ELSEVIER

Journal of Computational and Applied Mathematics 85 (1997) 287–313

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

Iterative solution of elliptic problems by approximate factorization

Eldar Giladi[a,*,1], Joseph B. Keller[b]

[a] *Scientific Computing and Computational Mathematics Program, Stanford University, Stanford, CA 94305, USA*

[b] *Departments of Mathematics and Mechanical Engineering, Stanford University, Stanford, CA 94305, USA*

Accepted 2 June 1997

Abstract

An iterative method for the numerical solution of singularly perturbed second-order linear elliptic problems is presented. It is a defect correction iteration in which the approximate operator is the product of two first-order operators, which is readily inverted numerically. The approximate operator is generated by formal asymptotic factorization of the original operator. Hence this is a QUasi Analytic Defect correction iteration (QUAD). Both its continuous and discrete versions are analyzed in one dimension. The scheme is extended to a variety of two dimensional operators and it is analyzed for a model advection-diffusion equation. Numerical calculations show the effectiveness of the scheme over a wide range of values of the small parameter.

Keywords: Defect correction iteration; Asymptotic factorization; Preconditioners

AMS classification: 65F10; 65N22; 35J25

1. Introduction

We develop an iterative method for the numerical solution of singularly perturbed linear elliptic problems of second order. It was originally designed for the computation of smooth solutions to amplitude equations [6, 7]. Our goal was to devise a stable numerical scheme which incorporates as much analytic information as possible.

The basis of the method is a defect correction iteration [9, 21]. This is a general technique for solving iteratively an equation $Lu = f$ by means of an operator L which "approximates" L but is much easier to invert. In the present work, L is generated by a formal asymptotic factorization

[*] Corresponding author.

[1] Present address: Applied Mathematics, 217-50 Caltech, Pasadena, CA 91125, USA.

PII S0377-0427(97)00132-5

of the original operator into two first-order operators. The approximate operator is readily inverted numerically by solving a sequence of initial value problems. Hence the scheme can be classified as a QUasi Analytic Defect correction method (QUAD).

The algorithm combines the best of the asymptotic and the numerical approaches. On the one hand the use of asymptotic methods yields an approximate operator which is tailored to the problem. This yields a fast rate of convergence. On the other hand, the numerical implementation of the scheme eliminates the calculational burden associated with the derivation of a large number of terms in an asymptotic series. QUAD is a stable and convergent algorithm, in contrast to a standard asymptotic scheme which is unstable and ill-suited for numerical implementation, in general. Moreover, in the one-dimensional case the approximate operator can be inverted analytically and the scheme yields power series solutions to linear second-order singularly perturbed boundary value problems.

The present results are discussed in detail in [6]. Other papers that combine numerical and asymptotic methods include [3–5, 8, 15–20] and the collections [10, 11]. A third-order approximate factorization of the convection diffusion equation was presented in [14].

In Section 2 we motivate our work by presenting a one-dimensional model problem. Then in Section 3 we introduce the method in one dimension and we demonstrate it numerically on the model problem. The convergence analysis of both the continuous and the discrete iteration, in one dimension, is presented in Section 4. Variations on the themes of QUAD are presented in Section 5. In Section 6 we extend the scheme to a class of two-dimensional singularly perturbed elliptic problems. We analyze it for a model advection-diffusion equation and we present numerical calculations. We conclude with future directions in Section 7.

2. A model problem

We first introduce a one-dimensional model problem which arises in the hybrid numerical-asymptotic method for solving singularly perturbed equations in [6, 7]. We consider the propagation of a time harmonic wave of unit amplitude, traveling along the x axis to the right. Its motion is described by the one-dimensional Helmholtz equation

$$\varepsilon^2 u'' + n(x)^2 u = 0, \tag{1}$$

subject to the conditions

$$u(0) = 1, \quad u \text{ outgoing at } x = +\infty. \tag{2}$$

Here, $\varepsilon = 1/k$ and k is the wave number. The index of refraction, $n(x)$, is assumed positive in $[0, \infty)$.

The solution to (1) is highly oscillatory as $\varepsilon \to 0$ and requires a very refined mesh for its numerical solution. Here, we seek solutions for $u(x)$ in (1) of the form

$$u(x) = K(x) \mathrm{e}^{\mathrm{i}S(x)/\varepsilon}, \tag{3}$$

where $S(x)$ and $K(x)$ are called the phase and the amplitude, respectively. Both functions are slowly varying and they can be resolved numerically on a grid which is substantially coarser than the one required for u.

In order to obtain equations for K and S we introduce expression (3) for u into (1) and we divide the resulting expression by $\mathrm{e}^{\mathrm{i}S(x)/\varepsilon}/\varepsilon^2$ to obtain

$$K(n^2 - S_x^2) + \mathrm{i}\varepsilon(2K_xS_x + KS_{xx}) + \varepsilon^2 K_{xx} = 0. \tag{4}$$

Upon equating the coefficient of the leading order term in (4) to zero we obtain the eiconal equation of geometrical optics

$$S_x(x)^2 = n(x)^2. \tag{5}$$

Hence, $S(x) = \int_0^x n(s)\,\mathrm{d}s$, in view of the outgoing condition at ∞. We substitute this expression for $S(x)$ into (4) to obtain

$$\varepsilon K_{xx} + \mathrm{i}(2nK_x + n_x K) = 0. \tag{6}$$

Now we assume that $n(x)$ is constant for $x > 2 - \delta$ with δ small. Then in view of the representation (3) and the outgoing condition we require $K_x(2) = 0$.

Our model problem is Eq. (6) for K subject to the conditions

$$K(0) = 1 \quad \text{and} \quad K_x(2) = 0. \tag{7}$$

In ray theory, for $\varepsilon \ll 1$, its solution is approximated by an asymptotic series. Instead we develop a convergent iteration to solve it numerically.

In summary, our approach to solving (1) numerically has two steps. First, we represent the solution as the product of a highly oscillatory factor and a smooth amplitude K. This step reduces dramatically the numerical complexity of solving the original problem because the number of grid points required to resolve K and S accurately is much smaller than the number of points required to resolve the original problem. The savings increase as $\varepsilon \to 0$. The second step is to develop a specialized numerical algorithm for solving the equation for K. Here, we focus on this aspect and we develop QUAD. Later [7] we shall discuss the first step, which is usually more complicated than in this one-dimensional example.

3. The QUAD algorithm

Let the operator L be defined by

$$L = \varepsilon \frac{\mathrm{d}^2}{\mathrm{d}x^2} + a_1(x)\frac{\mathrm{d}}{\mathrm{d}x} + a_0(x), \tag{8}$$

where $0 < \varepsilon < 1$, $a_0(x)$, $a_1(x)$, are continuous functions in $[a,b]$, $\mathrm{Re}[a_1(x)] \leqslant 0$ and $a_1(x) \neq 0$ for $x \in [a,b]$. In this section we present the QUAD iteration for the problem

$$Lu = f, \tag{9}$$

$$u(a) = 1, \qquad \left[\alpha\left(\frac{\mathrm{d}}{\mathrm{d}x} - A\right) + (1-\alpha)\right] u(b) = \beta. \tag{10}$$

The case $\alpha = 0$ and $\alpha = 1$ in (10) corresponds to Dirichlet and robin boundary conditions at $x = b$, respectively. We assume that this problem has a unique solution.

3.1. The basic iteration

Suppose that there exists a sequence of functions $\{S_m(x,\varepsilon)\}_{m=1}^{\infty}$ such that the associated operators

$$L_m = \varepsilon\left(\frac{\mathrm{d}}{\mathrm{d}x} + \frac{a_1(x)}{\varepsilon} + S_m(x,\varepsilon)\right)\left(\frac{\mathrm{d}}{\mathrm{d}x} - S_m(x,\varepsilon)\right), \quad m = 1,2,\ldots \tag{11}$$

satisfy

$$L - L_m = \varepsilon^m r_m(x,\varepsilon), \tag{12}$$

and that

$$\lim_{\varepsilon\to 0} S_m(x,\varepsilon) = S_m(x,0), \tag{13}$$

$$\lim_{\varepsilon\to 0} r_m(x,\varepsilon) = r_m(x,0). \tag{14}$$

The operators L_m possess two important properties. First they approximate L for $\varepsilon \ll 1$, as indicated by (12) and (14). Second, they are in factored form. Therefore the differential equation

$$L_m d = r, \tag{15}$$

can be solved explicitly, as we shall see in Section 3.2. In view of these properties we develop a defect correction iteration [9] to solve the problem (9), (10). It computes an approximation of u to an arbitrary tolerance η.

Algorithm 1 (QUAD)

```
Compute the initial iterate $u_0$ by solving
  $L_m u_0 = f$
  $u_0(a) = 1, \quad [\alpha(\frac{\mathrm{d}}{\mathrm{d}x} - A) + (1-\alpha)]u_0(b) = \beta.$
For $i = 1,2,\ldots$
  Compute the residual
    $r_{i-1} = f - Lu_{i-1}.$
  If $\|r_{i-1}\|_\infty / \|r_0\|_\infty < \eta$
    STOP
  Else solve the following problem for the defect $d_i$
    $L_m d_i = r_{i-1}$
    $d_i(a) = [\alpha(\frac{\mathrm{d}}{\mathrm{d}x} - A) + (1-\alpha)]d_i(b) = 0.$
  Compute the next iterate
    $u_i = u_{i-1} + d_i.$
End
```

In the next section we present Algorithm 3 for the explicit solution of Eq. (15) for d_i, at each step of Algorithm 1. The existence and uniqueness of the solution d_i for all $0 < \varepsilon < \varepsilon_0$ is guaranteed under a mild condition, as is discussed in Appendix A.

