

THEORETICAL OPTIMIZATION OF FINITE DIFFERENCE SCHEMES

CLAIRE DAVID ¹, PIERRE SAGAUT ¹

¹ Université Pierre et Marie Curie-Paris 6

Institut Jean Le Rond d'Alembert, UMR CNRS 7190

Boîte courrier n^o162, 4 place Jussieu, 75252 Paris, cedex 05, France

ABSTRACT. The aim of this work is to develop general optimization methods for linear finite difference schemes used to approximate linear differential equations, on the basis of a matrix equation, which enables to determine the optimal value of a parameter for a given scheme.

1. Introduction: Scheme classes. Finite difference schemes used to approximate linear differential equations induce numerical errors, which partly depend on given parameters (as, for instance, the number of cells per wavelength). The usual process consists in testing several values for those coefficients, until one finds the optimal one.

We here propose a completely different approach, which consists in determining the optimal value of a given parameter, which minimizes the L_2 norm of the error for a given finite difference scheme.

Consider the transport equation:

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0 \quad (1)$$

with the initial condition $u(x, t = 0) = u_0(x)$.

A linear finite difference scheme for this equation can be written under the form:

$$\alpha u_i^{n+1} + \beta u_i^n + \gamma u_i^{n-1} + \delta u_{i+1}^n + \varepsilon u_{i-1}^n + \zeta u_{i+1}^{n+1} + \eta u_{i-1}^{n-1} + \theta u_{i-1}^{n+1} + \vartheta u_{i+1}^{n-1} = 0 \quad (2)$$

where:

$$u_l^m = u(lh, m\tau) \quad (3)$$

$l \in \{i-1, i, i+1\}$, $m \in \{n-1, n, n+1\}$, $j = 0, \dots, n_x$, $n = 0, \dots, n_t$, h , τ denoting respectively the mesh size and time step.

The Courant-Friedrichs-Lewy number (cfl) is defined as $\sigma = c\tau/h$.

A numerical scheme is specified by selecting appropriate values of the coefficients α , β , γ , δ , ε , ζ , η , θ and ϑ in equation (2). Values corresponding to numerical schemes retained for the present works are given in Table 1.

The number of time steps will be denoted n_t , the number of space steps, n_x . In general, $n_t \gg n_x$.

TABLE 1. Numerical scheme coefficient.

Name	α	β	γ	δ	ϵ	ζ	η	θ	ϑ
Leapfrog	$\frac{1}{2\tau}$	0	$\frac{-1}{2\tau}$	$\frac{c}{2h}$	$\frac{-c}{2h}$	0	0	0	0
Lax	$\frac{1}{\tau}$	0	0	$\frac{-1}{2\tau} + \frac{c}{2h}$	$\frac{-1}{2\tau} - \frac{c}{2h}$	0	0	0	0
Lax-Wendroff	$\frac{1}{\tau}$	$\frac{-1}{\tau} + \frac{c^2\tau}{h^2}$	0	$\frac{(1-\sigma)c}{2h}$	$\frac{-(1+\sigma)c}{2h}$	0	0	0	0
Crank-Nicolson	$\frac{1}{\tau} + \frac{c}{h^2}$	$\frac{-1}{\tau} + \frac{c}{h^2}$	0	$\frac{-c}{h^2}$	$\frac{-c}{h^2}$	0	$\frac{-c}{h^2}$	$\frac{-c}{h^2}$	0

The paper is organized as follows. The equivalent matrix equation is exposed in section 2. The determination of the optimal value of a given parameter presented in section 3.

2. The Sylvester equation.

2.1. Matricial form of the finite differences problem. Let us introduce a rectangular matrix defined by:

$$U = [u_i^n]_{1 \leq i \leq n_x - 1, 1 \leq n \leq n_t} \quad (4)$$

The finite difference scheme (2) can be written under the following matricial form:

$$M_1 U + U M_2 + \mathcal{L}(U) = M_0 \quad (5)$$

where M_1 , M_2 and M_0 are square matrices respectively $n_x - 1$ by $n_x - 1$, n_t by n_t , given by:

