COMMUNICATIONS IN NUMERICAL METHODS IN ENGINEERING
Commun. Numer. Meth. Engng 2000; **00**:1–6 *Prepared using* cnmauth.cls *[Version: 2000/03/22 v1.0]*

Modified Mass Matrices and Positivity Preservation for Hyperbolic and Parabolic PDEs

M. Berzins*

Computational PDEs Unit, School of Computing, The University of Leeds, Leeds LS2 9JT, UK.

SUMMARY

Modifications to the standard finite element mass matrix are considered with the aim of preserving the positivity of the discrete solution. The approach is used in connection with calculating the initial time derivative values for parabolic equations and in connection with nonlinear Petrov-Galerkin schemes for hyperbolic equations in one space dimension. The extension of the ideas to unstructured meshes in two and three space dimensions is indicated. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: *Positivity Preservation* ; Mass Matrices, Finite Element Methods

1. Introduction

There are many situations in the numerical solution of partial differential equations in which the computed solution values should, on physical grounds, remain non-negative. One the simplest examples is that of the simple advection equation with non-negative initial data while other cases are those of concentrations of chemical compounds in reacting flow calculations. In the latter case preserving positivity is essential to avoid the numerical calculation becoming meaningless. Consider the solution of the advection equation with appropriate initial and boundary condition by using the standard Galerkin method with linear basis (hat) functions $\phi_i(x)$ on a uniformly spaced mesh $x_i, i = 1,\ldots,N$ to get

$$\int_{x_{i-1}}^{x_{i+1}} \frac{\partial U}{\partial t}\phi_i(x)\, dx = \int_{x_{i-1}}^{x_{i+1}} -\frac{\partial U}{\partial x}\phi_i(x)\, dx, i = 1,\ldots,N, \tag{1}$$

where the approximate solution to this p.d.e. as defined by $U(x,t) = \sum_{i=1}^{N} \phi_i(x)U_i(t)$ where $\phi_i(x_j) = \delta_{ij}$. Evaluating the integrals gives rise to the numerical scheme defined by

$$\frac{1}{6}\left[\dot{U}_{i-1} + 4\dot{U}_i + \dot{U}_{i+1}\right] = \frac{-1}{2\delta x}(U_{i+1} - U_{i-1}) \tag{2}$$

*Correspondence to: M.Berzins, School of Computing, The University of Leeds, Leeds LS2 9JT, UK.

Received 10 January 2001
Revised 22 March 2001

where δx is the uniform mesh spacing in this case and where $\dot{U}_i = \frac{dU_i}{dt}$. Defining the time-dependent vector $\underline{U}$ by $\underline{U} = [U_1, \ldots, U_N]^T$ allows this system of equations to be rewritten in the form

$$A\underline{\dot{U}}(t) = \underline{F}(\underline{U}(t)) \tag{3}$$

where the matrix A is referred to as the mass matrix.

It is well-known that this scheme is unsatisfactory in a very similar way to that of linear central difference schemes, [8]. Many modified Galerkin methods have been proposed to remedy this situation. A survey of such methods is given in [8] and includes Streamline Upwind Petrov-Galerkin (SUPG) methods [7] in which the test functions are modified to improve the behaviour of the method and Discontinuous Galerkin (DG) methods [4, 7] in which discontinuous basis functions are used. There are many other approaches such as the modified Petrov-Galerkin method of Cardle [3] in which the test function is modified differently for the spatial and temporal terms. In this case the numerical scheme that results is given by

$$\dot{U}_i + \frac{1}{6}(1-\beta)\left[\dot{U}_{i-1} - 2\dot{U}_i + \dot{U}_{i+1}\right] = \frac{-1}{2\delta x}(U_{i+1} - U_{i-1}) + \frac{\alpha}{2\delta x}\left[U_{i-1} - 2U_i + U_{i+1}\right] \tag{4}$$

where β and α are the constants multiplying the Petrov-Galerkin additional polynomials in time (cubic polynomial) and space (quadratic polynomial), see [3].

