Computing 20, 229—240 (1978)

Computing

A Numerical Method for Solving Parabolic Equations With Opposite Orientations

W. Hackbusch, Köln

Received August 4, 1977

Abstract — Zusammenfassung

A Numerical Method for Solving Parabolic Equations With Opposite Orientations. The solution of parabolic control problems is characterized by a system of two equations parabolic with respect to opposite orientations. In this paper a fast iterative method for solving such problems is proposed.

Ein numerisches Verfahren zur Lösung parabolischer Differentialgleichungen mit entgegengesetzter Orientierung. Die Lösung parabolischer Kontrollprobleme wird durch ein System von zwei Differentialgleichungen charakterisiert, die bezüglich entgegenlaufender Richtungen parabolisch sind. In diesem Beitrag wird ein schnelles Iterationsverfahren zur Lösung derartiger Probleme vorgeschlagen.

1. Introduction

The boundary value problem for a system of two parabolic equations with opposite orientations is formulated and discussed in Chapter 2. Such problems, e.g., arise from control problems governed by parabolic differential equations. An important tool of solving this problem is the decoupling technique of Chapter 3. We describe decoupling and Gauß-Seidel iteration in Chapter 4, since elements of both processes are included in our iteration method.

While decoupling is a direct method requiring much computational work, Gauß-Seidel iterations are simply performed but in general do not converge. Introducing a partition of the time interval we apply Gauß-Seidel iterations for local problems. For the connection of the local solutions we need auxiliary matrices that are related to those of the decoupling method. However, there is an important difference between the decoupling and our method. The former requires an exact computation of the matrices, whereas for the latter only approximate matrices are necessary.

The scheme of this method is explained in Chapter 5. For finding suitable initial and final values the rest of Chapter 5 prepares an expansion which is valid for a local application of Gauß-Seidel iteration.

0010-485X/78/0020/0229/$ 02.40

In Chapter 6 auxiliary operators are constructed that allow us to complete the algorithm. The Theorem shows that the rate of convergence is as fast as desired provided that the partition is fine enough. Formulating the algorithm for the discrete case, we investigate in Chapter 7 how the computational work can be reduced.

Numerical examples are reported in Chapter 8.

2. The Problem

Let $L(t)$ denote an elliptic differential operator of second order defined on $H_B^2 \subset H^2(\Omega)$[1], $\Omega \subset \mathbb{R}^r$, $r \geq 1$, with coefficients depending on $t \in [0, T]$. The homogeneous boundary values are assumed to be incorporated into the subset H_B^2 of the Sobolev space $H^2(\Omega)$. E.g., $u \in H_B^2 = H_0^2(\Omega)$ = closure of $H^2(\Omega)$-functions with support in Ω is used for Dirichlet boundary values $u|_{\partial\Omega} = 0$.

Consider the system

$$\frac{d}{dt} y(t) = L(t)\, y(t) - D(t)\, p(t) + f(t) \qquad (2.1\text{ a})$$

$$(0 < t < T;\ y(t), p(t) \in H_B^2).$$

$$\frac{d}{dt} p(t) = -L^*(t)\, p(t) - E(t)\, y(t) + g(t) \qquad (2.1\text{ b})$$

$L^*(t)$ is the adjoint operator of $L(t)$; $D(t)$ and $E(t)$ are bounded operators of $L_2(\Omega)$ into $L_2(\Omega)$; $f(t)$ and $g(t)$ belong to $L_2(\Omega)$. D, E, f, g are assumed to be Hölder continuous with respect to t (cf. [5]).

In addition, *initial* and *final values* are prescribed for $t=0$ and $t=T$,

$$y(0) = B_0\, p(0) + b_0, \qquad (2.2\text{ a})$$

$$p(T) = A_T\, y(T) + a_T, \qquad (2.2\text{ b})$$

where $b_0, a_T \in L_2(\Omega)$, while B_0 and A_T are bounded operators of $L_2(\Omega)$ into $L_2(\Omega)$.

In the sequel we assume that D, E, A_T, B_0 are self-adjoint, e.g., these operators may represent multiplication by a real-valued function. Nevertheless, the following considerations hold for arbitrary D, E, A_T, B_0 and for any elliptic operators L_1, L_2 $(L_1 \neq L_2^*)$ instead of L, L^*, if the existence of a solution of (2.1, 2.2) is ensured. Moreover, each equation in (2.1) may represent a parabolic system of vector functions y, p.

The existence of a solution of (2.1, 2.2) requires additional assumptions. E.g., an unique solution exists if A_T, $-B_0$, D and E are positive (or negative) semi-definite. These properties are fulfilled for all systems arising from optimal control problems of systems governed by a parabolic equation $dy/dt = L(t)\, y + B(t)\, u(t) + f$ (u: *control*) (cf. Lions [8, 9] or [5]).

[1] For the notation of $H^m(\Omega)$, $H_0^m(\Omega)$ cf. [9]. Ω is assumed to be sufficiently smooth. Otherwise, the spaces $H^2(\Omega)$ and $L_2(\Omega)$ are to be replaced by $H^1(\Omega)$ and $H^{-1}(\Omega)$.

Another problem leading to (2.1, 2.2) is the periodic parabolic equation $du/dt = Lu + f$, $u(0) = u(2T)$ (cf. [5]). In this case a unique solution exists if $L = L^*$ is negative definite.

