# Differentiability of Slope Limiters on Unstructured Grids

Krzysztof Michalak and Carl F. Ollivier-Gooch

*Department of Mechanical Engineering, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada*

Email: *michalak@mech.ubc.ca*

## ABSTRACT

Despite emerging alternatives such as essentially nonoscillatory schemes, the use of slope limiters remains a standard means of eliminating oscillations in the second-order accurate solution of hyperbolic equations. However, slope limiters for unstructured grids are known to cause convergence problems. We demonstrate that fully-implicit solutions using Newton-Krylov methods are particularly affected. The root cause of this problem is shown to be the lack of differentiability of the limiting procedure. A modification to the limiting procedure is introduced and shown to improve convergence.

## 1 INTRODUCTION

A common means of achieving second-order accurate schemes for hyperbolic transport equations on unstructured grids is through the use of the MUSCL[5] finite-volume approach. The flux is obtained by first computing a piecewise linear reconstruction of the solution within each control volume and then using an approximate Riemann solver at Gauss points on the edges. However, unless special precautions are taken, the scheme produces oscillations near solution discontinuities. The use of a slope-limiter is a simple and efficient means of rectifying this problem. However, the use of limiters often hampers convergence to a steady-state solution. This has been observed with the use of explicit and approximately-factored implicit methods. As will be discussed herein, this problem is exacerbated when a fully-implicit method is used.

Newton-Krylov methods are gaining acceptance as an efficient means of obtaining steady-state solutions to CFD problems. The linear system arising at each Newton step is solved approximately using a Krylov-subspace method. Linear solvers such as GMRES[4] allow solutions of matrices with arbitrary fill and in fact can be implemented in a "matrix-free" fashion. This, in turn, means that the fully-coupled nature of the flow equations can be preserved in the discretization. The full benefit of the quadratically-converging Newton method can therefore be achieved when the objective function remains differentiable.

This work discusses the non-differentiability of the Barth and Jespersen[1] limiter and how the work of Venkatakrishnan[6] only partially addresses the problem. A modification to the standard limiting procedure is introduced and shown to improve convergence to steady-state.

## 2 FLOW SOLVER

In the present work, compressible inviscid flow modeled by the Euler equations in integral form is solved in two dimensions on a triangular unstructured grid. The governing equations for a control volume $\Omega$ with boundary $\partial\Omega$ are

$$\frac{\partial}{\partial t}\int_{\Omega}\mathbf{u}\,dV+\oint_{\partial\Omega}(\mathbf{f}\cdot\mathbf{n})\,dS=0$$

The conserved variables $\mathbf{u}$ are the density, two components of momentum and total energy. The vector $\mathbf{f}$ represents the inviscid flux. A median-dual control-volume is associated with each vertex in the triangular mesh. The contour integral is discretized using Gauss-quadrature over the dual-edges. To achieve second-order accuracy, a three-stage approach is taken to construct the discrete approximation to the flux $\mathbf{f}$:

1. A piecewise linear reconstruction of the primitive variables (density, velocity, and pressure) is performed by solving a constrained least-squares system. The resulting polynomial conserves the mean value of the solution in the control-volume. When extrapolated it also approximates nearby control-volume averages.

2. The reconstructed solution is computed at the Gauss quadrature points on the control-volume

edges. The limiter is then applied such that solution monotonicity is enforced.

3. The limited reconstructed solution on either side of the control volume edges are used as input to the approximate Riemann solver. In the present work, the AUSM+[3] scheme is used.

The discretization leads to a system of ordinary differential equations:

$$\frac{dU}{dt} + R(U) = 0$$

where $U$ are the control volume averages of the solution. $R$ is the discrete flux integral which we will call the residual. An implicit scheme is obtained by the following discretization :

$$\begin{aligned} \left(\frac{I}{\Delta t} + \frac{\partial R}{\partial U}\right)\delta U &= -R \\ U^{n+1} &= U^n + \delta U \end{aligned}$$

Since we are strictly concerned with the steady-state solution, $\Delta t$ could be taken as infinity and the equation simplified to Newton's method. In most practical cases this is not possible due to the highly non-linear nature of the residual. However, the time-step can be gradually increased to infinity as the solution process proceeds. A local time-step is used where $\Delta t$ is taken as proportional to the control-volume size. The linear system at each time-step is solved approximately using a preconditioned GMRES[4] solver. ILU-p with four levels of fill is used as the preconditioner.

