Optimal Petrov-Galerkin Methods through Approximate Symmetrization

J.W. Barrett and K.W. Morton

Numerical Analysis Report 4/80

Optimal Petrov-Galerkin Methods through Approximate Symmetrization

J.W. Barrett and K.W. Morton

Department of Mathematics, University of Reading, Reading RG6 2AX

A technique of approximate symmetrization is used to derive a test space from a given trial space for a Petrov-Galerkin method. This is applied to one-dimensional diffusion-convection problems to give approximations which are near optimal in an energy norm. Rigorous and precise error bounds are derived to demonstrate the uniformly good behaviour and near optimality of the procedure over all values of the mesh Péclet number.

1. Introduction

The success of finite element methods with elliptic problems $Lu = f$ stems largely from their optimality properties in the self-adjoint case. Of all members of a trial space S^h, the approximation obtained through the Galerkin formulation is the closest in the energy norm to the true solution - as a general reference see Strang & Fix (1973). Several valuable consequences follow from this fact: the coarsest possible mesh may be used with confidence; superconvergence phenomena allow more accurate and detailed information about the solution to be recovered, as in the work of Douglas & Dupont (1973a, b), Wheeler (1974), Dupont (1976), Bramble & Schatz (1977), Zienkiewicz (1977), Moan (1979), Lesaint & Zlámal (1979); and localised mesh refinement can be introduced, see Babuška & Rheinboldt (1978) and Rheinhardt (1980).

However, a lack of self-adjointness in the operator L erodes these

properties and eventually the Galerkin approximation becomes useless. Several techniques have been proposed to overcome this difficulty, most of them based on some sort of Petrov-Galerkin method. That is, one sets to zero the projection of the error $LU - f$ into some test space, which is in general different from the trial space. The rapidly growing literature on these techniques centres on the solution of convection dominated flow problems - see, for example, Guymon et al (1970), Christie et al (1976), Heinrich et al (1977), Hemker (1977), Barrett (1977), Hughes (1978) and the survey by Heinrich & Zienkiewicz (1979). Acceptable solutions can often be obtained to such problems but there is a lack of rigorous error estimates, certainly of the degree of sharpness possible in the self-adjoint case. Much of the existing analysis, and a good deal of the motivation in deriving the schemes, is based on viewing the discrete equations as finite difference approximations: as such, they are related to the schemes of Allen & Southwell (1955), Il'in (1969) and others, for which extensive analysis in the one-dimensional case has been carried out by Miller and his associates (see Miller, 1978).

In any Petrov-Galerkin method the main problem is the selection of the test-space and, ideally, one would like to select this without having to analyse the resulting discrete equations for truncation error. The present authors proposed in an earlier paper (Barrett & Morton, 1978) a Petrov-Galerkin technique based on approximately symmetrizing the bilinear form associated with a problem. In effect, the symmetrizing operator defines a test space T^h by a mapping from the trial space $N_\varepsilon : S^h \to T^h$ so that, when the original problem is approximated by a Petrov-Galerkin method using this test space, the symmetrized problem is treated by the Galerkin method: the subscript ε here denotes the extent

to which the symmetrizing operator N_ε is only approximate. If the symmetrization were exact, the resulting approximation would be optimal in an energy norm derived from the symmetric operator $N_0^* L$, where N_0^* is the formal adjoint of the operator N_ε with $\varepsilon = 0$. Numerical experiments on one dimensional problems presented in the earlier paper showed that effectively optimal approximations could be achieved in practice even for some turning-point problems. The task of recovering superconvergent information was also addressed there. In a later paper (Morton & Barrett, 1980) preliminary consideration was given to two-dimensional problems.

The aim of the present paper is to derive rigorous and precise error estimates when the procedure is applied to the one-dimensional diffusion-convection problem for $u(x)$:

$$-\big(a(x)u'(x)\big)' + \big(b(x)u(x)\big)' = f(x) \quad \text{on} \quad (0, 1) \qquad (1.1)$$

$$u(0) = g_L, \quad \gamma u(1) + (1 - \gamma)u'(1) = \gamma g_R, \qquad (1.2)$$

where either $\gamma = 1$, giving a Dirichlet problem, or $\gamma = 0$, giving a homogeneous Neumann condition at $x = 1$. These error estimates will take various forms but the particular aim is to establish the extent to which the proposed procedure achieves optimality and to do this with a bound which is uniform with regard to the ratio b/a so that the results are valid for singular perturbation problems. Only linear elements are studied in detail: though the analytical framework is presented for general elements and similar detailed results could be derived for higher order elements, it could be argued that there is less need to exploit the property of optimality in such cases and a weaker property such as uniform stability of the approximation, is adequate.

The arrangement of the paper is as follows. In the next section the sym-

metrization technique is described and basic error estimates derived. Then in section 3 the special case in which the coefficients a and b are positive constants is studied in detail. The precise analysis possible then shows how to choose the perturbation function $\varepsilon(x)$ involved in the symmetrization: it turns out that an appropriately positioned δ-function (in this case at $x = 0$) gives both practical advantages over the choice used in the earlier paper as well as sharper error bounds, though the method then lies outside the usual Petrov-Galerkin framework. An a posteriori bound on the difference between the nodal values U_j^N of the approximation U^N and those of the optimal approximation U^* is given in the form

$$|U_j^N - U_j^*| \le \frac{3}{2b}\left[a\left|\frac{U_1^N - g_L}{h} - \frac{f(0)}{b}\right| + \int_0^1 e^{-bx/a}|f(x) - f(0)|\right] + O\left(e^{-b/a}\right). \tag{1.3}$$

Terms which are $O\left(e^{-b/a}\right)$ are entirely negligible for the problems considered since, as shown by Schatz (1974), Galerkin methods still have optimal order of accuracy even for non-self-adjoint problems, the loss of absolute optimality for (1.1) being just proportional to b/a: thus Petrov-Galerkin methods need only be used when b/a is quite large, typically when the "mesh Péclet number" bh/a is of the order of unity. Typical a priori error bounds in section 3 show the extent to which optimality is lost with the present technique: in terms of the energy norm

$$\|v\|_S^2 = \int_0^1 (a^2 v'^2 + b^2 \bar{v}^2)\,dx \tag{1.4}$$

one such bound is given by

$$\|u - U^N\|_S^2 \le \|u - U^*\|_S^2 + a^2[u'(0) - U^{*\prime}(0)]^2. \tag{1.5}$$

For singular perturbation problems, normalisation is carried out relative to we suppose
b and $'/a \to 0$: as boundary layers occur at $x = 1$ rather than $x = 0$, (1.5) becomes increasingly sharp.

Finally, in section 4 the variable coefficient problems are considered, firstly non-turning-point problems and then problems involving a single turning point, with $b(x) \le 0$ for $x < \xi$ and $b(x) \ge 0$ for $x > \xi$ - the so-called 'accelerating flow' case. The form of the error bounds is very little changed and, most importantly, the uniformly good behaviour of the approximation as $b/a \to \infty$ is retained. Some precision in the estimates is lost of course so attention is concentrated on the a posteriori bounds with the a priori bounds merely related to those for associated self-adjoint problems.

2. The Symmetrization Technique and Basic Error Estimates

Weak Formulation.

For the solution of the problems (1.1), (1.2) we introduce the following sets of functions:

$$H_0^1 = \{v \in H^1(0, 1) \mid v(0) = 0, \text{ and } v(1) = 0 \text{ if } \gamma = 1\} \quad (2.1a)$$

$$H_E^1 = \{v \in H^1(0, 1) \mid v(0) = g_L, \text{ and } v(1) = g_R \text{ if } \gamma = 1\}, \quad (2.1b)$$

where $H^m(0, 1)$ is the usual Sobolev space of functions with m^{th} derivatives square integrable. The bilinear form associated with (1.1) and (1.2) is given by

$$B(w_1, w_2) = \langle aw_1', w_2' \rangle + \langle (bw_1)', w_2 \rangle, \quad (2.2)$$

where the inner product $\langle w_1, w_2 \rangle$ denotes the integral $\int_0^1 w_1 w_2 dx$.

Then the weak form of the problems is: find $u \in H_E^1$ such that

$$B(u, v) = \langle f, v \rangle, \qquad \forall\, v \in H_0^1. \quad (2.3)$$

The existence and uniqueness of a solution to (2.3) is guaranteed by the generalised Lax-Milgram theorem (Babuška & Aziz, 1972) under the following conditions:

Assumptions 1

(i) $a(x) > 0 \quad \forall\, x \in [0, 1]$, (2.4a)

(ii) $a, b \in H^1(0, 1)$, (2.4b)

(iii) $f \in L_2(0, 1)$. (2.4c)

Symmetrization

We define a symmetric bilinear form with a positive weight function $\rho(x)$ as

$$B_S(w_1, w_2) = \langle [aw_1' - bw_1], \rho[aw_2' - bw_2] \rangle + [\rho abw_1w_2](1) \tag{2.5a}$$

$$= \langle \rho a^2 w_1', w_2' \rangle + \langle [\rho b^2 + (\rho ab)']w_1, w_2 \rangle, \tag{2.5b}$$

$$\forall\, w_1 \in H^1(0, 1),\ w_2 \in H_0^1.$$

Then the symmetrizing operator $N : H_0^1 \to H_0^1$ has to satisfy

$$B(w_1, Nw_2) = B_S(w_1, w_2), \qquad \forall\, w_1 \in H^1(0, 1),\ w_2 \in H_0^1, \tag{2.6}$$

and the mapping has to be onto for (2.3) to be equivalent to the following symmetric problem: find $u \in H_E^1$ such that

$$B_S(u, v) = \langle f, Nv \rangle, \qquad \forall\, v \in H_0^1. \tag{2.7}$$

Unfortunately this is not always possible. Comparing (2.2) with (2.5a) we see that it requires that

$$(Nv)' = \rho[av' - bv] \quad \text{in} \quad (0, 1), \tag{2.8}$$

together with the two boundary conditions

$$(Nv)(0) = 0, \qquad (Nv)(1) = (\rho av)(1). \tag{2.9}$$

This can be achieved only with an operator N consisting of multiplication by an exponential of large argument and leading to an energy norm with this

exponential as the weighting factor ρ - see Barrett & Morton (1978) for details. Though exponential test functions are used with some methods - for instance, with those of Hemker (1977) and Barrett (1977) which directly seek accurate nodal values - this choice would be unduly restrictive in the present context: for the exponentially weighted norm will often lead to an ill-conditioned recovery problem from a best fit which has concentrated information away from the main regions of interest, though it should be noted that Dixon et al (1979) have used this choice for certain model problems. We therefore consider a general weight function and, for a problem with no turning points, introduce a non-negative perturbation function $\varepsilon(x)$, normalised by $\int_0^1 \varepsilon \, dx = 1$, and a corresponding operator $N_\varepsilon : H_0^1 \to H_0^1$ given by:

$$(N_\varepsilon v)' = \rho[av' - bv] + \langle \alpha, v \rangle \varepsilon \quad \text{in} \quad (0, 1), \tag{2.10}$$

where setting

$$\alpha = \rho b + (\rho a)' \tag{2.11}$$

will allow the imposition of

$$(N_\varepsilon v)(0) = 0, \qquad (N_\varepsilon v)(1) = (\rho a v)(1). \tag{2.12}$$

Thus we have

$$B(w_1, N_\varepsilon w_2) = B_S(w_1, w_2) + \langle aw_1' - bw_1, \varepsilon \rangle - \langle \alpha, w_2 \rangle, \qquad \forall\, w_1 \in H^1(0, 1),\ w_2 \in H_0^1. \tag{2.13}$$

It is also convenient to introduce the operator N_0 given by

$$(N_0 v)' = \rho[av' - bv] \quad \text{in} \quad (0, 1), \quad (N_0 v)(1) = (\rho a v)(1) \tag{2.14}$$

for which (2.13) holds with $\varepsilon(x)$ set equal to the Dirac delta function at $x = 0$.