3.2. Solution of the approximate equation

Eq. (15) is readily solved with Algorithm 2 below when the boundary conditions are the following special case of (10):

$$\left(\frac{\mathrm{d}}{\mathrm{d}x} - S_m(b,\varepsilon)\right) d(b;\gamma) = \gamma, \quad d(a;\gamma) = 1. \tag{16}$$

Algorithm 2

- Solve the initial value problem (17) for w:

$$\left(\frac{\mathrm{d}}{\mathrm{d}x} + \frac{a_1(x)}{\varepsilon} + S_m(x,\varepsilon)\right) w(x;\gamma) = \frac{r(x)}{\varepsilon}, \qquad w(b;\gamma) = \gamma. \tag{17}$$

- Solve the initial value problem (18) for d:

$$\left(\frac{\mathrm{d}}{\mathrm{d}x} - S_m(x,\varepsilon)\right) d(x;\gamma) = w(x;\gamma), \qquad d(a;\gamma) = 1. \tag{18}$$

Since the problem (15) and (16) is linear, and γ enters linearly in (16), it follows that

$$d(x;\gamma) = U(x) + \gamma V(x). \tag{19}$$

In order to solve Eq. (15) subject to the general condition (10), we use a technique analogous to the "shooting" method for first-order systems of differential equations (see [12, 13, 22]). We replace (10) by (16), where γ is to be determined so that (10) is satisfied. It follows that (10) will be satisfied if γ is a root of

$$\alpha\gamma - \beta + [\alpha(S_m - A) + (1-\alpha)]d(b;\gamma) = 0. \tag{20}$$

To determine γ we use (19) in (20) to obtain

$$\{\alpha + [\alpha(S_m - A) + (1-\alpha)]V(b)\}\gamma = \beta - [\alpha(S_m - A) + (1-\alpha)]U(b). \tag{21}$$

If $\alpha + [\alpha(S_m - A) + (1-\alpha)]V(b) \neq 0$ there is a unique solution to this equation; otherwise there is no solution, unless the right-hand side is zero in which case there is an infinite number of solutions.

We summarize the scheme for the solution of problem (15), (10) in Algorithm 3 below.

Algorithm 3

(1) Evaluate $d(x;0) = U(x)$ using Algorithm 2.
(2) Evaluate $d(x;1) = U(x) + V(x)$ using Algorithm 2.
(3) Determine $\gamma(\beta)$ from

$$\gamma(\beta) = \frac{\beta - [\alpha(S_m - A) + (1-\alpha)]U(b)}{\alpha + [\alpha(S_m - A) + (1-\alpha)]V(b)}. \tag{22}$$

(4) Determine $d[x;\gamma(\beta)]$

$$d[x;\gamma(\beta)] = d(x;0) + \gamma(\beta)[d(x;1) - d(x;0)]. \tag{23}$$

3.3. Approximate factorization of L

We now present a scheme for obtaining the approximate factorization of L. The approximation of order $m \geqslant 0$ requires that a_0 and a_1 have $m-1$ derivatives.

In analogy with the exact factorization of second-order operators (see [1, p. 22]), we write

$$\varepsilon \frac{\mathrm{d}^2}{\mathrm{d}x^2} + a_1(x)\frac{\mathrm{d}}{\mathrm{d}x} + a_0(x) = \varepsilon\left(\frac{\mathrm{d}}{\mathrm{d}x} + \frac{a_1(x)}{\varepsilon} + S(x,\varepsilon)\right)\left(\frac{\mathrm{d}}{\mathrm{d}x} - S(x,\varepsilon)\right), \tag{24}$$

where the function $S(x,\varepsilon)$ is to be determined. We expand the right-hand side of this equation and subtract it from the left-hand side to obtain the following Riccati equation for S:

$$R(S,x) = 0, \tag{25}$$

where

$$R(S,x) = \varepsilon S^2 + \varepsilon S_x + a_1(x)S + a_0(x). \tag{26}$$

Since we do not know how to solve (25)explicitly, we seek an asymptotic approximation to S of the form

$$S(x,\varepsilon) \sim \frac{1}{\varepsilon}\sum_{j=0}^{\infty} s_j\varepsilon^j \quad \text{as } \varepsilon \to 0. \tag{27}$$

To determine the functions s_j we substitute (27) into (25), collect coefficients of equal powers of ε and equate the resulting coefficients of each power of ε to zero. This yields the following recursive system of algebraic equations for s_j:

$$a_1 s_0 + s_0^2 = 0, \tag{28}$$

$$a_1 s_1 + 2s_0 s_1 + s_{0,x} + a_0 = 0, \tag{29}$$

$$a_1 s_m + 2s_0 s_m + s_{m-1,x} + \sum_{k=1}^{m-1} s_k s_{m-k} = 0, \quad m \geqslant 2. \tag{30}$$

This system is solved readily and yields

$$s_0 = 0 \quad \text{or} \quad s_0 = -\, a_1(x), \tag{31}$$

$$s_1 = -\frac{s_{0,x} + a_0}{a_1 + 2s_0}, \tag{32}$$

$$s_m = -\frac{s_{m-1,x} + \sum_{k=1}^{m-1} s_k s_{m-k}}{a_1 + 2s_0}, \quad m \geqslant 2. \tag{33}$$

Now we define $S_m(x,\varepsilon)$ by

$$S_m(x,\varepsilon) = \frac{1}{\varepsilon}\sum_{j=0}^{m} s_j(x)\varepsilon^j. \tag{34}$$

Upon substituting S_m for S in (25), we find that

$$R(S_m, x) = \varepsilon^m r_m(x, \varepsilon), \tag{35}$$

where

$$r_m(x, \varepsilon) = \sum_{j=1}^{m} \varepsilon^{j-1} \sum_{k=j}^{m} s_k s_{m+j-k} + s_{m,x}, \quad m \geqslant 1. \tag{36}$$

It follows that the family of operators

$$L_m = \varepsilon \left(\frac{\mathrm{d}}{\mathrm{d}x} + \frac{a_1(x)}{\varepsilon} + S_m(x, \varepsilon) \right) \left(\frac{\mathrm{d}}{\mathrm{d}x} - S_m(x, \varepsilon) \right), \quad m \geqslant 1, \tag{37}$$

where $S_m(x, \varepsilon)$ is defined by (34) with $a_0 = 0$ satisfy the conditions (12)–(14).

3.4. Numerical calculations

We now present a few numerical calculations with QUAD for Eq. (6) of Section 2. Hence, we shall solve (9) in $[0, 2]$. The operator L is defined in (8) with

$$a_1 = \mathrm{i}2n(x), \qquad a_0 = \mathrm{i}n_x, \tag{38}$$

and the index of refraction $n(x)$ is chosen to be

$$n(x) = \mathrm{e}^{-20(x-1)^2} + 1. \tag{39}$$

The boundary conditions are (16) with $\gamma = 0$.

In our calculations the mesh parameter $h = 1/1000$ and the order of the approximate operator is $m = 1$. The approximate operator is inverted with the trapezoidal method and the exact operator is discretized with second-order central differences. The small parameter ε is varied from one calculation to the other. Fig. 1 indicates the infinity norm of the error, i.e., the difference between the exact solution of the central difference discretization of problem (6), (7) and the QUAD iterate. We note that even for $\varepsilon = 1$ the algorithm converges. As ε becomes smaller, the rate of convergence increases. The algorithm works well for $\varepsilon > h$ and $\varepsilon < h$. To demonstrate further the stability of the scheme, we performed additional calculations in which 2500 iterations of QUAD where performed. The results are displayed in Fig. 2.

4. Convergence analysis

4.1. The continuous iteration

We shall first show that the continuous version of QUAD converges. We present the analysis of Algorithm 1 for problem (9), (10) with $\alpha = 0$. The analysis for other values of α is analogous.

Theorem 1. *Suppose that problem* (9), (10) *with* $\alpha = 0$ *has a unique solution. Moreover suppose that* L_m *satisfies conditions* (12)–(14) *and that one of the existence conditions of Lemma 1 in*

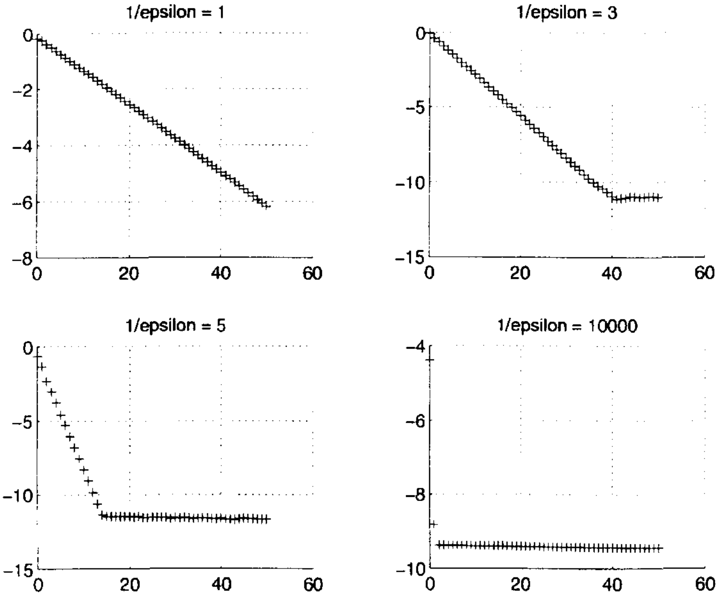


Fig. 1. The ordinate is the $\log_{10}$ of the infinity norm of the difference between the QUAD iterate and the solution to the finite difference discretization of (9). The abscissa is the number of iterations. The value of ε is indicated at the top of each graph.

