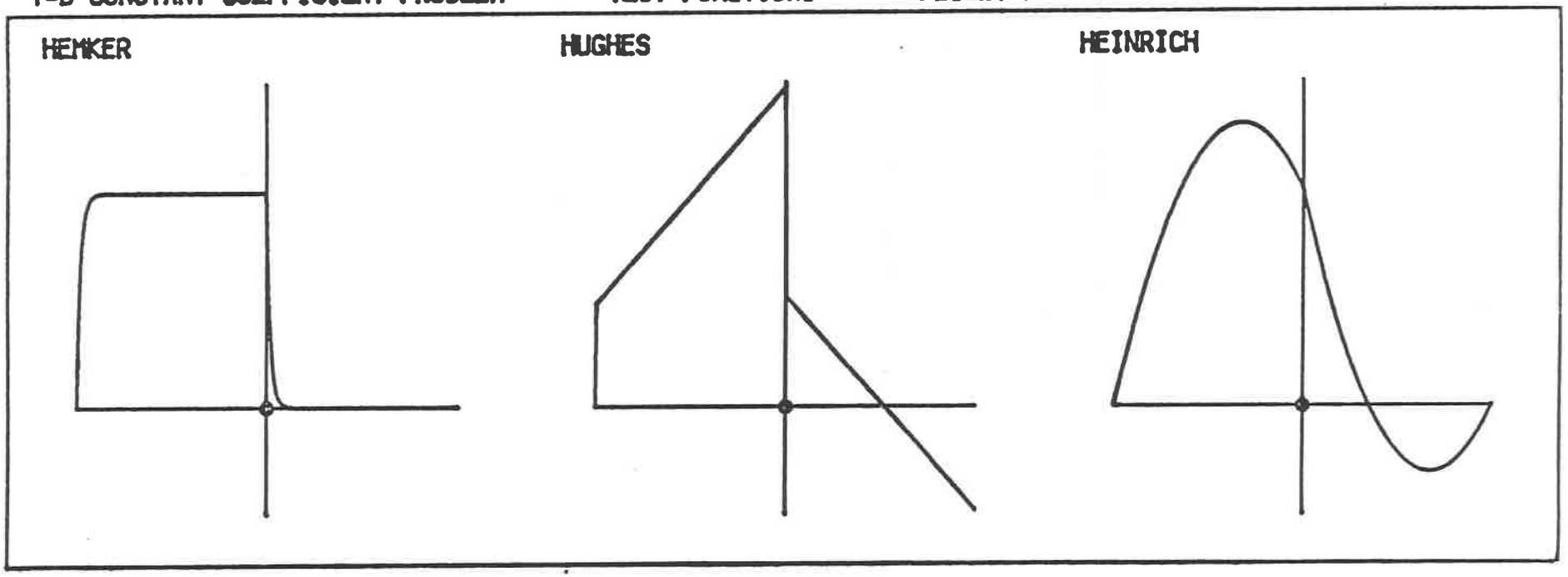


Figure 3.6b

4. NUMERICAL EXPERIMENTS IN TWO DIMENSIONS

In the first part of this section generalisations of two of the 'upwind' Petrov-Galerkin methods in Section 3.6 are used to solve a two-dimensional test problem.

(a) Heinrich et al (1977)

If a piecewise linear trial space is used for the solution of the one dimensional problem (3.4), the test space employed by Heinrich et al has a set of basis functions given by (3.36), (3.37) and (3.80). The degree of upwinding is determined by the parameter σ : if $\sigma = 0$, the method reduces to the Galerkin approximation; if $\sigma \geq 1 - 2/\beta$, where $\beta = bh/a$ is the mesh Péclet number, the method produces a set of difference equations for problem (3.4) with a discrete maximum principle and hence a solution which is not oscillatory; the choice $\sigma = \sigma_{AS}$ reproduces the Allen & Southwell difference operator, (see Christie et al, 1976).

When the problem (1.1) with the operator L taking the form given in (2.5) is considered in two dimensions, the trial space is generalised to become the space of piecewise bilinear functions on a regular mesh. The trial functions are then tensor products of the piecewise linear trial functions used for the one dimensional problem (3.4); similarly the test functions are tensor products of those used in one dimension.

Consider the test function associated with node i over the shaded element in Figure 4.1(a)

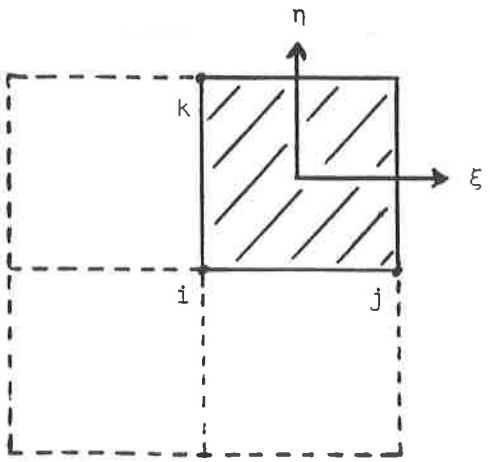


Figure 4.1(a)

Using local (ξ,η) co-ordinates with origin at the centre of this element and normalised between -1 and +1, the bilinear trial function ϕ_i is given by

$$\phi_i = (1-\xi)/2 \cdot (1-\eta)/2 \ .$$

Over this region the corresponding test function ψ_i is given by

$$\psi_i = \psi_i(\xi) \cdot \psi_i(\eta) \tag{4.1}$$

where

$$\psi_i(\xi) = \phi_i(\xi) + \sigma_{ij}\alpha(\xi) \tag{4.2}$$

and

$$\psi_i(\eta) = \phi_i(\eta) + \sigma_{ik}\alpha(\eta). \tag{4.3}$$

The quadratic perturbation α is given by

$$\alpha(s) = -3(1-s)(1+s)/4 \ . \tag{4.4}$$

The coefficient σ_{ij} along the edge ij is given by

$$\sigma_{ij} = \coth(\beta_{ij}/2) - 2/\beta_{ij} \ , \tag{4.5}$$

with a similar expression for σ_{ik}. Here β_{ij} is the mesh Péclet number calculated using local nodal velocity values along the edge ij. That is, $\beta_{ij} = b_{ij}h/a$ where b_{ij} is the average velocity in the ξ direction calculated from $b_{ij} = (\underline{b}_i+\underline{b}_j)\cdot\underline{e}_{ij}/2$: $\underline{e}_{ij}$ is a unit vector in the direction from node i to node j, and $\underline{b}_i$ and $\underline{b}_j$ are values of the velocity field at nodes i and j respectively. β_{ik}, which is required for the coefficient σ_{ik} along the edge ik, is calculated similarly.

(b) Streamline Upwind. (Hughes & Brooks, 1979)

With the introduction of an artificial diffusion tensor $\underline{\underline{K}}$, the weak formulation of problem (1.1) with L as in (2.5) becomes that of finding $U \in S_E^h(\Omega)$ such that

$$\langle \underline{\nabla} U, (a + \underline{\underline{K}})\underline{\nabla}\phi\rangle + \langle \underline{b}\cdot\underline{\nabla}U, \phi\rangle = \langle f,\phi\rangle \ \forall\phi \in S_0^h(\Omega). \tag{4.6}$$

The tensor $\underline{\underline{K}}$ acts in the direction of flow, and hence takes the form

$$\underline{\underline{K}} = \tilde{K} \begin{pmatrix} b_1^2 & b_1 b_2 \\ b_1 b_2 & b_2^2 \end{pmatrix} \tag{4.7}$$

where $\underline{b} = (b_1, b_2)^T$.

This is the formulation introduced in Hughes & Brooks (1979). There, and in Brooks (1981), the parameter $\tilde{K}$ is based on the choice of parameter σ_{AS} in (3.89). In the case of a general bilinear element with a (ξ,η) local co-ordinate system having its origin at the centre of the element, and normalised between -1 and +1 (see Fig. 4.1(b)), the parameter $\tilde{K}$ is given by

$$\tilde{K} = \tfrac{1}{2}(\tilde{\xi} b_\xi h_\xi + \tilde{\eta} b_\eta h_\eta), \tag{4.8}$$

where h_ξ and h_η are element size parameters, and b_ξ and b_η are the velocity field components, in the ξ and η directions respectively.

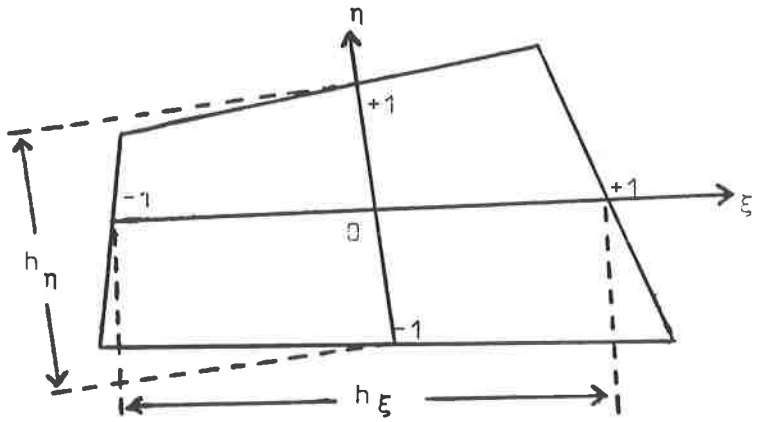


Fig. 4.1(b)

The parameters $\tilde{\xi}$ and $\tilde{\eta}$ are defined by

$$\tilde{\xi} = \coth(\beta_\xi/2) - 2/\beta_\xi$$

and

$$\tilde{\eta} = \coth(\beta_\eta/2) - 2/\beta_\eta \quad , \tag{4.9}$$

where

$$\beta_\xi = b_\xi h_\xi / a \qquad \text{and} \qquad \beta_\eta = b_\eta h_\eta / a. \tag{4.10}$$

The components b_1 and b_2 of the convective velocity field are treated throughout as continuous functions, and since 2x2 Gaussian integration will be used in the evaluation of the inner products in (4.6), their values will be required at the Gauss points.

To reduce the cost of calculation, in (4.9) one might make use of the limits

$$\coth(\alpha) - 1/\alpha \to \alpha/3 \qquad \text{as} \qquad \alpha \to 0$$

and

$$\coth(\alpha) - 1/\alpha \to 1 \qquad \text{as} \qquad \alpha \to +\infty$$

matched at $\alpha = 3$ to give the approximation

$$\coth(\alpha) - 1/\alpha \simeq \begin{cases} \alpha/3 , & |\alpha| \le 3 \\ \text{sign } \alpha , & |\alpha| > 3 \end{cases} , \tag{4.11}$$

though this was not used in the computations described below.

Owing to the form (4.7) of the tensor $\underline{\underline{K}}$, we may write

$$\langle \underline{\nabla} U, (a+\underline{\underline{K}})\ \underline{\nabla}\phi \rangle = \langle a\underline{\nabla}U, \underline{\nabla}\phi\rangle + \langle (\tilde{K}/b^2)\underline{b}\cdot\underline{\nabla}U,\ \underline{b}\cdot\underline{\nabla}\phi\rangle \quad , \tag{4.12}$$

and hence (4.6) becomes

$$\langle \underline{\nabla} U,\ a\underline{\nabla}\phi\rangle + \langle \underline{b}\cdot\underline{\nabla}U,\ \phi + (\tilde{K}/b^2)\underline{b}\cdot\underline{\nabla}\phi\rangle = \langle f, \phi\rangle \quad \forall \phi \in S_0^h(\Omega). \tag{4.13}$$

If a is constant and the trial space is piecewise bilinear, then $\underline{\nabla}\cdot(a\underline{\nabla}U) = a\nabla^2 U = 0$ on each element. Provided therefore that the term $\langle a\nabla^2 U,\ \underline{b}\cdot\underline{\nabla}\phi\rangle$ is evaluated in this way, the left hand side of (4.13) may be written as

$$\langle a\nabla^2 U - \underline{b}\cdot\underline{\nabla}U, \psi\rangle_e$$

where

$$\psi = \phi + (\tilde{K}/b^2)\underline{b}\cdot\underline{\nabla}\phi \quad , \tag{4.14}$$

and $\langle \cdot,\cdot\rangle_e$ denotes that the inner product is calculated by element-by-element integration. Hence if on the right hand side of (4.13) f is integrated

against ψ rather than ϕ, the streamline upwind method may be regarded as a consistent Petrov-Galerkin method using the test functions given in (4.14), (see Hughes & Brooks, 1981). This approach was used in the numerical calculations carried out using this method. An analysis of the streamline upwind method appears in Johnson & Nävert (1981).