To benefit from the quadratic convergence properties of the Newton method, $R$ needs to be differentiable and the left-hand-side term $\frac{\partial R}{\partial U}$ needs to be accurately represented. Due to the reconstruction procedure it would be very difficult to explicitly form this Jacobian matrix. Furthermore, the matrix will have more fill and a larger condition number than the first-order accurate flux Jacobian. For this reason some previous works[7] have used the first-order accurate flux Jacobian on the left-hand-side and the second-order residual on the right-hand-side. The alternative is to use a matrix-free[2] approach where the matrix-vector products are replaced by finite-difference approximations. The resulting scheme requires one residual evaluation per inner GMRES iteration. Each time-step is therefore considerably more expensive, but Newton-like convergence can be obtained when the time-step is increased. Since the preconditioner requires an explicit representation of the matrix, the first-order Jacobian is still formed for this.

Optimum performance is achieved by using a combination of these methods. At the beginning of the solution process, when small values of $\Delta t$ are needed and quadratic convergence cannot be expected, the first-order Jacobian is used. After a certain level of convergence has been reached the matrix-free approach is used to obtain quadratic-convergence.

For robustness, our scheme adjusts the magnitude of the update vector $\delta U$ by using a cubic line-search. If the result indicates that the magnitude did not need to be reduced the time-step at the next iteration is doubled. Otherwise, the next time-step is reduced based on the adjusted relative size of the update vector. This scheme requires additional residual evaluations at each time-step due to the line-search. However, since the time-step is aggressively increased when appropriate, the overall number of residual evaluations is not adversely affected. Furthermore, the convergence process is very robust and does not require manual adjustments to the startup procedure to obtain good performance on a variety of test cases.

In the present work, we use the first-order Jacobian for up to 100 time-steps. If during this process a relative residual drop of $5 \times 10^{-3}$ is achieved, we immediately switch to a full-Newton scheme using the second-order matrix-free Jacobian. Otherwise, after 100 iterations, timestepping is continued with the left-hand-side based on the matrix-free Jacobian.

## 3 BARTH-JESPERSEN LIMITER

Barth and Jespersen[1] introduced the first multidimensional limiter which allowed oscillation-free solutions of transonic flows on irregular triangular meshes. The scheme consists of finding a value $\Phi$ that will limit the gradient in the piecewise-linear reconstruction of the solution in each control-volume.

$$u(x,y)_i = \bar{u}_i + \Phi_i \bigtriangledown u_i \cdot \triangle r_i,\ \Phi \in [0,1]$$

The goal is to find the largest $\Phi_i$ which prevents the formation of local extrema at the flux integration Gauss points. The following procedure is used by Barth and Jespersen :

1. Find the largest negative ($\delta u_i^{min} = \min(\bar{u} - \bar{u}_i)$) and positive ($\delta u_i^{max} = \max(\bar{u} - \bar{u}_i)$) difference between the solution in the immediate neighbours and the current control volume.

2. Compute the unconstrained reconstructed value at each Gauss point ($u_{ij} = u(x_j, y_j)$).

3. Compute a maximum allowable value of $\Phi_{ij}$ for

each Gauss point $j$.

$$\Phi_{ij} = \begin{cases} \min(1, \frac{\delta u_i^{max}}{u_{ij} - \bar{u}_i}), & \text{if } u_{ij} - \bar{u}_i > 0 \\ \min(1, \frac{\delta u_i^{min}}{u_{ij} - \bar{u}_i}), & \text{if } u_{ij} - \bar{u}_i < 0 \\ 1, & \text{if } u_{ij} - \bar{u}_i = 0 \end{cases}$$

4. Select $\Phi_i = \min(\Phi_{ij})$ .

Clearly, steps 1, 3, and 4 introduce non-differentiability in the computation of the reconstructed function. Consequently, the second-order flux is also non-differentiable.