$$M_1 = \begin{pmatrix} \beta & \delta & 0 & \dots & 0 \\ \epsilon & \beta & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \beta & \delta \\ 0 & \dots & 0 & \epsilon & \beta + \delta \end{pmatrix} \quad M_2 = \begin{pmatrix} 0 & \gamma & 0 & \dots & 0 \\ \alpha & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \gamma \\ 0 & \dots & 0 & \alpha & 0 \end{pmatrix} \quad (6)$$

and where \mathcal{L} is a linear matricial operator which can be written as:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 \quad (7)$$

where \mathcal{L}_1 , \mathcal{L}_2 , \mathcal{L}_3 and \mathcal{L}_4 are given by:

$$\mathcal{L}_1(U) = \zeta \begin{pmatrix} u_2^2 & u_2^3 & \dots & u_2^{n_t} & 0 \\ u_3^2 & u_3^3 & \dots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ u_{n_x-1}^2 & u_{n_x-1}^3 & \dots & u_{n_x-1}^{n_t} & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix} \quad \mathcal{L}_2(U) = \eta \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & u_1^1 & u_1^2 & \dots & u_1^{n_t-1} \\ 0 & u_1^0 & u_1^1 & \dots & u_2^{n_t-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & u_{n_x-2}^1 & u_{n_x-2}^2 & \dots & u_{n_x-2}^{n_t-1} \end{pmatrix} \quad (8)$$

$$\mathcal{L}_3(U) = \theta \begin{pmatrix} 0 & \dots & \dots & \dots & 0 \\ u_1^2 & u_1^3 & \dots & u_1^{n_t} & 0 \\ u_2^2 & u_2^3 & \dots & u_2^{n_t} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ u_{n_x-2}^2 & u_{n_x-2}^3 & \dots & u_{n_x-2}^{n_t} & 0 \end{pmatrix} \quad \mathcal{L}_4(U) = \vartheta \begin{pmatrix} 0 & u_2^1 & u_2^2 & \dots & u_2^{n_t-1} \\ 0 & u_3^1 & u_3^2 & \dots & u_3^{n_t-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & u_{n_x-1}^1 & \dots & \dots & u_{n_x-1}^{n_t-1} \\ 0 & 0 & \dots & \dots & 0 \end{pmatrix} \quad (9)$$

The second member matrix M_0 bears the initial conditions, given for the specific value $n = 0$, which correspond to the initialization process when computing loops, and the boundary conditions:

$$M_0 = \begin{pmatrix} -\gamma u_1^0 - \varepsilon u_0^1 - \eta u_0^0 - \theta u_0^2 - \vartheta u_2^0 & -\varepsilon u_0^2 - \eta u_0^1 - \theta u_0^3 & \dots & \dots & -\varepsilon u_0^{n_t} - \eta u_0^{n_t-1} \\ -\gamma u_2^0 - \eta u_1^0 - \vartheta u_3^0 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\gamma u_{n_x-2}^0 - \eta u_{n_x-2}^0 - \vartheta u_{n_x-1}^0 & 0 & \dots & \dots & 0 \\ -\gamma u_{n_x-1}^0 - \eta u_{n_x-2}^0 - \zeta u_{n_x}^2 - \vartheta u_{n_x}^0 & -\zeta u_{n_x}^3 - \vartheta u_{n_x}^1 & \dots & \dots & -\vartheta u_{n_x}^{n_t-1} \end{pmatrix} \quad (10)$$

When non-Dirichlet boundary conditions are taken into account, and when the solution at $i = n_x$ is obtained by interpolation, setting $u_{n_x}^n = u_{n_x-1}^n$, M_0 has to be changed in:

$$M_0^{mod} = \begin{pmatrix} -\gamma u_1^0 - \varepsilon u_0^1 - \eta u_0^0 - \theta u_0^2 - \vartheta u_2^0 & -\varepsilon u_0^2 - \eta u_0^1 - \theta u_0^3 & \dots & \dots & -\varepsilon u_0^{n_t} - \eta u_0^{n_t-1} \\ -\gamma u_2^0 - \eta u_1^0 - \vartheta u_3^0 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\gamma u_{n_x-2}^0 - \eta u_{n_x-2}^0 - \vartheta u_{n_x-1}^0 & 0 & \dots & \dots & 0 \\ -\gamma u_{n_x-1}^0 - \eta u_{n_x-2}^0 - \vartheta u_{n_x}^0 & 0 & \dots & \dots & 0 \end{pmatrix} \quad (11)$$

and a corrective term, of the form: $\mathcal{D}.U.M_3$, has to be taken into account in the matrix equation, where:

$$\mathcal{D} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix}, \quad M_3 = \begin{pmatrix} 0 & \vartheta & 0 & \dots & 0 \\ \zeta & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vartheta \\ 0 & \dots & 0 & \zeta & 0 \end{pmatrix} \quad (12)$$

