Relationship between Eigenvalues and Size of Time Step in Computer Simulation of Thermomechanics Phenomena

Elzbieta Gawronska, Norbert Sczygiol

Abstract—The main idea of this article is the use of eigenvalues of matrices to determine the size of time step in modeling of solidification. As far as numerical simulations are concerned it is very important to obtain solutions which are stable and physically correct. It is acquired by fulfilling many assumptions and conditions when building a numerical model and carrying out computer simulations. One of the conditions is a proper selection of time step. The size of time step has a great impact on the stability of used schemes of time integration (e.g. explicit scheme), or on a proper image of physical phenomena occurring during the simulation (e.g. implicit scheme). The eigenvalues of matrices in main equations influence the appropriate selection of size of time step in computer simulations.

Index Terms—eigenvalues, numerical methods, computer simulation, solidification processing

I. INTRODUCTION

A NUMERICAL modeling of solidification is known to be a very time consuming task. The constantly increasing demand for efficient and precise computational solvers becomes the factor that decides about usability of a given solidification simulation software. In many cases practitioners require multiple scenarios to be tested, e.g. for different input parameters, before they make a final decision about the setup of a given technological process. At the same time increasing size of computer memory makes it possible to consider problems with increasing size, which in turn results in increased precision of simulations. There are several possible ways to tackle this kind of problems. For instance, one can use parallel computers or accelerated architectures such as GGPUs or FPGAs. This solution however requires another level of expertise in both, parallel hardware and software, which very often is not easily available. In papers [1], [2] we proposed another method, which relies on the application of the technique called mixed time partitioning. Our approach exploits the fact that physical processes inside a mould are of different nature than those in a solidifying casting. As a result different time steps can be used to run computations within both sub-domains. Because processes that are modeled in the casting subdomain are more dynamic they require very fine-grained time step. On the other hand a heat transfer within the mould sub-domain is less intense, and thus coarse-grained step is sufficient to guarantee desired precision of computations. Obviously, increasing length of a single time step results in decreased computational load, which in turn greatly improves performance of our approach. In this paper we put emphasis on determination of stability criteria for the selected integration method. Mathematical apparatus of the chosen stability analysis method is applied for the homogeneous form of the semidiscretized equation of solidification, as the stability is independent of the inhomogeneous part. The analysis of numerical stability of mixed time integration methods for the dynamics of structures and for heat conduction problem was adopted to the solidification problem with temperature-dependent material properties.

II. SOLIDIFICATION IN TERMS OF THE FINITE ELEMENT METHOD

The finite element method facilitates the modeling of many complex problems. Its wide application for founding comes from the fact that it permits an easy adaptation of many existing solutions and technics of modeling of solidification. As it comes to computer calculations there is a need to use discrete models, which means problems must be formulated by introducing time-space mesh. These methods convert given physical equations into matrix equations (algebraic equations). These equations usually contain many thousands of unknowns, that is why the efficiency of method applied to solve them is crucial.

After essential transformations we obtain an ordinary differential equation containing only the time derivative [3]

\[ M(T) \dot{T} + K(T)T = b(T), \]

where \( M \) is the capacity matrix, \( K \) is the conductivity matrix, \( T \) is temperature vector and \( b \) is right-hand side vector values of which are calculated using boundary conditions. The global form of these matrices is obtained by adding up components through all the finite elements. The components are defined for a single finite element as

\[ M = \sum_c \int_{\Omega} c^e N^T N \, d\Omega, \]

\[ K = \sum_c \int_{\Omega} \lambda \nabla N^T \nabla N \, d\Omega, \]

\[ b = \sum_c \int_{\Gamma} N^T_{\Gamma} q^T n \, d\Gamma, \]

