VKI March 2003 Lecture Series

Numerical Methods for Conservation Laws on Structured and Unstructured Meshes

Timothy Barth[1]
NASA Ames Research Center
Moffett Field, California 94035 USA
barth@nas.nasa.gov

Contents

Chapter 1

Basic Design Principles in the Construction of Upwind Finite Volume Methods

Abstract: Finite volume methods are a class of discretization schemes that have proven highly successful in approximating the solution of a wide variety of conservation law systems. They are extensively used in fluid mechanics and have seen great popularity in other areas such as meteorology, electromagnetics, models of biological processes, semiconductor device simulation, financial options pricing and many other engineering areas governed by conservative systems that can be written in integral control volume form.

This lecture reviews elements of the foundation and analysis of modern finite volume methods. Throughout this lecture, specific attention is given to scalar nonlinear hyperbolic conservation laws and the development of high order accurate schemes for discretizing them. When compared to other discretization methods such as finite elements or finite differences, the primary attraction of finite volume methods is numerical robustness through the obtention of discrete maximum (minimum) principles, applicability on very general unstructured meshes, and the intrinsic local conservation properties of the resulting schemes. A key tool in the design and analysis of finite volume schemes suitable for non-oscillatory discontinuity capturing is discrete maximum principle analysis. Consequently, the emphasis of this lecture concerns maximum principle analysis and a discussion of how finite volume schemes are constructed based on this concept.

1.1 Finite volume (FV) methods for nonlinear conservation laws

This lecture reviews selected elements of the foundation and analysis of modern finite volume methods. Finite volume methods are a class of discretization schemes that have proven highly successful in approximating the solution of a wide variety of conservation law systems such as those occurring in fluid mechanics, meteorology, electromagnetics, models of biological processes, semi-conductor device simulation, financial options pricing and many other engineering areas governed by conservative systems that can be written in integral control volume form. In the remainder of this lecture, basic design principles used in the construction of upwind finite volume methods are presented. In a few representative cases, theorems are given with more or less complete proofs to illustrate the

typical mathematical tools and techniques employed in finite volume algorithm design and analysis.

We begin by considering the scalar Cauchy initial value problem

$$\partial_t u + \nabla \cdot f(u) = 0 \qquad \text{in } \mathbb{R}^d \times \mathbb{R}^+ \ , \tag{1a}$$

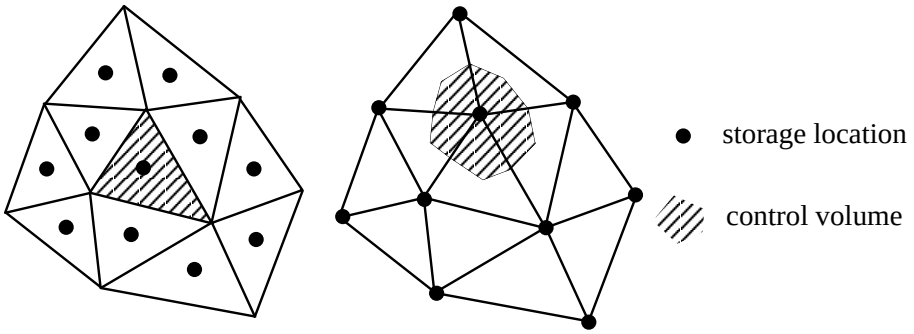
$$u(x,0) = u_0(x) \quad \text{in } \mathbb{R}^d. \tag{1b}$$

Here $u(x,t) : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}$ denotes the dependent solution variable, $f(u) \in C^1(\mathbb{R})$ denotes the flux function, and $u_0(x) : \mathbb{R}^d \to \mathbb{R}$ the initial data. In the finite volume method, the computational domain, $\Omega \subset \mathbb{R}^d$, is first tessellated into a collection of non overlapping control volumes that completely cover the domain. Let $\mathcal{T}$ denote a tessellation of the domain Ω with control volumes $T \in \mathcal{T}$ such that $\cup_{T \in \mathcal{T}} \overline{T} = \overline{\Omega}$. Let h_T denote a length scale associated with each control volume T, e.g. $h_T \equiv \text{diam(T)}$. For two distinct control volumes T_i and T_j in $\mathcal{T}$, the intersection is either an oriented edge (2-D) or face (3-D) e_{ij} with oriented normal ν_{ij} or else a set of measure at most $d-2$. In each control volume, an *integral conservation law* statement is then imposed.

Definition 1.1.1 (Integral conservation law) *An integral conservation law asserts that the rate of change of the total amount of a substance with density u in a fixed control volume T is equal to the flux f of the substance through the boundary ∂T*

$$\frac{\partial}{\partial t} \int_T u \, dx + \int_{\partial T} f(u) \cdot d\nu = 0 \ . \tag{2}$$

This integral conservation law statement is readily obtained upon spatial integration of the divergence equation (1a) in the region T and application of the divergence theorem. The choice of control volume tessellation is flexible in the finite volume method. For example,



a. Cell-centered b. Vertex-centered

Figure 1.1: Control volume variants used in the finite volume method: (a) cell-centered and (b) vertex-centered control volume tessellation.

Fig. 1.1 depicts a 2-D triangle complex and two typical control volume tessellations (among many others) used in the finite volume method. In the *cell-centered* finite volume method shown in Fig. 1.1a, the triangles themselves serve as control volumes with solution unknowns stored on a per triangle basis. In the *vertex-centered* finite volume method shown in Fig. 1.1b, control volumes are formed as a geometric dual to the triangle complex and solution unknowns stored on a per triangulation vertex basis.

1.1.1 Godunov finite volume discretizations

Fundamental to finite volume methods is the introduction of the control volume cell average for each $T_j \in \mathcal{T}$

$$u_j \equiv \frac{1}{|T_j|} \int_{T_j} u \; dx \;\; . \tag{3}$$

For stationary meshes, the finite volume method can be interpreted as producing an evolution equation for cell averages

$$\frac{\partial}{\partial t} \int_{T_j} u \, dx = |T_j| \frac{\partial}{\partial t} u_j \;\; . \tag{4}$$

Godunov [God59] pursued this interpretation in the discretization of the gas dynamic equations by assuming piecewise constant solution representations in each control volume with value equal to the cell average. However, the use of piecewise constant representations renders the numerical solution multivalued at control volume interfaces thereby making the calculation of a single solution flux at these interfaces ambiguous. The second aspect of Godunov's scheme and subsequent variants was the idea of supplanting the true flux at interfaces by a *numerical flux function*, $g(u,v) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$, a Lipschitz continuous function of the two interface states u and v. A single unique numerical flux was then calculated from an exact or approximate local solution of the Riemann problem in gas dynamics posed at these interfaces. Figure 1.2 depicts a representative 1-D solution profile in Godunov's method. For a given control volume $T_j = [x_{j-1/2}, x_{j+1/2}]$, Riemann problems are solved at each interface $x_{j\pm 1/2}$. For example, at the interface $x_{j+1/2}$ the Riemann problem counterpart of (1a-1b)

$$\partial_\tau w_{j+1/2}(\xi,\tau) + \partial_\xi f(w_{j+1/2}(\xi,\tau)) = 0 \quad \text{ in } \mathbb{R} \times \mathbb{R}^+$$

for $w_{j+1/2}(\xi,\tau) \in \mathbb{R}$ with initial data

$$w_{j+1/2}(\xi,0) = \begin{cases} u_j & \text{if } \xi < 0 \\ u_{j+1} & \text{if } \xi > 0 \end{cases}$$

is solved either exactly or approximately. From this local solution, a single unique numerical flux at $x_{j+1/2}$ is computed from $g(u_j, u_{j+1}) = f(w_{j+1/2}(0, \mathbb{R}^+))$. This construction utilizes the fact that the solution of the Riemann problem at $\xi = 0$ is a constant for all time $\tau > 0$.

In higher space dimensions, the flux integral appearing in (2) is similarly approximated by

$$\int_{\partial T_j} f(u) \cdot d\nu \approx \sum_{\forall e_{jk} \in \partial T_j} g_{jk}(u_j, u_k) \tag{5}$$

where the numerical flux is assumed to satisfy the properties:

- (Conservation) This property insures that fluxes from adjacent control volumes sharing a mutual interface exactly cancel when summed. This is achieved if the numerical flux satisfies the identity

$$g_{jk}(u,v) = -g_{kj}(v,u) \;\; . \tag{6a}$$

- (Consistency) Consistency is obtained if the numerical flux with identical state arguments reduces to the true flux of that same state, i.e.

$$g_{jk}(u,u) = \int_{e_{jk}} f(u) \cdot d\nu \;\; . \tag{6b}$$

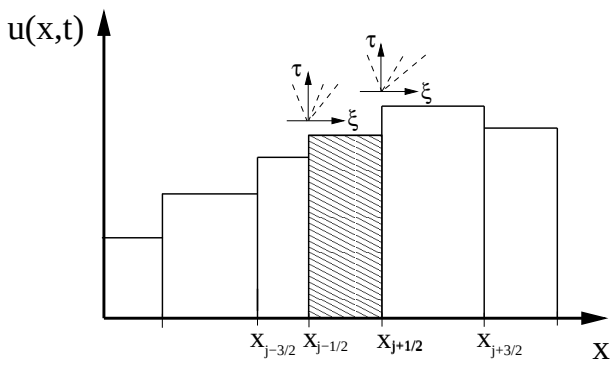


Figure 1.2: 1-D control volume, $T_j = [x_{j-1/2}, x_{j+1/2}]$, depicting Godunov's interface Riemann problems, $w_{j\pm1/2}(\xi, \tau)$, from piecewise constant interface states.

Combining (4) and (5) yields perhaps the simplest finite volume scheme in semi-discrete form. Let V_h^0 denote the space of piecewise constants, i.e.

$$V_h^0 = \{v \mid v|_T \in \chi(T),\ \forall T \in \mathcal{T}\} \tag{7}$$

with $\chi(T)$ a characteristic function in the control volume T.

Definition 1.1.2 (Semi-discrete finite volume method) *The semi-discrete finite volume approximation of (1a-1b) utilizing continuous in time solution representation, $t \in [0, \tau]$, and piecewise constant solution representation in space, $u_h(t) \in V_h^0$, such that*

$$u_j(t) = \frac{1}{|T_j|} \int_{T_j} u_h(x,t)\ dx$$

with initial data

$$u_j(0) = \frac{1}{|T_j|} \int_{T_j} u_0(x)\ dx$$

and numerical flux function $g_{jk}(u_j, u_k)$ is given by the following system of ordinary differential equations

$$\frac{d}{dt} u_j + \frac{1}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} g_{jk}(u_j, u_k) = 0 \ , \quad \forall T_j \in \mathcal{T} \ . \tag{8}$$

This system of ordinary differential equations can be marched forward using a variety of explicit and implicit time integration methods. Let u_j^n denote a numerical approximation of the cell average solution in the control volume T_j at time $t^n \equiv n\Delta t$. A particularly simple time integration method is the forward Euler scheme

$$\frac{d}{dt} u_j \approx \frac{u_j^{n+1} - u_j^n}{\Delta t}$$

thus producing the fully-discrete finite volume form.

Definition 1.1.3 (Fully-discrete finite volume method) *The fully-discrete finite volume approximation of (1a-1b) for the time slab interval $[t^n, t^n + \Delta t]$ utilizing the piecewise constant solution representation in space, $u_h^n \in V_h^0$, such that*

$$u_j^n = \frac{1}{|T_j|} \int_{T_j} u_h^n(x)\ dx$$

with initial data

$$u_j^0 = \frac{1}{|T_j|}\int_{T_j} u_0(x)\ dx$$

and numerical flux function $g_{jk}(u_j^n, u_k^n)$ *is given by the following fully-discrete system*

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j} g_{jk}(u_j^n, u_k^n)\ \ , \quad \forall T_j \in \mathcal{T}\ \ . \tag{9}$$

Monotone schemes

Unfortunately, the numerical flux conditions (6a) and (6b) are insufficient to guarantee convergence to entropy satisfying weak solutions and additional numerical flux restrictions are necessary. Two classes of numerical fluxes that guarantee such convergence for piecewise constant numerical solution data are *monotone fluxes* and *E-fluxes*. Specifically, Harten [HHL76] provides the following result concerning convergence of the fully-discrete one-dimensional scheme to weak solutions which was later generalized to (9) and irregular grids by Cockburn et al. [CCL94].

Theorem 1.1.4 (Monotone schemes and weak solutions,[HHL76]) *Consider a 1-D finite volume discretization of (1a-1b) with* $2k+1$ *stencil on a uniformly spaced mesh in both time and space with corresponding mesh spacing parameters* Δt *and* Δx

$$\begin{aligned} u_j^{n+1} &= H_j(u_{j+k}, \ldots, u_j, \ldots, u_{j-k}) \\ &= u_j^n - \frac{\Delta t}{\Delta x}(g_{j+1/2} - g_{j-1/2}) \end{aligned} \tag{10}$$

and consistent numerical flux of the form

$$g_{j+1/2} = g(u_{j+k}, \ldots, u_{j+1}, u_j, \ldots, u_{j-k+1})$$

that is monotone in the sense

$$\frac{\partial H_j}{\partial u_{j+l}} \geq 0\ \ , \quad \forall\ |l| \leq k\ \ . \tag{11}$$

Then as Δt *and* Δx *tend to zero with* $\Delta t/\Delta x =$ constant, u_j^n *converges boundedly almost everywhere to* $u(x,t)$*, an entropy satisfying weak solution of (1a-1b).*

The monotonicity condition (11) motivates the introduction of Lipschitz continuous monotone fluxes satisfying

$$\frac{\partial g_{j+1/2}}{\partial u_l} \geq 0 \ \text{ if } \ l = j \tag{12a}$$

$$\frac{\partial g_{j+1/2}}{\partial u_l} \leq 0 \ \text{ if } \ l \neq j \tag{12b}$$

together with a CFL (Courant-Friedrichs-Levy) like condition

$$1 - \frac{\Delta t}{\Delta x}\left(\frac{\partial g_{j+1/2}}{\partial u_j} - \frac{\partial g_{j-1/2}}{\partial u_j}\right) \geq 0$$

so that (11) is satisfied. Some examples of monotone fluxes for (1a) include

- (Godunov flux)

$$g^{\mathrm{G}}_{j+1/2} = \begin{cases} \min\limits_{u\in[u_j,u_{j+1}]} f(u) & \text{if } u_j < u_{j+1} \\ \max\limits_{u\in[u_j,u_{j+1}]} f(u) & \text{if } u_j > u_{j+1} \end{cases} \tag{13}$$

- (Lax-Friedrichs flux)

$$g^{\mathrm{LF}}_{j+1/2} = \frac{1}{2}\left(f(u_j) + f(u_{j+1})\right) - \frac{1}{2}\sup_{u\in[u_j,u_{j+1}]} |f'(u)|\,(u_{j+1} - u_j) \ . \tag{14}$$

E-flux schemes

A more general class of numerical fluxes was introduced and analyzed by Osher [Osh84] that still guarantees convergence to weak entropy solutions when used in (9) or (10). These fluxes are called E-fluxes, $g_{j+1/2} = g^{\mathrm{E}}(u_{j+k}, \ldots, u_{j+1}, u_j, \ldots, u_{j-k+1})$, due to the relationship to Olienick's well-known E-condition which characterizes entropy satisfying discontinuities. E-fluxes satisfy the inequality

$$\frac{g^{\mathrm{E}}_{j+1/2} - f(u)}{u_{j+1} - u_j} \le 0 \ , \quad \forall u \in [u_j, u_{j+1}] \ . \tag{15}$$

E-fluxes can be characterized by their relationship to Godunov's flux. Specifically, E-fluxes are precisely those fluxes such that

$$g^{\mathrm{E}}_{j+1/2} \le g^{\mathrm{G}}_{j+1/2} \quad \text{if } u_{j+1} < u_j \tag{16a}$$

$$g^{\mathrm{E}}_{j+1/2} \ge g^{\mathrm{G}}_{j+1/2} \quad \text{if } u_{j+1} > u_j \ . \tag{16b}$$

Viewed another way, note that any numerical flux can be written in the form

$$g_{j+1/2} = \frac{1}{2}\left(f(u_j) + f(u_{j+1})\right) - \frac{1}{2}Q(u_{j+1} - u_j) \tag{17}$$

where $Q(\cdot)$ denotes a viscosity for the scheme. When written in this form, E-fluxes are those fluxes that contribute at least as much viscosity as Godunov's flux, i.e.

$$Q^{\mathrm{G}}_{j+1/2} \le Q_{j+1/2} \ . \tag{18}$$

The most prominent E-flux is the Enquist-Osher flux

$$g^{\mathrm{EO}}_{j+1/2} = \frac{1}{2}\left(f(u_j) + f(u_{j+1})\right) - \frac{1}{2}\int_{u_j}^{u_{j+1}} |f'(s)|\,ds \ , \tag{19}$$

although other fluxes such as certain forms of Roe's flux with entropy fix fall into this category. From (16a-16b), the monotone fluxes of Godunov $g^{\mathrm{G}}_{j+1/2}$ and Lax-Friedrichs $g^{\mathrm{LF}}_{j+1/2}$ are also E-fluxes.

1.1.2 Discrete maximum principles and stability

A compelling motivation for the use of monotone and E-fluxes in the finite volume schemes (8) and (9) is the obtention of discrete maximum principles in the resulting numerical solution of nonlinear conservation laws (1a). A standard analysis technique is to first construct local discrete maximum principles which can then be applied successively to obtain global maximum principles and stability results.

The first result concerns the boundedness of local extrema in time for semi-discrete finite volume schemes that can be written in nonnegative coefficient form.

Lemma 1.1.5 (LED Property) *The semi-discrete scheme for each $T_j \in \mathcal{T}$*

$$\frac{du_j}{dt} = \frac{1}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} C_{jk}(u_h)(u_k - u_j), \tag{20}$$

is Local Extremum Diminishing (LED), i.e. local maxima are non-increasing and local minima are nondecreasing, if

$$C_{jk}(u_h) \geq 0 \ , \ \forall e_{jk} \in \partial T_j \ . \tag{21}$$

Proof: Assume a discrete maximum located in cell T_j so that $u_k - u_j < 0$. For $C_{jk}(u_h) \geq 0$ this implies that

$$\frac{du_j}{dt} \leq 0$$

so that this maximum decreases with time. Repeating the argument for a discrete minimum yields the stated lemma. ∎

Rewriting the semi-discrete finite volume scheme (8) in terms of monotone fluxes or E-fluxes

$$\begin{aligned} \frac{du_j}{\partial t} &= -\frac{1}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} \frac{g_{jk}(u_j, u_k) - f(u_j) \cdot \nu_{jk}}{u_k - u_j}(u_k - u_j) \\ &= -\frac{1}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} \frac{\partial g_{jk}}{\partial u_k}(u_j, \tilde{u}_{jk})(u_k - u_j) \end{aligned} \tag{22}$$

for appropriately chosen $\tilde{u}_{jk} \in [u_j, u_k]$ together with the monotone flux conditions (12a-12b) or the E-flux condition (15) reveals that monotone flux and E-flux finite volume schemes (8) are LED. In order to obtain local space-time maximum principle results for the fully-discrete discretization (9) requires the introduction of an additional CFL-like condition for non-negativity of coefficients in space-time.

Lemma 1.1.6 (Local space-time discrete maximum principle) *The fully-discrete scheme for the time slab increment $[t^n, t^{n+1}]$ and each $T_j \in \mathcal{T}$*

$$u_j^{n+1} = u_j^n + \frac{\Delta t}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} C_{jk}(u_h^n)(u_k^n - u_j^n) \tag{23}$$

exhibits a local space-time discrete maximum principle

$$\min_{\forall e_{jk} \in \partial T_j} (u_k^n, u_j^n) \leq u_j^{n+1} \leq \max_{e_{jk} \in \partial T_j} (u_k^n, u_j^n) \tag{24}$$

if

$$C_{jk}(u_h^n) \geq 0 \ , \quad \forall e_{jk} \in \partial T_j \tag{25}$$

and satisfies the CFL-like condition

$$1 - \frac{\Delta t}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} C_{jk}(u_h^n) \geq 0 \ . \tag{26}$$

Again noting that the flux terms in the fully-discrete finite volume scheme (9) can be written in the form (22) reveals that the monotone flux conditions (12a-12b) or the E-flux condition (15) together with a local CFL-like condition obtained from (26) imply a local space-time discrete maximum principle. By successive application of Lemma 1.1.6, a global L^∞-stability bound is obtained for the scalar initial value problem (1a-1b) in terms of initial data $u_0(x)$.

Theorem 1.1.7 (L^∞-stability) *Assume a fully-discrete finite volume scheme (9) for the scalar initial value problem (1a-1b) utilizing monotone fluxes or E-fluxes that satisfy a local CFL-like condition as given in Lemma 1.1.6 for each time slab increment $[t^n, t^{n+1}]$. Under these conditions, the finite volume scheme is L^∞-stable and the following estimate holds:*

$$\inf_{x\in\mathbb{R}^d} u_0(x) \le u_j^n \le \sup_{x\in\mathbb{R}^d} u_0(x), \quad \textit{for all } (T_j, t^n) \in \mathcal{T} \times [0,\tau]. \tag{27}$$

Consider now steady-state solutions, $u^{n+1} = u^n = u^*$, using monotone flux or E-flux schemes in the fully-discrete finite volume scheme (9). At steady state, non-negativity of the coefficients $C(u_h)$ in (23) implies a discrete maximum principle.

Lemma 1.1.8 (Local discrete maximum principle in space) *The fully-discrete scheme (23) exhibits a local discrete maximum principle at steady state, u_h^*, for each $T_j \in \mathcal{T}$*

$$\min_{\forall e_{jk}\in\partial T_j} u_k^* \le u_j^* \le \max_{\forall e_{jk}\in\partial T_j} u_k^* \tag{28}$$

if

$$C_{jk}(u_h^*) \ge 0 \ , \quad \forall e_{jk} \in \partial T_j \ .$$

Once again by virtue of (12a-12b) and (15), the conditions for a local discrete maximum principle at steady state are fulfilled by monotone flux and E-flux finite volume schemes (9). Global maximum principles for characteristic boundary valued problems are readily obtained by successive application of the local maximum principle result.

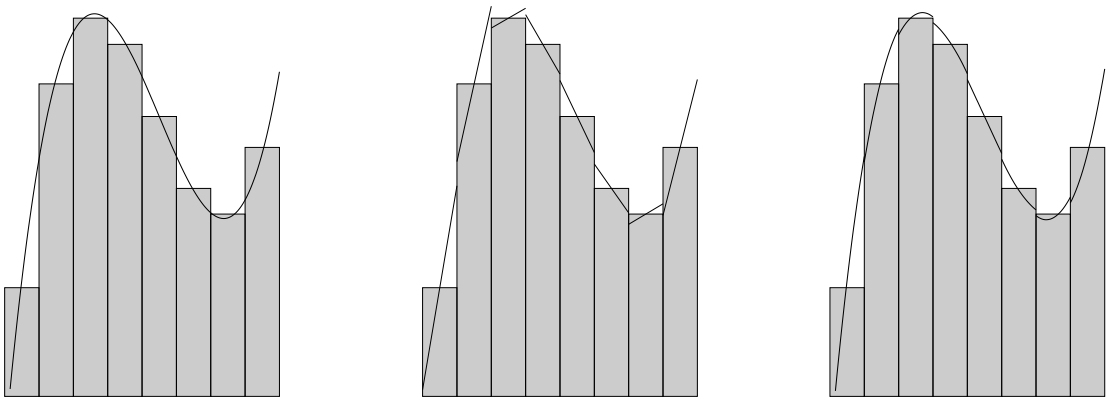
The local maximum principles given in (24) and (28) preclude the introduction of spurious extrema and $\mathcal{O}(1)$ Gibbs-like oscillations that occur near solution discontinuities computed using many numerical methods (even in the presence of grid refinement). For this reason, discrete maximum principles of this type are a highly sought after design principle in the development of numerical schemes for nonlinear conservation laws.

1.2 Higher order accurate FV generalizations

Although an $\mathcal{O}(h^{1/2})$ L_1-norm error bound for the monotone and E-flux schemes of Sect. 1.1 is known to be sharp ([Pet91]), an $\mathcal{O}(h)$ solution error is routinely observed in numerical experiments with convex flux functions. Even so, first order accurate schemes are generally considered too inaccurate for most quantitative calculations unless the mesh spacing is made excessively small thus rendering the schemes inefficient. Godunov [God59] has shown that all *linear* schemes that preserve solution monotonicity are at most first order accurate. The low order accuracy of these monotonicity preserving linear schemes has motivated the development of higher order accurate schemes with the important distinction that these new schemes utilize essential *nonlinearity* so that monotone resolution of discontinuities and high order accuracy away from discontinuities are simultaneously attained.

1.2.1 Higher order accurate FV schemes in 1-D

A significant step forward in the generalization of Godunov's finite volume method to higher order accuracy is due to van Leer [vL79]. In the context of Lagrangian hydrodynamics with Eulerian remapping, van Leer generalized Godunov's method by employing linear solution *reconstruction* in each cell (see Fig. 1.3b). Let N denote the number of con-



a. Cell averaging of data b. Linear reconstruction c. Quadratic reconstruction

Figure 1.3: Piecewise polynomial approximation used in the finite volume method: (a) cell averaging of analytic data, (b) piecewise linear reconstruction from cell averages and (c) piecewise quadratic reconstruction from cell averages.

trol volume cells in space so that the j-th cell extends over the interval $T_j = [x_{j-1/2}, x_{j+1/2}]$ with length Δx_j such that $\cup_{1\le j\le N} T_j = [0,1]$ with $T_i \cap T_j = \emptyset, i \neq j$. In a purely Eulerian setting, the higher order accurate schemes of van Leer are of the form

$$\frac{du_j}{dt} + \frac{1}{\Delta x_j}(g(u^-_{j+1/2}, u^+_{j+1/2}) - g(u^-_{j-1/2}, u^+_{j-1/2})) = 0$$

where $g(u,v)$ is a numerical flux function utilizing states $u^-_{j\pm1/2}$ and $u^+_{j\pm1/2}$ obtained from evaluation of the linear solution reconstructions from the left and right cells surrounding the interfaces $x_{j\pm1/2}$, By altering the slope of the linear reconstruction in cells, non-oscillatory resolution of discontinuities can be obtained. Note that although obtaining the exact solution of the scalar nonlinear conservation law with linear initial data is a formidable task, the solution at each cell interface location for small enough time is the same as the solution of the Riemann problem with piecewise constant data equal to the linear solution approximation evaluated at the same interface location. Consequently, the numerical flux functions used in Sect. 1.1 can be once again used in the generalized schemes of van Leer. This single observation greatly simplifies the construction of higher order accurate generalizations of Godunov's method. The ideas of van Leer have been extended to quadratic approximations in each cell (see Fig. 1.3c) by Colella and Woodward [CP84]. Although these generalizations of Godunov's method and further generalizations given later can be interpreted in 1-D as finite difference discretizations, concepts originally developed in 1-D such as solution monotonicity, positive coefficient discretization and discrete maximum principle analysis are often used in the design of finite volume methods in multiple space dimensions and on unstructured meshes where finite difference discretization is problematic.