The new matrix equation is then:

$$M_1 U + U M_2 + \mathcal{L}(U) + \mathcal{D}.U.M_3 = M_0^{mod} \quad (13)$$

Denote by u_{exact} the exact solution of the transport equation (1). The U_{exact} corresponding matrix will be:

$$U_{exact} = [U_{exact_i}^n]_{0 \leq i \leq n_x-1, 0 \leq n \leq n_t} \quad (14)$$

where:

$$U_{exact_i}^n = U_{exact}(x_i, t_n) \quad (15)$$

with $x_i = i h$, $t_n = n \tau$.

U is then the solution of:

$$M_1 U + U M_2 + \mathcal{L}(U) = M_0 \quad (16)$$

We will call *error matrix* the matrix defined by:

$$E = U - U_{exact} \quad (17)$$

Let us consider the matrix F defined by:

$$F = M_1 U_{exact} + U_{exact} M_2 + \mathcal{L}(U_{exact}) - M_0 \quad (18)$$

The *error matrix* E satisfies then:

$$M_1 E + E M_2 + \mathcal{L}(E) = F \quad (19)$$

2.2. The matrix equation.

2.2.1. *Theoretical formulation.* Minimizing the error due to the approximation induced by the numerical scheme is equivalent to minimizing the norm of the matrices E satisfying (19).

Since the linear matricial operator \mathcal{L} appears only in the Crank-Nicholson scheme, we will restrain our study to the case $\mathcal{L} = 0$. The generalization to the case $\mathcal{L} \neq 0$ can be easily deduced: denote

$$\Sigma = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \dots & 0 & \dots & 0 \end{pmatrix} \quad (20)$$

we have:

$$\mathcal{L}_1(U) = \zeta \Sigma.U^T \Sigma, \quad \mathcal{L}_2(U) = \eta^T \Sigma.U \Sigma, \quad \mathcal{L}_3(U) = \theta^T \Sigma.U^T \Sigma, \quad \mathcal{L}_4(U) = \vartheta \Sigma.U \Sigma \quad (21)$$

where T denotes the transposed matrix, which yields:

$$\mathcal{L}(U) = (\zeta \Sigma + \theta^T \Sigma).U^T \Sigma + (\eta^T \Sigma + \vartheta \Sigma).U \Sigma \quad (22)$$

The problem is then the determination of the minimum norm solution of:

$$M_1 E + E M_2 + (\zeta \Sigma + \theta^T \Sigma).E^T \Sigma + (\eta^T \Sigma + \vartheta \Sigma).E \Sigma = F \quad (23)$$

which is a specific form of the generalized Sylvester equation:

$$AX + BXC = D \quad (24)$$

where A and B are m by m matrices, C and F n by n matrices, and X , a m by n matrix.

The solving of the Sylvester equation is generally based on Schur decomposition: for a given square n by n matrix A , n being an even number of the form $n = 2p$, there exists a unitary matrix U and an upper triangular block matrix T such that:

$$A = U^*TU \quad (25)$$

where U^* denotes the (complex) conjugate matrix of the transposed matrix TU . The diagonal blocks of the matrix T correspond to the complex eigenvalues λ_i of A :

$$T = \begin{pmatrix} T_1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & T_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & T_p \end{pmatrix} \quad (26)$$

where the block matrices T_i , $i = 1, \dots, p$ are given by:

$$\begin{pmatrix} \mathcal{R}e[\lambda_i] & \mathcal{I}m[\lambda_i] \\ -\mathcal{I}m[\lambda_i] & \mathcal{R}e[\lambda_i] \end{pmatrix} \quad (27)$$

$\mathcal{R}e$ being the real part of a complex number, and $\mathcal{I}m$ the imaginary one.