This limiting procedure is akin to reducing the magnitude of the gradient vector such that it falls within a region of maximum allowable gradient. This is shown in Figure 1. The monotonicity requirement at each Gauss point is represented by an inequality constraint. Graphically, the angle the constraint line forms is based purely on the geometry of the mesh. However its distance from the origin is based on the results of steps 1 and 3 of the limiting procedure. If the solution is perturbed the constraint line moves smoothly only as long as the result of the switches in steps 1 and 3 remains the same. The non-differentiability of step 4 can be visualized by considering the consequence of the solution being perturbed gradually from Figure 1 to Figure 2. The limited gradient has a singularity at the corner of the region of allowable gradient where the active constraint is changed.

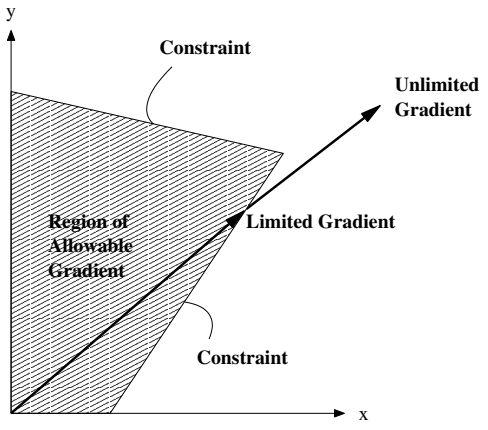


Figure 1: The limiting procedure can be visualized in terms of a region of maximum allowable gradient

## 4 VENKATAKRISHNAN LIMITER

Venkatakrishnan[6] introduces a smooth alternative to step 3 of the Barth-Jespersen procedure. For the case

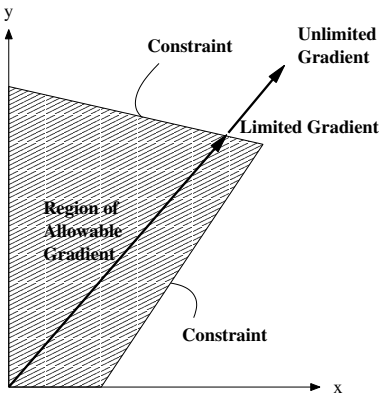


Figure 2: The result of the limiting procedure after a small perturbation in the solution

$u_{ij} - \bar{u}_i > 0$ the limiter becomes

$$\Phi_{ij} = \frac{1}{\Delta_-}\left[\frac{(\Delta_+^2 + \varepsilon^2)\Delta_- + 2\Delta_-^2\Delta_+}{\Delta_+^2 + 2\Delta_-^2 + \Delta_-\Delta_+ + \varepsilon^2}\right]$$

In this equation $\Delta_- = u_{ij} - \bar{u}_i$, $\Delta_+ = \delta u_i^{max}$ and $\varepsilon^2 = (k\Delta x)^3$. With $k = 0$, this modification maintains strict-monotonicity. Graphically, on the maximum allowable gradient plot, this limiter brings the constraints slightly closer to the origin than the Barth-Jespersen limiter. The benefit is that the constraint lines move smoothly as long as the result of the switch in step 1 of the limiting process remains the same.

The introduction of the tunable $k$ term serves to increase the allowable gradient by moving all the constraint lines outward by some small constant. This allows monotonicity to be violated slightly. The singularity due to step 4 (change in active constraint) occurs most frequently where the flow is nearly uniform. When this is the case, the constraint lines can be thought as all being very close to the origin. Therefore, even small changes in the solution will cause frequent changes in the active-constraint. The tunable term $k$ helps to alleviate this problem. In addition to convergence benefits this also decreases the diffusivity and increases the accuracy of the scheme in smooth flows. However large values of $k$ introduce significant oscillations near shocks.