TVD schemes

In considering the scalar nonlinear conservation law (1a-1b), Lax [Lax73] made the following basic observation:

> *"the total increasing and decreasing variations of a differentiable solution between any pair of characteristics are conserved".*

Furthermore, in the presence of shock wave discontinuities, information is lost and the total variation *decreases*. For the 1-D nonlinear conservation law with compactly supported (or periodic) solution data $u(x,t)$, integrating along the constant time spatial coordinate at times t_1 and t_2 yields

$$\int_{-\infty}^{\infty} |du(x,t_2)| \leq \int_{-\infty}^{\infty} |du(x,t_1)|, \qquad t_2 \geq t_1 \ . \tag{29}$$

This motivated Harten [Har83] to consider the discrete total variation

$$\mathrm{TV}(u_h) \equiv \sum_j |\Delta_{j+1/2} u_h| \ , \quad \Delta_{j+1/2} u_h \equiv u_{j+1} - u_j$$

and the discrete total variation non-increasing (TVNI) bound counterpart to (29)

$$\mathrm{TV}(u_h^{n+1}) \leq \mathrm{TV}(u_h^n) \tag{30}$$

in the design of numerical discretizations for nonlinear conservation laws. A number of simple results relating TVNI schemes and monotone schemes follow from simple analysis.

Theorem 1.2.1 (TVNI and monotone scheme properties, [Har83]) *(i) Monotone schemes are TVNI. (ii) TVNI schemes are monotonicity preserving, i.e. the number of solution extrema is preserved in time.*

Property (i) follows from the L_1 contraction property of monotone schemes [CM80]. Property (ii) is readily shown using a proof by contradiction by assuming a TVNI scheme with monotone initial data that produces new solution data at a later time with interior solution extrema present. Using the notion of discrete total variation, Harten [Har83] then constructed sufficient algebraic conditions for achieving the TVNI inequality (30).

Theorem 1.2.2 (Harten's explicit TVD criteria, [Har83]) *The fully discrete explicit 1-D scheme*

$$u_j^{n+1} = u_j^n + \Delta t \, \left(C_{j+1/2}(u_h^n) \, \Delta_{j+1/2} u_h^n + D_{j+1/2}(u_h^n) \, \Delta_{j-1/2} u_h^n \right) , \quad j = 1, \ldots, N \tag{31}$$

is total variation non-increasing if for each j

$$C_{j+1/2} \geq 0 \ , \tag{32a}$$

$$D_{j+1/2} \leq 0 \ , \tag{32b}$$

$$1 - \Delta t \, \left(C_{j-1/2} - D_{j+1/2} \right) \geq 0 \ . \tag{32c}$$

Note that although the inequality constraints (32a-32c) in Theorem 1.2.2 insure that the total variation is non-increasing, these conditions are often referred to as total variation diminishing (TVD) conditions. Also note that inequality (32c) implies a CFL-like time step restriction that may be more restrictive than the time step required for stability of the numerical method. The TVD conditions are easily generalized to wider support stencils written in incremental form, see for example [JL86] and their corrected result in [JL87].

Theorem 1.2.3 (Generalized explicit TVD criteria, [JL86]) *The fully discrete explicit 1-D scheme*

$$u_j^{n+1} = u_j^n + \Delta t \sum_{l=-k}^{k-1} C_{j+1/2}^{(l)}(u_h^n)\Delta_{j+l+1/2}u_h^n, \quad j = 1, \ldots, N \tag{33}$$

with stencil width parameter k is total variation non-increasing if for each j

$$C_{j+1/2}^{(k-1)} \geq 0 \ , \tag{34a}$$

$$C_{j+1/2}^{(-k)} \leq 0 \ , \tag{34b}$$

$$C_{j+1/2}^{(l-1)} - C_{j-1/2}^{(l)} \geq 0 \ , \quad -k+1 \leq l \leq k-1, \ l \neq 0 \ , \tag{34c}$$

$$1 - \Delta t \left(C_{j-1/2}^{(0)} - C_{j+1/2}^{(-1)} \right) \geq 0 \ . \tag{34d}$$

The extension to implicit methods follows immediately upon rewriting the implicit scheme in terms of the solution spatial increments $\Delta_{j+l+1/2}u_h$ and imposing sufficient algebraic conditions such that the implicit matrix acting on spatial increments is an M-matrix and thus has a nonnegative inverse.

Theorem 1.2.4 (Generalized implicit TVD criteria) *The fully discrete implicit 1-D scheme*

$$u_j^{n+1} - \Delta t \sum_{l=-k}^{k-1} C_{j+1/2}^{(l)}(u_h^{n+1})\Delta_{j+l+1/2}u_h^{n+1} = u_j^n, \quad j = 1, \ldots, N \tag{35}$$

with stencil width parameter k is total variation non-increasing if for each j

$$C_{j+1/2}^{(k-1)} \geq 0 \ , \tag{36a}$$

$$C_{j+1/2}^{(-k)} \leq 0 \ , \tag{36b}$$

$$C_{j+1/2}^{(l-1)} - C_{j-1/2}^{(l)} \geq 0 \ , \quad -k+1 \leq l \leq k-1, \ l \neq 0 \ . \tag{36c}$$

Theorems 1.2.3 and 1.2.4 provide sufficient conditions for non-increasing total variation of explicit (33) or implicit (35) numerical schemes written in incremental form. These incremental forms do not imply *discrete conservation* unless additional constraints are imposed on the discretizations. A sufficient condition for discrete conservation of the discretizations (33) and (35) is that these discretizations can be written in a finite volume flux balance form

$$g_{j+1/2} - g_{j-1/2} = \sum_{l=-k}^{k-1} C_{j+1/2}^{(l)}(u_h)\Delta_{j+l+1/2}u_h$$

where $g_{j\pm1/2}$ are the usual numerical flux functions. Section 1.2.1 provides an example of how the discrete TVD conditions and discrete conservation can be simultaneously achieved. A more comprehensive overview of finite volume numerical methods based on TVD constructions can be found the books by Godlewski and Raviart [GR91] and LeVeque [LeV02].

MUSCL schemes

A general family of TVD discretizations with 5-point stencil is the Monotone Upstream-centered Scheme for Conservation Laws (MUSCL) discretization of van Leer [vL79, vL85]. MUSCL schemes utilize a κ-parameter family of interpolation formulas with *limiter function* $\Psi(R) : \mathbb{R} \mapsto \mathbb{R}$

$$\begin{aligned} u^-_{j+1/2} &= u_j + \frac{1+\kappa}{4}\Psi(R_i)\Delta_{j-1/2}u_h \quad + \frac{1-\kappa}{4}\Psi(1/R_j)\Delta_{j+1/2}u_h \\ u^+_{j-1/2} &= u_j - \frac{1+\kappa}{4}\Psi(1/R_j)\Delta_{j+1/2}u_h - \frac{1-\kappa}{4}\Psi(R_j)\Delta_{j-1/2}u_h \end{aligned} \tag{37}$$

where R_j is a ratio of successive solution increments

$$R_j \equiv \frac{\Delta_{j+1/2}u_h}{\Delta_{j-1/2}u_h} \ . \tag{38}$$

The technique of incorporating limiter functions to obtain non-oscillatory resolution of discontinuities and steep gradients dates back to Boris and Book [BB73]. For convenience, the interpolation formulas (37) have been written for a uniformly spaced mesh although the extension to irregular mesh spacing is straightforward. The unlimited form of this interpolation is obtained by setting $\Psi(R) = 1$. In this unlimited case, the truncation error for the conservation law divergence in (1a) is given by

$$\text{Truncation Error} = -\frac{\left(\kappa - \frac{1}{3}\right)}{4}(\Delta x)^2 \frac{\partial^3}{\partial x^3} f(u) \ .$$

This equation reveals that for $\kappa = 1/3$, the 1-D MUSCL formula yields an overall spatial discretization with $\mathcal{O}(\Delta x^3)$ truncation error. Using the MUSCL interpolation formulas given in (37), sufficient conditions for the discrete TVD property are easily obtained.

Theorem 1.2.5 (MUSCL TVD scheme) *The fully discrete 1-D scheme*

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x_j}\left(g^n_{j+1/2} - g^n_{j-1/2}\right) \ , \quad j = 1, \ldots, N$$

with monotone Lipschitz continuous numerical flux function

$$g_{j+1/2} = g(u^-_{j+1/2}, u^+_{j+1/2})$$

utilizing the κ-parameter family of MUSCL interpolation formulas (37) and (38) is total variation non-increasing if there exists a $\Psi(R)$ such that $\forall R \in \mathbb{R}$

$$0 \le \Psi(R) \le \frac{3-\kappa}{1-\kappa} - (1+\alpha)\frac{1+\kappa}{1-\kappa} \tag{39a}$$

and

$$0 \le \frac{\Psi(R)}{R} \le 2 + \alpha \tag{39b}$$

with $\alpha \in [-2, 2\,(1-\kappa)/(1+\kappa)]$ *under the time step restriction*

$$1 - \frac{\Delta t}{\Delta x_j} \frac{2-(2+\alpha)\kappa}{1-\kappa} \left|\frac{\partial g}{\partial u}\right|_j^{\max} \geq 0$$

where

$$\left|\frac{\partial g}{\partial u}\right|_j^{\max} \equiv \sup_{\substack{\widetilde{u}\in[u^-_{j-1/2}, u^-_{j+1/2}] \\ \widetilde{\widetilde{u}}\in[u^+_{j-1/2}, u^+_{j+1/2}]}} \left(\frac{\partial g}{\partial u^-}(\widetilde{u}, u^+_{j+1/2}) - \frac{\partial g}{\partial u^+}(u^-_{j-1/2}, \widetilde{\widetilde{u}}) \right) \quad .$$

Proof: The first step is to introduce the mean value flux function linearization states, $\widetilde{u} \in [u^-_{j-1/2}, u^-_{j+1/2}]$ and $\widetilde{\widetilde{u}} \in [u^+_{j-1/2}, u^+_{j+1/2}]$, such that

$$\begin{aligned} u_j^{n+1} &= u_j^n - \frac{\Delta t}{\Delta x_j} \Big(\underbrace{\frac{\partial g}{\partial u^-}}_{(+)}(\tilde{u}, u^+_{j+1/2,k})\,(u^-_{j+1/2,k} - u^-_{j-1/2,k}) \\ &\qquad + \underbrace{\frac{\partial g}{\partial u^+}}_{(-)}(u^-_{j-1/2,k}, \tilde{\tilde{u}})\,(u^+_{j+1/2,k} - u^+_{j-1/2,k})\Big) \end{aligned}$$

The assumption of a monotone flux function fixes the signs of $\partial g/\partial u^{\pm}$. With rearrangement, the remaining terms simplify to

$$u^-_{j+1/2} - u^-_{j-1/2} = \left(1 + \frac{1+\kappa}{4}\left(\Psi(R_j) - \frac{\Psi(R_{j-1})}{R_{j-1}}\right) + \frac{1-\kappa}{4}\left(\Psi(1/R_j)R_j - \Psi(1/R_{j-1})\right)\right)\Delta_{j-1/2}u,$$

and

$$u^+_{j+1/2} - u^+_{j-1/2} = \left(1 - \frac{1+\kappa}{4}\left(\Psi(1/R_{j+1})R_{j+1} - \Psi(1/R_j)\right) - \frac{1-\kappa}{4}\left(\Psi(R_{j+1}) - \frac{\Psi(R_j)}{R_j}\right)\right)\Delta_{j+1/2}u.$$

Appealing to Harten's TVD theorem 1.2.2, a sufficient condition for the TVD condition is that the limiter function $\Psi(\cdot)$ satisfies

$$1 + \frac{1+\kappa}{4}\left(\Psi(R_j) - \frac{\Psi(R_{j-1})}{R_{j-1}}\right) + \frac{1-\kappa}{4}\left(\Psi(1/R_j)R_j - \Psi(1/R_{j-1})\right) \geq 0$$

and

$$1 - \frac{1+\kappa}{4}\left(\Psi(1/R_{j+1})R_{j+1} - \Psi(1/R_j)\right) - \frac{1-\kappa}{4}\left(\Psi(R_{j+1}) - \frac{\Psi(R_j)}{R_j}\right) \geq 0$$

together with a CFL-like time step restriction. Both inequalities are simultaneously satisfied if $\forall Q, R, S, T \in \mathbb{R}$

$$1 + \frac{1+\kappa}{4}\left(\Psi(Q) - \Psi(R)/R\right) + \frac{1-\kappa}{4}\left(\Psi(S)/S - \Psi(T)\right) \geq 0 \quad . \tag{40}$$

Next assume a limiter function $\Psi(\cdot)$ satisfying the interval constraints

$$\begin{aligned} 0 &\leq \Psi(R) \leq f(\kappa, \alpha) \\ 0 &\leq \Psi(R)/R \leq 2 + \alpha \quad . \end{aligned}$$

Inserting these interval limits into the inequality (40) yields an explicit bound for $f(\kappa, \alpha)$ as indicated in the stated theorem. The time step restriction then follows directly from theorem 1.2.2. ∎

For accuracy considerations away from extrema, it is desirable that the unlimited form of the discretization is obtained. Consequently, the constraint $\Psi(1) = 1$ is also imposed upon the limiter function. This constraint together with the algebraic conditions (39a-b) are readily achieved using the well known *MinMod* limiter, Ψ^{MM}, with compression parameter β determined from the TVD analysis

$$\Psi^{\mathrm{MM}}(R) = \max(0, \min(R, \beta)) \ , \quad \beta \in [1, (3-\kappa)/(1-\kappa)] \ .$$

Table 1.1 summarizes the MUSCL scheme and maximum compression parameter for a number of familiar discretizations. Another limiter due to van Leer that meets the tech-

Table 1.1: Members of the MUSCL TVD family of schemes.

κ	Unlimited Scheme	$\beta_{\max}$	Truncation Error
1/3	Third-Order	4	0
-1	Fully Upwind	2	$\frac{1}{3}(\Delta x)^2 \frac{\partial^3}{\partial x^3} f(u)$
0	Fromm's	3	$\frac{1}{12}(\Delta x)^2 \frac{\partial^3}{\partial x^3} f(u)$
1/2	Low Truncation Error	5	$-\frac{1}{24}(\Delta x)^2 \frac{\partial^3}{\partial x^3} f(u)$

nical conditions of Theorem 1.2.5 and also satisfies $\Psi(1) = 1$ is given by

$$\Psi^{\mathrm{VL}}(R) = \frac{R + |R|}{1 + |R|} \ .$$

This limiter exhibits differentiability away from $R = 0$ which improves the iterative convergence to steady state for many algorithms. Numerous other limiter functions are considered and analyzed in Sweby [Swe84].

Unfortunately, TVD schemes locally degenerate to piecewise constant approximations at smooth extrema which locally degrades the accuracy. This is an unavoidable consequence of the strict TVD condition.

Theorem 1.2.6 (TVD critical point accuracy, [Osh84]) *The TVD discretizations (31), (33) and (35) all reduce to at most first order accuracy at non-sonic critical points, i.e. points u^* at which $f'(u^*) \neq 0$ and $u^*_x = 0$.*

ENO/WENO schemes

To circumvent the degradation in accuracy of TVD schemes at critical points, weaker constraints on the solution total variation were devised. To this end, Harten proposed the following abstract framework for generalized Godunov schemes in operator composition form (see [HOEC86, HOEC87, Har89])

$$u_h^{n+1} = A \cdot E(\tau) \cdot R_p^0(\cdot; u_h^n) \ . \tag{41}$$

In this equation, $u_h^n \in V_h^0$ denotes the global space of piecewise constant cell averages as defined in (7), $R_p^0(x)$ is a reconstruction operator which produces a cell-wise discontinuous

p-th order polynomial approximation from the given solution cell averages, $E(\tau)$ is the evolution operator for the PDE (including boundary conditions), and A is the cell averaging operator. Since A is a nonnegative operator and $E(\tau)$ represents exact evolution in the small, the control of solution oscillations and Gibbs-like phenomena is linked directly to oscillation properties of the reconstruction operator, $R_p^0(x)$. One has formally in one space dimension

$$\mathrm{TV}(u_h^{n+1}) = \mathrm{TV}(A \cdot E(\tau) \cdot R_p^0(\cdot; u_h^n)) \leq \mathrm{TV}(R_p^0(x; u_h^n))$$

so that the total variation depends entirely upon properties of the reconstruction operator $R_p^0(x; u_h^n)$. The requirements of high order accuracy for smooth solutions and discrete conservation give rise to the following additional design criterion for the reconstruction operator

- $R_p^0(x; u_h) = u(x) + e(x)\ \Delta x^{p+1} + O(\Delta x^{p+2})$ whenever u is smooth (42a)
- $A|_{T_j} R_p^0(x; u_h) = u_h|_{T_j} = u_j, \quad j = 1, \ldots, N$ to insure discrete conservation (42b)
- $\mathrm{TV}(R(x; u_h^n)) \leq \mathrm{TV}(u_h^n) + O(\Delta x^{p+1})$ an essentially non-oscillatory reconstruction. (42c)

Harten ([HOEC86, HOEC87, Har89]) then proposed a family of Essentially Non-Oscillatory Approximations (ENO) which allow $\mathcal{O}(h^p)$ violations of the discrete maximum principles outline above. The ENO construction was later simplified and improved by Shu and coworkers [JS96, Shu99] in the context of Weighted Essentially Non-Oscillatory Approximations (WENO) schemes. The construction of ENO and WENO schemes is somewhat technical and deferred to the lectures of Prof. Shu in this lecture series.

1.2.2 Higher order accurate FV schemes in multi-dimensions.

Although the one-dimensional TVD operators may be readily applied in multi-dimensions on a dimension-by-dimension basis, a result of Goodman and LeVeque [GV85] shows that TVD schemes in two or more space dimensions are only first order accurate.

Theorem 1.2.7 (Accuracy of TVD schemes in multi-dimensions) *Any two-dimensional finite volume scheme of the form*

$$u_{i,j}^{n+1} = u_{i,j}^n - \frac{\Delta t}{|T|_{i,j}}(g_{i+1/2,j}^n - g_{i-1/2,j}^n) - \frac{\Delta t}{|T|_{i,j}}(h_{i,j+1/2}^n - h_{i,j-1/2}^n)\ , \quad 1 \leq i \leq M,\ 1 \leq j \leq N$$

with Lipschitz continuous numerical fluxes for integers p, q, r, s

$$\begin{aligned} g_{i+1/2,j} &= g(u_{i-p,j-q}, \ldots, u_{i+r,j+s}), \\ h_{i,j+1/2} &= h(u_{i-p,j-q}, \ldots, u_{i+r,j+s}), \end{aligned}$$

that is total variation non-increasing in the sense

$$TV(u_h^{n+1}) \leq TV(u_h^n)$$

where

$$TV(u) \equiv \sum_{i,j} \left[\Delta y_{i+1/2,j}\, |u_{i+1,j} - u_{i,j}| + \Delta x_{i,j+1/2}\, |u_{i,j+1} - u_{i,j}|\right]$$

is at most first-order accurate.

Motivated by the negative results of Goodman and LeVeque, weaker conditions yielding solution monotonicity preservation have been developed from discrete maximum principle analysis. These alternative constructions have the positive attribute that they extend to unstructured meshes as well.

Positive coefficient schemes on structured meshes

Theorem 1.1.6 considers schemes of the form

$$u_j^{n+1} = u_j^n + \frac{\Delta t}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} C_{jk}(u_h^n)(u_k^n - u_j^n) \ , \quad \forall T_j \in \mathcal{T}$$

and provides a local space-time discrete maximum principle

$$\min_{\forall e_{jk} \in \partial T_j} (u_k^n, u_j^n) \leq u_j^{n+1} \leq \max_{\forall e_{jk} \in \partial T_j} (u_k^n, u_j^n)$$

$\forall T_j \in \mathcal{T}$ under a CFL-like condition on the time step parameter if all coefficients C_{jk} are nonnegative. Schemes of this type are often called *positive coefficient schemes* or more simply *positive schemes*. To circumvent the negative result of Theorem 1.2.7, Spekreijse [Spe87] developed a family of high order accurate positive coefficient schemes on two-dimensional structured $M \times N$ meshes. For purposes of positivity analysis, these schemes are written in incremental form on a $M \times N$ logically rectangular 2-D mesh

$$\begin{aligned} u_{i,j}^{n+1} = u_{i,j}^n + \Delta t \ (\quad & A_{i+1,j}^n \, (u_{i+1,j}^n - u_{i,j}^n) + B_{i,j+1}^n \, (u_{i,j+1}^n - u_{i,j}^n) \\ + \quad & C_{i-1,j}^n \, (u_{i-1,j}^n - u_{i,j}^n) + D_{i,j-1}^n \, (u_{i,j-1}^n - u_{i,j}^n) \quad) \ , \quad 1 \leq i \leq M, \ 1 \leq j \leq N \end{aligned} \tag{43}$$

where the coefficients are nonlinear functions of the solution

$$\begin{aligned} A_{i+1,j}^n &= A(..., u_{i-1,j}^n, u_{i,j}^n, u_{i+1,j}^n, ...) \\ B_{i,j+1}^n &= B(..., u_{i,j-1}^n, u_{i,j}^n, u_{i,j+1}^n, ...) \\ C_{i-1,j}^n &= C(..., u_{i-1,j}^n, u_{i,j}^n, u_{i+1,j}^n, ...) \\ D_{i,j-1}^n &= D(..., u_{i,j-1}^n, u_{i,j}^n, u_{i,j+1}^n, ...) \ . \end{aligned}$$

Once written in incremental form, the following theorem follows from standard positive coefficient maximum principle analysis.

Theorem 1.2.8 (Positive coefficient schemes in multi-dimensions) *The discretization (43) is a positive coefficient scheme if for each* $1 \leq i \leq M, \ 1 \leq j \leq N$ *and time slab increment* $[t^n, t^{n+1}]$

$$A_{i+1,j}^n \geq 0, \ B_{i,j+1}^n \geq 0, \ C_{i-1,j}^n \geq 0, \ D_{i,j-1}^n \geq 0, \tag{44}$$

and

$$1 - \Delta t \ (A_{i+1,j}^n + B_{i,j+1}^n + C_{i-1,j}^n + D_{i,j-1}^n) \geq 0 \tag{45}$$

with discrete space-time maximum principle

$$\min(u_{i,j}^n, u_{i-1,j}^n, u_{i+1,j}^n, u_{i,j-1}^n, u_{i,j+1}^n) \leq u_{i,j}^{n+1} \leq \max(u_{i,j}^n, u_{i-1,j}^n, u_{i+1,j}^n, u_{i,j-1}^n, u_{i,j+1}^n)$$

and discrete maximum principle at steady state

$$\min(u_{i-1,j}^*, u_{i+1,j}^*, u_{i,j-1}^*, u_{i,j+1}^*) \leq u_{i,j}^* \leq \max(u_{i-1,j}^*, u_{i+1,j}^*, u_{i,j-1}^*, u_{i,j+1}^*)$$

where u^* *denotes the numerical steady state.*

Using a procedure similar to that used in the development of MUSCL TVD schemes in 1-D, Spekrijse [Spe87] developed a family of monotonicity preserving MUSCL approximations in multi-dimensoins from the positivity conditions of Theorem 1.2.8.