3. Decoupling of the System

As known from ordinary differential equations (cf., e.g., Taufer [13]), the boundary value problem of type (2.1, 2.2) can be decoupled by introducing new unknowns which are solutions of initial value problems. Seeking operators $A(t)$, $B(t)$: $L_2(\Omega) \to L_2(\Omega)$ and functions $a(t), b(t) \in L_2(\Omega)$ satisfying

$$p(t) = A(t)\,y(t) + a(t), \quad y(t) = B(t)\,p(t) + b(t) \quad (0 \le t \le T), \tag{3.1}$$

we obtain evolution equations for A, B, a, b (cf. Da Prato [3], Temam [14], [5]):

$$\frac{dA}{dt} = -(L^*A + AL) + ADA - E, \qquad A(T) = A_T, \tag{3.2}$$

$$\frac{dB}{dt} = BL^* + LB - D + BEB, \qquad B(0) = B_0, \tag{3.3}$$

$$\frac{da}{dt} = -(L^* - AD)\,a + g - Af, \qquad a(T) = a_T, \tag{3.4}$$

$$\frac{db}{dt} = (L + BE)\,b + f - Bg, \qquad b(0) = b_0, \tag{3.5}$$

where $A_T, B_0, a_T, b_0, D, E, f, g, L, L^*$ are the coefficients of (2.1, 2.2).

Let $U(t, s)$ $(0 \le s \le t \le T)$ be the *solution operator* (*fundamental solution*) of $dy/dt = L(t)\,y$, i.e. $y(t) = U(t, s)\,y(s)$ or

$$\frac{d}{dt} U(t, s) = L(t)\,U(t, s), \quad U(s, s) = I \quad (0 \le s \le t \le T) \tag{3.6}$$

(cf. e.g. Ladas and Lakshmikantham [6]). Using the properties of $U(t, s)$ we are able to construct the solutions A, B of the operator Riccati equations (3.2, 3.3) by means of a nonlinear Volterra integral equation (cf. [5]). The solution $B(t)$ is bounded in some interval $[0, t_0]$ $(t_0 > 0)$. For positive t the operator $B(t)$ maps $L_2(\Omega)$ into H_B^2 and is strongly differentiable. Furthermore, $A(t)$ and $B(t)$ are self-adjoint.

In the sequel we assume

$$\sup_{0 \le t \le T} \{\|A(t)\|, \|B(t)\|, \|[I - A(t)B(t)]^{-1}\|\} < \infty \quad (\|\cdot\| = \|\cdot\|_{L_2(\Omega) \to L_2(\Omega)}). \tag{3.7}$$

Therefore, the solution of (2.1, 2.2) can be represented by

$$y(t) = [I - B(t)A(t)]^{-1}[b(t) + B(t)a(t)], \; p(t) = [I - A(t)B(t)]^{-1}[a(t) + A(t)b(t)]. \tag{3.8}$$

If A_T, $-B_0, D, E$ are positive semi-definite (cf. Chapter 2), then $A(t)$ and $-B(t)$ are positive semi-definite and uniformly bounded for all $t \leq T$ ($t \geq 0$ resp.). Moreover, $\| [I - AB]^{-1} \| \leq 1 + \| A \| \; \| B \|$ holds (cf. [5]).

4. Usual Numerical Methods

In the case of control problems, where (2.1, 2.2) corresponds to a variational problem, special numerical methods are applicable (e.g., compare Leroy [7], Casti and Ljung [1]).

The *decoupling technique* can be used in several ways. We may start with a discretization of (2.1, 2.2). Then the system of difference equations admits a decoupling into initial value problems analogous to the treatment of the differential equation in Chapter 3 (cf. Nedelec [11]). The operators $A(t)$, $B(t)$ $(0 \leq t \leq T)$ are to be replaced by $(N \times N)$-matrices $A_h(t)$, $B_h(t)$ ($t/\Delta t$ integer), where $h = (\Delta t, \Omega_h)$ indicates the "time-step" Δt and the grid $\Omega_h \subset \Omega$ used for the discretization of (2.1, 2.2). N is the number of grid points belonging to Ω_h. If, e.g., a regular grid of step size Δx is chosen, N is proportional to Δx^{-r} $(r: \Omega \subset \mathbb{R}^r)$. After computing and storing $A_h(t)$ and $a_h(t)$ for $t = T, T - \Delta t, \ldots, 0$, we are able to solve the discretized initial value problem (2.1 a, 2.2 a) by eliminating p_h $(= A_h y_h + a_h)$. Because of the Riccati type the computation of A_h or B_h involves a multiplication of full matrices. Thus, the computational work amounts to const $\cdot \Delta t^{-1} \cdot \Delta x^{-3r}$.

For a second approach one may discretize (2.1 a, 2.2 a) as well as (3.2, 3.4). Then p_h is obtained by $p_h = A_h y_h + a_h$.

The method we propose will need auxiliary matrices, too. But we want to use these matrices only for generating a fast iteration method. Therefore, only approximations of these matrices are needed. This fact can be utilized for reducing the computational work. Our method includes elements of the decoupling technique as well as an application of the following iteration.

The *Gauß-Seidel iteration method* requires no matrix operations. We start with some $y^0(t)$ and integrate in some discretized form

$$\frac{d}{dt} p^\mu = -L^* p^\mu - E y^{\mu-1} + g \qquad (\mu \geq 1), \tag{4.1}$$

$$\frac{d}{dt} y^\mu = L y^\mu - D p^\mu + f \qquad (\mu \geq 1) \tag{4.2}$$

for the initial data

$$p^\mu(T) = A_T y^{\mu-1}(T) + a_T, \quad y^\mu(0) = B_0 p^\mu(0) + b_0$$

(cf. Miellou [10], Leroy [7]).

5. Gauß-Seidel Iteration

The very simple iteration (4.1, 4.2) converges only if the coupling of p and y (by A_T, B_0, D, E) is weak enough. This is ensured if the t-interval is small enough and if the initial and final data are prescribed in a suitable form. Therefore, we introduce a partition $\Delta := \{0 = t_0 < t_1 < \ldots < t_n = T\}$ $(n \geq 1)$ of $[0, T]$ and write $|\Delta| := \max_{\nu} (t_\nu - t_{\nu-1})$.

