ELSEVIER

Applied Numerical Mathematics 18 (1995) 489–501

APPLIED NUMERICAL MATHEMATICS

An effective numerical technique for solving a special class of ordinary difference equations

V.S. Ryaben'kii [a,*], S.V. Tsynkov [b,1]

[a] *Keldysh Institute for Applied Mathematics, Russian Academy of Sciences, Miusskaya sq. 4, 125047 Moscow, Russia*
[b] *NASA Langley Research Center, Mail Stop 128, Hampton, VA 23681-0001, USA*

Abstract

We consider a system of ordinary difference equations with constant coefficients, which is defined on an infinite one-dimensional mesh. The right-hand side (RHS) of the system is compactly supported, therefore, the system appears to be homogeneous outside some finite mesh interval. At infinity, we impose certain boundary conditions, e.g., conditions of boundedness or decay of the solution, so that the resulting boundary-value problem is uniquely solvable and well posed.

We also consider a truncation of this infinite-domain problem to some finite mesh interval that entirely contains the support of the RHS. We require that the solution to this truncated problem, which is the one we are going to actually calculate, coincides on the finite mesh interval where it is defined with the corresponding fragment of the solution to the original (infinite) problem.

This requirement necessitates setting some special boundary conditions at the ends of the aforementioned finite interval. In so doing, one should guarantee an exact transfer of boundary conditions from infinity through the (semi-infinite) intervals of homogeneity of the original system. It turns out that the desired boundary conditions at the ends of the finite interval can be naturally formulated in terms of the eigen subspaces of the system operator. This, in turn, enables us to develop an effective numerical algorithm for solving the system of ordinary difference equations on the finite mesh interval. This algorithm can be referred to as a version of the well-known successive substitution technique but without its final ("inverse" or "resolving") stage.

The special class of systems described in this paper appears to be most useful when constructing highly accurate artificial boundary conditions (ABCs) for the numerical treatment of problems initially formulated on unbounded domains. Therefore, an effective numerical algorithm for solving such systems becomes an important issue.

1. Formulation of the problem

We consider a two-point system of ordinary difference equations,

$$\boldsymbol{A}\boldsymbol{v}_{m+1} + \boldsymbol{B}\boldsymbol{v}_m = \boldsymbol{f}_{m+1/2}, \quad -\infty < m < +\infty, \tag{1}$$

* Corresponding author. Fax: ++7-095-972-0737. E-mail: ryab@applmat.msk.su.
[1] Fax: +1-804-864-8816. E-mail: s.v.tsynkov@larc.nasa.gov.

Elsevier Science B.V.
SSDI 0168-9274(95)00081-X

which is defined on an infinite one-dimensional mesh $-\infty < m < +\infty$. Here, v_m, $-\infty < m < +\infty$, is an n-dimensional vector function representing unknowns, $f_{m+1/2}$, $-\infty < m < +\infty$, is an n-dimensional vector function representing the right-hand side (RHS), A and B are the constant square matrices of order n. None of the matrices A and B is expected to be singular, so that both $Q = A^{-1}B$ and $Q^{-1} = B^{-1}A$ exist. The entries of A and B, as well as the components of v_m and $f_{m+1/2}$ may, generally speaking, be complex.

The only specific assumption we do regarding the RHS $f_{m+/12}$, $-\infty < m < +\infty$, of system (1), is that $f_{m+1/2}$ is compactly supported. More precisely, we henceforth allow $f_{m+1/2}$ to differ from zero nowhere except for some finite mesh interval $\{m + 1/2 \mid m = 0, \ldots, M-1\}$, M being a positive integer.

We first assume that none of the eigenvalues μ_s, $s = 1, \ldots, n$, of the matrix Q has an absolute value equal to unity, $|\mu_s| \neq 1$, $s = 1, \ldots, n$. Moreover, we assume that both eigenvalues μ_s with the absolute values less than unity, $|\mu_s| < 1$, and eigenvalues μ_s with the absolute values greater than unity, $|\mu_s| > 1$, always exist in the spectrum of Q. Therefore, we can require that the solution v_m to (1) vanishes at infinity, i.e.,

$$v_m \to 0, \quad \text{as } m \to \pm\infty. \tag{2}$$

As will be seen from further consideration, the problem (1)–(2) is uniquely solvable for any compactly supported RHS. In Section 4, we describe some practical applications in which formulations like (1) and (2) may appear to be most useful.

We also consider a truncated version of (1)

$$Av_{m+1} + Bv_m = f_{m+1/2}, \quad m = 0, \ldots, M-1, \tag{3}$$

which is defined on a finite one-dimensional mesh, $m = 0, \ldots, M$. This new finite mesh, $m = 0, \ldots, M$, is contained in the original one, $-\infty < m < +\infty$, as a subset. Moreover, we emphasize that the finite mesh $m = 0, \ldots, M$ is chosen so that to entirely contain the support of the RHS, supp $f_{m+1/2} \subset \{m + 1/2 \mid m = 0, \ldots, M-1\}$.

