European Conference on Computational Fluid Dynamics
ECCOMAS CFD 2006
P. Wesseling, E. Oñate and J. Périaux (Eds)

CONSTRAINED LEAST-SQUARES METHODS FOR PARTIAL DIFFERENTIAL EQUATIONS

Bart De Maerschalck* **and Marc I. Gerritsma**†

*Von Karman Institute for Fluid Dynamics
Waterloosesteenweg 72, 1640 Sint-Genesius-Rode, Belgium
e-mail: demaersc@vki.ac.be
web page: http://www.vki.ac.be/∼ demaersc
†Delft University of Technology, Faculty L&R,
Kluyverweg 1, 2629 HS Delft, The Netherlands
e-mail: M.I.Gerritsma@LR.TUDelft.nl

Key words: Least-squares, spectral elements, space-time, entropy condition

Abstract. *Least-squares methods for partial differential equations are based on a norm-equivalence between the error norm and the residual norm. The resulting algebraic system of equations, which is symmetric positive definite, can also be obtained by solving a weighted collocation scheme using least-squares to solve the resulting algebraic equations. Furthermore, least-squares allows to ad extra constraints to the system. In the present work the entropy is added as an extra inequality constraint to ensure only physical solutions for the one-dimensional inviscid Burgers equation are obtained.*

1 INTRODUCTION

In the least-squares method one tries to find an approximate solution to a system of equations – algebraic equations or partial differential equations –, by minimizing the residual in a certain norm. Once a norm-equivalence between the residual and error can be established, one can show that minimizing the residual is equivalent to minimize the error.

Least-squares methods are well embedded in linear algebra and statistics where they are mainly used to solve overdetermined algebraic systems. The same mathematical base can be used to approximate the solutions of a system of partial differential equations. In order to be able to apply the least-squares approach to systems of partial differential equations on an arbitrary mesh, the least-squares approach is applied in a finite element sense. This method was first proposed by Jiang[1]. Simultaneously, Gerritsma and Proot[2,3] and Pontaza[4] extended this approach to the least-squares spectral element method. Higher order polynomials are used to approach the exact solution within each element. This approach allows for high order convergence rates, when the underlaying exact solution is sufficiently smooth. In this case polynomial enrichment, so called p-type refinement, leads to exponential convergence speed. In all other cases the convergence is algebraic and the convergence rate depends on the polynomial degree of the approximating solution and the smoothness of the underlying exact solution.

In the present paper the least-squares spectral element formulation is applied to the one-dimensional Burgers equation in a space-time formulation. In recent work it has been demonstrated that it is possible to render a fully converged solution without the need to add any artificial diffusion[5,6]. The least-squares approach allows to add more constraints which the solution is desired to satisfy. In the present paper a weak entropy condition will be added as an extra constraint in order to ensure that only physical solutions are obtained.

2 LEAST SQUARES PRINCIPLES

When using least squares methods one tries to find an approximate solution in space by minimizing the residual in a certain norm. Consider the following abstract system of equations:

$$A\,x = b\;, \tag{1}$$

with $x \in X$, $b \in Y$ and A a matrix or operator projecting the solution space X to the *residual space* Y, $A : X \to Y$. This may be a system of linear algebraic equations, partial differential equations or even a combination of both. In general the solution space X is infinitely large. In numerical methods one has to restrict the search to a finite dimensional subspace of X, $X^h \in X$. In general the exact solution is not a member of this subspace X^h. Therefore one tries to find a solution in X^h so that the difference with the exact solution measured in a certain norm is as small as possible. However, when the exact solution is unknown it is not possible to measure the error. What can be measured is the error in the residual space:

$$\left\|A\,x^h - A\,x_{exact}\right\|_Y = \left\|A\,x^h - b\right\|_Y\;. \tag{2}$$

Therefore, if a norm equivalence between the error in the solutions space and the error in the residual space can be established, one can find an approximate solution by minimizing the residual in the residual space:

$$x^h = \arg\min_{x^h \in X^h} \left\|A\,x^h - b\right\|_Y\;. \tag{3}$$

In linear algebra the residual can be minimized in the L^2-norm. Minimizing (3) is then equivalent to setting up the normal equations:

$$\min_{x^h \in X^h} \left\|A\,x^h - b\right\|_{L^2} \quad \Leftrightarrow \quad A^{\mathrm{T}} A\,x^h = A^{\mathrm{T}} b \tag{4}$$

If A has full rank, these normal equations are symmetric positive definite.