Analogously to the multiple shooting method for ordinary two-point boundary value problems (cf. Stoer and Bulirsch [12]), we may solve (4.1, 4.2) in all intervals $[t_{\nu-1}, t_\nu]$ for some initial and final data at t_ν $(0 \leq \nu \leq n)$. The solution is discontinuous at t_ν. Therefore, we have to compute a correction which can be obtained from the residuals by means of decoupling operators (or matrices resp.). Assume that approximate values of these matrices are known. Then we may proceed as follows: In a primary iteration the initial and final data at t_ν $(1 \leq \nu \leq n)$ are improved, while during a secondary iteration (4.1, 4.2) is applied.

We reject the method sketched above because of the two stacked iterations. The desired iteration method should be of the following form:

1. Choose any starting function $y^0(t)$, e.g., $y^0(t) = 0$;
2. for $\nu = n, n-1, \ldots, 1$ solve (4.1) in $[t_{\nu-1}, t_\nu]$ with a suitable final value $p^\mu(t_\nu) = \bar{p}^\mu_\nu$;
3. for $\nu = 1, 2, \ldots, n$ solve (4.2) in $[t_{\nu-1}, t_\nu]$ with a suitable initial value $y^\mu(t_{\nu-1}) = \bar{y}^\mu_\nu$;
4. go to 2.; (5.1)

where $\bar{p}^\mu_\nu$, $\bar{y}^\mu_\nu$ are to be chosen in such a manner that all information obtained from the previous steps are included.

In the remaining part of this paragraph the Gauß-Seidel iterations will be analysed. Without loss of generality consider the homogeneous equations (4.1, 4.2) restricted to the first subinterval $[t_0, t_1]$:

$$\frac{d}{dt} \pi^\mu = -L^* \pi^\mu - E \eta^{\mu-1}, \; \pi^\mu(t_1) = \bar{\pi}^\mu \qquad (\mu \geq 1), \tag{5.2}$$

$$\frac{d}{dt} \eta^\mu = L \eta^\mu - D \pi^\mu, \qquad \eta^\mu(t_0) = \bar{\eta}^\mu \qquad (\mu \geq 1). \tag{5.3}$$

The solutions $\pi^\mu(t)$ and $\eta^\mu(t)$ depend linearly on η^0 and $\bar{\pi}^\chi$, $\bar{\eta}^\chi$ $(1 \leq \chi \leq \mu)$:

$$\pi^\mu(t) = \alpha_0(t)\, \bar{\pi}^\mu + \sum_{\chi=1}^{\mu-1} \left(\alpha_{\mu-\chi}(t)\, \bar{\pi}^\chi + \beta_{\mu-\chi}(t)\, \bar{\eta}^\chi\right) + \gamma_\mu(t)\, \eta^0 \qquad (t_0 \leq t \leq t_1). \tag{5.4}$$

A similar expansion holds for $\eta^\mu(t)$. Using the solution operator $U(t, s)$ of (3.6), we construct integral operators α, β by

$$(\alpha\eta)(t) = \int_t^{t_1} U(s, t)^* E(s)\, \eta(s)\, ds, \quad (\beta\pi)(t) = -\int_{t_0}^{t} U(t, s)\, D(s)\, \pi(s)\, ds.$$

Then the coefficients of (5.4) are

$$\alpha_\chi = (\alpha\,\beta)^\chi\, U(t_1,\cdot)^*, \quad \beta_\chi = \alpha\,(\beta\,\alpha)^{\chi-1}\, U(\cdot, t_0), \quad \gamma_\mu = \beta\,(\gamma\,\beta)^{\mu-1}.$$

Thus, the leading terms [2] of $\pi^\mu(t_0)$ and $\eta^\mu(t_1)$ are

$$\pi^\mu(t_0) = \mathscr{U}^*\,\bar\pi^\mu + \mathscr{E}\,\bar\eta^{\mu-1} + \mathscr{F}^*\,\bar\pi^{\mu-1} + \dots \qquad (\mu \ge 2), \tag{5.5 a}$$

$$\eta^\mu(t_1) = \mathscr{U}\,\bar\eta^\mu + \mathscr{D}\,\bar\pi^{\mu-1} + \mathscr{F}\,\bar\eta^{\mu-1} + \dots \qquad (\mu \ge 2), \tag{5.5 b}$$

where

$$\begin{aligned}
\mathscr{U} &= U(t_1, t_0),\\
\mathscr{D} &= \mathscr{D}^* = -\int_{t_0}^{t_1} U(t_1, s)\, D(s)\, U(t_1, s)^*\, ds,\\
\mathscr{E} &= \mathscr{E}^* = \int_{t_0}^{t_1} U(s, t_0)^*\, E(s)\, U(s, t_0)\, ds,\\
\mathscr{F} &= -\iint_{t_0 \le s \le \sigma \le t_1} U(t_1, s)\, D(s)\, U(\sigma, s)^*\, E(\sigma)\, U(\sigma, t_0)\, d\sigma\, ds.
\end{aligned} \tag{5.6}$$

The convergence of (5.2, 5.3) for fixed $\bar\pi^\mu = \bar\pi^1$, $\bar\eta^\mu = \bar\eta^1$ follows from the estimates

$$\|\alpha_\mu\| \le (K_D\, K_E)^\mu\, K_U^{2\mu+1}\, (t_1 - t_0)^{2\mu}; \quad \|\beta_\mu\|, \|\gamma_\mu\| \le K_E^\mu\, K_D^{\mu-1}\, K_U^{2\mu}\, (t_1 - t_0)^{2\mu-1} \tag{5.7}$$

if $t_1 - t_0 < 1/(\sqrt{K_D\, K_E}\, K_U)$ (cf. [5]). K_D, K_E and K_U are bounds for D, E and U.

