THE UNIVERSITY OF READING

A Note on Duality in a Scalar Hyperbolic Equation

by

M.J. Baines

Numerical Analysis Report 8/97

DEPARTMENT OF MATHEMATICS

A Note on Duality in a Scalar Hyperbolic Equation

M.J.Baines

Abstract

We describe a solution/grid duality in the approximate solution of a scalar hyperbolic equation on triangles. Solution mechanisms for a Least Squares minimisation approach are discussed.

1. Form of the Fluctuation

Suppose that a function $u(x,y)$ is aproximated by a linear function $U(x,y)$ in each triangle T of an unstructured grid $\{T\}$. Then, for the PDE

$$\mathbf{a}.\nabla u = 0 \tag{1.1}$$

let the *fluctuation* in the triangle T be defined [1] as

$$\phi_T = -\int_T (\mathbf{a}.\nabla U)\, d\Omega. \tag{1.2}$$

Since∇U is constant in T we have

$$\phi_T = -S_T\,(\mathbf{a}_T.\nabla U) \tag{1.3}$$

where S_T is the area of the triangle and

$$\mathbf{a}_T = \frac{1}{S_T}\int_T \mathbf{a} d\Omega. \tag{1.4}$$

.Let the corners of the triangle be labelled $1,2,3$: then we may write

$$\phi_T = \sum_{\nu=1}^{3} k_\nu U_\nu \tag{1.5}$$

where

$$k_\nu = -\frac{1}{2}\mathbf{a}_T.\mathbf{n}_\nu \tag{1.6}$$

$\mathbf{n}_\nu$ being a vector in the direction of the inward normal to the side ν of magnitude equal to the length of the side. Note that k_3 for example can be written

$$k_3 = \frac{1}{2}\mathbf{a}_T.\begin{pmatrix} (Y_2 - Y_1) \\ -(X_2 - X_1) \end{pmatrix} \tag{1.7}$$

so that, if $\mathbf{a}_T = (a,b)$

$$k_3 = \frac{1}{2}a(Y_2 - Y_1) - \frac{1}{2}b(X_2 - X_1) \tag{1.8}$$

$$= \frac{|\mathbf{a}_T|}{2}\left((\mathbf{r}_2 - \mathbf{r}_1).\widehat{\mathbf{N}}\right) \tag{1.9}$$

$$= \frac{|\bar{\mathbf{a}}_T|}{2}\left((\mathbf{r}_2 - \mathbf{r}_1).\widehat{\mathbf{N}}\right) \tag{1.10}$$

where $\mathbf{r}_2 - \mathbf{r}_1$ is the vector length from vertex 1 to vertex 2 and $\widehat{\mathbf{N}}$ is a unit vector perpendicular to $\mathbf{a}_T$. Therefore k_3 is proportional to the projection of $\mathbf{r}_2 - \mathbf{r}_1$ in the direction of $\widehat{\mathbf{N}}$, denoted by $N_2 - N_1$, say. We may therefore write

$$k_3 = \frac{|\bar{\mathbf{a}}_T|}{2}(N_2 - N_1) \tag{1.11}$$

and hence

$$\phi_T = \frac{|\bar{\mathbf{a}}_T|}{2}\sum U_3 (N_2 - N_1) = \frac{|\bar{\mathbf{a}}_T|}{2}\mathbf{U}_T^t (P_T \mathbf{N}_T) = -\frac{|\bar{\mathbf{a}}_T|}{2}(P_T \mathbf{U}_T)^t \mathbf{N}_T \tag{1.12}$$

where $\mathbf{U}_T$ and $\mathbf{N}_T$ are 3-vectors of U and N values at the cornersof triangle T and P is an antisymmetric permutation matrix which in 2-D takes the form

$$P_T = \frac{1}{2}\begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}. \tag{1.13}$$

Equally via transposition we obtain the dual forms

$$\phi_T = -|\bar{\mathbf{a}}_T|\,\mathbf{N}_T^t (P_T \mathbf{U}_T) = |\bar{\mathbf{a}}_T|\,(P_T \mathbf{N}_T)^t \mathbf{U}_T. \tag{1.14}$$