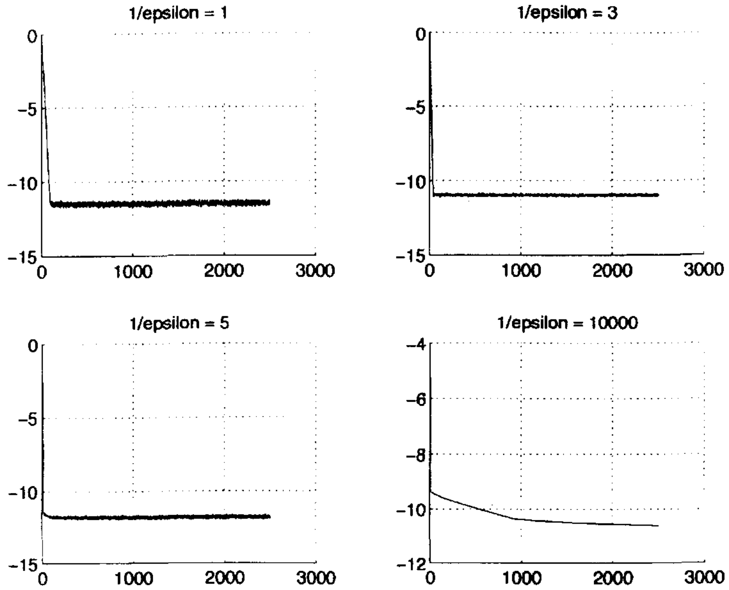


Fig. 2. Long-term behavior of the QUAD algorithm. The ordinate is $\log_{10}$ of the infinity norm of the difference between the QUAD iterate and the solution to the finite difference discretization of (9). The abscissa is the number of iterations. The value of ε is indicated at the top of each graph.

Appendix A is satisfied. Then there is an $\varepsilon_0 > 0$ such that if $\varepsilon < \varepsilon_0$, Algorithm 1 *converges and the error is reduced at each iteration by a factor which is* $O(\varepsilon^m)$.

Proof. In order to determine the error $e_i = u - u_i$ at step i, we substitute $d_i = u_i - u_{i-1}$ and $r_{i-1} = Lu - Lu_{i-1}$ in the equation for the defect d_i. Then we add and subtract $L_m u$ from the left of that equation to obtain after some manipulation

$$L_m e_i = (L_m - L)e_{i-1}, \tag{40}$$

$$e_i(a) = 0, \qquad e_i(b) = 0. \tag{41}$$

Upon substituting (12) in (40) we obtain

$$L_m e_i(x) = -\varepsilon^m r_m(x,\varepsilon)e_{i-1}(x). \tag{42}$$

We compute the solution to (42) and (41) with Algorithm 3 to obtain

$$e_i(x) = \varepsilon^m \left(G_i(x,\varepsilon) - G_i(b,\varepsilon)\frac{H(x,\varepsilon)}{H(b,\varepsilon)} \right), \tag{43}$$

where the homogeneous and particular solutions, H and G_i, are given by

$$H(x,\varepsilon) = \int_a^x \exp\left(\int_z^x S_m(t,\varepsilon)\,\mathrm{d}t + \int_z^b S_m(t,\varepsilon) + \int_z^b \frac{a_1(t)}{\varepsilon}\,\mathrm{d}t \right) \mathrm{d}z, \tag{44}$$

$$G_i(x,\varepsilon) = -\frac{1}{\varepsilon} \exp\left[\int_a^x S_m(t,\varepsilon)\,\mathrm{d}t \right] \int_a^x \exp\left[-\int_a^z \frac{a_1(t)}{\varepsilon} + 2S_m(t,\varepsilon)\,\mathrm{d}t \right]$$
$$\times \int_b^z \exp\left[\int_a^y \frac{a_1(t)}{\varepsilon} + S_m(t,\varepsilon)\,\mathrm{d}t \right] r_m(y,\varepsilon)e_{i-1}(y)\,\mathrm{d}y\,\mathrm{d}z. \tag{45}$$

We will show that there exist constants C_1, C_2 and η such that for all $\varepsilon > 0$

$$|G_i(x,\varepsilon)| \leqslant C_1 \|e_{i-1}\|_\infty, \tag{46}$$

and for all $0 < \varepsilon < \eta$

$$|H(x,\varepsilon)/H(b,\varepsilon)| < C_2. \tag{47}$$

We use these bounds in (43) to obtain

$$\|e_i\|_\infty \leqslant \varepsilon^m C_1(1 + C_2)\|e_{i-1}\|_\infty. \tag{48}$$

It follows that the iteration converges if $\varepsilon < \varepsilon_0$ where

$$\varepsilon_0 = \min\{\eta, (1/C_1(1 + C_2))^{1/m}\}. \tag{49}$$

In order to derive the bound (46) we divide and multiply (45) by

$$a_1(z) + \varepsilon 2S_m(z,\varepsilon),$$

and then we integrate by parts to obtain

$$G_i(x,\varepsilon)=\frac{\exp[\int_z^x S_m(t,\varepsilon)\,\mathrm{d}t]}{a_1(z)+2\varepsilon S_m(z,\varepsilon)}\int_2^z \exp\left[\int_z^y \frac{a_1(t)}{\varepsilon}+S_m(t,\varepsilon)\,\mathrm{d}t\right] r_m(y,\varepsilon)e_{i-1}(y)\,\mathrm{d}y\Bigg|_0^x$$
$$-\int_0^x \exp\left[\int_z^x S_m(t,\varepsilon)\,\mathrm{d}t\right] v(z)\,\mathrm{d}z, \tag{50}$$

where

$$v(z)=\frac{r_m(z,\varepsilon)e_{i-1}(z)}{a_1(z)+2\varepsilon S_m(z,\varepsilon)}-\frac{a_{1,z}(z)+2\varepsilon S_{m,z}(z)}{(a_1(z)+2\varepsilon S_m(z,\varepsilon))^2}$$
$$\times\int_2^z \exp\left[\int_z^y \frac{a_1(t)}{\varepsilon}+S_m(t,\varepsilon)\,\mathrm{d}t\right] r_m(y,\varepsilon)e_{i-1}(y)\,\mathrm{d}y. \tag{51}$$

The expression $\exp[\int_z^y (a_1(t)/\varepsilon)\,\mathrm{d}t]$ in (50) and (51) is uniformly bounded by 1 for all $\varepsilon>0$, because $z\leqslant y$ and $\mathrm{Re}[a_1]\leqslant 0$. We use this fact and relations (14) and (13) to obtain after some manipulation the bound (46).

In order to derive (47) we use (A.15) of Appendix A:

$$H(x,\varepsilon)\sim\varepsilon\left(\frac{f(a,x)}{a_1(a)}\mathrm{e}^{\phi(a)/\varepsilon}-\frac{f(x,x)}{a_1(x)}\mathrm{e}^{\phi(x)/\varepsilon}\right),\quad \varepsilon\to 0, \tag{52}$$

where f is defined in (A.16). Therefore

$$\left|\frac{H(x,\varepsilon)}{H(b,\varepsilon)}\right|\sim\left|\frac{f(a,x)\exp(\int_a^b \frac{a_1(t)}{\varepsilon}\,\mathrm{d}t)/a_1(a)-f(x,x)\exp(\int_x^b \frac{a_1(t)}{\varepsilon}\,\mathrm{d}t)/a_1(x)}{f(a,b)\exp(\int_a^b \frac{a_1(t)}{\varepsilon}\,\mathrm{d}t)/a_1(a)-1/a_1(b)}\right|,\quad \varepsilon\to 0. \tag{53}$$

The numerator in this expression is uniformly bounded for all ε because $\mathrm{Re}\,[a_1]\leqslant 0$. For ε sufficiently small the denominator is uniformly bounded away from 0 as indicated in Remark A.3 of Appendix A. Thus the right-hand side of (53) is bounded for ε sufficiently small. Then (53) yields the bound (47) and this completes the proof of Theorem 1. □

For future reference we introduce the *error amplification operator*

$$M_m=L_m^{-1}(L_m-L), \tag{54}$$

which determines the evolution of the error via the relation

$$e_i=M_m e_{i-1}, \tag{55}$$

$$e_i(a)=0,\quad e_i(b)=0, \tag{56}$$

in view of (40). Eq. (48) yields the bound

$$\|M_m\|_\infty<C\varepsilon^m, \tag{57}$$

where $C=C_1(1+C_2)$. An explicit formula for C can be found in [6].

4.2. The discrete iteration

4.2.1. Discretization

We next analyze the numerical implementation of Algorithm 1. We consider a mesh of $N+1$ points: $x_j = jh$, $j = 0, \dots, N$, $h = (b-a)/N$. We discretize the operator L in (8), with second-order finite difference schemes and we denote the discrete operator by L_h

$$L_h = \varepsilon D_+ D_- + a_1(x_j) D_0 + a_0(x_j), \tag{58}$$

$$D_+ D_- K_j = \frac{K_{j+1} - 2K_j + K_{j-1}}{h^2}, \quad D_0 K_j = \frac{K_{j+1} - K_{j-1}}{2h}. \tag{59}$$

The boundary conditions are incorporated in a standard fashion. At each step of Algorithm 1, we solve the first-order equations in Algorithm 2 using the trapezoidal or the backward Euler method [22]. Here we analyze the algorithm for an implementation with the trapezoidal method. The results for backward Euler are similar and may be found in [6].

The error in the discrete iteration evolves according to the relation

$$e_{h,j+1} = M_{m,h} e_{h,j}, \tag{60}$$

where $M_{m,h}$ is the discretization of M_m in (54). This relation follows from the continuous relation (55). To analyze relation (60) we introduce the modal representation

$$e_{h,j} = \sum_{\omega=0}^{N} \hat{e}_\omega e^{i2\pi\omega m/(N+1)}, \tag{61}$$

into that equation and we study the expression

$$M_h e^{i2\pi\omega m/(N+1)}, \tag{62}$$

for each mode of the grid $\omega = 0, \dots, n$, separately.