We require that the solution to (3) defined for $m = 0, \ldots, M$ coincides on this finite mesh interval with the corresponding fragment of the solution to (1)–(2). In other words, we require that the solution to (3) found for $m = 0, \ldots, M$ admits such a unique complement to the entire line (i.e., to the infinite mesh) that solves problem (1)–(2). Obviously, to meet this requirement, one needs to set some special boundary conditions for system (3) at the edges of the finite mesh interval, i.e., at $m = 0$ and at $m = M$. We will do that analyzing the behavior of the solution to problem (1)–(2) for $m \geqslant M$ and $m < 0$. We, however, emphasize that problem (1)–(2) itself is not intended for the numerical solution since the corresponding domain of definition is infinite. The system we will actually solve is (3) with such boundary conditions at $m = 0$ and $m = M$ that ensure its equivalence to (1)–(2) on $m = 0, \ldots, M$; the corresponding computational procedure is described in Section 2.

To construct the desired boundary conditions for (3), we first recall that system (1) is homogeneous outside the interval $m = 0, \ldots, M$. Therefore, each solution v_m (which satisfies $Av_{m+1} + Bv_m = 0$ for $m \geqslant M$ and for $m < 0$) can be uniquely represented as a sum of two vector functions, $v_m = v_m^+ + v_m^-$. Either of these vector functions, v_m^+ and v_m^-, is also a solution to the homogeneous part of (1), $Av_{m+1}^+ + Bv_m^+ = 0$ and $Av_{m+1}^- + Bv_m^- = 0$, moreover, $v_m^+ \to 0$ as $m \to \infty$ and $v_m^- \to 0$ as $m \to -\infty$.

Let us now introduce the space C of all complex n-dimensional vectors. This space can be represented as a direct sum of two its subspaces, $C = C^+ \oplus C^-$, $C^+ \cap C^- = \mathbf{0}$. Here, both C^+ and C^- are the eigen subspaces of operator Q, which means that $\forall v^+ \in C^+$ $Qv^+ \in C^+$ and $\forall v^- \in C^-$ $Qv^- \in C^-$. Moreover, $\forall v^+ \in C^+$ $Q^m v^+ \to \mathbf{0}$ as $m \to +\infty$ and $Q^m v^+ \to \infty$ as $m \to -\infty$; analogously, $\forall v^- \in C^-$ $Q^m v^- \to \mathbf{0}$ as $m \to -\infty$ and $Q^m v^- \to \infty$ as $m \to +\infty$ (here, Q^m are the integer powers of Q). The subspaces C^+ and C^- obviously depend on the structure of the spectrum of Q. Namely, C^- is a linear span of all those eigen and adjoint vectors of Q that correspond to the eigenvalues μ_s, $|\mu_s| > 1$; C^+ is a linear span of all those eigen and adjoint vectors of Q that correspond to the eigenvalues μ_s, $|\mu_s| < 1$. (Clearly, all eigen and adjoint vectors of Q as a whole form a full system of n linearly independent vectors because Q is nonsingular, Ker $Q = \mathbf{0}$.) Note, since the spectrum of Q always contains the eigenvalues μ_s, $|\mu_s| < 1$, as well as the eigenvalues μ_s, $|\mu_s| > 1$, none of the subspaces C^+ and C^- is trivial. Clearly, the representation of an arbitrary solution v_m to the homogeneous part of (1) as a sum of two terms, $v_m = v_m^+ + v_m^-$ (see above), actually implies that $v_m^+ \in C^+$ for each m ($m \geqslant M$ or $m \leqslant 0$) and $v_m^- \in C^-$ for each m ($m \geqslant M$ or $m \leqslant 0$).

Let us now recall that the solution to system (1) we are looking for should meet boundary conditions (2). It means that for both semi-infinite intervals of homogeneity, $m \geqslant M$ and $m \leqslant 0$, we have to select only decreasing components from the general representation $v_m = v_m^+ + v_m^-$. In other words, the solution v_m to (1) satisfies (2) iff $v_m \in C^+$ for all $m \geqslant M$ and $v_m \in C^-$ for $m \leqslant 0$, or equivalently, $v_m^- = \mathbf{0}$ for all $m \geqslant M$ and $v_m^+ = \mathbf{0}$ for all $m \leqslant 0$. The latter statement, in particular, implies that boundary conditions (2) are noncontradictory because the subspaces C^+ and C^- are complementary, $C^+ \cap C^- = \mathbf{0}$ and $C^+ \oplus C^- = C$. (Note, the two parts of the solution v_m that correspond to $m \leqslant 0$ and $m \geqslant M$, respectively, are different, both of them are driven by the RHS $f_{m+1/2}$, which is concentrated on $\{m + 1/2 \mid m = 0, \ldots, M-1\}$. Therefore, the requirements $v_m^+ = \mathbf{0}$ for $m \leqslant 0$ and $v_m^- = \mathbf{0}$ for $m \geqslant M$ should not cause misunderstandings since they, of course, do not imply that v_m^+ and v_m^- are also zero for $m \geqslant M$ and $m \leqslant 0$, respectively.)

Clearly, since C^+ and C^- are the eigen subspaces of Q, then the following boundary conditions at $m = 0$,

$$v_0 \in C^-, \tag{4a}$$

and at $m = M$,

$$v_M \in C^+, \tag{4b}$$

should supplement system (3) in order to guarantee that its solution found for $m = 0, \ldots, M$ will coincide on this finite interval with the corresponding fragment of the solution to (1)–(2).