4.1 Test Problem 1. (Hutton, 1981)

The problem is illustrated in Figure 4.2. Calculations were carried out using a regular square grid of 10 x 20 elements using a piecewise bilinear trial space. Diffusivity was varied from 0.1 to 10^{-6} whilst the velocity field and mesh size remained unchanged in order to increase the mesh Péclet number. The velocity field $\underline{b} = (b_1, b_2)^T$ is incompressible and is given by

$$\begin{aligned} b_1 &= 2y(1 - x^2) \\ b_2 &= -2x(1 - y^2) \quad . \end{aligned} \tag{4.15}$$

The boundary conditions are given by

$$u = 1 + \tanh(10(2x+1)) \quad \text{on} \quad y = 0, \; -1 \le x \le 0$$

$$u = 0 \text{ on } \begin{cases} x = -1, & 0 < y < 1 \\ y = 1, & -1 \le x \le 1 \\ x = 1, & 0 \le y < 1 \end{cases} \tag{4.16}$$

and $\partial u/\partial n = 0$ on $y = 0, \quad 0 < x < 1$.

Note that on the outflow boundary $y = 0, \; 0 < x < 1$ $\underline{n} \cdot \underline{b} > 0$. The boundary data is therefore such as to guarantee existence and uniqueness of a solution to the approximate problem.

Some typical streamlines are shown in Figure 4.2.

The results for this problem are shown in the form of outlet profiles, that is for $y = 0$ and $0 < x < 1$, in Figure 4.3 for the Streamline Upwind method and in Figure 4.4 for the Heinrich et al method. For the case of pure

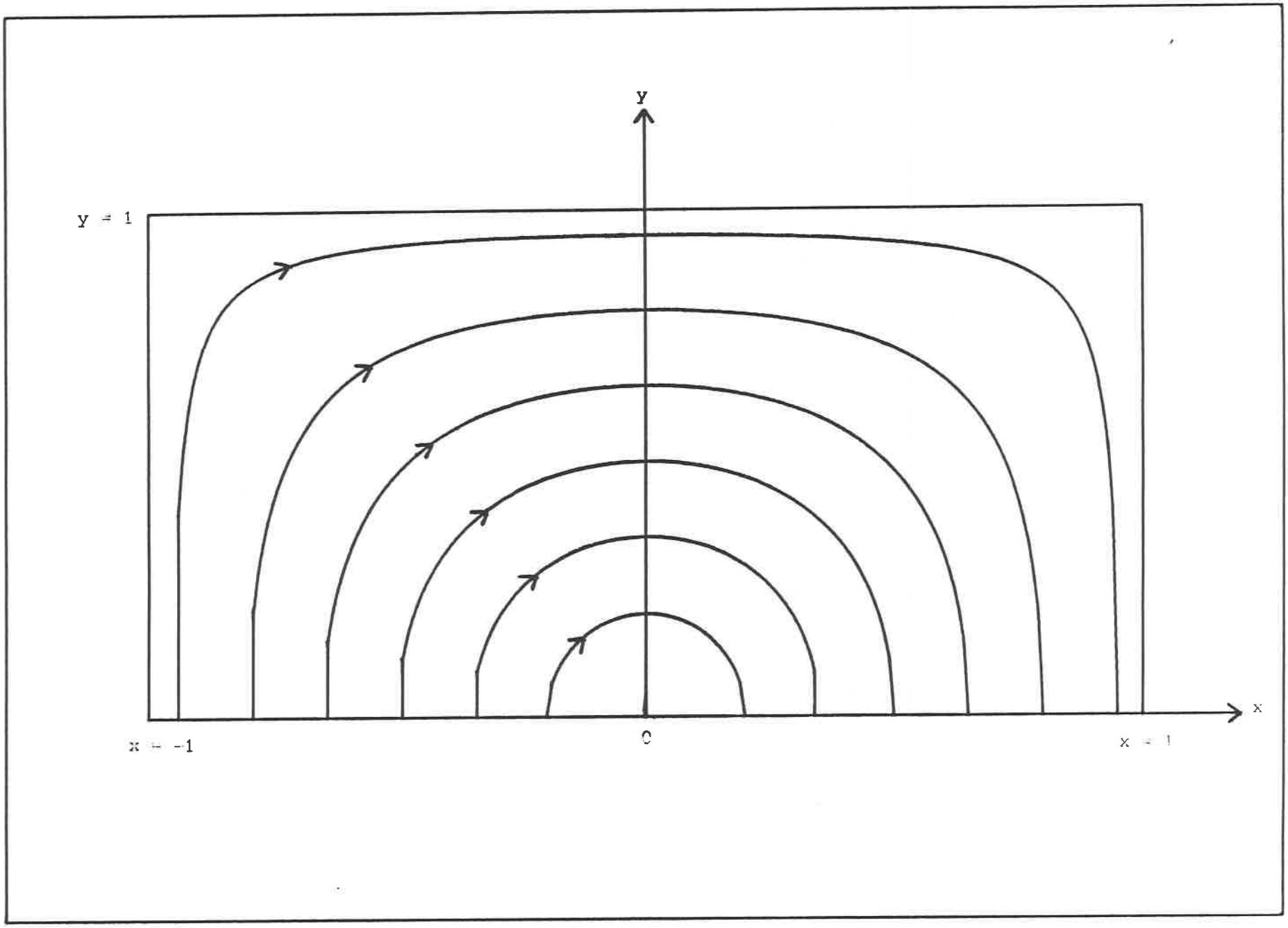


Figure 4.2

convection, nodal values of the solutions at the outlet are given in Table 4.1.

x	Exact	HB	Error	% Error	HHMZ	Error	% Error
0.0	2.00000	2.00000	-	-	2.00000	-	-
0.1	2.00000	1.97929	-0.02071	1.0	1.99248	-0.00752	0.4
0.2	1.99999	2.02345	0.02346	1.2	2.02335	0.02336	1.2
0.3	1.99933	2.09622	0.09689	4.8	2.09511	0.09578	4.8
0.4	1.96403	1.77387	-0.19016	9.7	1.75906	-0.20497	10.4
0.5	1.00000	0.96107	-0.03893	3.9	0.95045	-0.04955	5.0
0.6	0.03597	0.19840	0.16243	-	0.21810	0.18213	-
0.7	0.00067	-0.06399	-0.06466	-	-0.04869	-0.04936	-
0.8	0.00001	-0.00752	-0.00753	-	-0.02230	-0.02231	-
0.9	0.00000	0.00558	0.00058	-	0.00734	0.00734	-
1.0	0.00000	0.00000	-	-	0.00000	-	-

Table 4.1

The results are shown in the column headed HB for the Hughes & Brooks Streamline Upwind Scheme, and in the column headed HHMZ for the upwind method of Heinrich et al. Neither scheme satisfies a discrete maximum principle. In both cases there is an overshoot of 4.8% at $x = 0.3$, whilst the maximum nodal error in both cases occurs at $x = 0.4$.

It should be pointed out that as the mesh is refined, although such overshoots are gradually eroded, the general trend of overshoot before the steep gradient persists. For example, the overshoot using the Streamline Upwind method at the point where the approximate solution takes on its highest value is shown in Table 4.2 for varying mesh sizes.

h	x	HB - Exact soln.
0.1	0.3	9.689×10^{-2}
0.05	0.35	4.580×10^{-2}
0.033	0.366	1.794×10^{-2}

Table 4.2

No such overshoot remains by the time h is reduced to 0.02.

Some results for this problem using high order finite difference formulae are presented in Thompson & Wilkes, 1982.

4.2 Mixed Method based on $B_2(\cdot,\cdot)$

We now consider a method, for problem (1.1) where the operator L is of the form (2.5), based on the formulation proposed by Barrett & Morton (1982). The aim is to produce a best fit $U^* \in S_E^h(\Omega)$ to the solution in the norm based on the operator T_2 in (2.10). For one dimensional problems direct Petrov-Galerkin methods with this aim are constructed in Barrett & Morton (1980) and analysed in Barrett & Morton (1981) and in Barrett (1980). For the two dimensional problem we define the bilinear form

$$B_2(w_1,w_2) = \langle \rho a \underline{\nabla} w_1, \underline{\nabla} w_2 \rangle + \langle w_1, a^{-1}(\rho \underline{b}\cdot\underline{b} + \underline{\nabla}\cdot(\rho a \underline{b}) w_2 \rangle \quad ,$$

$$\forall\, w_1,w_2 \in H^1(\Omega)$$

where $\rho(\underline{x})$ is a positive weight function which may be chosen. The bilinear form on $H^1_{E_0}(\Omega)$ in (2.14) is then

$$B_2(w_1,w_2) = \langle a^{\frac{1}{2}}\underline{\nabla} w_1 - \underline{b} a^{-\frac{1}{2}} w_1, \rho(a^{\frac{1}{2}}\underline{\nabla} w_2 - \underline{b} a^{-\frac{1}{2}} w_2)\rangle + \int_{\partial\Omega_2} \rho \underline{b}\cdot\underline{n} w_1 w_2 \, ds \qquad (4.17)$$

$$\forall\, w_1,w_2 \in H^1_{E_0}(\Omega).$$

Then $B_2(w_1,w_2)$ is $H^1_{E_0}(\Omega)$ - elliptic provided that

$$\rho a > 0 \quad \text{and} \quad \rho b^2 + \underline{\nabla}\cdot(\rho a \underline{b}) \geq 0, \qquad (4.18)$$

conditions which can simply be satisfied by choosing

$$\rho > 0 \quad \text{and} \quad \underline{b}\cdot\underline{\nabla}(\rho a) \geq 0.$$

Noting that for $w_1,w_2 \in H^1(\Omega)$,

$$-\langle \underline{\nabla}\cdot(a\underline{\nabla} w_1 - \underline{b} w_1),\, \rho a w_2 \rangle = \langle a\underline{\nabla} w_1 - \underline{b} w_1,\, \underline{\nabla}(\rho a w_2)\rangle$$

$$- \int_{\partial\Omega} (a\underline{\nabla} w_1 - \underline{b} w_1) \rho a w_2 \cdot \underline{n} \, ds$$

$$= \langle a^{\frac{1}{2}}\underline{\nabla}w_1 - \underline{b}a^{-\frac{1}{2}}w_1,\ \rho a(a^{\frac{1}{2}}\underline{\nabla}w_2 - \underline{b}a^{-\frac{1}{2}}w_2)\rangle$$

$$+ \langle a\underline{\nabla}w_1 - \underline{b}w_1,\ (\rho\underline{b} + \underline{\nabla}(\rho a))w_2\rangle - \int_{\partial\Omega} (a\underline{\nabla}w_1 - \underline{b}w_1)\rho a w_2 \cdot \underline{n}\ ds, \qquad (4.19)$$

we see that when a is constant the weak formulation (1.3) may be considered as finding $u \in H_E^1(\Omega)$ such that

$$B_2(u,v) = \langle f,\rho v\rangle + \langle \underline{b}u - a\underline{\nabla}u, a^{-1}(\rho\underline{b} + \underline{\nabla}(\rho a))v\rangle \qquad (4.20)$$

$$\forall\ v \in H_{E_0}^1(\Omega)\ .$$

If $\underline{V}$ is an approximation to $\underline{v} = \underline{b}u - a\underline{\nabla}u$, we may solve for $U \in S_E^h(\Omega)$ such that

$$B_2(U,\phi) = \langle f,\rho\phi\rangle + \langle \underline{V}, a^{-1}(\rho\underline{b} + \underline{\nabla}(\rho a))\phi\rangle \qquad \forall\ \phi \in S_0^h(\Omega). \qquad (4.21)$$

The problem now is to obtain an equation for $\underline{V}$ which can be used alternately with (4.21) to give a convergent iteration. A preliminary technique, which works well for the constant coefficient problem considered by Raithby (1976) but not for the Test Problem 1 above, was put forward by Barrett & Morton (1982). This was based on solving

$$\langle \underline{b} \times \underline{V},\ \chi_e\rangle = \langle -a\underline{b} \times \underline{\nabla}U,\ \chi_e\rangle$$

for each element, where χ_e is the characteristic function having the value unity in the interior of element e and zero elsewhere, together with an explicit difference scheme to solve

$$\underline{\nabla}\cdot\underline{v} = 0$$

in the limit of pure convection.