The above described approach can also be applied to systems of partial differential equations. Consider therefore the following system with boundary conditions:

$$\mathcal{L}u = f \qquad \text{in } \Omega\;, \tag{5}$$

$$\mathcal{R}u = g \qquad \text{on } \Gamma\;, \tag{6}$$

with $u \in X$ and $(\mathcal{L}, \mathcal{R}) : X \to Y(\Omega) \times Y(\Gamma)$. If one can derive the an a priori error estimate of the following kind:

$$C_1 \left\|u - u_{exact}\right\|_X \leq \left\|\mathcal{L}u - f\right\|_{Y(\Omega)} + \left\|\mathcal{R}u - g\right\|_{Y(\Gamma)} \leq C_2 \left\|u - u_{exact}\right\|_X\;, \tag{7}$$

the norm equivalence between the error norm and the residual norm is established. Therefore an approximating solution for the system of partial differential equations can be found by minimizing the residual. Minimizing the residual means minimizing the following functional,

$$\mathcal{J}(u) = \frac{1}{2} \|\mathcal{L}u - f\|_{Y(\Omega)} + \|\mathcal{R}u - g\|_{Y(\Gamma)} \ , \tag{8}$$

by means of variational analysis we get the associated Euler-Lagrange equations

$$\lim_{\epsilon \to 0} \frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathcal{J}(u + \epsilon v) = 0 \, , \qquad \forall \, v \in X \ , \tag{9}$$

which yields the following equation:

$$\mathcal{B}(u, v) = \mathcal{F}(v) \ , \qquad \forall \, v \in X \ , \tag{10}$$

with:

$$\mathcal{B}(u, v) \ = \ (\mathcal{L}u \, , \, \mathcal{L}v)_{Y(\Omega)} + (\mathcal{R}u \, , \, \mathcal{R}v)_{Y(\Gamma)} \tag{11}$$

$$\mathcal{F}(v) \ = \ (f \, , \, \mathcal{L}v)_{Y(\Omega)} + (g \, , \, \mathcal{R}v)_{Y(\Gamma)} \tag{12}$$

In the present work the L^2-norm is taken for the residual norm. The test functions are chosen in such a way that they satisfy the boundary conditions. If the boundary conditions are enforced strongly, the boundary terms in (7), (8), (12) and (12) can be omitted. Therefore one gets:

$$(\mathcal{L}u \, , \, \mathcal{L}v)_{L^2(\Omega)} = (f \, , \, \mathcal{L}v)_{L^2(\Omega)} \ , \qquad \forall \, v \in X \tag{13}$$

with for the inner-products:

$$(\mathcal{L}u \, , \, \mathcal{L}v)_{L^2(\Omega)} = \int_\Omega \mathcal{L}u \, \mathcal{L}v \, \mathrm{d}\Omega \tag{14}$$

$$(f \, , \, \mathcal{L}v)_{L^2(\Omega)} = \int_\Omega f \, \mathcal{L}v \, \mathrm{d}\Omega \tag{15}$$

As for the normal equations the system in (13) is symmetric and positive definite.

3 BASIS FUNCTIONS AND SPECTRAL ELEMENT MATRICES

If the solution space is infinitely large one has to define a discrete solution space $X^h \in X$. In standard finite and spectral element methods this is done by dividing the domain in a finite number of non-overlapping sub-domains. On each sub-domain or element a set of continuous basis functions is defined. In the present work nodal basis functions are used. In one dimension these basis functions are Langange polynomials based on the zeros of the first derivative of the Legendre polynomial of degree P extended with the boundary nodes of the element. These nodes are well known as the Gauss-Lobatto-Legendre roots. The $P + 1$ polynomials on a standard element $[-1, 1]$ are then defined as:

$$h_i(\xi) = \frac{(\xi^2 - 1) \ \frac{\mathrm{d}\, L_P(\xi)}{\mathrm{d}\xi}}{P\,(P + 1)\, L_P(\xi) \ (\xi - \xi_i)} \tag{16}$$