Trial and test spaces.

We use standard trial spaces S^h, where h is a mesh parameter, which satisfy the usual approximation properties, as given by Babuška & Aziz (1972):

(i) $S^h \subset H^1(0, 1), \quad 0 < h < 1,$

(ii) for each $w \in H^r(0, 1), \quad \exists\, W \in S^h$ such that

$$\|w - W\|_{H^\ell(0, 1)} \leq Ch^{r-\ell} \|w\|_{H^r(0, 1)} \quad \text{for } \ell = 0, 1 \qquad (2.15)$$

where $r \geq 2$ is a given integer.

To apply the Petrov-Galerkin method we then construct test spaces by setting $T^h = N_\varepsilon S^h$. From a practical viewpoint the set of functions $\{N_\varepsilon \phi_j\}$ does not form the most convenient basis for T^h: appropriate linear combinations of successive test functions have the advantage of better localisation as well as being independent of the perturbation $\varepsilon(x)$. If we define

$$\alpha_j = \langle \alpha, \phi_j \rangle \qquad (2.16)$$

then we have from (2.10)

$$\begin{aligned} \chi_j(x) &\equiv \alpha_{j-1}(N_\varepsilon \phi_j)(x) - \alpha_j(N_\varepsilon \phi_{j-1})(x) \\ &= \alpha_{j-1}(N_0 \phi_j)(x) - \alpha_j(N_0 \phi_j)(x). \end{aligned} \qquad (2.17)$$

Depending on the problem being considered, and assuming that it has no turning points, T^h can be based on $\{\chi_j\}$ for an appropriate set of j-values plus just one function $N_\varepsilon \phi_{j_0}$ for some choice of j_0. Thus the perturbation function affects only one element of the basis and the numerical examples in Barrett & Morton (1978) demonstrate how appropriate choices of $\varepsilon(x)$ and j_0 can achieve an almost optimal approximation U^N to u. This will also be clear from the error analysis below.

For turning point problems, the interval is subdivided into subinter-

vals in which either $b(x) \geq 0$ or $b(x) \leq 0$ and an appropriate perturbation function, together with the attendent set of test functions, constructed for each.

Basic error estimates.

For the error analysis it is more convenient to work directly with $\{N_\varepsilon \phi_j\}$. Then the Petrov-Galerkin method becomes: find $U^N \in S_E^h$ such that

$$B(U^N, N_\varepsilon \phi_j) = \langle f, N_\varepsilon \phi_j \rangle, \qquad \forall\, \phi_j \in S_0^h. \tag{2.18}$$

We introduce the following assumptions for later use:

Assumptions 2

$$\text{(i)} \quad \rho(x) > 0 \quad \forall\, x \in [0, 1] \text{ and } \int_0^1 \rho dx = 1; \tag{2.19a}$$

$$\text{(ii)} \quad \rho \in H^1(0, 1); \tag{2.19b}$$

$$\text{(iii)} \quad [\rho b^2 + (\rho a)'](x) > 0, \quad \forall\, x \in [0, 1]; \tag{2.19c}$$

$$\text{(iv)} \quad \|\rho b + (\rho a)'\|_{-S} \equiv \sup_{v \in H_0^1} \frac{|\langle \rho b + (\rho a)', v \rangle|}{\|v\|_S} \leq K_\rho, \tag{2.19d}$$

where by (2.5b), (2.19a) and (2.19c) we can deduce that $B_S(u, v)$ is coercive and define the energy norm $\|\cdot\|_S$ on $H^1(0, 1)$ by

$$\|v\|_S^2 = B_S(v, v) = \langle \rho(av' - bv), (av' - bv) \rangle + [\rho a b v^2](1) \tag{2.20a}$$

$$= \langle \rho a^2 v', v' \rangle + \langle [\rho b^2 + (\rho a b)'] v, v \rangle, \qquad v \in H_0^1. \tag{2.20b}$$

Assumption (2.19c) is really only necessary at this stage when $\gamma = 0$ and even then by (2.20a) it can be obviated by assuming that $b(1) \geq 0$: note that in the context of diffusion convection problems this latter condition corresponds to the common assumption that Neumann conditions are imposed only at an outflow boundary. Assumption (2.19d) is also only necessary for the Neumann problem and when ρa or $\rho a b$ are non-constant: in the Dirichlet case we have

$$|\langle \rho b + (\rho a)', v\rangle| = |\langle \rho^{\frac{1}{2}}, \rho^{\frac{1}{2}}(bv - av')\rangle|$$

$$\leq \|\rho^{\frac{1}{2}}(bv - av')\|_2 = \|v\|_S$$

i.e. $$\|\rho b + (\rho a)'\|_{-S} \leq 1, \quad \text{for} \quad \gamma = 1; \tag{2.21}$$

and when $(\rho a)' = (\rho ab)' = 0$ comparison of (2.19d) with (2.20b) shows immediately that again we have $K_\rho \leq 1$.

Assumptions (2.4b) and (2.19b) guarantee that $N_\varepsilon \phi_j \in H_0^1$ and hence, for non-turning point problems, from (2.13) and (2.16) we have

$$0 = B(u - U^N, N_\varepsilon \phi_j) = B_S(u - U^N, \phi_j) + \alpha_j \ell(u - U^N), \qquad \forall\, \phi_j \in S_0^h \tag{2.22}$$

where

$$\ell(v) = \langle av' - bv, \varepsilon\rangle. \tag{2.23}$$

We also introduce $U^* \in S_E^h$, the optimal approximation to u in the energy norm, given by

$$B_S(u - U^*, \phi_j) = 0, \qquad \forall\, \phi_j \in S_0^h, \tag{2.24}$$

and decompose the error into an approximation error $e^A \in H_0^1$ and a perturbation error $e^\varepsilon \in S_0^h$:

$$u - U^N = e^A + e^\varepsilon, \tag{2.25}$$

where $e^A = u - U^*$, $e^\varepsilon = U^* - U^N$.

Combining (2.22) and (2.24) we obtain

$$B_S(e^\varepsilon, \phi_j) = -\alpha_j[\ell(e^A) + \ell(e^\varepsilon)], \qquad \forall\, \phi_j \in S_0^h. \tag{2.26}$$

Suppose now we define $V^* \in S_0^h$ by

$$B_S(V^*, \phi_j) = \alpha_j, \qquad \forall\, \phi_j \in S_0^h: \tag{2.27}$$

from the definition of α in (2.11) we see immediately that $\|V^*\|_S \leq K_\rho$. Then we have the following results:

Theorem 1 Suppose Assumptions 1 and 2 hold, and u, U^* and V^* are defined by (2.3), (2.24) and (2.27). Then if $1 + \ell(V^*) > 0$, there exists a unique solution U^N to (2.18) which satisfies

$$U^N - U^* = [\ell(u - U^N)]V^* \tag{2.28}$$

and

$$\|u - U^N\|_S^2 \leq \|u - U^*\|_S^2 + K_\rho[1 + \ell(V^*)]^{-2}[\ell(u - U^*)]^2. \tag{2.29}$$

Proof: As (2.18) generates a finite system of linear equations, existence follows from uniqueness, for which we have to show that $U^{(0)} \equiv 0$ is the only solution of (2.18) with $f \equiv 0$. Substituting from (2.27) into (2.22) with $u \equiv 0$ gives

$$\begin{aligned} B_S(U^{(0)}, \phi_j) &= -\alpha_j \ell(U^{(0)}) \\ &= -\ell(U^{(0)})B_S(V^*, \phi_j). \end{aligned}$$

The coercivity of $B_S(\cdot, \cdot)$ ensures that this system is non-singular and hence that

$$U^{(0)} = -\ell(U^{(0)})V^*,$$

so that

$$\ell(U^{(0)})[1 + \ell(V^*)] = 0.$$

Thus if $1 + \ell(V^*) > 0$, we have $\ell(U^{(0)}) = 0$ and $U^{(0)} \equiv 0$.

By a similar manipulation (2.28) follows from (2.22) and (2.27). Thence by operating on (2.28) with ℓ and using the decomposition (2.25) we have

$$\ell(e^\varepsilon) + \ell(V^*)[\ell(e^A) + \ell(e^\varepsilon)] = 0, \tag{2.30}$$

$$e^\varepsilon \equiv U^* - U^N = -[1 + \ell(V^*)]^{-1}\ell(e^A)V^* \tag{2.31}$$

and from (2.26)

$$\begin{aligned} \|e^\varepsilon\|_S^2 &= -[1 + \ell(V^*)]^{-1}\ell(e^A) \ \langle \alpha, e^\varepsilon \rangle \\ &\leq [1 + \ell(V^*)]^{-1}|\ell(e^A)| \ \|\alpha\|_{-S} \|e^\varepsilon\|_S. \end{aligned} \tag{2.32}$$

From (2.24) and (2.25) we have

$$\|u - U^N\|_S^2 = \|e^A\|_S^2 + \|e^\varepsilon\|_S^2 \qquad (2.33)$$

and the final result (2.29) therefore follows from (2.32) by using the bound on $\|\alpha\|_{-S} \equiv \|\rho b + (\rho a)'\|_{-S}$ assumed in (2.19d). ■

A posteriori and a priori estimates.

It is relatively easy to estimate V^* and in any case it can be calculated from (2.27) for the same effort as calculating U^N. From (2.28), we therefore see that U^N is close to U^* if $\ell(u - U^N)$ is small. The choice of the perturbation function is guided by this aim and a check on the success of a given choice is provided by a calculation of $\ell(U^N)$, with an a posteriori error bound following if $\ell(u)$ is also estimated. To obtain an a priori estimate from (2.29) we need also to estimate $\ell(V^*)$ and $\ell(e^A)$. These estimates will be given in the next two sections.

Turning point problems.

To complete this section we return to general turning point problems and consider the effect of using several different perturbation functions $\varepsilon_m(x)$, $m = 1, 2, \ldots, M$, each normalised as before. We define the M-column matrix A by

$$A_{jm} = \begin{cases} \langle \rho b + (\rho a)', \phi_j \rangle, & \text{if } N_\varepsilon \phi_j \text{ is defined with } \varepsilon_m \\ 0, & \text{otherwise,} \end{cases} \qquad (2.34)$$

and the M-row vector $\underline{\ell}(v)$ by

$$\ell_m(v) = \langle av' - bv, \varepsilon_m \rangle. \qquad (2.35)$$

Then (2.26) generalises to

$$B_S(e^\varepsilon, \phi_j) = -\{A[\underline{\ell}(e^A) + \underline{\ell}(e^\varepsilon)]\}_j \qquad \forall\, \phi_j \in S_0^h. \qquad (2.36)$$

Suppose now we define the M-row vector function $\underline{V}^* \in [S_0^h]^M$ by

$$B_S(\{\underline{V}^*\}_m, \phi_j) = A_{jm}, \qquad \forall\ \phi_j \in S_0^h \tag{2.37}$$

and denote by L the $M \times M$ matrix

$$L_{mn} = \ell_m(\{\underline{V}^*\}_n). \tag{2.38}$$

Then we have the following generalisation of Theorem 1.