As mentioned above, the subspaces C^+ and C^+ depend on the eigenvalues of matrix Q. One can explicitly determine these subspaces using the following equalities,

$$S^- v_0 = \mathbf{0}, \tag{5a}$$

$$S^+ v_M = \mathbf{0}, \tag{5b}$$

where

$$S^- = \prod_{|\mu_s|>1} (Q - \mu_s I), \tag{6a}$$

$$S^+ = \prod_{|\mu_s|<1} (Q - \mu_s I), \tag{6b}$$

here I is the identity matrix of order n and the matrix products in (6) are calculated in accordance with the multiplicities of eigenvalues. Indeed, it was proven in [6] that the inclusion (4a) is equivalent to the fulfilment of equality (5a) (when S^- is defined in (6a)); analogously, the inclusion (4b) is equivalent to the fulfilment of equality (5b) (when S^+ is defined in (6b)). Note, we do not impose any special restrictions on Q in [6], in particular, Q may not have a basis composed of eigenvectors (or equivalently, may contain Jordan blocks of order more than 1 in its canonical form).

Based on the equivalence of (4) and (5), one can affirm that the solution of problem (3) and (5) and the solution of problem (1)–(2) coincide for $m = 0, \ldots, M$. It is easy to make sure that the problem (3) and (5) is well posed in the sense of [3], i.e.,

$$\max_{0 \leqslant m \leqslant M} \| v_m \| \leqslant c \cdot \max_{0 \leqslant m \leqslant M-1} \| f_{m+1/2} \|, \tag{7}$$

where the constant c does not depend on M. In the next section, we describe an effective numerical algorithm for solving (3) and (5).

2. Numerical algorithm

Let us first introduce the projection operators, P^+ and P^-, onto the subspaces C^+ and C^-, respectively, so that any $v \in C$ is uniquely represented as a sum of two terms, $v = v^+ + v^-$, where $v^+ = P^+ v \in C^+$ and $v^- = P^- v \in C^-$. The existence of such a unique representation for any $v \in C$ is simply a reformulation of the fact that $C = C^+ \oplus C^-$.

Note that both matrices S^- and S^+ (see (6)) are rank-deficient,

$$\operatorname{rank} S^- \overset{\text{def}}{=} n^-, \qquad \operatorname{rank} S^+ \overset{\text{def}}{=} n^+,$$

and clearly $n^- + n^+ = n$. Clearly,

$$\dim C^- = n - n^- = n^+, \qquad \dim C^+ = n - n^+ = n^-.$$

Let us select any n^- linearly independent rows of S^-, orthonormalize them, and complement the resulting set of n^- n-component orthonormal vectors up to the full orthonormal system (of n vectors). For example, one can use any n^+ linearly independent rows of S^+ to construct this complement. Now, the complex conjugate vectors to those which constitute the above complement form the basis in C^-. Let us designate this basis $\{e_1, \ldots, e_{n^+}\}$. Analogously, starting from the orthonormalization of any n^+ linearly independent rows of S^+ and then complementing up to the full system we obtain the basis in C^+ and designate it $\{d_1, \ldots, d_{n^-}\}$. Let

$$v^- = \sum_{i=1}^{n^+} \alpha_i e_i, \qquad v^+ = \sum_{j=1}^{n} \beta_j d_j;$$

we need to find the coefficients α_i, $i = 1, \ldots, n^+$, and β_j, $j = 1, \ldots, n^-$. Since $v^+ + v^- = v$, these coefficients solve the following linear algebraic system of order n

$$\begin{bmatrix} e_1^{(1)} & \ldots & e_{n^+}^{(1)} & d_1^{(1)} & \ldots & d_{n^-}^{(1)} \\ e_1^{(2)} & \ldots & e_{n^+}^{(2)} & d_1^{(2)} & \ldots & d_{n^-}^{(2)} \\ \vdots & & \vdots & \vdots & & \vdots \\ e_1^{(n)} & \ldots & e_{n^+}^{(n)} & d_1^{(n)} & \ldots & d_{n^-}^{(n)} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{n^+} \\ \beta_1 \\ \vdots \\ \beta_{n^-} \end{bmatrix} = \begin{bmatrix} v^{(1)} \\ \vdots \\ v^{(n)} \end{bmatrix}. \tag{8}$$

The coordinate representation of $\{e_1, \ldots, e_{n^+}\}$, $\{d_1, \ldots, d_{n^-}\}$ in (8) is written with respect to the same ("third") basis as the coordinate representation of v. Now, let D_α be a rectangular matrix of n^+ rows and n columns and D_β be a rectangular matrix of n^- rows and n columns, so that the square $(n \times n)$ matrix $\begin{bmatrix} D_\alpha \\ D_\beta \end{bmatrix}$ coincides with the inverse system matrix from (8),

$$\begin{bmatrix} D_\alpha \\ D_\beta \end{bmatrix} = \begin{bmatrix} e_1^{(1)} & \ldots & e_{n^+}^{(1)} & d_1^{(1)} & \ldots & d_{n^-}^{(1)} \\ e_1^{(2)} & \ldots & e_{n^+}^{(2)} & d_1^{(2)} & \ldots & d_{n^-}^{(2)} \\ \vdots & & \vdots & \vdots & & \vdots \\ e_1^{(n)} & \ldots & e_{n^+}^{(n)} & d_1^{(n)} & \ldots & d_{n^-}^{(n)} \end{bmatrix}^{-1}. \tag{9}$$

Then, the operators P^- and P^+ are given by

$$P^- = \begin{bmatrix} e_1^{(1)} & \ldots & e_{n^+}^{(1)} \\ e_1^{(2)} & \ldots & e_2^{(n)} \\ \vdots & & \vdots \\ e_1^{(n)} & \ldots & e_{n^+}^{(n)} \end{bmatrix} D_\alpha, \tag{10a}$$

$$P^+ = \begin{bmatrix} d_1^{(1)} & \ldots & d_{n^-}^{(1)} \\ d_1^{(2)} & \ldots & d_{n^-}^{(2)} \\ \vdots & & \vdots \\ d_1^{(n)} & \ldots & d_{n^-}^{(n)} \end{bmatrix} D_\beta. \tag{10b}$$

It is easy to see from (9)–(10) that $P^{+2} = P^+$ and $P^{-2} = P^-$, i.e., the operators P^+ and P^- are projections. We will use projecting onto the subspaces C^+ and C^- while solving problem (3) and (5).