6. Determination of Initial and Final Data

In (5.6) the operators $\mathscr{U}$, $\mathscr{D}$, $\mathscr{E}$, $\mathscr{F}$ are constructed for the first interval $[t_0, t_1]$. Generally we denote the operators corresponding to $[t_{\nu-1}, t_\nu]$ by $\mathscr{U}_\nu$, $\mathscr{D}_\nu$, $\mathscr{E}_\nu$, $\mathscr{F}_\nu$ $(1 \le \nu \le n)$.

Note: *It is convenient to compute* $\mathscr{U}_\nu$, $\mathscr{D}_\nu$, $\mathscr{E}_\nu$, $\mathscr{F}_\nu$ *by*

$$\begin{aligned}
&\mathscr{U}(t_{\nu-1}) = I, && d\mathscr{U}(t)/dt = L(t)\,\mathscr{U}(t), \quad \mathscr{U}_\nu := \mathscr{U}(t_\nu),\\
&\mathscr{D}(t_{\nu-1}) = 0, && d\mathscr{D}(t)/dt = L(t)\,\mathscr{D}(t) + \mathscr{D}(t)\, L^*(t) - D(t), \quad \mathscr{D}_\nu := \mathscr{D}(t_\nu),\\
&\mathscr{F}(t_{\nu-1}) = 0, && d\mathscr{F}(t)/dt = L(t)\,\mathscr{F}(t) + \mathscr{D}(t)\, E(t)\,\mathscr{U}(t), \quad \mathscr{F}_\nu := \mathscr{F}(t_\nu),\\
&\mathscr{E}(t_\nu) = 0, && d\mathscr{E}(t)/dt = -L(t)^*\,\mathscr{E}(t) - \mathscr{E}(t)\, L(t) - E(t), \quad \mathscr{E}_\nu := \mathscr{E}(t_{\nu-1}).
\end{aligned} \tag{6.1}$$

The equations for $\mathscr{U}$, $\mathscr{D}$, $\mathscr{F}$ *can be solved simultaneously.*

Let $A_n := A_T$ and B_0 be the operators appearing in (2.2). The auxiliary operators A_ν, B_ν, c_ν, d_ν $(1 \le \nu \le n)$ are defined recursively by

$$\begin{aligned}
&c_\nu = (\mathscr{U}_\nu + \mathscr{F}_\nu)^*\, (I - A_\nu\, \mathscr{D}_\nu)^{-1}, \quad d_\nu = (\mathscr{U}_\nu + \mathscr{F}_\nu)\, (I - B_{\nu-1}\, \mathscr{E}_\nu)^{-1},\\
&A_{\nu-1} = \mathscr{E}_\nu + c_\nu\, A_\nu\, (\mathscr{U}_\nu + \mathscr{F}_\nu), \quad B_\nu = \mathscr{D}_\nu + d_\nu\, B_{\nu-1}\, (\mathscr{U}_\nu + \mathscr{F}_\nu)^*.
\end{aligned} \tag{6.2}$$

[2] The restriction to three leading terms is arbitrary. Taking into account more terms we improve the rate of convergence, but the method would require more complicated formulae in Chapter 6.

Lemma (cf. [5]): *A_ν, B_ν $(0 \leq \nu \leq n)$ are self-adjoint. A_ν and B_ν differ from $A(t_\nu)$ and $B(t_\nu)$ by $0(|\Delta|^2)$. Moreover*

$$c_{\nu\rho} = c_{\nu+1}\, c_{\nu+2} \dots c_\rho \; (\nu \leq \rho), \quad d_{\nu\rho} = d_\nu\, d_{\nu-1} \dots d_{\rho+1} \; (\nu \geq \rho) \tag{6.3}$$

approximate the solution operators of (3.4), (3.5) *by the same order. Since $\mathscr{D}_\nu = 0(|\Delta|)$ and $\mathscr{E}_\nu = 0(|\Delta|)$ hold, c_ν and d_ν are well-defined for a sufficiently fine partition Δ (or if A_ν, $-\mathscr{D}_\nu$, $-B_\nu$, $\mathscr{E}_\nu$ are semi-definite).*

At the intermediate values t_ν additional *transition conditions* are to be fulfilled:

$$(y - B_\nu\, p)|_{t_\nu + 0} = (y - B_\nu\, p)|_{t_\nu - 0}, \quad (p - A_\nu\, y)|_{t_\nu + 0} = (p - A_\nu\, y)|_{t_\nu - 0} \quad (1 \leq \nu < n). \tag{6.4}$$

(6.4) ensures continuity if $I - A_\nu B_\nu$ is a bijection. This holds by (3.7) and the previous Lemma, if $|\Delta|$ is small enough (or if A_ν and $-B_\nu$ are semi-definite).

In (5.1) the functions $\bar{y}_\nu^\mu$, $\bar{p}_\nu^\mu$ are introduced. The magnitudes[3]

$$\hat{y}_\nu^\mu = y^\mu(t_\nu - 0) + \mathscr{F}_\nu\, (\bar{y}_\nu^\mu - \bar{y}_\nu^{\mu-1}), \quad \check{p}_\nu^\mu = p^\mu(t_{\nu-1} + 0) + \mathscr{F}_\nu^*\, (\bar{p}_\nu^\mu - \bar{p}_\nu^{\mu-1}),$$

$$\hat{\hat{y}}_\nu^\mu = \hat{y}_\nu^\mu + \mathscr{D}_\nu\, (\bar{p}_\nu^{\mu+1} - \bar{p}_\nu^\mu), \qquad \check{\check{p}}_\nu^\mu = \check{p}_\nu^\mu + \mathscr{E}_\nu\, (\bar{y}_\nu^\mu - \bar{y}_\nu^{\mu-1})$$

are to be thought of as values of y^μ, p^μ at $t = t_{\nu-1}$, t_ν improved by the knowledge of the history of the iterative process. We define the *residuals*