Theorem 1.2.9 (MUSCL positive coefficient scheme,[Spe87]) *The fully discrete 2-D finite volume scheme*

$$u_{i,j}^{n+1} = u_{i,j}^n - \frac{\Delta t}{|T|_{i,j}}(g_{i+1/2,j}^n - g_{i-1/2,j}^n) - \frac{\Delta t}{|T|_{i,j}}(h_{i,j+1/2}^n - h_{i,j-1/2}^n) \ , \quad 1 \le i \le M, \ 1 \le j \le N$$

utilizing monotone Lipschitz continuous numerical flux functions

$$\begin{aligned} g_{i+1/2,j} &= g(u_{i+1/2,j}^-, u_{i+1/2,j}^+) \\ h_{i,j+1/2} &= h(u_{i,j+1/2}^-, u_{i,j+1/2}^+) \end{aligned}$$

and MUSCL extrapolation formulas

$$\begin{aligned} u_{i+1/2,j}^- &= u_{i,j} + \frac{1}{2}\Psi(R_{i,j})\,(u_{i,j} - u_{i-1,j}) \\ u_{i-1/2,j}^+ &= u_{i,j} - \frac{1}{2}\Psi(1/R_{i,j})\,(u_{i+1,j} - u_{i,j}) \\ u_{i,j+1/2}^- &= u_{i,j} + \frac{1}{2}\Psi(S_{i,j})\,(u_{i,j} - u_{i,j-1}) \\ u_{i,j-1/2}^+ &= u_{i,j} - \frac{1}{2}\Psi(1/S_{i,j})\,(u_{i,j+1} - u_{i,j}) \end{aligned}$$

where

$$R_{i,j} \equiv \frac{u_{i+1,j} - u_{i,j}}{u_{i,j} - u_{i-1,j}} \ , \quad S_{i,j} \equiv \frac{u_{i,j+1} - u_{i,j}}{u_{i,j} - u_{i,j-1}}$$

satisfies the local maximum principle properties of Lemma 1.2.8 and is second order accurate if the limiter $\Psi = \Psi(R)$ has the properties that there exist constants $\beta \in (0, \infty)$, $\alpha \in [-2, 0]$ such that $\forall R \in \mathbb{R}$

$$\alpha \le \Psi(R) \le \beta \ , \quad -\beta \le \frac{\Psi(R)}{R} \le 2 + \alpha \tag{46}$$

with the constraint $\Psi(1) = 1$ and the smoothness condition $\Psi(R) \in C^2$ near $R = 1$ together with a time step restriction for stability

$$1 - (1+\beta)\,\frac{\Delta t}{|T_{i,j}|}\,\left(\left|\frac{\partial g}{\partial u}\right|_{i,j}^{n,\max} + \left|\frac{\partial h}{\partial u}\right|_{i,j}^{n,\max} \right) \ge 0$$

where

$$\begin{aligned} \left|\frac{\partial g}{\partial u}\right|_{i,j}^{\max} &\equiv \sup_{\substack{\widetilde{u} \in [u_{i-1/2,j}^-, u_{i+1/2,j}^-] \\ \widetilde{\widetilde{u}} \in [u_{i-1/2,j}^+, u_{i+1/2,j}^+]}} \left(\frac{\partial g}{\partial u^-}(\widetilde{u}, u_{i+1/2,j}^+) - \frac{\partial g}{\partial u^+}(u_{i-1/2,j}^-, \widetilde{\widetilde{u}}) \right) \ge 0 \\ \left|\frac{\partial h}{\partial u}\right|_{i,j}^{\max} &\equiv \sup_{\substack{\widehat{u} \in [u_{i,j-1/2}^-, u_{i,j+1/2}^-] \\ \widehat{\widehat{u}} \in [u_{i,j-1/2}^+, u_{i,j+1/2}^+]}} \left(\frac{\partial h}{\partial u^-}(\widehat{u}, u_{i,j+1/2}^+) - \frac{\partial h}{\partial u^+}(u_{i,j-1/2}^-, \widehat{\widehat{u}}) \right) \ge 0 \ . \end{aligned}$$

Proof: The first step is to introduce the mean value flux function linearization states $\widetilde{u} \in [u_{i-1/2,j}^-, u_{i+1/2,j}^-]$, $\widetilde{\widetilde{u}} \in [u_{i-1/2,j}^+, u_{i+1/2,j}^+]$, $\widehat{u} \in [u_{i,j-1/2}^-, u_{i,j+1/2}^-]$ and $\widehat{\widehat{u}} \in [u_{i,j-1/2}^+, u_{i,j+1/2}^+]$ such that

$$u_{i,j}^{n+1} = u_{i,j}^n \ - \ \frac{\Delta t}{\Delta x}(\frac{\partial g}{\partial u^-}(\tilde{u}, u_{i+1/2,j}^+)\,(u_{i+1/2,j}^- - u_{i-1/2,j}^-)$$

$$
\begin{aligned}
&+\frac{\partial g}{\partial u^+}(u^-_{i-1/2,j},\tilde{\tilde{u}})\,(u^+_{i+1/2,j}-u^+_{i-1/2,j}))\\
-\quad &\frac{\Delta t}{\Delta x}(\frac{\partial h}{\partial u^-}(\tilde{u},u^+_{i,j+1/2})\,(u^-_{i,j+1/2}-u^-_{i,j-1/2})\\
&+\frac{\partial h}{\partial u^+}(u^-_{i,j-1/2},\tilde{\tilde{u}})\,(u^+_{i,j+1/2}-u^+_{i,j-1/2}))
\end{aligned}
$$

or equivalently by

$$
\begin{aligned}
u^{n+1}_{i,j}=u^n_{i,j}\quad - &\quad \frac{\Delta t}{\Delta x}(\underbrace{\frac{\partial g}{\partial u^-}}_{(+)}(\tilde{u},u^+_{i+1/2,j})\,\frac{u^-_{i+1/2,j}-u^-_{i-1/2,j}}{u_{i,j}-u_{i-1,j}}\,(u_{i,j}-u_{i-1,j})\\
&+\underbrace{\frac{\partial g}{\partial u^+}}_{(-)}(u^-_{i-1/2,j},\tilde{\tilde{u}})\,\frac{u^+_{i+1/2,j}-u^+_{i-1/2,j}}{u_{i+1,j}-u_{i,j}}\,(u_{i+1,j}-u_{i,j}))\\
- &\quad \frac{\Delta t}{\Delta y}(\underbrace{\frac{\partial h}{\partial u^-}}_{(+)}(\hat{u},u^+_{i,j+1/2})\,\frac{u^-_{i,j+1/2}-u^-_{i,j-1/2}}{u_{i,j}-u_{i,j-1}}\,(u_{i,j}-u_{i,j-1})\\
&+\underbrace{\frac{\partial h}{\partial u^+}}_{(-)}(u^-_{i,j-1/2},\hat{\tilde{u}})\,\frac{u^+_{i,j+1/2}-u^+_{i,j-1/2}}{u_{i,j+1}-u_{i,j}}\,(u_{i,j+1}-u_{i,j}))\quad .
\end{aligned}
$$

Sufficient conditions for a positive coefficient discretization are then given by

$$
\begin{aligned}
\frac{u^-_{i+1/2,j}-u^-_{i-1/2,j}}{u_{i,j}-u_{i-1,j}}\quad &\geq\quad 0\quad , \quad \frac{u^+_{i+1/2,j}-u^+_{i-1/2,j}}{u_{i+1,j}-u_{i,j}}\geq 0\\
\frac{u^-_{i,j+1/2}-u^-_{i,j-1/2}}{u_{i,j}-u_{i,j-1}}\quad &\geq\quad 0\quad , \quad \frac{u^+_{i,j+1/2}-u^+_{i,j-1/2}}{u_{i,j+1}-u_{i,j}}\geq 0\quad .
\end{aligned}
\tag{47}
$$

Define the following ratios of successive differences

$$
R_{i,i}\equiv\frac{u_{i+1,j}-u_{i,j}}{u_{i,j}-u_{i-1,j}}\quad ,\quad S_{i,i}\equiv\frac{u_{i,j+1}-u_{i,j}}{u_{i,j}-u_{i,j-1}}
$$

and the limited extrapolation formulas

$$
\begin{aligned}
u^-_{i+1/2,j}\quad &=\quad u_{i,j}+\frac{1}{2}\Psi(R_{i,j})\,(u_{i,j}-u_{i-1,j})\\
u^+_{i-1/2,j}\quad &=\quad u_{i,j}-\frac{1}{2}\Psi(1/R_{i,j})\,(u_{i+1,j}-u_{i,j})\\
u^-_{i,j+1/2}\quad &=\quad u_{i,j}+\frac{1}{2}\Psi(S_{i,j})\,(u_{i,j}-u_{i,j-1})\\
u^+_{i,j-1/2}\quad &=\quad u_{i,j}-\frac{1}{2}\Psi(1/S_{i,j})\,(u_{i,j+1}-u_{i,j})\quad .
\end{aligned}
$$

By forming the ratios (47), a sufficient condition for positivity of the ratios is that $\forall R,S\in\mathbb{R}$

$$
1+\frac{1}{2}\Psi(R)-\frac{1}{2}\Psi(S)/S\geq 0\quad .
$$

Boundedness of the coefficients is obtained by requiring that

$$
\Psi(R)-\Psi(S)/S\leq 2\beta\quad .
$$

The scheme is therefore monotonicity preserving under a time step restriction if

$$\alpha \leq \Psi(R) \leq \beta \ .$$

and

$$-\beta \leq \Psi(R)/R \leq 2 + \alpha$$

where we assume $\alpha \in [-2, 0]$. The time step restriction is then directly obtained from Theorem 1.2.8. ∎

A similar proof follows immediately assuming that the numerical flux is an E-flux. Many limiter functions satisfy the technical conditions (46) of Theorem 1.2.9. Some examples include

- the van Leer limiter
$$\Psi^{\mathrm{VL}}(R) = \frac{R + |R|}{1 + |R|} \ ,$$
- the van Albada limiter
$$\Psi^{\mathrm{VA}}(R) = \frac{R + R^2}{1 + R^2} \ .$$

In addition, Koren [Kor88] has constructed the limiter

$$\Psi^{\mathrm{K}}(R) = \frac{R + 2R^2}{2 - R + 2R^2}$$

which also satisfies the technical conditions (46) and corresponds for smooth solutions in 1-D to the most accurate $\kappa = 1/3$ MUSCL scheme of van Leer.

FV schemes on unstructured meshes utilizing linear reconstruction

Higher order finite volume extensions of Godunov discretization to unstructured meshes are of the general form

$$\frac{du_j}{dt} = -\frac{1}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} g_{jk}(u^-_{jk}, u^+_{jk}) \ , \quad \forall T_j \in \mathcal{T} \tag{48}$$

with the numerical flux $g_{jk}(u, v)$ given by the quadrature rule

$$g_{jk}(u^-_{jk}, u^+_{jk}) \equiv \sum_{q=1}^{Q} \omega_q \, g(\nu_{jk}(x_q); u^-_{jk}(x_q), u^+_{jk}(x_q)) \ , \tag{49}$$

where $\omega_q \in \mathbb{R}$ and $x_q \in e_{jk}$ represent quadrature weights and locations, $q = 1, \ldots, Q$. Given the global space of piecewise constant cell averages, $u_h \in V_h^0$, the extrapolated states $u^-_{jk}(x)$ and $u^+_{jk}(x)$ are evaluated using a p-th order polynomial reconstruction operator, $R_p^0 : V_h^0 \mapsto V_h^p$,

$$\begin{aligned} u^-_{jk}(x) &\equiv \lim_{\epsilon \downarrow 0} R_p^0(x - \epsilon \, \nu_{jk}(x); u_h) \\ u^+_{jk}(x) &\equiv \lim_{\epsilon \downarrow 0} R_p^0(x + \epsilon \, \nu_{jk}(x); u_h) \end{aligned}$$

for $x \in e_{jk}$. In addition, it is assumed that the reconstruction satisfies the property $\frac{1}{|T_j|} \int_{T_j} R_p^0(x; u_h) \, dx = u_j$. In the general finite volume formulation, the control volume

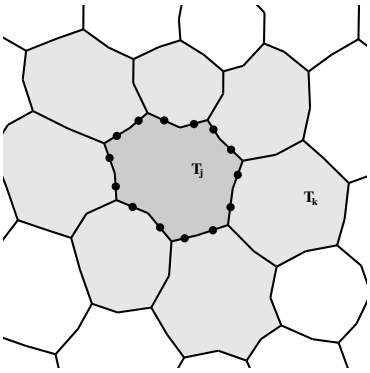


Figure 1.4: Polygonal control volume cell T_j and perimeter quadrature points (solid circles).

shapes need not be convex, see for example Fig. 1.4. Even so, the solution accuracy and maximum stable time step for explicit schemes may depend strongly on the shape of individual control volumes. In the special case of linear reconstruction, $R_1^0(x; u_h)$, the impact of control volume shape on stability of the scheme can be quantified more precisely. Specifically, the maximum principle analysis presented later for the scheme (48) reveals an explicit dependence on the geometrical shape parameter

$$\Gamma^{geom} = \sup_{0 \le \theta \le 2\pi} \alpha^{-1}(\theta) \tag{50}$$

where $0 < \alpha(\theta) < 1$ represents the smallest fractional perpendicular distance from the gravity center to one of two minimally separated parallel hyperplanes with orientation θ and hyperplane location such that all quadrature points in the control volume lie between or on the hyperplanes as shown in Fig. 1.5. Table 1.2 lists Γ^{geom} values for various

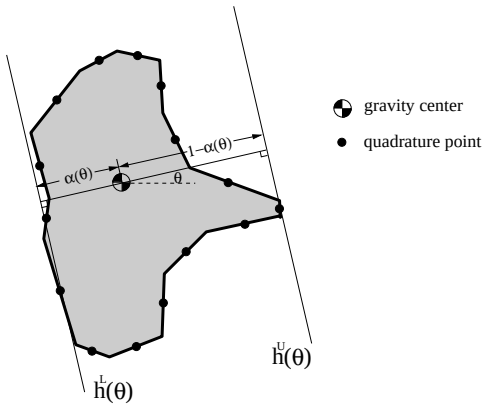


Figure 1.5: Minimally separated hyperplanes $h^L(\theta)$ and $h^U(\theta)$ and the fractional distance ratio $\alpha(\theta)$ for use in the calculation of Γ^{geom}.

control volume shapes in $\mathbb{R}^1$, $\mathbb{R}^2$, $\mathbb{R}^3$, and $\mathbb{R}^d$. As might be expected, those geometries that have exact quadrature point symmetry with respect to the control volume gravity center have geometric shape parameters Γ^{geom} equal to 2 regardless of the number of space dimensions involved. The following lemma and subsequent theorem build upon several techniques set forth in [Osh84, Liu93, Wie94, BLC96] that are now extended to arbitrary linear reconstruction and general control volume shapes.

Table 1.2: Reconstruction geometry factors for various control volume shapes utilizing midpoint quadrature rule.

control volume shape	space dimension	$\Gamma^{\rm geom}$
segment	1	2
triangle	2	3
parallelogram	2	2
regular n-gon	2	$n/\left\lceil \frac{n-1}{2} \right\rceil$
tetrahedron	3	4
parallelepiped	3	2
simplex	d	d+1
hyper-parallelepiped	d	2
polytope	d	Eqn. (50)

Lemma 1.2.10 (Finite volume interval bounds on unstructured meshes, $\mathbf{R_1^0(x;u_h)}$)

The fully discrete finite volume scheme

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} g_{jk}(u_{jk}^{-,n}, u_{jk}^{+,n}) \ , \quad \forall T_j \in \mathcal{T} \tag{51}$$

with monotone Lipschitz continuous numerical flux function, nonnegative quadrature weights, and linear reconstructions

$$\begin{aligned} u_{jk}^-(x) &\equiv \lim_{\epsilon \downarrow 0} R_1^0(x - \epsilon\, \nu_{jk}(x); u_h) \ , \quad x \in e_{jk} \ , \quad u_h \in V_h^0 \\ u_{jk}^+(x) &\equiv \lim_{\epsilon \downarrow 0} R_1^0(x + \epsilon\, \nu_{jk}(x); u_h) \ , \quad x \in e_{jk} \ , \quad u_h \in V_h^0 \ , \end{aligned}$$

with extremal trace values at control volume quadrature points

$$U_j^{\min} \equiv \min_{\substack{\forall e_{jk} \in \partial T_j \\ 1 \le q \le Q}} u_{jk}^{\pm}(x_q) \ , \quad U_j^{\max} \equiv \max_{\substack{\forall e_{jk} \in \partial T_j \\ 1 \le q \le Q}} u_{jk}^{\pm}(x_q) \ , \quad x_q \in e_{jk} \tag{52}$$

exhibits the local interpolated interval bound

$$\sigma_j U_j^{\min,n} + (1 - \sigma_j) u_j^n \le u_j^{n+1} \le (1 - \sigma_j)\, u_j^n + \sigma_j U_j^{\max,n} \tag{53}$$

with the time step proportional interpolation parameter σ_j defined by

$$\sigma_j \equiv \frac{\Delta t}{|T_j|}\, \Gamma^{\rm geom} \sum_{\substack{\forall e_{jk} \in \partial T_j \\ 1 \le q \le Q}} \sup_{\substack{\widetilde{u} \in [U_j^{\min,n}, U_j^{\max,n}] \\ \widetilde{\widetilde{u}} \in [U_j^{\min,n}, U_j^{\max,n}]}} \left| \frac{\partial g_{jk}}{\partial u^+}(\nu_{jk}(x_q); \widetilde{u}, \widetilde{\widetilde{u}}) \right| \tag{54}$$

that depends on the shape parameter $\Gamma^{\rm geom}$ defined in (50).

Proof: Let u_A and u_B denote two arbitrary states. The scheme in semi-discrete form is readily manipulated into the following equivalent forms

$$(u_j)_t = -\frac{1}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} g_{jk}(u_{jk}^-, u_{jk}^+)$$

$$
\begin{aligned}
&= -\frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j}\left(g_{jk}(u_A,u_B)+\left(g_{jk}(u_{jk}^-,u_{jk}^+)-g_{jk}(u_A,u_B)\right)\right)\\
&= -\frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j}\Big(g_{jk}(u_A,u_B)+\left(g_{jk}(u_{jk}^-,u_{jk}^+)-g_{jk}(u_A,u_{jk}^+)\right)\\
&\qquad\qquad +\left(g_{jk}(u_A,u_{jk}^+)-g_{jk}(u_A,u_B)\right)\Big)\\
&= -\frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j}\Big(g_{jk}(u_A,u_B)+\sum_{1\le q\le Q}\omega_q\underbrace{\frac{\partial g}{\partial u^-}}_{(+)}(\nu_{jk}(x_q);\tilde{u}_{jkq},u_{jk}^+)\left(u_{jk}^-(x_q)-u_A\right)\\
&\qquad\qquad +\sum_{1\le q\le Q}\omega_q\underbrace{\frac{\partial g}{\partial u^+}}_{(-)}(\nu_{jk}(x_q);u_A,\tilde{\tilde{u}}_{jkq})\left(u_{jk}^+(x_q)-u_B\right)\Big)
\end{aligned}
$$

for assumed nonnegative quadrature weights ω_q and chosen mean value states $\tilde{u}_{jkq}\in[u_A,u_{jk}^-(x_q)]$ and $\tilde{\tilde{u}}_{jkq}\in[u_B,u_{jk}^+(x_q)]$. Next, define the extreme trace values at control volume quadrature points

$$
\mathcal{U}_j^{\min}\equiv\min_{\substack{\forall e_{jk}\in\partial T_j\\1\le q\le Q}}u_{jk}^+(x_q)\ ,\quad \mathcal{U}_j^{\max}\equiv\max_{\substack{\forall e_{jk}\in\partial T_j\\1\le q\le Q}}u_{jk}^+(x_q)\ ,\quad x_q\in e_{jk}
$$

$$
u_j^{\min}\equiv\min_{\substack{\forall e_{jk}\in\partial T_j\\1\le q\le Q}}u_{jk}^-(x_q)\ ,\quad u_j^{\max}\equiv\max_{\substack{\forall e_{jk}\in\partial T_j\\1\le q\le Q}}u_{jk}^-(x_q)\ ,\quad x_q\in e_{jk}
$$

so that upon setting

$$
u_A=u_j^{\min}\text{ and }u_B=U_j^{\max}\equiv\max(u_j^{\max},\mathcal{U}_j^{\max})
$$

the following inequality from above holds

$$
(u_j)_t\ \le\ -\frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j}g_{jk}(u_j^{\min},U_j^{\max})\ .
$$

Similarly, upon setting

$$
u_A\ =\ u_j^{\max}\text{ and }u_B=U_j^{\min}\equiv\min(u_j^{\min},\mathcal{U}_j^{\min})
$$

the following inequality from below holds

$$
(u_j)_t\ \ge\ -\frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j}g_{jk}(u_j^{\max},U_j^{\min})\ .
$$

In both cases, the numerical flux of a constant state over the support of the local discretization can be added since its contribution vanishes when summed over a closed control volume

$$
\begin{aligned}
(u_j)_t\ &\le\ -\frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j}g_{jk}(u_j^{\min},U_j^{\max})\\
&=\ -\frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j}\left(g_{jk}(u_j^{\min},U_j^{\max})-g_{jk}(u_j^{\min},u_j^{\min})\right)\\
&=\ \frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j}\left|\frac{\partial g_{jk}}{\partial u^+}(u_j^{\min},\widehat{u}_j)\right|\left(U_j^{\max}-u_j^{\min}\right)
\end{aligned}
\tag{55}
$$

and the second case

$$\begin{aligned}(u_j)_t &\geq -\frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j} g_{jk}(u_j^{\max}, U_j^{\min}) \\ &= -\frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j} \left(g_{jk}(u_j^{\max}, U_j^{\min}) - g_{jk}(u_j^{\max}, u_j^{\max})\right) \\ &= \frac{1}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j} \left|\frac{\partial g_{jk}}{\partial u^+}(u_j^{\max}, \widehat{\widehat{u}}_j)\right| (U_j^{\min} - u_j^{\max}) \end{aligned} \tag{56}$$

for chosen mean value states $\widehat{u} \in [u_j^{\min}, U_j^{\max}]$ and $\widehat{\widehat{u}} \in [u_j^{\max}, U_j^{\min}]$. The next task is to bound $U_j^{\max} - u_j^{\min}$ assuming a linear reconstruction with cell average u_j located at the centroid of T_j. From the definition of Γ^{geom} in (50), a bound on the extremal interior states and cell average of the form

$$u_j - u_j^{\min} \leq \gamma^{\text{geom}} \left(u_j^{\max} - u_j^{\min}\right) \ , \quad \gamma^{\text{geom}} \equiv \frac{\Gamma^{\text{geom}} - 1}{\Gamma^{\text{geom}}}$$

for $\Gamma^{geom} \in (1, \infty)$ is readily obtained for any linear reconstruction operator with cell average equal to u_j. By definition, $u_j^{\max} \leq U_j^{\max}$ so that the sequence of inequalities follows straightforwardly

$$\begin{aligned} U_j^{\max} - u_j^{\min} &= (U_j^{\max} - u_j) + (u_j - u_j^{\min}) \\ &\leq (U_j^{\max} - u_j) + \gamma^{\text{geom}} \left(u_j^{\max} - u_j^{\min}\right) \\ &\leq (U_j^{\max} - u_j) + \gamma^{\text{geom}} \left(U_j^{\max} - u_j^{\min}\right) \\ (1 - \gamma^{\text{geom}})(U_j^{\max} - u_j^{\min}) &\leq (U_j^{\max} - u_j) \\ U_j^{\max} - u_j^{\min} &\leq \Gamma^{\text{geom}} \ (U_j^{\max} - u_j) \ . \end{aligned}$$

Using similar arguments, a bound inequality on $U_j^{\min} - u_j^{\max}$ is obtained

$$U_j^{\min} - u_j^{\max} \geq \Gamma^{\text{geom}} \ (U_j^{\min} - u_j) \ .$$

Inserting these inequalities into (55) and (56)

$$\frac{\Gamma^{\text{geom}}}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j} \left|\frac{\partial g_{jk}}{\partial u^+}(u_j^{\max}, \widehat{\widehat{u}}_j)\right| (U_j^{\min} - u_j) \leq (u_j)_t \leq \frac{\Gamma^{\text{geom}}}{|T_j|}\sum_{\forall e_{jk}\in\partial T_j} \left|\frac{\partial g_{jk}}{\partial u^+}(u_j^{\min}, \widehat{u}_j)\right| (U_j^{\max} - u_j)$$

or in slightly weaker form

$$\delta_j \ (U_j^{\min} - u_j) \leq (u_j)_t \leq \delta_j \ (U_j^{\max} - u_j)$$

with

$$\delta_j \equiv \frac{1}{|T_j|}\Gamma^{\text{geom}} \sum_{\substack{\forall e_{jk}\in\partial T_j \\ 1\leq q\leq Q}} \sup_{\substack{\widetilde{u}\in[U_j^{\min,n}, U_j^{\max,n}] \\ \widetilde{\widetilde{u}}\in[U_j^{\min,n}, U_j^{\max,n}]}} \left|\frac{\partial g_{jk}}{\partial u^+}(\nu_{jk}(x_q); \widetilde{u}, \widetilde{\widetilde{u}})\right| \ .$$

Replacing the time derivative with a discrete forward Euler integration together with rearrangement of terms yields the stated lemma. ∎

Given the two-sided bound of Lemma 1.2.10, a discrete maximum principle is obtained under a CFL-like time step restriction if the limits $U_j^{\max}$ and $U_j^{\min}$ can be bounded from above and below respectively by the neighboring cell averages. This idea is given more precisely in the following theorem.

Theorem 1.2.11 (Finite volume maximum principle on unstructured meshes, $\mathbf{R}_1^0$) *Let $u_j^{\min}$ and $u_j^{\max}$ denote the minimum and maximum value of solution cell averages for a given cell T_j and corresponding adjacent cell neighbors, i.e.*

$$u_j^{\min} \equiv \min_{\forall e_{jk} \in \partial T_j} (u_j, u_k) \;\; \textit{and} \;\; u_j^{\max} \equiv \max_{\forall e_{jk} \in \partial T_j} (u_j, u_k) \;\; . \tag{57}$$

The fully discrete finite volume scheme

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} g_{jk}(u_{jk}^{-,n}, u_{jk}^{+,n}) \;\; , \quad \forall T_j \in \mathcal{T} \tag{58}$$

with monotone Lipschitz continuous numerical flux function, nonnegative quadrature weights, and linear reconstructions

$$\begin{aligned} u_{jk}^-(x) &\equiv \lim_{\epsilon \downarrow 0} R_1^0(x - \epsilon\, \nu_{jk}(x); u_h) \;\; , \quad x \in e_{jk} \;\; , \quad u_h \in V_h^0 \\ u_{jk}^+(x) &\equiv \lim_{\epsilon \downarrow 0} R_1^0(x + \epsilon\, \nu_{jk}(x); u_h) \;\; , \quad x \in e_{jk} \;\; , \quad u_h \in V_h^0 \end{aligned} \tag{59}$$

exhibits the local space-time maximum principle for each $T_j \in \mathcal{T}$

$$\min_{\forall e_{jk} \in \partial T_j} (u_j^n, u_k^n) \le u_j^{n+1} \le \max_{\forall e_{jk} \in \partial T_j} (u_j^n, u_k^n)$$

as well as the local spatial maximum principle at steady state ($u^{n+1} = u^n = u^*$)

$$\min_{\forall e_{jk} \in \partial T_j} u_k^* \le u_j^* \le \max_{\forall e_{jk} \in \partial T_j} u_k^*$$

if the linear reconstruction satisfies $\forall e_{jk} \in \partial T_j$ and $x_q \in e_{jk}, q = 1, \ldots, Q$

$$\max(u_j^{\min,n}, u_k^{\min,n}) \le u_{jk}^{-,n}(x_q) \le \min(u_j^{\max,n}, u_k^{\max,n}) \tag{60}$$

under the time step restriction

$$1 - \frac{\Delta t}{|T_j|}\, \Gamma^{\text{geom}} \sum_{\substack{\forall e_{jk} \in \partial T_j \\ 1 \le q \le Q}} \sup_{\substack{\widetilde{u} \in [u_j^{\min,n}, u_j^{\max,n}] \\ \widetilde{\widetilde{u}} \in [u_j^{\min,n}, u_j^{\max,n}]}} \left| \frac{\partial g_{jk}}{\partial u^+}(\nu_{jk}(x_q); \widetilde{u}, \widetilde{\widetilde{u}}) \right| \ge 0$$

with Γ^{geom} defined in (50).