Due to this decomposition, the Sylvester equation require, to be solved, that the

dimensions of the matrices be even numbers. We will therefore, in the following, restrain our study to n_x and n_t being even numbers. So far, it is interesting to note that the Schur decomposition being more stable for higher order matrices, it perfectly fits finite differences problems.

Complete parametric solutions of the Sylvester equation (24) is given in [2], [3].

As for the determination of the solution of the Sylvester equation, it is a major topic in control theory, and has been the subject of numerous works (see [1], [6], [8], [9], [10], [11], [12]).

In [1], the method is based on the reduction of the observable pair (A, C) to an observer-Hessenberg pair (H, D) , H being a block upper Hessenberg matrix. The reduction to the observer-Hessenberg form (H, D) is achieved by means of the staircase algorithm (see [4]).

In [9], in the specific case of B being a companion form matrix, the authors propose a very neat general complete parametric solution, which is expressed in terms of the controllability of the matrix pair (A, B) , a symmetric matrix operator, and a parametric matrix in the Hankel form.

We recall that a companion form, or Frobenius matrix is one of the following kind:

$$B = \begin{pmatrix} 0 & \dots & \dots & \dots & 0 & -b_0 \\ 1 & 0 & \dots & \dots & 0 & -b_1 \\ 0 & 1 & 0 & \dots & \vdots & \vdots \\ \vdots & 0 & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & -b_{p-1} \end{pmatrix} \quad (28)$$

These results can be generalized through matrix block decomposition to a block companion form matrix, which happens to be the case of our matrix M_2 in the

specific case of n_x and n_t being even numbers:

$$M_2 = \begin{pmatrix} M_2^{B^1} & 0 & \dots & \dots & 0 \\ 0 & M_2^{B^2} & 0 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & M_2^{B^k} \end{pmatrix} \quad (29)$$

the $M_2^{B^p}$, $1 \leq p \leq k$ being companion form matrices.

Another method is presented in [14], where the determination of the minimum-norm solution of a Sylvester equation is specifically developed.

The accuracy and computational stability of the solutions is examined in [5].

2.2.2. Existence condition of the solution. Equation (24) has a unique solution if and only if A and B have no common eigenvalues, which is presently the case. (19) has thus a unique solution, which accounts for the consistency of the given problem.

3. Optimization: determination of the optimal value of a specific parameter for a given scheme. Advection of a sinusoidal signal

$$u = \text{Cos} \left[\frac{2\pi}{\lambda} (x - ct) \right] \quad (30)$$

through the Lax scheme, where:

$$\lambda = n_\lambda dx \quad (31)$$

n_λ denotes the number of cells per wavelength.

Let n_λ remain unknown.

Equation (23) can thus be normalized as:

$$\overline{M}_1 E + E \overline{M}_2 = \overline{F} \quad (32)$$

where

$$\begin{cases} \overline{M}_1 &= \frac{h c fl}{c} M_1 \\ \overline{M}_2 &= \frac{h c fl}{c} M_2 \\ \overline{F} &= \frac{h c fl}{c} F \end{cases} \quad (33)$$

We deliberately choose a small value for the number of steps: $n_t = n_x = 20$, starting from the point that if the error is minimized for a small value of this number, it will be the same as this number increases.

Figure 1 displays the L_2 norm of the isovalues of the error as a function of n_λ (maximums are in white, minimums in black; larger values are shown lighter). It clearly appears that, if one chooses a value of n_λ in a darker zone, the error is minimized. Calculations have been done with the *LyapunovSolve* package of *Mathematica*, which presents the advantage of dealing with symbolic parameters, as it is the case for our a priori unknown coefficient n_λ .

The L_2 norm of the error, for two significative values of the number of cells per wavelength n_p (one corresponding to a minimum of the error, and one corresponding to a maximum of the error), is displayed in Figure 2:

The above results ensure the faster convergence of the scheme.

4. Conclusion. The matrix form of the linear finite difference problem enables us to minimize the symbolic expression of the L_2 norm of the error as a function of one of the scheme coefficient (as the number of cells per wavelength, presented as an example). Thus, the optimal value of a given coefficient, i. e. one that minimizes the error, can be easily determined.