In the case of many of these methods it is clear that the magnitude of unphysical values is not as large as with the standard Galerkin method and in the case of DG methods the mass matrix is the identity matrix; this makes it much easier to prove properties such as positivity preservation. The definition used here for a positivity preserving scheme for the advection equation is one (see [1]) for which the numerical solution at time t_{n+1} may be written in terms of the numerical solution at time t_n in the form

$$U_i(t_{n+1}) = \sum_j a_j U_j(t_n) \text{ where } \quad \sum_j a_j = 1, \text{ and } a_j \geq 0\,. \tag{5}$$

The key observation with regard to preserving positivity is due to Godunov [6] who proved that any scheme of better than first order which preserves positivity for the advection equation must be nonlinear. That is the coefficients a_j in (5) above must depend on the numerical solution to the p.d.e. This means that α and β in (4) must also depend on the solution.

In investigating positivity preserving mass matrices and Galerkin finite element methods for transient problems the starting point will be to rewrite the mass matrix as a positivity preserving matrix. This will be applied to the solution of a parabolic equation. The same idea will then be applied to hyperbolic equations and linked to the work of Cardle, [3] and to work on nonlinear finite difference schemes . Finally the extension of the approach to unstructured triangular and tetrahedral meshes will be considered.

2. Modified Mass Matrices and the Initialisation of Parabolic Equations

In trying to solve parabolic equations using a Galerkin method-of-lines approach Skeel and Berzins [9] showed that the initial time derivatives may have the wrong sign. This is because the inverse of the mass matrix A may have have negative entries. Suppose that the parabolic equation is discretised in space to get a system of equations of the form of (3) with $F_i \geq 0$. The initial values of the time derivatives are given by solving the equations (3) for the initial values of the time derivatives $\underline{\dot{U}}(0)$. In computational experiments time derivatives with the wrong sign slow down the the integration and

may give physically misleading solution values. For these reasons Skeel and Berzins devised a scheme that may be viewed as a lumped finite element scheme in which the mass matrix is replaced by the identity matrix.

The issue of when a matrix may have an inverse consisting of positive entries is considered in a large body of work on M matrices. See, for example, [2] who show that if A is a diagonally dominant M matrix with negative off diagonal entries then its inverse A^{-1} has only positive entries. The task is thus to modify the mass matrix so that it has negative off-diagonal entries.

2.1. Derivation of Modifed Mass Matrix

For simplicity consider the case when linear basis functions are used on a uniform spatial mesh as in equations (1). The jth row of the mass matrix is then given by

$$\dot{U}_j \left(\phi_j, \phi_j\right) + \dot{U}_{j+1} \left(\phi_j, \phi_{j+1}\right) + \dot{U}_{j-1} \left(\phi_j, \phi_{j-1}\right) = F_j. \tag{6}$$

Using the identity that on $[x_{j-1}, x_{j+1}]$ $\phi_j + \phi_{j-1} + \phi_{j+1} = 1$ gives:

$$\dot{U}_j \left(\phi_j, 1\right) + \left(\dot{U}_{j+1} - \dot{U}_j\right) \left(\phi_j, \phi_{j+1}\right) + \left(\dot{U}_{j-1} - \dot{U}_j\right) \left(\phi_j, \phi_{j-1}\right) = F_j. \tag{7}$$

Defining the ratio $s_j = \frac{\dot{U}_{j+1} - \dot{U}_j}{\dot{U}_j - \dot{U}_{j-1}}$ allows the jth row of the mass matrix to be rewritten as

$$\dot{U}_j \left(\phi_j, 1\right) + \left(\dot{U}_j - \dot{U}_{j+1}\right) \left[\frac{\left(\phi_j, \phi_{j-1}\right)}{s_j} - \left(\phi_j, \phi_{j+1}\right)\right] = F_j. \tag{8}$$