Venkatakrishnan's limiter therefore provides a complete solution to the differentiability problem of step 3 and a partial solution to the differentiability problem of step 4. However, as the original[6] work shows, convergence is still a problem for transonic flows unless $k$ is chosen to be large enough to cause substantial oscillations to occur near the shock.

## 5 THE NEW LIMITING PROCEDURE

To achieve complete differentiability, we seek a differentiable approximation to the min and max functions used in steps 1 and 4 of the limiting procedure. The new functions, which we will call softmin and softmax, take a set of values $x = \{x_1, x_2, \ldots, x_n\}$ as input and return a value that is near the minimum and maximum respectively. The softmax function is selected such that it meets the following requirements.

- At the limit when $\max(x)$ is much greater than all other $x$, $\text{softmax}(\text{x}) = \max(\text{x})$.
- If $x_1 = x_2 = \max(x)$ then $\frac{\partial \text{softmax(x)}}{\partial x_1} = \frac{\partial \text{softmax(x)}}{\partial x_2}$.
- $\text{softmax}(\text{x})$ is differentiable at all points with respect to all $x$.
- To ensure strict monotonicity of the scheme, we require $\text{softmax}(\text{x}) \leq \max(x)$.

Similarly, the requirement for the function softmin are:

- At the limit when $\min(x)$ is much smaller than all other $x$, $\text{softmin}(\text{x}) = \min(\text{x})$.
- If $x_1 = x_2 = \min(x)$ then $\frac{\partial \text{softmin(x)}}{\partial x_1} = \frac{\partial \text{softmin(x)}}{\partial x_2}$.
- $\text{softmin}(\text{x})$ is differentiable at all points with respect to all $x$.
- To ensure strict monotonicity in step 1 we require $\text{softmin}(\text{x}) \geq \min(x)$.

The use of the softmin function in step 4 of the limiting procedure will cause strict monotonicity to be violated due to the property $\text{softmin}(\text{x}) \geq \min(x)$. However, the experimental results indicate that no significant oscillation result from this.

The functions are constructed by weighing the contribution of each member of the set based on how close it is to being the minimum or maximum value. The proposed softmax function is :

$$\text{softmax}(\text{x}) = \frac{\sum(\text{W}_\text{i}\text{x}_\text{i})}{\sum \text{W}_\text{i}}$$

$$W_i = \exp\Big(-\alpha \frac{\max(x) - x_i}{\text{scale}(\text{x})}\Big)$$

The softmin function is formed by replacing the max with min in weight equation. In step 1 of the limiting procedure, $\text{scale}(x)$ is taken as the $L_2$ norm of $x$. On the other hand, in step 4, $\text{scale}(x)$ can be fixed at 1. In

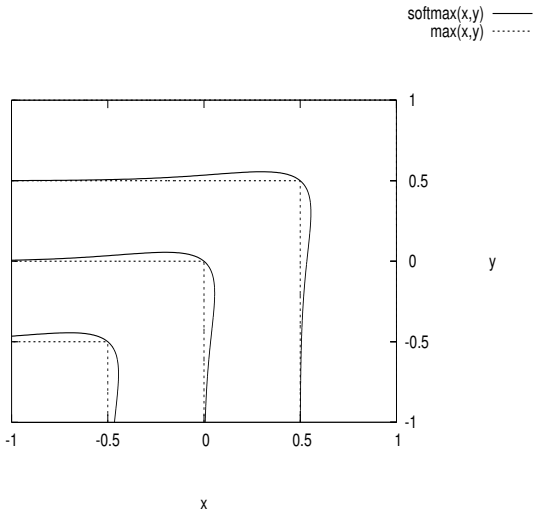


Figure 3: Isolines of the softmax function compared to the max function

Figure 3, the softmax function isolines are plotted with $\alpha = 5$ and $\text{scale}(\text{x}) = 1$.