$$R_\nu^\mu = \check{p}_{\nu+1}^\mu - \bar{p}_\nu^\mu - A_\nu\, (\bar{y}_{\nu+1}^{\mu-1} - \hat{\hat{y}}_\nu^{\mu-1}), \quad \mathring{R}_\nu^\nu = \check{p}_{\nu+1}^\mu - \bar{p}_\nu^{\mu-1} - A_\nu\, (\bar{y}_{\nu+1}^{\mu-1} - \hat{y}_\nu^{\mu-1}) \quad (1 \leq \nu < n),$$

$$S_\nu^\mu = \hat{y}_\nu^\mu - \bar{y}_{\nu+1}^\mu - B_\nu\, (\bar{p}_\nu^\mu - \check{\check{p}}_{\nu+1}^\mu), \qquad \mathring{S}_\nu^\mu = \hat{y}_\nu^\mu - \bar{y}_{\nu+1}^{\mu-1} - B_\nu\, (\bar{p}_\nu^\mu - \check{p}_\nu^\mu) \qquad (1 \leq \nu < n).$$

Theorem: *The iteration* (5.1) *with*

$$p_\nu^\mu(t_\nu) = \bar{p}_\nu^\mu := \bar{p}_\nu^{\mu-1} + (I - A_\nu\, B_\nu)^{-1} \left\{ \sum_{\rho=\nu+1}^{n} c_{\nu\rho}\, R_\rho^\mu + \mathring{R}_\nu^\mu + A_\nu \sum_{\rho=0}^{\nu-1} d_{\nu\rho}\, S_\rho^{\mu-1} \right\}, \tag{6.5 a}$$

$$y_\nu^\mu(t_{\nu-1}) = \bar{y}_\nu^\mu := \bar{y}_\nu^{\mu-1} + (I - B_{\nu-1}\, A_{\nu-1})^{-1} \left\{ \sum_{\rho=0}^{\nu-2} d_{\nu-1,\rho}\, S_\rho^\mu + \mathring{S}_{\nu-1}^\mu + B_{\nu-1} \sum_{\rho=\nu}^{n} c_{\nu-1,\rho}\, R_\rho^\mu \right\} \tag{6.5 b}$$

yields the estimates

$$\begin{aligned} \max_\nu \|\bar{p}_\nu^\mu - p(t_\nu)\| &\leq 0(|\Delta|^2) \max_{0 \leq t \leq T} \{\|y^{\mu-2}(t)\|, \|p^{\mu-2}(t)\|, \|y^{\mu-3}(t)\|\}, \\ \max_\nu \|\bar{y}_\nu^\mu - y(t_{\nu-1})\| &\leq 0(|\Delta|^2) \max_{0 \leq t \leq T} \{\|p^{\mu-1}(t)\|, \|y^{\mu-2}(t)\|, \|p^{\mu-2}(t)\|\}. \end{aligned} \tag{6.6}$$

The rate of convergence is at least proportional to $|\Delta|^{6/5}$:

$$\|y_\nu^\mu - y(t_{\nu-1})\| \leq K_1\, s^\mu \; (\mu \geq 0), \quad \|p_\nu^\mu - p(t_\nu)\| \leq K_1\, s^{\mu - \frac{1}{2}} \; (\mu \geq 1), \quad s \leq K_2\, |\Delta|^{6/5}. \tag{6.7}$$

Proof: (6.6) is obtained by reordering the right-hand sides of (6.5), applying (5.5) and estimating by (5.7). (6.7) follows from (6.6), (5.4) and (5.7) (cf. [5]).

[3] Define $y_\nu^{-1} = p_\nu^0 = 0$, since these values are additional starting values.

7. Application of the Method in the Discretized Form

In Chapter 6 the iteration method was considered for the differential equation itself. Now we discretize (2.1, 2.2) by the implicit scheme using a step size Δt and a grid $\Omega_h \subset \Omega$, both indicated by $h = (\Delta t, \Omega_h)$. The intermediate values t_ν of the partition Δ must be multiples of Δt. For each subinterval the analysis of the discretized version of (5.2) proves an expansion corresponding to (5.5) with $\mathscr{U}_\nu$, $\mathscr{D}_\nu$, $\mathscr{E}_\nu$, $\mathscr{F}_\nu$ replaced by some $(N \times N)$-matrices $\mathscr{U}_{h\nu}$, $\mathscr{D}_{h\nu}$, $\mathscr{E}_{h\nu}$, $\mathscr{F}_{h\nu}$ (for N cf. Chapter 4). The further auxiliary matrices $c_{h\nu}$, $d_{h\nu}$, $A_{h\nu}$, $B_{h\nu}$ are constructed analogously to (6.2). In many cases the following Note applies.

Note: If the coefficients L, D, E of (2.1) are time-independent and if the partition Δ is equidistant, then $\mathscr{U}_{h\nu}$, $\mathscr{D}_{h\nu}$, $\mathscr{E}_{h\nu}$ and $\mathscr{F}_{h\nu}$ do not depend on ν. $\mathscr{D}_{h\nu} = \mathscr{F}_{h\nu} = 0$ holds for $D = 0$, whereas $E = 0$ implies $\mathscr{E}_{h\nu} = \mathscr{F}_{h\nu} = 0$.

In the following we want to reduce the computational work. Contrary to the decoupling method mentioned in Chapter 4, the iteration (5.1) tends to the discrete solution of (2.1, 2.2), which is independent of perturbations of $\mathscr{U}_{h\nu}$, $\mathscr{D}_{h\nu}$ etc. Therefore, we look for approximate matrices $\tilde{\mathscr{U}}_{h\nu}, \ldots$ that are easier to compute and which result in almost the same rate of convergence.