This matrix is an M matrix if $[\ldots]$ is positive (on a uniform mesh this requires $0 < s_j < 1$) as then the matrix is diagonally dominant with negative off-diagonal entries. In the case when $s_j > 1$ on a uniform mesh the jth row is written as

$$\dot{U}_j \left(\phi_j, 1\right) + \left(\dot{U}_j - \dot{U}_{j-1}\right) \left[\left(\phi_j, \phi_{j+1}\right) s_j - \left(\phi_j, \phi_{j-1}\right)\right] = F_j \tag{9}$$

which again is a row of an M matrix as $[\ldots]$ is positive. In the case when $s_j < 0$ there appears no alternative but to diagonalise (lump) the matrix as $\dot{U}_j \left[\left(\phi_j, \phi_j\right) + \left(\phi_j, \phi_{j-1}\right) + \left(\phi_j, \phi_{j+1}\right)\right]$. The modified matrix may also be written as (with appropriate modifications if $s_j = 0$) :

$$\dot{U}_j + \frac{(s_j + |s_j|)}{12|s_j|} \left[\dot{U}_{j-1} - 2\dot{U}_j + \dot{U}_{j+1}\right] = \frac{F_j}{\delta x}. \tag{10}$$

In solving the equations (8) and (9) for the initial values of the time derivatives it is thus necessary to iteratively solve nonlinear equations. Let $\dot{U}_j^m$ and s_j^m be the values calculated at iteration m. The equations solved in the case when $s_j^m > 1$ are then given by

$$\dot{U}_j^{m+1} + \gamma \left(\dot{U}_j^{m+1} - \dot{U}_{j-1}^m\right) = \frac{F_j}{\delta x}, \text{ where } \gamma = \left[\frac{s_j^m - 1}{6}\right]. \tag{11}$$

In the case when $0 < s_j^m < 1$ the iteration is defined by

$$\dot{U}_j^{m+1} + \gamma \left(\dot{U}_j^{m+1} - \dot{U}_{j+1}^m\right) = \frac{F_j}{\delta x}, \text{ where } \gamma = \left[\frac{1/s_j^m - 1}{6}\right]. \tag{12}$$

In order to illustrate that this procedure produces initial values of the time derivatives with the right sign the following example is used. Skeel and Berzins [9] consider test examples such as the case when the right side of equation (3) is given by $F_j \approx U_t(x_j,0) = \frac{C}{(Cx_j+1.1)}$ where $C = 0.1$ if $x < 0$ and $C = 1.0$ otherwise. For a uniform mesh of 11 and 21 points across the interval $[-1,1]$ Figure 1 shows that the

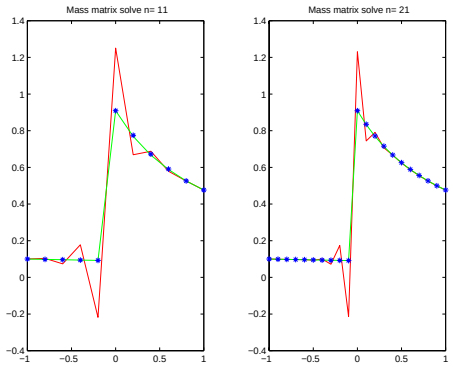

Figure 1. Mass Matrix Calculation for Time Derivatives - is true * is new, – is original

method does result in time derivatives of the correct sign without the overshoots and undershoots of a standard Galerkin approach. The issue of preserving positivity for the diffusion equation has been considered in very recent work by Farago and Horvath [5]. They show that for a d dimensional problem using the θ time integration method, it is necessary to restrict the choice of the parameter θ and the timestep δt by $\frac{d}{6\theta} < \frac{\delta t}{\delta x^2} < \frac{1}{3d(1-\theta)}$, where $d = 1,2$ Their results also extends to three space dimensions.

3. Modified Mass Matrices and Hyperbolic Equations

In general the time derivatives may not have constant sign and it is the non-negativity of the solution that must be considered. In the case of the advection equation it is possible to rewrite equation (4) as a simple explicit method for hyperbolic equations:

$$A\underline{U}(t_{n+1}) = A\underline{U}(t_n) + \delta t\, \underline{F}(\underline{U}(t_n)) \tag{13}$$