Let us specify $v_0^l = \mathbf{0}$ (superscript "l" means *left*) and "integrate" (3) from left to right,

$$v_{m+1}'^l = -Qv_m^l + A^{-1} f_{m+1/2}, \quad m = 0, \ldots, M-1, \tag{11a}$$

implementing the projection,

$$v_{m+1}^l = P^+ v_{m+1}'^l \tag{11b}$$

onto the subspace C^+ at each step. Analogously, let $v_M^r = \mathbf{0}$ (superscript "r" means *right*) and "integrate" (3) from right to left,

$$v_m'^r = -Q^{-1}v_{m+1}^r + B^{-1}f_{m+1/2}, \quad m = M-1,\dots,0, \tag{12a}$$

projecting onto C^- at each step,

$$v_m^r = P^- v_m'^r. \tag{12b}$$

Then, we can justify the following:

Proposition 1. *The vector function* $v_m \overset{\text{def}}{=} v_m^r + v_m^l$, $m = 0,\dots,M$, *is the solution to* (3) *that satisfies* (5).

Proof. It is easy to see that if we "integrate" (3) from left to right,

$$\tilde{v}_{m+1}^l = -Q\hat{v}_m^l + A^{-1}f_{m+1/2}, \qquad \tilde{v}_0^l = \mathbf{0},$$

and from right to left,

$$\tilde{v}_m^r = -Q^{-1}\tilde{v}_{m+1}^r + B^{-1}f_{m+1/2}, \qquad \tilde{v}_M^r = \mathbf{0},$$

without projecting onto C^+ and C^-, respectively, at each step, then $\tilde{v}_m^l - v_m^l \in C^-$ (i.e., $v_m^l = P^+\tilde{v}_m^l$) and $\tilde{v}_m^r - v_m^r \in C^+$ (i.e., $v_m^r = P^-\tilde{v}_m^r$). In other words, $v_m'^l$ and $\tilde{v}_m^l$ have the same projection v_m^l onto C^+; analogously, $v_m'^r$ and $\tilde{v}_m^r$ have the same projection v_m^r onto C^-. Let us designate

$$\bar{v}_m^l = \tilde{v}_m^l - v_m^l \in C^-, \qquad \bar{v}_m^r = \tilde{v}_m^r - v_m^r \in C^+$$

and also

$$A^{-1}f_{m+1/2} = g_{m+1/2} = g_{m+1/2}^+ + g_{m+1/2}^-,$$

$$g_{m+1/2}^+ = P^+ g_{m+1/2} \in C^+, \qquad g_{m+1/2}^- = P^- g_{m+1/2} \in C^-,$$

$$B^{-1}f_{m+1/2} = h_{m+1/2} = h_{m+1/2}^+ + h_{m+1/2}^-,$$

$$h_{m+1/2}^+ = P^+ h_{m+1/2} \in C^+, \qquad h_{m+1/2}^- = P^- h_{m+1/2} \in C^-.$$

Then,

$$v_{m+1}^l = -Qv_m^l + g_{m+1/2}^+, \qquad \bar{v}_{m+1}^l = -Q\bar{v}_m^l + g_{m+1/2}^-, \tag{13}$$

$$v_m^r = -Q^{-1}v_{m+1}^r + h_{m+1/2}^-, \qquad \bar{v}_m^r = -Q^{-1}\bar{v}_{m+1}^r + h_{m+1/2}^+. \tag{14}$$

One can easily see that $Ag_{m+1/2}^+ = Bh_{m+1/2}^+$ and $Ag_{m+1/2}^- = Bh_{m+1/2}^-$. Indeed,

$$g_{m+1/2} = A^{-1}f_{m+1/2} = A^{-1}Bh_{m+1/2} = Qh_{m+1/2} = Qh_{m+1/2}^+ + Qh_{m+1/2}^-$$

and, consequently, $g^+_{m+1/2} = Qh^+_{m+1/2}$ and $g^-_{m+1/2} = Qh^-_{m+1/2}$ (since C^+ and C^- are the eigen subspaces of Q), which implies the above statement. Then, multiplying the first of Eqs. (13) by A and the first of Eqs. (14) by B and adding them, we obtain,

$$A\left(v^l_{m+1} + v^r_{m+1}\right) + B\left(v^l_m + v^r_m\right) = Ag^+_{m+1/2} + Bh^-_{m+1/2}$$
$$= Ag^+_{m+1/2} + Ag^-_{m+1/2} = Ag_{m+1/2} = f_{m+1/2},$$

which means that $v_m = v^l_m + v^r_m$ solves (3). Moreover, since $v_0 = v^l_0 + v^r_0 = v^r_0 \in C^-$ and $v_M = v^l_M + v^r_M = v^l_M \in C^+$, the boundary conditions (5) are satisfied. □

Note, the algorithm (11)–(12) for solving problem (3) and (5) can be referred to as a variant of the well-known approach based on the idea of successive substitution. We, however, see that (11)–(12) does not require the final, i.e., "resolving" stage, which is generally relevant to the successive substitution techniques. Instead, we can simply add two quantities, $v^l_m + v^r_m = v_m$, and thus obtain the desirable solution. This is an essential simplification in comparison with the general (matrix) successive substitution, it is accounted for by the special form of boundary conditions (5). The latter are formulated in accordance with the structure of the "growing" and "decaying" eigen subspaces of the operator Q.

