ATTI ACCADEMIA NAZIONALE LINCEI CLASSE SCIENZE FISICHE MATEMATICHE NATURALI

Rendiconti Lincei Matematica e Applicazioni

Ewa Turska, Bernardo A. Schrefler

On consistency, stability and convergence of staggered solution procedures

Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Matematica e Applicazioni, Serie 9, Vol. 5 (1994), n.3, p. 265–271.

Accademia Nazionale dei Lincei

<http://www.bdim.eu/item?id=RLIN_1994_9_5_3_265_0>

L'utilizzo e la stampa di questo documento digitale è consentito liberamente per motivi di ricerca e studio. Non è consentito l'utilizzo dello stesso per motivi commerciali. Tutte le copie di questo documento devono riportare questo avvertimento.

Articolo digitalizzato nel quadro del programma bdim (Biblioteca Digitale Italiana di Matematica) SIMAI & UMI http://www.bdim.eu/

Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Matematica e Applicazioni, Accademia Nazionale dei Lincei, 1994.

Meccanica dei continui. — On consistency, stability and convergence of staggered solution procedures. Nota di Ewa Turska e Bernardo A. Schrefler, presentata (*) dal Corrisp. G. Maier.

ABSTRACT. — The simultaneous and staggered procedures of solving a partitioned form of a coupled system of ordinary differential equations are presented. Formulas for errors are compared. Counter-examples for convergence with a constant number of iterations at each time step are given.

KEY WORDS: Coupled problems; Systems of linear simultaneous equations; Numerical solution procedures.

RIASSUNTO. — Sulla consistenza, stabilità e convergenza di procedure di soluzione con partizione matriciale. Vengono discussi due metodi di soluzione di un sistema di equazioni differenziali ordinarie che descrivono problemi di interazione nel campo dell'ingegneria. Si presentano formule per la valutazione dell'errore e contro esempi concernenti la convergenza ottenuta con un numero costante di iterazioni per passo temporale.

INTRODUCTION

Many engineering problems involve some time dependent interacting fields. Typical examples relating to slow phenomena are thermomechanical coupling and isothermal or non-isothermal consolidation. For their quantitative solution numerical procedures are often applied which consist of finite element discretization in space and finite difference discretization in time. The effectiveness of them depends to a great extent on algorithms used for the solution of the algebraic system of equations resulting from the discretization process. Usually the solutions of these equations are obtained by iteration.

One of the iterative methods is the staggered procedure. It allows to solve large problems of coupled fields using available numerical codes made for single field problems, and is easily transferable from linear to non-linear equations. From a large number of papers on the topic it can be concluded that the staggered scheme is accurate and efficient, but in some cases can cause difficulties.

The main aim of the paper is to find satisfactory criteria to test the staggered procedure and to compare them with ones for the simultaneous procedure and a direct solution process. The properties shall be illustrated by examples in which the generalised mid-point (GM-P) method is used as time discretization, thus the presented counterexamples cover a large range of finite difference schemes.

We present a detailed study of linear problems, without assuming symmetry or positive definiteness of matrices appearing in the semi-discrete problem. This allows to extend the conclusions to non-linear problems treating them as linear at each time instant.

^(*) Nella seduta del 12 febbraio 1994.

CONDITIONS SUFFICIENT FOR GLOBAL CONVERGENCE

Let us consider the following differential equation. It may represent *e.g.* a thermomechanical problem or a semi-saturated isothermal or non isothermal consolidation problem (see [1]):

(1)

$$B\dot{X} + CX = F$$

where matrices $B = [b_{ij}]$, $C = [c_{ij}]$, $F = [f_i]$ have been obtained by the *FE* discretization, the nodal variables vector X being a point in a vector space with Euclidean norm. Symmetry or positive definiteness of matrices B and C is not assumed.