The dependence of the approximation U in (4.21) on the accuracy of the approximation $\underline{V}$ is through the following estimate : suppose that $U^* \in S_E^h(\Omega)$ is the best fit in the norm $B_2(\cdot,\cdot)$ to the solution u to problem (1.3) so that

$$B_2(U^*,\phi) = B_2(u,\phi) \qquad \forall\ \phi \in S_0^h(\Omega). \qquad (4.22)$$

Writing $\underline{\alpha} = a^{-1}[\rho\underline{b} + \underline{\nabla}(\rho a)]$, subtracting (4.20) from (4.21) gives

$$B_2(U - u,\phi) = \langle \underline{\alpha}\cdot(\underline{V} - \underline{v}),\phi\rangle \qquad \forall\ \phi \in S_0^h(\Omega),$$

and hence using (4.22) we obtain

$$B_2(U - U^*,\phi) = \langle \underline{\alpha}\cdot(\underline{V} - \underline{v}),\phi\rangle \qquad \forall\ \phi \in S_0^h(\Omega). \tag{4.23}$$

Since $U - U^* \in S_0^h(\Omega)$ we have

$$\| U - U^* \|_2^2 = \langle \underline{\alpha}\cdot(\underline{V} - \underline{v}),\ U - U^*\rangle$$

$$\leq \| \hat{\underline{\alpha}}\cdot(\underline{V} - \underline{v}) \|_{L_2} \| |\underline{\alpha}|\ (U - U^*) \|_{L_2},$$

where $\hat{\underline{\alpha}} = \underline{\alpha} / |\underline{\alpha}|$. Then since

$$\langle U - U^*,\ a^{-1}(\rho \underline{b}\ \underline{b} + \underline{\nabla}\cdot(\rho a \underline{b}))(U - U^*)\rangle = \| U - U^* \|_2^2 - \langle \rho a \underline{\nabla}(U - U^*), \underline{\nabla}(U - U^*)\rangle,$$

we have that

$$\| U - U^* \|_2 \leq \gamma \| \hat{\underline{\alpha}}\cdot(\underline{V} - \underline{v}) \|_{L_2}, \tag{4.24}$$

where γ is a constant such that

$$|\rho \underline{b} + \underline{\nabla}(\rho a)| \leq \gamma a^{\frac{1}{2}}(\rho \underline{b}\ \underline{b} + \underline{\nabla}\cdot(\rho a \underline{b}))^{\frac{1}{2}} \quad \text{uniformly.} \tag{4.25}$$

Note in particular that if a is constant, $\underline{\nabla}\cdot\underline{b} = 0$ and the weight function $\rho(\underline{x}) \equiv 1$, then $\gamma = a^{-\frac{1}{2}}$. (Note that there is a factor $a^{-\frac{1}{2}}$ in the weighted L_2 part of the norm $\| \cdot \|_2$.)

From (4.24) it is clear that only the component of $\underline{v}$ in the direction of $\underline{\alpha}$ is important, and consequently we now construct an equation for $s = a\underline{\alpha}\cdot\underline{v}$ which may be used in an iterative scheme with (4.21).

We suppose that a is constant and let $\hat{\underline{s}}$ denote the unit vector parallel to the convective velocity field $\underline{b}$, and $\hat{\underline{n}}$ a unit vector perpendicular to it. Then if $\underline{b}$ has components b_1 and b_2 in the x and y co-ordinate directions respectively, we may write

$$\hat{\underline{s}} = (1/b)(b_1,b_2) \quad \text{and} \quad \hat{\underline{n}} = (1/b)(-b_2,b_1), \tag{4.26}$$

where $b^2 = b_1^2 + b_2^2$. If we take $\rho(\underline{x}) \equiv 1$, then $\underline{\alpha} = a^{-1}\underline{b}$, and

$$\underline{v} = (s/b)\,\hat{\underline{s}} - a\ \partial u/\partial n\ \hat{\underline{n}}\ , \tag{4.27}$$

where $\partial u/\partial n$ denotes the derivative in the direction of $\hat{\underline{n}}$.

Taking $\underline{i}$ and $\underline{j}$ to be unit vectors in the x and y directions respectively, (4.27) may be rewritten as

$$\begin{aligned}\underline{v} = {} & (sb_1/b^2 + (b_2 a/b^2)(-b_2\ \partial u/\partial x + b_1\ \partial u/\partial y))\ \underline{i} \\ & + (sb_2/b^2 - (b_1 a/b^2)(-b_2\ \partial u/\partial x + b_1\ \partial u/\partial y))\ \underline{j}\ .\end{aligned} \tag{4.28}$$

Using the divergence operator on (4.28), and the fact that $\underline{\nabla}\cdot\underline{b} = 0$ enables us to write

$$\underline{\nabla}\cdot\underline{v} = (\underline{b}\cdot\underline{\nabla})(s/b^2) - a\underline{\nabla}\cdot(\partial u/\partial n\ \hat{\underline{n}})\ . \tag{4.29}$$

Since $Lu = 0$, with L as in (2.5), can be written as $\underline{\nabla}\cdot\underline{v} = 0$ we have

$$(\underline{b}\cdot\underline{\nabla})(s/b^2) = a\underline{\nabla}\cdot(\partial u/\partial n\ \hat{\underline{n}}). \tag{4.30}$$

We use a finite difference approximation to equation (4.30) in an iterative scheme with (4.21), namely

$$(\underline{b}\cdot\underline{\nabla})_h(S/b^2) = a\underline{\nabla}_h\cdot((\hat{\underline{n}}\cdot\underline{\nabla}_h U)\hat{\underline{n}}), \tag{4.31}$$

where the values of U on the right hand side are obtained from the finite element solution to (4.21) at the previous iteration step.

The operator $\underline{b}\cdot\underline{\nabla}$ is discretised to give $(\underline{b}\cdot\underline{\nabla})_h$ using a directionally upwinded finite difference operator. We shall use finite difference stencil notation in which the stencil represents a discrete operator on the nodal value at its centre, for example,

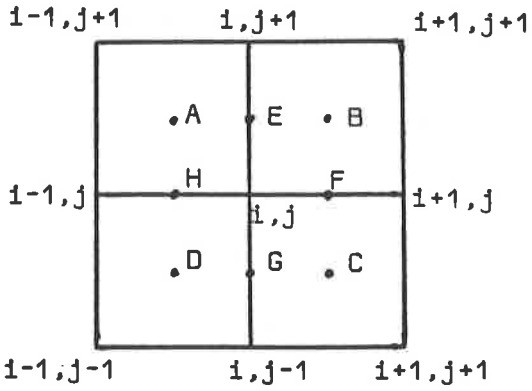
$$\begin{vmatrix} 0 & 0 & 0 \\ \gamma_2 & \gamma_1 & 0 \\ \gamma_3 & 0 & 0 \end{vmatrix} \qquad \theta_{ij} = \gamma_1\theta_{ij} + \gamma_2\theta_{i-1,j} + \gamma_3\theta_{i-1,j-1}.$$

On a regular square mesh of size h with, for example, $b_1 \geq 0$, $b_2 \geq 0$ and $b_1 \geq b_2$, the stencil representing $(\underline{b}\cdot\underline{\nabla})_h$ is

$$(1/h)\begin{vmatrix} 0 & 0 & 0 \\ b_2-b_1 & b_1 & 0 \\ -b_2 & 0 & 0 \end{vmatrix} .$$

Generalisation to other velocity directions is straightforward.

To interpret the right hand side of (4.31) we make use of the bilinear nature of the approximation U. Consider the four elements surrounding node i,j :



The points A, B, C and D are the centres of each of the elements, and since U is bilinear we may write

$$U(A) = \tfrac{1}{4}(U_{i-1,j+1} + U_{i,j+1} + U_{i-1,j} + U_{i,j}),$$

with similar expressions for U(B), U(C) and U(D). Then at the mid-points of the interior sides of the elements, i.e. E,F, G and H, we may compute $\underline{\nabla}U$ using

$$\partial U/\partial x\ (E) = (U(B) - U(A))/h$$

and

$$\partial U/\partial y\ (E) = (U_{i,j+1} - U_{i,j})/h \quad ,$$

with similar expressions at the points F, G and H. To approximate the right hand side of (4.31) we then use

$$(a/h)\Big[\{((b_2^2/b^2)\ \partial U/\partial x - (b_1b_2/b^2)\ \partial U/\partial y)\ (F)$$
$$-((b_2^2/b^2)\ \partial U/\partial x - (b_1b_2/b^2)\ \partial U/\partial y)\ (H)$$
$$+((b_1^2/b^2)\ \partial U/\partial y - (b_1b_2/b^2)\ \partial U/\partial x)\ (E)$$
$$-((b_1^2/b^2)\ \partial U/\partial y - (b_1b_2/b^2)\ \partial U/\partial x)\ (G)\ \Big] \ .$$

Values of S/b^2 are set on the boundary $\partial\Omega_1$ using a finite difference approximation to

$$S/b^2 = U - (a\underline{b}\cdot\underline{\nabla}U)/b^2 \ .$$

From (4.23) it is clear that if our approximation to S is chosen from the same trial space as U, then it should attempt to be a least squares best fit to S. Consequently we next interpolate the nodal values of S produced by the difference scheme for equation (4.31) to obtain S_I and then project onto the trial space in a least squares sense to obtain S_{II}. With the trial space piecewise bilinear spanned by basis functions $\{\phi_j,\ j = 1,\ldots,N\}$, suppose we interpolate S_I in the form

$$S_I = \sum_{i=1}^{N} S_{I_i}\ \psi_i \ ,$$

where $\{\psi_j,\ j = 1,\ldots,N\}$ forms a basis for a space of piecewise biquadratic functions over the domain. Then the projection S_{II} onto the trial space is obtained by solving

$$\sum_{j=1}^{N} \langle (S_{II_j}\ \phi_j - S_{I_j}\ \psi_j),\ \phi_i \rangle = 0 \ \forall\ \phi_i \in S_0^h(\Omega). \qquad (4.32)$$

In particular if the domain is partitioned using a regular square mesh, equations (4.32) correspond to inverting the difference stencil operator

$$(1/36)\begin{vmatrix} 1 & 4 & 1 \\ 4 & 16 & 4 \\ 1 & 4 & 1 \end{vmatrix} S_{II_j} = (1/144)\begin{vmatrix} 1 & 10 & 1 \\ 10 & 100 & 10 \\ 1 & 10 & 1 \end{vmatrix} S_{I_j} \qquad (4.33)$$

for all j such that $\phi_j \in S_0^h(\Omega)$.

S_{II} is then put into the right hand side of equation (4.21), and we solve for $U \in S_E^h(\Omega)$ such that

$$B_2(U,\phi) = \langle f,\phi\rangle + \langle a^{-1}S_{II},\phi\rangle \qquad \forall\ \phi \in S_0^h(\Omega) \tag{4.34}$$

as the next step in the iteration. Since the bilinear form $B_2(\cdot,\cdot)$ is symmetric, equation (4.34) leads to the inversion of a symmetric matrix.

Results for the Mixed Method applied to Test Problem 1 are shown in Figure 4.5 in the form of outlet profiles at $y = 0$, $0 < x < 1$.