where \( N \) is a shape vector in the area \( \Omega \), \( N_{\Gamma} \) is a shape vector on the edge \( \Gamma \), \( n \) is an ordinary vector towards the edge \( \Gamma \), and \( q \) is vector of nodal fluxes.
The system of ordinary differential equations (1) containing time derivative was obtained as a result of space integration and it may only be solved in approximation. In order to receive an approximate solution it is needed to use the division of time interval \((0, t_{max})\) into subintervals \((t_k, t_{k+1})\) with the length of \(\Delta t_k = t_{k+1} - t_k\), and time integration is performed by the use of single or multisteps methods [4]. One of the single step schemes which can be used is the scheme \(\Theta\) of the form

\[
T^{n+1} = T^n + (-M^{n+\Theta})^{-1}K^{n+\Theta}T^n + (M^{n+\Theta})^{-1}b^{n+\Theta})(1 - \Theta)\Delta t + (-M^{n+\Theta})^{-1}K^{n+\Theta}T^{n+1} + (M^{n+\Theta})^{-1}b^{n+\Theta}\Theta\Delta t.
\]

(5)

Due to a possible dependence of materials properties from the temperature, namely \(M\), \(K\) and \(b\) from unknown, for \(\Theta \neq 0\) it is a system of nonlinear equations. To solve the system, iterative methods must be used.

The front Euler scheme

\[
M^nT^{n+1} = M^nT^n - \Delta tK^nT^n + \Delta tb^n,
\]

(6)

is obtained for one extreme value \(\Theta = 0\) and the backward Euler scheme

\[
(M^{n+\Delta t}K^{n+1})T^{n+1} = M^{n+1}T^n + \Delta tb^{n+1},
\]

(7)

is obtained for the other extreme value \(\Theta = 0\). And if the values of matrices coefficients \(M\) i \(K\) in the equation (7) are evaluated on the level of previous time step then a modified backward Euler scheme is obtained [4].

\[
(M^n + \Delta tK^n)T^{n+1} = M^nT^n + \Delta t b^{n+1}.
\]

(8)

### III. Basic Equations

In the experimental part explicit capacitive formulation of solidification is used

\[
\nabla \cdot (\lambda \nabla T) = c' T \frac{\partial T}{\partial t}.
\]

(9)

The equation (9) is solved by mixed time partitioning method considering:

1) semi-discretization,
2) initial-value problem which consists of given functions \(T = T(t)\) satisfying the equation (9) and being the part of initial conditions \(T(t = 0) = T_0\) for \(t \in (0,T)\), \(T > 0\),
3) one step scheme \(\Theta\) of time integration.

The finite elements mesh consists of two groups: \(A\) - connected with a mould and \(B\) - connected with a casting. Each of these can be integrated with the use of different schemes of time integration. Finite elements lying by the dividing line have no common nodes. This fact simplifies finding of the maximum admissible time step and the stability analysis. If this division is assumed then it may be written as

\[
M_A = \sum_{e=1}^{8} M_e, \quad K_A = \sum_{e=6}^{8} K_e, \quad b_A = \sum_{e=6}^{8} b_e,
\]

\[
M_B = \sum_{e=1}^{5} M_e, \quad K_B = \sum_{e=1}^{5} K_e, \quad b_B = \sum_{e=1}^{5} b_e.
\]

(10)

All vectors are also divided into parts according to finite elements mesh division \(T = (T_A T_B)^T\), \(\dot{T} = (\dot{T}_A \dot{T}_B)^T\), the upper index \(T\) represents transportation. As above, vector \(\dot{T}\) may be written in this scheme

\[
\dot{T} = v_A + v_B, \quad v_A = M^{-1}(b_A - K_A T), \quad v_B = M^{-1}(b_B - K_B T).
\]

(11)

In the area connected with a mould the integration is carried out with a bigger time step \((m\Delta t)\) whereas in the area connected with a casting with a smaller time step \((\Delta t)\). This allows to build a system of equations on the basis of (1) and solving it separately for the sub-domain \(B\) elements, that is to carry out calculations for it more often than for the whole mesh.