In the local framework the functions U_T and N_T are defined element by element and from a global point of view there is no reason for them to be continuous. We may however enforce nodal continuity by introducing an assembly matrix M which has the effect of adding the contributions to the local matrix entries from triangles around a node. Then the vectors $\{\mathbf{U}_T\}$ of $\mathbf{U}_T$'s and $\{\mathbf{N}_T\}$ of $\mathbf{U}_T$'s satisfy

$$\{\mathbf{U}_T\}{=}M\mathbf{U} \qquad \{\mathbf{N}_T\}{=}M\mathbf{N}. \tag{1.15}$$

The matrix M may be thought of as node based and is simplest to write down in a node based numbering. Unlike $\mathbf{U}_T$ the direction of $\mathbf{N}_T$ and its zero positionl changes from element to element but $\mathbf{N}$ is still a local average at a node.

2. Fluctuation Norm Minimisation

Solving the set of equations $\phi_T = 0\ \forall T$ for the unknowns $\mathbf{U}_T$ and $\mathbf{N}_T$ is an underdetermined problem even when continuity of $\mathbf{U}$ and $\mathbf{N}$ are enforced. We

seek a solution by least squares using a weighted l_2 norm corresponding to the square of the average residual R_T of (1.1).

Define the square of the local l_2 norm of the average residual of (1.1) in the triangle T to be

$$J_T = \frac{1}{2} S_T R_T^2 = \frac{1}{2S_T} \int_T (\bar{\mathbf{a}}.\nabla U)^2 \, d\Omega = \frac{1}{2} S_T \left(\bar{\mathbf{a}}.\nabla U\right)^2 = \frac{1}{2} \frac{\phi_T^2}{S_T}. \tag{2.1}$$

Using (1.12) and (1.14) we have

$$J_T = \frac{1}{2} \mathbf{U}_T^t \left(P_T \mathbf{N}_T\right) w_T \left(P_T \mathbf{N}_T\right)^t \mathbf{U}_T = \frac{1}{2} \mathbf{N}_T^t \left(P_T \mathbf{U}_T\right) w_T \left(P_T \mathbf{U}_T\right)^t \mathbf{N}_T \tag{2.2}$$

$$= \frac{1}{2} \mathbf{U}_T^t E_{NT} \mathbf{U}_T = \frac{1}{2} \mathbf{N}_T^t E_{UT} \mathbf{N}_T \tag{2.3}$$

where

$$E_{NT} = \left(P_T \mathbf{N}_T\right) w_T \left(P_T \mathbf{N}_T\right)^t, \qquad E_{UT} = \left(P_T \mathbf{U}_T\right) w_T \left(P_T \mathbf{U}_T\right)^t \tag{2.4}$$

and

$$w_T = \frac{\left|\bar{\mathbf{a}}_T\right|^2}{S_T}. \tag{2.5}$$

Note that if w_T is held constant (2.2) is quadratic in each of the variables $\mathbf{U}_T$ and $\mathbf{N}_T$ separately.

The global functional

$$J = \sum_T J_T = \sum_T \frac{1}{2} \frac{\phi_T^2}{S_T} = \frac{1}{2} \sum_T \mathbf{U}_T^t E_{NT} \mathbf{U}_T$$

can then be assembled, using (2.2) and (2.4), as

$$= \frac{1}{2} \mathbf{U}^t M^t \{E_{NT}\} M \mathbf{U} = \frac{1}{2} \mathbf{U}^t M^t E_N M \mathbf{U} = \frac{1}{2} \mathbf{U}^t A_N \mathbf{U} \tag{2.6}$$

where

$$E_N = \{E_{NT}\} \qquad A_N = M^t E_N M. \tag{2.7}$$

Also, interchanging the roles of $\mathbf{U}$ and $\mathbf{N}$,

$$J = \frac{1}{2} \sum_T \mathbf{N}_T^t E_{UT} \mathbf{N}_T$$

$$= \frac{1}{2}\mathbf{N}^t M^t \{E_{UT}\} M\mathbf{N} = \frac{1}{2}\mathbf{N}^t M^t E_U M\mathbf{N} = \frac{1}{2}\mathbf{N}^t A_U \mathbf{N} \tag{2.8}$$