Proof: The proof follows immediately from (53) in Lemma 1.2.10 by considering an interface shared by two control volumes since the condition (60) places a bound on the extremal trace values (52). ∎

Note that a variant of this theorem also holds if the definition of $u^{\max}$ and $u^{\min}$ are expanded to include more control volume neighbors. Two alternative definitions frequently used when the control volume shape is a simplex are given by

$$u_j^{\min} \equiv \min_{\substack{T_k \in \mathcal{T} \\ T_j \cap T_k \neq \emptyset}} u_k \;\; \text{and} \;\; u_j^{\max} \equiv \max_{\substack{T_k \in \mathcal{T} \\ T_j \cap T_k \neq \emptyset}} u_k \;\; . \tag{61}$$

These expanded definitions include adjacent cells whose intersection with T_j in $\mathbb{R}^d$ need only be a set of measure zero or greater.

Slope limiters for linear reconstruction. Given a linear reconstruction $R_1^0(x;u_h)$ that does not necessarily satisfy the requirements of Theorem 1.2.11, it is straightforward to modify the reconstruction so that the new modified reconstruction does satisfy the requirements of Theorem 1.2.11. For each control volume $T_j \in \mathcal{T}$ a modified reconstruction operator $\widetilde{R}_1^0(x;u_h)$ of the form

$$\widetilde{R}_1^0(x;u_h)|_{T_j} = u_j + \alpha_{T_j}(R_1^0(x;u_h)|_{T_j} - u_j)$$

is assumed for $\alpha_{T_j} \in [0,1]$. By construction, this modified reconstruction correctly reproduces the control volume cell average for all values of α_{T_j}, i.e.

$$\frac{1}{|T_j|}\int_{T_j} \widetilde{R}_1^0(x;u_h)\,dx = u_j \quad . \tag{62}$$

The most restrictive value of α_{T_j} for each control volume T_j is then computed based on the Theorem 1.2.11 constraint (60), i.e.

$$\alpha_{T_j}^{\mathrm{MM}} = \min_{\substack{\forall e_{jk}\in\partial T_j \\ 1\le q\le Q}} \begin{cases} \frac{\min(u_j^{\max},u_k^{\max})-u_j}{R_1^0(x_q;u_h)|_{T_j}-u_j} & \text{if } R_1^0(x_q;u_h)|_{T_j} > \min(u_j^{\max},u_k^{\max}) \\ \frac{\max(u_j^{\min},u_k^{\min})-u_j}{R_1^0(x_q;u_h)|_{T_j}-u_j} & \text{if } R_1^0(x_q;u_h)|_{T_j} < \max(u_j^{\min},u_k^{\min}) \\ 1 & \text{otherwise} \end{cases} \tag{63}$$

where $u^{\max}$ and $u^{\min}$ are defined in (57). When the resulting modified reconstruction operator is used in the extrapolation formulas (59), the discrete maximum principle of Theorem 1.2.11 is attained under a CFL-like time step restriction. By utilizing the inequalities

$$\max(u_j,u_k) \le \min(u_j^{\max},u_k^{\max}) \text{ and } \quad \min(u_j,u_k) \ge \max(u_j^{\min},u_k^{\min})$$

it is straightforward to construct a simpler but more restrictive limiter function

$$\alpha_{T_j}^{\mathrm{LM}} = \min_{\substack{\forall e_{jk}\in\partial T_j \\ 1\le q\le Q}} \begin{cases} \frac{\max(u_j,u_k)-u_j}{R_1^0(x_q;u_h)|_{T_j}-u_j} & \text{if } R_1^0(x_q;u_h)|_{T_j} > \max(u_j,u_k) \\ \frac{\min(u_j,u_k)-u_j}{R_1^0(x_q;u_h)|_{T_j}-u_j} & \text{if } R_1^0(x_q;u_h)|_{T_j} < \min(u_j,u_k) \\ 1 & \text{otherwise} \end{cases} \tag{64}$$

that yields modified reconstructions satisfying the technical conditions of Theorem 1.2.11. This simplified limiter (64) introduces additional slope reduction when compared to (63). This can be detrimental to the overall accuracy of the discretization. The limiter strategy (64) and other variants for simplicial control volumes are discussed further in [Liu93, Wie94, BLC96].

In Barth [BJ89], a variant of (63) was proposed

$$\alpha_{T_j}^{\mathrm{BJ}} = \min_{\substack{\forall e_{jk}\in\partial T_j \\ 1\le q\le Q}} \begin{cases} \frac{u_j^{\max}-u_j}{R_1^0(x_q;u_h)|_{T_j}-u_j} & \text{if } R_1^0(x_q;u_h)|_{T_j} > u_j^{\max} \\ \frac{u_j^{\min}-u_j}{R_1^0(x_q;u_h)|_{T_j}-u_j} & \text{if } R_1^0(x_q;u_h)|_{T_j} < u_j^{\min} \\ 1 & \text{otherwise} \end{cases} \quad . \tag{65}$$

Although this limiter function does not produce modified reconstructions satisfying the requirements of Theorem 1.2.11, using Lemma 1.2.10 it can be shown that the Barth and Jespersen limiter yields finite volume schemes (51) possessing a global extremum diminishing property, i.e. that the solution maximum is non-increasing and the solution minimum is nondecreasing between successive time levels. This limiter function produces the least amount of slope reduction when compared to the limiter functions (63) and (64). Note that in practical implementation, all three limiters (63), (64) and (65) require some modification to prevent near zero division for nearly constant solution data.

Linear reconstruction operators on simplicial control volumes

Linear reconstruction operators on simplicial control volumes that satisfy the cell averaging requirement (42b) often exploit the fact that the cell average is also a pointwise value of any valid linear reconstruction evaluated at the gravity center of a simplex. This reduces the reconstruction problem to that of gradient estimation given pointwise samples at the gravity centers. In this case, it is convenient to express the reconstruction in the form

$$R_1^0(x;u_h)|_{T_j} = u_j + (\nabla u_h)_{T_j} \cdot (x - x_j^{\bullet}) \tag{66}$$

where $x_j^{\bullet}$ denotes the gravity center for the simplex T_j and $(\nabla u_h)_{T_j}$ is the gradient to be determined. Figure 1.6 depicts a 2-D simplex $\triangle_{123}$ and three adjacent neighboring simplices. Also shown are the corresponding four pointwise solution values $\{A, B, C, O\}$ located at gravity centers of each simplex. By selecting any three of the four pointwise solution values, a set of four possible gradients are uniquely determined, i.e. $\{\ \nabla(ABC)$, $\nabla(ABO)$, $\nabla(BCO)$, $\nabla(CAO)\ \}$. Using the example of Fig. 1.6, a number of slope limited

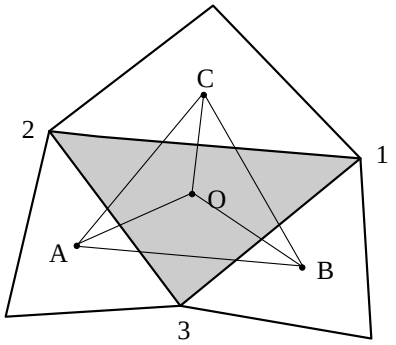


Figure 1.6: Triangle control volume $\triangle_{123}$ (shaded) with three adjacent cell neighbors.

reconstruction techniques are possible for use in the finite volume scheme (58) that meet the technical conditions of Theorem 1.2.11.

1. Choose $(\nabla u_h)_{T_{123}} = \nabla(ABC)$ and limit the resulting reconstruction using (63) or (64). This technique is pursued in Barth [BJ89] but using the limiter (65) instead.

2. Limit the reconstructions corresponding to gradients $\nabla(ABC), \nabla(ABO), \nabla(BCO)$ and $\nabla(CAO)$ using (63) or (64) and choose the limited reconstruction with largest gradient magnitude. This technique is a generalization of that described in Batten et al. [BLC96] wherein limiter (64) is used.

3. Choose the unlimited reconstruction $\nabla(ABC), \nabla(ABO), \nabla(BCO)$ and $\nabla(CAO)$ with largest gradient magnitude that satisfies the maximum principle reconstruction bound inequality (60). If all reconstructions fail the bound inequality, the reconstruction gradient is set equal to zero, see Liu [Liu93].

Linear reconstruction operators on general control volumes shapes

In the case of linear reconstruction on general volume shapes, significant simplification is possible when compared to the general p-exact reconstruction formulation given in Sect. 1.2.2. It is again convenient to express the reconstruction in the form

$$R_1^0(x;u_h)|_{T_j} = u_j + (\nabla u_h)_{T_j} \cdot (x - x_j^{\bullet}) \tag{67}$$

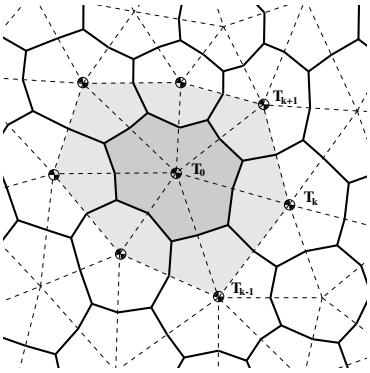


Figure 1.7: Triangulation of gravity center locations showing a typical control volume T_0 associated with the triangulation vertex v_0 with cyclically indexed graph neighbors $T_k, k = 1, \ldots, N_0$.

where $x_j^{\bullet}$ denotes the gravity center for the control volume T_j and $(\nabla u_h)_{T_j}$ is the gradient to be determined. Two common techniques for simplified linear reconstruction include a simplified least squares technique and a Green-Gauss integration technique.

Simplified least squares linear reconstruction. As was exploited in the linear reconstruction techniques for simplicial control volumes, linear reconstructions satisfying (42b) on general control volume shapes are greatly simplified by exploiting the fact that the cell average value is also a pointwise value of all valid linear reconstructions evaluated at the gravity center of a general control volume shape. This again reduces the linear reconstruction problem to that of gradient estimation given pointwise values. In the simplified least squares reconstruction technique, a triangulation (2D) or tetrahedralization (3D) of gravity centers is first constructed as shown in Fig. 1.7. Referring to this figure, for each edge of the simplex mesh incident to the vertex v_0, an edge projected gradient constraint equation is constructed subject to a pre-specified nonzero scaling w_k

$$w_k \, (\nabla u_h)_{T_0} \cdot (x_k^{\bullet} - x_0^{\bullet}) = w_k (u_k - u_0) \ .$$

The number of edges incident to a simplex mesh vertex in $\mathbb{R}^d$ is greater than or equal to d thereby producing the following generally non-square matrix of constraint equations

$$\begin{bmatrix} w_1 \Delta x_1^{\bullet} & w_1 \Delta y_1^{\bullet} \\ \vdots & \vdots \\ w_{N_0} \Delta x_{N_0}^{\bullet} & w_{N_0} \Delta y_{N_0}^{\bullet} \end{bmatrix} (\nabla u_h)_{T_0} = \begin{pmatrix} w_1 (u_1 - u_0) \\ \vdots \\ w_{N_0} (u_{N_0} - u_0) \end{pmatrix}$$

or in abstract form

$$\begin{bmatrix} \vec{L}_1 & \vec{L}_2 \end{bmatrix} \nabla u = \vec{f} \ .$$

This abstract form can be symbolically solved in a least squares sense using an orthogonalization technique yielding the closed form solution

$$\nabla u = \frac{1}{l_{11} l_{22} - l_{12}^2} \begin{pmatrix} l_{22} (\vec{L}_1 \cdot \vec{f}) - l_{12} (\vec{L}_2 \cdot \vec{f}) \\ l_{11} (\vec{L}_2 \cdot \vec{f}) - l_{12} (\vec{L}_1 \cdot \vec{f}) \end{pmatrix} \tag{68}$$

with $l_{ij} = \vec{L}_i \cdot \vec{L}_j$. The form of this solution in terms of scalar dot products over incident edges suggests that the least squares linear reconstruction can be efficiently computed via an edge data structure without the need for storing a non-square matrix.

Green-Gauss linear reconstruction. This reconstruction technique specific to simplicial meshes assumes nodal solution values at vertices of the mesh which uniquely describes a C^0 linear interpolant, u_h. Gradients are then computed from the mean value approximation

$$|\Omega_0|\,(\nabla u_h)_{\Omega_0} \approx \int_{\Omega_0} \nabla u_h\, dx = \int_{\partial\Omega_0} u_h\, d\nu \ . \tag{69}$$

For linear interpolants, the right-hand side term can be written in the following equivalent

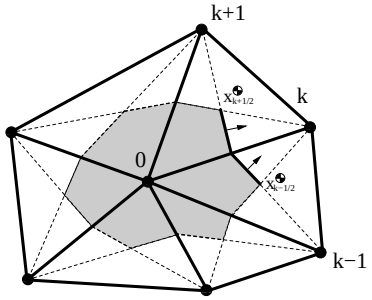


Figure 1.8: Median dual control volume T_0 demarcated by median segments of triangles incident to the vertex v_0 with cyclically indexed adjacent vertices $v_k, k = 1, \ldots N_0$.

form using the configuration depicted in Fig. 1.8

$$\int_{\Omega_0} \nabla u_h\, dx = \sum_{k=1}^{N_0} \frac{3}{2}(u_0 + u_k)\, \nu_{0k}$$

where ν_{0k} represents *any* path integrated normal connecting pairwise adjacent simplex gravity centers, i.e.

$$\nu_{0k} = \int_{x_{k-1/2}}^{x_{k+1/2}} d\nu \ . \tag{70}$$

A particularly convenient path is one that traces out portions of median segments as shown in Fig. 1.8. These segments demarcate the so-called *median dual* control volume. By construction, the median dual volume $|T_0|$ is precisely equal to $|\Omega_0|/3$ in 2-D. Consequently, a linear reconstruction operator on non-overlapping median dual control volumes is given by

$$|T_0|(\nabla u_h)_{T_0} \approx \sum_{k=1}^{N_0} \frac{1}{2}(u_0 + u_k)\, \nu_{0k} \ . \tag{71}$$

The gradient calculation is exact whenever the numerical solution varies linearly over the support of the reconstruction. Since mesh vertices are not located at the gravity centers of median dual control volumes, the cell averaging property (42b) and the bounds of Theorem 1.2.11 are only approximately satisfied using the Green-Gauss technique.

A number of slope limited linear reconstruction strategies for general control volume shapes are possible for use in the finite volume scheme (58) that satisfy the technical conditions of Theorem 1.2.11. Using the example depicted in Fig. 1.7, let $\nabla_{k+1/2} u_h$ denote the unique linear gradient calculated from the cell average set $\{u_0, u_k, u_{k+1}\}$. Three slope limiting strategies that are direct counterparts of the simplex control volume case are:

1. Compute $(\nabla u_h)_{T_0}$ using the least squares linear reconstruction or any other valid linear reconstruction technique and limit the resulting reconstruction using (63) or (64).

2. Limit the reconstructions corresponding to the gradients $\nabla_{k+1/2} u_h, k = 1, \ldots, N_0$ and $(\nabla u_h)_{T_0}$ using (63) or (64) and choose the limited reconstruction with largest gradient magnitude.

3. Choose the unlimited reconstruction from $\nabla_{k+1/2} u_h, k = 1, \ldots, N_0$ and $(\nabla u_h)_{T_0}$ with largest gradient magnitude that satisfies the maximum principle reconstruction bound inequality (60). If all reconstructions fail the bound inequality, the reconstruction gradient is set equal to zero.

General p-exact reconstruction operators on unstructured meshes

Abstractly, the reconstruction operator serves as a finite-dimensional pseudo inverse of the cell averaging operator A whose j-th component A_j computes the cell average of the solution in T_j

$$A_j u = \frac{1}{|T_j|} \int_{T_j} u \, dx \ .$$

The development of a general polynomial reconstruction operator, R_p^0, that reconstructs p-degree polynomials from cell averages on unstructured meshes follows from the application of a small number of simple properties.

1. (Conservation of the mean) Given solution cell averages u_h, the reconstruction $R_p^0 u_h$ is required to have the correct cell average, i.e.

   $$\text{if } v = R_p^0 u_h \text{ then } u_h = Av \ .$$

   More concisely,

   $$AR_p^0 = I$$

   so that R_p^0 is a right inverse of the averaging operator A.

2. (p-exactness) A reconstruction operator R_p^0 is p-exact if $R_p^0 A$ reconstructs polynomials of degree p or less exactly. Denoting by $\mathcal{P}_p$ the space of all polynomials of degree p,

   $$\text{if } u \in \mathcal{P}_p \text{ and } v = Au \text{ then } R_p^0 v = u \ .$$

   This can be written succinctly as

   $$R_p^0 A|_{\mathcal{P}_p} = I$$

   so that R_p^0 is a left inverse of the averaging operator A restricted to the space of polynomials of degree at most p.

3. (Compact support) The reconstruction in a control volume T_j should only depend of cell averages in a relatively small neighborhood surrounding T_j. Recall that a polynomial of degree p in $\mathbb{R}^d$ contains $\binom{p+d}{d}$ degrees of freedom. The support set for T_j is required to contain at least this number of neighbors. As the support set becomes even larger for fixed p, not only does the computational cost increase, but eventually the accuracy decreases as less valid data from further away is brought into the calculation.

Practical implementations of polynomial reconstruction operators fall into two classes:

- Fixed support stencil reconstructions. These methods choose a fixed support set as a preprocessing step. Various limiting strategies are then employed to obtain non-oscillatory approximation, see for example Barth [BF90] and Delanaye [Del96] for further details.

- Adaptive support stencil reconstructions. These ENO-like methods dynamically choose reconstruction stencils based on solution smoothness criteria, see for example [HC91, Van93, Abg94, Son97, Son98] for further details.

Positive coefficient schemes on unstructured meshes

Several related positive coefficient schemes have been proposed on multi-dimensional simplicial meshes based on one-dimensional interpolation. The simplest example is the *upwind triangle scheme* as introduced by Billey et al. [BPPS87], Desideri and Dervieux [DD88], Rostand and Stoufflet[RS88] with later improved variants given by Jameson [Jam93] and Cournede et al. [CCDD98]. These schemes are not Godunov methods in the sense that a single multi-dimensional gradient is not obtained in each control volume. The basis for these methods originates from the gradient estimation formula (71) generalized to the calculation of flux divergence on a median dual tessellation. In deriving this flux divergence formula, the assumption has been made that flux components vary linearly within a simplex yielding the discretization formula

$$\int_{T_j} \mathrm{div}(f)\, dx = \int_{\partial T_j} f \cdot d\nu = \sum_{\forall e_{jk} \in \partial T_j} \frac{1}{2} \left(f(u_j) + f(u_k) \right) \cdot \nu_{jk}$$

where ν_{jk} is computed from a median dual tessellation using (70). This discretization is the unstructured mesh counterpart of central differencing on a structured mesh. Schemes using this discretization of flux divergence lack sufficient stability properties for computing solutions of general nonlinear conservation laws. This lack of stability can be overcome by adding suitable diffusion terms. One of the simplest modifications is motivated by upwind domain of dependence arguments yielding the numerical flux

$$g_{jk}(u_j, u_k) = \frac{1}{2} \left(f(u_j) + f(u_k) \right) \cdot \nu_{jk} - \frac{1}{2} |a|_{jk} \, \Delta_{jk} u \quad , \quad \Delta_{jk} u \equiv u_k - u_j \tag{72}$$

with a_{jk} a mean value (a.k.a. Murman-Cole) linearization satisfying

$$\nu_{jk} \cdot \Delta_{jk} f = a_{jk} \, \Delta_{jk} u \quad .$$

Away from sonic points where $f'(u^*) = 0$ for $u^* \in [u_j, u_{j+1}]$, this numerical flux is formally an E-flux satisfying (15). With suitable modifications of a_{jk} near sonic points, it is then possible to produce a modified numerical flux that is an E-flux for all data, see Osher [Osh84]. Theorems 1.1.6, 1.1.7 and 1.1.8 show that schemes such as (9) using E-fluxes exhibit local discrete maximum principles and L_∞ stability.

Unfortunately, schemes based on (72) are too dissipative for most practical calculations. The main idea in the upwind triangle scheme is to add anti-diffusion terms to the numerical flux function (72) such that the sum total of added diffusion and anti-diffusion terms in the numerical flux function vanish entirely whenever the numerical solution varies linearly over the support of the flux function. In all remaining situations, the precise amount of anti-diffusion is determined from maximum principle analysis.

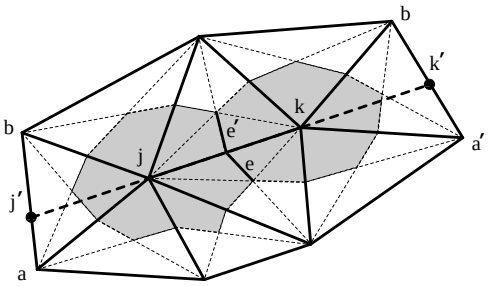


Figure 1.9: Triangle complex used in the upwind triangle schemes showing the linear extension of e_{jk} into neighboring triangle for the determination of points $x_{j'}$ and $x_{k'}$.

Theorem 1.2.12 (Maximum Principles for the Upwind Triangle Scheme) *Let $\mathcal{T}$ denote the median dual tessellation of an underlying simplicial mesh. Also let u_j denote the nodal solution value at a simplex vertex in one-to-one dual correspondence with the control volume $T_j \in \mathcal{T}$ such that a C^0 linear solution interpolant is uniquely specified on the simplicial mesh. Let $g_{jk}(u_{j'}, u_j, u_k, u_{k'})$ denote the numerical flux function with limiter function $\Psi(\cdot) : \mathbb{R} \mapsto \mathbb{R}$*

$$g_{jk}(u_{j'}, u_j, u_k, u_{k'}) \equiv \frac{1}{2}(f(u_j) + f(u_k)) \cdot \nu_{jk} \quad - \quad \frac{1}{2} a^+_{jk} \left(1 - \Psi\left(\frac{h_{jk}\ \Delta_{j'j} u}{h_{j'j}\ \Delta_{jk} u}\right)\right) \Delta_{jk} u$$
$$+ \quad \frac{1}{2} a^-_{jk} \left(1 - \Psi\left(\frac{h_{jk}\ \ \Delta_{kk'} u}{h_{kk'}\ \Delta_{jk} u}\right)\right) \Delta_{jk} u \ ,$$

utilizing the mean value speed a_{jk} satisfying

$$\nu_{jk} \cdot \Delta_{jk} f = a_{jk}\, \Delta_{jk} u \tag{73}$$

and variable spacing parameter $h_{jk} = |\Delta_{jk} x|$. The fully discrete finite volume scheme

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{|T_j|} \sum_{\forall e_{jk} \in \partial T_j} g_{jk}(u^n_{j'}, u^n_j, u^n_k, u^n_{k'}) \ , \quad \forall T_j \in \mathcal{T}$$

with linearly interpolated values $u_{j'}$ and $u_{k'}$ as depicted in Fig. 1.9 exhibits the local space-time maximum principle

$$\min_{\forall e_{jk} \in \partial T_j} (u^n_j, u^n_k) \le u^{n+1}_j \le \max_{\forall e_{jk} \in \partial T_j} (u^n_j, u^n_k)$$

under a time step restriction and the local spatial maximum principle at steady state ($u^{n+1} = u^n = u^*$)

$$\min_{\forall e_{jk} \in \partial T_j} u^*_k \le u^*_j \le \max_{\forall e_{jk} \in \partial T_j} u^*_k$$

if the limiter $\Psi(R)$ satisfies $\forall R \in \mathbb{R}$

$$0 \le \Psi(R)/R \ , \ \Psi(R) \le 2 \ .$$

Proof: Utilizing the mean value linearization 73, the numerical flux is rewritten in the following equivalent forms:

$$g_{jk}(u_{j'}, u_j, u_k, u_{k'}) = \left(\frac{1}{2} a^+_{jk}\, \Psi\left(\frac{h_{jk}\ \Delta_{j'j} u}{h_{j'j}\ \Delta_{jk} u}\right) + \frac{1}{2} a^-_{jk} \left(2 - \Psi\left(\frac{h_{jk}\ \ \Delta_{kk'} u}{h_{kk'}\ \Delta_{jk} u}\right)\right)\right) \Delta_{jk} u$$

$$= \frac{1}{2} a_{jk}^{+} \Psi\left(\frac{h_{jk}\Delta_{j'j}u}{h_{j'j}\Delta_{jk}u}\right)\left(\frac{\Delta_{jk}u}{\Delta_{j'j}u}\right)\Delta_{j'j}u + \frac{1}{2}|a_{jk}^{-}|\left(2 - \Psi\left(\frac{h_{jk}\ \Delta_{kk'}u}{h_{kk'}\Delta_{jk}u}\right)\right)\Delta_{kj}u$$

By construction (see Fig. 1.9), the state value $u_{k'}$ is a positive weighted (convex) combination of $u_{a'}$ and $u_{b'}$ and similarly $u_{j'}$ is a positive weighted combination of u_a and u_b. Focusing on control volume T_j, the stated theorem follows immediately upon the requirement that the discretization for control volume T_j be a positive coefficient discretization, i.e. that the discretization for control volume T_j depends positively on u_j and nonpositively on surrounding solution values subject to a time step restriction. ∎

Some standard limiter functions that satisfy the requirements of Theorem 1.2.12 include

- the MinMod limiter with maximum compression parameter equal to 2

$$\Psi^{\mathrm{MM}}(R) = \max(0, \min(R, 2))$$

- the van Leer limiter

$$\Psi^{\mathrm{VL}}(R) = \frac{R + |R|}{1 + |R|} \ .$$

Other limiter formulations involving three successive one-dimensional slopes are given in [Jam93, CCDD98].

1.2.3 Extension to systems of nonlinear conservation laws

A positive attribute of finite volume methods is the relative ease in which the numerical discretization schemes of Sects. 1.1 and 1.2 can be algorithmically extended to systems of nonlinear conservation laws of the form

$$\begin{aligned} \partial_t u + \nabla \cdot f(u) &= 0 \qquad \text{in } \mathbb{R}^d \times \mathbb{R}^+ \ , & \text{(74a)} \\ u(x,0) &= u_0(x) \quad \text{in } \mathbb{R}^d & \text{(74b)} \end{aligned}$$

where $u(x,t) : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^m$ denotes the vector of dependent solution variables, $f(u) : \mathbb{R}^m \mapsto \mathbb{R}^{m\times d}$ denotes the flux vector, and $u_0(x) : \mathbb{R}^d \to \mathbb{R}^m$ denotes the initial data vector at time $t = 0$. It is assumed that this system is strictly hyperbolic, i.e. the eigenvalues of the flux jacobian $A(\nu; u) \equiv \partial f/\partial u \cdot \nu$ are real and distinct for all bounded $\nu \in \mathbb{R}^d$.