A first possibility is the application of a *larger step size* $\tilde{\Delta} t = \varkappa \cdot \Delta t$ ($\varkappa > 1$, $t_\nu / \tilde{\Delta} t$ integers) for the numerical integration of (6.1). Let $L_h(t)$ be a difference analogue of $L(t)$. $I - \tilde{\Delta} t\, L_h(t)$ is assumed to be invertible. Then $\tilde{\mathscr{U}}_{h\nu}$, $\tilde{\mathscr{D}}_{h\nu}$, $\tilde{\mathscr{E}}_{h\nu}$, $\tilde{\mathscr{F}}_{h\nu}$ may be defined by

$$\begin{aligned}
&\tilde{\mathscr{U}}_h(t_{\nu-1}, t_{\nu-1}) = I,\ \tilde{\mathscr{U}}_h(t + \tilde{\Delta} t, t_{\nu-1}) = \left(I - \tilde{\Delta} t\, L_h(t)\right)^{-1} \tilde{\mathscr{U}}_h(t, t_{\nu-1}),\\
&\tilde{\mathscr{U}}_{h\nu} := \tilde{\mathscr{U}}_h(t_\nu, t_{\nu-1}),
\end{aligned} \tag{7.1 a}$$

$$\begin{aligned}
&\tilde{\mathscr{D}}_h(t_{\nu-1}, t_{\nu-1}) = 0,\\
&\tilde{\mathscr{D}}_h(t + \tilde{\Delta} t, t_{\nu-1}) = \left(I - \tilde{\Delta} t\, L_h(t)\right)^{-1} \left(\tilde{\mathscr{D}}_h(t, t_{\nu-1}) - \tilde{\Delta} t\, D_h(t)\right) \left(I - \tilde{\Delta} t\, L_h(t)^*\right)^{-1},\\
&\tilde{\mathscr{D}}_{h\nu} := \tilde{\mathscr{D}}_h(t_\nu, t_{\nu-1}),
\end{aligned} \tag{7.1 b}$$

$$\begin{aligned}
&\tilde{\mathscr{E}}_h(t_\nu, t_\nu) = 0,\ \tilde{\mathscr{E}}_h(t - \tilde{\Delta} t, t_\nu) = \left(I - \tilde{\Delta} t\, L_h(t)^*\right)^{-1} \left(\tilde{\mathscr{E}}_h(t, t_\nu) + \tilde{\Delta} t\, E_h(t)\right) \left(I - \tilde{\Delta} t\, L_h(t)\right)^{-1},\\
&\tilde{\mathscr{E}}_{hv} := \mathscr{E}_h(t_{\nu-1}, t_\nu),
\end{aligned} \tag{7.1 c}$$

$$\begin{aligned}
&\tilde{\mathscr{F}}_h(t_{\nu-1}, t_{\nu-1}) = 0,\\
&\tilde{\mathscr{F}}_h(t + \tilde{\Delta} t, t_{\nu-1}) = \left(I - \tilde{\Delta} t\, L_h(t)\right)^{-1} \left(\tilde{\mathscr{F}}_h(t, t_{\nu-1}) + \tilde{\Delta} t\ \tilde{\mathscr{D}}_h(t, t_{\nu-1})\, E_h(t)\, \tilde{\mathscr{U}}_h(t, t_{\nu-1})\right),\\
&\tilde{\mathscr{F}}_{h\nu} := \tilde{\mathscr{F}}_h(t_\nu, t_{\nu-1}).
\end{aligned}$$

where E_h, D_h denote approximations of the operators E, D. Again (7.1 a, b, c) are to be solved simultaneously.

The most work arises from the product $\tilde{\mathscr{D}}_h E_h \tilde{\mathscr{U}}_h$ of two full matrices appearing in (7.1 d). (E_h is assumed to be diagonal.) Consider $\mathscr{U}_\nu \delta_P$, where δ_P approximates the Dirac function for $P \in \Omega$. The function $\mathscr{U}_\nu \delta_P$ is concentrated in a small neighbourhood of P if $|\Delta|$ or the coefficients of the differential operator L are small. Expecting similar properties for $\mathscr{U}_h$ and $\mathscr{D}_h$, we may omit all entries corresponding to points $P, Q \in \Omega_h$ with $|P - Q| > \rho_0$ for suitable ρ_0. Thus, $\tilde{\mathscr{U}}_h$ and $\tilde{\mathscr{D}}_h$ involved in

(7.1 d) become *sparse matrices.* For $r=1$ these matrices are band matrices. In any case the work for computing $\tilde{\mathcal{D}}_h E_h \tilde{\mathcal{U}}_h$ is reduced.

Now consider the opposite case and assume that δ_P is smoothed out by applying $\mathcal{U}_\nu$. Recalling that the computational work is proportional to Δt^{-1} Δx^{-3r} (cf. Chapter 4), the replacement of Ω_h by a *coarser spatial grid* $\Omega_{\tilde{h}} \subset \Omega_h$ seems promising. Therefore, we introduce two mappings φ and ψ. By φ the $\Omega_{\tilde{h}}$-grid functions are interpolated on Ω_h, while ψ maps functions defined on Ω_h into functions defined on $\Omega_{\tilde{h}}$. Furthermore, $\psi \circ \varphi = I$ is assumed. If $\Omega_{\tilde{h}} \subset \Omega_h \subset \Omega = (0, 1)$ are equidistant grids we may, e.g., define ψ by truncation of the trigonometric interpolating polynomial, whereas φ describes the trigonometric interpolation in Ω_h.