On use of a one-step finite difference operator endowed with consistency, for the time derivative (*e.g.* the backward or forward Euler, arbitrary Runge-Kutta etc.) we obtain the discrete equation:

(2)
$$\mathfrak{A}X_{n+1} - \mathfrak{B}X_n - \mathfrak{F} = 0$$

where $\mathfrak{A} = [\mathfrak{a}_{ij}]$, $\mathfrak{B} = [\mathfrak{b}_{ij}]$. Equation (2) will be called the monolithic equation. Subscripts *n* and *n* + 1 refer to the subsequent time instants.

In the following we shall focus our attention on three solution algorithms of eq. (1): a direct one [in which X_{n+1} is directly evaluated from the monolithic equation (2)] and two iteration schemes – simultaneous and staggered ones – resulting from a partitioning of the monolithic equation (2).

Direct solution scheme

The error introduced when solving eq. (2) directly is of form: $\mathcal{E}_{n+1} = X(t_{n+1}) - \widetilde{X}_{n+1}$, where $X(t_{n+1})$ is the exact solution of eq. (1) and \widetilde{X}_{n+1} is the numerical solution of eq. (2). It satisfies the following equation:

(3)
$$(I+G) \mathcal{E}_{n+1} = H \mathcal{E}_n + r_{n+1}$$

The stability condition $\|(I + G)^{-1}H\| < 1$ and consistency property, $r_{n+1} = O(\Delta t^2)$ (see [3]) are sufficient for the direct process to be globally convergent (see [2]). $\|\cdot\|$ represents the spectral norm. The matrices G and H are defined in the following and r_{n+1} is the local truncation error.

Partitioned solution schemes

The partitioning of matrix \mathfrak{A} is chosen in such a way that the staggered scheme can be used:

(4)
$$\mathfrak{A} = \mathfrak{A}^{L} + \mathfrak{A}^{R},$$

(5)
$$\mathfrak{A}^{L} = \begin{bmatrix} \mathfrak{a}_{11} & \mathbf{0} \\ \mathbf{0} & \mathfrak{a}_{22} \end{bmatrix}, \qquad \mathfrak{A}^{R} = \begin{bmatrix} \mathbf{0} & \mathfrak{a}_{12} \\ \mathfrak{a}_{21} & \mathbf{0} \end{bmatrix}.$$

Then eq. (2) can be written as:

(6)
$$\mathfrak{A}^{L}X_{n+1} = \mathfrak{B}X_{n} + \hat{\mathfrak{f}} - \mathfrak{A}^{R}X_{n+1}.$$

ON CONSISTENCY, STABILITY AND CONVERGENCE ...

After some rearrangements we obtain the partitioned form of eq. (2): (7) $x_{n+1} = -G_1y_{n+1} + H_1X_n + \mathcal{J}_1$, $y_{n+1} = -G_2x_{n+1} + H_2X_n + \mathcal{J}_2$ where:

(8)
$$\begin{cases} \mathbf{X}_{n+1} = [\mathbf{x}_{n+1}, \mathbf{y}_{n+1}]^T, \\ \widehat{\mathbf{f}} = [\widehat{\mathbf{f}}_1, \widehat{\mathbf{f}}_2]^T, \quad \mathbf{f}_1 = \mathbf{a}_{11}^{-1} \widehat{\mathbf{f}}_1, \quad \mathbf{f}_2 = \mathbf{a}_{22}^{-1} \widehat{\mathbf{f}}_2, \\ G_1 = \mathbf{a}_{11}^{-1} \mathbf{a}_{12}, \quad G_2 = \mathbf{a}_{22}^{-1} \mathbf{a}_{21}, \\ H_1 = \mathbf{a}_{11}^{-1} [\mathbf{b}_{11}, \mathbf{b}_{12}], \quad H_2 = \mathbf{a}_{22}^{-1} [\mathbf{b}_{21}, \mathbf{b}_{22}]. \end{cases}$$

Matrices G_1 , G_2 depend on the time increment Δt .