### IV. General Outline of Numerical Method Stability

Numerical method is stable when a little error in any solution stage moves further with a decreasing value. An error appearing on time level \(n\) may be defined as \(\varepsilon^n\), on time level \(n + 1\) as \(\varepsilon^{n+1}\), whereas values of this error may be determined with equation [1]

\[
\varepsilon^{n+1} = g\varepsilon^n,
\]

(12)

where \(g\) is amplification factor connected with integral operator \(T(\Delta t, \Delta)\) that is time integration scheme. The amplificatory factor refers to a method error connected with this scheme. That is why for stability of the method one conditions must be fulfilled: the value of an error on time level \(n + 1\) must not be bigger than value of an error on time level \(n\).

That may be written in this formula

\[
|\varepsilon^{n+1}| \leq |\varepsilon^n|,
\]

(13)

and using the definition of amplification factor (12)

\[
|g\varepsilon^n| \leq |\varepsilon^n|.
\]

(14)

It follows that numerical stability may be achieved if condition

\[
|g| \leq 1.
\]

(15)

is fulfilled. This condition is limited to issues leading to finite solutions.

For the system \(N\) of ordinary differential first-order equations an error vector is defined as \(\varepsilon^n\). Each coordinate of this vector is an error connected with an appropriate dependent variable of the system. For each time step an error is multiplied by amplification matrix \(G\) in order to obtain an error vector for a newer time step.

\[
\varepsilon^{n+1} = G\varepsilon^n.
\]

(16)

Amplification matrix is connected to an integral operator which couples solutions in consecutive time steps. It means that if an error \(\varepsilon^n\) appeared in a solution \(T^n\) on time level \(n\) then after some necessary transformations is obtained

\[
T^{n+1} + \varepsilon^{n+1} = T\left(T^n + \varepsilon^n\right).
\]

(17)

Assuming that an error vector has a small amplitude, the equation (17) may be expanded into the Taylor series, taking into account only its two first terms. After some transformations an expression joining together two time levels is obtained

\[
\varepsilon^{n+1} = \left\{ \frac{\partial}{\partial T}(T^T) \right\}^n \varepsilon^n.
\]

(18)
This expression also defines the amplification matrix in the equation (16). The operator on the right-hand side of this equation is a linear matrix operator. Using the given integration scheme it is possible to determine a amplification matrix for it.

Thanks to the knowledge of a amplification matrix referring to amplitudes of each other by appropriate eigenvalues $\varepsilon_i$ and $\lambda_i$, connected to each other by appropriate eigenvalues $g_i$ of matrix strength may be written as

$$\varepsilon_i^{n+1} = g_i \varepsilon_i^n. \quad (19)$$

Stability condition must be used separately for amplitudes of each error eigenvector

$$|\varepsilon_i^{n+1}| \leq |\varepsilon_i^n|,$$

for all $i$, that is

$$|g_i| \leq 1. \quad (21)$$

Stability criterion defined in a given way is limited to a demand that each eigenvalue $g_i$ of a matrix strength $G$ was smaller or equal to a unit. In the paper this condition is used for the stability analysis of the mixed time partitioning method of solidification issues.

V. RELATIONSHIP BETWEEN THE EIGENVALUES OF THE MATRICES AND THE SIZE OF TIME STEP

It is essential to find the criterion to determine the size of time step for the explicit scheme. $\Theta = 0$ is accepted and the equation (33) is reduced to the form

$$T^{n+1} = (I - \Delta tM^{-1}K)T^n. \quad (22)$$

The equation (22) is called the equation of evolution, because it gives the possibility to obtain the value of searched size $T$ at the time level $n + 1$ from appropriate values of nodal quantities at the time level $n$.