Theorem 2 Suppose Assumptions 1 and 2 hold and u, U^* and $\underline{V}^*$ are defined by (2.3), (2.24) and (2.37). Then if the $M \times M$ matrix $I + L$ is non-singular, there exists a unique solution U^N to (2.18) which satisfies

$$U^N - U^* = [\underline{\ell}(u - U^N)]^T \underline{V}^* \tag{2.39}$$

and

$$\|u - U^N\|_S^2 \le \|u - U^*\|_S^2 + K_\rho \|(I + L)^{-1} \underline{\ell}(u - U^*)\|_\infty^2 . \tag{2.40}$$

Proof: With the definitions and relations (2.34) - (2.38), the proof follows the same lines as in Theorem 1. When $f \equiv 0$, we have

$$U^{(0)} = -[\underline{\ell}(U^{(0)})]^T \underline{V}^*,$$

which implies that

$$(I + L)\underline{\ell}(U^{(0)}) = 0$$

and hence $U^{(0)} \equiv 0$ under the given hypothesis. The result (2.39) follows immediately from (2.36) and (2.37) by the coercivity of $B_S(\cdot,\cdot)$ and hence

$$(I + L)\underline{\ell}(e^\varepsilon) = -L\underline{\ell}(e^A). \tag{2.41}$$

Substituting for $\ell(e^\varepsilon)$ in (2.36) then gives

$$\|e^\varepsilon\|_S^2 = -\sum_j e_j^\varepsilon \{A(I + L)^{-1} \underline{\ell}(e^A)\}_j$$

$$\le \|(I + L)^{-1} \underline{\ell}(e^A)\|_\infty \langle \alpha, e^\varepsilon \rangle, \tag{2.42}$$

since there is only one non-zero element in each row of A, because

each $N_\varepsilon\phi_j$ is defined with one and only one ε_m. The final argument to obtain (2.40) is the same as in Theorem 1. ■

Further assumptions are needed to turn the basic error estimates of Theorems 1 and 2 into precise error bounds, and this will be done in the next two sections. As implied by the characterisation of S^h in (2.15), quite general basis functions can be used on a non-uniform mesh but in these next sections we shall consider only linear elements on a uniform mesh. It should be noted here too that the arguments used in these two theorems may be used to establish the uniform stability of the approximations U^N as $b/a \to \infty$ when the discrete Green's function associated with $B_S(\cdot,\cdot)$ can be estimated. Thus for the non-turning point problem of Theorem 1, suppose $Z \in S_E^h$ satisfies the equation $B_S(Z, \phi_j) = \langle f, N_\varepsilon\phi_j\rangle$, $\forall\, \phi_j \in S_0^h$. Then from (2.18), (2.22) and (2.27)

$$U^N + \ell(U^N)V^* = Z \,. \tag{2.43}$$

$$\ell(U^N) = \ell(Z)/[1 + \ell(V^*)]. \tag{2.44}$$

Since $\|V^*\|_S \le K_\rho$, we have

$$\|U^N\|_S \le \|Z\|_S + K_\rho|\ell(Z)|/[1 + \ell(V^*)] \tag{2.45}$$

and bounding $\|Z\|_S$ and $|\ell(Z)|$ in terms of f and the boundary data establishes the stability of the approximation when $1 + \ell(V^*)$ has been bounded from zero.

3. Constant Coefficient Problems

In this section we establish the choice of $\varepsilon(x)$ such that $1 + \ell(V^*) > 0$ and develop more precise error bounds under the assumption that the coefficients a and b in (1.1) are constants; there is no loss of generality in assuming they are positive. The problem for $u(x)$ then

reduces to

$$-au'' + bu' = f \quad \text{on} \quad (0, 1); \quad a, b > 0; \quad f \in L_2(0, 1) \tag{3.1}$$

$$u(0) = g_L, \quad (1 - \gamma)u'(1) + \gamma u(1) = \gamma g_R, \tag{3.2}$$

where $\gamma = 0$ or 1. We shall also assume that the weight function $\rho(x)$ is constant, in fact $\rho(x) \equiv 1$, and that S^h is spanned by piecewise linear functions over a uniform mesh with interval $h = 1/J$. First of all we examine the method when the perturbation function is taken to be constant over a single element.

Properties of V^*.

The subsidiary function V^* defined in (2.27) can be constructed explicitly with the following results.

Lemma 3.1 (i) For constants a, b the function V^* is the Galerkin approximation in S^h to $v(x)$ defined by

$$-a^2v'' + b^2v = b \quad \text{in} \quad (0, 1); \quad v(0) = 0, \quad (1 - \gamma)v'(1) + \gamma v(1) = 0; \tag{3.3}$$

(ii) with linear elements on a uniform mesh,

$$V_j^* = \frac{1}{b} \frac{(1 - r^j)(1 - r^{J(2-\gamma)-j})}{1 + r^{J(2-\gamma)}}, \qquad j = 0, 1, \ldots, J \tag{3.4}$$

where

$$r = \nu/[1 + (1 - \nu^2)^{\frac{1}{2}}], \quad \nu = (6a^2 - b^2h^2)/(6a^2 + 2b^2h^2);$$

(iii) and for the perturbation function $\epsilon_k(x) = h^{-1}\chi_{I_k}(x)$, $I_k = [kh, (k + 1)h]$,

$$\ell(V^*) = \frac{1}{2h}[(2a - bh)V_{k+1}^* - (2a + bh)V_k^*]. \tag{3.5}$$

Proof: The definition of V^* in (2.27) and α_j in (2.16) gives (3.3) immediately. Then a straight-forward computation gives (3.4) and the integral (2.23) gives (3.5). ■

It is clear from (3.4) that, if $b^2h^2 = 6a^2$ so that $r = 0$, then $V_j^* = b^{-1}$ for $0 < j < J$. Hence, if $0 < k < J - 1$, for this value of bh/a we have $\ell(V^*) = -1$ and the existence and uniqueness of U^N breaks down in Theorem 1, as the governing equations become singular. This indicates that the perturbation should be made near the ends of the interval and, indeed, the uniform stability of the equations for U^N will be lost unless the integral of $\varepsilon(x)$ over the two end sub-intervals together is uniformly bounded away from zero. Although V^* itself only distinguishes between the two ends in the Neumann problem, $1 + \ell(V^*)$ is always largest if we take $k = 0$. We now consider a class of perturbation functions contained in the first interval.

Lemma 3.2 Under the assumptions of Lemma 3.1 but with

$$\varepsilon(x) = \begin{cases} (\lambda h)^{-1}, & 0 \le x \le \lambda h \le h \\ 0, & \lambda h < x \le 1 \end{cases} \tag{3.6}$$

we have

$$1 + \ell(V^*) \ge 1 - \frac{3}{4}\lambda, \quad \text{for all } a, b \text{ and } h > 0. \tag{3.7}$$

Proof: From (2.23) and (3.6), we have

$$\ell(V^*) = \left[\frac{a}{h} - \frac{\lambda b}{2}\right] V_1^*. \tag{3.8}$$

Writing $K = (2 - \gamma)J$, we obtain from (3.4)

$$V_1^* = \frac{1}{b}\frac{(1 - r)(1 - r^{K-1})}{1 + r^K} = \frac{1}{b}\left[1 - \frac{r(1 + r^{K-2})}{1 + r^K}\right].$$

Since $|r| \le 1$ and $K \ge 2$ we have

$$(1 + r^{K-2})/(1 + r^K) \le 2/(1 + r^2),$$

and from the definitions of r and ν,

$$2r/(1 + r^2) = \nu \quad \text{and} \quad -\tfrac{1}{2} < \nu \le 1.$$

We therefore have the following bound

$$0 \le V_1^* < 3/2b, \tag{3.9}$$

and the desired bound (3.7) then follows immediately from (3.8): it is attained for $K = 2$ as $bh/a \to \infty$. ■

We can see from (3.7) that the sharpest results will be obtained for λ as small as possible. In fact from a practical and theoretical point of view it is desirable to pass to the limit $\lambda = 0$, when $\varepsilon(x)$ becomes the Dirac delta function at 0, denoted by $\delta(x)$. The resulting test functions for $a, b > 0$ and $\rho \equiv 1$ are then given by

$$N_\varepsilon \phi_j(x) = a\phi_j(x) + \int_x^1 b\phi_j(t)dt. \tag{3.10}$$

They satisfy the differential equation (2.8) but not the boundary conditions (2.9), and thus have stepped outside the standard Petrov-Galerkin framework since $N_\varepsilon \phi_j \notin H_0^1$. It is a simple exercise to check that all the previous results and definitions hold as we pass to this limiting choice of $\varepsilon(x)$. Below we draw together the assumptions used throughout the remainder of this section:

Assumptions 3

(i) a and b are positive constants, and $f \in L_2(0, 1)$

(ii) $\rho \equiv 1$

(iii) S^h is the space of piecewise linear functions on a uniform mesh

(iv) $\varepsilon(x) \equiv \delta(x)$ giving $\ell(v) = av'(0) - bv(0)$

Estimates of $\ell(u)$.

The solution $u(x)$ of (3.1) and (3.2) can be expressed explicitly in terms of integrals of $f(x)$ and, as b/a becomes large, $\ell(u)$ is seen to depend principally on $f(x)$ for values of x for which $\varepsilon(x) \neq 0$.

Lemma 3.3 Under Assumptions 3 we have:

$$\ell(u - g_L) = \int_0^1 e^{-bx/a} f(x)dx + O(e^{-b/a}) \tag{3.11}$$

$$= (a/b)f(0) + \int_0^1 e^{-bx/a}[f(x) - f(0)]dx + O(e^{-b/a}). \tag{3.12}$$

Proof: From Assumption 3(iv) and the boundary condition (3.2) at $x = 0$,

$$\ell(u - g_L) = au'(0), \tag{3.13}$$

so we need to calculate $u'(0)$. Putting $w(x) = a[e^{-bx/a}u(x)]'$ in (3.1) gives

$$au'(0) - bg_L = w(0) = w(x)e^{bx/a} + \int_0^x f(y)dy \tag{3.14}$$

and the boundary condition (3.2) at $x = 1$ leads to

$$[\gamma + (1 - \gamma)b/a][g_L + \int_0^1 a^{-1}wdx] + (1 - \gamma)a^{-1}w(1) = e^{-b/a}\gamma g_R. \tag{3.15}$$

From (3.14) we obtain

$$b\int_0^1 a^{-1}wdx = (au'(0) - bg_L)(1 - e^{-b/a}) - \int_0^1 e^{-bx/a}fdx + e^{-b/a}\int_0^1 fdx, \tag{3.16}$$

and also that

$$w(1) = e^{-b/a}[au'(0) - bg_L - \int_0^1 fdx] = O(e^{-b/a}). \tag{3.17}$$

It follows from (3.15) that $g_L + \int_0^1 a^{-1}wdx = O(e^{-b/a})$ and thence from (3.16) that

$$au'(0) = \int_0^1 e^{-bx/a}fdx + O(e^{-b/a}),$$

giving (3.11): (3.12) follows directly. ■

A posteriori error bounds.