As the mesh Péclet number decreases it is found that convergence in the iteration scheme between equations (4.21) and (4.31) may not be reached. At low mesh Péclet numbers the oscillatory nature of the best fit U^* in the norm based on $B_2(\cdot,\cdot)$ is less severe, and consequently the need for a recovery procedure is diminished. In these circumstances we consider making use of the directionally upwinded difference stencil for the operator $\underline{b}\cdot\underline{\nabla}$ in a difference scheme applied directly to approximate the equation $Lu = 0$ with L as in (2.5). We will consider the case where at node i,j we have $b_1 \geq 0$, $b_2 \geq 0$, and $b_1 \geq b_2$, though generalisation to other velocity directions is straightforward. Combining the upwind stencil

$$(1/h)\begin{vmatrix} 0 & 0 & 0 \\ b_2-b_1 & b_1 & 0 \\ -b_2 & 0 & 0 \end{vmatrix}$$

with the central difference operator for $\underline{b}\cdot\underline{\nabla}$ we obtain

$$(\alpha/h)\begin{vmatrix} 0 & 0 & 0 \\ b_2-b_1 & b_1 & 0 \\ -b_2 & 0 & 0 \end{vmatrix} + ((1-\alpha)/2h)\begin{vmatrix} 0 & b_2 & 0 \\ -b_1 & 0 & b_1 \\ 0 & -b_2 & 0 \end{vmatrix} \tag{4.35}$$

as an approximation to the operator $\underline{b}\cdot\underline{\nabla}$, where $\alpha \in [0,1]$. Combining (4.35) with the discrete Laplacian operator

$$(a/h^2)\begin{vmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{vmatrix}$$

produces the following stencil to represent the operator L_h:

$$a/2h^2\begin{vmatrix} 0 & (1-\alpha)\beta_y-2 & 0 \\ 2\alpha\beta_y-(1+\alpha)\beta_x-2 & 8+2\alpha\beta_x & (1-\alpha)\beta_x-2 \\ -2\alpha\beta_y & -(1-\alpha)\beta_y-2 & 0 \end{vmatrix} \qquad (4.36)$$

where $\beta_x = b_1h/a$ and $\beta_y = b_2h/a$.

Thus L_h has a discrete maximum principle if we choose

$$\alpha \geq 1 - 2/\beta_x \quad \text{and} \quad \alpha \geq 1 - 2/\beta_y \ . \qquad (4.37)$$

Further, the choice

$$\alpha = \coth(\tfrac{1}{2}(\beta_x + \beta_y)) - 2/(\beta_x + \beta_y) \ , \qquad (4.38)$$

whilst giving a maximum principle, produces a scheme which is exponentially fitted when the flow is along the co-ordinate directions.

A discrete maximum principle is used by Kellogg (1980) to establish a uniform error bound of the type

$$|u(x_j) - U_j| \leq kh,$$

where k is independent of h and of the Péclet number b/a, for a two-dimensional generalisation of the exponentially-fitted difference scheme due to Allen & Southwell (1955). Such a scheme, however, suffers from excessive "crosswind diffusion" (see Griffiths & Mitchell (1979)). An advantage of the difference scheme represented by (4.36) is that the "crosswind diffusion", representing artificial diffusivity, is zero not only when the flow is along a co-ordinate direction, but also in the case where the flow is at 45^o to the mesh.

Results using the difference scheme (4.36) for Test Problem 1 compare favourably with those produced by the high-order difference schemes presented in Thompson & Wilkes (1982).

4.3 Test Problem 2

The problem has the same flow field as Test Problem 1 but the boundary condition along $x = 1,\ 0 \leq y < 1$ becomes $u = 100$, and along the inflow boundary $y = 0,\ -1 \leq x \leq 0$, as well as along the other tangential boundaries, we have $u = 0$. This thus represents a cool fluid being convected tangentially past a hot plate. Again the calculations were carried out using a regular square grid of 10×20 elements using a piecewise bilinear trial space and the mesh Péclet number varied by varying the diffusivity parameter only.

Figures 4.6 - 4.9 show cross-sections of the solutions between $x = 0$ and $x = 1$ at various values of y for a range of Péclet numbers produced by the Streamline Upwind scheme and by the upwind scheme of Heinrich et al. Figures 4.10 and 4.11 show the same cross-sections produced by the Mixed Method described in Section 4.2.

In the case where the diffusivity coefficient, a, tends to zero, the oscillations produced by the Mixed Method are compared in Table 4.3 with those in the weighted least squares best fit to the asymptotic approximation to the solution (described in Section 4.4.1) in the norm derived from (4.17). In Section 4.4.2 a recovery procedure, as used in one dimension by Barrett (1980), is described for recovering information from the oscillatory approximation resulting from the Mixed Method.

	y = 0.0		y = 0.5	
x	Mixed Method	Best Fit	Mixed Method	Best Fit
1.0	100.000	100.000	100.000	100.000
0.9	-27.422	-28.269	-27.312	-28.145
0.8	8.302	8.098	7.944	7.930
0.7	-2.556	-2.415	-2.278	-2.246
0.6	0.805	0.628	0.643	0.554
0.5	-0.261	-0.568	-0.178	-0.322

TABLE 4.3

4.4 Recovery of Boundary Layer Information

First of all an asymptotic approximation to the solution is presented for the boundary layer region near $x = 1$; in particular the half-width of the boundary layer is calculated, where the half-width $\delta(y)$ is defined by

$$u(1 - \delta(y)) = \tfrac{1}{2} u(1) \quad .$$

The recovery procedure is then employed, and a comparison between the half-width thus predicted and that of the asymptotic approximation is made.

4.4.1 Asymptotic Approximation

The equation $Lu = 0$ with L as in (2.5) and the convective velocity field as given by (3.15) becomes

$$-a\partial^2u/\partial x^2 - a\partial^2u/\partial y^2 + 2y(1-x^2)\partial u/\partial x - 2x(1-y^2)\partial u/\partial y = 0 \; . \tag{4.39}$$

Since the boundary layer is at $x = 1$, setting $x = 1 - k\xi$ gives

$$-(a/k^2)\partial^2u/\partial\xi^2 - a\partial^2u/\partial y^2 - 2y\xi(2-k\xi)\partial u/\partial\xi - 2(1-k\xi)(1-y^2)\partial u/\partial y = 0 \; . \tag{4.40}$$

Following, for example, Lamb (1932), Wilson (1959), Milne-Thomson (1960), we make the choice of $k = a^{\frac{1}{2}}$, producing a stretching of the x axis which causes the boundary layer to lie in the region $0 \le \xi \le 1$. Away from the region where y is close to 1, we may neglect the terms of order k and k^2 in equation (4.40) to give

$$\partial^2u/\partial\xi^2 + 4y\xi\partial u/\partial\xi + 2(1-y^2)\partial u/\partial y = 0. \tag{4.41}$$

Using the fact that $\underline{\nabla}\cdot\underline{b} = 0$ we may rewrite (4.41) as

$$\partial^2u/\partial\xi^2 + \partial/\partial\xi(4y\xi u) + \partial/\partial y(2(1-y^2)u) = 0. \tag{4.42}$$

Integrating (4.42) with respect to ξ from 0 to $h(y)$ gives

$$[\partial u/\partial\xi]_0^h + [4y\xi u]_0^h = -2(\partial/\partial y)\int_0^h (1-y^2)u\,d\xi. \tag{4.43}$$

Treating (4.42) as a free boundary problem, we define $h(y)$ by $u(\xi, y) = 0$ and $\partial u/\partial\xi(\xi,y) = 0$ when $\xi = h(y)$, and obtain

$$(\partial/\partial y)\left((1-y^2)\int_0^h u\,d\xi\right) = \tfrac{1}{2}[\partial u/\partial\xi]_{\xi=0} \quad . \tag{4.44}$$

As in the references above we choose the functional form for u proposed by Von Kármán:

$$u(\xi,y) = 100\,(1 - \sin(\pi\xi/2h)), \tag{4.45}$$

which satisfies the boundary conditions

$$u(0,y) = 100,\ u(h,y) = 0 \quad \text{and} \quad \partial u/\partial\xi(h,y) = 0.$$

This then allows $h(y)$ to be solved for by substituting (4.45) into (4.44). Since

$$\int_0^h u\,d\xi = \left[100(\xi + (2h/\pi)\cos(\pi\xi/2h))\right]_0^h$$

$$= 100\,(\pi-2)h/\pi\ ,$$

and

$$[\partial u/\partial\xi]_{\xi=0} = -100\pi/2h,$$

we obtain

$$(\partial/\partial y)\,((1-y^2)h) = -\pi^2/4h(\pi-2). \tag{4.46}$$

Multiplying (4.46) by $2h(y)(1-y^2)$ we obtain

$$(\partial/\partial y)\,((1-y^2)^2h^2) = -\tfrac{1}{2}\pi^2(1-y^2)/(\pi-2)\ ,$$

and hence

$$h^2(y) = d^2/(1-y^2)^2 - \tfrac{1}{2}\pi^2(y-y^3/3)/(\pi-2)(1-y^2)^2\ , \tag{4.47}$$

where $h = d$ at $y = 0$.

In order to relate the half-width of the boundary layer, $\delta(y)$ to the quantity $h(y)$, consider the form of $u(\xi,y)$ in (4.45). If $\xi_{\frac{1}{2}} = a^{-\frac{1}{2}}\delta$, then

$$u(\xi_{\frac{1}{2}},y) = 100\,(1 - \sin\,(\pi\xi_{\frac{1}{2}}/2h(y))) = 50,$$

and so solving for $\xi_{\frac{1}{2}}$ gives

$$\xi_{\frac{1}{2}} = \tfrac{1}{3}h(y),$$

and hence

$$\delta(y) = \frac{1}{3}\,a^{\frac{1}{2}}h(y). \tag{4.48}$$

4.4.2 Recovery Procedure

For each value of y, recovery is carried out in one dimension by assuming that the solution in the boundary layer region can be represented by the form

$$u_R(x) = Ae^{B(x-1)} + C, \tag{4.49}$$

where A,B and C are functions of y, and are determined by solving the system

$$B_2(U,\phi_j) = B_2(u_R,\phi_j), \qquad j = J-1,\ J-2 \tag{4.50}$$

and

$$u(x_J) = u_R(x_J) \quad .$$

Here, node J is on the boundary $x = 1$ at, say, $y = Y$, nodes J-1 and J-2 are adjacent nodes at the same y value, and x_j denotes the x-position of node j.

Along the line $y = Y$, U is taken as the expansion

$$U = \sum_j U_j\phi_j\ (x,Y),$$

where U_j is the nodal solution parameter at node j produced by the Mixed Method. In (4.50) $B_2(w_1,w_2)$ is the one-dimensional analogue of (4.17) in the case of $\underline{\nabla}\cdot\underline{b} = 0$ and a is constant, namely

$$B_2(w_1,w_2) = \int_0^1 aw_1'w_2'dx + \int_0^1 a^{-1}b^2w_1w_2dx, \tag{4.51}$$

where ' denotes differentiation with respect to x.

Equations (4.50) lead to the system

$$\int_0^1 a\,U'\phi'_{J-i}\,dx + \int_0^1 a^{-1}b^2U\phi_{J-i}\,dx = R_i, \quad i = 1,2$$

and

$$\int_0^1 a\,ABe^{B(x-1)}\phi'_{J-i}\,dx + \int_0^1 a^{-1}b^2(Ae^{B(x-1)}+C)\phi_{J-i}dx = R_i, \quad i = 1,2 \tag{4.52}$$

and

$$C = u_R(x_J) - A. \tag{4.53}$$

Rearrangement of (4.52) gives

$$A\left(\int_0^1 a\, Be^{B(x-1)}\phi'_{J-i}\, dx + \int_0^1 a^{-1}b^2 e^{B(x-1)}\phi_{J-i}\, dx - D_i\right) = R_i - D_i u_R(x_J), \quad i = 1,2 \tag{4.54}$$

where

$$D_i = \int_0^1 a^{-1} b\, \phi_{J-i}\, dx \quad , \qquad i = 1,2$$

and hence using (4.54) for $i = 1$ and $i = 2$ we may solve for B. A and C may then be obtained from (4.54) and (4.53) respectively, and the recovered half-width δ_R calculated using

$$Ae^{-B\delta_R} + C = 50.$$

In Table 4.4 positions of the recovered half-width δ_R are shown for various diffusivities and compared with the half-width, δ, obtained by using equations (4.47) and (4.48). For equation (4.47), d is calculated from the recovered value of δ_R at $y = 0$, that is $d = 3a^{-\frac{1}{2}}\delta_R$, so that the asymptotic and the recovered boundary layer half-widths are matched at $y = 0$.

Figure 4.12 compares graphically the positions of δ and δ_R. The success of such a recovery procedure is crucial in assessing the value of the Mixed Method approach.