In multiple dimensions the basis functions are obtained by means of tensor products. Arbitrarily shaped elements are mapped to the standard element by a parametric mapping. Within each element the solution is approximated by a finite sum of these basis functions:

$$u_e^P(\xi) = \sum_{i=1}^{P+1} \boldsymbol{u}_i^h \, h_i(\xi) \; . \tag{17}$$

With $\boldsymbol{u}_i^h$ the unknown coefficients. For each element this approximation can be substituted into (13) and for the testfunction v the same basis function can be used:

$$\left(\mathcal{L}u_e^P(\xi) \, , \, \mathcal{L}h_i(\xi)\right)_{L^2(\Omega)} = \left(f \, , \, \mathcal{L}h_i(\xi)\right)_{L^2(\Omega)} \; , \qquad \forall \, 1 \leq i \leq P+1 \; . \tag{18}$$

Therefore, for each element:

$$K \, \boldsymbol{u} = \boldsymbol{F} \; , \tag{19}$$

with for the components of K and $\boldsymbol{f}$:

$$K_{ij} \;=\; \left(\mathcal{L}h_j \, , \, \mathcal{L}h_i\right)_{L^2(\Omega)} \tag{20}$$

$$\boldsymbol{F}_i \;=\; \left(f \, , \, \mathcal{L}h_i\right)_{L^2(\Omega)} \tag{21}$$

The integrals in (20) and (21) are evaluated numerically by means of Gauss-Lobatto integration rule. Therefore, the element matrices and force vectors can also be written as the following matrix multiplications:

$$K \;=\; L^{\mathrm{T}} W L \; , \tag{22}$$

$$\boldsymbol{F} \;=\; L^{\mathrm{T}} W \boldsymbol{f} \; , \tag{23}$$

with W a diagonal matrix with in the diagonal the Gauss-Lobatto-Legendre weights and for L and $\boldsymbol{f}$:

$$L_{ij} \;=\; \mathcal{L}h_j\big|_{\xi=\xi_i} \; , \tag{24}$$

$$\boldsymbol{f}_i \;=\; f(\xi_i) \; . \tag{25}$$

L is an $m \times n$ matrix with m the number of integration points used for the Guass-integration and n the number of unknown coefficients for the element, equals $P+1$ in (17). In most spectral element methods m is chosen equal to n. However, this is not necessary and in a number of cases it is even advised to choose $m > n$ as it is demonstrated by De Maerschalck and Gerritsma[6].

4 CONSTRAINED LEAST SQUARES METHODS FOR PDE'S

For each element the local system of equations in (19) can be written as:

$$L^{\mathrm{T}} W L \, \boldsymbol{u} = L^{\mathrm{T}} W \boldsymbol{f} \qquad \Leftrightarrow \qquad \left(\sqrt{W} L\right)^{\mathrm{T}} \left(\sqrt{W} L\right) \boldsymbol{u} = \left(\sqrt{W} L\right)^{\mathrm{T}} \left(\sqrt{W} \boldsymbol{f}\right) \; . \tag{26}$$

Note that this system is exactly what one gets when setting up the normal equations for

$$\left(\sqrt{W}L\right)\,\boldsymbol{u} = \left(\sqrt{W}\boldsymbol{f}\right)\ . \tag{27}$$

When m is chosen larger then n, this system is overdetermined.

As in standard finite element methods, once the element matrices and vectors are calculated they can be gathered into the stiffness matrix A^h and force vector $\boldsymbol{f}^h$. In standard least-squares methods for PDE's the element matrices are calculated as in (20) and (21) and gathered into the symmetric positive definite stiffness matrix A^h which can then be solved with an efficient solver like the preconditioned conjugate gradient method. However, one can also calculate the element matrices as in (27), gather the element matrices and vectors into the stiffness matrix and right hand side vector and then solve the overdetermined system using a least-squares approach for algebraic systems. This has been demonstrated before by Hoitinga and de Groot[7]. Actually, this is in fact using a least-squares approach to solve a weighted collocation scheme. The collocation points are the m Gauss-Lobatto roots for the numerical integration and the collocation weights are exactly the squares rootes of the Gauss-Lobatto weights. Heinrichs[8,9] does something similar by applying a collocation scheme without the weights. After applying a weighted collocation method and gathering the element matrices, the remaining least-squares problem is of the form:

$$\boldsymbol{u}^h = \arg\min_{\boldsymbol{u}\in X^h} \left\| A^h\,\boldsymbol{u^h} - \boldsymbol{f}^h \right\|_{L^2}\ . \tag{28}$$

This latter formulation allows one to add extra constraints to the solution of the minimization problem. One can try to minimize (28) with an inequality constraint for $\boldsymbol{u}^h$:

$$\boldsymbol{u}^h = \arg\min_{\boldsymbol{u}\in X^h} \left\| A^h\,\boldsymbol{u^h} - \boldsymbol{f}^h \right\|_{L^2}, \qquad \text{with } C\boldsymbol{u}^h \leq d\ . \tag{29}$$

In the next sections an example will be given where the constrained is a weak entropy condition.