The Petrov-Galerkin method is needed to overcome the deficiencies of a Galerkin procedure only when the mesh Péclet number $\beta = bh/a$ becomes fairly large: for example, the linear Galerkin approximation starts to

oscillate for $\beta > 2$. The estimate (3.11) for $\ell(u - g_L)$ is then adequate for incorporation in an a posteriori error bound based on (2.28). From (3.4) and the bounds given in the proof of Lemma 3.2, we have

$$0 \le bV_j^* < \frac{3}{2}, \qquad j = 1, 2, \ldots, J. \tag{3.18}$$

Also, if U_j^N are the nodal parameters of $U^N(x)$, we have from (2.23)

$$\ell(U^N - g_L) = \frac{a}{h}(U_1^N - g_L). \tag{3.19}$$

The following a posteriori result then follows immediately from substituting (3.12), (3.18) and (3.19) into (2.28):

Theorem 3 Under Assumptions 3, the differences between the nodal parameters for the Petrov-Galerkin approximation U^N and those for the optimal approximation U^* satisfy

$$|U_j^N - U_j^*| \le \frac{3}{2}\beta^{-1}|U_1^N - g_L - hb^{-1}f(0)| + \frac{3}{2}b^{-1}\mathcal{E}_h^{\beta}[f] + O(e^{-b/a}), \tag{3.20}$$

$$j = 1, 2, \ldots, J$$

where

$$\mathcal{E}_h^{\mu}[f] = \int_0^1 e^{-\mu x/h}|f(x) - f(0)|dx. \tag{3.21}$$

In a typical situation where there is a sharp boundary layer near $x = 1$ and the differential equation is well approximated near $x = 0$ this bound will be quite small. In particular, when f is a constant the second term in (3.20) drops out and later in this section the first term will be shown to be $O(r^J)$. More explicit bounds can be obtained if it is assumed that $f \in S^h$: this is not unreasonable on the assumption that S^h can as well approximate $f(x)$ as $u(x)$.

Theorem 4 Suppose $f(x) \in S^h$, with nodal values f_j, and Assumptions 3 are satisfied. Then we have for $j = 1, 2, \ldots, J$

$$|U_j^N - U_j^*| < \frac{3}{2}\beta^{-1}|(U_1^N - g_L) - (hb^{-1})\sum_{j=0}^{J}\eta_j f_j| + O(e^{-b/a}), \qquad (3.22)$$

where

$$\eta_0 = 1 + \beta^{-1}(e^{-\beta} - 1),$$

and $$\eta_j = 4\beta^{-1}e^{-j\beta}\sinh^2 \tfrac{1}{2}\beta, \qquad j = 1, 2, \ldots, J.$$

Proof: In view of (3.18) and (3.19), it is necessary only to evaluate the integrals in (3.11):

$$\int_0^1 e^{-\beta(t+j)}[f_j + t(f_{j+1} - f_j)]dt = \beta^{-2}[e^{-j\beta}(e^{-\beta} - 1 + \beta)f_j + e^{-(j+1)\beta}(e^{\beta} - 1 - \beta)f_{j+1}].$$

The difference from η_J in the coefficient of f_J can be absorbed in the term $O(e^{-b/a})$ and we note that $\sum_{j=0}^{J}\eta_j = 1 + O(e^{-b/a})$. ∎

A priori error bounds.

All a priori error bounds for U^N require local estimations of U^N or U^* in order to calculate $\ell(U^N)$ or $\ell(U^*)$. This requires estimation of the discrete Green's function corresponding to the operator in (3.3). A calculation similar to that yielding the expression in (3.4) for V^* gives the following:

Lemma 3.4 Under Assumptions 3, the solution $G^m \in S_0^h$ of

$$B_S(G^m, \phi_j) = \langle\delta(x - mh), \phi_j\rangle, \qquad \forall\, \phi_j \in S_0^h, \qquad (3.23)$$

is given by

$$G_j^m = \frac{1}{b^2h}\left[\frac{1 - r}{1 + r}\right]\frac{1}{1 + (-1)^{\gamma}r^{2J}}\begin{cases}(r^m + (-1)^{\gamma}r^{2J-m})(r^{-j} - r^j), & j \le m\\ (r^{-m} - r^m)(r^j + (-1)^{\gamma}r^{2J-j}), & m \le j,\end{cases} \qquad (3.24)$$

where r is as defined in Lemma 3.1.

Then the error bound which is simplest to obtain is in terms of the solution u and its derivatives.

Theorem 5 Suppose $|u''(x)| \le M_{j-\frac{1}{2}}$ for $(j-1)h \le x \le jh$ with $M = \max\{M_{j-\frac{1}{2}},\ j = 1, 2, \ldots, J\}$ and $\overline{M}^2 = \sum_1^J h M_{j-\frac{1}{2}}^2$. Then under Assumptions 3, we have

$$\|u - U^N\|_S^2 \le \|e^A\|_S^2 + [\ell(e^A)]^2, \tag{3.25}$$

where

$$\|e^A\|_S \equiv \|u - U^*\|_S \le (1/\pi)[a^2 + (bh/\pi)^2]^{\frac{1}{2}}\overline{M}h, \tag{3.26}$$

and

$$|\ell(e^A)| \le (ah/4)[2M_{\frac{1}{2}} + 3(\sqrt{3} - 1)M]. \tag{3.27}$$

Proof: Substituting the uniform bound for $1 + \ell(V^*)$ given by (3.7) with $\lambda = 0$ into (2.29) and recalling that $K_\rho = 1$ in the constant coefficient case yields (3.25). We then bound both $\ell(e^A)$ and $\|e^A\|_S$ by introducing the piecewise linear interpolate U^I of u. We note first that

$$B_S(u - U^I, \phi_j) = -a^2h^{-1}\delta^2[u(jh) - U_j^I] + b^2\int_0^1 (u - U^I)\phi_j dx$$

$$= b^2\int_0^1 (u - U^I)\phi_j dx,$$

where δ^2 is the usual central difference operator. So

$$|B_S(u - U^I, \phi_j)| \le b^2h^3M/8. \tag{3.28}$$

We also have

$$\ell(e^A) = a(u - U^I)'(0) + \ell(U^I - U^*),$$

and hence

$$|\ell(e^A)| \le (ah/2)M_{\frac{1}{2}} + (a/h)|U_1^* - U_1^I|. \tag{3.29}$$

Now

$$U_1^* - U_1^I = \langle \delta(x - h),\ U^* - U^I \rangle = B_S(G^1,\ U^* - U^I) = B_S(G^1,\ u - U^I)$$

$$= \sum_{j=0}^{J} G_j^1 B_S(u - U^I,\ \phi_j),$$

and substituting for G_j^1 from (3.24) and using the bound (3.28), we have

$$|U_1^* - U_1^I| \le \frac{1}{8}h^2 M(1 - r)^2 \sum_{j=1}^{J} \left[\frac{1 + (-1)^\gamma r^{2(J-j)}}{1 + (-1)^\gamma r^{2J}}\right] |r|^{j-1}$$

$$\le \frac{1}{8}h^2(2 - \gamma)M(1 - r)^2/(1 - |r|),$$

and using bounds similar to those in Lemma 3.2, we obtain

$$|U_1^* - U_1^I| \le \frac{3}{4}(\sqrt{3} - 1)h^2 M. \tag{3.30}$$

Therefore (3.29) and (3.30) completes the estimation of $\ell(e^A)$ given by (3.27). Finally, by bounding $\|e^A\|_S$ by $\|u - U^I\|_S$, (3.26) is easily established using simple approximation theory (see Strang & Fix, 1973). ■

It should be noted that, since we are considering problems in which bh/a is fairly large, the norm $\|\cdot\|_S$ is most useful if preliminary scaling sets $b = 1$. Then $a = O(h)$ and all the bounds in (3.25), (3.26) and (3.27) are $O(h^2)$. The bound (3.30) is pessimistic since $M_{j-\frac{1}{2}}$ will usually be largest near $x = 1$. To improve it we need to consider estimating U_1^N in terms of the forcing function f.

Lemma 3.5 Under Assumptions 3, we have

$$\ell(U^N - g_L) = [1 + \ell(V^*)]^{-1}(a/h)\langle f, N_\varepsilon G^1 \rangle + O(\gamma r^J), \tag{3.31}$$

$$= (a/b)f(0) + E_h^\beta[f] + O(r^J), \tag{3.32}$$

where

$$E_h^{\beta}[f] = [1 + \ell(V^*)]^{-1}(a/h) \quad \langle f - f(0), N_\varepsilon G^1 \rangle. \tag{3.33}$$

Proof: From (2.18), (2.13) and the fact that $B_S(1, W) = -\langle \alpha, W \rangle \ell(1)$, $\forall\ W \in S_0^h$, we have

$$B_S(U^N - g_L, W) + \langle \alpha, W \rangle \ell(U^N - g_L) = \langle f, N_\varepsilon W \rangle \qquad \forall\ W \in S_0^h. \tag{3.34}$$

Firstly consider the case $\gamma = 0$: then $U^N - g_L \in S_0^h$ and applying (3.34) with $W \equiv G^1$ and using the definitions of V^* in (2.27) and G^1 in (3.23) yields

$$U_1^N - g_L = -V_1^* \ell(U^N - g_L) + \langle f, N_\varepsilon G^1 \rangle. \tag{3.35}$$

Multiplying both sides by (a/h) and rearranging, we obtain (3.31). In the case $\gamma = 1$, the Dirichlet case, $[U^N - g_L - x(g_R - g_L)] \in S_0^h$, so by subtracting $(g_R - g_L)B_S(x, W)$ from both sides of (3.34) and with the choice $W \equiv G^1$ we have in place of (3.35)

$$U_1^N - g_L - h(g_R - g_L) = -V_1^* \ell(U^N - g_L) + \langle f, N_\varepsilon G^1 \rangle - (g_R - g_L)B_S(x, G^1). \tag{3.36}$$

From the expression for G^1 in (3.24), we see that

$$B_S(x, G^1) = h + O(r^J), \tag{3.37}$$

since $x \notin S_0^h$; and hence (3.36) leads to (3.31).

Splitting $f(x)$ into $f(0) + [f(x) - f(0)]$ in expression (3.31), we obtain

$$\ell(U^N - g_L) = [1 + \ell(V^*)]^{-1}(a/h)\langle f(0), N_\varepsilon G^1 \rangle + E_h^{\beta}[f] + O(\gamma r^J).$$

Therefore to obtain the result (3.32), it is required to show that

$$\langle 1, N_\varepsilon G^1 \rangle = (h/b)[1 + \ell(V^*)] + O(r^J). \tag{3.38}$$

Now

$$\langle 1, N_\varepsilon G^1\rangle = \langle a + bx, G^1\rangle, \tag{3.39}$$

and from the definitions of α, V^* and G^1 in (2.16), (2.27) and (3.23) respectively, we have

$$\langle a, G^1\rangle = (a/b)\langle b, G^1\rangle = (a/b)V_1^*. \tag{3.40}$$

Also from (3.23) and (3.24) we obtain

$$\begin{aligned} B_S(x, G^1) &\equiv \langle a^2, G^{1'}\rangle + \langle b^2 x, G^1\rangle \\ &= \langle b^2 x, G^1\rangle + O(r^J). \end{aligned} \tag{3.41}$$

Therefore combining (3.37) and (3.41) yields

$$\langle bx, G^1\rangle = (h/b) + O(r^J), \tag{3.42}$$

and combining (3.39), (3.40) and (3.42) with the definition of $\ell(\cdot)$ in (2.23), we obtain (3.38). ■

Comparing Lemmas 3.3 and 3.5 when f is a constant, we obtain $\ell(u - U^N)' = O(e^{-b/a} + r^J)$ and hence $U^N - U^* = O(e^{-b/a} + r^J)$. For general f, we have the following a priori error bound.