Diffusivity	1.0×10^{-2}			5.0×10^{-3}			1.0×10^{-3}		
y	δ	δ_R	e	δ	δ_R	e	δ	δ_R	e
0.0	.0569	.0569	-	.0408	.0408	-	.0227	.0227	-
0.1	.0531	.0529	0.4	.0381	.0383	0.5	.0218	.0219	0.5
0.2	.0499	.0494	1.0	.0359	.0362	0.8	.0214	.0212	0.9
0.3	.0472	.0462	2.1	.0341	.0343	0.6	.0213	.0205	3.8
0.4	.0449	.0433	3.6	.0326	.0325	0.3	.0217	.0199	8.3
0.5	.0430	.0405	5.8	.0316	.0308	2.5	.0229	.0193	15.7

TABLE 4.4

e is the % error $\left(\frac{\delta_R - \delta}{\delta}\right) \times 100.$

ACKNOWLEDGEMENT

I would like to express my gratitude to Professor K.W. Morton for his invaluable guidance and supervision during the course of this work.

REFERENCES

[1] ALLEN, D.N. de G., & SOUTHWELL, R.V., 1955. Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. Quart. J. Mech. Appl. Math. 8, pp. 129-145.

[2] AUBIN, J.P., 1972. Approximation of Elliptic Boundary Value Problems. John Wiley & Sons, New York.

[3] AXELSSON, O., 1981. Stability and Error Estimates of Galerkin finite element approximations for convection-diffusion equations. I.M.A. J. Num. Anal. 1, No. 3, July 1981, pp. 329-346.

[4] BABUSKA, I. & AZIZ, A.K., 1972. Survey lectures on the mathematical foundation of the finite element method. The Mathematical Foundations of the Finite Element Method with applications to Partial Differential Equation (Ed. A.K. Aziz) Academic Press, New York, pp. 3-363.

[5] BARRETT, J.W., 1980. Optimal Petrov-Galerkin methods. Ph.D. Thesis, University of Reading.

[6] BARRETT, J.W. & MORTON, K.W., 1980. Optimal finite element solutions to diffusion-convection problems in one dimension. Int. J. Num. Meth. Eng., 15, pp. 1457-1474.

[7] BARRETT, J.W. & MORTON, K.W., 1981. Optimal Petrov-Galerkin methods through approximate symmetrization. I.M.A. J. Num. Anal., 1, pp. 439-468.

[8] BARRETT, J.W. & MORTON, K.W., 1982. Optimal finite element approximation for diffusion-convection problems. The Mathematics of Finite Elements and Applications. Proc. MAFELAP 1981. (Ed. J.R. Whiteman), Academic Press, London, pp. 403-411.

[9] BARRETT, K.E., 1977. Finite element analysis for flow between rotating discs using exponentially weighted basis functions. Int. J. Num. Meth. Eng., 11, pp. 1809-1817.

[10] BROOKS, A.N., 1981. A Petrov-Galerkin finite element formulation for convectio dominated flows. Ph.D. Thesis, California Institute of Technology, Pasadena, California.

[11] CHRISTIE, I., GRIFFITHS, D.F., MITCHELL, A.R. & ZIENKIEWICZ, O.C., 1976. Finite element methods for second order differential equations with significant first derivatives. Int. J. Num. Meth. Eng., 10, pp. 1389-1396.

[12] CIARLET, P.G., 1978. The Finite Element Method for Elliptic Problems. North Holland Publ. Comp., Amsterdam.

[13] DIXON, L.C.W., HARRISON, D. & MORGAN, J.V., 1979. On singular cases arising from Galerkin's method. The Mathematics of Finite Elements and Applications III. (Ed. J.R. Whiteman), Academic Press, pp. 217-225.

[14] GRIFFITHS, D.F., & MITCHELL, A.R., 1979. On generating upwind finite element methods. Finite Element Methods for Convection Dominated Flows. AMD- Vol. 34, A.S.M.E. (Ed. T.J.R. Hughes), New York, pp. 91-104.

[15] HEINRICH, J.C., HUYAKORN, P.S., MITCHELL, A.R. & ZIENKIEWICZ, O.C., 1977. An upwind finite element scheme for two-dimensional convective transport equations. Int. J. Num. Meth. Eng., 11, pp. 131-143.

[16] HEINRICH, J.C. & ZIENKIEWICZ, O.C., 1979. The finite element method and 'upwinding' techniques in the numerical solution of convection dominated flow problems. Finite Element Methods for Convection Dominated Flows. AMD - Vol 34, A.S.M.E. (Ed. T.J.R. Hughes), New York, pp. 105-136.

[17] HEMKER, P.W., 1977. A numerical study of stiff two-point boundary problems. Thesis, Math. Cent. Amsterdam.

[18] HUGHES, T.J.R., 1978. A simple scheme for developing 'upwind' finite elements. Int. J. Num. Meth. Eng., 12, pp. 1359-1365.

[19] HUGHES, T.J.R., & BROOKS, A.N., 1979. A multi-dimensional upwind scheme with no crosswind diffusion. Finite Element Methods for Convection Dominated Flows. AMD - Vol. 34, A.S.M.E. (Ed. T.J.R. Hughes), New York, pp. 19-35.

[20] HUGHES, T.J.R., & BROOKS, A.N., 1981. A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions : application to the streamline-upwind procedure. Finite Elements in Fluids, Vol. 4 (Ed. R.H. Gallagher), J. Wiley & Sons, New York.

[21] HUTTON, A.G., 1981. The numerical representation of convection. IAHR Working Group Meeting, May 1981.

[22] IL'IN, A.M., 1969. Differencing scheme for a differential equation with a small parameter affecting the highest derivative. Math. Notes Acad. Sci. USSR 6, pp. 596-602.

[23] JOHNSON, C., & NAVERT, U., 1981. An Analysis of some finite element methods for advection-diffusion problems. Conf. on Analytical and Numerical Approaches to Asymptotic Problems in Analysis. (Eds. O. Axelsson, L.S. Frank and A. van der Sluis), North Holland.

[24] KELLOGG, R.B., 1980. Analysis of a difference approximation for a singular perturbation problem in two dimensions. Boundary and Interior Layers - Computational and Asymptotic Methods (Ed. J.J.H. Miller), Boole Press, Dublin, pp. 113-117.

[25] KELLOGG, R.B. & TSAN, A., 1978. Analysis of some difference approximations for a singular perturbation problem without turning points. Math. Comp., 32, pp. 1025-1039.

[26] LAMB, H., 1932. Hydrodynamics. Sixth edition. Cambridge University Press.

[27] MILNE-THOMSON, L.M., 1960. Theoretical Hydrodynamics. Fourth edition. Macmillan & Co. Ltd., London.

[28] MORTON, K.W., 1981. Finite element methods for non-self-adjoint problems. University of Reading, Num. Anal. Report 3/81.

[29] RAITHBY, G.D., 1976. Skew upstream differencing schemes for problems involving fluid flow. Comp. Meth. Appl. Mech. & Eng., 9, pp. 153-164.

[30] STRANG, G., & FIX, G.J., 1973. An Analysis of the Finite Element Method. Prentice-Hall, New York.

[31] THOMPSON, C.P., & WILKES, N.S., 1982. Experiments with Higher-Order Finite Difference Formulae. AERE - R10493, U.K.A.E.A., Harwell.

[32] WAKIL, M.M. EL, 1962. Nuclear Power Engineering. McGraw-Hill, New York.

[33] WILKINSON, J.H., 1965. The Algebraic Eigenvalue Problem. O.U.P.

[34] WILSON, D.H., 1959. Hydrodynamics. Edward Arnold (Publ.) Ltd., London.

[35] ZIENKIEWICZ, O.C., 1977. The Finite Element Method. Third edition. McGraw-Hill, London.

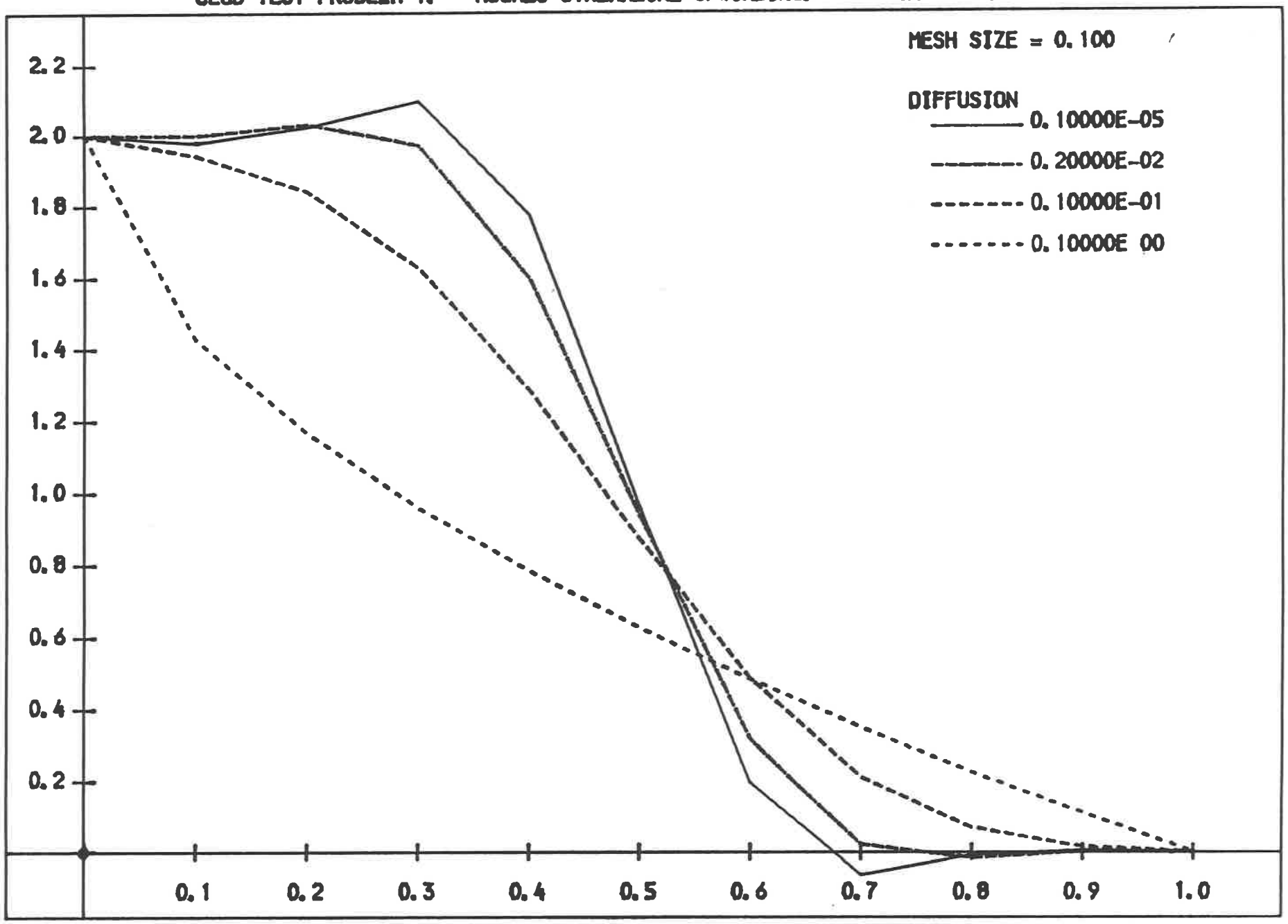

CEGB TEST PROBLEM 1. HUGHES STREAMLINE UPWINDING. OUTLET PROFILE
MESH SIZE = 0.100
DIFFUSION
0.10000E-05
0.20000E-02
0.10000E-01
0.10000E 00
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 4.3