We will now address the issue of computational stability of the numerical procedure (11)–(12). Namely, we will show that the spurious (e.g., roundoff) errors arising at all stages of the computational process cannot be accumulated and therefore cannot cause large errors in the final solution. Instead of (11), let us consider a real computational process

$$\begin{aligned}
&\check{v}^l_0 = \lambda^l_0,\\
&\check{v}^l_{m+1} = -Q\check{v}^l_m + A^{-1}f_{m+1/2} + \delta^l_{m+1}, \quad m = 0,\dots,M-1,\\
&\check{v}^l_{m+1} = P^+\check{v}'^l_{m+1} + \lambda^l_{m+1},
\end{aligned} \tag{15}$$

where δ^l_m and λ^l_m are the computational errors involved at each step. Generally, these errors may be caused by the non-precise specification of the RHSs, as well as by the non-precise specification of the entries of matrices. Note, representation (15) is somewhat rough since we assume that the computations are conducted in accordance with the exact formulae and then, the results are altered by introducing the perturbations of order ε,

$$\max_{0\leqslant m\leqslant M} \|\delta^l_m\| \leqslant \varepsilon, \qquad \max_{0\leqslant m\leqslant M} \|\lambda^l_m\| \leqslant \varepsilon.$$

Comparing (11) and (15), one can easily derive:

$$\begin{aligned}
&\check{v}^l_m - v^l_m = \lambda^l_m, \quad m = 0,\\
&\check{v}^l_m - v^l_m = P^+\delta^l_m + \lambda^l_m + \sum_{t=1}^{t=m} (-Q)^t P^+\left(\delta^l_{m-t} + \lambda^l_{m-t}\right), \quad m = 1,\dots,M.
\end{aligned} \tag{16}$$

To obtain (16), we take into account that P^+ is a projection, $P^{+2} = P^+$, and consequently, the operators P^+ and Q commute, $QP^+ = P^+Q$, since C^+ is the eigen subspace of Q; moreover, we formally let $\delta^l_0 = 0$. Recall now that the subspace C^+, which is the image of the projection P^+,

corresponds to those eigenvalues μ_s of the matrix Q that satisfy $|\mu_s| < 1$. Designate $|\mu^l| = \max_{|\mu_s|<1} |\mu_s|$; obviously, $|\mu^l| < 1$. Then,

$$\max_{0 \leqslant m \leqslant M} \| v_m^l - \check{v}_m^l \| \leqslant const \cdot \varepsilon \cdot \| P^+ \| \frac{1 - |\mu^l|^{M+1}}{1 - |\mu^l|}. \tag{17}$$

Here, we use the usual maximum norm for the n-component vectors v, $\| v \| = \max_{1 \leqslant i \leqslant n} |v^i|$. The value of *const* in (17) does not depend on M. It, however, may depend on the structure of the canonical form of Q (presence or absence of the nontrivial Jordan blocks), as well as on the transformation that maps Q from its original to the canonical form. Clearly, since the dimension n of the matrix Q is fixed we can disregard the last-mentioned dependence and from (17) conclude that the numerical process (11) is weakly sensitive to the spurious computational errors.

Analogously, we write instead of (12),

$$\begin{aligned} &\check{v}_M^r = \lambda_M^r, \\ &\check{v}_m'^r = -Q^{-1}\check{v}_{m+1}^r + B^{-1}f_{m+1/2} + \delta_m^r, \quad m = M-1, \ldots, 0, \\ &\check{v}_m^r = P^- \check{v}_m'^r + \lambda_m^r, \end{aligned} \tag{18}$$

and obtain:

$$\begin{aligned} &\check{v}_m^r - v_m^r = \lambda_M^r, \quad m = M, \\ &\check{v}_m^r - v_m^r = P^- \delta_m^r + \lambda_m^r + \sum_{t=1}^{t=M-m} (-Q)^{-t} P^- (\delta_{m+t}^r + \lambda_{m+t}^r), \quad m = M-1, \ldots, 0, \end{aligned} \tag{19}$$

which immediately yields the following estimate,

$$\max_{0 \leqslant m \leqslant M} \| v_m^r - \check{v}_m^r \| \leqslant const \cdot \varepsilon \cdot \| P^- \| \frac{1 - |\mu^r|^{-(M+1)}}{1 - |\mu^r|^{-1}}, \tag{20}$$

here $|\mu^r| = \min_{|\mu_s|>1} |\mu_s|$, $|\mu^r| > 1$. To obtain (20), we take into account that $P^{-2} = P^-$, $Q^{-1}P^- = P^-Q^{-1}$, and assume that the perturbations λ_m^r and δ_m^r are of order ε. Again, the value of *const* in (20) does not depend on M but may depend on Q, the last-named dependence is not essential for our consideration. Recall now that the final solution is given by $v_m = v_m^l + v_m^r$. Designating $\check{v}_m = \check{v}_m^l + \check{v}_m^r$ and combining (17) and (20), we obtain,