5 INVISCID BURGERS EQUATION

In this section the one-dimensional inviscid Burgers equation in conservative form will be considered:

$$\left\{ \begin{array}{l} \frac{\partial q}{\partial t} + \frac{\partial f}{\partial x} = 0\ , \\ -q^2 + f = 0\ . \end{array} \right. \tag{30}$$

With a prescribed initial condition at $t = 0$, and Dirichlet boundary conditions at $x = 0$ and $x = L$.

Before applying the least-squares principles the nonlinear algebraic equation in (30) is linearised using Newtons method:

$$\left\{ \begin{array}{l} \frac{\partial q}{\partial t} + \frac{\partial f}{\partial x} = 0\ , \\ -2\,q_0 q + f = -q_0^2\ , \end{array} \right. \tag{31}$$

with q_0 the solution of the previous iteration step.

For this problem a space-time formulation has been used. Therefore, the time is considered as an extra spatial variable and the dimension of the problem has been augmented by one. In this case this requires a two-dimensional grid of quadrilateral space-time elements. The scheme is semi-implicit in that sense that for each time level the solution is solved on a single space-time strip of one element *high*. The solution on the lower boundary of each strip is then prescribed by the solution on the upper boundary of the previous time level or by the initial condition for the first time level. For each space-time element, the element matrix and vector can be calculated as in (22) and (23) with for L and $\mathbf{f}$:

$$L = \begin{pmatrix} D_t & D_x \\ -J_0 & B \end{pmatrix} \tag{32}$$

$$\boldsymbol{f} = \begin{pmatrix} 0 \\ \boldsymbol{f}_q \end{pmatrix} \tag{33}$$

with:

$$[D_x]_{ij} = \left.\frac{\partial h_j}{\partial x}\right|_{\boldsymbol{\xi}_i} , \tag{34}$$

$$[D_t]_{ij} = \left.\frac{\partial h_j}{\partial t}\right|_{\boldsymbol{\xi}_i} , \tag{35}$$

$$[J_0]_{ij} = 2\, q_0(\boldsymbol{\xi}_i)\, h_j(\boldsymbol{\xi}_i) , \tag{36}$$

$$B_{ij} = h_j(\boldsymbol{\xi}_i) , \tag{37}$$

$$\left[\boldsymbol{f}_q\right]_i = -\left(q_0(\boldsymbol{\xi_i})\right)^2 \tag{38}$$

Figure 1 shows the solution and the flux when ten cells are used with with polynomial degree $N = 5$ in space and time. The initial condition is a single cosine hill. The solution is plotted at $t = 2$ and twelve time-steps are used. Note that no artificial diffusion terms need to be added to the formulation. The wiggles that occur are due to the Gibbs effect and indicate regions of limited smoothness of the solution. However, they do not have the intention to pollute the whole domain and are restricted to the direct neighborhood of the discontinuity. The apace-time least-squares formulation is inherently stable. When one would increase the number of elements, the polluted area around the singularity would be more compressed. Therefore, it is possible to correct for this numerical oscillations in a postprocessing reconstruction step[5,10,11].

6 WEAK ENTROPY CONSTRAINT

Consider the following *etropy function* $Q(q)$ and the *entropy flux function* $F(q)$:

$$Q(q) = q^2 , \tag{39}$$

$$F(q) = \frac{4}{3}q^3 . \tag{40}$$

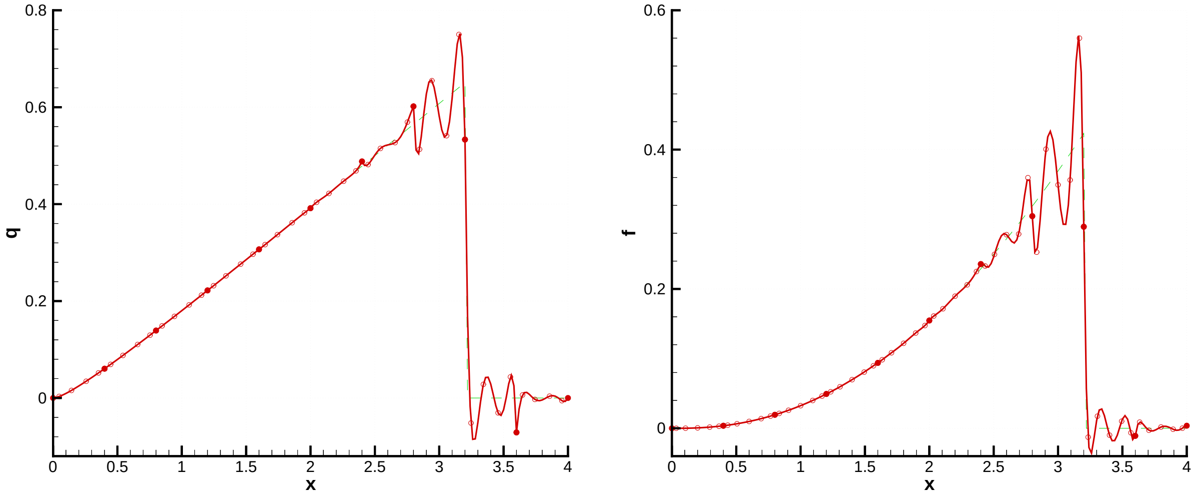

Figure 1: Solution (left) and flux (right) of the 1D Burgers equation, $N_{cells} = 10$, Order 5 in space and time.

If $q(x,t)$ is a solution of (30) then one can show that[12]:

$$\frac{\partial Q(q)}{\partial t} + \frac{\partial F(q)}{\partial x} = 0 \; , \tag{41}$$

in the smooth regions of the domain, and:

$$\frac{\partial Q(q)}{\partial t} + \frac{\partial F(q)}{\partial x} < 0 \; , \tag{42}$$

across the shock. Or in general: all weak solutions of (30) satisfy:

$$\frac{\partial Q(q)}{\partial t} + \frac{\partial F(q)}{\partial x} \leq 0 \; . \tag{43}$$

Consider for a moment one time level $[T, T + \Delta T]$ and integrate (43) over the space-time domain $[0, L] \times [T, T + \delta t]$, $\delta t \leq \Delta T$, as indicated in Figure 2:

$$\int_T^{T+\delta t} \int_0^L \left(\frac{\partial Q(q)}{\partial t} + \frac{\partial F(q)}{\partial x} \right) \mathrm{d}x \, \mathrm{d}t \leq 0 \; . \tag{44}$$

Note that on the lower, the left and right boundary of the space-time strip Dirichlet boundary conditions have been prescribed. When one integrates equation (44) by parts and brings the boundary values to the other side, then:

$$\int_0^L (q(x, T + \delta t))^2 \, \mathrm{d}x - \int_0^L (q(x,T))^2 \, \mathrm{d}x + c\,\delta t \leq 0 \qquad \forall \; \delta t \leq \Delta T \; , \tag{45}$$

with $c = \frac{4}{3}\left((q(L))^3 - (q(0))^3\right)$ constant per time level. Figure 3 shows the entropy according

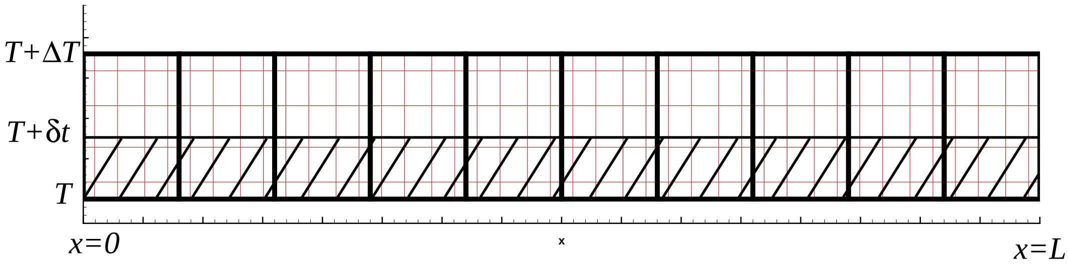

Figure 2: Space-time mesh with area of integration indicated.