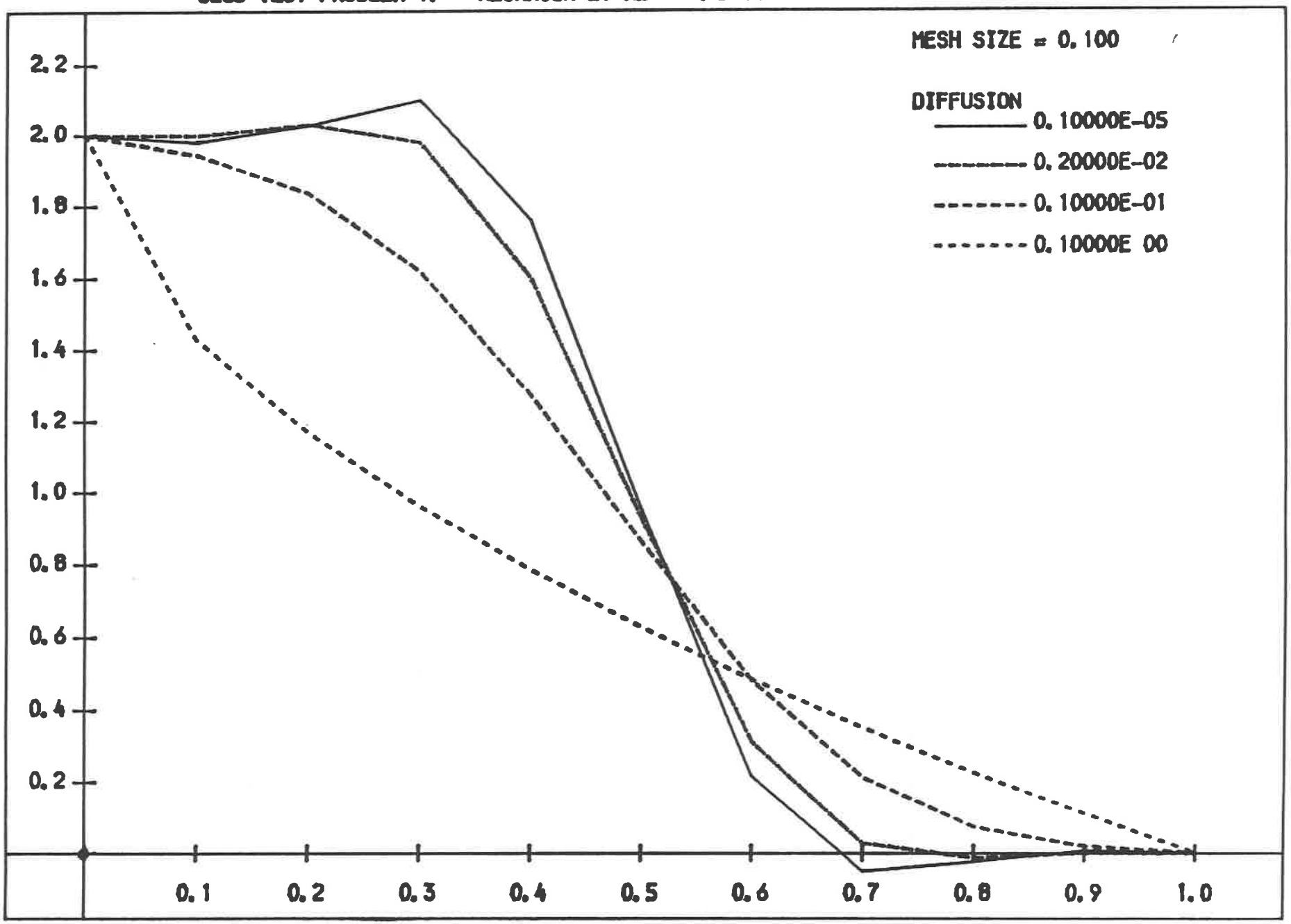

CEGB TEST PROBLEM 1. HEINRICH ET AL UPWIND METHOD. OUTLET PROFILE
MESH SIZE = 0.100
DIFFUSION
0.10000E-05
0.20000E-02
0.10000E-01
0.10000E 00
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 4.4

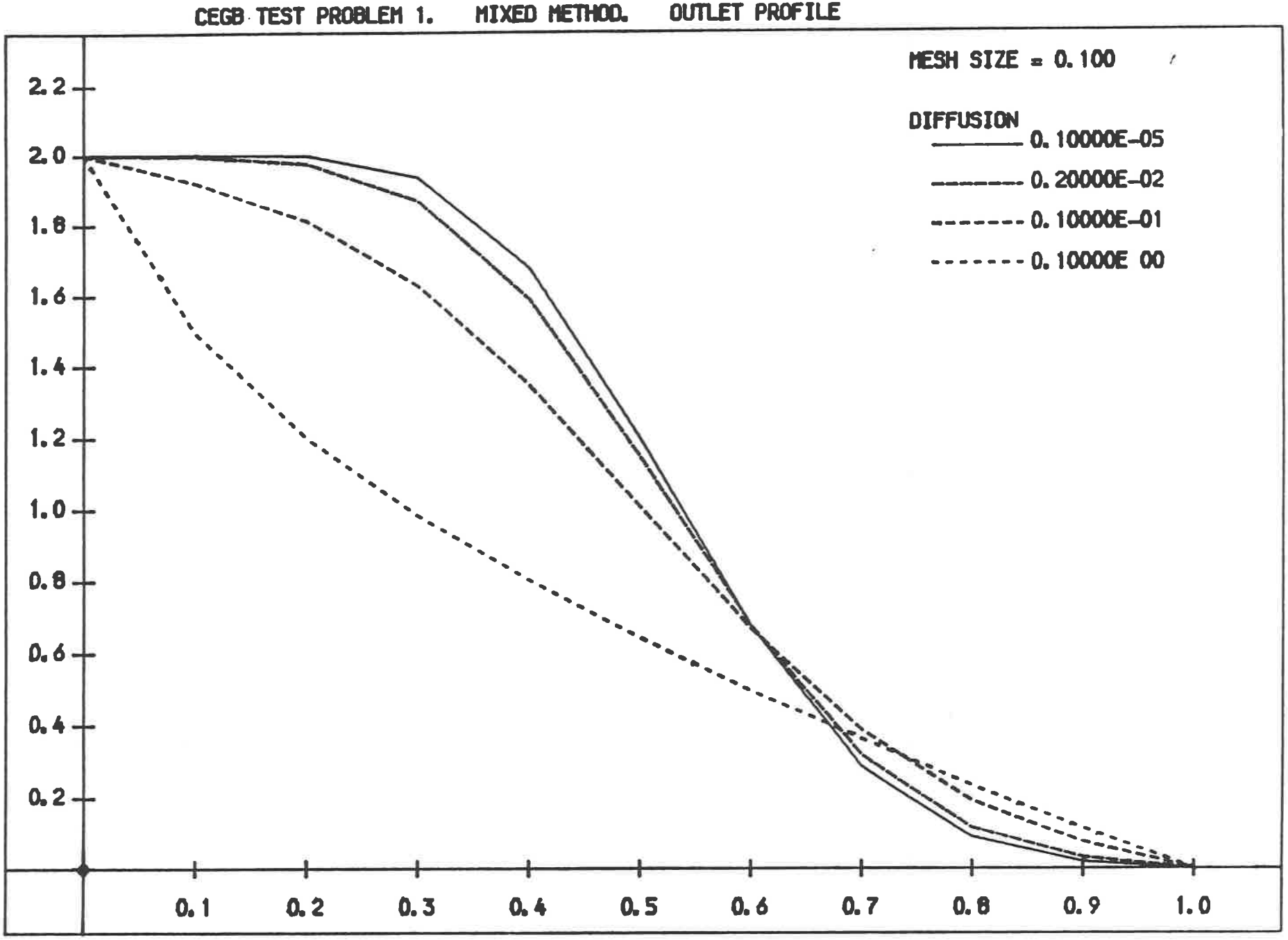

CEGB TEST PROBLEM 1. MIXED METHOD. OUTLET PROFILE
MESH SIZE = 0.100
DIFFUSION
0.10000E-05
0.20000E-02
0.10000E-01
0.10000E 00
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 4.5

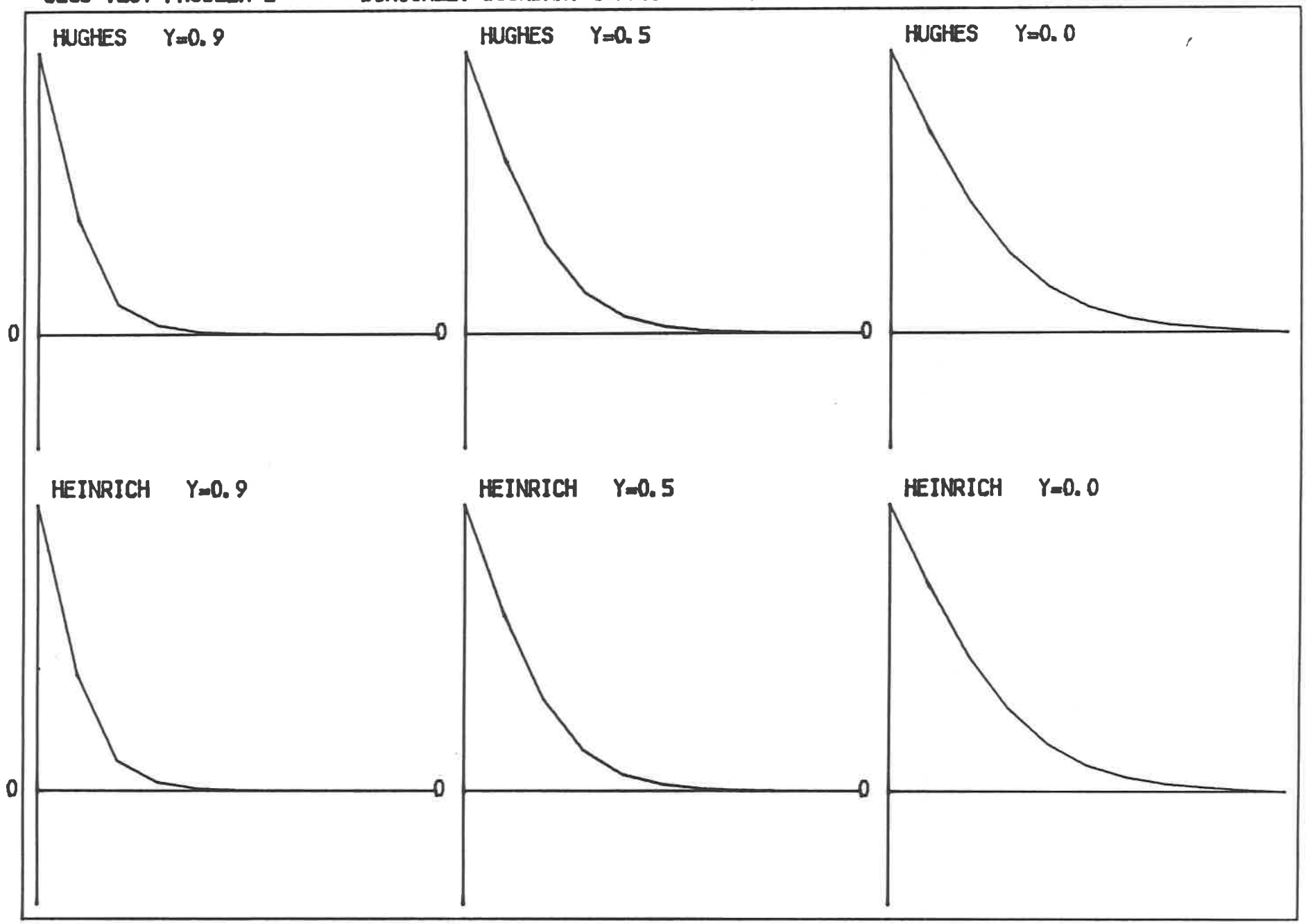

CEGB TEST PROBLEM 2
DIRICHLET BOUNDARY CONDS.
MESH PECLET NUMBER = 0.100E 01
HUGHES Y=0.9
HUGHES Y=0.5
HUGHES Y=0.0
HEINRICH Y=0.9
HEINRICH Y=0.5
HEINRICH Y=0.0
0
0
0
0
0
0