where

$$E_U = \{E_{UT}\} \qquad A_U = M^t E_U M. \tag{2.9}$$

The entries in the matrix A_N are

$$a_{Nij} = \sum_T (P_T \mathbf{N}_T)_i \, w_T \, (P_T \mathbf{N}_T)_j^t = \sum_T w_T \, (N_{i2} - N_{i1}) \, (N_{j2} - N_{j1}) \tag{2.10}$$

when the nodes i and j share a common triangle T and zero otherwise, where N_{i2}, N_{i1} are the N values at the nodes of the side opposite node i taken counter-clockwise. Similarly

$$a_{Uij} = \sum_T (P_T \mathbf{U}_T)_i \, w_T \, (P_T \mathbf{U}_T)_j^t = \sum_T w_T \, (U_{i2} - U_{i1}) \, (U_{j2} - U_{j1}) \, . \tag{2.11}$$

When essential boundary conditions are overwritten J is augmented by additional terms of the form

$$\mathbf{U}^t \mathbf{b} + \mathbf{N}^t \mathbf{c} \tag{2.12}$$

with the number of unknowns in $\mathbf{U}$ and $\mathbf{N}$ reduced accordingly.

A related simpler norm is

$$J_D = \frac{1}{2}\mathbf{N}^t D_U \mathbf{N} = \frac{1}{2}\mathbf{U}^t D_N \mathbf{U} \tag{2.13}$$

where

$$D_U = diag\{A_U\}, \qquad D_N = diag\{A_N\}. \tag{2.14}$$

Since the diagonal matrices D_N, D_U of A_N, A_U are assemblies of local matrices on each triangle, in this case their diagonals,

$$D_U = M^t diag\{E_{UT}\} M \qquad D_N = M^t diag\{E_{NT}\} M \tag{2.15}$$

the diagonal entries of D_U, D_N (and A_U, A_N) are as in (2.10),(2.11) with $i = j$.

3. Rayleigh Quotient

If D_U is non-singular, then for any vector $\mathbf{y}$ the Rayleigh Quotient

$$\frac{\mathbf{y}^t A_U \mathbf{y}}{\mathbf{y}^t D_U \mathbf{y}} = \frac{\mathbf{y}^t M^t E_U M \mathbf{y}}{\mathbf{y}^t M^t diag\{E_U\} M \mathbf{y}} = \frac{\mathbf{z}^t E_U \mathbf{z}}{\mathbf{z}^t diag\{E_U\} \mathbf{z}} \tag{3.1}$$

$$= \frac{\sum_T \mathbf{z}_T^t \left(E_{UT}\right) \mathbf{z}_T}{\sum_T \mathbf{z}_T^t diag\left(E_{UT}\right) \mathbf{z}_T} \tag{3.2}$$

where $\mathbf{z} = M\mathbf{y}$ and $\mathbf{z} = \{\mathbf{z}_T\}$. By expanding $\mathbf{z}_T$ in terms of the normalised eigenvectors $\mathbf{e}_j$ of $|\mathbf{a}_T|\, diag\left(E_{UT}\right)^{-1} E_{UT}$ with coefficients z_j the Rayleigh Quotient becomes

$$\frac{\sum_T^t \sum_j \sum_k z_j \lambda_k diag\left(E_{UT}\right) z_k}{\sum_T^t \sum_j \sum_k z_j diag\left(E_{UT}\right) z_k}$$

from which it can be shown (see e.g. [3], [4]) that it lies between the maximum and minimum values of the λ_k's, i.e within the roots of the determinantal equation

$$\begin{vmatrix} 1-\lambda & 1 & 1 \\ 1 & 1-\lambda & 1 \\ 1 & 1 & 1-\lambda \end{vmatrix} = 0 \tag{3.3}$$

which are 0 (twice) and 3. With boundary conditions applied the zero root becomes positive (ε) though small. Thus, provided that D_U is non-singular, the diagonal norm J_D stays close to J irrespective of the values of $\mathbf{U}$ (or of $\mathbf{a}_T$ or S_T). A similar argument holds for A_N.

Equivalently, the eigenvalues of $D_U^{-1} A_U$ and $D_N^{-1} A_N$ lie between ε and 3.