$$\max_{0 \leqslant m \leqslant M} \| v_m - \check{v}_m \| \leqslant const \cdot \varepsilon \cdot \left(\| P^+ \| \frac{1 - |\mu^l|^{M+1}}{1 - |\mu^l|} + \| P^- \| \frac{1 - |\mu^r|^{-(M+1)}}{1 - |\mu^r|^{-1}} \right), \tag{21}$$

which implies that in solving problem (3) and (5) by means of (11)–(12) the effect of spurious computational errors does not grow when the number of steps M of the numerical procedure (11)–(12) increases. Therefore, one can successively use the algorithm (11)–(12) for solving the well-posed (see (7)) problem (3) and (5). We additionally emphasize here that the projecting stage of the described algorithm (see (11b) and (12b), respectively) is essential since it enables to always keep the error in the corresponding "decaying" subspace, see (16) and (19). Otherwise, the error may (exponentially) grow and the estimate (21) would not hold.

To conclude this section, we will comment on the question of numerical efficiency of the algorithm (11)–(12). Obviously, either of the two parts of this algorithm, (11) and (12), requires two n-order matrix–vector multiplications per node. Indeed, for practical computing we may combine the stages (11a) and (11b), as well as (12a) and (12b), by calculating in advance the matrix products $\boldsymbol{P}^{+}\boldsymbol{Q}$, $\boldsymbol{P}^{+}\boldsymbol{A}^{-1}$ and $\boldsymbol{P}^{-}\boldsymbol{Q}^{-1}$, $\boldsymbol{P}^{-}\boldsymbol{B}^{-1}$, respectively. Therefore, the entire algorithm costs $O(4Mn^2)$ floating-point operations. We emphasize here that if the dimension n is fixed, then the total amount of computations required for solving problem (3) and (5) by means of (11)–(12) linearly increases with the growth of the number of grid nodes M. On the other hand, system (3) and (5) can generally be thought of as a linear algebraic system of $(M+1)\cdot n$ equations with $(M+1)\cdot n$ unknowns. Therefore, we can implement some elimination procedure to obtain its exact solution. (Note, since (11)–(12) provides an exact solution for problem (3) and (5), any other method for solving (3) and (5) that we are going to compare with (11)–(12) should not be iterative but should also provide an exact solution.) Clearly, an original Gauss method being applied to problem (3) and (5) would require $O((M\cdot n)^3)$ operations and therefore, would be essentially less effective than (11)–(12). We, however, know that the best direct methods for solving linear algebraic systems may generally require as little as $O(N\cdot \ln N)$ operations, here N is the dimension of the system (in our case $N=(M+1)\cdot n$). Such methods usually take into account the sparseness of the system matrix and/or its specific structure. In our case, we actually consider a $(M+1)\cdot n \times (M+1)\cdot n$ square matrix, which has two nonzero block diagonals composed of the standard $n\times n$ square blocks with somewhat non-standard upper-left and lower-right corners. Though we do not know if there is an exact $O(N\cdot \ln N)$ algorithm for solving the linear algebraic systems of the type (3) and (5), such an algorithm would at any rate be asymptotically (for large M and fixed n) less effective than the process (11)–(12) since the latter is linear with respect to M. Note, linear dependence of the required amount of computations on the dimension of the mesh is relevant to the scalar (one-dimensional) successive substitution techniques.

3. More general formulation

We now admit that some eigenvalue(s) μ_s of the matrix $\boldsymbol{Q}$ may have an absolute magnitude equal to unity, $|\mu_s| = 1$. If such an eigenvalue is not multiple, then we may expect system (1) would have a constant or oscillating eigensolution on its intervals of homogeneity. Otherwise (i.e., when an eigenvalue μ_s, $|\mu_s| = 1$, is multiple), the behavior of the corresponding eigensolutions depends on the structure of the set of eigenvectors of $\boldsymbol{Q}$. We will further restrict ourselves by considering only that case when there exist as many linearly independent eigenvectors corresponding to the multiple eigenvalue μ_s, $|\mu_s| = 1$, as its multiplicity is. In other words, we do not admit nontrivial Jordan blocks with the unitary (by the absolute magnitude) entries on the main diagonal in the canonical form of $\boldsymbol{Q}$. Note, we do not generally require that $\boldsymbol{Q}$ always has a full system of linearly independent eigenvectors (i.e., n linearly independent eigenvectors); the requirement of having as many different eigenvectors as the multiplicity of the corresponding eigenvalue is applies only to those μ_s that $|\mu_s| = 1$. Later on, in Section 4, we will see that the generalized formulation we are considering here actually originates from some important practical applications. In the meantime, we only note that

admitting the nontrivial Jordan blocks that correspond to the multiple eigenvalues μ_s, $|\mu_s| = 1$, (if any) does not, generally speaking, lead to any essential complications; we simply avoid this situation since we do not have it in the computational practice.