Figure 4.6

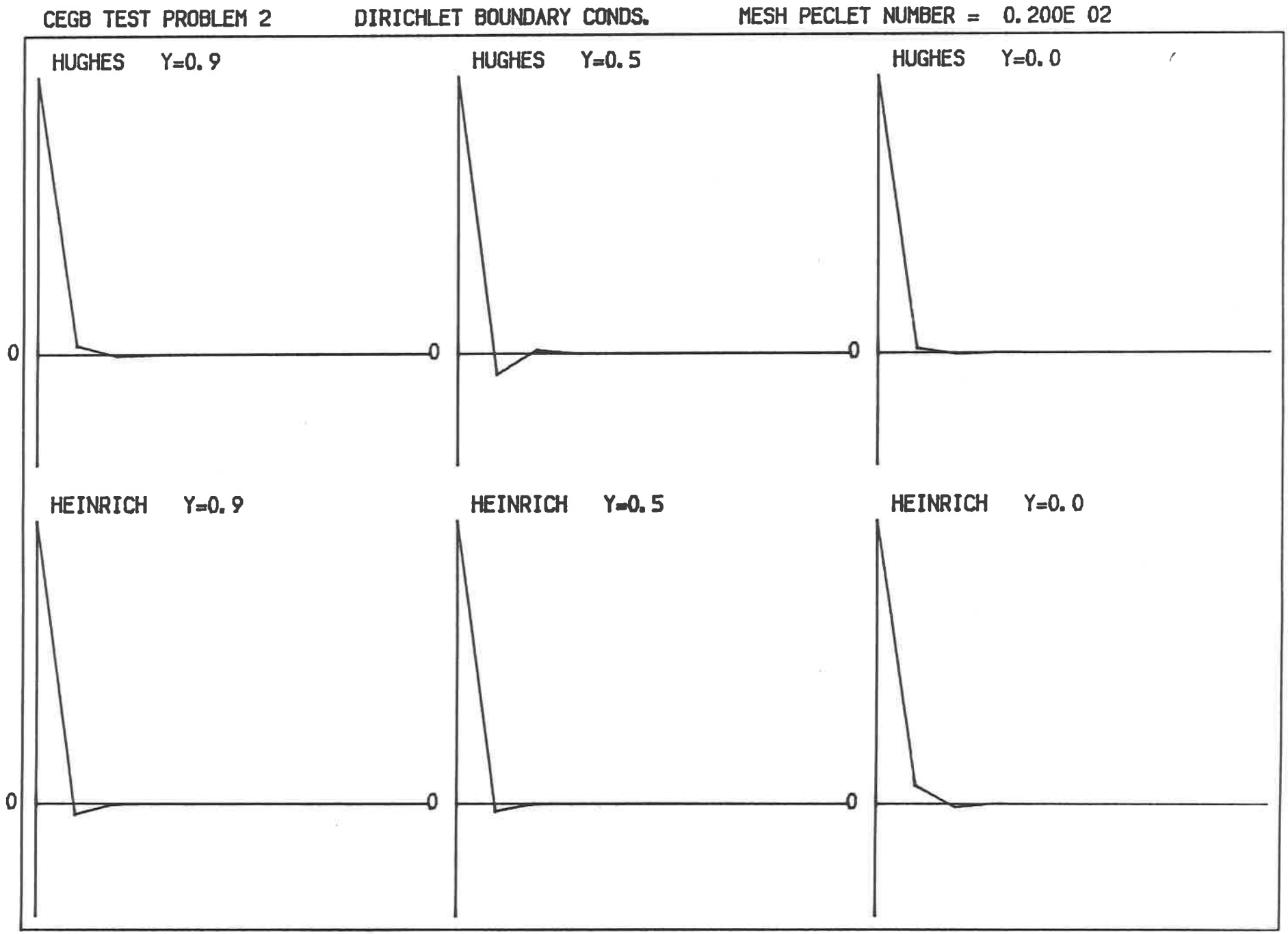

CEGB TEST PROBLEM 2
DIRICHLET BOUNDARY CONDS.
MESH PECLET NUMBER = 0.200E 02
HUGHES Y=0.9
HUGHES Y=0.5
HUGHES Y=0.0
HEINRICH Y=0.9
HEINRICH Y=0.5
HEINRICH Y=0.0
0
0
0
0
0
0

Figure 4.7

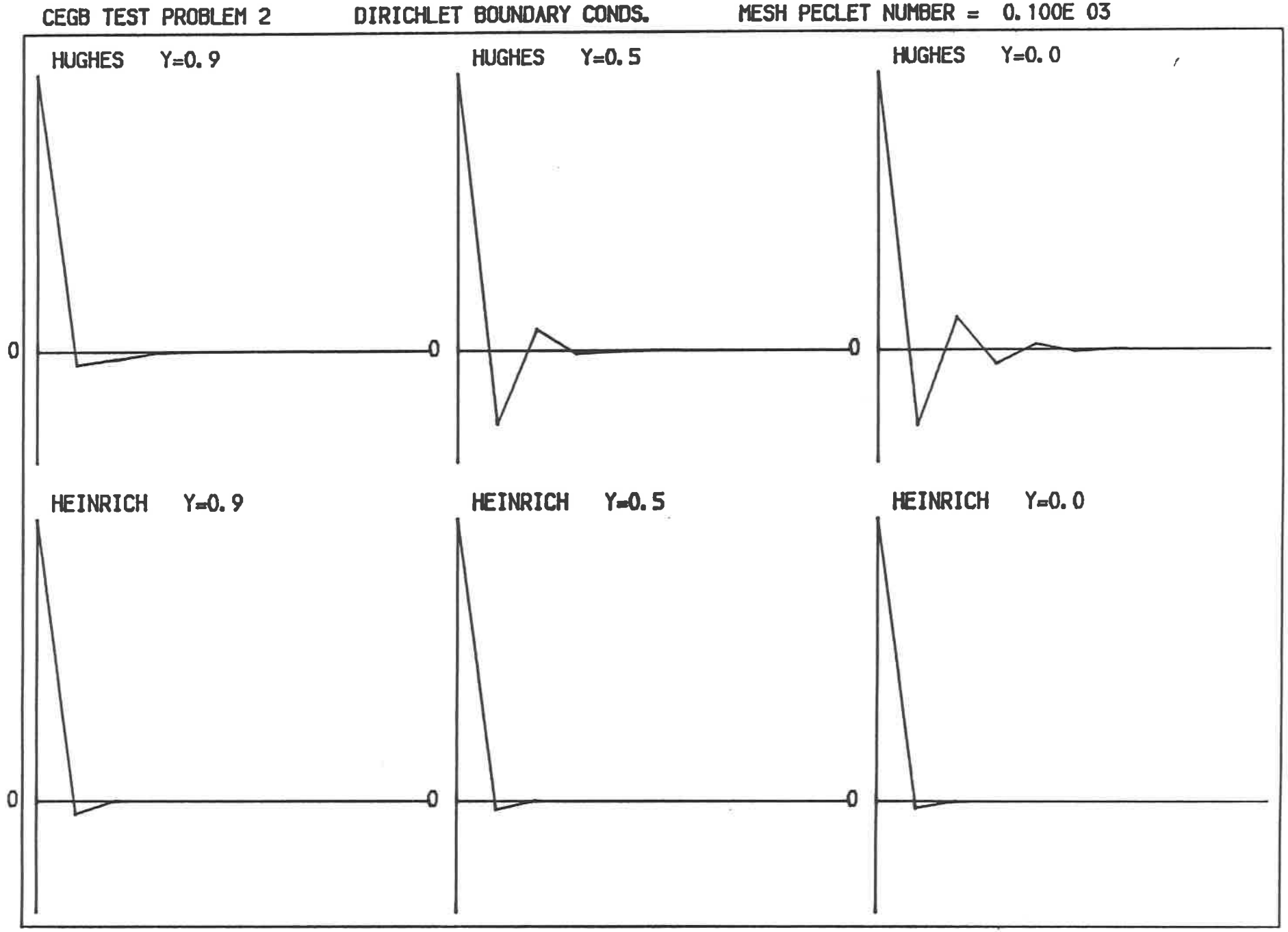

CEGB TEST PROBLEM 2
DIRICHLET BOUNDARY CONDS.
MESH PECLET NUMBER = 0.100E 03
HUGHES Y=0.9
HUGHES Y=0.5
HUGHES Y=0.0
HEINRICH Y=0.9
HEINRICH Y=0.5
HEINRICH Y=0.0
0
0
0
0
0
0

Figure 4.8

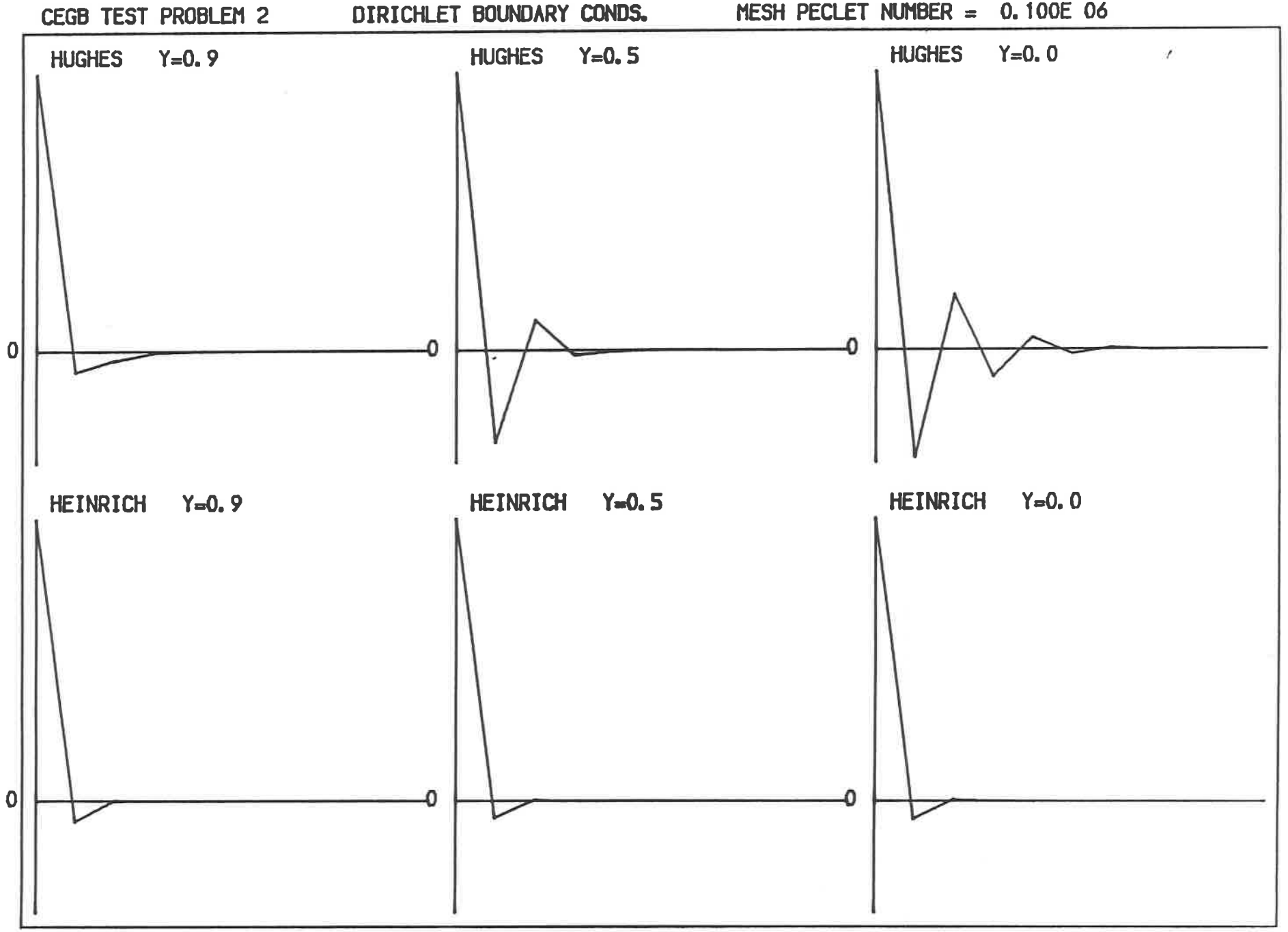

CEGB TEST PROBLEM 2
DIRICHLET BOUNDARY CONDS.
MESH PECLET NUMBER = 0.100E 06
HUGHES Y=0.9
HUGHES Y=0.5
HUGHES Y=0.0
HEINRICH Y=0.9
HEINRICH Y=0.5
HEINRICH Y=0.0
0
0
0
0
0
0