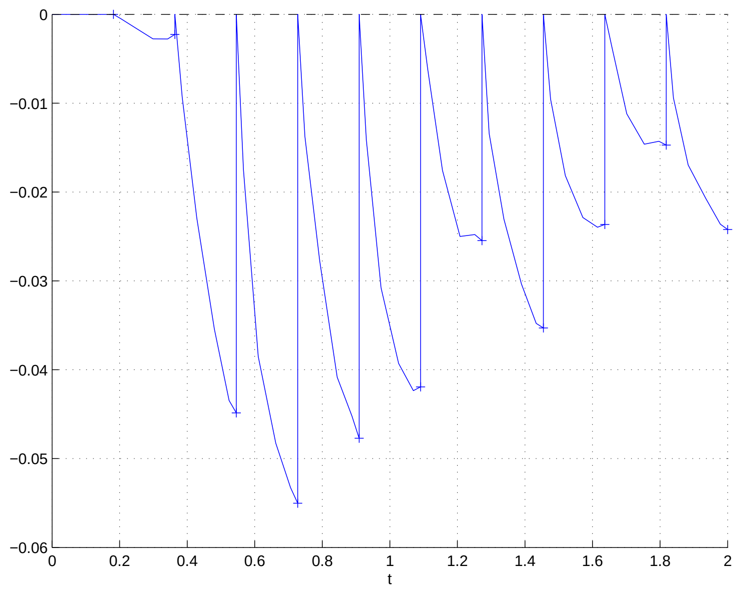

Figure 3: Entropy inequality for the test case presented in Figure 1

to the right hand side of equation (45) for the test case presented in Section 5. Note that at each moment in time, the solution satisfies the inequality of (45). Consider next the same test case, but with a discontinuous formulation. In this formulation $q(x,t)$ is continuous in time, but discontinuous in space across the element interfaces, while the flux is continuous in space and discontinuous in time[13]. The results for the solution and for the flux are plotted in Figure 4. It is obvious that this is not a physical solution of (30). The entropy is plotted in Figure 5. Note that in the last three time steps the weak entropy condition is no longer satisfied.

In Section 4 is was stated that it is possible to add extra constraints to the minimization problem. One can rewrite (45) by:

$$\int_0^L f(x,T+\delta t))\,\mathrm{d}x \leq \int_0^L \left(q(x,T)\right)^2\,\mathrm{d}x - c\,\delta t \qquad \forall\ \delta t \leq \Delta T\ . \tag{46}$$

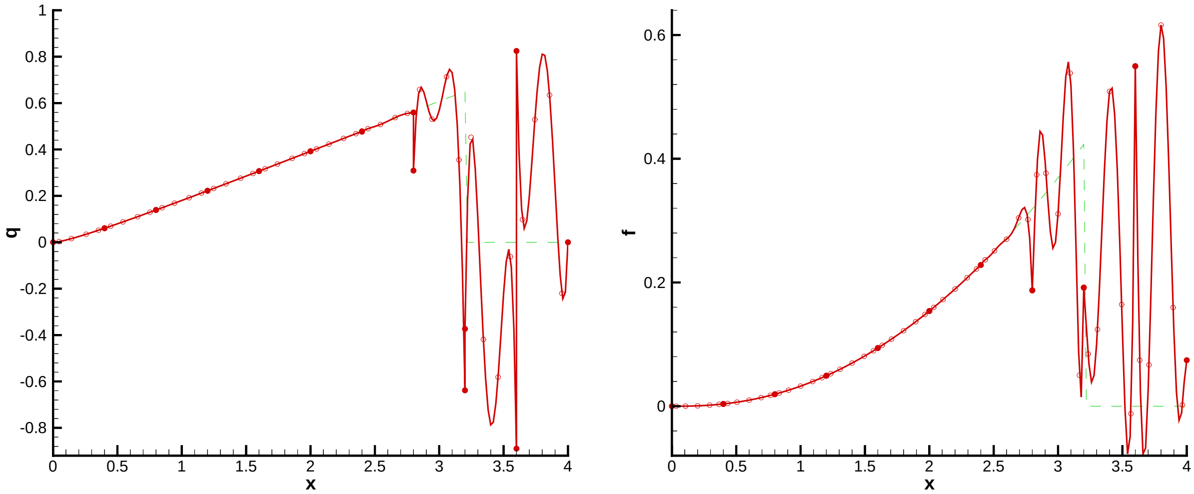

Figure 4: Solution (left) and flux (right) of the 1D Burgers equation, $N_{cells} = 10$, Order 5 in space and time, discontinuous LS-SEM.