Standard numerical analysis arguments show that $M_h e^{i2\pi\omega m/(N+1)}$ is a good approximation to $M e^{i2\pi\omega m/(N+1)}$ for modes ω of the error satisfying the low frequency condition

$$\omega/N < v < 1, \tag{63}$$

where v is a small parameter. Hence the low frequency modes are damped by the discrete iteration in view of Theorem 1.

In the remainder of this section we evaluate (62) for the high-frequency modes. To simplify the analysis and assume that the coefficients of L_h and of $L_{h,m}$ are constant, and that the boundary conditions are the homogeneous version of (16). Hence in the remainder of this section S_m, a_1 and

$$p = a_1 + \varepsilon S_m, \tag{64}$$

are constants.

4.2.2. Numerical stability

To evaluate (62) we first apply L_h to mode ω and obtain that for all interior points of the grid

$$L_h e^{i2\pi\omega m/(N+1)} = \mu(\omega) e^{i2\pi\omega m/(N+1)}, \tag{65}$$

where

$$\mu(\omega) = \varepsilon\lambda^2(\omega) + a_1\lambda(\omega)\cos(\omega\pi/(N+1)) + a_0, \tag{66}$$

and

$$\lambda(\omega) = \mathrm{i}\pi\omega \ \mathrm{sinc}\left(\frac{\omega}{N+1}\right)\frac{N}{N+1}. \tag{67}$$

Then, we apply L_h^{-1} to the right-hand side of (65). We perform this calculation explicitly using the discrete version of Algorithm 2. Indeed, the discretization of a differential equation by the trapezoidal method yields a first-order recurrence equation, which can be solved explicitly. We find that

$$M_h \mathrm{e}^{\mathrm{i}2\pi m\omega/(N+1)} = (1 - \mu c_1)\mathrm{e}^{\mathrm{i}2\pi\omega m/(N+1)} - \mu c_2 \theta_1^{-m} + \mu(c_1 + c_2)\theta_2^m, \tag{68}$$

where

$$c_1(\omega) = \frac{\cos^2(\pi\omega/(N+1))}{\varepsilon\lambda^2(\omega) + a_1\lambda(\omega)\cos(\pi\omega/(N+1)) - pS\cos^2(\pi\omega/(N+1))}, \tag{69}$$

$$c_2(\omega) = \left(\frac{1 + \frac{1}{2}\frac{hp}{\varepsilon}}{1 - \frac{1}{2}\frac{hp}{\varepsilon}}\right)^N \frac{\mathrm{e}^{-\mathrm{i}2\pi\omega/(N+1)}\cos(\pi\omega/(N+1))}{(a_1 + 2\varepsilon S)\lambda(\omega) + (\frac{p^2}{\varepsilon} + Sp)\cos(\pi\omega/(N+1))}. \tag{70}$$

The values of θ_1 and θ_2 are

$$\theta_1 = \frac{1 + \frac{1}{2}\frac{hp}{\varepsilon}}{1 - \frac{1}{2}\frac{hp}{\varepsilon}}, \quad \theta_2 = \frac{1 + \frac{1}{2}hS}{1 - \frac{1}{2}hS}, \tag{71}$$

and p is defined in (64).

Expression (68) is a linear combination of three functions: mode ω, θ_1^{-m} and θ_2^m, the solutions to the discretized homogeneous form of (17) and (18), respectively. We analyze the magnitude of this linear combination for the high-frequency modes of the grid, for which the discrete iteration is not a good approximation to the continuous one. Specifically we consider this expression under two limits

- *Limit* I: $h \to 0$ with ε fixed. This is the "standard" numerical analysis limit.
- *Limit* II: $h \to 0$ and $h/\varepsilon \to \infty$. Here we assume that the solution to the equation is smooth and that it is well resolved with a mesh parameter $h \gg \varepsilon$.

In both limits we keep $\omega/(N+1)$ fixed and therefore

$$|\lambda(\omega)| \to \infty. \tag{72}$$

Here $\lambda(\omega)$ is defined in (67).

We use (69), (70) and (66) for $c_1(\omega)$, $c_2(\omega)$ and $\mu(\omega)$, respectively, in (68) to obtain for the trapezoidal method

$$1 - \mu c_1 = \frac{(\varepsilon\lambda^2 + a_1\lambda\cos(\alpha))(1 - \cos^2(\alpha)) - (a_0 + pS)\cos^2(\alpha)}{\varepsilon\lambda^2 + a_1\lambda\cos(\alpha) - pS\cos^2(\alpha)}, \tag{73}$$

where

$$\alpha = \pi\omega/(N+1). \tag{74}$$

In the limit $h \to 0$, we find that

$$\lim_{h\to 0} 1 - \mu c_1 = 1 - \cos^2(\pi\omega/(N+1)), \tag{75}$$

in view of relation (72). Hence, the coefficient of $e^{i2\pi m\omega/(N+1)}$ in (68) satisfies

$$|1 - \mu c_1| \leqslant 1. \tag{76}$$

We now analyze the remaining terms in (68): θ_1^m and θ_2^m where θ_i are defined in (71).

Limit 1: With ε fixed,

$$\lim_{h\to 0} \theta_1^{-m} = e^{-px/\varepsilon}, \tag{77}$$

$$\lim_{h\to 0} \theta_2^m = e^{Sx}. \tag{78}$$

Both of these terms are *smooth* in the sense that they are resolved by the numerical scheme and will be damped by additional iterations, in view of Theorem 1.

Limit 2: The term θ_2^m converges to the limit (78) which is smooth and therefore is of no concern. We evaluate the coefficient of θ_1^{-m} using (66) and (70) to obtain after some manipulation

$$\mu c_2 = \frac{\theta_1^N e^{-i2\pi\omega/(N+1)} \cos(\pi\omega/(N+1))\varepsilon\lambda(\varepsilon\lambda + a_1\cos(\pi\omega/(N+1)) + a_0/\lambda)}{(a_1 + 2\varepsilon S)\varepsilon\lambda + (p^2 + \varepsilon S p)\cos(\pi\omega/(N+1))}. \tag{79}$$

If $\cos(\pi\omega/(N+1)) = 0$, then $\mu c_2 = 0$. Otherwise, we use the relations

$$a_1\cos(\pi\omega/(N+1)) \gg a_0/\lambda, \quad \lambda \to \infty, \tag{80}$$

$$a_1 \gg 2\varepsilon S \quad \text{and} \quad p^2 + \varepsilon S p \sim a_1^2, \quad \varepsilon \to 0 \tag{81}$$

in the numerator and denominator of (79) to obtain that for large λ and small ε

$$|c_2\mu| \sim \left| \frac{\theta_1^N \cos(\pi\omega/(N+1))\varepsilon\lambda}{a_1} \right|. \tag{82}$$

This expression is arbitrarily small because $\varepsilon\lambda \to 0$ as $h \to 0$. We also note that for the highest modes of the grid, $\omega \approx (N+1)/2$, $\cos(\pi\omega/(N+1)) \approx 0$.

In summary, under the first limit, expression (73) consists of mode ω with a coefficient smaller than one and an additional term which is smooth and therefore is damped by additional iterations of the scheme, in view of Theorem 1. Under the second limit, (73) consists of mode ω with a coefficient smaller than one, a smooth term which is damped by additional iterations of the scheme, and a high-frequency term. The high-frequency term is of small magnitude and will be damped by additional iterations, in view of the analysis above. Hence the numerical implementation of QUAD is a stable numerical algorithm.

5. Exact factorization of an approximate operator

In Section 3.3 we derived L_m by approximately factoring L. Here we choose L_m such that it can be factored exactly. We do this to solve problems locally, in the neighborhood of singular points or singular lines, such as turning points or caustics. This approach also works when there is no small parameter in the equation.

In order to illustrate the scheme we return to the model equation of Section 2:

$$\varepsilon u'' + n(x)u = 0, \tag{83}$$

in $[-\delta, \delta]$ subject to

$$u(-\delta) = \alpha, \quad u(\delta) = \beta. \tag{84}$$

We denote the operator in this equation by L and we assume that $n(0) = 0$ is the only zero of n in $[-\delta, \delta]$. Moreover we assume that the solution

$$\mathrm{Ai}\left(\left[\frac{n'(0)}{\varepsilon}\right]^{1/3} x\right)$$

to the Airy equation

$$\varepsilon u'' + n'(0)xu = 0, \tag{85}$$

is nonzero in $[-\delta, \delta]$.

The representation (3) for u is not valid in a neighborhood of 0 because the nature of u changes from oscillatory to exponentially decaying there (see [1]). Hence in a small neighborhood of 0 we solve (83) directly by iteration. Indeed, the operator in (85) can be factored exactly as follows (see discussion in [1, p. 22])

$$\varepsilon\frac{\mathrm{d}^2}{\mathrm{d}x^2} + n'(0)x = \varepsilon\left(\frac{\mathrm{d}}{\mathrm{d}x} + \left[\frac{n'(0)}{\varepsilon}\right]^{1/3}\frac{\mathrm{Ai}'}{\mathrm{Ai}}\right)\left(\frac{\mathrm{d}}{\mathrm{d}x} - \left[\frac{n'(0)}{\varepsilon}\right]^{1/3}\frac{\mathrm{Ai}'}{\mathrm{Ai}}\right). \tag{86}$$

We denote this operator by $\hat{L}$ and we note that Algorithm 1 with $\hat{L}$ substituted for L_m converges to the solution of (83), provided δ is sufficiently small. This can be shown with an analysis analogous to the one used in the proof of Theorem 1, in view of the fact that

$$L - \hat{L} = n(x) - n'(0)x = n''(\xi)x^2/2, \quad \xi, x \in [-\delta, \delta].$$

The convergence factor is $\mathrm{O}(\delta^2)$ and it does not depend on ε.