The tunable parameter $\alpha$ determines the trade-off between smoothness and accuracy of the softmax function. Large values of $\alpha$ help convergence by reducing the second derivatives $\frac{\partial^2 \text{softmax(x)}}{\partial x_i^2}$. However, small values of $\alpha$ are desirable so that $\text{softmax}(\text{x}) \approx \max(\text{x})$.

In step 1 of the limiting procedure, if $\bar{u}_i$ is included in the set $\bar{u}_j$, we expect $\max(\bar{u}_j - \bar{u}_i) \geq 0$ . However, the softmax function does not share this property. Failing to enforce $\text{softmax}(\bar{u}_j - \bar{u}_i) \geq 0$ would potentially lead to negative values of $\Phi_{ij}$. Therefore, we seek a smooth approximation to a binary function $\max(x, 0)$ so that it can be applied in tandem with the proposed approach ($\text{softpositive}(\text{softmax}(\text{x}))$). We define this new function as :

$$\text{softpositive}(\text{x}) = \frac{\text{x} + \sqrt{\text{x}^2 + \varepsilon^2}}{2}$$

Similarly, when finding the minimum neighbouring value in step 1 we make sure the value is negative by using :

$$\text{softnegative}(x) = -\text{softpositive}(-x)$$

The $\varepsilon$ in these equations plays a similar role to the $\varepsilon$ in the Venkatakrishnan limiter. In addition to providing differentiability in areas of nearly-constant flow, it allows small reconstruction gradients in areas of nearly homogeneous solution to remain unlimited. This new

value of $\varepsilon$ is taken to be proportional to the cell size ($C \cdot \Delta x$). The introduction of this new procedure makes the use of the $\varepsilon$ parameter in Venkatakrishnan's limiter unnecessary.

# 6 Results and Discussion

## 6.1 Test Case 1

The first test-case consists of transonic flow over a NACA 0012 airfoil with free-stream Mach number 0.8 and angle of attack 1.25 degrees. In this configuration, a strong shock is present on the top of the airfoil and a weak shock is present on the bottom. The computational grid consists of 6197 control volumes and is shown in Figure 4. The solution is converged to a

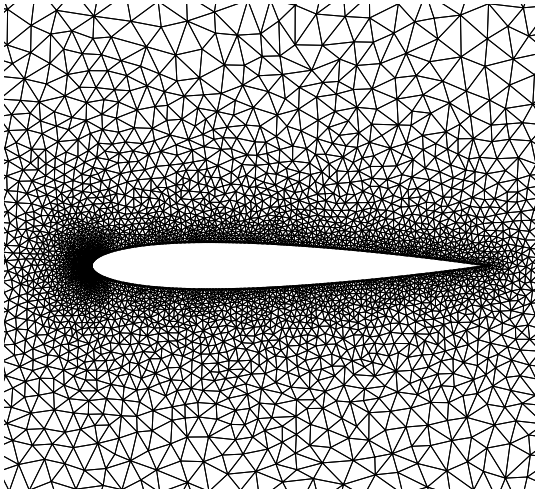

Figure 4: Computational mesh with 6197 vertices

relative residual drop of $5 \times 10^{-3}$ using pre-iterations consisting of time-stepping with the first-order Jacobian. The scheme then uses full Newton steps and the finite-difference matrix-vector products.

The convergence properties for Venkatakrishnan's limiter is compared with that of the new limiter ($\alpha = 25$) in Table 1. A convergence plot for select cases is also shown in Figure 5. From these results, we note some convergence improvement due to the new limiter. In particular, both the pre-iteration and Newton iteration count has been reduced. In the case of the Venkatakrishnan limiter, convergence appears to be very sensitive to the choice of the parameter $k$. The convergence of the new scheme is only slightly dependant on the choice of parameter $C$.

Next, the quality of the solution is assessed in Figure 6. The results using the new limiter are virtually indistinguishable from those of the Venkatakrishnan limiter with $k = 6$. With $k = 12$, the base scheme forms substantial oscillations near the shock.