4. Solution Methods

Descent methods for the minimisation of J are of the form

$$\mathbf{U}^{p+1} = \mathbf{U}^p + \tau^p \left(H^p\right)^{-1} \mathbf{d}^p \tag{4.1}$$

where H approximates the Hessian of J and $\mathbf{d}$ is a descent direction. The steepest descent method may be written as

$$\mathbf{U}^{p+1} = \mathbf{U}^p - \tau^p \nabla_{\mathbf{U}} J^p = \left(I - \tau A_N\right)^p \mathbf{U}^p + \tau^p \mathbf{b} \tag{4.2}$$

or

$$\mathbf{N}^{p+1} = \mathbf{N}^p - \sigma^p \nabla_{\mathbf{N}} J^p = \left(I - \sigma A_U\right)^p \mathbf{N}^p + \sigma^p \mathbf{c} \tag{4.3}$$

(see (2.6), (2.8) and (2.12)) where τ and σ are the stepsizes and w_T is held constant. This is a descent method with $\mathbf{d} = -\nabla J$ and H approximated by I. A modified steepest descent method is obtained by replacing H^p by D^p in (4.1) giving

$$\mathbf{U}^{p+1} = \left(I - \tau D_N^{-1} A_N\right)^p \mathbf{U}^p + \tau^p \left(D_N^{-1}\right)^p \mathbf{b} \tag{4.4}$$

or

$$\mathbf{N}^{p+1}=\left(I-\sigma D_U^{-1}A_U\right)^p\mathbf{N}^p+\sigma^p\left(D_U^{-1}\right)^p\mathbf{c}. \tag{4.5}$$

In any sweep each U_i^{p+1}, N_i^{p+1} component is a weighted average of its surrounding values plus a boundary term where appropriate.

Since the eigenvalues of $D^{-1}A$ lie between ε and 3 irrespective of their arguments (where is ε small) the eigenvalues of $I-\tau D^{-1}A$ lie between $1-\tau\varepsilon$ and $1-3\tau$. So choosing $\tau<2/3$ ensures that $\left|\rho\left(I-\tau D_N^{-1}A_N\right)\right|<1$ with a similar result for σ. The only proviso is that the matrix D is non-singular which in practice means that the modulus of the diagonal entries is not less than a certain tolerance TOL. If it is less than TOL the update is set to zero, which from (2.10) and (2.11) occurs when

$$\left|\sum_{T_i} w_T\left(N_{i2}-N_{i1}\right)^2\right|<TOL \tag{4.6}$$

or

$$\left|\sum_{T_i} w_T\left(U_{i2}-U_{i1}\right)^2\right|<10^{-10}. \tag{4.7}$$

A well-behaved algorithm can therefore be constructed in which the solution is found from a suitable initial guess by descent methods. An argument for this approach is that each variable depends on the other in the minimisation, so they should be iterated simultaneously. Direct Methods such as CG or GMRES may be used to improve the initial guess for $\mathbf{U}$.

For the J_D norm (2.13) the corresponding iterations, using (2.10) and (2.11), are $\forall i$

$$U_i^{p+1}=\left(1-\tau_i^p\sum_{T_i} w_T\left(N_{i2}-N_{i1}\right)^2\right)U_i^p+\tau_i^p b_i \tag{4.8}$$

or

$$N_i^{p+1}=\left(1-\sigma_i^p\sum_{T_i} w_T\left(U_{i2}-U_{i1}\right)^2\right)N_i^p+\sigma_i^p c_i. \tag{4.9}$$

In the conjugate gradient method the directions $\mathbf{d}$ in the descent method are made A-orthogonal to each other. The updates are

$$\mathbf{U}^{p+1}=\mathbf{U}^p-\tau^p\mathbf{d}_U^p \quad \text{or} \quad \mathbf{N}^{p+1}=\mathbf{N}^p-\sigma^p\mathbf{d}_N^p \tag{4.10}$$

where

$$\mathbf{d}_U^p=-\mathbf{r}_U^p+\beta^{p-1}\mathbf{d}_U^{p-1} \quad \text{or} \quad \mathbf{d}_N^p=-\mathbf{r}_N^p+\gamma^{p-1}\mathbf{d}_U^{p-1} \tag{4.11}$$