In the evolution equation the capacity matrix $M$ can be full or diagonal. It depends from the form of such a matrix in the evolution equation, conducting the analysis of the numerical stability is connected with executing some operations on matrices. In case of the capacity matrix is diagonal matrix, the calculation of the inverse matrix, namely $M^{-1}$, is very simple and then finding its eigenvalues, necessary to determine the critical value of time step, is not difficult. However in case of full capacity matrix which is symmetric and positively definite, in order to determine the inverse matrix to the matrix $M$ use the distribution $M = LL^T$ or homothetic transformation which keep the eigenvalues of full matrix, for which the inversion process and searching eigenvalue, which decides about the maximum, acceptable value of time step, is less complicated. The evolution equation (22) after converting can be written as follows

$$T^{n+1} = GT^n, \quad (23)$$

where amplification matrix $G$ datum is given as

$$G = I - \Delta tM^{-1}K. \quad (24)$$

The algorithm is explicit if the size $T^{n+1}$ can be received from the equation (33), without solving the system of algebraic equations and if the updates of searched size can be repeated $m$-times according to the formula [5]

$$T^{n+m} = G^mT^n. \quad (25)$$

A sufficient condition of numerical stability is to find the maximum eigenvalues of amplification matrix

$$Gx = \lambda x, \quad (26)$$

where $G$ is the matrix of $N$ degree, and $N$ is the number of nodes of sub-domain of the mesh for the elements connected with casting region or mould. The analysis of numerical stability is conducted separately for each sub-domain on the basis of finite elements inside the sub-domain.

Using the theory of eigenvalues and eigenvectors and the operations on matrices it is known that the size $G^mT^n \to 0$, if $m \to \infty$ for any $T^n \in R^N$, if $|\lambda_i| < 1$ for $i = 1 \ldots N$. Moreover, the size $G^mT^n$ is limited, if $m \to \infty$, if $|\lambda_i| \leq 1$ for $i = 1 \ldots N$, if there are linearly independent eigenvectors $x_i$ for each $|\lambda_i| = 1$, which is satisfied because of symmetry of matrices $G$.

After substituting $G$ from the equation (24) to the equation (26), multiplying this equation by $M$ and doing transformations the formula for generalized problem of eigenvalues is received

$$Kx = \frac{1 - \lambda}{\Delta t} Mx, \quad (27)$$

where $(1 - \lambda)/\Delta t$ is an eigenvalue of couple of matrices $K$ and $M$. From the equation (27) it is known that if $\lambda$ is equal to the unity then $Kx = 0$ only if $x = 0$.

There is true notation

$$\lambda_i = 1 - \Delta t\mu_i. \quad (28)$$

As $|\lambda_i| \leq 1$, the size of time step, which can be used to solve the system of equations (33), to be numerical stable and is limited by the inequality

$$\Delta t \leq \frac{2}{\mu_{\text{max}}}. \quad (29)$$

The most restrictive limitation of the size of time step, which assures about the stability is the case in which $\mu_i$ is the maximum eigenvalue $\mu_{\text{max}}$ of the matrix of equation (37). Taking into account the way of assembly of capacity and conductivity matrices, the equation (37) may be written for a definite element $e$ of the given area

$$K^{(e)}x^{(e)} = \mu^{(e)}M^{(e)}x^{(e)}, \quad (30)$$

wheras the limitation of a size of time step may be written as follows

$$\Delta t \leq \frac{2}{\mu_{(e)}^{(e)}}. \quad (31)$$

In order to find a maximum acceptable size of time step for the casting and mould regions it is necessary to determine, for all the elements, their biggest eigenvalues and create from them double inequality (details in section VII). This inequality is limited from the bottom by the smallest and from the top by the biggest of them

$$\mu_{\text{min}}^{(e)} \leq \mu \leq \mu_{\text{max}}^{(e)}, \quad (32)$$

where $e = 1 \ldots ne$, and $ne$ is the number of elements in the considered region.
VI. CRITERION OF DESIGNATING CRITICAL TIME STEP

In order to determine the criterion of stability of numerical method, operations converting this equation into general problem of the eigenvalues are conducted. The analysis of stability is carried out to determine the maximum size of time step, which exceeding may be cause of unsteady solutions.

The one step scheme of time integration of the equation obtained after space discretization is presented by the formula

$$\begin{align*}
(M + \Theta \Delta t K)T^{n+1} &= (M - (1 - \Theta) \Delta t K)T^n.
\end{align*}$$

(33)

The right-hand side vector is not taken into consideration because the homogenous equation is only essential for the numerical stability. If the homogenous expression is stable so the inhomogenous one is also stable [4].