By means of $\phi(M) := \varphi M \psi$, matrices M corresponding to $\Omega_{\tilde{h}}$ are mapped into Ω_h-matrices. Moreover, the "interpolation" ϕ is multiplicative: $\phi(M_1 M_2) = \phi(M_1)\,\phi(M_2)$. Let $\tilde{\mathcal{U}}_{\tilde{h}\nu}$, $\tilde{\mathcal{D}}_{\tilde{h}\nu}$, etc., denote the results of (7.1) for $\tilde{h} = (\tilde{\Delta} t, \Omega_{\tilde{h}})$. Formally[4] we define $\tilde{\mathcal{U}}_{h\nu} := \phi(\tilde{\mathcal{U}}_{\tilde{h}\nu})$, etc. The above assumption implies that the matrices $\mathcal{U}_{h\nu}, \dots$ have good approximations belonging to the subspace range (ϕ).

In order to analyse this situation, we consider a model problem. Let $\Omega_{\tilde{h}} \subset \Omega_h \subset \Omega$ be as mentioned above in the example for φ, ψ. Furthermore, the grid functions $(\sin \chi \mu \pi \Delta x)_{\mu=1,\dots,N}$ $(\Delta x = 1/(N+1),\ 1 \leq \chi \leq N)$ are assumed to be eigenfunctions of the discretized version of (2.1, 2.2). Then Fourier analysis gives the following result. The difference of the eigenvalues of $\mathcal{U}_{h\nu}$ and $\mathcal{U}_{\tilde{h}\nu}$ is very small for small wave numbers χ. Since the first N eigenvalues of $\mathcal{U}_{\tilde{h}\nu}$ and $\tilde{\mathcal{U}}_{h\nu} := \phi(\mathcal{U}_{\tilde{h}\nu})$ coincide, $\tilde{\mathcal{U}}_{h\nu}$ is a good approximation of $\mathcal{U}_{h\nu}$ with respect to the first part of their spectra. By our assumption the remaining components of the Fourier expansion are diminished by the difference equation itself, provided that the coupling coefficients D and E are not extremely large.

Finally we touch upon nonlinear problems. E.g., if we apply Newton's method, a sequence of linear problems has to be solved. The auxiliary matrices $\mathcal{U}_{h\nu}, \dots$ are to be computed for each step of Newton's method (primary iteration). However, since we need only approximate values $\tilde{\mathcal{U}}_{h\nu}, \dots$, we may use the same matrices for several iteration steps. As soon as the convergence rate of the secondary iteration becomes too slow, the auxiliary matrices are to be computed again.

8. Numerical Examples

Two one-dimensional examples are added to give an idea of the rate of convergence. The first example has also been computed by Nedelec [11]:

$$\Omega = (0, 1),\ H_B^2 = H_0^2(\Omega),\ L(t) = \frac{1}{400}\left(\frac{\partial}{\partial x}\right)^2,\ D(t) = 1 \text{ or } 100,\ E(t) = 0,$$

$$B_0 = 0,\ A_T = 1,\ T = 1.$$

[4] The right-hand sides of (6.5) can be expressed without using $\tilde{\mathcal{U}}_{h\nu}, \dots$ explicitly. Only $\tilde{\mathcal{U}}_{\tilde{h}\nu}, \dots$ and φ, ψ and $\varphi \circ \psi$ are necessary.

f, g, a_T and b_0 are such that

$$y(t)=\sin(\pi x)/\sqrt{10}-\sqrt{\frac{3}{2}}\,x(x-1),\quad p(t)=-\frac{3}{\sqrt{40}}\sin(\pi x)\exp\left(\frac{\pi^2 t}{400}\right)$$

satisfy (2.1), (2.2). The grid sizes are $\Delta t=\frac{1}{40}$ and $\Delta x=\frac{1}{20}$. The iteration starts with $u^{(0)}=0$. The CPU time[5] t_I is needed for one iteration, i.e. for the integration of p and y from T to 0 and 0 to T, respectively, without calculating the initial values (6.5). t_P corresponds to the CPU time of the preprocessing phase, while t_C denotes the time that is necessary for the computation of the initial values (6.5) during one iteration. Therefore, the numerical solution of the problem by v iterations requires the time $t_P+v(t_C+t_I)$. The time for the evaluation of the coefficients of the differential equations is not taken into account. The number of intervals is $n=1$.

Table 1. *First example* ($\delta y^{(\mu)}=\|y^{(\mu)}-y^{(\mu-1)}\|_\infty$, $\delta p^{(\mu)}=\|p^{(\mu)}-p^{(\mu-1)}\|_\infty$, w: *maximal band width*)

D	1		1		1		100		100	
	discretization parameters									
$\tilde{\Delta} t$	1/40		1/4		1/4		1/40		1/4	
$\tilde{\Delta} x$	1/20		1/20		1/10		1/20		1/20	
w	18		10		4		18		16	
	maximum norm of corrections									
μ	$\delta y^{(\mu)}$	$\delta p^{(\mu)}$	$\delta y^{(\mu)}$	$\delta p^{(\mu)}$	$\delta y^{(\mu)}$	$\delta p^{(\mu)}$	$\delta y^{(\mu)}$	$\delta p^{(\mu)}$	$\delta y^{(\mu)}$	$\delta p^{(\mu)}$
1	7.0(−1)	5.6(−1)	7.0(−1)	5.6(−1)	7.0(−1)	5.6(−1)	4.6(+1)	1.1(−2)	4.6(+1)	1.1(−2)
2	7.4(−2)	7.5(−2)	7.6(−2)	7.7(−2)	7.6(−2)	7.7(−2)	4.6(+1)	4.8(−1)	4.7(+1)	4.8(−1)
3	2.6(−5)	3.1(−5)	3.0(−4)	3.4(−4)	2.9(−4)	3.6(−4)	2.8(−2)	2.9(−4)	2.8(−1)	2.9(−3)
4	6.6(−8)	1.5(−7)	5.6(−6)	8.8(−6)	5.4(−6)	2.1(−5)	1.7(−5)	1.8(−7)	1.7(−3)	1.8(−5)
5		1.1(−9)	1.8(−7)	3.6(−7)	8.0(−7)	3.6(−6)	1.7(−7)	8.0(−9)	1.8(−5)	8.6(−7)
	additional computational work									
t_P/t_I	11		1.8		0.5		11		2.2	
t_C/t_I	0.11		0.10		0.10		0.11		0.10	

The first and fourth columns of Table 1 correspond to $\tilde{\Delta} t=\Delta t$, $\tilde{\Delta} x=\Delta x$. Therefore, the respective ratio t_P/t_I is similar to that of the decoupling method.