$$\mathbf{r}_U^p = A_U\mathbf{U} - \mathbf{b} \quad \text{or} \quad \mathbf{r}_N^p = A_N\mathbf{N} - \mathbf{c} \tag{4.12}$$

with

$$\mathbf{d}^0 = -\mathbf{r}^0. \tag{4.13}$$

Here

$$\tau^p = \frac{(\mathbf{r}_U^p)^t\,\mathbf{d}_U^p}{(\mathbf{d}_U^p)^t\,A_U\mathbf{d}_U^p}, \qquad \beta^p = \frac{\left(\mathbf{r}_U^{p+1}\right)^t A_U\mathbf{d}_U^p}{(\mathbf{d}_U^p)^t\,A_U\mathbf{d}_U^p} \tag{4.14}$$

or

$$\sigma^p = \frac{(\mathbf{r}_N^p)^t\,\mathbf{d}_N^p}{(\mathbf{d}_N^p)^t\,A_N\mathbf{d}_N^p}, \qquad \gamma^p = \frac{\left(\mathbf{r}_N^{p+1}\right)^t A_N\mathbf{d}_N^p}{(\mathbf{d}_N^p)^t\,A_N\mathbf{d}_N^p}. \tag{4.15}$$

The residuals are

$$\mathbf{r}_U = A_N\mathbf{U} - \mathbf{b} \quad \text{or} \quad \mathbf{r}_N = A_U\mathbf{N} - \mathbf{c}. \tag{4.16}$$

The method may be preconditioned using the diagonal matrices D_U or D_N.

5. Role of $\mathbf{w}_T$

When w_T is not held fixed the functional J is no longer quadratic in $\mathbf{N}$ and there are extra terms in the gradient of J of the form (see (2.3))

$$-\frac{1}{4}\sum_T \frac{\phi^2}{S_T^2}\nabla_{\mathbf{X},\mathbf{Y}} S_T = -\frac{1}{4}\sum_T \frac{\phi^2}{S_T^2}\,(P\mathbf{Y}_T, -P\mathbf{X}_T) \tag{5.1}$$

since

$$S_T = \frac{1}{2}\sum X_T\Delta Y_T = \mathbf{X}_T^t P\mathbf{Y}_T = -\frac{1}{2}\sum Y_T\Delta X_T = -\mathbf{Y}_T^t P\mathbf{X}_T. \tag{5.2}$$

Hence (4.3) receives extra terms

$$-\frac{\sigma}{4}\sum_{T_i}\frac{\phi^2}{S_T^2}\frac{\bar{\mathbf{a}}.\mathbf{n}_i}{|\bar{\mathbf{a}}|} \tag{5.3}$$

where $\mathbf{n}_i$ is the scaled inward normal to the side opposite node i, and there is also an additional steepest descent equation for the coordinates C in the characteristic direction of the form

$$\mathbf{C}^{p+1} = \mathbf{C}^p + \frac{\varrho}{4}\sum_{T_i}\frac{\phi^2}{S_T^2}P\mathbf{N}. \tag{5.4}$$

These additional factors represent nodal movements in which the elements compete for area and as such are smoothing effects. The same can be said for introducing diagonal swapping with area as the criterion for swapping. On the other hand the minimisation over **N** is directed towards making the residual zero. This is equivalent to aligning particular sides of the triangles with the local characteristic direction and as such is a directional rather than a smoothing operation. When both effects occur together there is inevitably some cancellation which reduces the effectiveness of either.

Alternatively the variation of J with respect to S_T can be used to construct separate iterative descent steps of the form

$$\begin{pmatrix} N \\ C \end{pmatrix}_i^{p+1} = \begin{pmatrix} N \\ C \end{pmatrix}_i^{p} + \frac{\rho^p}{2} \left(\sum_{\{T_i\}} S_T^{-2} \phi_T^2 \begin{pmatrix} PN \\ PC \end{pmatrix}_i \right)^p \tag{5.5}$$

$\forall i$, i.e.