Simultaneous iterations

The simultaneous iteration scheme for eq. (7) is defined by: (9) $x_{n+1}^k = -G_1 y_{n+1}^{k-1} + H_1 X_{n,K} + \mathcal{J}_1$, $y_{n+1}^k = -G_2 x_{n+1}^{k-1} + H_2 X_{n,K} + \mathcal{J}_2$ where k is the index of iteration, K is the last performed iteration K = K(n),

(10)
$$[\mathbf{x}_{n+1}^0, \mathbf{y}_{n+1}^0]^T = \mathbf{X}_{n+1}^0 = \sum_{i=0}^m \beta_i \mathbf{X}_{n-i}$$

and β_i are appropriate weighting coefficients.

Usually, the last obtained solution is used as the predictor. The error $E_{n+1,K} = X(t_{n+1}) - \tilde{X}_{n+1}^{K}$, where \tilde{X}_{n+1}^{K} is the numerical solution of eq. (9), can be written as (see [3]):

(11)
$$E_{n+1,K} =$$

= $(-G)^{K}E_{n+1,0} + (I - (-G)^{K})(I + G)^{-1}HE_{n,K} + (I - (-G)^{K})(I + G)^{-1}r_{n+1}$

where:

(12)
$$G = \begin{bmatrix} \mathbf{0} & G_1 \\ G_2 & \mathbf{0} \end{bmatrix}, \quad H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix}$$

and r_{n+1} is the local truncation error.

Sufficient conditions for the simultaneous process to be globally convergent are: stability $\|(I - (-G)^K)(I + G)^{-1}H\| < 1$, iteration convergence $\|G\| < 1$ and consistency $r_{n+1} = O(\Delta t^2)$.

Staggered iterations

The staggered iteration process is defined as follows. First we predict the value of x_{n+1} appearing on the r.h.s. of the second of eq. (7) $x_{n+1}^{(P)} = x_{n+1,0}$, and from this equation we calculate the value of $y_{n+1,0}$, then substitute it into the first equation of (7), which is solved for $x_{n+1,1}$. Now $x_{n+1,1}$ is used as a predictor in the second of eq. (7) and the whole process is repeated until a required tolerance τ is attained; we denote K as the number of performed iterations, K = K(n). The last step is to calculate

 $y_{n+1,K}$, then $X_{n+1} = [x_{n+1,K}, y_{n+1,K}]^T$. The predictor $x_{n+1,0}$ has the form:

(13)
$$\mathbf{x}_{n+1,0} = \sum_{i=0}^{m} \beta_i \, \mathbf{x}_{n-i,K} \, .$$

The following two alternative sets of equations represent the staggered scheme:

(14)
$$\mathbf{x}_{n+1,k} = -\mathbf{G}_1 \mathbf{y}_{n+1,k-1} + \mathbf{H}_1 \mathbf{X}_{n,K} + \mathbf{f}_1$$

(15)
$$y_{n+1,k-1} = -G_2 x_{n+1,k-1} + H_2 X_{n,K} + \mathcal{f}_2,$$

or:

(16)
$$x_{n+1,k} = G_1 G_2 x_{n+1,k-1} + (-G_1 H_2 + H_1) X_{n,K} - G_1 \mathcal{J}_2 + \mathcal{J}_1,$$

(17)
$$y_{n+1,K} = -G_2 x_{n+1,K} + H_2 X_{n,K} + \mathcal{f}_2$$

where eq. (16) is iterated and eq. (17) is used only once to evaluate $y_{n+1,K}$. The global error $e_{n+1,K} = X(t_{n+1}) - \tilde{X}_{n+1,K}$, where $\tilde{X}_{n+1,K}$ is the numerical solution of eq. (14) and eq. (15) satisfies (see [3]):

(18)
$$e_{n+1,K} = G^{2K}e_{n+1,0} + (I - G^{2K})(I + G)^{-1}He_{n,K} + (I - G^{2K})(I + G)^{-1}r_{n+1}$$
.
The sufficient conditions for global convergence are: iteration convergence $||G|| < 1$, stability $||Q_n|| = ||(I - G^{2K})(I + G)^{-1}H|| < 1$, and consistency $r_{n+1} = O(\Delta t^2)$. In the limit case $K \to \infty$ for $||G|| < 1$ the equations (11) and (18) have the same forms as the expression for the error \mathcal{E}_{n+1} of the direct method equation (3).