Figure 4.9

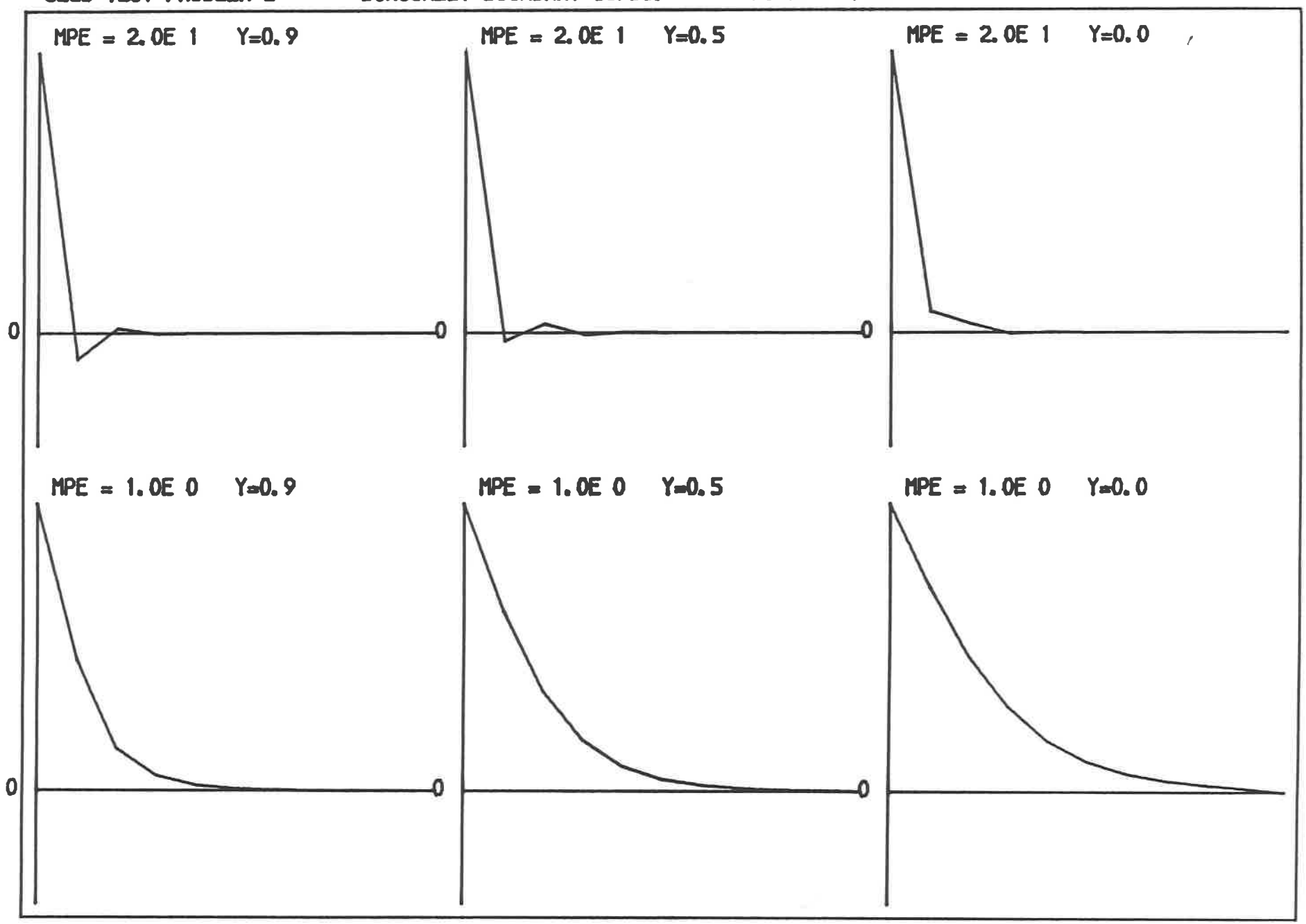

CEGB TEST PROBLEM 2
DIRICHLET BOUNDARY CONDS.
MIXED METHOD
MPE = 2.0E 1 Y=0.9
MPE = 2.0E 1 Y=0.5
MPE = 2.0E 1 Y=0.0
MPE = 1.0E 0 Y=0.9
MPE = 1.0E 0 Y=0.5
MPE = 1.0E 0 Y=0.0
0
0
0
0
0
0

Figure 4.10

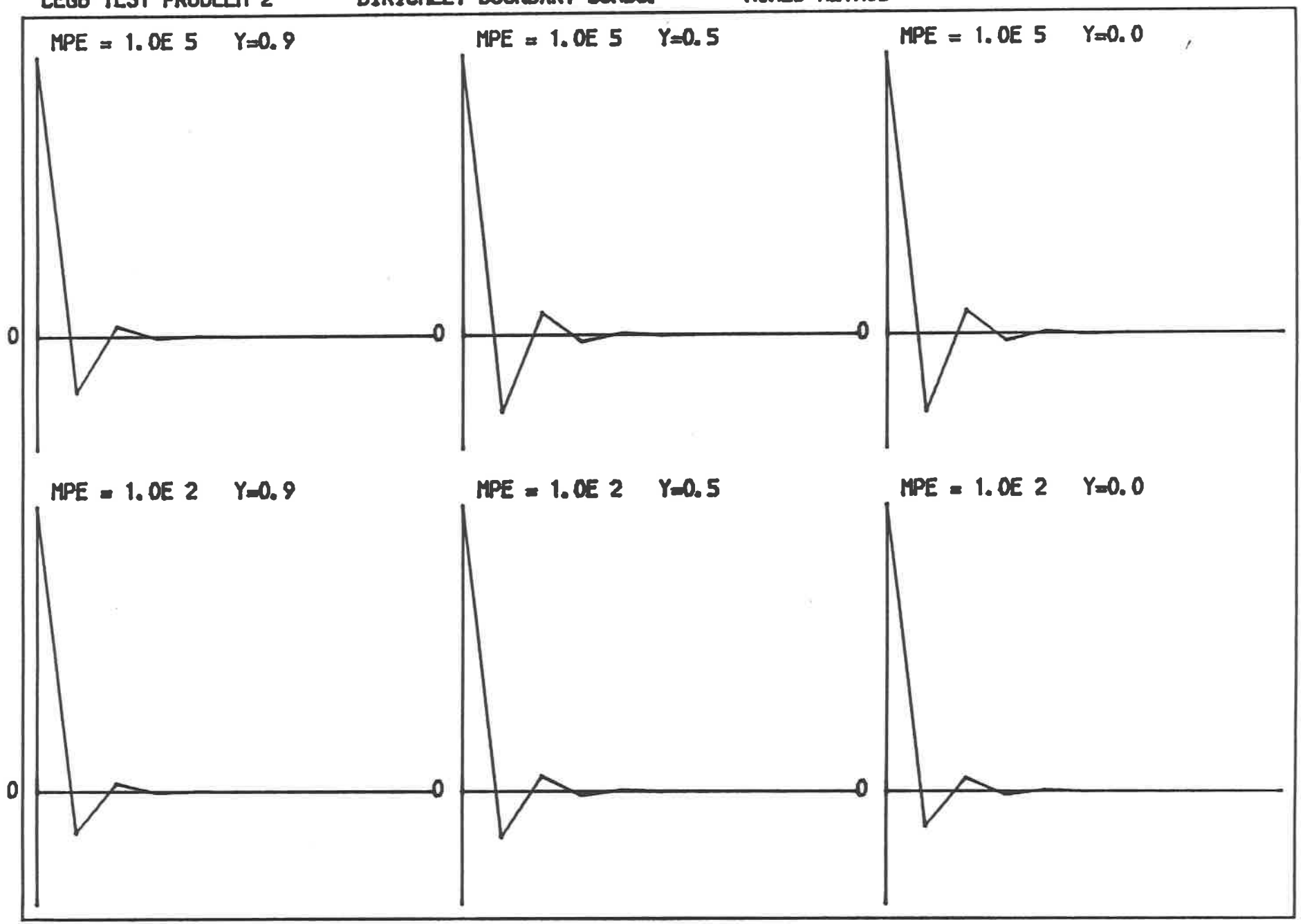

CEGB TEST PROBLEM 2
DIRICHLET BOUNDARY CONDS.
MIXED METHOD
MPE = 1.0E 5 Y=0.9
MPE = 1.0E 5 Y=0.5
MPE = 1.0E 5 Y=0.0
MPE = 1.0E 2 Y=0.9
MPE = 1.0E 2 Y=0.5
MPE = 1.0E 2 Y=0.0
0
0
0
0
0
0

Figure 4.11

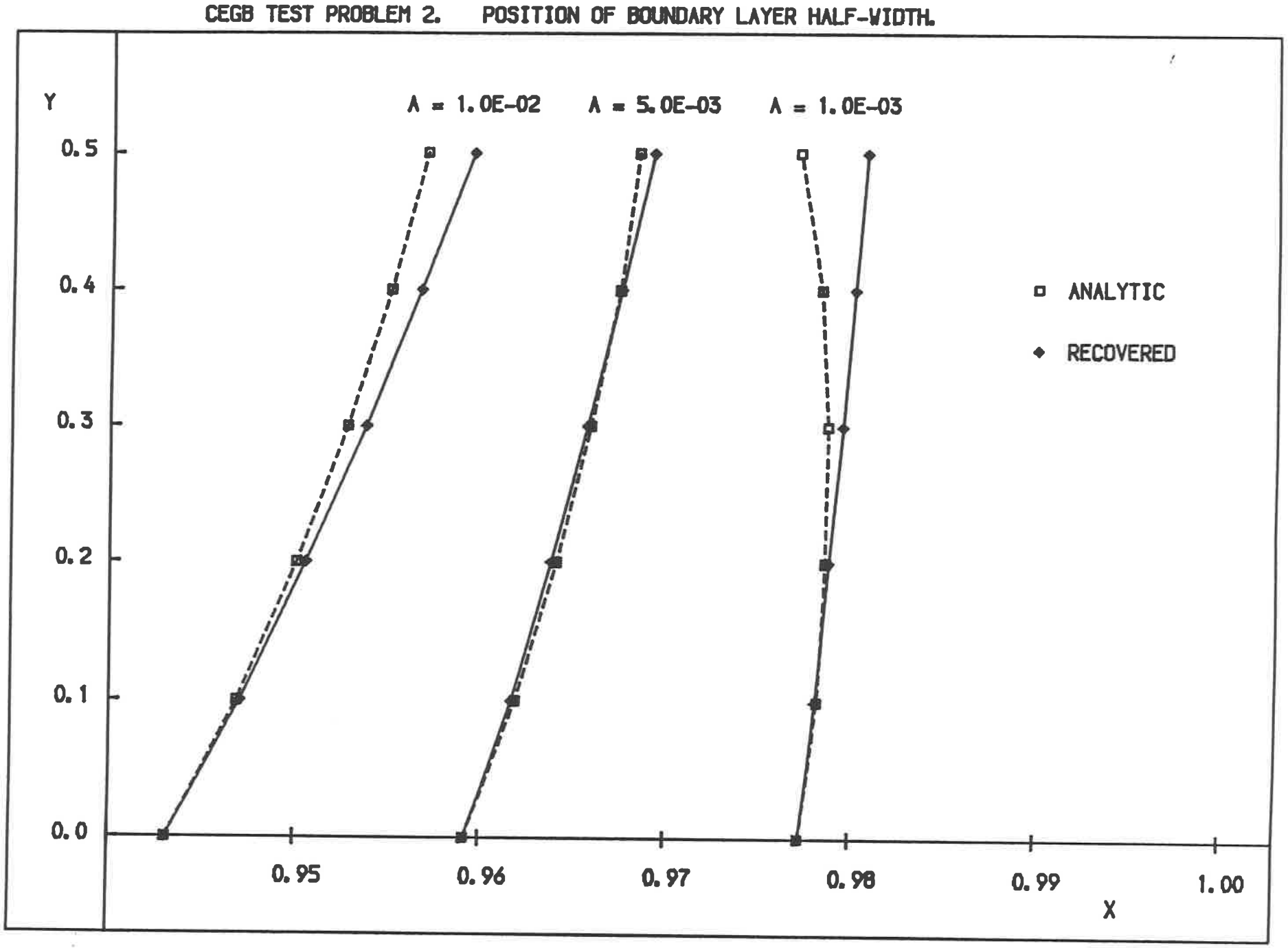

CEGB TEST PROBLEM 2. POSITION OF BOUNDARY LAYER HALF-WIDTH.
Y
A = 1.0E-02
A = 5.0E-03
A = 1.0E-03
0.5
0.4
0.3
0.2
0.1
0.0
ANALYTIC
RECOVERED
0.95
0.96
0.97
0.98
0.99
1.00
X

Figure 4.12