6. Generalization to higher dimensions

6.1. A formal factorization scheme

We now describe a formal scheme for the approximate factorization of a class of singularly perturbed partial differential operators. This scheme is a natural generalization of the one-dimensional scheme of Section 3.3. We consider operators of the form

$$L = \varepsilon\left(\frac{\partial^2}{\partial\sigma^2} + b_1\frac{\partial^2}{\partial\beta\partial\sigma} + b_0\frac{\partial^2}{\partial\beta^2} + c_1\frac{\partial}{\partial\sigma} + c_0\frac{\partial}{\partial\beta}\right) + a_1\frac{\partial}{\partial\sigma} + a_0, \tag{87}$$

in a domain $\Omega \in \mathbb{R}^2$. The coefficients in (87) are functions of σ and β and may be complex valued, and $a_1 \neq 0$. To formally factor (87) we rewrite this operator emphasizing the σ direction

$$L = \varepsilon\left(\frac{\partial^2}{\partial\sigma^2} + \left(\frac{a_1}{\varepsilon} + c_1 + b_1\frac{\partial}{\partial\beta}\right)\frac{\partial}{\partial\sigma} + \left(\frac{a_0}{\varepsilon} + b_0\frac{\partial^2}{\partial\beta^2} + c_0\frac{\partial}{\partial\beta}\right)\right). \tag{88}$$

Now in analogy with the one-dimensional case we seek an approximate factorization of L of the form

$$L = \varepsilon\left(\frac{\partial}{\partial\sigma} + \left(\frac{a_1}{\varepsilon} + c_1 + b_1\frac{\partial}{\partial\beta}\right) + S\right)\left(\frac{\partial}{\partial\sigma} - S\right). \tag{89}$$

In (89) we have introduced the unknown operator $S(\sigma, \beta, \varepsilon, \partial/\partial\beta)$.

We expand the right-hand side of (89) and find that it is equal to

$$L - \varepsilon R, \tag{90}$$

were the residual R is given by

$$R = b_0\frac{\partial^2}{\partial\beta^2} + c_0\frac{\partial}{\partial\beta} + \frac{a_0}{\varepsilon} + S_\sigma + b_1 S_\beta + \left(c_1 + \frac{a_1}{\varepsilon}\right)S + b_1 S\frac{\partial}{\partial\beta} + S^2. \tag{91}$$

The expressions S_σ and S_β in (91) represent the operators obtained by differentiating the coefficients of S by σ and β, respectively. We now equate R to 0 and obtain an equation for S. This equation is a quadratic in S with first-order derivatives and it is similar to the Riccati equation for the function S of Section 3.3.

We construct a formal asymptotic expansion in powers of ε for S of the form

$$S \sim \frac{1}{\varepsilon}\sum_{j=0}^{\infty} s_j\left(\sigma, \beta, \frac{\partial}{\partial\beta}\right). \tag{92}$$

The operators $s_j(\sigma, \beta, \partial/\partial\beta)$ are linear operators in $\partial^n/\partial\beta^n$. Upon substituting (92) for S into (91), collecting terms with equal powers of ε and multiplying by ε^2 we obtain

$$R = \varepsilon^2\left(b_0\frac{\partial^2}{\partial\beta^2} + c_0\frac{\partial}{\partial\beta}\right) + \varepsilon a_0$$

$$+ \sum_{j=0}^{\infty}\left(s_{j-1,\sigma} + b_1 s_{j-1,\beta} + b_1 s_{j-1}\frac{\partial}{\partial\beta} + c_1 s_{j-1} + a_1 s_j + \sum_{k=0}^{j} s_k s_{j-k}\right)\varepsilon^j. \tag{93}$$

Then we equate the coefficient of each power of ε to 0 and obtain the following recursive system of equations for s_j:

$$a_1 s_0 + s_0^2 = 0, \tag{94}$$

$$a_1 s_1 + s_0 s_1 + s_1 s_0 = -\left(a_0 + s_{0,\sigma} + b_1 s_{0,\beta} + b_1 s_0\frac{\partial}{\partial\beta} + c_1 s_0\right), \tag{95}$$

$$a_1 s_2 + s_0 s_2 + s_2 s_0 = -\left(s_1^2 + b_0\frac{\partial^2}{\partial\beta^2} + c_0\frac{\partial}{\partial\beta} + s_{1,\sigma} + b_1 s_{1,\beta} + b_1 s_1\frac{\partial}{\partial\beta} + c_1 s_1\right), \tag{96}$$

$$a_1 s_j + s_0 s_j + s_j s_0 = -\left(s_{j-1,\sigma} + b_1 s_{j-1,\beta} + b_1 s_{j-1}\frac{\partial}{\partial\beta} + c_1 s_{j-1} + \sum_{k=1}^{j-1} s_k s_{j-k}\right). \tag{97}$$

Eq. (94) is readily solved and yields

$$s_0 = 0 \quad \text{or } s_0 = -a_1. \tag{98}$$

Therefore, as in the one-dimensional case, there are two possible solutions and we denote them by Case 1 and 2, respectively. We shall base our factorization on the solution with $s_0 = 0$. However, for the sake of completeness we shall also obtain the second solution in Appendix B. A solution of (94)–(97) is

Case 1

$$s_0 = 0,$$

$$s_1 = -a_0/a_1, \tag{99}$$

$$s_2 = -\frac{b_0}{a_1}\frac{\partial^2}{\partial\beta^2} + \left(\frac{a_0 b_1}{a_1^2} - \frac{c_0}{a_1}\right)\frac{\partial}{\partial\beta} + d(\sigma,\beta), \tag{100}$$

where $d(\sigma,\beta)$ is defined by

$$d(\sigma,\beta) = (a_1 a_{0,\sigma} - a_0 a_{1,\sigma} + b_1(a_1 a_{0,\beta} - a_0 a_{1,\beta}) + c_1 a_0 a_1 - a_0^2)/a_1^3, \tag{101}$$

$$s_j = -\frac{1}{a_1}\left(s_{j-1,\sigma} + b_1 s_{j-1,\beta} + b_1 s_{j-1}\frac{\partial}{\partial\beta} + c_1 s_{j-1} + \sum_{k=1}^{j-1} s_k s_{j-k}\right). \tag{102}$$

6.2. Application to the steady-state convection diffusion equation

We now apply (99)–(102) to the steady-state convection diffusion equation with sub-characteristics parallel to the x axis

$$\varepsilon\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)u - a(x,y)\frac{\partial u}{\partial x} = f. \tag{103}$$

To make (103) agree with (87), we identify x with σ and y with β. We find that

$$a_0 = 0, \qquad b_1 = 0, \qquad c_1 = 0, \qquad c_0 = 0, \qquad a_1 = -a(x,y), \qquad b_0 = 1. \tag{104}$$

We introduce these coefficients in (99)–(102) to obtain

$$s_0 = 0, \qquad s_1 = 0, \qquad s_2 = \frac{1}{a}\frac{\partial^2}{\partial y^2}, \qquad s_3 = -\frac{a_x}{a^3}\frac{\partial^2}{\partial y^2}. \tag{105}$$

Then

$$L_2 = \varepsilon\left(\frac{\partial}{\partial x} - \frac{a}{\varepsilon} + \frac{\varepsilon}{a}\frac{\partial^2}{\partial y^2}\right)\left(\frac{\partial}{\partial x} - \frac{\varepsilon}{a}\frac{\partial^2}{\partial y^2}\right), \tag{106}$$

$$L_3 = \varepsilon\left(\frac{\partial}{\partial x} - \frac{a}{\varepsilon} + \left(\frac{\varepsilon}{a} - \frac{\varepsilon^2 a_x}{a^3}\right)\frac{\partial^2}{\partial y^2}\right)\left(\frac{\partial}{\partial x} - \left(\frac{\varepsilon}{a} - \frac{\varepsilon^2 a_x}{a^3}\right)\frac{\partial^2}{\partial y^2}\right). \tag{107}$$

For $m \geqslant 2$ we find that

$$L_m - L = -R_m, \tag{108}$$

where

$$R_m = \varepsilon^m\left(s_{m,x} + \sum_{j=1}^{m}\left(\sum_{k=j}^{m} s_k s_{m+j-k}\right)\varepsilon^{j-1}\right). \tag{109}$$

Therefore

$$R_2 = -\varepsilon^2\frac{a_x}{a^2}\frac{\partial^2}{\partial y^2} + \frac{\varepsilon^3}{a}\frac{\partial^2}{\partial y^2}\frac{1}{a}\frac{\partial^2}{\partial y^2}, \tag{110}$$

$$\begin{aligned} R_3 = {} & \varepsilon^3\left(\frac{1}{a}\frac{\partial^2}{\partial y^2}\frac{1}{a}\frac{\partial^2}{\partial y^2} - \left(\frac{a_x}{a^3}\right)_x\frac{\partial^2}{\partial y^2}\right) \\ & - \varepsilon^4\left(\frac{1}{a}\frac{\partial^2}{\partial y^2}\frac{a_x}{a^3}\frac{\partial^2}{\partial y^2} + \frac{a_x}{a^3}\frac{\partial^2}{\partial y^2}\frac{1}{a}\frac{\partial^2}{\partial y^2}\right) + \varepsilon^5\frac{a_x}{a^3}\frac{\partial^2}{\partial y^2}\frac{a_x}{a^3}\frac{\partial^2}{\partial y^2}. \end{aligned} \tag{111}$$

The results (107), (111) agree with a third-order approximate factorization of the convection diffusion operator, described in [14]. That factorization is obtained by modifying one for the constant case by a method designed for the convection diffusion operator. The present method gives factorizations of all orders.

6.3. Convergence analysis

We now analyze the convergence of QUAD applied to (103) in $[0,1]\times[0,1]$ with L_2 given by the second-order approximation (106). Here we consider the special case in which the coefficient $a(x,y)$ in (103) depends on x alone and is continuously differentiable. We also assume Dirichlet boundary conditions at $x=0$, $y=0$ and $y=1$ and the special condition

$$\left(\frac{\partial}{\partial x} - \frac{\varepsilon}{a}\frac{\partial^2}{\partial y^2}\right)u = 0, \tag{112}$$

at $x = 1$.