In concise form the error for the staggered scheme equation (18) can be written as:

(19)
$$e_{n+1,K} = P_{n+1}e_{n+1,0} + Q_{n+1}e_{n,K} + R_{n+1}r_{n+1}.$$

Recursively for n we have from eq. (18) that:

(20)
$$\boldsymbol{e}_{n+1,K} = \boldsymbol{P}_{n+1}\boldsymbol{e}_{n+1,0} + \sum_{l=0}^{n} \boldsymbol{Q}_{n+1} \dots \boldsymbol{Q}_{n+1-l} \boldsymbol{P}_{n-l} \boldsymbol{e}_{n-l,0} + \\ + \boldsymbol{Q}_{n+1} \dots \boldsymbol{Q}_{1} \boldsymbol{e}_{0,0} + \sum_{l=0}^{n} \boldsymbol{Q}_{n+1} \dots \boldsymbol{Q}_{n+1-l} \boldsymbol{R}_{n-l} \boldsymbol{r}_{n-l} + \boldsymbol{R}_{n+1} \boldsymbol{r}_{n+1} .$$

We can recognise that $e_{0,0}$ is the initial round-off error made at the starting step. Notice that the local truncation error r_{n+1} is the same for all the above errors.

The conditions of convergence and stability for each of the solution schemes are independent and it is generally not possible to replace one by the other. In [3] it has been shown that if G has real positive eigenvalues, and the iterations are convergent, since ||G|| < 1, then the condition of stability for the staggered scheme can be replaced by the stability condition for the simultaneous iteration scheme with K = 1, *i.e.* ||H|| < 1.

For the simultaneous scheme, the assumption of real and positive eigenvalues of *G* permits us to check the stability condition for one iteration K = 1, instead of checking it for the performed number of iterations, because then $||(I - (-G)^K)(I + G)^{-1}|| < 1$ so also $||(I - (-G)^K)(I + G)^{-1}H|| < 1$.

For the staggered scheme, even for G with real and positive eigenvalues and $\|G\| < 1$, the stability condition for one iteration *i.e.* $\|(I - G)H\| < 1$, does not give

stability for arbitrary K. This is caused by the fact that $\|(I - G^{2K})(I - G^2)^{-1}\|$ may be larger than 1, so $\|(I - G^{2K})(I + G)^{-1}H\| = \|(I - G^{2K})(I - G^2)^{-1}(I - G)H\|$ may also be larger than 1.

Other convergence conditions

In the case of partitioned solutions some authors (see [4-6]) verify stability in a different way than the one presented above. First they assume the form of the predictor, then they check the stability of an equation which is obtained from the partitioned equation (6) by substituting into it the specific form of the predictor.

Let us illustrate this on an example, with the predictor taken as the last obtained solution $X_{n+1}^0 = \widetilde{X}_n^K$. Examples of this type of predictor can be found in [1, 4-7]. Equation (7) with the above predictor becomes:

(21)
$$X_{n+1} = -GX_n + HX_n + \mathcal{J}$$

and the stability condition is:

$$\|-G+H\|<1$$

It is easily seen, that if G and H have real and positive eigenvalues and satisfy eq. (22) and ||G|| < 1, then also the condition for stability is fulfilled for the staggered scheme *i.e.* $||(I - G^{2K})(I + G)^{-1}H|| < 1$. For the simultaneous scheme in addition an even number of iterations K is required, because then $||(I - (-G)^K)(I + G)^{-1}H|| < 1$. Generally, for arbitrary G and H, the above facts are not true.