The main task in extending finite volume methods to systems of nonlinear conservation laws is the construction of a suitable numerical flux function. To gain insight into this task, consider the one-dimensional *linear* Cauchy problem for $u(x,t) : \mathbb{R} \times \mathbb{R}^+ \mapsto R^m$ and $u_0(x) : R \mapsto \mathbb{R}^m$

$$\begin{aligned} \partial_t u + \partial_x (A\, u) &= 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}^+ \ , \\ u(x,0) &= u_0(x) \quad \text{in } \mathbb{R} \end{aligned} \qquad (75)$$

where $A \in \mathbb{R}^{m\times m}$ is a *constant* matrix. Assume the matrix A has m real and distinct eigenvalues, $\lambda_1 < \lambda_1 < \cdots < \lambda_m$, with corresponding right and left eigenvectors denoted by $r_k \in \mathbb{R}^m$ and $l_k \in \mathbb{R}^m$ respectively for $k = 1, \ldots, m$. Furthermore, let $X \in \mathbb{R}^{m\times m}$ denote the matrix of right eigenvectors, $X = [r_1, \ldots, r_m]$, and $\Lambda \in \mathbb{R}^{m\times m}$ the diagonal matrix of eigenvalues, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ so that $A = X\Lambda X^{-1}$. The one-dimensional system (75) is readily decoupled into scalar equations via the transformation into characteristic variables $\alpha = X^{-1}u$

$$\partial_t \alpha + \partial_x (\Lambda\, \alpha) = 0 \qquad \text{in } \mathbb{R} \times \mathbb{R}^+ \ ,$$

$$\alpha(x,0) \;=\; \alpha_0(x) \text{ in } \mathbb{R} \tag{76}$$

and component-wise solved exactly

$$\alpha^{(k)}(x,t) = \alpha_0^{(k)}(x-\lambda_k\, t) \;\;, \quad k=1,\ldots,m$$

or recombined in terms of the original variables

$$u(x,t) = \sum_{k=1}^{m} l_k \cdot u_0(x-\lambda_k\, t)\, r_k \;\;.$$

Using this solution, it is straightforward to solve exactly the associated Riemann problem for $w(\xi,\tau) \in R^m$

$$\partial_\tau w + \partial_\xi (A\, w) = 0 \qquad \text{in } \mathbb{R}\times\mathbb{R}^+$$

with initial data

$$w(\xi,0) = \begin{cases} u & \text{if } \xi < 0 \\ v & \text{if } \xi > 0 \end{cases}$$

thereby producing the following Godunov-like numerical flux function

$$\begin{aligned} g(u,v) &= Aw(0,\mathbb{R}^+) \\ &= \frac{1}{2}(Au+Av) - \frac{1}{2}|A|\,(v-u) \end{aligned} \tag{77}$$

with $|A| \equiv X|\Lambda|X^{-1}$. When used in one-dimensional discretization together with piecewise constant solution representation, the linear numerical flux (77) produces the well-known Courant-Isaacson-Rees (CIR) upwind scheme for linear systems of hyperbolic equations

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\left(A^+\,(u_j^n - u_{j-1}^n) + A^-\,(u_{j+1}^n - u_j^n)\right)$$

where $A^\pm = X\Lambda^\pm X^{-1}$. Note that higher order accurate finite volume methods with slope limiting procedures formally extend to this linear system via component wise slope limiting of the characteristic components $\alpha^{(k)}, k=1,\ldots m$ for use in the numerical flux (77).

Numerical flux functions for systems of conservation laws

In Godunov's original work (see Godunov [God59]), exact solutions of the one-dimensional *nonlinear* Riemann problem of gas dynamics were used in the construction of a similar numerical flux function

$$g^{\mathrm{G}}(u,v) = f(w(0,\mathbb{R}^+))\cdot\nu \tag{78}$$

where $w(\xi,\tau) \in \mathbb{R}^m$ is now a solution of a nonlinear Riemann problem

$$\partial_\tau w + \partial_\xi f^{(\nu)}(w) = 0 \qquad \text{in } \mathbb{R}\times\mathbb{R}^+$$

with initial data

$$w(\xi,0) = \begin{cases} u & \text{if } \xi < 0 \\ v & \text{if } \xi > 0 \end{cases} \;\;.$$

Recall that solutions of the Riemann problem for gas dynamic systems are a composition of shock, contact and rarefaction wave family solutions. For the gas dynamic equations considered by Godunov, a unique solution of the Riemann problem exists for general states

u and v except those states producing a vacuum. Even so, the solution of the Riemann problem is both mathematically and computationally nontrivial. Consequently, a number of alternative numerical fluxes have been proposed that are more computationally efficient. These alternative numerical fluxes can be sometimes interpreted as approximate Riemann solvers. A partial list of alternative numerical fluxes is given here. A more detailed treatment of this subject is given in Godlewski and Raviart [GR91], Kröner [Krö97], and LeVeque [LeV02].

- Osher-Solomon flux ([OS82]). This numerical flux is a system generalization of the Enquist-Osher flux of Sect. 1.1. All wave families are approximated in state space as rarefaction or inverted rarefaction waves with Lipschitz continuous partial derivatives. The Osher-Solomon numerical flux is of the form

$$g^{\mathrm{OS}}(u,v) = \frac{1}{2}(f(u)+f(v))\cdot\nu - \frac{1}{2}\int_u^v |A(\nu;w)|\,dw$$

  where $|A|$ denotes the usual matrix absolute value. By integrating on m rarefaction wave integral subpaths that are each parallel to a right eigenvector, a system decoupling occurs on each subpath integration. Furthermore, for the gas dynamic equations with ideal gas law, it is straightforward to construct $m-1$ Riemann invariants on each subpath thereby eliminating the need for path integration altogether. This reduces the numerical flux calculation to purely algebraic computations with special care taken at sonic points, see Osher [OS82].

- Roe flux ([Roe81]). Roe's numerical flux can be a interpreted as approximating all waves families as discontinuities. The numerical flux is of the form

$$g^{\mathrm{Roe}}(u,v) = \frac{1}{2}(f(u)+f(v))\cdot\nu - \frac{1}{2}|A(\nu;u,v)|\,(v-u)$$

  where $A(\nu;u,v)$ is the "Roe matrix" satisfying the matrix mean value identity

$$(f(v)-f(u))\cdot\nu = A(\nu;u,v)\,(v-u)$$

  with $A(\nu;u,u) = A(\nu;u)$. For the equations of gas dynamics with ideal gas law, the Roe matrix takes a particularly simple form. Steady discrete mesh-aligned shock profiles are resolved with one intermediate point. The Roe flux does not preclude the formation of entropy violating expansion shocks unless additional steps are taken near sonic points.

- Steger-Warming flux vector splitting ([SW81]). Steger and Warming considered a splitting of the flux vector for the gas dynamic equations with ideal gas law that exploited the fact that the flux vector is homogeneous of degree one in the conserved variables. From this homogeneity property, Euler's identity then yields that $f(u)\cdot\nu = A(\nu;u)\,u$. Steger and Warming then considered the matrix splitting

$$A = A^+ + A^- \quad , \quad A^\pm \equiv X\Lambda^\pm X^{-1}$$

  where $\Lambda^\pm$ is computed component wise. From this matrix splitting, the final upwind numerical flux function was constructed as

$$g^{\mathrm{SW}}(u,v) = A^+(\nu;u)\,u + A^-(\nu;v)\,v \quad .$$

Although not part of their explicit construction, for the gas dynamic equations with ideal gas law, the jacobian matrix $\partial g^{\mathrm{SW}}/\partial u$ has eigenvalues that are all nonnegative and the jacobian matrix $\partial g^{\mathrm{SW}}/\partial v$ has eigenvalues that are all nonpositive whenever the ratio of specific heats γ lies in the interval $[1, 5/3]$. The matrix splitting leads to numerical fluxes that do not vary smoothly near sonic and stagnation points. Use of the Steger-Warming flux splitting in the schemes of Sect. 1.1 and 1.2 results in rather poor resolution of linearly degenerate contact waves and velocity slip surfaces due to the introduction of excessive artificial diffusion for these wave families.

- Van Leer flux vector splitting. Van Leer [vL82] provided an alternative flux splitting for the gas dynamic equations that produces a numerical flux of the form

$$g^{\mathrm{VL}}(u, v) = f^{-}(u) + f^{+}(v)$$

using special Mach number polynomials to construct fluxes that remain smooth near sonic and stagnation points. As part of the splitting construction, the jacobian matrix $\partial g^{\mathrm{SW}}/\partial u$ has eigenvalues that are all nonnegative and the matrix $\partial g^{\mathrm{SW}}/\partial v$ has eigenvalues that are all nonpositive. The resulting expressions for the flux splitting are somewhat simpler when compared to the Steger-Warming splitting. The van Leer splitting also introduces excessive diffusion in the resolution of linearly degenerate contact waves and velocity slip surfaces.

- System Lax-Friedrichs flux. This numerical flux is the system equation counterpart of the scalar Lax-Friedrichs flux (14). For systems of conservation laws the Lax-Friedrichs flux is given by

$$g^{\mathrm{LF}}(u, v) = \frac{1}{2}(f(u) + f(v)) \cdot \nu - \frac{1}{2}\alpha(\nu)\,(v - u)$$

where $\alpha(\nu)$ is given through the eigenvalues $\lambda_k(\nu; w)$ of $A(\nu; w)$

$$\alpha(\nu) = \max_{1 \le k \le m} \sup_{w \in [u,v]} |\lambda_k(\nu; w)| \ .$$

The system Lax-Friedrichs flux is usually not applied on the boundary of domains since it generally requires an over specification of boundary data. The system Lax-Friedrichs flux introduces a relatively large amount of artificial diffusion when used in the schemes of Sect. 1.1. Consequently, this numerical flux is typically only used together with relatively high order reconstruction schemes where the detrimental effects of excessive artificial diffusion are mitigated.

- Harten-Lax-van Leer flux ([HLvL83]). The Harten-Lax-van Leer numerical flux originates from a simplified two wave model of more general m wave systems such that waves associated with the smallest and largest characteristic speeds of the m wave system are always accurately represented in the two wave model. The following numerical flux results from this simplified two wave model

$$g^{\mathrm{HLL}}(u, v) = \frac{1}{2}(f(u)+f(v))\cdot\nu - \frac{1}{2}\frac{\alpha_{\max}+\alpha_{\min}}{\alpha_{\max}-\alpha_{\min}}(f(v)-f(u))\cdot\nu + \frac{\alpha_{\max}\alpha_{\min}}{\alpha_{\max}-\alpha_{\min}}\,(v-u)$$

where

$$\alpha_{\max}(\nu) = \max_{1 \le k \le m}(0, \sup_{w \in [u,v]} \lambda_k(\nu; w)) \ , \quad \alpha_{\min} = \min_{1 \le k \le m}(0, \inf_{w \in [u,v]} \lambda_k(\nu; w)) \ .$$

When compared to the Lax-Friedrichs flux, this flux can be considerably more accurate in flow situations where $0 < |(\alpha_{\max} + \alpha_{\min})/(\alpha_{\max} - \alpha_{\min})| < 1$. Using this flux, full upwinding is obtained for supersonic flow. Modifications of this flux are suggested in Einfeldt [EMRS92] to improve the resolution of intermediate waves as well.

Further examples of numerical fluxes (among others) include the kinetic flux vector splitting due to Deshpande [Des86], the advection upstream splitting flux (AUSM) of Liou [LS93], and the convective upwind and split pressure (CUSP) flux of Jameson [Jam93, Jam95].

Chapter 2

A Posteriori Error Estimation for Higher Order Godunov Methods

Abstract: In this lecture, *A posteriori* error estimates for high order Godunov finite volume methods are presented which exploit the two solution representations inherent in the method, viz. as piecewise constants u_0 and cellwise p-th order reconstructed functions $R_p^0 u_0$. Using standard duality arguments, an exact error representation formula for user specified functionals is derived that is tailored to the class of high order Godunov finite volume methods with data reconstruction as first described in Barth and Larson [BL02]. From this error representation formula, computable error estimates are then devised that exploit the structure of Godunov finite volume methods. The present theory applies directly to a wide range of finite volume methods based on cellwise reconstruction (see Chapter 1) including MUSCL, TVD, UNO, and ENO methods [vL79, Har83, HOEC87, Har89, SO88, BJ89, BF90, DOE90, Bar98, Abg94, Van93]. Practical issues such as the treatment of nonlinearity and the post-processing of dual problem data are considered. Numerical results using the schemes of Chapter 1 for linear advection and nonlinear conservation laws at steady-state are presented to validate the analysis.

2.1 Overview

A frequent objective in numerically solving partial differential equations is the subsequent calculation of certain derived quantities of particular interest, e.g., aerodynamic lift and drag coefficients, stress intensity factors, mean temperatures, etc. Consequently, there is considerable interest in constructing *a posteriori* error estimates for such derived quantities (mathematically described as functionals) so as to improve the reliability and efficiency of numerical computations. For an introduction to *a posteriori* error analysis see the articles by Becker and Rannacher [BR98], Eriksson et al. [EEHJ95], Giles et al. [GLLS97, GP99], Johnson et al. [JRB95], Parashivoiu et al. [PPP97], Prudhomme and Oden [PO99, OP99], Süli [S̈98], the collected NATO lecture notes [BD02], and a previous version of this work in Barth and Larson [BL02].

This lecture revisits the topic of *a posteriori* error estimation of user prescribed functionals with specific consideration given to finite volume methods that are extensions of Godunov's original method [God59] to high order accuracy via various forms of data reconstruction, e.g. MUSCL in [vL79], TVD in [Har83], UNO in [HOEC87], ENO in [Har89, SO88] with faithful generalizations of Godunov's method to unstructured meshes given in [BJ89, BF90, DOE90, Bar98, Abg94, Van93]. Recall from Chapter 1 that these

methods can be viewed abstractly in the following operator composition form for a first-order conservation law in d space dimensions and time

$$u_0^{n+1} = A \cdot E(\tau) \cdot R_p^0(\cdot) u_0^n \ . \tag{1}$$

In this lecture, u_0^n denotes the space of piecewise constant cell-averages of the conservation law solution $u(x,t)$ at time t_n, $R_p^0(x)$ is a reconstruction operator which produces a cellwise discontinuous p-th order polynomial approximation of the solution given cell-averages, $E(t)$ is the evolution operator for the PDE (including boundary conditions), and A is the cell-averaging operator such that $A|_T$ performs cell-averaging for each control volume T in the mesh $\mathcal{T}$. The requirements of high order accuracy for smooth solutions and discrete conservation give rise to the following additional design criterion for the reconstruction operator (see Harten [HOEC87, Har89])

$$\bullet \ R_p^0(x) u_0 = u(x) + O(h^{p+1}) \text{ whenever } u \text{ is smooth} \tag{2}$$

$$\bullet \ A|_T R_p^0(x) u_0 = u_0|_T, \ \forall T \in \mathcal{T} \text{ to insure discrete conservation} \tag{3}$$

As we will see, it is possible (see Barth and Larson [BL02]) to construct an exact error representation formula and simple *a posteriori* error estimation theory without knowing the precise details of a particular reconstruction operator beyond the requirements of Eqns. (2) and (3). In constructing this *a posteriori* error estimation theory for finite volume methods, it is convenient to utilize the notion of a mesh dependent *broken space* $\mathcal{V}_p^{\mathrm{B}}$ consisting of discontinuous piecewise polynomials of at most degree p in each control volume. Using this space, consider the Discontinuous Galerkin (DG) finite element method introduced by Reed and Hill [RH73] as analyzed by Johnson and Pitkäranta [JP86] and further refined for nonlinear conservation laws by Cockburn et al. [CLS89, CS97]:
DG FEM. Find $u_p \in \mathcal{V}_p^{\mathrm{B}}$ such that

$$\mathcal{B}_{\mathrm{DG}}(u_p, v) = F(v), \quad \forall v \in \mathcal{V}_p^{\mathrm{B}} \tag{4}$$

where $\mathcal{B}_{\mathrm{DG}}(\cdot,\cdot)$ denotes an abstract variational form corresponding to a weak integrated-by-parts form of the conservation law and $F(v)$ a functional possibly including boundary conditions and any external forcing terms. Precise forms of these operators will be given later. It is well-known that in the case $p = 0$, the DG method reduces to the lowest order accurate Godunov finite volume method. As will be shown later, the underpinning of the present error estimation theory comes from the simple observation that the higher order Godunov methods can be expressed as a Petrov-Galerkin variant of the basic DG method:
Higher Order Godunov FVM. Find $u_0 \in \mathcal{V}_0^{\mathrm{B}}$ such that

$$\mathcal{B}_{\mathrm{DG}}(R_p^0 u_0, v) = F(v), \quad \forall v \in \mathcal{V}_0^{\mathrm{B}}, \quad R_p^0 : \mathcal{V}_0^{\mathrm{B}} \mapsto \mathcal{V}_p^{\mathrm{B}} \ . \tag{5}$$

Here R_p^0 represents the same reconstruction operator described in Chapter 1 which maps one broken space into another. Using these constructions, it will be shown that the *a posteriori* error estimation theory previously developed for the DG method can be modified for use in higher order Godunov methods with a modicum of effort by appealing directly to the Petrov-Galerkin form given in Eqn. (5).

Remark 2.1.1 *The idea of abstractly representing finite volume methods as a Petrov-Galerkin variational method has been used previously in a priori error estimates for finite volume methods discretizing elliptic problems in [BR87, Cai91, Sül91, LMV96, VHMD98, Cha99, CL00, Her00, EGH00, ELL02].*

Remark 2.1.2 *Although time dependent terms are present in portions of the presentation and the theory applies to full space-time formulations, the final analysis as well as calculated numerical results will present error estimates for steady-state calculations.*

2.2 Higher Order Godunov Finite Volume Methods in Petrov-Galerkin Form

Let Ω be a domain in $\mathbb{R}^d$ and $\mathcal{T}$ a tessellation of Ω into control volumes, $T \in \mathcal{T}$. Further let $\mathcal{V}^{\rm B}$ be the mesh dependent broken space of discontinuous piecewise Sobolev H^s functions defined on $\mathcal{T}$, i.e.,

$$\mathcal{V}^{\rm B} = \{v \,:\, v|_T \in H^s(T),\ \forall T \in \mathcal{T}\}\ . \tag{6}$$

Similarly, let $\mathcal{V}_p^{\rm B}$ denote the finite dimensional space consisting of discontinuous piecewise polynomial functions of degree p defined on the tessellation $\mathcal{T}$

$$\mathcal{V}_p^{\rm B} = \{v \,:\, v|_T \in \mathcal{P}_p(T),\ \forall T \in \mathcal{T}\} \tag{7}$$

with $\mathcal{P}_p(T)$ the space of polynomials of degree $\leq p$ defined in a control volume T.

Next consider the following prototype scalar nonlinear conservation law in a domain Ω with boundary Γ with solution $u(x,t) : \Omega \times \mathbb{R} \mapsto \mathbb{R}$ and flux vector $f(u) : \mathbb{R} \mapsto \mathbb{R}^d$

$$\begin{aligned} u_{,t} + \mathrm{div} f(u) &= 0, && \text{in } \Omega \times [0,\tau] \\ u(x,0) &= u_0(x), && \text{in } \Omega \\ a^-(n;u)(g-u) &= 0, && \text{on } \Gamma \ \text{ with } a(n;u) \equiv f_{,u}\cdot n\ . \end{aligned}$$

Let I_n denote the time slab increment, $I_n \equiv [t_n, t_{n+1}]$, with $[0,\tau] = \cup_{n=0,N-1} I_n$. In addition, let T and T' denote two control volumes adjacent to an interface e so that $u_\pm(\partial T \cap e)$ denotes the trace restrictions of functions on that interface segment such that $u_-(x)$ is the restriction from T and $u_+(x)$ is the restriction from T' for $x \in e$. Using this compact notation, the Godunov finite volume method and discontinuous Galerkin method for a single time slab increment are written succinctly as

Godunov Finite Volume. Find $u_0 \in \mathcal{V}_0^{\rm B}$ such that for each $T \in \mathcal{T}$

$$\frac{d}{dt}\int_{I_n} u_0|_T\, dt + \int_{I_n\times\partial T\setminus\Gamma} h(n;(R_p^0 u_0)_-,(R_p^0 u_0)_+)\, ds\, dt + \int_{I_n\times\partial T\cap\Gamma} h(n;(R_p^0 u_0)_-, g)\, ds\, dt = 0 \tag{8}$$

Discontinuous Galerkin. Find $u_p \in \mathcal{V}_p^{\rm B}$ for all $v \in \mathcal{V}_p^{\rm B}$ (implied sum on i)

$$\begin{aligned} \sum_{T\in\mathcal{T}}\Big(\int_{I_n\times T} v\ (u_p)_{,t}\, dx\, dt - \int_{I_n\times T} v_{,x_i}\ f^i(u_p)\, dx\, dt + \int_{I_n\times\partial T\setminus\Gamma} v_-\, h(n;(u_p)_-,(u_p)_+)\, ds\, dt \\ + \int_{I_n\times\partial T\cap\Gamma} v_-\, h(n;(u_p)_-, g)\, ds\, dt\Big) = 0 \end{aligned} \tag{9}$$

where $h(n;u_-,u_+)$ is a numerical flux function such that $f(u)\cdot n = h(n;u,u)$ and $h(n;u_-,u_+) = -h(-n;u_+,u_-)$. In these formulations, we have omitted (for sake of simplicity) those terms that would arise from discontinuous *in time* approximation since our final objective here are error estimates at steady-state. Also observe that Eqn. (9) is consistent with our abstract variational representation given earlier for DG in Eqn. (4)

Find $u_p \in \mathcal{V}_p^{\rm B}$ *such that*

$$\mathcal{B}_{\rm DG}(u_p, v) = F(v), \quad \forall v \in \mathcal{V}_p^{\rm B}\ . \tag{10}$$

Close comparison of Eqns. (8) and (9) suggests the following lemma of importance in *a posteriori* error estimation for Godunov finite volume methods.

Lemma 2.2.1 *Let R_p^0 denote a reconstruction operator $R_p^0 : \mathcal{V}_0^{\mathrm{B}} \mapsto \mathcal{V}_p^{\mathrm{B}}$ on a nondeforming space-time tessellation $\mathcal{T} \times I_n$ satisfying the cell-averaging condition for $u_0 \in \mathcal{V}_0^{\mathrm{B}}$ and all $T \in \mathcal{T}$*

$$(R_p^0 u_0, v)|_T = (u_0, v)|_T, \quad \forall v \in \mathcal{V}_0^{\mathrm{B}} \tag{11}$$

where $(\cdot,\cdot)|_T$ denotes an inner product integration on Ω restricted to a control volume T. The Godunov finite volume method (8) is written equivalently as the following Petrov-Galerkin variant of the discontinuous Galerkin method (9):
Find $u_0 \in \mathcal{V}_0^{\mathrm{B}}$

$$\mathcal{B}_{\mathrm{DG}}(R_p^0 u_0, v) = F(v), \quad \forall v \in \mathcal{V}_0^{\mathrm{B}} \ . \tag{12}$$

Proof: The proof follows immediately from term-by-term inspection of Eqns. (8) and (9) together with the cell-averaging condition (11). ∎

Remark 2.2.2 *Observe that the cell-averaging condition given here in Eqn. (11) is identical to that given earlier in Eqn. (3).*

2.3 *A Posteriori* Error Estimation of Functionals

Using Lemma 2.2.1, an exact error representation formula and computable *a posteriori* error estimates will be derived for user specified functionals tailored to Godunov finite volume methods as first described in Barth and Larson [BL02]. The development given here follows closely the previous work of Becker and Rannacher [BR98] and Süli [SÖ98] as well as previous *a posteriori* error estimation work by the present author in [Bar99, BL99] for the DG method.

2.3.1 Functionals

The objective is to estimate the error in a user specified functional $M(u)$ which can be expressed as a weighted integration over the domain Ω

$$M_\psi(u) = \int_\Omega \psi \, N(u) \, dx$$

or a weighted integration on the boundary Γ

$$M_\psi(u) = \int_\Gamma \psi \, N(u) \, dx$$

for some user specified ψ and function $N(u) : \mathbb{R} \mapsto \mathbb{R}$. Examples of functionals used in later calculations are:
Example 1: Outflow functional, $u_{,t} + \lambda \cdot \nabla u = 0$

$$M_\psi(u) = \int_\Gamma \psi \, (\lambda \cdot n)^+ \, u \, dx, \quad x \in \mathbb{R}^d \ . \tag{13}$$

Example 2: Solution average functional

$$M_{\mathrm{ave}}(u) = \int_\Omega u \, dx, \quad x \in \mathbb{R}^d \ . \tag{14}$$

Example 3: Mollified pointwise functional

$$M_\psi(u) = \int_\Omega \psi(r_0; |x - x_0|)\, u\, dx, \quad x \in \mathbb{R}^2 \tag{15}$$

$$\psi(r_0; r) = \begin{cases} 0 & r \geq r_0 \\ \frac{e^{1/(r^2/r_0^2 - 1)}}{2\pi \int_0^{r_0} e^{1/(\xi^2/r_0^2 - 1)} \xi\, d\xi} & r < r_0 \end{cases} .$$

2.3.2 Error Representation Formulas

In this section, exact error representation formulas are derived for three abstract formulations with
(1) $B(\cdot,\cdot)$ a bilinear form with $M(\cdot)$ a linear functional
(2) $\mathcal{B}(\cdot,\cdot)$ a semilinear form (nonlinear in the first argument and linear in the second argument) with $\mathcal{M}(\cdot)$ a nonlinear functional
(3) $\mathcal{B}(R_p^0 \cdot, \cdot)$ a nonlinear semilinear form (nonlinear in the first argument and linear in the second argument) with $\mathcal{M}(\cdot)$ a nonlinear functional
In these derivations, π_p denotes any suitable projection operator (e.g. interpolation, L_2 projection) into $\mathcal{V}_p^{\mathrm{B}}$.