The evolution of the error is determined by the relation

$$L_2 e_{j+1} = (L_2 - L)\, e_j, \tag{113}$$

$$e_{j+1}(x,0)=0, \qquad e_{j+1}(x,1)=0, \tag{114}$$

$$e_{j+1}(0,y)=0, \qquad \left(\frac{\partial}{\partial x}-\frac{\varepsilon}{a}\frac{\partial^2}{\partial y^2}\right)e_{j+1}(1,y)=0. \tag{115}$$

Relation (113) is derived in the same way as relation (40).

We now use (108) in (113) and we solve the resulting problem (113)–(115) in two steps. First we solve

$$\varepsilon\left(\frac{\partial}{\partial x}-\frac{a}{\varepsilon}+\frac{\varepsilon}{a}\frac{\partial^2}{\partial y^2}\right)w(x,y)=-R_2e_j(x,y), \tag{116}$$

$$w(1,y)=w(x,0)=w(x,1)=0, \tag{117}$$

for w and then we solve

$$\left(\frac{\partial}{\partial x}-\frac{\varepsilon}{a}\frac{\partial^2}{\partial y^2}\right)e_{j+1}(x,y)=w(x,y), \tag{118}$$

$$e_{j+1}(0,y)=e_{j+1}(x,0)=e_{j+1}(x,1)=0, \tag{119}$$

for e_{j+1}.

We use (110) for R_2 in problem (116), (117) which we then solve to obtain

$$w(x,y)=\sum_{\omega=1}^{\infty}\hat{w}(x,\omega)\sqrt{2}\sin(\omega\pi y), \tag{120}$$

where

$$\hat{w}(x,\omega)=\int_1^x \exp\left(-\int_x^t p(s,\omega)\,\mathrm{d}s\right)\frac{\hat{r}(t,\omega)}{\varepsilon}\,\mathrm{d}t, \tag{121}$$

$$p(s,\omega)=\frac{a(s)}{\varepsilon}+\frac{\varepsilon}{a(s)}[\omega\pi]^2, \tag{122}$$

$$\hat{r}(t,\omega)=-\left(\frac{\varepsilon\omega\pi}{a(t)}\right)^2[a_x(t)+\varepsilon(\omega\pi)^2]\hat{e}_j(t,\omega), \tag{123}$$

and $\hat{e}_j(x,\omega)$ is the ω Fourier coefficient of $e_j(x,y)$

$$e_j(x,y)=\sum_{\omega=1}^{\infty}\hat{e}_j(x,\omega)\sqrt{2}\sin(\omega\pi y). \tag{124}$$

We now define r^*, p^*, x_p, and x_r as

$$r^*=\max_{x\in[0,1]}|\hat{r}(x,\omega)|=\left|\left(\frac{\varepsilon\omega\pi}{a(x_r)}\right)^2[a_x(x_r)+\varepsilon(\omega\pi)^2]\,\hat{e}(x_r,\omega)\right|, \tag{125}$$

$$p^*=\min_{x\in[0,1]}p(x,\omega)=\frac{a(x_p)}{\varepsilon}+\frac{\varepsilon}{a(x_p)}[\omega\pi]^2. \tag{126}$$

Then we manipulate (121) to obtain

$$|\hat{w}(x,\omega)| \leqslant \frac{r^*}{\varepsilon p^*}[1-\exp(p^*(x-1))]. \tag{127}$$

We now solve the problem (118), (119) to obtain

$$e_{j+1}(x,y) = \sum_{\omega=1}^{\infty} \hat{e}_{j+1}(x,\omega)\sqrt{2}\sin(\omega\pi y), \tag{128}$$

$$\hat{e}_{j+1}(x,\omega) = \int_0^x \exp\left(-\int_t^x q(s,\omega)\,\mathrm{d}s\right)\hat{w}(t,\omega)\,\mathrm{d}t, \tag{129}$$

$$q(s,\omega) = \frac{\varepsilon}{a(s)}[\omega\pi]^2. \tag{130}$$

We note that the series (120), (124) and (128) are uniformly convergent in view of the homogeneous conditions at $y=0$ and $y=1$ that w, e_j and e_{j+1} satisfy.

We define q^* and x_q as

$$q^* = \min_{x\in[0,1]} q(x,\omega) = \frac{\varepsilon}{a(x_q)}[\omega\pi]^2, \tag{131}$$

and we manipulate (129) to obtain

$$|\hat{e}_{j+1}(x,\omega)| \leqslant \int_0^x \exp(q^*(t-x))|\hat{w}(t,\omega)|\,\mathrm{d}t. \tag{132}$$

Then we use (127) in (132) and manipulate the resulting expression to obtain

$$|\hat{e}_{j+1}(x,\omega)| \leqslant \frac{r^*}{\varepsilon p^* q^*}[1-\exp(-q^*x)]. \tag{133}$$

We use the definitions of r^*, p^*, and q^* in (125)–(131) to obtain

$$\frac{\|\hat{e}_{j+1}(x,\omega)\|_\infty}{\|\hat{e}_j(x,\omega)\|_\infty} \leqslant \frac{a(x_q)a(x_p)}{a(x_r)^2}\frac{|\varepsilon a_x(x_r)+(\varepsilon\omega\pi)^2|}{a(x_p)^2+(\varepsilon\omega\pi)^2}. \tag{134}$$

We now define the convergence factor of mode ω, ρ_ω, by

$$\rho_\omega = \frac{\|\hat{e}_{j+1}(x,\omega)\|_\infty}{\|\hat{e}_j(x,\omega)\|_\infty}, \tag{135}$$

and we note from (134) that

$$\rho_\omega \leqslant A^2\frac{|\varepsilon a_x(x_r)+(\varepsilon\omega\pi)^2|}{a(x_p)^2+(\varepsilon\omega\pi)^2}, \tag{136}$$

where

$$A = \frac{\max_{x\in[0,1]} a(x)}{\min_{x\in[0,1]} a(x)}. \tag{137}$$

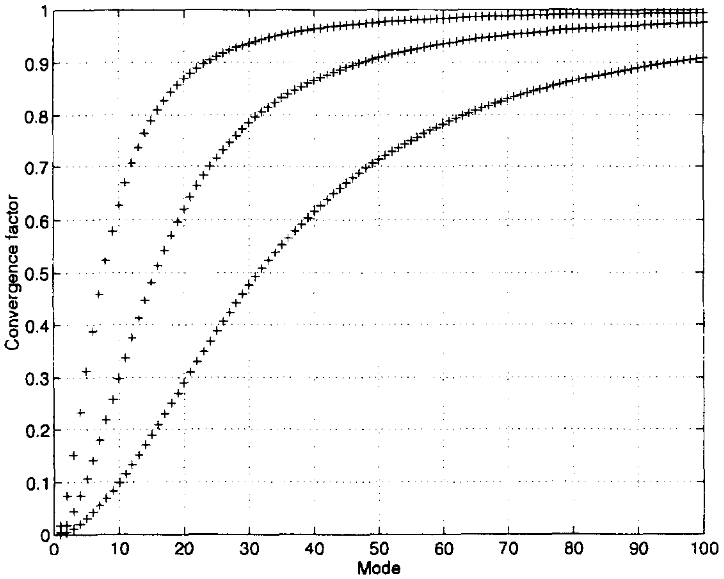


Fig. 3. The bound for ρ_ω as a function of ω for $\varepsilon=\frac{1}{25}$, $\frac{1}{50}$ and $\frac{1}{100}$, from left to right.

Fig. 3 shows the bound (136) normalized by A^2, as a function of ω for $\varepsilon=\frac{1}{25}$, $\frac{1}{50}$ and $\frac{1}{100}$. The bound is evaluated with $a(x_p)=a_x(x_r)=1$. We see that modes $\omega \ll \varepsilon^{-1}$ are rapidly damped and that ρ_ω is uniformly bounded for all modes. It follows that the numerical implementation of QUAD is particularly effective when the mesh parameter required to accurately resolve the problem satisfies $h>\varepsilon$. Then all the modes of the grid are in the rapid convergence range. This occurs for the amplitude equations generated by the hybrid numerical asymptotic method described in Section 2.

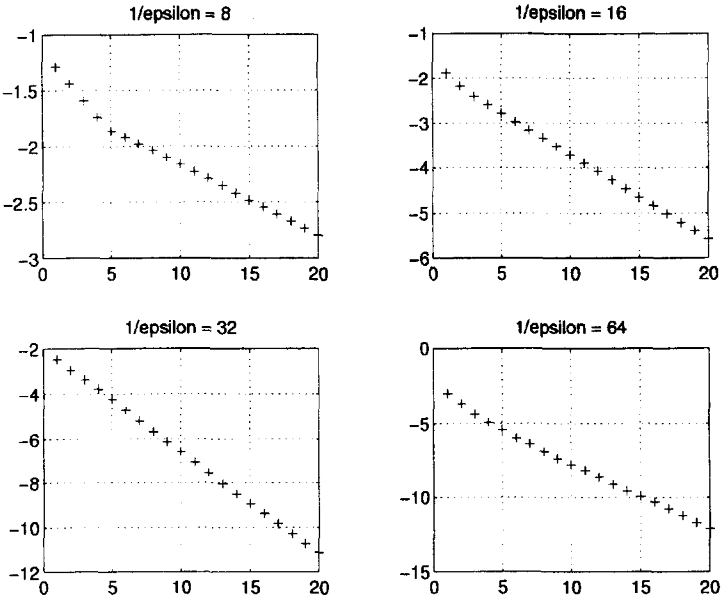
In order to enhance the rate of decay of the high frequency modes of the grid, an additional smoothing step [2] can be used at each iteration. Moreover, we are currently seeking an approximate factorization scheme which generates operators that effectively damp modes in intervals $[\omega_1,\omega_2]$, with $\omega_1>1/\varepsilon$.