Since we henceforth assume that all those eigenvalues μ_s of Q, for which $|\mu_s| = 1$, have as many linearly independent eigenvectors as their multiplicity is, then the corresponding eigensolutions of (1) (on the semi-infinite intervals of homogeneity, $m \geqslant M$ and $m \leqslant 0$) are either constant or oscillatory but always bounded. (Otherwise, we would have the polynomially growing eigensolutions.) Therefore, boundary conditions (2) can no longer be satisfied; instead, we will require

$$\|v_m\| \leqslant const, \quad \text{as } m \to \pm\infty. \tag{22a}$$

More specifically, we can always require that the solution to (1) vanishes as $m \to -\infty$ and be bounded as $m \to +\infty$, i.e.,

$$\begin{aligned} &\|v_m\| \to 0, && \text{as } m \to -\infty, \\ &\|v_m\| \leqslant const, && \text{as } m \to +\infty. \end{aligned} \tag{22b}$$

The reason for such an asymmetry originates from practical applications, it is delineated in [1] and will also be addressed in Section 4. For the time being, we simply point out that problem (1) and (22b) is uniquely solvable; in case there are actually no eigenvalues μ_s, $|\mu_s| = 1$, in the spectrum of Q the formulation (1) and (22b) still remains valid and becomes equivalent to (1)–(2). Note, there are, of course, other possible correct formulations of the infinite-mesh problem for the case when the spectrum of Q contains the eigenvalue(s) μ_s, $|\mu_s| = 1$.

We now reformulate our definitions of the spaces C^+ and C^-. Again, we require that $C^+ \cap C^- = \mathbf{0}$, $C^+ \oplus C^- = C$. The space C^-, in fact, remains the same as before, i.e., it is the "decaying to the left" eigen subspace of Q. The space C^+ is also the eigen subspace of Q but we now weaken our above condition and only require that the positive (integer) powers of Q applied to any $v^+ \in C^+$ always remain bounded. Clearly, C^- is still a linear span of all those eigen and adjoint vectors of Q that correspond to the eigenvalues μ_s, $|\mu_s| > 1$; C^+ is now a linear span of all those eigen and adjoint vectors of Q that correspond to the eigenvalues μ_s, $|\mu_s| \leqslant 1$.

Then, we construct a finite-mesh analogue to problem (1) and (22b) (as done in Section 1). More precisely, we supplement system (3) by special boundary conditions at $m = 0$ and at $m = M$ that would ensure that the solution of this system found on the finite mesh coincides for $m = 0, \ldots, M$ with the corresponding fragment of the solution to (1) and (22b). Such boundary conditions for the finite-mesh problem may again be formulated in the form (5). However, we have to give here a new definition of the matrix S^+. Namely, instead of (6) we now write

$$S^- = \prod_{|\mu_s| > 1} (Q - \mu_s I), \tag{23a}$$

$$S^+ = \prod_{|\mu_s| \leqslant 1} (Q - \mu_s I). \tag{23b}$$

Note, equalities (6a) and (23a) are the same; equalities (6b) and (23b) are different.

The new finite-mesh problem, i.e., problem (3), (5), and (23), remains well-posed, however, the value of c in (7) can no longer be independent of M. In our specific case, when Q has the eigenvalue(s) μ_s, $|\mu_s| = 1$, but does not have the corresponding nontrivial Jordan blocks in its canonical form, it is possible to show that c becomes proportional to M, $c = \tilde{c} \cdot M$.

The algorithm for solving (3), (5), and (6) described in Section 2, i.e., the computational process (11)–(12), applies to solving (3), (5), and (23) without any formal changes, except for the new definition of S^+ (compare (23b) to (6b)). All estimates of the numerical efficiency of this algorithm (see Section 2) also remain valid. The only natural difference occurs in justification of the computational stability of the process (11)–(12) being applied to solving (3), (5), and (23). Namely, instead of (17) we now have a new estimate,

$$\max_{0 \leqslant m \leqslant M} \| v_m^l - \check{v}_m^l \| \leqslant const_1 \cdot \varepsilon \cdot \| P^+ \| \left(\frac{1 - |\mu^l|^{M+1}}{1 - |\mu^l|} + const_2 \cdot M \right). \tag{24}$$

Here, the definition of μ^l remains the same as above, inequality (24) simply takes into account that there are also μ_s, $|\mu_s| = 1$. To obtain the final estimate of type (21), we should now combine (20) and (24), which yields,

$$\max_{0 \leqslant m \leqslant M} \| v_m - \check{v}_m \| \leqslant \varepsilon \cdot (c_1 + c_2 \cdot M). \tag{25}$$

The quantities c_1 and c_2 in (25) are constants, which depend on $\| P^+ \|$, $\| P^- \|$, $|\mu^l|$, $|\mu^r|$ but do not depend on M. Estimate (25) naturally agrees with the above-made comment on the dependence of c on M in the (7)-type inequality for the case when the eigenvalue(s) μ_s, $|\mu_s| = 1$, exist(s) in the spectrum of Q.

To conclude this section, we would like to note that the idea of calculating the "growing" and "decaying" subspaces for a homogeneous one-dimensional system and then projecting the nonhomogeneous contribution onto the corresponding subspace was effectively employed by Godunov in [2]. Our algorithm, however, markedly differs from Godunov's orthogonal successive substitution (see [2]) since we essentially use the constancy of coefficients, as well as the special form of boundary conditions (5) and (6) or (5) and (23).

4. Practical origins and applications

The one-dimensional problems like (3), (5), and (6) and/or (3), (5), and (23) arise in practical computations when constructing artificial boundary conditions (ABCs) for the numerical solution of boundary-value problems initially formulated on infinite domains. These boundary conditions are typically set at the external (artificial) boundary of the finite computational domain once the latter is obtained from the original unbounded domain by means of truncation. Implementation of the ABCs enables completing the "truncated problem" and, therefore, making it available for solution on the computer.