Theorem 2.3.1 (Galerkin error representation, $B(\cdot,\cdot)$ bilinear and $M(\cdot)$ linear) *Let $B(\cdot,\cdot)$ denote a bilinear form and $M(\cdot)$ a linear functional. Assume the finite-dimensional primal numerical problem*

Find $u_p \in \mathcal{V}_p^{\mathrm{B}}$ such that

$$B(u_p, v) = F(v) \quad \forall\ v \in \mathcal{V}_p^{\mathrm{B}} \ , \tag{16}$$

and the infinite-dimensional auxiliary dual problem

Find $\Phi \in \mathcal{V}^{\mathrm{B}}$ such that

$$B(v, \Phi) = M(v) \quad \forall\ v \in \mathcal{V}^{\mathrm{B}} \ .$$

Then, the numerical error in a linear functional $M(u) - M(u_p)$ is given by the following error representation formula:

$$M(u) - M(u_p) = F(\Phi - \pi_p \Phi) - B(u_p, \Phi - \pi_p \Phi) \ . \tag{17}$$

Proof: An exact error representation formula for a given functional $M(\cdot)$ results from the following steps:

$$\begin{aligned} M(u) - M(u_p) &= M(u - u_p) && \text{(linearity of } M\text{)} \\ &= B(u - u_p, \Phi) && \text{(dual problem)} \\ &= B(u - u_p, \Phi - \pi_p \Phi) && \text{(orthogonality)} \\ &= B(u, \Phi - \pi_p \Phi) - B(u_p, \Phi - \pi_p \Phi) && \text{(linearity of } B\text{)} \\ &= F(\Phi - \pi_p \Phi) - B(u_p, \Phi - \pi_p \Phi) && \text{(variational problem)} \end{aligned}$$

∎

Next, we consider a semilinear form $\mathcal{B}(\cdot,\cdot)$ (nonlinear in the first argument and linear in the second argument) and a nonlinear functional $\mathcal{M}(\cdot)$. To cope with nonlinearity, it is convenient to introduce the mean value linearizations

$$\begin{aligned}\mathcal{B}(u,v) &= \mathcal{B}(u_p,v) + \overline{\mathcal{B}}(u_p,u;u-u_p,v) \quad \forall\ v \in \mathcal{V}^{\mathrm{B}} \\ \mathcal{M}(u) &= \mathcal{M}(u_p) + \overline{\mathcal{M}}(u_p,u;u-u_p)\ .\end{aligned}$$

For example, if $\mathcal{B}(u,v) = (Lu,v)$ for some nonlinear differential operator L then for $v \in \mathcal{V}^{\mathrm{B}}$

$$\begin{aligned}\mathcal{B}(u,v) &= \mathcal{B}(u_p,v) + \Big(\int_0^1 L_{,u}(\tilde{u}(\theta))\,d\,\theta\ (u-u_p), v\Big) \\ &= \mathcal{B}(u_p,v) + (\overline{L}_{,u}\,(u-u_p), v) \\ &= \mathcal{B}(u_p,v) + \overline{\mathcal{B}}(u_p,u;u-u_p,v).\end{aligned}$$

with $\tilde{u}(\theta) \equiv u_p + (u-u_p)\ \theta$. A simple error representation formula then results for nonlinear Galerkin variational forms.

Theorem 2.3.2 (Galerkin error representation, $\mathcal{B}(\cdot,\cdot)$ semilinear and $\mathcal{M}(\cdot)$ nonlinear) *Let $\mathcal{B}(\cdot,\cdot)$ denote a semilinear form and $\mathcal{M}(\cdot)$ a nonlinear functional. Assume the finite-dimensional primal numerical problem*

Find $u_p \in \mathcal{V}_p^{\mathrm{B}}$ such that

$$\mathcal{B}(u_p,v) = F(v) \quad \forall\ v \in \mathcal{V}_p^{\mathrm{B}} \tag{18}$$

and the infinite-dimensional auxiliary mean value linearized dual problem

Find $\Phi \in \mathcal{V}^{\mathrm{B}}$ such that

$$\overline{\mathcal{B}}(u_p,u;v,\Phi) = \overline{\mathcal{M}}(u_p,u;v) \quad \forall\ v \in \mathcal{V}^{\mathrm{B}}\ . \tag{19}$$

Then, the numerical error in a nonlinear functional $\mathcal{M}(u) - \mathcal{M}(u_p)$ is given by the following error representation formula:

$$\mathcal{M}(u) - \mathcal{M}(u_p) = F(\Phi - \pi_p\Phi) - \mathcal{B}(u_p, \Phi - \pi_p\Phi)\ . \tag{20}$$

Proof: An exact error representation formula for a given nonlinear functional $\mathcal{M}(\cdot)$ then results from the following steps

$$\begin{aligned}\mathcal{M}(u) - \mathcal{M}(u_p) &= \overline{\mathcal{M}}(u_p,u;u-u_p) && \text{(mean value } \mathcal{M}) \\ &= \overline{\mathcal{B}}(u_p,u;u-u_p,\Phi) && \text{(dual problem)} \\ &= \overline{\mathcal{B}}(u_p,u;u-u_p,\Phi-\pi_p\Phi) && \text{(orthogonality)} \\ &= \mathcal{B}(u,\Phi-\pi_p\Phi) - \mathcal{B}(u_p,\Phi-\pi_p\Phi) && \text{(mean value } \mathcal{B}) \\ &= F(\Phi-\pi_p\Phi) - \mathcal{B}(u_p,\Phi-\pi_p\Phi), && \text{(variational problem)}\end{aligned}$$

∎

Remark 2.3.3 *Note that although Eqns. (17) and (20) appear identical, mean value linearization introduces a subtle right-hand side dependency on the exact solution in Eqn. (20). This complication is addressed in Sect. 2.4.2.*

Next, we consider Godunov finite volume methods with $\mathcal{B}(R_p^0\cdot,\cdot)$ semilinear and $\mathcal{M}(\cdot)$ nonlinear. Mean value linearizations are again introduced

$$\begin{aligned}\mathcal{B}(u,v) &= \mathcal{B}(R_p^0 u_0, v) + \overline{\mathcal{B}}(R_0^p u_0, u; u - R_p^0 u_0, v) \quad \forall\ v \in \mathcal{V}^{\mathrm{B}} \\ \mathcal{M}(u) &= \mathcal{M}(R_p^0 u_0) + \overline{\mathcal{M}}(R_p^0 u_0, u; u - R_p^0 u_0)\ .\end{aligned}$$

Theorem 2.3.4 (Godunov finite volume error representation, $\mathcal{B}(R_p^\cdot,\cdot)$ semilinear and $\mathcal{M}(\cdot)$ nonlinear, [BL02]) *Let $\mathcal{B}(\mathbb{R}_p^0\cdot,\cdot)$ denote a Godunov finite volume semilinear form and $\mathcal{M}(\cdot)$ a nonlinear functional. Assume the finite-dimensional primal Godunov finite volume numerical problem*

Find $u_0 \in \mathcal{V}_0^{\mathrm{B}}$ such that

$$\mathcal{B}(R_p^0 u_0, v) = F(v) \quad \forall\ v \in \mathcal{V}_0^{\mathrm{B}} \tag{21}$$

and the infinite-dimensional auxiliary mean value linearized dual problem

Find $\Phi \in \mathcal{V}^{\mathrm{B}}$ such that

$$\overline{\mathcal{B}}(R_p^0 u_0, u; v, \Phi) = \overline{\mathcal{M}}(v) \quad \forall\ v \in \mathcal{V}^{\mathrm{B}}\ . \tag{22}$$

Then, the numerical error in a nonlinear functional $\mathcal{M}(u) - \mathcal{M}(R_p^0 u_0)$ is given by the following error representation formula:

$$\mathcal{M}(u) - \mathcal{M}(R_p^0 u_0) = F(\Phi - \pi_0\Phi) - \mathcal{B}(R_p^0 u_0, \Phi - \pi_0\Phi)\ . \tag{23}$$

Proof: An exact error representation formula for a given nonlinear functional $\mathcal{M}(\cdot)$ for the class of Godunov finite volume methods results from the following steps

$$\begin{aligned}\mathcal{M}(u) - \mathcal{M}(R_p^0 u_0) &= \overline{\mathcal{M}}(u - R_p^0 u_0) && \text{(mean value } \mathcal{M}\text{)} \\ &= \overline{\mathcal{B}}(u - R_p^0 u_0, \Phi) && \text{(dual problem)} \\ &= \overline{\mathcal{B}}(u - R_p^0 u_0, \Phi - \pi_0\Phi) && \text{(orthogonality)} \\ &= \mathcal{B}(u, \Phi - \pi_0\Phi) - \mathcal{B}(R_p^0 u_0, \Phi - \pi_0\Phi) && \text{(mean value } \mathcal{B}\text{)} \\ &= F(\Phi - \pi_0\Phi) - \mathcal{B}(R_p^0 u_0, \Phi - \pi_0\Phi), && \text{(Godunov FV problem)}\end{aligned}$$

∎

This final form for the Godunov finite volume method serves as a progenitor for the remaining derivations given below.

2.4 Computable Error Estimates

Computationally, the error representation formulas (17), (20) and (23) are not suitable for obtaining computable *a posteriori* error estimates and use in mesh adaptation.

- $\Phi \in \mathcal{V}^B$, the solution of the infinite-dimensional problem is not generally known.
- The mean value linearization used in the linearized dual problems (19) and (22) requires knowledge of the exact solution u.
- The error representation formulas do not suggest any simple strategy for control volume refinement/coarsening.

2.4.1 Approximating $\Phi - \pi_0\Phi$

We list several strategies for approximating $\Phi - \pi_0\Phi$ for Godunov finite volume methods. The first two techniques seek to exploit the two scale structure of Godunov methods, i.e. that as a weighted residual method of Petrov-Galerkin type, the residual is orthogonal to test functions in $\mathcal{V}_0^{\mathrm{B}}$ and not to test functions in $\mathcal{V}_p^{\mathrm{B}}$.
Inherent two scale approximation. Compute the linearized dual problem:
Find $\Phi_0 \in \mathcal{V}_0^{\mathrm{B}}$ *such that*

$$\overline{\mathcal{B}}(R_p^0 u_0, u; v, R_p^0\Phi_0) = \overline{\mathcal{M}}(R_p^0 u_0, u; v), \quad \forall\, v \in \mathcal{V}_0^{\mathrm{B}}$$

and approximate

$$\Phi - \pi_0\Phi \approx R_p^0\Phi_0 - \Phi_0 \ . \tag{24}$$

Remark 2.4.1 *This strategy fails in standard Galerkin finite element methods since any approximation of* $\Phi \in \mathcal{V}_p^{\mathrm{B}}$ *is orthogonal to the residual, hence with Galerkin finite element methods the contribution is identically zero and no error estimate is obtained.*

Patch recovery post-processing. Compute the linearized dual problem:
Find $\Phi_0 \in \mathcal{V}_0^{\mathrm{B}}$ *such that*

$$\overline{\mathcal{B}}(R_p^0 u_0, u; v, R_p^0\Phi_0) = \overline{\mathcal{M}}(R_p^0 u_0, u; v), \quad \forall\, v \in \mathcal{V}_0^{\mathrm{B}}$$

and approximate using a patch recovery technique $\overline{R}_q^p : \mathcal{V}_p^{\mathrm{B}} \mapsto \mathcal{V}_q^{\mathrm{B}}$ for $q \geq p$

$$\Phi - \pi_0\Phi \approx \overline{R}_q^p R_p^0\Phi_0 - \Phi_0 \ . \tag{25}$$

The patch recovery is motivated by the original work of Zienkiewicz and Zhu [ZZ92]. In the present computations, the least squares reconstruction operator discussed in Section 2.6 is also used as a patch recovery operator so that

$$\overline{R}_q^p R_p^0 u_0 = R_q^0 u_0 \ . \tag{26}$$

Global higher order solves. Solve the linearized dual problem global using a higher order method:
Find $\Phi_0 \in \mathcal{V}_0^{\mathrm{B}}$ *such that*

$$\overline{\mathcal{B}}(R_p^0 u_0, u; v, R_q^0\Phi_0) = \overline{\mathcal{M}}(R_p^0 u_0, u; v), \quad \forall\, v \in \mathcal{V}_0^{\mathrm{B}}$$

for some $q > p$. While conceptually straightforward, this technique typically makes solving the linearized dual problem more computationally expensive than the primal problem in terms of computer memory and arithmetic operations. This can be prohibitive in three space dimensions.

2.4.2 Approximating the Mean Value Linearized Dual Problem

The mean value linearization requires knowledge of the exact solution u. Two computable approximate linearizations are considered
Jacobian derivative linearization. The mean value linearization is supplanted by the Jacobian linearization so that the computable linearized dual problem for the Godunov method is obtained

Find $\Phi_0 \in \mathcal{V}_0^{\mathrm{B}}$ *such that*

$$\overline{\mathcal{B}}(R_p^0 u_0, R_p^0 u_0; v, \Phi_0) = \overline{M}(R_p^0 u_0, R_p^0 u_0; v) \quad \forall\ v \in \mathcal{V}^{\mathrm{B}} \ . \tag{27}$$

Mean value linearization via post-processing and numerical quadrature. The Godunov FV method provides easy access to post-processed approximations of the solution, i.e. $\overline{R}_q^p R_p^0 u_0$ as $R_q^0 u_0$ for $q > p$, thus suggesting the improved computable approximation of the mean value linearized dual problem

Find $\Phi_0 \in \mathcal{V}_0^{\mathrm{B}}$ *such that for* $q > p$

$$\overline{\mathcal{B}}(R_p^0 u_0, R_q^0 u_0; v, \Phi_0) = \overline{M}(R_p^0 u_0, R_q^0 u_0; v) \quad \forall\ v \in \mathcal{V}^{\mathrm{B}} \tag{28}$$

where numerical quadrature (i.e. trapezoidal quadrature) is employed to approximate the mean value path integration.

2.4.3 Direct Estimates

Given the error representation formula (23) for the Godunov finite volume method, error estimates suitable for adaptive meshing are easily obtained

$$\begin{aligned}
|\mathcal{M}(u) - \mathcal{M}(R_p^0 u_0)| &= \left|\mathcal{B}(R_p^0 u_0, \Phi - \pi_0\Phi) - F(\Phi - \pi_0\Phi)\right| \quad \text{(error representation)} \\
&= |\sum_{T\in\mathcal{T}} \big(\mathcal{B}_T(R_p^0 u_0, \Phi - \pi_0\Phi) - F_T(\Phi - \pi_0\Phi)\big)| \qquad \text{(element assembly)} \\
&\le \sum_{T\in\mathcal{T}} \left|\big(\mathcal{B}_T(R_p^0 u_0, \Phi - \pi_0\Phi) - F_T(\Phi - \pi_0\Phi)\big)\right| \qquad \text{(triangle inequality)}
\end{aligned} \tag{29}$$

where $\mathcal{B}_T(\cdot,\cdot)$ and $F_T(\cdot)$ are restrictions of $\mathcal{B}(\cdot,\cdot)$ and $F(\cdot)$ to the control volume T.

Note that the element assembly representation is not unique. For example strong and weak forms of the variational operator $\mathcal{B}(\cdot,\cdot)$ yield differing assembly representations. For the Godunov finite volume method with time terms omitted, the error representation formula (23) yields

$$\begin{aligned}
\mathcal{B}(R_p^0 u_0, \Phi - \pi_0\Phi) - F(\Phi - \pi_0\Phi) = \sum_{T\in\mathcal{T}} \Big(&- \int_T f^i(R_p^0 u_0)\,(\Phi - \pi_0\Phi)_{,x_i}\,dx \\
&+ \int_{\partial T\setminus\Gamma} (\Phi - \pi_0\Phi)_-\, h(n; (R_p^0 u_0)_-, (R_p^0 u_0)_+)\,ds \\
&+ \int_{\partial T\cap\Gamma} (\Phi - \pi_0\Phi)_-\, h(n; (R_p^0 u_0)_-, g)\,ds \Big).
\end{aligned} \tag{30}$$

The present numerical computations utilize the numerical flux formula

$$h(n; u_-, u_+) = \frac{1}{2}\left(f(n; u_-) + f(n; u_+)\right) - \frac{1}{2}|a(n; \overline{u}(u_-, u_+))|\,[u]_-^+ \tag{31}$$

with $\overline{u}(u_-, u_+)$ chosen so that

$$[f(n;u)]_-^+ = a(n; \overline{u}(u_-, u_+))\,[u]_-^+ \tag{32}$$

with $f(n;u) = (f(u_-)\cdot n)$ and $a(n;u) = \partial f(n;u)/\partial u$. Using this particular numerical flux, the following weighted residual (strong) form can be obtained upon integration by parts

$$\begin{aligned}\mathcal{B}(R_p^0 u_0, \Phi - \pi_0\Phi) - F(\Phi - \pi_0\Phi) = \sum_{T\in\mathcal{T}} \Big(&\int_T (\Phi - \pi_0\Phi)\ \mathrm{div} f(R_p^0 u_0)\, dx \\ &+ \int_{\partial T\setminus\Gamma} (\Phi - \pi_0\Phi)_-\, a^-(n; (R_p^0 u_0)_-, (R_p^0 u_0)_+)\, [R_p^0 u_0]_-^+\, ds \\ &+ \int_{\partial T\cap\Gamma} (\Phi - \pi_0\Phi)_-\, a^-(n; (R_p^0 u_0)_-, g)\, (g - (R_p^0 u_0)_-)\, ds \Big). \end{aligned} \tag{33}$$

This latter weighted residual form and the implied element assembly form $\sum_T \mathcal{B}_T(\cdot,\cdot) - F_T(\cdot)$ is preferred in the error estimates (29) since the individual terms represent residual components that vanish individually when the exact solution is inserted into the variational form and a slightly sharper approximation is obtained after application of the triangle inequality in (29).

2.5 Adaptive Meshing

The error estimates of the previous section motivate a simple strategy for mesh adaptation. Defining for each control volume T

$$\eta_T \equiv \mathcal{B}_T(R_p^0 u_0, \Phi - \pi_0\Phi) - F_T(\Phi - \pi_0\Phi) \tag{34}$$

we have a candidate *adaptation element indicator* $|\eta_T|$ such that

$$\big|\mathcal{M}(u) - \mathcal{M}(R_p^0 u_0)\big| \le \sum_{T\in\mathcal{T}} |\eta_T| \tag{35}$$

and an accurate *adaptation stopping criteria*

$$\big|\mathcal{M}(u) - \mathcal{M}(R_p^0 u_0)\big| = |\sum_{T\in\mathcal{T}} \eta_T| \ . \tag{36}$$

These quantities suggest a simple mesh adaptation strategy in common use with other indicator functions:

Mesh Adaptation Algorithm

(1) Construct an initial mesh $\mathcal{T}$.
(2) Compute a numerical approximation of the primal problem on the current mesh $\mathcal{T}$ using Godunov's method with p-th order reconstruction yielding $R_p^0 u_0$.
(3) Compute a numerical approximation of the dual problem on the current mesh $\mathcal{T}$ using Godunov's method with p-th order reconstruction yielding $R_p^0 \Phi_0$.
(4) Optionally improve the accuracy of the numerically computed dual problem via a post-processing recovery operator $\overline{R}_q^p$ for $q \ge p$ yielding $\overline{R}_q^p R_p^0 \Phi_0$.
(5) Compute η_T for all control volumes in $\mathcal{T}$ using $R_p^0 u_0$ and the approximation

$$\Phi - \pi_0\Phi \approx R_p^0\Phi_0 - \Phi_0 \quad \text{or} \quad \Phi - \pi_0\Phi \approx \overline{R}_q^p R_p^0 \Phi_0 - \Phi_0 \ .$$

(6) If($|\sum_{T\in\mathcal{T}} \eta_T| < TOL$) STOP
(7) Otherwise, refine and coarsen a user specified fraction of the total number of control volumes according to the size of $|\eta|_T$, generate a new mesh $\mathcal{T}$ and GOTO 2

2.6 Least Squares Reconstruction on Patches

The reconstruction operator used in calculations is based on a least squares approximation given cell-averages on patches of control volumes. Let $\mathcal{N}(T) \subset \Omega$ denote a patch of control volumes containing the control volume T. The global reconstruction operator $R_p^0 : \mathcal{V}_0^{\mathrm{B}} \to \mathcal{V}_p^{\mathrm{B}}$ is constructed piecewise on a local patch-by-patch basis with

$$(R_{p,\mathcal{N}(T)}^0 u_0)|_T = (R_p^0 u_0)|_T, \quad \forall T \in \mathcal{T}$$

for $u_0 \in \mathcal{V}_0^{\mathrm{B}}$ so that the task reduces to that of finding the local patch reconstruction operator $R_{p,\mathcal{N}(T)}^0$ for each $T \in \mathcal{T}$. To do so, first define the L_2 projection $\Pi_0 : \mathcal{V}^{\mathrm{B}} \mapsto \mathcal{V}_0^{\mathrm{B}}$, i.e. for each $u \in \mathcal{V}^{\mathrm{B}}$

$$(u - \Pi_0 u, v) = 0, \quad \forall v \in \mathcal{V}_0^{\mathrm{B}} \ .$$

The local reconstruction operator $R_{p,\mathcal{N}(T)}^0$ is then constructed from the following two conditions

1. **Exact Π_0 projection in T.** The $\Pi_{0,T}$ projection of $R_{p,\mathcal{N}(T)}^0 u_0$ is exact in T, i.e., it holds that
$$\Pi_{0,T} R_{p,\mathcal{N}(T)}^0 u_0 = u_{0,T} \quad \text{for each } u_0 \in \mathcal{V}_0^{\mathrm{B}}. \tag{37}$$
where $\Pi_{0,T}$ and $u_{0,T}$ denote restrictions of Π_0 and u_0 to the control volume T. This condition is equivalent to the cell-averaging property given in Eqn. (3).

2. **Constrained least squares fitting on patch $\mathcal{N}(T)$.** The L_2 deviation of the $\Pi_{0,T'}$ projection of $R_{p,\mathcal{N}(T)}^0 u_0$ from given cell-averaged data in patch control volumes $T' \in \mathcal{N}(T)$ is minimized subject to the constraint (37)
$$\|u_0 - \Pi_0 R_{p,\mathcal{N}(T)}^0 u_0\|_{\mathcal{N}(T)} = \min_{w \in \mathcal{Q}_p(\mathcal{N}(T))} \|u_0 - \Pi_0 w\|_{\mathcal{N}(T)}, \tag{38}$$
for all $u_0 \in \mathcal{V}_0^{\mathrm{B}}$. Here $\mathcal{Q}_p(\mathcal{N}(T))$ is the subspace of polynomials in $\mathcal{P}_p(\mathcal{N}(T))$ such that (37) holds.

Remark 2.6.1 (p-th Order Exactness) *Note that the patch cardinality* $\mathrm{card}(\mathcal{N}(T))$ *is always chosen sufficiently large (e.g. by increasing graph distance) so that there exists a unique solution to the constrained least squares problem and the local reconstruction operator $R_{p,\mathcal{N}(T)}^0$ is fully determined. As a consequence, it follows that for $r \leq p$*

$$R_{p,\mathcal{N}(T)}^0 \Pi_0 u_r = u_r \quad \forall u_r \in \mathcal{P}_r(\mathcal{N}(T)),$$

which simply asserts that if the given data u_r is in the space of polynomials of degree $r \leq p$ in the patch neighborhood $\mathcal{N}(T)$, then the cellwise projection to cell-averages followed by p-th order patch reconstruction exactly reproduces the given data.

2.7 Slope Limiting for Discontinuous Solutions

For solutions containing discontinuities such as the Burgers' equation example of Sect. 2.8, a slope limiter is employed following the analysis of Chapter 1 so that non-oscillatory solutions are obtained. The following particular solution ansatz for each $T \in \mathcal{T}$

$$U(x)_T \equiv u_{0,T} + \Psi_T \cdot (R_p^0(x) u_0 - u_0)_T \tag{39}$$

is chosen with $\Psi_T \in [0,1]$ so that the cell-average property of the reconstruction (37) is maintained regardless of the particular value of Ψ_T. Next for each $T \in \mathcal{T}$ compute the minimum and maximum of all adjacent neighbor cell-averages

$$u_T^{\min} = \min_{T' \in \mathcal{N}(T)} u_{0,T'}, \quad u_T^{\max} = \max_{T' \in \mathcal{N}(T)} u_{0,T'}$$

and determine the largest value of $\Psi_T \in [0,1]$ such that

$$u_T^{\min} \leq U(x)_T \leq u_T^{\max}$$

when evaluated at the quadrature points used in the flux integral computation. To achieve this, compute the extrapolated state $U(x_q)$ at each quadrature point location x_q in the flux integral and determine the most restrictive Ψ_T

$$\Psi_T = \begin{cases} \min(1, \frac{u_T^{\max} - u_{0,T}}{U(x_q) - u_{0,T}}), & \text{if } U(x_q) - u_{0,T} > 0 \\ \min(1, \frac{u_T^{\min} - u_{0,T}}{U(x_q) - u_{0,T}}), & \text{if } U(x_q) - u_{0,T} < 0 \\ 1 & \text{if } U(x_q) - u_{0,T} = 0 \end{cases} \tag{40}$$

across all quadrature points. Unfortunately, the convergence of nonlinear iterative methods can be erratic using this type of non-differentiable limiter. Consequently, an additional quadratic dissipation term is added to the numerical flux function for discontinuous solution problems to enhance the convergence of nonlinear iterative methods

$$\begin{aligned} h^{\text{mod}}(n, \epsilon; (R_p^0 u_0)_-, (R_p^0 u_0)_+) &= h(n; (R_p^0 u_0)_-, (R_p^0 u_0)_+) \\ &+ \epsilon \sup_{\Omega} |f'| \left(\frac{[R_p^0 u_0]_-^+}{[u_0]_-^+} \right)^2 [u_0]_-^+ \end{aligned} \tag{41}$$

with $\epsilon = .01$ used in the Burgers' equation calculations given below.

2.8 Numerical Results for Scalar Conservation Laws

To validate and assess the *a posteriori* error estimation theory for the Godunov finite volume method, numerical solutions for linear advection and nonlinear Burgers' equation are computed.