In using the modified mass matrix approach as part of a method for hyperbolic equations it is instructive to note other similar approaches. The approximation of U_x by a standard Galerkin method is identical to that of central finite differences. A nonlinear central difference method for hyperbolic equations is given by Swanson and Turkel [10] as :

$$\dot{U}_j = \frac{-1}{2\delta x}(U_{j+1} - U_{j-1}) + \frac{1}{2\delta x}\left[L_{j+1}(U_{j+1} - U_j) - L_j(U_j - U_{j-1})\right] \tag{14}$$

where $L_j = \frac{|1-\hat{r}_j|}{1+|\hat{r}_j|}$ and $\hat{r}_j = \frac{U_{j-1}-U_{j-2}}{U_j - U_{j-1}}$. The right side of this may also be interpreted as a nonlinear Petrov-Galerkin method in which the test function is $\hat{\phi}(x)$ where

$$\hat{\phi}(x) = \phi(x) + \delta x\, L^* \frac{d\phi(x)}{dx} \tag{15}$$

where $L^* = L_j$ if $x_{j-1} < x < x_{j+1}$ and $L^* = L_{j+1}$ otherwise. Although this scheme is positivity preserving it is quite diffuse. A less diffuse scheme is one in which L^* is defined by $L^* = 1 - V(r_j)/r_j$ if $x_{j+1} > x > x_j$ and $L^* = 1 - V(r_{j-1})/r_{j-1}$ if $x_j > x > x_{j-1}$ where $V(r_j) = \frac{r_j + |r_j|}{1+|r_j|}$ and $r_j = \frac{U_{j+1} - U_j}{U_j - U_{j-1}}$. Routine manipulation shows that this definition gives the well-known van Leer scheme, e.g [1]:

$$\dot{U}_j(t) = \frac{-1}{\delta x}\beta_j(U_j(t) - U_{j-1}(t)), \quad \text{where } \beta_j = \left[1 + \frac{V(r_j)}{2} - \frac{V(r_{j-1})}{2r_{j-1}}\right]. \tag{16}$$

An alternative view of this scheme is thus as a nonlinear Petrov-Galerkin method. Requiring positivity for forward Euler timestepping requires the CFL type restriction $0 \le \beta_i < \frac{\delta x}{\delta t}$. The nonlinear extension of the type of Petrov-Galerkin method given by (3) is thus given by equation (10) with F_j defined by the right side of (16). An outline proof that, when combined with forward Euler timestepping, this is positivity preserving follows from a modified version of (11). (The proof for the case in (12) being similar). The iteration from (11) may be written as:

$$\dot{U}_j^{m+1}(t) = \frac{1}{1+\gamma}\left[\gamma \dot{U}_{j-1}^m(t) + \frac{F_j}{\delta x}\right], \quad m = 0, 1, \ldots$$

where γ is defined as in equation (11) and so depends on s_j^m and where $\frac{F_j}{\delta x}$ is defined by the righthand side of equation (16). Hence this equation may also be written as

$$\dot{U}_j^{m+1}(t) = \frac{1}{1+\gamma}\left[\gamma \dot{U}_{j-1}^m(t) - \frac{\beta_j}{\delta x}(U_j(t) - U_{j-1}(t))\right], \quad m = 0, 1, \ldots \tag{17}$$

The predicted values, $\dot{U}_{j-1}^0(t_{n+1})$, are given by equation(16). An outline of the approach used to define solution positivity in terms of equation (5) may now be given. The initial guesses for the time drivatives are given by equation (16):

$$\dot{U}_{j-1}^0(t_{n+1}) = -\frac{1}{\delta x}\left[U_{j-1}(t_n) - U_{j-2}(t_n)\right]\beta_{j-1} \tag{18}$$

where β_{j-1} is also defined as in (16). Substituting for $\dot{U}_{j-1}^0$ in (17) and applying forward Euler timestepping gives:

$$U_j^1(t_{n+1}) = U_j(t_n) - \frac{\delta t}{\delta x(1+\gamma)}\left[\beta_j(U_j(t_n) - U_{j-1}(t_n)) + \gamma\beta_{j-1}(U_{j-1}(t_n) - U_{j-2}(t_n))\right] \tag{19}$$

where $U_j^1(t_{n+1})$ is the first iteration estimate for $U_j(t_{n+1})$. This may be rewritten as

$$U_j^1(t_{n+1}) = (1 - \beta_j E)U_j(t_n) + (\beta_j - \gamma\beta_{j-1})E\, U_{j-1}(t_n) + (\gamma\beta_{j-1})E\, U_{j-2}(t_n) \tag{20}$$