To achieve a given accuracy it has often been recommended to cut the time step Δt instead of performing more iterations K. This is permissible only if the procedure is globally convergent for a constant number of performed iterations K. We recall that the process is globally convergent when the global error $e_{n,K} = X(t_n) - \tilde{X}_n$ tends to zero *i.e.*

(23)
$$\|e_{n,K}\| \xrightarrow{\Delta t \to 0}_{\substack{\Delta t \to 0 \\ n \to \infty}} 0 \text{ with } n \Delta t = \text{const}, \quad K = \text{const}$$

and for all solutions \widetilde{X}_n satisfying the starting conditions, $\widetilde{X}_{0,0} = X(t_0)$. In $(2^{\frac{1}{3}}) \|e_{n,K}\|$ is a norm of the global error vector.

In paper [3] it has been shown that the global error is bounded but does not tend to zero. This fact suggested that the partitioned procedure is convergent for a fixed number of iterations, but to a solution of another equation. In the following we find the equation in the case of the last solution predictor. Following the same procedure such equations can be found for other predictors of type eqs. (10), (13).

For the time discretization we have chosen the generalised mid-point (GM-P) method because it is a member of many classes of one-step algorithms, *i.e.* implicit single step methods [2], SSPJ methods (SSPJ-single step *p*-order polynomial, j = 1, the order of the equation) [8], Runge-Kutta methods (see [9]) or a degenerate case of collocation schemes [10]. It also has good stability properties, it is unconditionally stable (A-stable) for $\theta \ge 1/2$ and is nonlinearly B-stable, contrary to the trapezoidal

method which is not B-stable (see [11]). The GM-P method is constructed with the use of:

(24)
$$\dot{X}_{n+\theta} = (X_{n+1} - X_n)/\Delta t,$$

(25)
$$X_{n+\theta} = (1-\theta)X_n + \theta X_{n+1}$$

where θ is a parameter, usually $0 < \theta < 1$. Substituting eq. (24), (25) into eq. (1) at time instant $t_{n+\theta}$ we obtain:

(26)
$$[\boldsymbol{B} + \theta \Delta t \boldsymbol{C}] \boldsymbol{X}_{n+1} = [\boldsymbol{B} - (1-\theta) \Delta t \boldsymbol{C}] \boldsymbol{X}_n + \Delta t \boldsymbol{F}_{n+\theta}$$

thus:

(27)
$$\mathfrak{A} = \mathbf{B} + \theta \Delta t \mathbf{C}$$
, $\mathfrak{B} = \mathbf{B} - (1 - \theta) \Delta t \mathbf{C}$, $\widehat{\mathfrak{f}} = \Delta t \mathbf{F}_{n+\theta}$.

Let us find the equation which is consistent with the equation representing the simultaneous scheme for one performed iteration K = 1 *i.e.* with eq. (6). For this purpose we expand $X(t_{n+1})$ and $X(t_n)$ in Taylor series about $t_{n+\theta}$:

(28)
$$X(t_{n+1}) = X(t_{n+\theta}) + (1-\theta)\Delta t \dot{X}(t_{n+\theta}) + O(\Delta t^2),$$

(29)
$$\mathbf{X}(t_n) = \mathbf{X}(t_{n+\theta}) - \theta \Delta t \, \dot{\mathbf{X}}(t_{n+\theta}) + \mathbf{O}(\Delta t^2)$$

and substitute eqs. (28), (29) into:

(30)
$$\mathfrak{A}^{L}X(t_{n+1}) = \mathfrak{B}X(t_{n}) + \widehat{\mathfrak{f}}_{n+\theta} - \mathfrak{A}^{R}X(t_{n+1}) + \mathfrak{A}^{L}r_{n+1}.$$

The obtained equation with condition $r_{n+1} = O(\Delta t^2)$, yields that eq. (6) is consistent with the following equation:

(31)
$$B^{L}\dot{X} + CX = F$$

where $B^{L} = \begin{bmatrix} b_{11} & \mathbf{0} \\ \mathbf{0} & b_{22} \end{bmatrix}$.