Theorem 6 Under Assumptions 3, we have

$$\|u - U^N\|_S^2 \le \|u - U^*\|_S^2 + [\ell(e)]^2, \tag{3.43}$$

where

$$|\ell(e)| \equiv |\ell(u - U^N)| \le \sum_{j=1}^{J} \int_{(j-1)h}^{jh} w_j(x)|f(x) - f(0)|dx + O(e^{-b/a} + r^J), \tag{3.44}$$

with

$$w_j(x) = e^{-bx/a} + 3(\sqrt{3} - 1)\beta^{-1}|r|^{\max(j-2,0)}, \qquad j = 1, 2, \ldots, J,$$

and

$$|r| \le 2 - \sqrt{3} \approx 0.27 \quad \text{if} \quad \beta \ge \sqrt{3/2}.$$

Proof: Expression (3.43) follows directly from (2.33) and (2.28) with $\|V^*\|_S \le K_\rho \le 1$. Combining (3.12), (3.21) and (3.32) we then obtain

$$\ell(e) = \ell(u - g_L) - \ell(U_N - g_L) = \mathcal{E}_h^{\beta}[f] - E_h^{\beta}[f] + O(e^{-b/a} + r^J) . \quad (3.45)$$

In bounding $E_h^{\beta}[f]$ we require the following estimates:

from (3.4) and Assumption 3(iv), we have

$$1 + \ell(V^*) = 1 + \beta^{-1}(1 - r) + O(r^J), \quad (3.46)$$

and from (3.24)

$$G_j^1 = (1/b^2h)(1 - r)^2 r^{j-1} + O(r^J). \quad (3.47)$$

Therefore substituting for G_j^1 from (3.47) and $1 + \ell(V^*)$ from (3.46) into (3.33) we obtain

$$E_h^{\beta}[f] = \frac{\beta^{-2}(1 - r)^2}{[1 + \beta^{-1}(1 - r)]} \sum_{j=1}^{J-1} r^{j-1} \langle f - f(0), \phi_j + \int_x^1 (\beta/h)\phi_j dt \rangle + O(r^J). \quad (3.48)$$

It can be shown easily that

$$\left| \sum_{j=1}^{J}{}' r^{j-1}[\phi_j + \int_x^1 (\beta/h)\phi_j dt] \right| \le \left[\frac{1 - |r| + \beta}{1 - |r|}\right] |r|^{\max(j-2,0)} \quad (3.49)$$

$$\text{for } x \in [(j - 1)h, jh], \quad j = 1, 2, \ldots, J,$$

and applying (3.49) to (3.48) we obtain

$$|E_h^{\beta}[f]| \le \beta^{-1} \frac{(1 - r)^2}{1 - |r|}\left[\frac{1 - |r| + \beta}{1 - r + \beta}\right] \sum_{j=1}^{J} \int_{(j-1)h}^{jh} w_j^h |f(x) - f(0)| dx + O(r^J), \quad (3.50)$$

where

$$w_j^h = |r|^{\max(j-2,0)}, \quad j = 1, 2, \ldots, J.$$

Using bounds in (3.50) similar to those in Lemma 3.2 and incorporating the bound (3.21) on $\mathcal{E}_h^{\beta}[f]$ yields the desired result (3.44) from (3.45). ■

The results obtained above can be used to establish the uniform stability of the approximation. For ease of presentation, we only consider the case

of homogeneous boundary data, that is $U^N \in S_0^h$. Then from (2.44), (2.45) and since $K_\rho = 1$ we have

$$\|U^N\|_S \leq \|Z\|_S + \ell(U^N), \tag{3.51}$$

where $Z \in S_0^h$ satisfies $B_S(Z, \phi_j) = \langle f, N_\varepsilon \phi_j \rangle, \quad \forall\, \phi_j \in S_0^h$.

Bounding $\|Z\|_S$ is straightforward, since with $F(x) \equiv \int_0^x f(t)dt$

$$\|Z\|_S^2 = \langle f, N_\varepsilon Z \rangle$$

$$\leq \|F\|_2 \|Z\|_S, \tag{3.52}$$

where $\|\cdot\|_2$ denotes the standard L_2 norm. To bound $\ell(U_N)$, we use (3.31), (3.33), (3.44) and hence

$$|\ell(U^N)| \leq \sum_{j=1}^{J} \int_{(j-1)h}^{jh} w_j^h |f(x)| dx + O(r^J), \tag{3.53}$$

where

$$w_j^h = 3(\sqrt{3} - 1)\beta^{-1} |r|^{\max(j-2,0)} \qquad j = 1, 2, \ldots, J. \tag{3.54}$$

By preliminary scaling, setting $b = 1$, and combining (3.51), (3.52), (3.53) we establish the uniform stability of the approximation U^N as $b/a \to \infty$ in the L_2 norm; that is

$$\|U^N\|_2 \leq \|U^N\|_S \leq \|F\|_2 + \sum_{j=1}^{J} \int_{(j-1)h}^{jh} w_j^h |f(x)| dx + O(r^J), \tag{3.55}$$

where w_j^h is defined by (3.54).

4. Variable Coefficient Problems

We return in this section to considering the more general problems in which a and b are variable coefficients and ρ a variable weighting factor.

We consider first problems with no turning points, i.e. $b(x) \geq 0$ for $x \in [0, 1]$, and then those with a single turning point (and Dirichlet boundary conditions) such that $b(x) \leq 0$ for $x \leq \xi$ and $b(x) \geq 0$ for $x \geq \xi$. In each case only one perturbation function is needed, $V^*(x)$ is therefore a scalar function and the first step is to establish that $1 + \ell(V^*) > 0$, indeed we shall give conditions under which $\ell(V^*) \geq 0$; hence by Theorems 1 and 2 the Petrov-Galerkin approximations are uniquely defined. Then we shall give a posteriori error bounds of the same form as in Theorem 3, though there will be some loss of precision through taking account of the variability of a, b and ρ. For a priori bounds, we shall confine ourselves to relating them to bounds for an associated, self-adjoint problem.

Assumptions 1, 2 are supposed to hold and we strengthen them with the following:

Assumptions 4

In addition to Assumptions 1, 2 we suppose that:

(i) S^h is the space of piecewise linear elements on a uniform mesh;

(ii) $\exists\ \xi \in [0, 1)$ such that $b(x) \leq 0$ for $0 \leq x \leq \xi$, $b(x) \geq 0$ for $\xi \leq x \leq 1$ and we take

$$\epsilon(x) \equiv \delta(x - \xi) \quad \text{giving} \quad \ell(v) = a(\xi)v'(\xi) - b(\xi)v(\xi); \tag{4.1}$$

(iii) $b(x_j) \neq 0, \quad \forall\ j,$ and

$$b(x_j)\alpha_j \equiv b(x_j) \int_0^1 [\rho b + (\rho a)']\phi_j dx > 0, \qquad \forall\ \phi_j \in S_0^h \tag{4.2}$$

To satisfy Assumptions 4 it is sufficient to choose $\rho(x)$ such that both $\rho b^2 + (\rho ab)' > 0$ and $b[\rho b + (\rho a)'] > 0$ except where $b(x) = 0$. This can be achieved for example by taking $(\rho a)' = 0$ where $b' \geq 0$ and $(\rho ab)' = 0$ where $b' \leq 0$. In addition we need certain smoothness

assumptions on a, b and ρ, or implicitly an upper bound on the mesh size h, which are best given in hypotheses needed to establish bounds on V^*.

Properties of V^*.

As in Lemma 3.1, V^* is the Galerkin approximation in S^h to the solution $v(x)$ of a self-adjoint problem for which we have the following lemmas, firstly when there is no turning point in the original problem. We identify the coefficient p with ρa^2, q with $\rho b^2 + (\rho ab)'$ and s with $\alpha \equiv \rho b + (\rho a)'$ and recall that $\ell(V^*) = a(0)V^{*\prime}(0) = a(0)V_1^*/h$ in this case. Therefore $1 + \ell(V^*) > 0$ if $V_1^* \geq 0$: clearly this is true for sufficiently small h when the system of equations for V^* satisfies a minimum principle, by assumption (4.2). However for a general choice of h, $V_1^* \geq 0$ still holds under mild smoothness conditions on p, q and s as given in the following lemma.

Lemma 4.1 Consider the following two-point boundary value problem on $[0, 1]$ for the function v and its Galerkin approximation $V^* \in S_0^h$:

$$-(pv')' + qv = s, \quad v(0) = 0, \quad \gamma v(1) + (1 - \gamma)v'(1) = 0; \tag{4.3}$$

and

$$A_{j-\frac{1}{2}}V_{j-1}^* + B_jV_j^* + A_{j+\frac{1}{2}}V_{j+1}^* = S_j, \qquad \forall\, \phi_j \in S_0^h, \tag{4.4}$$

where

$$A_{j-\frac{1}{2}} = \langle p\phi_{j-1}', \phi_j' \rangle + \langle q\phi_{j-1}, \phi_j \rangle, \quad B_j = \langle p\phi_j', \phi_j' \rangle + \langle q\phi_j, \phi_j \rangle$$
$$S_j = \langle s, \phi_j \rangle.$$

For each j, define r_j as the smaller root in absolute value of

$$A_{j+\frac{1}{2}}r_j^2 + B_jr_j + A_{j-\frac{1}{2}} = 0. \tag{4.5}$$

Suppose that:

(i) $p \in H^1[0, 1]$ is a positive function and $q, s \in L_2[0, 1]$ are

non-negative functions such that $S_j > 0 \quad \forall j$, and

$$\langle q\phi_j, \phi_j - \phi_{j-1} - \phi_{j+1} \rangle \geq 0, \quad \langle q\phi_j, \phi_j + \phi_{j-1} + \phi_{j+1} \rangle > 0 \quad \forall j; \tag{4.6}$$

(ii) there are real numbers r_-, r_+, σ in (0, 1), with $r = \max(r_-, r_+)$ such that

$$-r_- \leq r_j \leq r_+, \qquad \forall j \tag{4.7}$$

$$|A_{j-\frac{1}{2}}| \leq (B_j - r|A_{j+\frac{1}{2}}|)\min(\sigma r_-, r_+) \quad \text{if} \quad A_{j-\frac{1}{2}}A_{j+\frac{1}{2}} < 0, \tag{4.8}$$

$$-r_- \leq -A_{j-\frac{1}{2}}/B_j \leq r_+ \quad \text{if} \quad \gamma = 0 \tag{4.9}$$

and

$$S_j \geq r_- \max[S_{j+1}, \ \sigma \sum_{m=j+1}^{J} r_+^{m-j-1} S_m] \quad \text{if} \quad A_{j+\frac{1}{2}} > 0. \tag{4.10}$$

Then we have

$$V_1^* \geq 0 \tag{4.11}$$

and

$$-r_- V_+ \leq V_j^* \leq V_+ \equiv (1 - r_-)^{-1} \max_{\phi_i \in S_0^h} \frac{\langle s, \phi_i \rangle}{\langle q\phi_i, \phi_i \rangle}, \qquad \forall j. \tag{4.12}$$

Proof: From hypothesis (i), (4.3) is a well-posed problem with a non-negative solution v. Suppose we solve the discrete equations (4.4) by direct LU-decomposition. Formally we have

$$V_j^* = E_j V_{j-1}^* + F_j, \qquad V_0^* = 0, \tag{4.13}$$

where

$$E_j = \frac{-A_{j-\frac{1}{2}}}{B_j + A_{j+\frac{1}{2}}E_{j+1}}, \qquad F_j = \frac{S_j - A_{j+\frac{1}{2}}F_{j+1}}{B_j + A_{j+\frac{1}{2}}E_{j+1}}$$

and the boundary conditions at $x = 1$ give

$\gamma = 1 : E_J = F_J = 0$

$\gamma = 0 : E_J = -A_{J-\frac{1}{2}}/B_J,\ F_J = S_J/B_J.$

We show first, by induction, that

$$-1 < -r_- < E_j < r_+ < 1, \qquad \forall\ j. \tag{4.14}$$

By hypothesis (4.9) this is satisfied at $j = J$ and the induction step falls into two cases: $E_jE_{j+1} > 0$ and $E_jE_{j+1} < 0$. Hypothesis (4.6) together with $p > 0$ clearly implies diagonal dominance of the system (4.4), i.e.