This inequality is linear in f. One can vary δt so that $T + \delta t$ coincides with the Gauss-Lobatto nodes in time:

$$\delta t = (\eta_i + 1)\,\frac{\Delta T}{2}\,, \qquad \text{with} \;\; -1 \leq \eta_i \leq 1\,. \tag{47}$$

If for this time levels the Gauss-integration rule is used to evaluate the integrals in (46) it is possible to write for each of these time levels the inequality in the following matrix form:

$$C\,\boldsymbol{u}^h \leq \boldsymbol{c}\,, \tag{48}$$

with $\boldsymbol{u}^h$ the vector with the unknown coefficients. One can solve now the constrained least-squares problem of equation (29).

Figure 6 and 7 show the results for the solution and the flux and the weak entropy equation for the test case as described above. Adding these inequality constraints to the least-squares problem does improve the solution.

7 CONSLUSIONS

A least-squares spectral element method has been applied to the one-dimensional inviscid Burgers equation. The Burgers equation is used in a conservative form and Newton linearization has been applied. It has been shown that no special treatment is required to render a stable solution. When a discontinuous least-squares spectral element method is used, this may result in a non-physical solution. The weak entropy condition is no longer satisfied at all time levels. To overcome this this inconsistency the entropy condition is added as an extra inequality constraint to the least-squares approach. Adding the extra constraints does improve the quality of the solution.

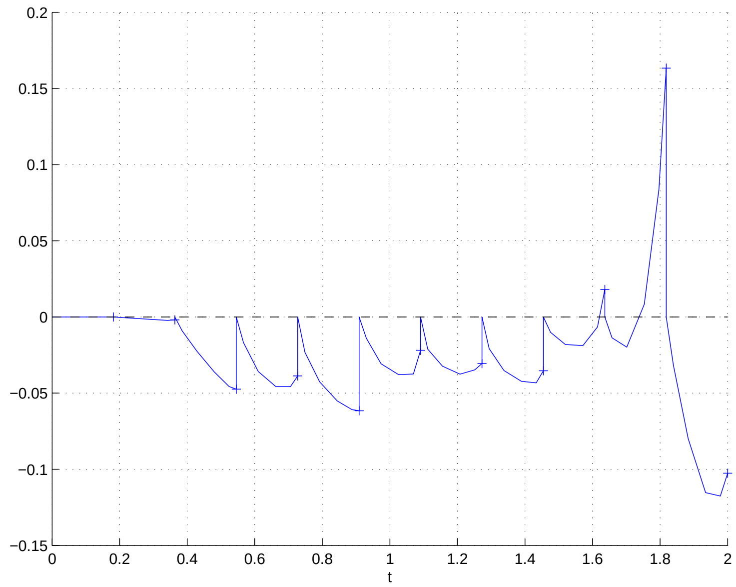

Figure 5: Entropy inequality for the test case presented in Figure 4

REFERENCES

[1] B.-N. Jiang. *The Least-Squares Finite Element Method*. Springer, Berlin Heidelberg, Germany, (1998).

[2] M.M.J. Proot and M.I. Gerritsma. Least-squares spectral elements applied to the Stokes problem. *J. Comp. Phys.*, **181**,454–477, (2002).

[3] M.M.J. Proot and M.I. Gerritsma. A least-squares spectral element formulation for the Stokes problem. *J. Sci. Comp.*, 17(1-4), 297–306, (2002).

[4] J.P. Pontaza. *Least-squares variational principles and the finite element method: theory, formulations, and models for solid and fluid mechanics*. PhD thesis, Texas A&M University, Massachusett, Texas, USA, December (2003).

[5] B. De Maerschalck and M. I. Gerritsma. The use of the Chebyshev Approximation in the Space-Time Least-Squares Spectral Element Method. *Num. Alg.*, **38**, 173–196, (2005).

[6] B. De Maerschalck and M. I. Gerritsma. Higher-Order Gauss-Lobatto Integration for Nonlinear Hyperbolic Equations. *J. Sci. Comput., to appear*, (2006).

[7] W. Hoitinga, R. de Groot, and M. I. Gerritsma. Direct Minimization of the Least Squares Spectral Element Functional - Part I: Direct Solver. *IJCES, submitted to*, (2005).

[8] W. Heinrichs. Least-squares collocation for discontinuous and singular perturbation problems. *J. Comput. Appl. Math.*, **157**, 329–345, (2003).