$$\text{where } E = \frac{\delta t}{\delta x}\frac{1}{(1+\gamma)}.$$

In proving positivity of this consider the worst case in equation (19) and suppose that $\beta_j(U_j(t_n) - U_{j-1}(t_n))$ has a different sign to $\beta_{j-1}(U_{j-1}(t_n) - U_{j-2}(t_n))$. It then follows that r_{j-1} is negative and hence that $\beta_j = \left[1 + \frac{V(r_j)}{2}\right]$ and $\beta_{j-1} = \left[1 - \frac{V(r_{j-2})}{2r_{j-2}}\right]$. Hence

$$\beta_j - \gamma\beta_{j-1} = 1 + \frac{V(r_j)}{2} - \gamma + \gamma\,\frac{V(r_{j-2})}{2r_{j-2}}$$

and in order to guarantee a positive solution at the end of the first iteration we need to impose the condition $\gamma < 1.0$. Thus, from equations (11) and (12), the iterative method is only applied if

$$\frac{1}{7} \; < \; s_j^m \; < \; 7$$

The same approach can then be used inductively to prove positivity for second and subsequent iterations. Figure 2 shows the numerical results obtained when the method is applied to the advection of a square pulse function from $x = 0.2$ to 0.7 using a spatial mesh of 51 equally spaced points. The figure compares the solution obtained with the van Leer method with the new approach and shows the effect of using a mass matrix is to give a still positive but more skewed profile than that obtained from the van Leer method.

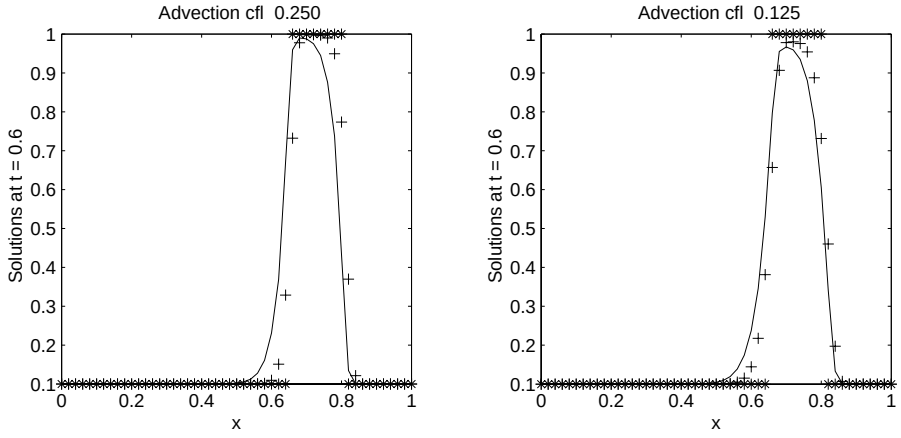

Figure 2. Numerical solution of advected square wave

4. Extension to Triangular and Tetrahedral Meshes

In the case of triangular mesh examples it is possible to use the same idea as in one space dimension. In this case, for example, the mass matrix for a mesh fragment consisting of three triangles with node i in common and a perimeter consisting of nodes j, k and l is given by:

$$\dot{U}_i \;\; (\phi_i, 1) + \left(\dot{U}_j - \dot{U}_i\right) \;\; \left(\phi_j, \phi_i\right) + \left(\dot{U}_k - \dot{U}_i\right) \;\; \left(\phi_k, \phi_i\right) + \left(\dot{U}_l - \dot{U}_i\right) \;\; \left(\phi_l, \phi_i\right) \tag{21}$$

where $\left(\phi_i, \phi_j\right) = \frac{1}{12} \;\; \left[A_{ijk} + A_{ijl}\right]$, $\left(\phi_j, \phi_j\right) = \frac{1}{12} \left[A_{ijk} + A_{ijl} + A_{ilk}\right]$ and where A_{ijk} is the area of triangle with nodes i, j and k. As the contribution of element ijk to the mass matrix is

$$\frac{A_{ijk}}{12} \left[4\dot{U}_i + \left(\dot{U}_j - 2\dot{U}_i - \dot{U}_k\right)\right]. \tag{22}$$