$$B_j > |A_{j-\frac{1}{2}}| + |A_{j+\frac{1}{2}}| \tag{4.15}$$

and thus ensures that $|E_j| < 1$ and also that E_j has the opposite sign to $A_{j-\frac{1}{2}}$. Thus the first case corresponds to $A_{j-\frac{1}{2}}A_{j+\frac{1}{2}} > 0$ and the recurrence for E_j can be regarded as one step in an iteration which we see from Figure 1(a) is monotonically convergent to the root r_j of (4.5) from any point on the same side of the other root, R_j say. However, hypothesis (4.6) ensures that $|R_j| > 1$ and hence $|E_{j+1}| < |R_j|$. Specifically, when $A_{j-\frac{1}{2}},\ A_{j+\frac{1}{2}} < 0$ and so $E_{j+1},\ r_j,\ E_j > 0$, (4.15) becomes $A_{j+\frac{1}{2}} + B_j + A_{j-\frac{1}{2}} > 0$, so that $r_j < 1 < R_j$, and

$$0 < r_j,\ E_{j+1} < r_+ < 1 \Rightarrow (E_{j+1} - r_j)/(E_j - r_j) > 1$$

$$\Rightarrow 0 < E_j < r_+.$$

The hypotheses (4.7), (4.9) ensure that the induction holds and a similar argument obtains when all these quantities have opposite signs. In the other case we have hypothesis (4.8) which is based on an assumption of sufficient smoothness on the part of p and q to ensure that when $A_{j-\frac{1}{2}}$ and $A_{j+\frac{1}{2}}$ have opposite signs at least one of them is small relative

to B_j. From this hypothesis,

$$|E_j| \le \frac{|A_{j-\frac{1}{2}}|}{B_j - r|A_{j+\frac{1}{2}}|} \le \min(\sigma r_-, r_+), \quad \text{when} \quad E_j E_{j+1} < 0, \tag{4.16}$$

see Figure 1(b). The induction is now complete.

Turning now to the recurrence for F_j we write

$$G_j = (B_j + A_{j+\frac{1}{2}} E_{j+1}) F_j \tag{4.17}$$

to obtain

$$G_j = S_j + E_{j+1} G_{j+1}, \quad G_J = (1 - \gamma) S_J \tag{4.18}$$

from which we establish by induction that

$$0 \le G_j \le \sum_{m=j}^{J} r_+^{m-j} S_m, \qquad \forall\, j. \tag{4.19}$$

The upper bound follows from the lower together with the bound (4.14) on E_j and (4.18), and the induction on the lower bound for G_j can break down only if $E_{j+1} < 0$. If $E_{j+2} > 0$, then bound (4.16) applies to E_{j+1} and hence $G_j \ge 0$ by hypothesis (4.10): on the other hand if $E_{j+2} < 0$, we have

$$\begin{aligned} G_j = S_j + E_{j+1} S_{j+1} + E_{j+1} E_{j+2} G_{j+2} &\ge S_j + E_{j+1} S_{j+1} \\ &\ge S_j - r_- S_{j+1} \ge 0, \end{aligned} \tag{4.20}$$

by hypothesis (4.10). Hence (4.19) is established and thence it follows that

$$F_j \ge 0, \qquad \forall\, j. \tag{4.21}$$

Since $V_1^* = F_1$, the first result (4.11) has been established. Now suppose that V_+ and $-V_-$ are respectively the largest and smallest values attained by V_j^*. From (4.13) and the fact that $F_j \ge 0$ it follows that $V_{j-1}^* < 0 \Rightarrow V_j^* > r_+ V_{j-1}^* > V_{j-1}^*$ and that $V_{j-1}^* > 0 \Rightarrow V_j^* > -r_- V_{j-1}^* \ge -r_- V_+$. Hence

$$V_- = \max_j(-V_j^*) \le r_- \max_j V_j^* = r_- V_+. \tag{4.22}$$

If now $V_m^* = V_+$, we have from (4.6), (4.22) and the fact that $A_{j-\frac{1}{2}} \le \langle q\phi_j, \phi_{j-1}\rangle$

$$\begin{aligned} S_m &= B_m V_+ + A_{m-\frac{1}{2}} V_{m-1}^* + A_{m+\frac{1}{2}} V_{m+1}^* \\ &\ge V_+\{B_m + A_{m-\frac{1}{2}} + A_{m+\frac{1}{2}} - (1 + r_-)[\max(A_{m-\frac{1}{2}}, 0) + \max(A_{m+\frac{1}{2}}, 0)]\} \\ &= V_+\{\langle q, \phi_m\rangle - (1 + r_-)[\max(A_{m-\frac{1}{2}}, 0) + \max(A_{m+\frac{1}{2}}, 0)]\} \qquad (4.23) \\ &\ge V_+\{\langle q, \phi_m\rangle - (1 + r_-)\langle q\phi_m, \phi_{m-1} + \phi_{m+1}\rangle\} \\ &= V_+\langle q\phi_m, \phi_m - r_-(\phi_{m-1} + \phi_{m+1})\rangle \\ &\ge V_+(1 - r_-)\langle q\phi_m, \phi_m\rangle. \end{aligned}$$

This gives the final result but it should be noted that the intermediate inequality (4.23) will be considerably sharper unless, as is implied by the problems considered in this paper, we have $qh^2 \ge p$. ■

In applying this result it should be noted that (4.6) implies a smoothness constraint on q, (4.8) a constraint on p and q, and (4.10) a constraint on s. These could all be expressed in terms of smoothness conditions on a, b and ρ. The parameters r_j correspond to the parameter r in the constant coefficient case which ranged from $r = 1$ when $b/a \to 0$ to $r = -2 + \sqrt{3}$ when $b/a \to \infty$. Thus the constraints on r_+ and r_- needed to satisfy (4.7)-(4.10) are rather mild.

Very similar results hold when there is a single turning point as allowed under Assumption 4. In this case we have

$$\ell(V^*) = a(\xi)V^{*\prime}(\xi) = a(\xi)(V_{k+1}^* - V_k^*)/h, \tag{4.24}$$

where $x_k < \xi < x_{k+1}$. Therefore, $1 + \ell(V^*) > 0$ if $V_{k+1}^* - V_k^* \ge 0$, and conditions under which this holds are given in the following lemma.

Lemma 4.2 Suppose the hypotheses for the two-point boundary value problem

of Lemma 4.1 with Dirichlet boundary conditions are modified as follows:

(i) $s(x) \geq 0$ for $x \geq \zeta$, $s(x) \leq 0$ for $x \leq \zeta$,

where $0 < x_k < \zeta < x_{k+1} < 1$, so that

$$S_j < 0 \quad \text{for} \quad j \leq k, \quad S_j > 0 \quad \text{for} \quad j \geq k+1; \tag{4.25}$$

and in addition to (4.6), for some $C > 1$ and $i = 0, 1$

$$\langle q\phi_{k-i}, \phi_{k-i} - c\phi_{k-i+1} - (2-c)\phi_{k-i-1} \rangle \geq 0$$

$$\langle q\phi_{k+i+1}, \phi_{k+i+1} - c\phi_{k+1} - (2-c)\phi_{k+i+2} \rangle \geq 0, \text{ for } 1 \leq c \leq C; \tag{4.26}$$

(ii) r_j is defined as the smaller root in absolute value of

$$A_{j\mp\frac{1}{2}} r_j^2 + B_j r_j + A_{j\pm\frac{1}{2}} = 0, \qquad \forall\, j, \tag{4.27}$$

with the upper sign being taken for $j \leq k$ and the lower for $j \geq k+1$, such that for r_-, r_+ and σ in $(0, 1)$ with $r_- \geq 2C^{-1} - 1$ and $r = \max(r_-, r_+)$, we have, with the same sign convention

$$-r_- \leq r_j \leq r_+, \qquad \forall\, j$$

$$|A_{j\pm\frac{1}{2}}| \leq (B_j - r|A_{j\mp\frac{1}{2}}|)\min(\sigma r_-, r_+) \quad \text{if} \quad A_{j-\frac{1}{2}} A_{j+\frac{1}{2}} < 0 \tag{4.28}$$

and

$$|S_j| \geq r_- \max\Big|S_{j\mp1}, \sigma \sum_{m=j\mp1}^{\frac{1}{2}(J\mp J)} r_+^{\mp(m-j)-1} S_m\Big| \quad \text{if} \quad A_{j\mp\frac{1}{2}} > 0.$$

Then we have

$$V_{k+1}^* - V_k^* \geq 0 \tag{4.29}$$

and

$$|V_j^*| \leq 2(1 - r_-)^{-1} \max_{\phi_i \in S_0^h} \frac{|\langle s, \phi_i \rangle|}{\langle q\phi_i, \phi_i \rangle}, \qquad \forall\, j. \tag{4.30}$$

Proof: We partition V^* formally into two parts, $V^* = V^- + V^+$,

where $\{V_j^-\} = \{V_1^*, \ldots, V_k^*, 0, \ldots, 0\}$ and $\{V_j^+\} = \{0, \ldots, 0, V_{k+1}^*, \ldots, V_{J-1}^*\}$. Then V^- and V^+ satisfy the partitioned equations given by (4.4) for V^*, but with the right-hand sides and boundary conditions changed as follows:

$$\begin{aligned} V^- &: \{S_j^-\} = \{S_1, \ldots, S_{k-1}, S_k - A_{k+\frac{1}{2}}V_{k+1}^+\}, \quad V_0^- = V_{k+1}^- = 0; \\ V^+ &: \{S_j^+\} = \{S_{k+1} - A_{k+\frac{1}{2}}V_k^-, S_{k+2}, \ldots, S_{J-1}\}, \quad V_k^+ = V_J^+ = 0. \end{aligned} \tag{4.31}$$

Suppose now we calculate E_j^+, F_j^+ and G_j^+ as in Lemma 4.1 starting from $E_J^+ = F_J^+ = G_J^+ = 0$. We obtain, with $B_j^+ = B_j + A_{j+\frac{1}{2}}E_{j+1}^+$,

$$F_{k+1}^+ = G_{k+1}^+/B_{k+1}^+ = (S_{k+1}^+ + E_{k+2}^+G_{k+2}^+)/B_{k+1}^+. \tag{4.32a}$$

Calculating E_j^-, F_j^- and G_j^- similarly, but starting from the left with $E_0^- = F_0^- = G_0^- = 0$ and with $B_j^- = B_j + A_{j-\frac{1}{2}}E_{j-1}^-$, we have

$$F_k^- = G_k^-/B_k^- = (S_k^- + E_{k-1}^-G_{k-1}^-)/B_k^-. \tag{4.32b}$$

Thence, since $V_{k+1}^* = V_{k+1}^+ = F_{k+1}^+$ and $V_k^* = V_k^- = F_k^-$, we obtain by substituting from (4.31) for S_{k+1}^+ and S_k^-

$$B_{k+1}^+V_{k+1}^* + A_{k+\frac{1}{2}}V_k^* = S_{k+1} + E_{k+2}^+G_{k+2}^+ \geq 0 \tag{4.33a}$$

$$A_{k+\frac{1}{2}}V_{k+1}^* + B_k^-V_k^* = S_k + E_{k-1}^-G_{k-1}^- \leq 0, \tag{4.33b}$$

the inequalities following from the induction argument on the $\{G_j\}$ as in Lemma 4.1. Multiplying (4.33a) by $B_k^- + A_{k+\frac{1}{2}}$ and (4.33b) by $B_{k+1}^+ + A_{k+\frac{1}{2}}$, the positivity of each factor being guaranteed by the diagonal dominance condition (4.6), and subtracting the two equations gives

$$(B_k^-B_{k+1}^+ - A_{k+\frac{1}{2}}^2)(V_{k+1}^* - V_k^*) \geq 0. \tag{4.34}$$

Moreover, $B_{k+1}^+ > |A_{k+\frac{1}{2}}|$ and $B_k^- > |A_{k+\frac{1}{2}}|$, so the multiplying factor in (4.34) is strictly positive and the main result (4.29) follows. This factor is just the determinant of the pair of equations (4.33a, b) which

are therefore non-singular so that V_k^*, V_{k+1}^* can be solved for and the equivalence of (4.31) to the equations for V^* established.