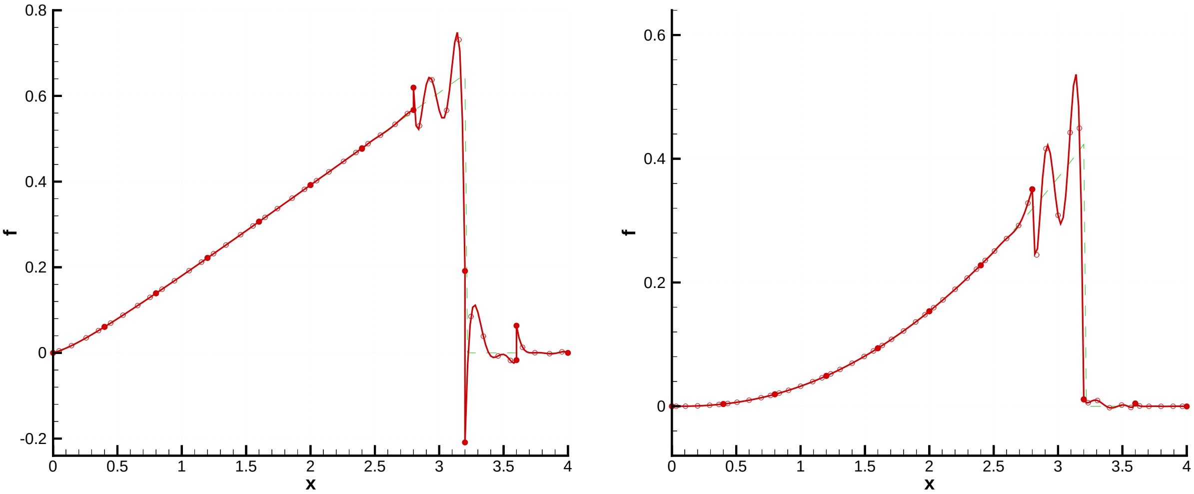

Figure 6: Solution (left) and flux (right) of the 1D Burgers equation, $N_{cells} = 10$, Order 5 in space and time, discontinuous LS-SEM with constrained least-squares.

[9] W. Heinrichs. Least-squares spectral collocation for for the Navier-Stokes equations. *J. Scient. Comput.*, **21**(1), 81–90, (2004).

[10] B. De Maerschalck. *Space-Time Least-Squares Spectral Element Method for Unsteady Flows - Application and Evaluation for linear and non-linear hyperbolic scalar equations.* http://www.hsa.lr.tudelft.nl/~bart, (2003).

[11] B. De Maerschalck, M. I. Gerritsma, and M.M.J. Proot. Space-Time Least-Squares Spectral Elements for Convection Dominated Unsteady Flows. *AIAA journal*, **44**, 558–565, (2005).

[12] C.B. Laney. *Computational Gasdynamics.* Cambridge University Press, Cambridge, (1998).

[13] B. De Maerschalck and M. I. Gerritsma. Least-Squares Spectral Element Method for Nonlinear Hyperbolic Differential Equations. *J. Comput. Appl. Math., to appear*, (2005).

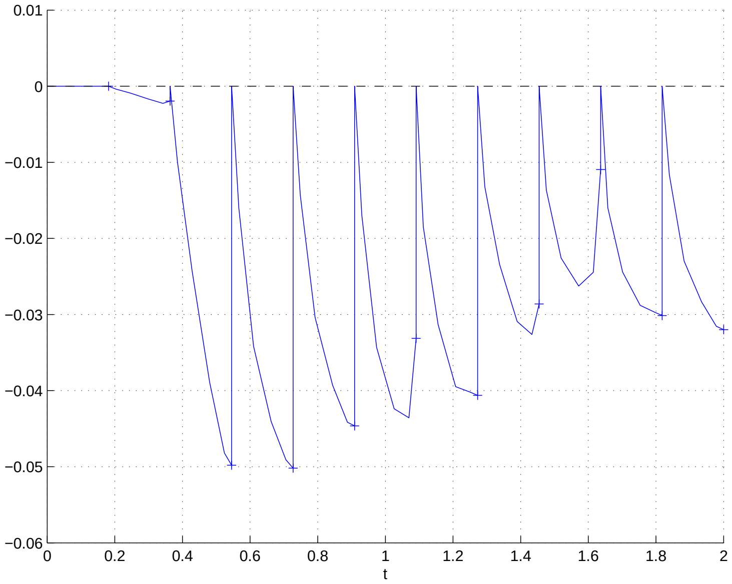

Figure 7: Entropy inequality for the test case presented in Figure 6