The same ideas as in one space dimension may be used to rewrite $\dot{U}_j - 2\,\dot{U}_i + \dot{U}_k$ by using the same approach as in Section 2.1. Consider the i, j, k triangle and let $\dot{U}_{(j+k)/2}$ be midpoint solution value on jk edge. The decomposition given by

$$\dot{U}_j - 2\,\dot{U}_i + \dot{U}_k = \left[\dot{U}_j - 2\,\dot{U}_{(j+k)/2)} + \dot{U}_k\right] + 2\left[\dot{U}_{(j+k)/2)} - \dot{U}_i\right] \tag{23}$$

allows the terms on the right side of this equation to be viewed as second order approximations to second and first space derivatives. Hence in discarding these terms we introduce second order errors as in one space dimension.

The same idea extends to tetrahedral mesh examples. Consider a single tetrahedron with its four nodes labelled as i, j, k, l and associated linear basis functions ϕ_i, ϕ_j, ϕ_k and ϕ_l. The mass matrix integral associated with this tetrahedron is

$$V_{ijkl}\left[\dot{U}_i\ (\phi_i, \phi_i) + \dot{U}_j\ \left(\phi_j, \phi_i\right) + \dot{U}_k\ (\phi_k, \phi_i) + \dot{U}_l\ (\phi_l, \phi_i)\right]. \tag{24}$$

Evaluating the volume integral gives:

$$\frac{V_{ijkl}}{20}\left[5\dot{U}_i + \left(\dot{U}_j - \dot{U}_i\right) + \left(\dot{U}_k - \dot{U}_i\right) + \left(\dot{U}_l - \dot{U}_i\right)\right] \tag{25}$$

which may be rewritten as three terms of the form $\frac{V_{ijkl}}{20}\left[\frac{5}{3}\dot{U}_i + \frac{1}{2}\left(\dot{U}_j - 2\dot{U}_i - \dot{U}_k\right)\right]$ and same ideas applied as in one and two space dimensions.

5. Summary

In this paper a novel approach to preserving positivity has been taken. The approach relies on using a nonlinear form of the mass matrix in conjunction with nonlinear Petrov-Galerkin type terms. The approach has applications in one,two and three space dimensional cases, but further work is clearly needed to assess its usefulness.

REFERENCES

1. Berzins M, Ware JM. Positive Cell Centered Finite Volume Discretisation Methods for Hyperbolic Equations on Irregular Meshes, *Applied Numerical Mathematics* 1995; **16**:417–438.
2. Berman A, Plemmons RJ. *Nonnegative Matrices in the Mathematical Sciences.* Acadedmic Press, 1979.
3. Cardle JA. A modification of the Petrov-Galerkin method for the transient convection-diffusion equation. *International Journal for Numerical Methods in Engineering* 1995; **38**:171–181.
4. Cockburn B, Karniadakis GE, Shu C-W. (eds). *Discontinuous Galerkin Methods, Theory Computations and Applications. Lecture Notes in Computataional Science and Engineering 11* Springer Berlin Heidelberg, 2000; 3–53.
5. Farago I, Horvath R. On the nonnegativity conservation of finite element solution of parabolic problems. In *3D Finite Element Conference Proceedings* Krizek M. , et al (eds). Gakkotosho Co., Tokyo 2001.
6. Godunov SK. Finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Math Sbornik* 1959; **47**:271–306.
7. Johnson C. *Numerical Solution of Partial Differential Equations by the Finite Element Method* Cambridge University Press, 1987.
8. Segal A. Finite element methods for advection-diffusion equations. In *Numerical Methods for Advection Diffusion Problems. Notes on Numerical Fluid Mechanics, Volume 45* Vreugdenhil CB, Koren B (eds). Vieweg: Braunscweig/Wiesbaden, 1993; 195–214.
9. Skeel RD, Berzins M. A Method for the Spatial Discretisation of Parabolic Equations, *SIAM Journal on Scientific Computing* 1990; **11**(1):1–32.
10. Swanson RC, Turkel E. On central-difference and upwind schemes. *Journal of Computational Physics* 1992; **101**:292–306.