Linear Advection. $u(x,y) : [0,1]^2 \mapsto \mathbb{R}$ with $\lambda = (-y, x)^T$.

$$\begin{aligned} \text{div}(\lambda u) &= 0, \quad \text{in } [0,1]^2 \\ u(x,0) &= g(x), \\ u(1,y) &= 0, \end{aligned}$$

with inflow profile data

$$g(x) = \begin{cases} \widetilde{\psi}(9/20; |x - 1/2|) \cdot (1 - \widetilde{\psi}(9/20; |x - 1/20|)) & \text{if } x \leq 1/2 \\ \widetilde{\psi}(9/20; |x - 1/2|) \cdot (1 - \widetilde{\psi}(9/20; |x - 19/20|)) & \text{if } x > 1/2 \end{cases}$$

where $\widetilde{\psi}(r;x) \equiv \psi(r;x)/\psi(r;0)$ and $\psi(r;x)$ the mollifier function defined in Eqn. 15. Figure 2.1 (left) shows isocontours of the primal numerical solution obtained using the Godunov method with linear reconstruction on a relatively fine mesh containing 6400 simplicies. Figure 2.1 (right) graphs global measures of the solution error on meshes

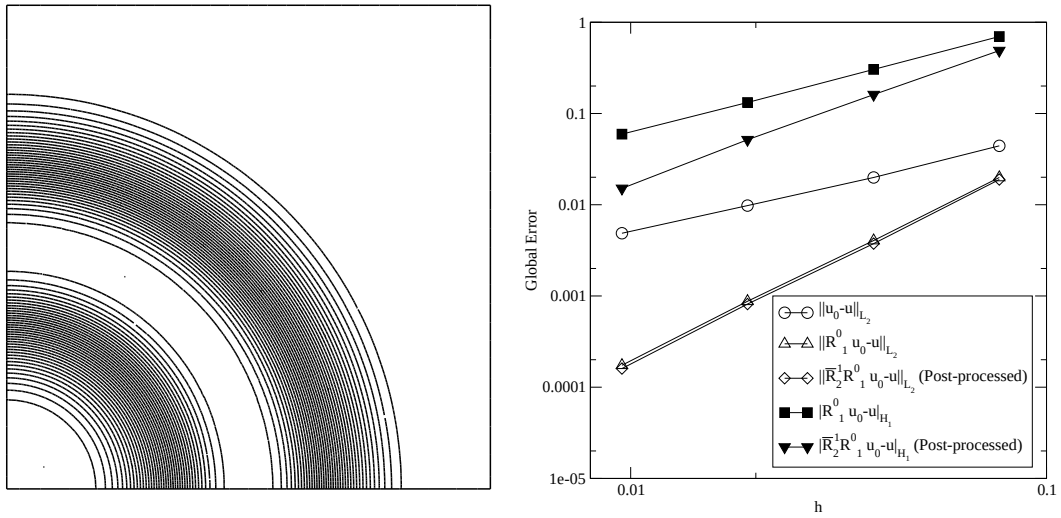


Figure 2.1: Circular advection problem. Isocontours of the primal numerical solution $R_1^0 u_0$ computed using the Godunov FVM method with linear reconstruction on a mesh containing 6400 simplicies (left) and global measures of solution error versus mesh spacing parameter h (right).

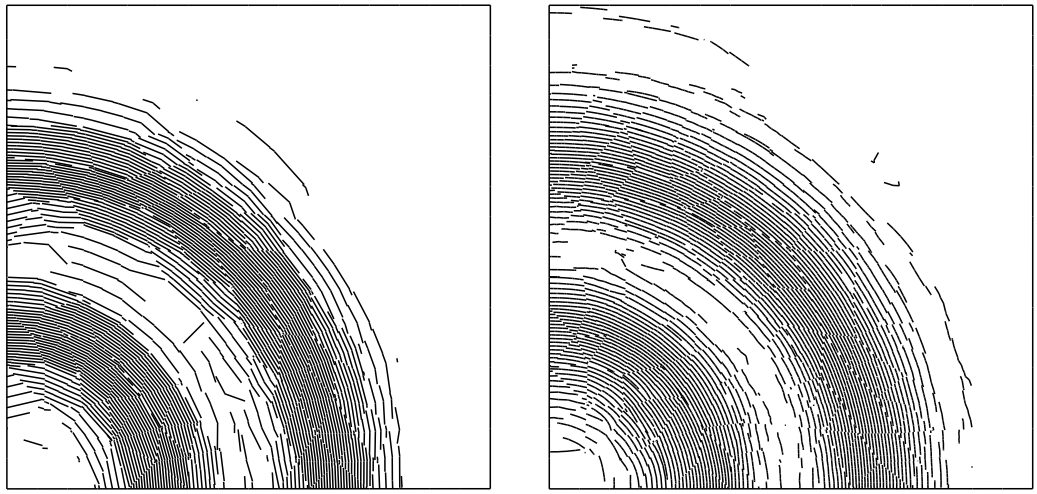


Figure 2.2: Isocontours of the coarsest mesh (400 simplicies) numerical solution $R_1^0 u_0$ obtained using the Godunov FVM with linear reconstruction (left) and isocontours of the post-processed solution $\overline{R}_2^1 R_1^0 u_0 = R_2^0 u_0$ (right).

containing 400, 1600, 6400 and 25600 simplices. The graphs show that the L_2 solution error in cell-averages $\|u_0 - u\|_{L_2}$ as well as the H_1 semi-norm of the linearly reconstructed solution $|R_1^0 u_0 - u|_{H_1}$ are both first order accurate as expected while the L_2 solution error in the linearly reconstructed solution $\|R_1^0 u_0 - u\|_{L_2}$ exceeds second order accuracy. Also included in these graphs is the effect of solution post-processing. Specifically, the quadratic post-processing recovery procedure $\overline{R}_2^1 R_1^0 u_0 = R_2^0 u_0$ was employed. Although the L_2 norm $\|R_2^0 u_0 - u\|_{L_2}$ shows no improvement in convergence rate (slope), a slight vertical shift in the graph of this data indicates a slight improvement in absolute accuracy that is somewhat obfuscated by the logarithmic scaling. More conspicuous is the improvement in the H_1 semi-norm. The effect of post-processing is to increase the convergence rate of $|R_2^0 u_0 - u|_{H_1}$ to second order. This indicates significant improvement in the accuracy of derivative information through least squares post-processing. This improvement is

visually seen in Fig. 2.2 which shows isocontours of the numerical solution obtained on the coarsest mesh and the effect of quadratic post-processing. Next, the *a posteriori* error estimates are evaluated. Specifically considered are

- The outflow functional Eqn. (13) with

$$\psi_{\text{outflow}}(x,y) = \begin{cases} \widetilde{\psi}(7/20;|y-3/5|)\cdot(1-\widetilde{\psi}(7/20;|y-1/4|)) & \text{if } y \le 3/5 \\ \widetilde{\psi}(7/20;|y-3/5|)\cdot(1-\widetilde{\psi}(7/20;|y-19/20|)) & \text{if } y > 3/5 \end{cases} .$$

- The solution average functional Eqn. (14).

- The mollified pointwise value functional Eqn. (15) with

$$\psi_{\text{mollified}}(x,y) = \widetilde{\psi}(1/20;\sqrt{(x-1/10)^2+(y-3/5)^2}) \ .$$

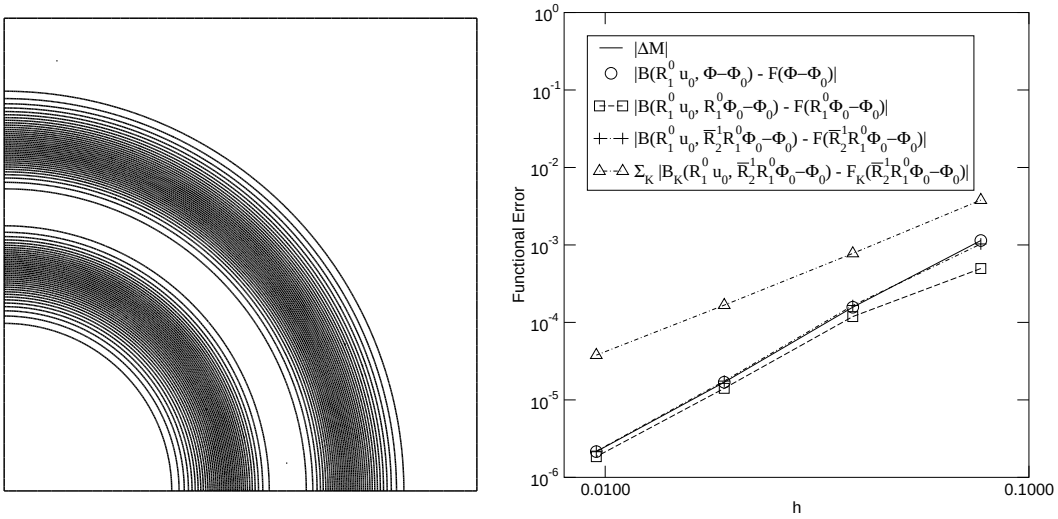


Figure 2.3: Outflow functional for the pure advection problem using the Godunov FVM with linear reconstruction. Isocontours of the dual problem solution (left) and functional error versus mesh parameter h (right).

Dual solutions and error estimates are shown in Figs. 2.3-2.5. Each of these figures shows isocontours for the numerical solution of the dual problem (left) and graphs of the functional error and the *a posteriori* error estimates (right). The graphs in Figs. 2.3-2.5 each contain five curves. The first curve depicts the exact functional error $|\Delta M|$ since the analytic primal solution is known. Observe the third order superconvergence in the outflow functional and the mollified pointwise value functional. For each functional the continuous dual solution Φ can be constructed either exactly via Green's function or to a specified accuracy using series expansion and/or adaptive quadrature. Using this Φ, the second curve depicts $|B(R_1^0 u_0, \Phi - \pi_0\Phi) - F(\Phi - \pi_0\Phi)|$ which according to (23) should be identical to the first curve $|\Delta M|$. This is verified for each functional. Curve number three graphs $|B(R_1^0 u_0, R_1^0\Phi_0 - \Phi_0) - F(R_1^0\Phi_0 - \Phi_0)|$ so that the effect of numerically approximating the continuous dual problem is assessed. Some noticable error is observed on coarse meshes but the performance on the finest meshes is quite acceptable. Curve number four shows the effect of post-processing of the numerically obtained dual data $|B(R_1^0 u_0, R_2^0\Phi_0 - \Phi_0) - F(R_2^0\Phi_0 - \Phi_0)|$. Using this post-processed dual data, good accuracy is obtained on all

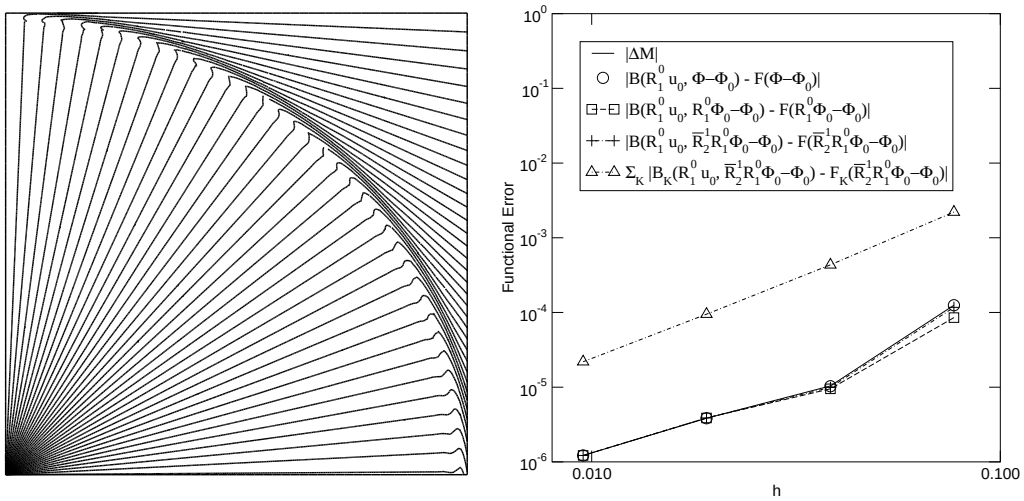


Figure 2.4: Integral solution average functional for the pure advection problem using the Godunov FVM with linear reconstruction. Isocontours of the dual problem solution (left) and functional error versus mesh parameter h (right).

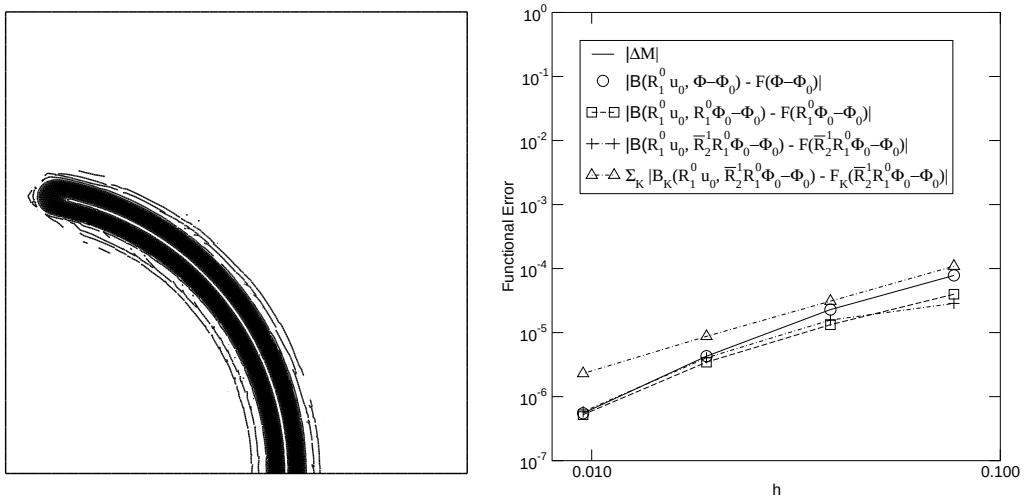
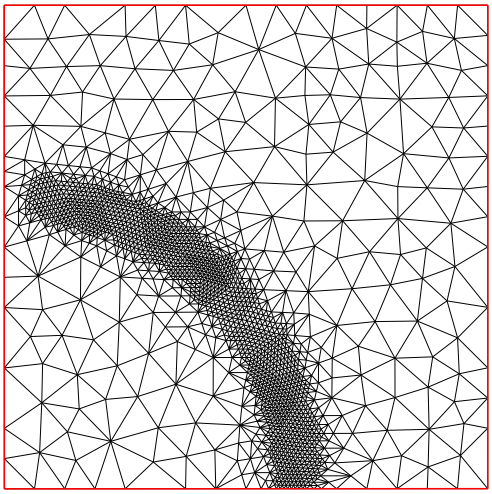


Figure 2.5: Mollified pointwise value functional for the pure advection problem using the Godunov FVM with linear reconstruction. Isocontours of the dual problem solution (left) and functional error versus mesh parameter h (right).

meshes for all functionals except perhaps the mollified pointwise value functional. In this latter case, the dual solution consists of a slightly smoothed ridge function that is not well-resolved on the coarsest meshes using linear or quadratic approximations. Even so, the estimates in curves three or four seem acceptable as an adaptive mesh stopping criteria. Curve number five graphs $\sum_{T\in\mathcal{T}} |B_T(R_1^0 u_0, R_2^0\Phi_0 - \Phi_0) - F_T(R_2^0\Phi_0 - \Phi_0)|$ for each functional. In this approximation, interelement error cancellation does not occur because of the application of the triangle inequality in Eqn. (29). Consequently, the third order superconvergence rates seen in the outflow and mollified functionals is absent and only second order rates of convergence are observed for all functionals. In addition, this last estimate over predicts the true error by factors of 3-1000 depending of the functional and the mesh size. Finally, in Fig. 2.6 (left) an adapted mesh obtained for the mollified pointwise value functional is plotted. The mesh has been adapted using the algorithm of Sect. 2.5 with quadratically post-processed numerical dual data. Figure 2.6 (right)

indicates the efficiency of the adaptation procedure by tabulating levels of adaptation, mesh sizes, and the functional error for each level of adaptation.



Adaptation Levels	#cells	$\lvert\Delta M\rvert$
0	400	1.9E-3
1	602	9.4E-4
2	1232	1.7E-5
3	3418	5.8E-8

Figure 2.6: Adapted mesh for mollified pointwise value functional (left) and tabulated mesh sizes for increasing levels of refinement (right).

Burgers' Equation. $u(x,y) : [0,1]^2 \mapsto \mathbb{R}$ with $\lambda = (u/2, 1)^T$.

$$\begin{aligned} \mathrm{div}(\lambda\, u) &= 0, \quad \text{in } [0,1]^2 \\ u(x,0) &= 5/4 - 2x, \\ u(1,y) &= -3/4, \\ u(0,y) &= 5/4 \ . \end{aligned}$$

As a final example, Burgers' equation is solved in a unit square as shown in Fig. 2.7. As mentioned earlier, the Jacobian linearization with post-processing is used as an approximation of the mean value linearization for the dual problem. Unfortunately, the limiter function Ψ_T in Eqn. (39) is highly non-differentiable and has not been linearized in the present computations. For this problem, error estimates for the solution average functional Eqn. (14) have been obtained. Isocontours of the dual solution are shown in Fig. 2.8 (left). It is observed that monotonicity of the primal shock profile is essential for obtaining meaningful numerical approximations of the dual problem. Figure 2.8 (right) graphs the functional error using various approximations. The first curve graphs the exact functional error $|\Delta M|$. Observe that this functional converges at a first order rate presumably due to the first order accuracy of the primal scheme in the shock region. The second curve graphs $|\mathcal{B}(R_1^0 u_0, \Phi - \pi_0\Phi) - F(\Phi - \pi_0\Phi)|$ but using an analytical Φ linearized about the exact solution. Consequently, this quantity only approximates $|\Delta M|$ and large differences are seen on the coarsest mesh. The third curve graphs $|\mathcal{B}(R_1^0 u_0, R_1^0\Phi_0 - \Phi_0) - F(R_1^0\Phi_0 - \Phi_0)|$ using a numerically approximated dual problem. Again large discrepancies are seen on the coarsest mesh. With mesh refinement the accuracy quickly becomes acceptable. Note we have not included post-processing of the dual data in these calculations. The fourth curve graphs $\sum_{T\in\mathcal{T}} |\mathcal{B}_T(R_1^0 u_0, R_1^0\Phi_0 - \Phi_0) - F_T(R_1^0\Phi_0 - \Phi_0)|$. The results show that this estimate over predicts the true error by an order of magnitude but show the same rate of convergence as the true error.

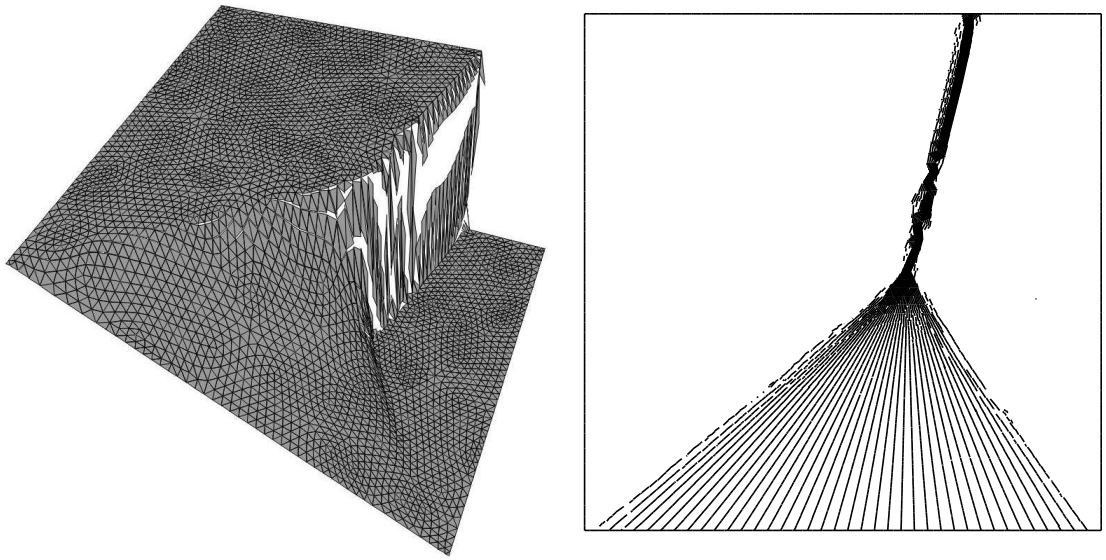


Figure 2.7: Primal numerical solution $R_1^0 u_0$ for Burgers' equation problem using the Godunov FVM with linear reconstruction. Carpet plot in 3D (left) and solution isocontours in 2D (right).

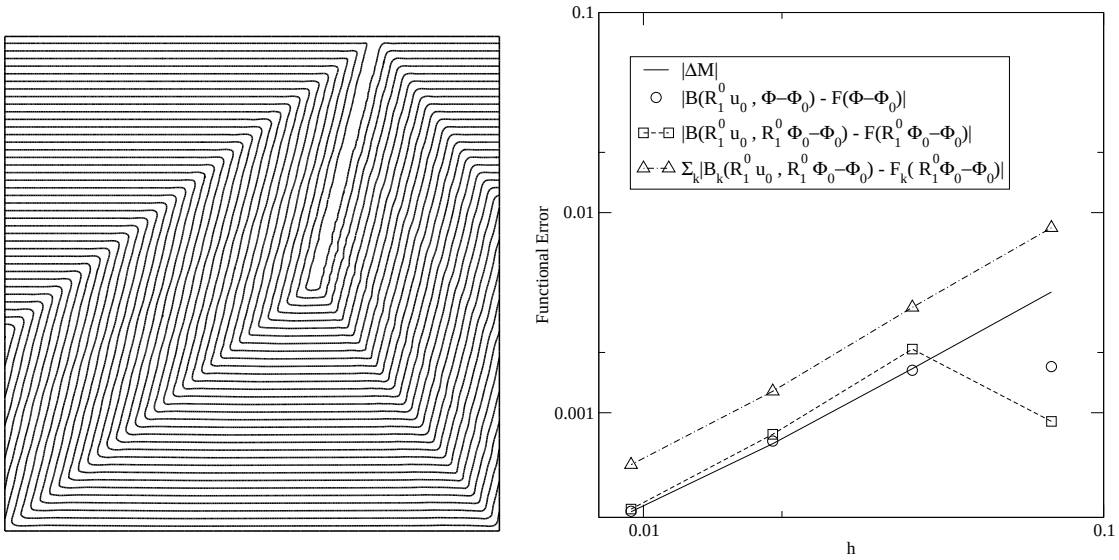


Figure 2.8: Integral solution average functional for Burgers' equation problem. Isocontours of the dual problem (left) and functional error versus mesh spacing parameter h (right).

2.8.1 Numerical Results for the Euler Equations

In the remainder of this section, further numerical results are computed for the Euler equations of gasdynamics using the finite volume scheme with linear reconstruction outlined in the previous section together with Roe flux.

Multi-element Airfoil Flow: To evaluate the accuracy of the Godunov finite volume error representation formula (23), Ringleb flow (an exact transonic solution of the 2-D Euler equations obtained via hodograph transformation, see [Chi85]) is computed in the

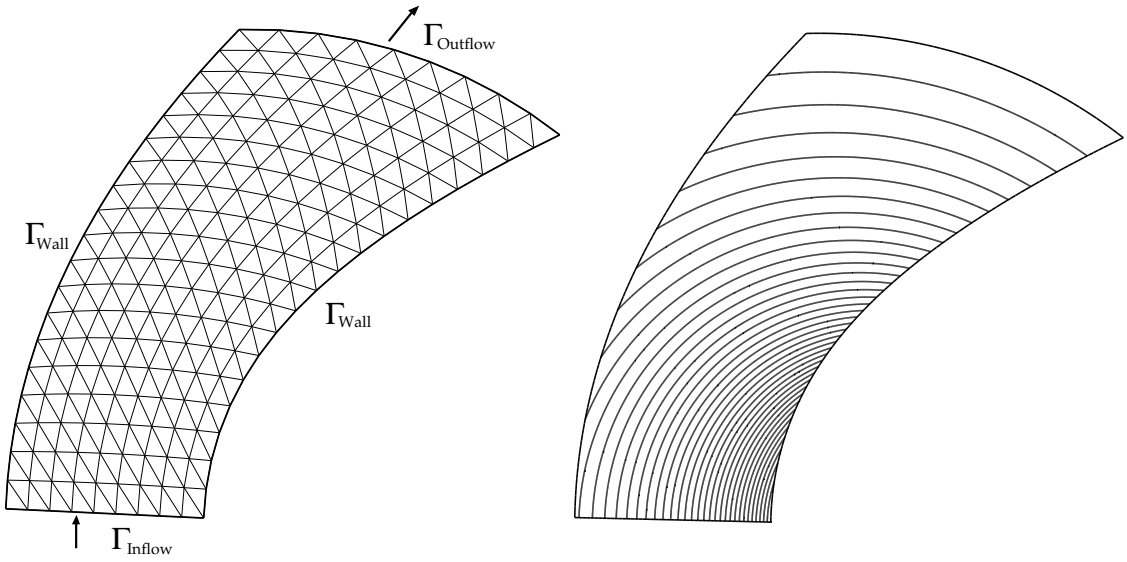


Figure 2.9: Ringleb channel geometry. Coarse mesh showing channel geometry with inflow, outflow and channel walls (left) and density contours of the primal solution (left).

channel geometry shown in Fig. 2.9. To test the error representation formula, the vertical force component exerted on the channel walls is computed from the functional

$$M_\Psi(u) = \int_{\Gamma_{wall}} (\mathbf{n} \cdot \psi)\ Pressure(u)\ d\,x\ , \quad \Psi = (0,1)^T\ .$$

Figure 2.9 shows density contours for the Ringleb solution. Note that the flow is actually supersonic for a small region including the right inflow boundary.

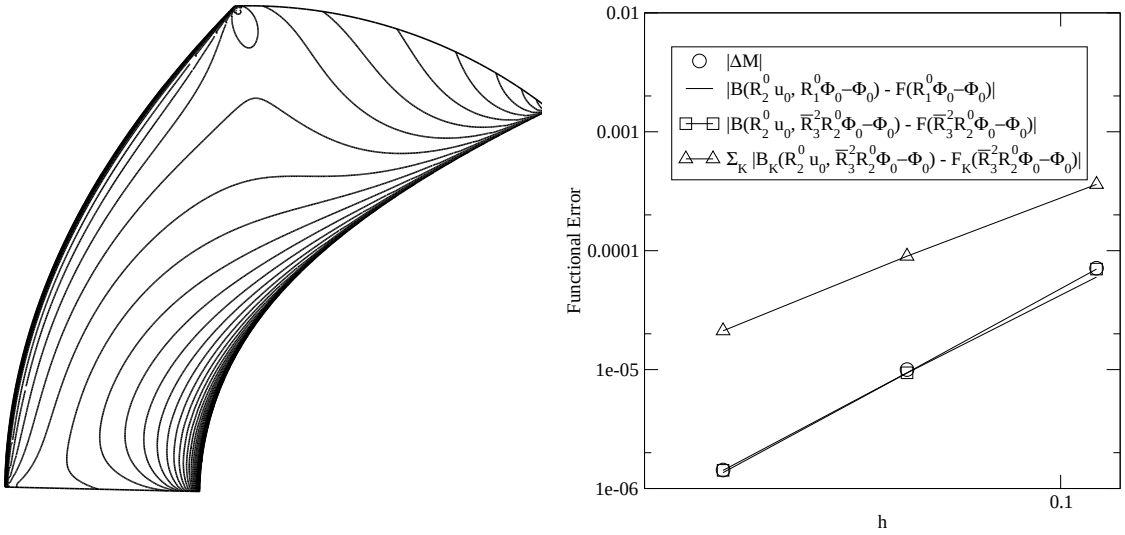


Figure 2.10: Ringleb channel geometry. Contours of the dual density solution (left) and functional errors (right) for the vertical force component functional.