Overall, a small improvement in convergence speed and robustness was achieved for this test case with no degradation of solution quality.

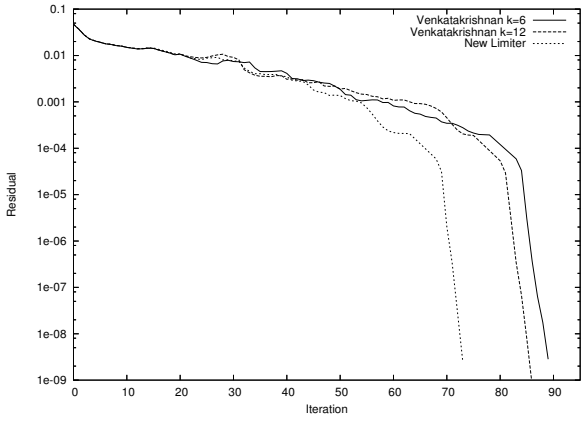


Figure 5: Convergence history for test case 1

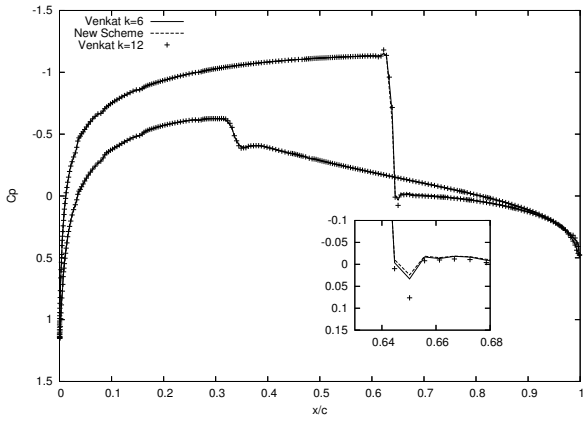


Figure 6: Surface pressure profile for test case 1. The New limiter was tuned with C=1.0 and $\alpha = 25$.

## 6.2 Test Case 2

The second test-case consists of transonic flow over a NACA 0012 airfoil with free-stream Mach number 0.85 and angle of attack 1 degree. The small perturbation of flow conditions relative to the first test case causes the formation of a strong shock on the bottom of the airfoil. Computationally, this test case is much

more demanding since the location of the shock on the top of the airfoil is strongly coupled to the location of the shock on the lower airfoil. For this reason, convergence using pre-iterations fails to yield a solution that is sufficiently converged to allow the use of full-Newton steps. Therefore, after 100 pre-iterations, time-stepping with the matrix-free Jacobian is used.

The convergence results are presented in Table 2 and in Figure 7. Substantial improvements are achieved with the new limiter. The number of fully-implicit time-steps and the total number of residual evaluations are both reduced by a factor of 2.

The pressure plot for this test case is presented in Figure 8. As in the first test case, the new limiter has a result comparable to that of the Venkatakrishnan limiter with $k = 6$ while the limiter with $k = 12$ shows excessive oscillations.