REFERENCES

- [1] Van Dooren, P., *Reduced order observers: A new algorithm and proof*, Systems Control Lett., Vol. 4, pp. 243-251 (1984).
- [2] Berman, A., Plemmons, R. J., *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, PA (1994).
- [3] Gail, H. R., Hantler, S. L., Taylor, B. A., *Spectral Analysis of M/G/1 and G/M/1 type Markov chains*, Adv. Appl. Probab., Vol. 28, pp. 114-165 (1996).

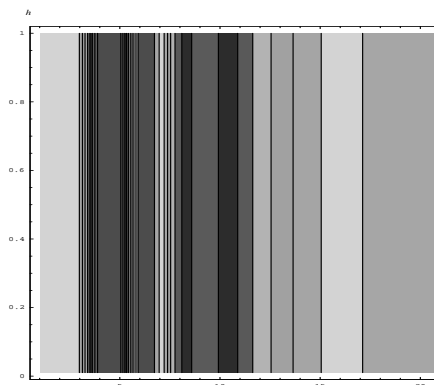


FIGURE 1. Isovalues of the L_2 norm of the error as a function of the number of cells per wavelength n_λ

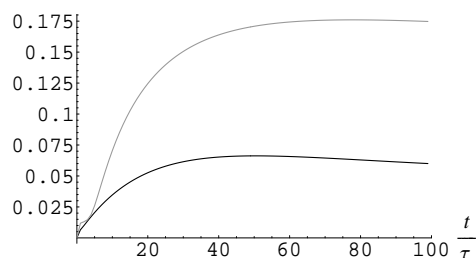


FIGURE 2. L_2 norm of the error for different values of the number of cells per wavelength n_λ . Case 1 (in black): $n_p = 6.75$. Case 2 (in grey): $n_p = 20$.

- [4] Boley, D. L., *Computing the Controllability algorithm / Observability Decomposition of a Linear Time-Invariant Dynamic System, A Numerical Approach*, PhD. thesis, Report STAN-CS-81-860, Dept. Comp. i, Sci., Stanford University (1981).
 - [5] Deif, A. S., Seif, N. P., Hussein, S. A., *Sylvester's equation: accuracy and computational stability*, Journal of Computational and Applied Mathematics, Vol. 61, pp. 1-11 (1995).
 - [6] Hearon, J. Z., *Nonsingular solutions of $TA - BT = C$* , Linear Algebra and its applications, Vol. 16, pp. 57-63 (1977).
 - [7] Huo, C. H., *Efficient methods for solving a nonsymmetric algebraic equation arising in stochastic fluid models*, Journal of Computational and Applied Mathematics, pp. 1-21 (2004).
 - [8] Tsui, C. C., *A complete analytical solution to the equation $TA - FT = LC$ and its applications*, IEEE Trans. Automat. Control AC, Vol. 32, pp. 742-744 (1987).
 - [9] Zhou, Bin, Duan, Guang-Den, S., *An explicit solution to the matrix equation $AX - XF = BY$* , Linear Algebra and its applications, Vol. 402, pp. 345-366 (2005).
 - [10] Duan, G. R., *Solution to matrix equation $AV + BW = EVF$ and eigenstructure assignment for descriptor systems*, Automatica, Vol. 28, pp. 639-643 (1992).
 - [11] Duan, G. R., *On the solution to Sylvester matrix equation $AV + BW = EVF$ and eigenstructure assignment for descriptor systems*, IEEE Trans. Automat. Control AC, Vol. 41 (4), pp. 276-280 (1996)
 - [12] Kirrinnis, P., *Fast algorithms for the Sylvester equation $AX - XB^T = C$* , Theoretical Computer Science, Vol.259, pp. 623-638 (2000).
 - [13] Konstantinov, M., Mehrmann, V., Petkov, P., *On properties of Sylvester and Lyapunov operators*, Linear Algebra and its applications, Vol. 312, pp. 35-71 (2000).
 - [14] Varga, A., *TA numerically reliable approach to robust pole assignment for descriptor systems*, Future Generation Computer Systems, Vol. 19 (7), pp. 1221-1230 (2003).
 - [15] Witham, G.B., *Linear and Nonlinear Wave*, Wiley-Interscience (1974).
 - [16] Wolfram, S., *The Mathematica book*, Cambridge University Press (1999).
- E-mail address:* david@lmm.jussieu.fr