The issue of setting the ABCs presents a particular interest in computational fluid dynamics, where a lot of interesting problems have an external nature (e.g., a finite body or system of bodies immersed in an unbounded fluid flow). In the previous work [6–8], we have constructed the ABCs for computation of a certain class of the external compressible viscous flows. The

ABCs [6–8] provide high accuracy of computations and can essentially improve the robustness of the entire numerical procedure since, contrary to many other techniques, these boundary conditions may basically be constructed as close to the exact ones as desired. In other words, the ABCs [6–8] meet the following fundamental requirement. *One should be able to uniquely complement the solution calculated inside the finite computational domain to its infinite exterior so that the original problem is solved within the prescribed accuracy.*

To construct the ABCs [6–8], we first linearize the governing (Navier–Stokes) equations in the far field (against the constant free-stream background) and then implement the Difference Potentials Method (DPM) [4,5]. Clearly, after the linearization we obtain a homogeneous system of linear PDEs with constant coefficients; the domain of definition for this linear system is an infinite exterior to the computational domain. The central idea of the approach [6–8] is to equivalently replace the linearized exterior problem by a certain operator equation formulated at the artificial boundary. The last-named equation may be referred to as being partially analogous to the Calderon boundary pseudodifferential equations [1]. Once appropriately resolved, this equation provides highly accurate ABCs for practical computing. The principle element of the entire construction is a special auxiliary problem (AP) whose Green's operator plays in our consideration the role analogous to the role of convolution with the fundamental solution in the classical potential theory. Namely, the solution of the AP is used for constructing the generalized potential [4,5]. Then, the solution to the exterior linear problem is represented in the form of this generalized potential and the above-mentioned operator equation is written with respect to the density of the generalized potential.

The AP is first formulated on the entire plane for the nonhomogeneous counterpart of the constant-coefficients system of PDEs obtained after the linearization, the RHS for the AP is always compactly supported. We require that the solution to the AP has the same behavior at infinity as the behavior of the solution to the original problem (linearized in the far field). Specifically, we require that all perturbations vanish at infinity, which is natural for the external viscous flows. Then, we truncate the AP and consider a new formulation of this problem on some rectangular domain; the solution of the new AP is in a certain sense (asymptotically) close to the solution of the original AP. Our algorithm for constructing the ABCs requires a repeated solution of the new AP, which is done by means of variables separation. In doing so, we implement the Fourier transform in the cross-stream direction and obtain a problem like (3), (5), and (23) for each wavenumber. Note, the asymmetry in boundary conditions (22b) (i.e., in boundary conditions (5) and (23)) is accounted for by the physical asymmetry: we require a strict decay of each Fourier mode upstream and admit selected non-decaying (but bounded) modes downstream (see [6] for more details). In the process of calculating the ABCs (i.e., calculating the matrices of the operators involved), problem (3), (5), and (23) should be solved many times (for different system matrices and for different RHSs), therefore, the issue of a fast numerical algorithm for solving such problems appears to be most important. Our computational experience corroborates the efficiency of the process (11)–(12) as applied to solving (3), (5), and (23).

Finally, we note that the above-mentioned approach for constructing the ABCs (see [6–8]) may, generally speaking, have an essentially more wide domain of applications than only external flow problems. Indeed, for many physical formulations, the consideration of a linear problem (at least) in the far field is reasonable. Since any constant-coefficients system of PDEs

admits variable separation in the Cartesian coordinates, we can appropriately modify the technique [6–8] for handling these problems as well. Of course, such a modification will be based on another governing system, however, the final problem which requires the actual numerical solution will have a form similar to (3), (5), and (6) or (3), (5), and (23). Therefore, the computational algorithm described above (see Section 2) can be applied to a new problem without any changes (the difference will be "hidden" in the structure of system matrices).

Acknowledgements

The work of the first author was partially supported by the Russian Fund for Fundamental Research, project no. 93-012-859.

The work of the second author started when he held a Postdoctoral Fellowship at the School of Mathematical Sciences, Tel-Aviv University, and was accomplished when he held a National Research Council Research Associateship at the NASA Langley Research Center.

References

[1] A.P. Calderon, Boundary-value problems for elliptic equations, in: *Proceedings Soviet-American Conference on the Partial Differential Equations at Novosibirsk* (Moscow, Fizmatgiz, 1963) 303–304.

[2] S.K. Godunov, A method of orthogonal successive substitution for the solution of systems of difference equations, *U.S.S.R. Comput. Math. and Math. Phys.* 2 (1963) 1151–1165.

[3] V.S. Ryaben'kii, Necessary and sufficient conditions for good definition of boundary value problems for systems of ordinary difference equations, *U.S.S.R. Comput. Math. and Math. Phys.* 4 (1964) 43–61.

[4] V.S. Ryaben'kii, Boundary equations with projections, *Rus. Math. Surveys* 40 (1985) 147–183.

[5] V.S. Ryaben'kii, *Difference Potentials Method for Some Problems of Continuous Media Mechanics* (Moscow, Nauka, 1987) (in Russian).

[6] V.S. Ryaben'kii and S.V. Tsynkov, Artificial boundary conditions for the numerical solution of external viscous flow problems, *SIAM J. Numer. Anal.* 32 (5) (1995).

[7] S.V. Tsynkov, An application of nonlocal external conditions to viscous flow computations *J. Comput. Phys.* 116 (1995) 212–225.

[8] S.V. Tsynkov, E. Turkel and S. Abarbanel, External flow computations using global boundary conditions, AIAA Paper 95-0562, 33rd AIAA Aerospace Sciences Meeting, Reno, NV (1995); also: *AIAA J.* (to appear).