$$\begin{pmatrix} N \\ C \end{pmatrix}_i^{p+1} = \begin{pmatrix} N \\ C \end{pmatrix}_i^{p} + \frac{\rho^p}{2} \left(\sum_{\{T_i\}} S_T^{-2} \phi_T^2 \begin{pmatrix} C_{i2} - C_{i1} \\ -(N_{i2} - N_{i1}) \end{pmatrix}_i \right)^p \tag{5.6}$$

or

$$\mathbf{r}_i^{p+1} = \mathbf{r}_i^p + \frac{\rho^p}{2} \left(\sum_{\{T_i\}} S_T^{-2} \phi_T^2 \mathbf{n}_i \right)^p \tag{5.7}$$

where $\mathbf{r}_i$ is the displacement of node i and $\mathbf{n}_i$ is the scaled inward normal (by side length) to the side opposite node i. Since $S_T = |\mathbf{n}_i| \, h_i$ where h_i is the height of the triangle measured from node i we have

$$\mathbf{r}_i^{p+1} = \mathbf{r}_i^p + \frac{\rho^p}{2} \left(\sum_{\{T_i\}} S_T^{-1} \phi_T^2 h_i \hat{\mathbf{n}}_i \right)^p . \tag{5.8}$$

where $\hat{\mathbf{n}}_i$ is a unit vector in the direction of $\mathbf{n}_i$. A suitable dimensional scaling is to divide the update by

$$\left(\sum_{\{T_i\}} S_T^{-2} \phi_T^2 \right)^p \tag{5.9}$$

giving

$$\mathbf{r}_i^{p+1} = \mathbf{r}_i^p + \chi^p \left(\frac{\sum_{\{T_i\}} R_T^2 h_i \hat{\mathbf{n}}_i}{\sum_{\{T_i\}} R_T^2} \right)^p \tag{5.10}$$

where $S_T^{-1} \phi_T = R_T$ is the average residual and χ is a new relaxation coefficient.

6. Upwind Minimisation

Suppose now that for an *interior* triangle the update to $\mathbf{U}_T$ is carried out over only those corner values of U which correspond to *downwind* nodes. Denote such an update by a superfix d. Then the local update becomes

$$\mathbf{U}_T^{p+1} - \mathbf{U}_T^p = -\tau^p E_{NT}^d \mathbf{U}_T^p - \mathbf{b}_T \tag{6.1}$$

where E_T^d is the part of E_T corresponding to the downwind nodes, the rest of E_{NT}^d consisting of zeros. After assembly we have (cf (4.2))

$$\mathbf{U}^{p+1} - \mathbf{U}^p = -\tau^p M^t E_N^d M \mathbf{U}_T^p. \tag{6.2}$$

However since the matrix $M^t E_N^d M$ is not symmetric, it is clear that (6.2) cannot be obtained from a least squares minimisation.

On the other hand, if we confine the calculation to the elements adjacent to the inflow boundary (as if the downwind corners were outflow points), overwriting of the inflow values ensures that each local matrix is symmetric with respect to the downwind points. For example, if the least squares method is applied only to the strip of elements adjacent to inflow, it is indeed equivalent to an upwind method.

Carrying out the minimisation over U in this way results in a scheme close to the LDA scheme [3]. The update is of the form

$$U_i^{p+1} - U_i^p = -\tau \sum k_i \phi_i \tag{6.3}$$

where the sum is taken over only upwind triangles, which corresponds to the LDA scheme provided that the non-zero k_i are normalised so that they sum to 1.

The question arises whether the upwind least squares approach can also usefully be applied to the coordinate N. If N is treated in the same way as U, namely obeying inflow conditions which enable N to be overwritten and leaving outflow conditions free, then there is no reason why it shouldn't. The key point is that the minimisation should be done a strip at a time, working forward in a frontal manner through the grid.

7. References

[1] P.L.Roe (1996), Compounded of Many Simples, in Proceedings of Workshop on Barriers and Challenges in CFD, ICASE, NASA Langley, August 1996, Kluwer (to appear).

[2] H.Paillere, H Deconinck and P L Roe (1995), Conservative Upwind Residual-Distribution Schemes based on the Steady Characteristics of the Euler Equations, AIAA 12th CFD Conference, 592-605.

[3] A.J.Wathen (1986), Mesh-independent Spectra in the Moving Finite Element Equations, SIAM J Numer An, 23, 797-814.

[4] K.Miller (1992), On the Matrix Spectrum Bounds of Wathen and the Local Moving Finite Elements of Baines, SIAM J Numer An 29, 89-106.