Suppose now that $-V_-^{\pm} \le V_j^{\pm} \le V_+^{\pm}$ and $V_+^+ \ge V_-^-$ so we consider $j \ge k+1$. If the lower limit $-V_-^+$ and the upper limit V_+^+ are attained at other than $j = k+1$, exactly the same argument holds as in Lemma 4.1 to show that $V_-^+ \le r_- V_+^+$ and hence the bound (4.12) obtained for V_+^+. If $V_{k+1}^+ = V_+^+ \ge V_-^-$, we can bound V_+^+ from the equation centred about $k+1$: using (4.29) to get $-V_+^+ \le V_k^* \le V_+^+$, the fact that $-r_- V_+^+ \le V_{k+2}^+ \le V_+^+$ and (4.26) we have

$$
\begin{aligned}
S_{k+1} &= B_{k+1}V_+^+ + A_{k+3/2}V_{k+2}^+ + A_{k+\frac12}V_k^* \\
&\ge V_+^+\{(B_{k+1} + A_{k+3/2} + A_{k+\frac12}) - (1 + r_-)\max(A_{k+3/2}, 0) - 2\max(A_{k+\frac12}, 0)\} \\
&\ge V_+^+\{\langle q, \phi_{k+1}\rangle - (1 + r_-)\langle q\phi_{k+1}, \phi_{k+2}\rangle - 2\langle q\phi_{k+1}, \phi_k\rangle\} \\
&= V_+^+\langle q\phi_{k+1}, \phi_{k+1} - r_-\phi_{k+2} - \phi_k\rangle \\
&\ge \tfrac12 V_+^+(1 - r_-)\langle q\phi_{k+1}, \phi_{k+1}\rangle.
\end{aligned}
$$

If $V_{k+1}^+ = -V_-^+$ then we still have $V_j^+ \ge -r_- V_+^+$ for $j \ge k+2$ but to get a similar bound when $V_{k+2}^+ = V_+^+$ we need (4.26) with $i = 1$ and $V_{k+1}^+ \ge V_k^- \ge -V_-^- \ge -V_+^+$. This establishes (4.30) under all circumstances. ■

It should be noted that the condition (4.26) requires a slightly stronger smoothness constraint on q about the turning point than (4.6).

At this stage it is worth noting that the bounding of $\ell(V^*)$ has reduced the problem of finding a priori bounds on $u - U^N$ to an approximation problem for u: and, moreover, estimates for u may be derived by considering it as the solution to an associated self-adjoint problem. Thus from Theorem 2 and Lemmas 4.1, 4.2, we have

$$
\|u - U^N\|_S^2 \le \|u - U^*\|_S^2 + K_\rho |\ell(u - U^*)|^2, \qquad (4.35)
$$

where U^* is the best fit to u in the $\|\cdot\|_S$ norm. That is, U^* is the Galerkin approximation to u which, with the above definitions of p,

q and s, satisfies

$$-(pu')' + qu = t, \tag{4.36}$$

where t is given by setting

$$\langle t, w\rangle = B_S(u, w) = B(u, N_\varepsilon w) - \langle s, w\rangle \ell(u)$$
$$= \langle f, N_\varepsilon w\rangle - \langle s, w\rangle \ell(u), \qquad \forall\, w \in H_0^1. \tag{4.37}$$

In particular, if $f \equiv 0$ and once $\ell(u)$ has been estimated (see below) then this problem reduces to that considered in the two lemmas above.

A Posteriori Error Estimates.

As in the constant coefficient case, the solution $u(x)$ of (1.1) and (1.2) can be expressed explicitly in terms of integrals of f, leading to the following generalisation of Lemma 3.3.

Lemma 4.3 Under Assumptions 4,

$$\ell(u \div g_L) = H\left(\int_0^x [f(t) - b'(t)g_L]dt\right)/H(1) + O(e^{-\lambda(1)}), \quad \text{if } \xi = 0; \tag{4.38a}$$

$$\ell(u) = H\left(\int_\xi^x f(t)dt\right)/H(1) + O(e^{-\lambda(0)}, e^{-\lambda(1)}), \quad \text{if } \xi \neq 0; \tag{4.38b}$$

where the functional $H(v) \equiv H(v(x))$ is given by

$$H(v) = \int_0^1 e^{-\lambda(x)}[v(x)/a(x)]dx, \tag{4.39}$$

and

$$\lambda(x) = \int_\xi^x [b(t)/a(t)]dt. \tag{4.40}$$

Proof: From the definition of $\lambda(x)$ in (4.40), setting $w(x) = a(x)[e^{-\lambda(x)}u(x)]'$ in equation (1.1) implies that

$$a(\xi)u'(\xi) - b(\xi)u(\xi) = w(\xi) = w(x)e^{\lambda(x)} + \int_\xi^x f(y)dy. \tag{4.41}$$

We also have

$$\int_0^1 [w(x)/a(x)]dx = e^{-\lambda(1)}u(1) - e^{-\lambda(0)}u(0), \qquad (4.42)$$

and combining this with (4.41), we obtain

$$\ell(u) = w(0) = \left[\int_0^1 e^{-\lambda(x)}/a(x)dx\right]^{-1}\left[\int_0^1 e^{-\lambda(x)}\{\int_\xi^x f(y)dy\}/a(x)dx - g_L\right]$$

$$+ O(e^{-\lambda(1)}), \quad \text{if} \quad \xi = 0; \qquad (4.43a)$$

$$\ell(u) = w(\xi) = \left[\int_0^1 e^{-\lambda(x)}/a(x)dx\right]^{-1}\left[\int_0^1 e^{-\lambda(x)}\{\int_\xi^x f(y)dy\}/a(x)dx\right]$$

$$+ O(e^{-\lambda(1)}, e^{-\lambda(0)}), \quad \text{if} \quad \xi \neq 0. \qquad (4.43b)$$

Therefore incorporating the definition of $H(v)$ from (4.39) with (4.43a, b) and noting that $H(b) = 1 + O(e^{-\lambda(1)})$ if $\xi = 0$ we obtain the desired results (4.38a, b). ■

These a priori estimates on $\ell(u)$ lead directly to the following a posteriori error bounds. First of all consider the case of a non-turning point problem, that is $\xi = 0$; then we have the following generalisation of Theorem 3.

Theorem 7 Under Assumptions 4 with $\xi = 0$ and assuming the hypotheses of Lemma 4.1 hold for V^*, the difference between the nodal parameters for the Petrov-Galerkin approximation U^N and those for the optimal approximation U^* satisfy, if $b(0) \neq a'(0)$,

$$|U_j^N - U_j^*| \leq M\left\{\left|\frac{U_1^N - g_L}{h} - \left[\frac{f(0) - b'(0)g_L}{b(0) - a'(0)}\right]\right| + \mathcal{E}[f]\right\} + O(e^{-\lambda(1)}),$$

$$j = 1, 2, \ldots, J; \qquad (4.44)$$

where

$$M = a(0)(1 - r_-)^{-1} \max_{\phi_i \in S_0^h}\left[\frac{\langle \rho b + (\rho a)', \phi_i\rangle}{\langle(\rho b^2 + (\rho a b)')\phi_i, \phi_i\rangle}\right], \qquad (4.45)$$

with r_- defined as in Lemma 4.1, and

$$\mathcal{E}[f] = H(|F|)/H(1), \tag{4.46}$$

with

$$F(x) = \int_0^x \{[f(t) - b'(t)g_L] - \frac{[b(t)t - a(t)]'}{b(0) - a'(0)}[f(0) - b'(0)g_L]\}dt. \tag{4.47}$$

Proof: In the constant coefficient problem, we made use of the fact that a constant forcing function f results in a linear solution u to (3.1), with certain boundary data. The generalisation of this to the variable coefficient problem is to note that a multiple of $[b(x)x - a(x)]'$ as the forcing function produces a linear solution to (1.1). This leads to the following splitting of $f(t) - b'(t)g_L$ in (4.38a):

$$f(t) - b'(t)g_L = [b(t)t - a(t)]'\left[\frac{f(0) - b'(0)g_L}{b(0) - a'(0)}\right] + F'(t), \tag{4.48}$$

where $F(x)$ is as defined in (4.47), the multiple of $[b(t)t - a(t)]'$ having been chosen to set $F'(0) = 0$. We then have

$$H\left(\int_0^x [f(t) - b'(t)g_L]dt\right) = [H(b(x)x - a(x)) + H(a(0))]\left[\frac{f(0) - b'(0)g_L}{b(0) - a'(0)}\right] + H(F). \tag{4.49}$$

Now from the definition of $H(v)$ in (4.39), we see that

$$H(b(x)x - a(x)) = O(e^{-\lambda(1)}), \tag{4.50}$$

and therefore (4.43a), (4.49) and (4.50) imply that

$$\ell(u - g_L) = \left[\frac{a(0)}{b(0) - a'(0)}\right][f(0) - b'(0)g_L] + H(F) + O(e^{-\lambda(1)}). \tag{4.51}$$

Combining the expression for $\ell(U^N - g_L)$ with (4.51) and using the bound on V_j^* from Lemma 4.1 in (2.28), the desired result (4.44) is obtained. ■

The bound (4.44) is a direct generalisation of the constant coefficient case (3.20). Note that the first term in (4.44) is a forward difference approximation to the differential equation (1.1), ignoring the second derivative term at the point $x = 0$ and this will be quite small in a typical situation where there is a boundary layer at $x = 1$. The second term will be negligible in many circumstances since it is an integral of a function with an exponentially decaying weight away from the origin, and the function with its first derivative are zero at the origin. The corresponding a posteriori error bound for the turning point problem is given in the following theorem.