[5] The computations were performed on the CDC computer Cyber 72 of the Rechenzentrum der Universität zu Köln.

Table 1 shows that the computational amount of the preprocessing phase can greatly be reduced by a suitable choice of $\tilde{\Delta} t$ and $\tilde{\Delta} x$ preserving fast convergence.

Because of $E(t)=0$, the expansions (5.5 a, b) consist of at most two terms. For the second example both coefficients D and E do not vanish:

$$\Omega=(0,1),\ H_B^2=H_0^2(\Omega),\ L(t)=\left(\frac{\partial}{\partial x}\right)^2,\ D(t)=10,\ E(t)=1,\ B_0=A_T=1,\ T=4.$$

f, g, a_T and b_0 correspond to the solution $y=p=x(1-x)$. The parameters of the discretization are $\Delta t=\frac{4}{80}$, $\Delta x=\frac{1}{20}$. The first column of Table 2 is obtained for $n=1$, $\tilde{\Delta} t=\Delta t$, $\tilde{\Delta} x=\Delta x$, while the second uses $n=1$, $\tilde{\Delta} t=\frac{4}{40}$, $\tilde{\Delta} x=\frac{1}{4}$.

Table 2. *Second example*

μ number of iterations	$\tilde{\Delta} t=4/80$, $\Vert y^{(\mu)}-y^{(\mu-1)}\Vert$	$\tilde{\Delta} x=1/20$ $\Vert p^{(\mu)}-p^{(\mu-1)}\Vert$	$\tilde{\Delta} t=4/40$, $\Vert y^{(\mu)}-y^{(\mu-1)}\Vert$	$\tilde{\Delta} x=1/4$ $\Vert p^{(\mu)}-p^{(\mu-1)}\Vert$
1	4.2(−1)	2.2(−1)	4.2(−1)	2.2(−1)
2	2.0(−1)	2.1(−1)	2.0(−1)	2.2(−1)
3	3.6(−2)	4.0(−2)	4.5(−2)	4.9(−2)
4	7.2(−3)	7.9(−3)	1.0(−2)	1.1(−2)
5	1.4(−3)	1.6(−3)	2.3(−3)	2.5(−3)
6	2.7(−4)	3.0(−4)	5.1(−4)	5.6(−4)
7	5.3(−5)	5.9(−5)	1.1(−4)	1.3(−4)
8	1.0(−5)	1.2(−5)	2.5(−5)	2.8(−5)
9	2.0(−6)	2.3(−6)	5.7(−6)	6.3(−6)
t_P/t_I	65		1.2	
t_C/t_I	0.13		0.04	

The replacement of $\tilde{\Delta} x=\frac{1}{20}$ by $\tilde{\Delta} x=\frac{1}{4}$ corresponds to the interpolation of (19 × 19)-matrices by (3 × 3)-matrices. Table 2 shows that nearly the same results are obtained with very much less computational work of the preprocessing phase.

References

[1] Casti, J., Ljung, L.: Some new analytic and computational results for operator Riccati equations. SIAM J. Control *13*, 817—826 (1975).

[2] Curtain, R. F.: The infinite-dimensional Riccati equation with applications to affine hereditary differential systems. SIAM J. Control *13*, 1130—1143 (1975).

[3] Da Prato. G.: Equation d'évolution dans des algèbres d'opérateurs et application à des équations quasi-linéaires. J. Math. Pures et Appl. *48*, 59—107 (1969).

[4] Gail, H.-P., Sedlmayr, E., Traving, G.: Non-LTE line formation in turbulent media. Astron. & Astrophys. *44*, 421—429 (1975).

[5] Hackbusch, W.: A numerical method for solving parabolic equations with opposite orientations, Part I: Description of the method. Report 77-6, Mathematisches Institut (Angewandte Mathematik), Universität zu Köln, 1977.

[6] Ladas, G. E., Lakshmikantham, V.: Differential equations in abstract spaces. New York-London: Academic Press 1972.

[7] Leroy, D.: Méthodes numériques en contrôle optimal, application à un problème d'échange thermique. Thèse, Paris, 1972.

[8] Lions, J. L.: Optimal control of systems governed by partial differential equations. Berlin-Heidelberg-New York: Springer 1971.

[9] Lions, J. L., Magenes, E.: Non-homogeneous boundary value problems and applications I, II. Berlin-Heidelberg-New York: Springer 1972.

[10] Miellou, J.-C.: Sur une notion de monotonie conduisant à une extension de l'application de la méthode variationelle dans l'étude des systèmes d'équations et d'inéquations aux dérivées partielles, opérateurs paramonotones. Thèse Sc. math., Grenoble, 1970.

[11] Nedelec, M.: Schémas d'approximation pour des équations intégro différentielles de Riccati. Thèse, Paris, 1970.

[12] Stoer, J., Bulirsch, R.: Einführung in die Numerische Mathematik II. Berlin-Heidelberg-New York: Springer 1973.

[13] Taufer, J.: On factorization method. Aplikace matematiky *11*, 427—451 (1966).

[14] Temam, R.: Sur l'équation de Riccati associée à des opérateurs non bornés, en dimension infinie. J. Functional Analysis *7*, 85—115 (1971).

Dr. W. Hackbusch
Mathematisches Institut
Universität zu Köln
Weyertal 86—90
D-5000 Köln 41
Federal Republic of Germany