The formal factorization scheme of Section 6.1 does not always yield a useful approximate operator for QUAD. Indeed, if we substitute $a(x_p)=\mathrm{i}b$ with b real in (136), we find that the resulting expression has a pole at $\omega=b/(\varepsilon\pi)$. Hence we expect that $\rho_\omega \gg 1$ for modes $\omega\approx b/(\varepsilon\pi)$ and QUAD is highly unstable in this case. We have observed this instability in numerical calculations for Eq. (103) in which $\mathrm{i}b(x,y)$ was substituted for $a(x,y)$, with b real.

6.4. Numerical calculations

We now apply QUAD to solve (103) in $[0,1]\times[0,1]$. The forcing function f is chosen such that the solution to the problem is

$$u(x,y)=(x-1)^2\sin(10\pi y)\cos(10\pi x). \qquad (138)$$


1/epsilon = 8

1/epsilon = 16

1/epsilon = 32

1/epsilon = 64

Fig. 4. The magnitude of the residual (141) for the QUAD iterates. In each graph the ordinate is $\log_{10} \|\text{residual}\|_\infty$ for the value of ε shown at the top. The abscissa is the number of iterations.

We impose Dirichlet conditions at $y=0$, $y=1$ and $x=0$ which are compatible with the solution (138). At $x=1$, we impose the boundary condition

$$\left(\frac{\partial}{\partial x}-\frac{\varepsilon}{a}\frac{\partial^2}{\partial y^2}\right)u=0. \tag{139}$$

The coefficient $a(x,y)$ in (103) is

$$a(x,y)=0.5\times(3-\cos(20\pi\sqrt{x^2+y^2})). \tag{140}$$

We use the second-order factorization (106) and the factors are inverted with the Crank Nicholson methods. In our calculation the mesh parameter is $h=\frac{1}{100}$. The magnitude of the residual

$$f_h-L_hu_k, \tag{141}$$

is shown in Fig. 4 for several values of ε. The rate of convergence increases as ε decreases and the iteration is stable in all cases.

6.5. Other applications

We now consider operators of the form

$$L=\varepsilon\left(\frac{\partial^2}{\partial x^2}+A\frac{\partial^2}{\partial x\partial y}+B\frac{\partial^2}{\partial y^2}+C\frac{\partial}{\partial x}+D\frac{\partial}{\partial y}\right)+E\frac{\partial}{\partial y}+F\frac{\partial}{\partial x}+G, \tag{142}$$

in $\Omega \subset \mathbb{R}^2$ where the functions A–G may be complex valued, and $[E,F] \neq 0$. The operators (142) can be formally factored using the scheme of Section 6.1, once they are transformed into the form (87). This is possible by first representing (142) in a characteristic coordinate system determined by curves $(X(\sigma,\tau), Y(\sigma,\tau))$ [23] satisfying

$$\frac{\partial Y}{\partial \sigma} = E(X,Y), \qquad \frac{\partial X}{\partial \sigma} = F(X,Y), \tag{143}$$

and then normalizing the resulting operator by the coefficient of $\partial^2/\partial\sigma^2$.

This transformation is possible only when the equations

$$x = X(\sigma,\tau), \qquad y = Y(\sigma,\tau), \tag{144}$$

can be uniquely inverted for σ and τ in Ω. Moreover, the transformation of variables requires derivatives of $\sigma(x,y)$ and $\tau(x,y)$ which in some cases can only be approximated numerically.

7. Future direction

The treatment of general boundary conditions in more than one dimension is possible with a "shooting method" for second order systems in factored form, analogous to the method of Section 3.2. We shall present this method in a future paper. We are also interested in the use of the factored operators as preconditioners for Krylov accelerators and the combination of QUAD with a smoothing step. Approximate operators that would complement the operators of Section 6.1 and enable QUAD to effectively damp modes of the error in other frequency ranges are desirable.

Appendix A. Existence and uniqueness of solution

We now study the existence and uniqueness of the solution to the boundary value problem

$$\varepsilon\left(\frac{\mathrm{d}}{\mathrm{d}x} + \frac{a_1(x)}{\varepsilon} + S_m(x,\varepsilon)\right)\left(\frac{\mathrm{d}}{\mathrm{d}x} - S_m(x,\varepsilon)\right) = f, \tag{A.1}$$

subject to either of the boundary conditions

$$u(b) = \beta, \qquad u(a) = 1, \tag{A.2}$$

$$\left(\frac{\mathrm{d}}{\mathrm{d}x} - A\right)u(b) = \beta, \qquad u(a) = 1. \tag{A.3}$$

Here we assume that

$$S_m(b,0) \neq A, \tag{A.4}$$

$a_1 \neq 0$ and $\mathrm{Re}[a_1] \leqslant 0$. We will now prove

Lemma A.1. (1) *The problem* (A.1)–(A.3) *has a unique solution for ε sufficiently small.*

(2) *The problem* (A.1), (A.2) *has a unique solution for* ε *sufficiently small provided* $\mathrm{Re}[a_1]$ *is not identically* 0 *in* $[a,b]$.

(3) *The problem* (A.1), (A.2) *has a unique solution for* ε *sufficiently small provided* $\mathrm{Re}[a_1]$ *is identically* 0 *in* $[a,b]$ *and*

$$\left|\exp\left(2\int_a^b S_m(t,0)\,\mathrm{d}t\right)\right| \neq |a_1(a)/a_2(b)|. \tag{A.5}$$

Proof. We first rewrite (A.1) as the first-order system

$$\begin{pmatrix} u \\ w \end{pmatrix}' = \begin{pmatrix} S_m & 1 \\ 0 & -\left(\dfrac{a_1}{\varepsilon}+S_m\right) \end{pmatrix}\begin{pmatrix} u \\ w \end{pmatrix} + \begin{pmatrix} 0 \\ f \end{pmatrix}, \tag{A.6}$$

subject to

$$B_a\begin{pmatrix} u(a) \\ w(a) \end{pmatrix} + B_b\begin{pmatrix} u(b) \\ w(b) \end{pmatrix} = \begin{pmatrix} 1 \\ \beta \end{pmatrix}, \tag{A.7}$$

where

$$B_a = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{A.8}$$

and B_b is given for conditions (A.2) and (A.3) respectively by

$$B_b = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad B_b = \begin{pmatrix} 0 & 0 \\ S_m - A & 1 \end{pmatrix}, \tag{A.9}$$

respectively. The existence theory for linear boundary value problems (see [13, Theorem 3.26]) guarantees a unique solution to these equations provided the matrix

$$Q(\varepsilon) = B_a Y(a,\varepsilon) + B_b Y(b,\varepsilon), \tag{A.10}$$

is nonsingular. Here Y is a fundamental solution matrix for the system (A.6) and we choose

$$Y(x,\varepsilon) = \begin{pmatrix} \exp(\int_b^x S_m\,\mathrm{d}t)\int_a^x \exp(\int_z^b [2S_m(t,\varepsilon) + \frac{a_1(t)}{\varepsilon}]\mathrm{d}t)\mathrm{d}z & \exp(\int_a^x S_m\,\mathrm{d}t) \\ \exp(\int_x^b [S_m + \frac{a_1(t)}{\varepsilon}]\mathrm{d}t) & 0 \end{pmatrix}. \tag{A.11}$$

The determinant of $Q(\varepsilon)$ is nonzero provided that

$$H(b,\varepsilon) \neq \mu, \tag{A.12}$$

where

$$H(x,\varepsilon) = \int_a^x \exp\left(\int_z^x S_m(t,\varepsilon)\,\mathrm{d}t + \int_z^b S_m(t,\varepsilon) + \int_z^b \frac{a_1(t)}{\varepsilon}\mathrm{d}t\right)\mathrm{d}z. \tag{A.13}$$

and the constant μ is given by

$$\mu=0, \qquad \mu=\frac{1}{A-S_m(b,\varepsilon)}, \tag{A.14}$$

for conditions (A.2) and (A.3), respectively. In order to verify when condition (A.12) holds, we first determine the asymptotic expansion of H as $\varepsilon\to 0$.

Remark A.2. H defined in (A.13) satisfies

$$H(x,\varepsilon)\sim\varepsilon\left(\frac{f(a,x)}{a_1(a)}\mathrm{e}^{\phi(a)/\varepsilon}-\frac{f(x,x)}{a_1(x)}\mathrm{e}^{\phi(x)/\varepsilon}\right), \quad \varepsilon\to 0. \tag{A.15}$$

Here we have introduced the functions

$$f(z,x)=\exp\left(\int_z^x S_m(t,\varepsilon)\,\mathrm{d}t+\int_z^b S_m(t,\varepsilon)\,\mathrm{d}t\right), \quad \phi(z)=\int_z^b a_1(t)\,\mathrm{d}t. \tag{A.16}$$

To prove (A.15) we rewrite H as the *Laplace* integral

$$H(x,\varepsilon)=\int_a^x f(z,x)\mathrm{e}^{\phi(z)/\varepsilon}\,\mathrm{d}z. \tag{A.17}$$

In view of the relation $\phi'(z)=-a_1(z)\neq 0$, and the differentiability of f and a_1, we integrate (A.17) by parts to obtain

$$H(x,\varepsilon)=\varepsilon\left(\frac{f(a,x)}{a_1(a)}\mathrm{e}^{\phi(a)/\varepsilon}-\frac{f(x,x)}{a_1(x)}\mathrm{e}^{\phi(x)/\varepsilon}\right)+\varepsilon\int_a^x\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{f(z,x)}{a_1(z)}\right]\mathrm{e}^{\phi(z)/\varepsilon}\,\mathrm{d}z. \tag{A.18}$$

The integral in (A.18) is o(1) as $\varepsilon\to 0$. Indeed, let

$$x_0=\max\left\{x\,|\,x\in[a,b],\ \mathrm{Re}[a_1]<0\right\}. \tag{A.19}$$

Then for all $x\in[a,x_0)$, $\mathrm{Re}[\phi(z)]<0$, and the integrand

$$\frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{f(z,x)}{a_1(z)}\right]\mathrm{e}^{\phi(z)/\varepsilon},$$

tends to zero as $\varepsilon\to 0$. Hence, the integral in (A.18) is o(1) as $\varepsilon\to 0$, for all $x\in[a,x_0]$. If $x_0<b$ then for all $x\in[x_0,b]$, $\mathrm{Re}[a_1]=0$ and $\mathrm{Im}[a_1]\neq 0$. Hence $\phi(z)$ is purely imaginary, continuously differentiable and $-\mathrm{i}\phi$ is monotone, because $\phi'(z)=-a_1(z)\neq 0$. The change of variables $u=\phi(z)/\mathrm{i}$ yields a Fourier integral which is o(1) as $\varepsilon\to 0$, in view of the Riemann Lebesgue Lemma. Thus (A.15) is proved.