This means that if we perform only one iteration and diminish Δt then we calculate the solution of eq. (31), not eq. (1).

Similarly, supposing that only two iterations of eq. (9) are performed and then the time step Δt is diminished, we obtain a convergent process (one with a decreasing residual), but to the following equation:

$$B^L \dot{X} + B(B^L)^{-1} C X = F.$$

Applying the same method we can construct an equation the solution of which is the limit of the staggered scheme eqs. (14), (15) with one performed iteration. This equation reads:

$$(33) \qquad (B-B^*)\dot{X}+CX=F$$

where $B^* = \begin{bmatrix} B_{12}B_{22}^{-1}B_{21} & 0\\ 0 & 0 \end{bmatrix}$.

As we see eq. (33) differs from eq. (1) by term B^* .

Conclusions

When matrix G has real and positive eigenvalues and the iterations are convergent, then the stability condition for the simultaneous scheme with K = 1 is equivalent to the staggered stability condition for arbitrary K or the simultaneous stability condition for arbitrary K.

For the staggered scheme when G has real and positive eigenvalues the stability condition for K = 1 does not give the stability condition for arbitrary K. For arbitrary B and C (*e.g.* in the non-linear case) both of the conditions: convergence and stability, must be verified for the K used in the calculations.

The stability condition obtained from the equation with the last solution predictor, for G with real and positive eigenvalues, gives convergence and stability for the simultaneous and staggered schemes with predictor of general form eqs. (10), (13).

For a fixed number of iterations the staggered and simultaneous schemes give a solution to another differential equation than the one originally considered.

References

- L. SIMONI B. A. SCHREFLER, A staggered F. E. solution for water and gas flow in deforming porous media. Commun. Appl. Num. Methods, 7, 1991, 213-223.
- [2] J. D. LAMBERT, Computational Methods in Ordinary Differential Equations. J. Wiley & Sons, London 1973.
- [3] E. TURSKA K. WISNIEWSKI B. A. SCHREFLER, Error propagation of staggered solution procedures for transient problems. Accepted for publication in Comp. Meth. in Appl. Mech. and Engng.
- [4] K. C. PARK, Stabilisation of partitioned solution procedure for pore fluid-soil interaction analysis. Int. J. Num. Meth. in Engng., 19, 1983, 1669-1673.
- [5] K. C. PARK, Partitioned transient analysis procedures for coupled-field problems: stability analysis. ASME J. Appl. Mech., 47, 1980, 370-376.
- [6] O. C. ZIENKIEWICZ D. K. PAUL A. H. C. CHAN, Unconditionally stable staggered solution procedure for soil-pore fluid interaction problems. Int. J. Num. Meth. in Engng., 26, 1988, 1039-1055.
- [7] K. C. PARK C. A. FELIPPA, Partitioned analysis of coupled systems. In: T. BELYTSCHKO T. J. R. HUGH-ES (eds.), Computational Methods for Transient Analysis. Elsevier Science Publishers B.V., 1983.
- [8] O. C. ZIENKIEWICZ R. L. TAYLOR, The Finite Element Method. Vol. 2, McGraw-Hill, London 1991.
- [9] C. W. GEAR, Numerical Initial Value Problems in Ordinary Differential Equations. Prentice-Hall Inc., Engelwood Cliffs, N.J. 1971.
- [10] T. J. R. HUGHES, The Finite Element Method. Prentice-Hall Inc., Engelwood Cliffs, N.J. 1987.
- [11] G. WANNER, A short proof on non-linear A-stability. BIT, 16, 1976, 226-227.

B. A. Schrefler: Istituto di Scienza e Tecnica delle Costruzioni Università degli Studi di Padova Via F. Marzolo, 9 - 35131 Padova

> E. Turska: Polish Academy of Sciences - IFTR ul. Swietokrzyska 00-049 WARSZAWA (Polonia)