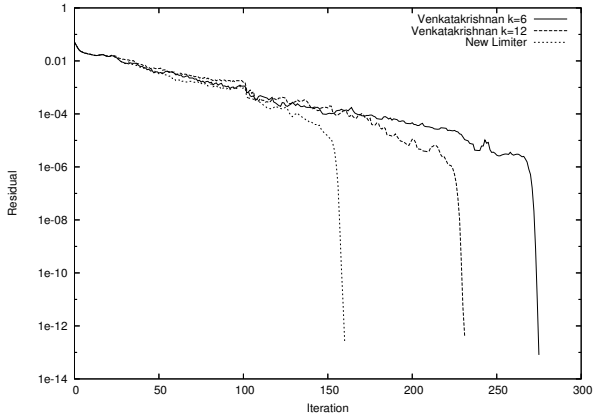


Figure 7: Convergence history for test case 2

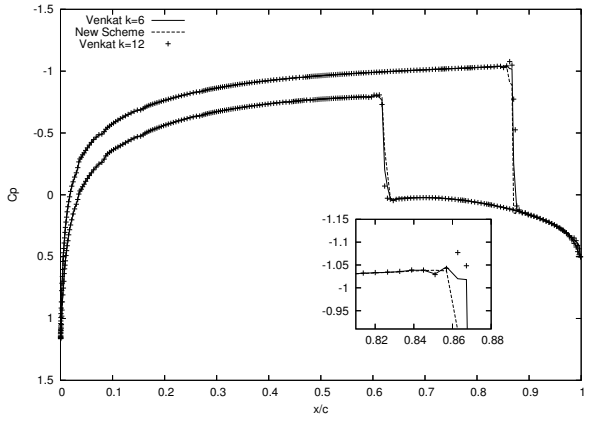


Figure 8: Surface pressure profile for test case 2. The New limiter was tuned with C=1.0 and $\alpha = 25$.

## 7 Conclusion

We have shown that lack of complete differentiability of existing limiters is of particular concern when using fully-implicit schemes on flows with strongly coupled shocks. In these cases, the new limiting procedure dramatically improves convergence without any negative effects on the solution quality. The new scheme is also less sensitive to its tuning parameters than the Venkatakrishnan limiter. This helps to provide additional robustness in the solution process.

## Acknowledgments

This work was supported by the Canadian Natural Sciences and Engineering Research Council under Operating Grant OPG-0194467 and Special Research Opportunities Grant SRO-299160.

## References

[1] T. J. Barth and D. C. Jespersen. The design and application of upwind schemes on unstructured meshes. In *27th Aerospace Sciences Meeting*, Jan. 1989. AIAA paper 89-0366.

[2] T. J. Barth and S. W. Linton. An unstructured mesh Newton solver for compressible fluid flow and its parallel implementation. AIAA Conference Paper 95-0221, 1995.

[3] M. S. Liou. A sequel to AUSM: AUSM+. *J. Computational Physics.*, 129:364–382, 1996.

[4] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving non-symmetric linear systems. *J. Sci. Stat. Comput.*, 7(3):856–869, 1986.

[5] B. van Leer. Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's Method. *J. Computational Physics*, 32:101–136, 1979.

[6] V. Venkatakrishnan. Convergence to steady state solutions of the Euler equations on unstructured grids with limiters. *Journal of Computational Physics*, 118(1):120–130, Apr. 1995.

[7] V. Venkatakrishnan and D. J. Mavripilis. Implicit solvers for unstructured meshes. *J. Computational Physics*, 105:83–91, 1993.

| Limiter | | | Convergence Results | | |
|---|---|---|---|---|---|
| Type | k | C | Pre-iterations | Newton iterations | Residual Evaluations |
| Venkat | 4 | | 83 | 12 | 564 |
| Venkat | 6 | | 79 | 7 | 350 |
| Venkat | 8 | | 66 | 7 | 303 |
| Venkat | 12 | | 76 | 7 | 384 |
| New | 0 | 0.25 | 65 | 7 | 353 |
| New | 0 | 0.5 | 67 | 6 | 332 |
| New | 0 | 1.0 | 64 | 6 | 297 |
| New | 0 | 2.0 | 64 | 5 | 332 |

Table 1: Convergence summary for test case 1

| Limiter | | | Convergence Results | | |
|---|---|---|---|---|---|
| Type | k | C | Pre-iterations | Implicit iterations | Residual Evaluations |
| Venkat | 4 | | 100 | 140 | 3332 |
| Venkat | 6 | | 100 | 175 | 2736 |
| Venkat | 8 | | 100 | 134 | 2053 |
| Venkat | 12 | | 100 | 131 | 2357 |
| New | 0 | 0.25 | 100 | 84 | 1743 |
| New | 0 | 0.50 | 100 | 71 | 1434 |
| New | 0 | 1.00 | 100 | 60 | 1144 |
| New | 0 | 2.00 | 100 | 56 | 1117 |

Table 2: Convergence summary for test case 2