The generalized problem of the eigenvalues is connected with the homogenous form of the equations for casting region $B$, connected with sub-cycle and mould $A$, associated with total cycle $[2]$, $[1]$, it can be written in the universal form

$$\begin{align*}
Ax_i &= \lambda_i Bx_i,
\end{align*}$$

(34)

where $N$ is a grade of the matrix $A$ i $B$ expressed by the formulas

$$\begin{align*}
A &= M - (1 - \Theta^{(c)}) \Delta t K, \\
B &= M + \Theta^{(c)} \Delta t K,
\end{align*}$$

(35)

(36)

for

$$\begin{align*}
\Theta^{(c)} &= \Theta_A \quad \text{dla } e \in A, \\
\Theta^{(c)} &= \Theta_B \quad \text{dla } e \in B.
\end{align*}$$

After substituting the equations (35), (36) into the formula (34) and doing the transformations the expression is received

$$\begin{align*}
Kx_i &= \mu_i Mx_i,
\end{align*}$$

(37)

in which $\mu_i$ is the eigenvalue of coupled matrices of $M$ i $K$ form

$$\begin{align*}
\mu_i &= \frac{1 - \lambda_i}{(1 - \Theta + \Theta \lambda_i) \Delta t}.
\end{align*}$$

(38)

Using previous formulas the condition was developed which is enough for the stability of single step method $\Theta$ used in the prime row equation (9). The homogenous equation (9) is obtained in result of space discretization, after transformation, it can be written as follows

$$\begin{align*}
\dot{T} + BT &= 0,
\end{align*}$$

(39)

where $B = M^{-1} K$. Naturally, such inversion of the matrix $M$ would cause its asymmetry, that is why such explicit inversion is not used, but for such a purpose e.g. Cholesky decomposition is used.

The one step method $\Theta$ is used in a scalar equation, which comes from the modal decomposition of system of equations (39), gives

$$\begin{align*}
T^{n+1} &= \lambda T^n,
\end{align*}$$

(40)

where the eigenvalue $\lambda$ is expressed by the formula

$$\begin{align*}
\lambda &= \frac{1 - (1 - \Theta) \mu \Delta t}{1 + \Theta \mu \Delta t}.
\end{align*}$$

(41)

As far as the equation (41) and the inequality $|\lambda| \leq 1$ are concerned, the stability of the method is obtained if the following condition is satisfied

$$\begin{align*}
2 + (2 \Theta - 1) \mu \Delta t &\geq 0.
\end{align*}$$

(42)

It arises from (42) that for $\Theta \geq 1/2$ the condition of the inequality is always satisfied, so the method is stable. Moreover, for $\Theta < 1/2$ the stability of the method depends on the size of quotient $\mu \Delta t$, because of that for the explicit scheme ($\Theta = 0$) the size of maximum and accessible time step is strictly connected with the maximum eigenvalue in a given area (the casting, the mould).

VII. RESTRICTIONS IMPOSED ON THE EIGENVALUES


$$\begin{align*}
MT + KT &= 0.
\end{align*}$$

(43)

Substitute $T = e^{-\lambda T} v$ into the equation (43), and receive equal system of equations

$$\begin{align*}
\lambda M v &= K v.
\end{align*}$$

(44)

Because of the semi-discretization the equation (44) is satisfied for $\lambda = \lambda_i$ and $v = v_i$.

The mass matrix $M$ is diagonal and it helps to reduce the analysis of stability. If this matrix is the full symmetric matrix, the analysis of stability of the equation is conducted in a different way, however, the effect of both operations is the same as the criterion limiting the size of time step in the explicit scheme of the integration.