Theorem 8 Under Assumptions 4 with $x_k < \xi < x_{k+1}$, and with Lemma 4.2 holding for V^* the difference between the nodal parameters for the Petrov-Galerkin approximation U^N and those for the optimal approximation U^* satisfy

$$|U_j^{N\prime} - U_j^{*\prime}| \le M\left\{\left|\frac{U_{k+1}^N - U_k^N}{h} - \left[\frac{b'(\xi)f'(\xi) - b''(\xi)f(\xi)}{b'(\xi)\{2b'(\xi) - a''(\xi)\} + a'(\xi)b''(\xi)}\right]\right|\right.$$
$$\left. + \mathcal{E}[f]\right\} + O(e^{-\lambda(0)} + e^{-\lambda(1)}), j = 1, 2, \ldots, J - 1; \quad (4.52)$$

where

$$M = 2a(\xi)(1 - r_-)^{-1} \max_{\phi_i \in S_0^h}\left[\frac{|\langle \rho b + (\rho a)', \phi_i\rangle|}{\langle(\rho b^2 + (\rho a b)')\phi_i, \phi_i\rangle}\right], \quad (4.53)$$

with r_- defined as in Lemma 4.2 and

$$\mathcal{E}[f] = H(|F|)/H(1), \quad (4.54)$$

with

$$F(x) = \int_\xi^x \{f(t) - \frac{[b(t)(t - \mu) - a(t)]'}{b'(\xi)(\xi - \mu) - a'(\xi)} f(\xi)\}dt, \quad (4.55)$$

and

$$\mu - \xi = \{a'(\xi)f'(\xi) + [a''(\xi) - 2b'(\xi)]f(\xi)\}/\{b''(\xi)f(\xi) - b'(\xi)f'(\xi)\}. \quad (4.56)$$

Proof: When $\xi \neq 0$, we have $H(b(x)) = O(e^{-\lambda(0)} + e^{-\lambda(1)})$ as well as $H(b(x)x - a(x)) = O(e^{-\lambda(0)} + e^{-\lambda(1)})$, this leads to considering the following splitting of $f(t)$ in (4.38b):

$$f(t) = \frac{[b(t)(t - \mu) - a(t)]'}{b'(\xi)(\xi - \mu) - a'(\xi)} f(\xi) + F'(t), \tag{4.57}$$

where $F(x)$ is as defined in (4.55) with an arbitrary choice of μ; the constant as before has been chosen so that $F'(\xi) = 0$. We then have

$$H\left(\int_{\xi}^{x} f(t)dt\right) = [H(b(x)(x - \mu) - a(x)) - H(b(\xi)(\xi - \mu) - a(\xi))] \frac{f(\xi)}{b'(\xi)(\xi - \mu) - a'(\xi)} + H(F). \tag{4.58}$$

The properties of $H(v)$ stated above and the fact that $b(\xi) = 0$ imply that

$$H\left(\int_{\xi}^{x} f(t)dt\right)/H(1) = a(\xi)f(\xi)/[b'(\xi)(\xi - \mu) - a'(\xi)] + \mathcal{E}[f], \tag{4.59}$$

where $\mathcal{E}[f]$ is defined by (4.54). The expression (4.59) holds for all $\mu \neq \xi$, we fix this degree of freedom by requiring $F''(\xi) = 0$, this leads to the choice of μ given in (4.56). Substituting for μ into (4.59) yields an expression for $\ell(u)$, combining this with $\ell(U^N)$ and using the bound on V_j^* from Lemma 4.2 in (2.28), the desired result (4.52) is obtained. ■

The first term in (4.53) is the difference between a difference and an asymptotic representation of $u'(\xi)$, and this will be small in a typical situation as the differential equation will be well approximated near $x = \xi$, whereas boundary layers may occur at $x = 0$ and 1.

To conclude, we have shown how the perturbation function, which determines the effect of approximate symmetrization, can be chosen to guarantee

existence and uniqueness for the Petrov-Galerkin method and can also be placed where the underlying solution to the differential equation is not rapidly varying. This leads to $\ell(u - U^N)$ being small and thus a near optimal approximation being obtained.

REFERENCES

ALLEN, D. & SOUTHWELL, R. 1955 Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. Quart. J. Mech. and Appl. Math. VIII, 129-145.

BABUŠKA, I. & AZIZ, A.K. 1972 Survey lectures on the mathematical foundations of the finite element method, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (ed. A.K. Aziz), New York: Academic Press, 3-363.

BABUŠKA, I & RHEINBOLDT, W.C. 1978 Error estimates for adaptive finite element computations. SIAM J. Numer. Anal. 15, 736-754.

BARRETT, J.W. & MORTON, K.W. 1978 Optimal finite element solutions to diffusion-convection problems in one dimension. U. of Reading, Num. Anal. Report 3/78, to appear in Int. J. Num. Meth. Engng.

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

BRAMBLE, J.H. & SCHATZ, A.H. 1977 Higher order local accuracy by averaging in the finite element method. Math. Comp. 31. 94-111.

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. Engng. 10, 1389-1396.

DIXON, L.C.W., HARRISON, D. & MORGAN, J.V. 1979 On singular cases arising from Galerkin's method. Proc. Conf. on The Mathematics of Finite Elements & Applications III (ed. J.R. Whiteman), Academic Press, 217-226.

DOUGLAS, J., Jr. & DUPONT, T. 1973a Some superconvergence results for Galerkin methods for the approximate solution of two-point boundary problems. Proc. Conf. on Topics in Numerical Analysis (ed. J.J.H. Miller), Dublin: Royal Irish Academy, 89-92.

DOUGLAS, J., Jr. & DUPONT, T. 1973b Superconvergence for Galerkin methods for the two-point boundary problem via local projections. Numer. Math. 21, 270-278.

DUPONT, T. 1976 A unified theory of superconvergence for Galerkin methods for two-point boundary problems. SIAM J. Numer. Anal. 13, 362-368.

GUYMON, G.L., SCOTT, V.H. & HERRMANN, L.R. 1970 A general numerical solution of the two-dimensional diffusion-convection equation by the finite element method. Water Resources 6, 1611-1617.

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. Engng. 11, 131-143.

HEINRICH, J.C. & ZIENKIEWICZ, O.C. 1979 The finite element method and 'upwinding' techniques in the numerical solution of convection dominated flow problems. In Finite Element Methods for Convection Dominated Flows (ed. T.J.R. Hughes) AMD Vol. 34, Am. Soc. Mech. Eng. (ASME)

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

HUGHES, T.J.R. 1978 A simple scheme for developing 'upwind' finite elements. Int. J. Num. Meth. Engng. 12, 1359-1365.

IL'IN, A.M. 1969 Differencing scheme for a differential equation with a small parameter affecting the highest derivative. Math. Notes Acad. Sci. U.S.S.R. 6, 596-602.

LESAINT, P. & ZLÁMAL, M. 1979 Superconvergence of the gradient of finite element solutions. R.I.A.R.O. Analyse numérique 13 No. 2, 139-166.

MILLER, J.J.H. 1978 Sufficient conditions for the convergence, uniformly in ε, of a three-point difference scheme for a singular perturbation problem. In Numerical Treatment of Differential Equations in Applications (eds. R. Ansorge & W.Tornig) Lect. Notes in Maths. 679, Berlin: Springer, 85-91.

MOAN, T. 1979 On the nature of spatial finite element approximations in structural mechanics. Proc. Conf. on The Mathematics of Finite Elements and Applications III (ed. J.R. Whiteman), Academic Press, 391-414.

MORTON, K.W. & BARRETT, J.W. 1980 Optimal finite element methods for diffusion-convection problems. Proc. Conf. Boundary and Interior Layers - Computational and Asymptotic Methods (ed. J.J.H. Miller), Dublin: Boole Press, 134-148.

RHEINHARDT, H.J. 1980 A posteriori error estimates for the finite element solution of a singularly perturbed linear ordinary differential equation. To appear in SIAM J. Numer. Anal.

SCHATZ, A.H. 1974 An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. Math. Comp. 28, 959-962.

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

WHEELER, M.F. 1974 A Galerkin procedure for estimating the flux for two-point boundary value problems. SIAM J. Numer. Anal. 11, 764-768.

ZIENKIEWICZ, O.C. 1977 The Finite Element Method, 3rd edn. London: McGraw-Hill.

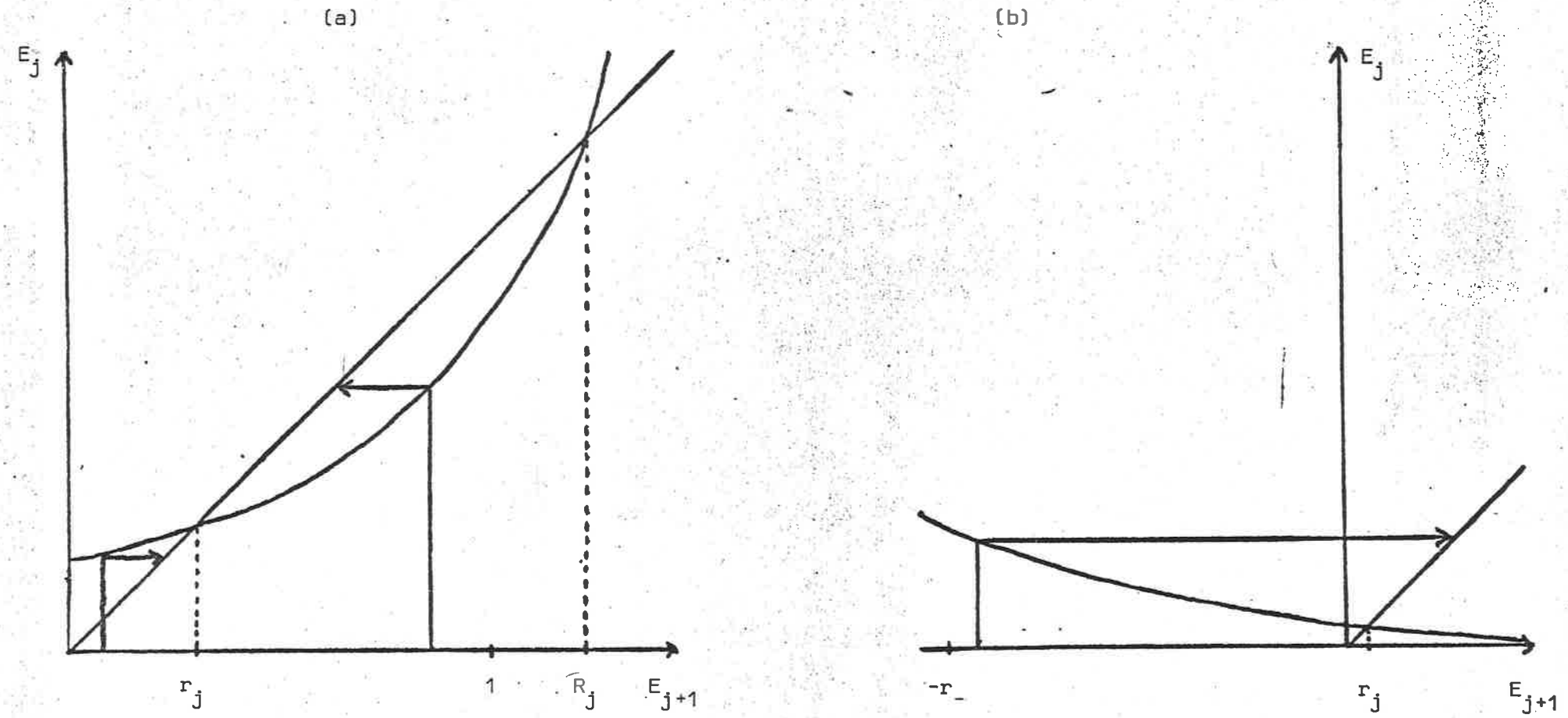

Figure 1: The iteration $E_{j+1} \to E_j$ when (a) $A_{j-\frac{1}{2}} \leq 0$, $A_{j+\frac{1}{2}} \leq 0$ and (b) $A_{j-\frac{1}{2}} \leq 0$, $A_{j+\frac{1}{2}} \geq 0$.