We now complete the proof of the lemma by verifying when condition (A.12) holds in each of the two cases given in (A.14).

Case I: $\mu\neq 0$

- In view of the expansion (A.15), $H\to 0$ as $\varepsilon\to 0$ and therefore condition (A.12) holds for all sufficiently small ε.

Case II: $\mu=0$

- If both a_1 and S_m are real then H in (A.13) is positive for $b>a$ and (A.12) holds for all $\varepsilon>0$.

- If $\mathrm{Re}[a_1]$ is not identically zero in $[a,b]$, then $\phi(a)$ has a negative real part and $H \sim -\varepsilon/a_1(b)$ as $\varepsilon \to 0$, in view of (A.18). Hence, (A.12) holds for all sufficiently small ε.
- If a_1 is purely imaginary so is $\phi(a)$ and (A.12) holds for all sufficiently small ε provided (A.5) holds, in view of (A.18).

It follows from the above discussion that

Remark A.3. When one of the conditions of Lemma A.1 is satisfied, the function H/ε is uniformly bounded away from 0 for all sufficiently small ε.

Appendix B. Case 2 of the asymptotic factorization

We now solve (95)–(97) when $s_0 = -a_1$. Specifically, we solve (95) in detail and describe how to extend the solution procedure to (96) and (97).

In order to solve (95) we substitute $-a_1$ for s_0 into that equation to obtain

$$s_1 a_1 = a_0 - a_{1,\sigma} - b_1 a_{1,\beta} - b_1 a_1 \frac{\partial}{\partial\beta} - c_1 a_1. \tag{B.1}$$

The right-hand side of (B.1) is a first-order differential operator and we conclude that s_1 has the form

$$s_1 = \alpha_0 + \alpha_1 \frac{\partial}{\partial\beta}. \tag{B.2}$$

The coefficients α_0 and α_1 in (B.2) are to be determined so that s_1 satisfies (B.1). Upon substituting (B.2) into (B.1), we obtain

$$a_1\alpha_0 + a_1\alpha_1 \frac{\partial}{\partial\beta} + \alpha_1 a_{1,\beta} = a_0 - a_{1,\sigma} - b_1 a_{1,\beta} - b_1 a_1 \frac{\partial}{\partial\beta} - c_1 a_1. \tag{B.3}$$

We equate the coefficient of $\partial/\partial\beta$ on each side of (B.3) to obtain

$$\alpha_1 = -b_1. \tag{B.4}$$

Then, we introduce the value of α_1 back into (B.3) and find that

$$\alpha_0 = \frac{a_0 - a_{1,\sigma} - c_1 a_1}{a_1}. \tag{B.5}$$

Combining (B.2), (B.4) and (B.5) we obtain

$$s_1 = \frac{a_0 - a_{1,\sigma} - c_1 a_1}{a_1} - b_1 \frac{\partial}{\partial\beta}. \tag{B.6}$$

The procedure we used to solve (95) can be generalized to solve (96) and (97). We now describe this extension to (97). In order to do so, we first show by induction that the right-hand side of that equation is a linear differential operator of order j. We denote this operator by

$$\beta_0 + \beta_1 \frac{\partial}{\partial\beta} + \cdots + \beta_j \frac{\partial^j}{\partial\beta^j}. \tag{B.7}$$

The coefficients β_j in (B.7) depend on the coefficients of the operators $s_0, \dots s_{j-1}$. The left-hand side of (97) is $-s_j a_1$. Hence in order for s_j to satisfy (97), it has to be a linear differential operator of order j

$$s_j = \alpha_0 + \alpha_1 \frac{\partial}{\partial\beta} + \cdots + \alpha_j \frac{\partial^j}{\partial\beta^j}. \tag{B.8}$$

The unknown coefficients α_l in (B.8) are to be determined so that s_j satisfies (97). Upon introducing the right-hand side of (B.8) into the left-hand side of (97) and then replacing the right-hand side of (97) with (B.7) we find that

$$a_1 \left(\alpha_1 \frac{\partial}{\partial\beta} + \cdots + \alpha_j \frac{\partial^j}{\partial\beta^j} \right) + \alpha_0 a_1 + \alpha_1 a_1^{(1)} + \cdots \alpha_j a_1^{(j)} = \beta_0 + \beta_1 \frac{\partial}{\partial\beta} + \cdots + \beta_j \frac{\partial^j}{\partial\beta^j}, \tag{B.9}$$

where $a_1^{(j)}$ denotes $\partial^j/\partial\beta^j a_1$. Upon equating coefficients of derivatives of the same order we find that

$$\alpha_j = \beta_j / a_1 \quad \text{for } j \geqslant 1, \tag{B.10}$$

and

$$\alpha_0 = \frac{1}{a_1} \left(\beta_0 - \frac{\beta_1 a_1^{(1)}}{a_1} - \cdots - \frac{\beta_1 a_1^{(j)}}{a_1} \right). \tag{B.11}$$

References

[1] C.M. Bender, S. Orszag, Advanced Mathematical Methods for Scientists and Engineers, McGraw-Hill, New York, 1978.

[2] W.L. Briggs, A Multigrid Tutorial, SIAM, Philadelphia, PA, 1987.

[3] R. Chin, A domain decomposition method for generating orthogonal polynomials, J. Comput. Phys. 99 (1992) 321–336.

[4] R.C.Y. Chin, G.W. Hedstrom, Domain decomposition: an instrument of asymptotic-numerical methods, in: H.G. Kaper, M. Garbey (Eds.), Asymptotic Analysis and the Numerical Solution of Partial Differential Equations, Dekker, New York, 1991.

[5] R.C.Y. Chin, R. Krasny, A hybrid asymptotic-finite element method for stiff two-point boundary value problems, SIAM J. Sci. Statist. Comput. 4 (1983) 229–243.

[6] E. Giladi, Hybrid numerical asymptotic methods, Ph.D. thesis, Stanford University, 1995.

[7] E. Giladi, J.B. Keller, A hybrid numerical asymptotic method for the solution of differential equations, in preparation.

[8] G.W. Hedstrom, F.A. Howes, A domain decomposition method for a convection diffusion equation with turning point, in: T.F. Chan, R. Glowinski, J. Periaux, O. Widlund (Eds.), Domain Decomposition Methods, 1988.

[9] P.W. Hemker, The defect correction principle, in: J.J.H. Miller (Ed.), An Introduction to Computational and Asymptotic Methods for Boundary and Interior Layers, Boole press, Dublin, 1982.

[10] H.G. Kaper, M. Garbey, editors. Asymptotic Analysis and the Numerical Solution of Partial Differential Equations, Marcel Dekker, New York, 1991.

[11] H.G. Kaper, M. Garbey (Eds.), NATO Advanced Workshop on Asymptotic-Induced Numerical Methods for Partial Differential Equations, Critical Parameters, and Domain Decomposition, Kluwer, Dordrecht, 1993.

[12] H.B. Keller, Numerical Methods for Two-Point Boundary-Value Problems, Dover, New York, 1992.

[13] R.M.M. Mattheij, U.M. Ascher, R.D. Russell, Numerical Solution of Boundary Value Problems for Ordinary Differential Equations, SIAM, Philadelphia, PA, 1995.

[14] F. Nataf, J.P. Loheac, M. Schatzman, Parabolic approximations of the convection-diffusion equation, Math. Comp. 60 (1993) 515–530.

[15] S.A. Pruess, Solving linear boundary value problems by approximating the coefficients, Math. Comp. 27 (1970) 551–561.

[16] G. Rodrigue, E. Reiter, A domain decomposition method for boundary layer problems, in: T.F. Chan, R. Glowinsky, J. Periaux, O. Widlund (Eds.), Domain Decomposition Methods, 1988.

[17] P.S. Saylor, S.F. Ashby, J.S. Scroggs, Physically motivated domain decomposition preconditioners, in: Proc. Copper Mountain Conf. on Iterative Methods, vol. 1, 1992.

[18] J.S. Scroggs, An iterative method for systems of nonlinear hyperbolic equations, Comput. Math. Appl. 21(5) (1989) 137–144.

[19] J.S. Scroggs, A physically motivated domain decomposition for singularly perturbed equations, SIAM J. Numer. Anal. 28(1) (1991) 168–178.

[20] J.S. Scroggs, J. Saltz, Distributed-memory computing of a physically-motivated domain decomposition method, in: Proc. of SIAM Conf. on Domain Decomposition, SIAM, Philadelphia, PA, 1990.

[21] H.J. Stetter, The defect correction principle and discretization methods, Numer. Math. 29 (1978) 425–443.

[22] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Springer, Berlin, 1993.

[23] E. Zauderer, Partial Differential Equations of Applied Mathematics, Wiley Interscience, New York, 1989.