If the matrix $M$ is positively definite, Cholesky decomposition can be executed, namely $M = LL^T$, where $L$ is lower triangular and nonsingular matrix. Using such distribution in the equation (44) and multiplying both sides of the equation by $L^{-1}$, it is obtained that

$$\begin{align*}
\lambda_i L^T v_i &= L^{-1} K (L^{-1})^T L^T v_i,
\end{align*}$$

(45)

where $L^T v_i$ is the eigenvector, and $\lambda_i$ is the eigenvalue of the symmetric matrix $P = L^{-1} K (L^{-1})^T$. The matrix $P$ has the set of linearly independent eigenvalues $v_i$.

If the matrix $V$ is composed of $v_i$, which are the columns of such a matrix and $L^T V$ is orthogonal, it can be written that

$$\begin{align*}
V^T L L^T V &= V^T M V = I.
\end{align*}$$

(46)

Moreover, on the basis of the equation (44) and the Cholesky decomposition process it can be written that

$$\begin{align*}
\lambda_i &= \lambda_i V_i^T M V_i = V_i^T K V_i.
\end{align*}$$

(47)

Substituting $T = V x$ into the equation (43) and left-multiplying both sides by $V^T$ it is obtained that

$$\begin{align*}
V^T M V x + V^T K V x &= 0,
\end{align*}$$

(48)

and then

$$\begin{align*}
I x + A x &= 0,
\end{align*}$$

(49)

where $A = \text{diag}(\lambda_i)$. Such distribution is known as modal decomposition and allows to write the system of equations in the scalar form.

$$\begin{align*}
\dot{x}_i + \lambda_i x_i &= 0.
\end{align*}$$

(50)
The problem of the stability is connected with some restrictions of the eigenvalues. For the problems described by the prime row equations, from the equation (44) the eigenvalues and the eigenvectors can be designated. However, the restrictions imposed on the eigenvalues in the equation (47) can be derived from Rayleigh quotient

\[ \lambda = \frac{v^T K v}{v^T M v}. \]  

(51)

Taking into consideration the way of matrix assembler K and M

\[ \lambda = \frac{\sum_i (v_i^T K_i v_i)}{\sum_i (v_i^T M_i v_i)}, \]  

(52)

where \( e \) is an element, and \( v_i \) is appropriate component of the eigenvector, Rayleigh quotient for an element can be written as follows

\[ \lambda_i^{(e)} = \frac{v_i^T K_i^{(e)} v_i}{v_i^T M_i^{(e)} v_i}. \]  

(53)

Inserting (53) into (52) and doing certain transformations it is obtained that

\[ \lambda = \frac{\sum_i \alpha_i \lambda_i^{(e)}}{\sum_i \alpha_i}, \]  

(54)

where \( \alpha_i = v_i^T M_i^{(e)} v_i > 0 \), because the capacity matrix is positively definite. It is resulted from the equation (54) that \( \lambda \) is determined as the weighted average from \( \lambda_i^{(e)} \) with positively weight, so the restrictions resulting from Rayleigh quotient can be written as follows

\[ \lambda_{\min}^{(e)} \leq \lambda \leq \lambda_{\max}^{(e)}. \]  

(55)

Estimation of the extreme values is received from the formulas

\[ \lambda_{\min}^{(e)} \leq \frac{\min \{v_i^T K_i^{(e)} v_i\}}{\max \{v_i^T M_i^{(e)} v_i\}}, \]  

\[ \lambda_{\max}^{(e)} \leq \frac{\max \{v_i^T K_i^{(e)} v_i\}}{\min \{v_i^T M_i^{(e)} v_i\}} \]  

(56)

(57)

VIII. CONCLUSION

The eigenvalues remain with close relation to the stability of numerical method and hence with the size of the time step. For explicit schemes of time integration such a step cannot exceed a certain critical value. For implicit schemes of time integration the size of the time step cannot be unlimited because exceeding certain limit can result in omission of important physical phenomena. The use of the analysis of the relation between the eigenvalues and the size of time step allows to designate the maximum permissible size of the time step and to conduct the computer simulations correctly. The problem of the eigenvalues of the matrices is a very extensive issue and the works have very deep scientific and practical justification.