Next, the linearized dual problem is computed using the Jacobian linearization (27) and the error representation formulas computed. Figure 2.10 shows the dual solution for the vertical force functional and the resulting computed errors. The graphs show essentially the same features seen for scalar conservation laws, i.e. the dual solution is adequately computed using the same order method as the primal problem with some small improvement using postprocessing. Application of the triangle inequality in constructing the adaptation indicators yields a loss in accuracy of about one order magnitude. To illustrate the use of these indicators in adaptation, the mollified delta functional ($x_0 = (-.64, 1.70)^T$, $rad = .05$) for the energy component of the solution has been implemented

$$M_\delta(u) = \int_\Omega \text{Energy}(u)\ \delta(x - x_0)\ d\,x \ , \quad x_0 = (-.63, 1.70)^T \ .$$

Using this functional, the corresponding dual problem has been computed and the mesh adapted using the adaptation algorithm of Sect. 2.5. Figure 2.11 shows the resulting dual solution and adapted mesh with three levels of refinement (15000 triangles). The adapted

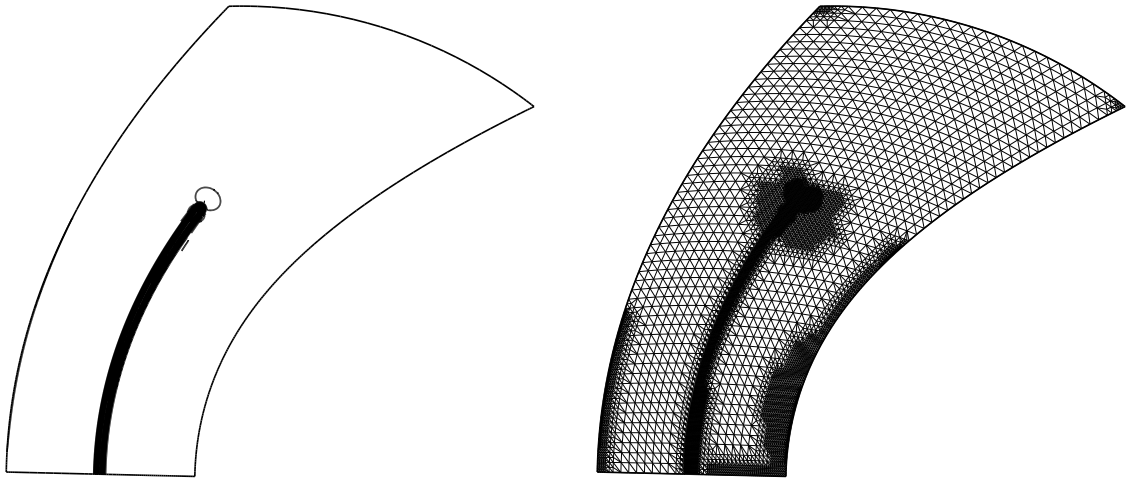


Figure 2.11: Ringleb channel geometry. Dual energy isocontours solution corresponding to mollified delta functional (left) and final adapted mesh (3 levels) (right).

mesh shows the upstream dependence on the numerical solution similar to that observed for the pure advection problem discussed earlier.

Multi-element Airfoil Flow: In the next example, subsonic $M = .13$ Euler flow over a multiple component airfoil geometry in computed. In this example, the lift force functional is used:

$$M_\Psi(u) = \int_{\Gamma_{wall}} (\mathbf{n} \cdot \psi)\ Pressure(u)\ d\,x \ , \quad \Psi \perp V_\infty \ .$$

Figure (2.12) show Mach number contours of the primal solution and dual solution contours associated with the x-component of the momentum. The dual solution, contains rather large singularities at the trailing edges of each element. Figure 2.13 shows the initial and final adapted mesh (3 levels refinement). The functional error during each level of refinement in shown in Fig. 2.14. In estimating the errors used in this figure, a mesh

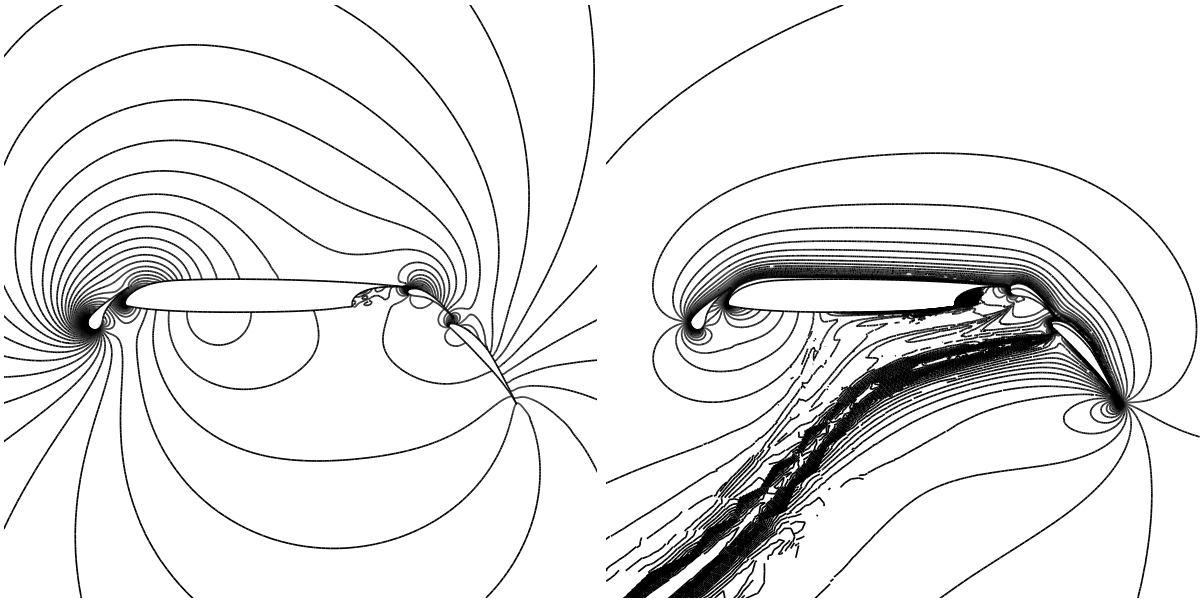


Figure 2.12: Multi-element airfoil geometry, $M_\infty = .13, \alpha = 15.0°$. Isomach contours of primal solution (left) and corresponding contours of the dual x-momentum solution (right) for the lift functional.

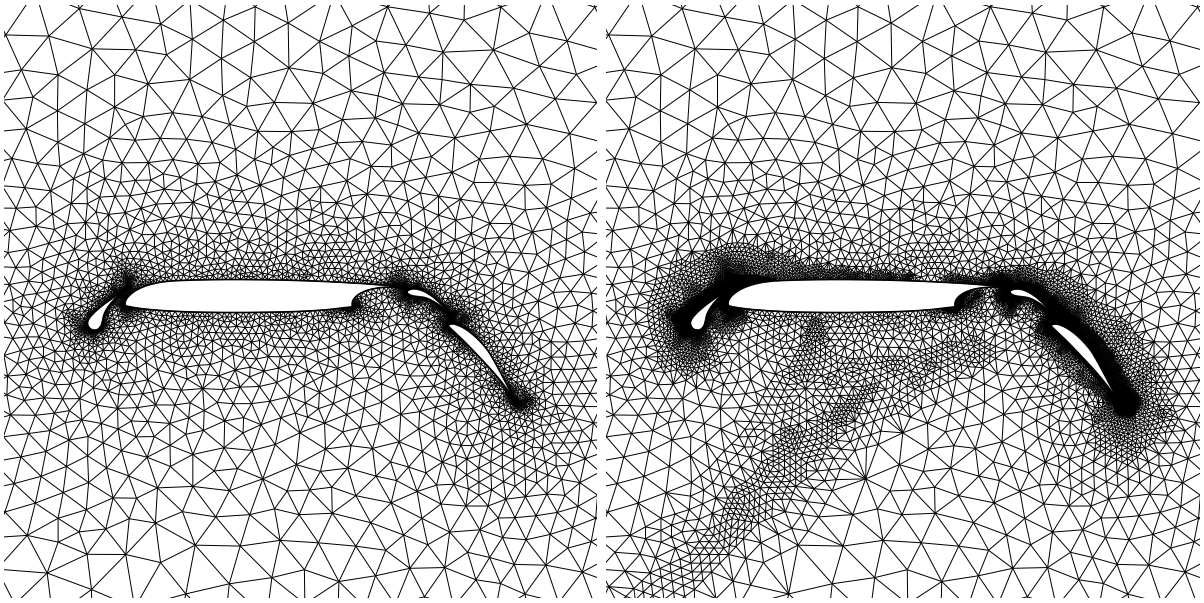


Figure 2.13: Multiple component airfoil meshes. Original mesh with 10k triangles and final adapted mesh with 50k triangles (right).

containing over 100k triangle elements has been used to estimate the mesh asympotic lift value.

Transonic Airfoil Flow: As a final example, transonic Euler flow ($M\infty = .85, \alpha = 2.0°$) is computed over the NACA 0012 airfoil geometry. One again, the lift functional is chosen for evaluation.

$$M_\Psi(u) = \int_{\Gamma_{wall}} (\mathbf{n} \cdot \psi)\ Pressure(u)\ d\,x \ , \quad \Psi \perp V_\infty \ .$$

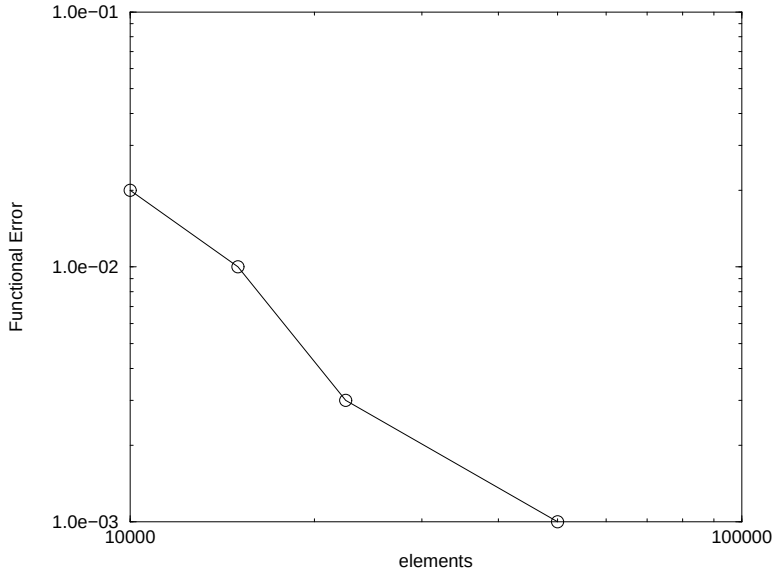


Figure 2.14: Multiple component airfoil problem. Functional error versus number of triangle elements during intermediate levels of refinement.

Figure 2.15 shows isodensity contours of the primal solution and dual iso-density contours of the dual solution corresponding to the lift functional. As also noted for the multiple component airfoil, this dual solution contains a singularity at the trailing edge of the airfoil. In addition, pronounced structures in the dual solution emanate from the leading edge stagnation point and base of the upper and lower shockwaves. These structures signify the sensitivity of the lift force to the location of the stagnation point and shockwaves. The final adapted mesh with 3 levels of refinement shown in Fig. 2.16 shows

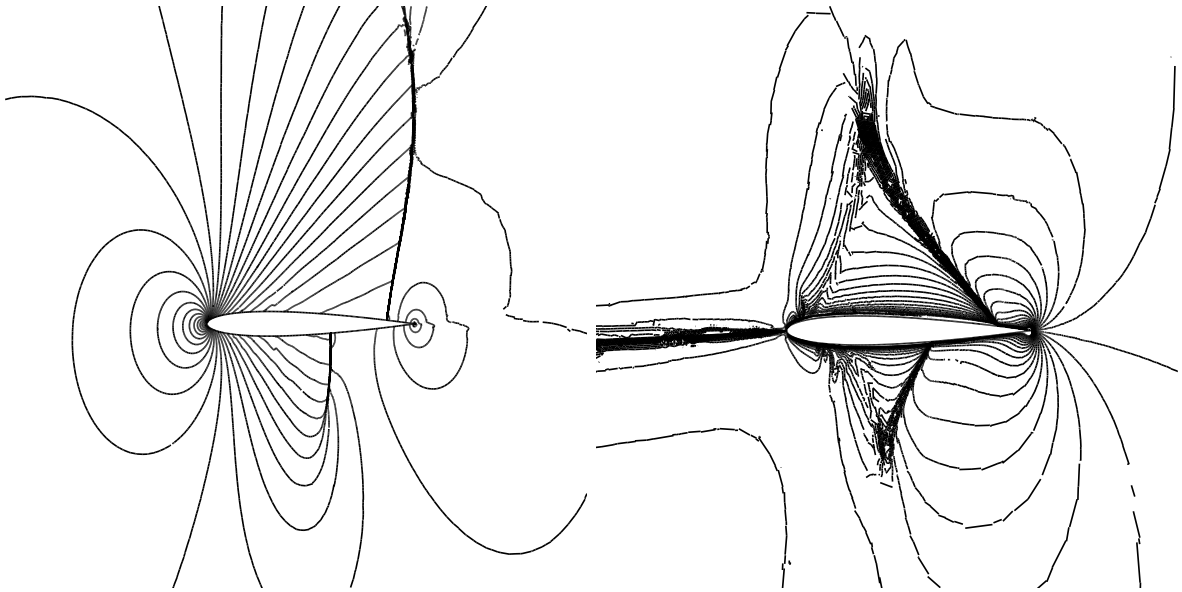


Figure 2.15: NACA airfoil geometry, $M_\infty = .85, \alpha = 2.0°$. Isodensity contours of primal solution (left) and corresponding contours of the dual density solution (right) for the lift functional.

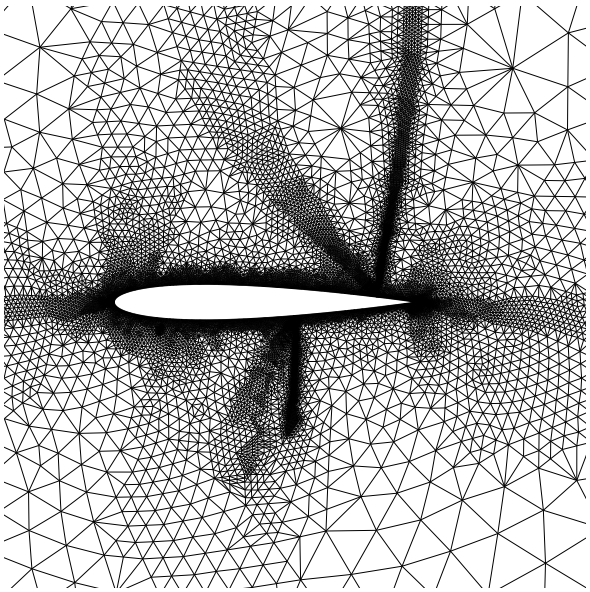


Figure 2.16: NACA airfoil, $M_\infty = .85, \alpha = 2.0°$. Final adapted mesh (3 levels refinement).

mesh refinement not achievable from heuristic methods utilizing either local gradient or local residual information.

Bibliography

[Abg94] R. Abgrall. An essentially non-oscillatory reconstruction procedure on finite-element type meshes. *Comp. Meth. Appl. Mech. Engrg.*, 116:95–101, 1994.

[Bar98] T.J. Barth. Numerical methods for gasdynamic systems on unstructured meshes. In Kröner, Ohlberger, and Rohde, editors, *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws*, volume 5 of *Lecture Notes in Computational Science and Engineering*, pages 195–285. Springer-Verlag, Heidelberg, 1998.

[Bar99] T.J. Barth. Simplified discontinuous Galerkin methods for systems of conservation laws with convex extension. In Cockburn, Karniadakis, and Shu, editors, *Discontinuous Galerkin Methods*, volume 11 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Heidelberg, 1999.

[BB73] J.P. Boris and D.L. Book. Flux corrected transport: Shasta, a fluid transport algorithm that works. *J. Comp. Phys.*, 11:38–69, 1973.

[BD02] T.J. Barth and H. Deconinck(eds). Error estimation and adaptive discretization methods in CFD. In Barth and Deconinck, editors, *Error Estimation and Adaptive Discretization Methods in CFD*, volume 25 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Heidelberg, 2002.

[BF90] T. J. Barth and P.O. Frederickson. Higher order solution of the Euler equations on unstructured grids using quadratic reconstruction. Technical Report 90-0013, AIAA, Reno, NV, 1990.

[BJ89] T. J. Barth and D. C. Jespersen. The design and application of upwind schemes on unstructured meshes. Technical Report 89-0366, AIAA, Reno, NV, 1989.

[BL99] T.J. Barth and M.G. Larson. A posteriori error estimation for adaptive discontinuous Galerkin approximations of hyperbolic systems. Technical Report NAS-99-010, NASA Ames Research Center, 1999.

[BL02] T.J. Barth and M.G. Larson. A-posteriori error estimation for higher order Godunov finite volume methods on unstructured meshes. In Herbin and Kröner, editors, *Finite Volumes for Complex Applications III*, pages 41–63. Hermes Science Pub., London, 2002.

[BLC96] P. Batten, C. Lambert, and D.M. Causon. Positively conservative high-resolution convection schemes for unstructured elements. *Int. J. Numer. Meth. Engrg.*, 39:1821–1838, 1996.

[BPPS87] V. Billey, J. Périaux, P. Perrier, and B. Stoufflet. 2-d and 3-d Euler computations with finite element methods in aerodynamics. volume 1270 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1987.

[BR87] R. Bank and D.J. Rose. Some error estimates for the box method. *SIAM J. Numer. Anal.*, 24:777–787, 1987.

[BR98] R. Becker and R. Rannacher. Weighted a posteriori error control in FE methods. In *Proc. ENUMATH-97, Heidelberg*. World Scientific Pub., Singapore, 1998.

[Cai91] Z. Cai. On the finite volume element method. *Numer. Math.*, 58:713–735, 1991.

[CCDD98] P.-H. Cournède, C., Debiez, and A. Dervieux. A positive MUSCL scheme for triangulations. Technical Report 3465, Institut National De Recherche En Informatique Et En Automatique (INRIA), 1998.

[CCL94] B. Cockburn, F. Coquel, and P.G. Lefloch. An error estimate for finite volume methods for multidimensional conservation laws. *Math. Comput.*, 63:77–103, 1994.

[Cha99] P. Chatzipantelidis. A finite volume method based on the crouzeix-raviart element for elliptic problems. *Numer. Math.*, 82:409–432, 1999.

[Chi85] G. Chiocchia. Exact solutions to transonic and supersonic flows. Technical Report AR-211, AGARD, 1985.

[CL00] S.H. Chou and Q. Li. Error estimates in l^2, h^1 and l^∞ in covolume methods for elliptic and parabolic problems: a unified approach. *Math. Comp.*, 69:103–120, 2000.

[CLS89] B. Cockburn, S.Y. Lin, and C.W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One dimensional systems. *J. Comp. Phys.*, 84:90–113, 1989.

[CM80] M. G. Crandall and A. Majda. Monotone difference approximations for scalar conservation laws. *Math. Comp.*, 34:1–21, 1980.

[CP84] P. Colella and P. Woodward P. The piecewise parabolic methods for gas-dynamical simulations. *J. Comp. Phys.*, 54:174–201, 1984.

[CS97] B. Cockburn and C.W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems. Technical Report 201737, Institite for Computer Applications in Science and Engineering (ICASE), NASA Langley R.C., 1997.

[DD88] J. Desideri and A. Dervieux. Compressible flow solvers using unstructured grids, March 1988. von Karman Institute Lecture Series 1988-05.

[Del96] M. Delanaye. *Polynomial Reconstruction Finite Volume Schemes for the Compressible Euler and Navier-Stokes Equations on Unstructured Adaptive Grids*. PhD thesis, University of Liége, Belgium, 1996.

[Des86] S. M. Deshpande. On the Maxwellian distribution, symmetric form, and entropy conservation for the Euler equations. Technical Report TP-2583, NASA Langley, Hampton, VA, 1986.

[DOE90] L. Durlofsky, S. Osher, and B. Engquist. Triangle based TVD schemes for hyperbolic conservation laws. Technical Report 90-10, Institite for Computer Applications in Science and Engineering (ICASE), NASA Langley R.C., 1990.

[EEHJ95] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. Introduction to numerical methods for differential equations. *Acta Numerica*, pages 105–158, 1995.

[EGH00] R. Eymard, T. Galluoët, and R. Herbin. *Finite Volume Methods*, volume 7 of *Handbook of Numerical Analysis*. North Holland, Amsterdam, 2000.

[ELL02] R.E. Ewing, T. Lin, and Y. Lin. On the accuracy of the finite volume element method based on piecwise linear polynomials. *SIAM J. Numer. Anal.*, 39(6):1865–1888, 2002.

[EMRS92] B. Einfeldt, C. Munz, P. Roe, and B. Sjögreen. On Godunov-type methods near low densities. *J. Comp. Phys.*, 92:273–295, 1992.

[GLLS97] M. Giles, M. Larson, M. Levenstam, and E. Süli. Adaptive error control for finite element approximations of the lift and drag coefficients in viscous flow. preprint NA-97/06, Comlab, Oxford University, 1997.

[God59] S. K. Godunov. A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Mat. Sb.*, 47:271–290, 1959.

[GP99] M. Giles and N.A. Pierce. Improved lift and drag estimates using adjoint Euler equations. Technical Report 99-3293, AIAA, Reno, NV, 1999.

[GR91] E. Godlewski and P-A. Raviart. *Hyperbolic Systems of Conservation Laws*. Mathematiques & Applications. Ellipses, 1991.

[GV85] J. D. Goodman and R. J. Le Veque. On the accuracy of stable schemes for 2D conservation laws. *Math. Comp.*, 45(171):15–21, 1985.

[Har83] A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comp. Phys.*, 49:357–393, 1983.

[Har89] A. Harten. ENO schemes with subcell resolution. *J. Comp. Phys.*, 83:148–184, 1989.

[HC91] A. Harten and S. Chakravarthy. Multi-dimensional ENO schemes for general geometries. Technical Report ICASE-91-76, Institite for Computer Applications in Science and Engineering (ICASE), NASA Langley R.C., 1991.

[Her00] F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.*, 160(2):481–499, 2000.

[HHL76] A. Harten, J. M. Hyman, and P. D. Lax. On finite-difference approximations and entropy conditions for shocks. *Comm. Pure and Appl. Math.*, 29:297–322, 1976.

[HLvL83] A. Harten, P. D. Lax, and B. van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Rev.*, 25:35–61, 1983.

[HOEC86] A. Harten, S. Osher, B. Engquist, and S. Chakravarthy. Some results on uniformly high order accurate essentially non-oscillatory schemes. *Appl. Num. Math.*, 2:347–377, 1986.

[HOEC87] A. Harten, S. Osher, B. Engquist, and S. Chakravarthy. Uniformly high-order accurate essentially nonoscillatory schemes III. *J. Comp. Phys.*, 71(2):231–303, 1987.

[Jam93] A. Jameson. Artificial diffusion, upwind biasing, limiters and their effect on accuracy and convergence in transonic and hypersonic flows. Technical Report 93-3359, AIAA, Reno, NV, 1993.

[Jam95] A. Jameson. Analysis and design of numerical schemes for gas dynamics. Technical Report TR 94-15, RIACS, NASA Ames R.C., Moffett Field, CA, 1995.

[JL86] A. Jameson and P.D. Lax. Conditions for the construction of multipoint variation diminishing difference schemes. *Appl. Numer. Math.*, 2(3-5):235–345, 1986.

[JL87] A. Jameson and P.D. Lax. Corrigendum: Conditions for the construction of multipoint variation diminishing difference schemes. *Appl. Numer. Math.*, 3(3):289, 1987.

[JP86] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46:1–26, 1986.

[JRB95] C. Johnson, R. Rannacher, and M. Boman. Numerics and hydrodynamics stability theory: towards error control in CFD. *SIAM J. Numer. Anal.*, 32:1058–1079, 1995.

[JS96] G. Jiang and C.-W. Shu. Efficient implementation of weighted ENO schemes. *J. Comp. Phys.*, 126:202–228, 1996.

[Kor88] B. Koren. Upwind schemes for the Navier-Stokes equations. In *Proceedings of the Second International Conference on Hyperbolic Problems*. Vieweg:Braunschweig, 1988.

[Krö97] D. Kröner. *Numerical Schemes for Conservation Laws*. Wiley–Teubner, 1997.

[Lax73] P. D. Lax. *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*. SIAM, Philadelphia, Penn., 1973.

[LeV02] R. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[Liu93] X.-D. Liu. A maximum principle satisfying modification of triangle based adaptive stencils for the solution of scalar hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 30:701–716, 1993.

[LMV96] R.D. Lazarov, I.D. Michev, and P.S. Vassilevsky. Finite volume methods for convection-diffusion problems. *SIAM J. Numer. Anal.*, 33:31–35, 1996.

[LS93] M.S. Liou and C.J. Steffen. A new flux-splitting scheme. *J. Comp. Phys.*, 107:23–39, 1993.

[OP99] J. T. Oden and S. Prudhomme. Goal-oriented error estimation and adaptivity for the finite element method. Technical Report 99-015, TICAM, U. Texas, Austin,TX, 1999.

[OS82] S. Osher and F. Solomon. Upwind difference schemes for hyperbolic systems of conservation laws. *Math. Comp.*, 38(158):339–374, 1982.

[Osh84] S. Osher. High resolution schemes and the entropy condition. *SIAM J. Numer. Anal.*, 21(5):955–984, 1984.

[Pet91] T. Peterson. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *SIAM J. Numer. Anal.*, 28(1):133–140, 1991.

[PO99] S. Prudhomme and J.T. Oden. On goal-oriented error estimation for elliptic problems: application to the control of pointwise errors. *Comp. Meth. Appl. Mech. and Eng.*, pages 313–331, 1999.

[PPP97] M. Parashivoiu, J. Peraire, and A. Patera. A posteriori finite element bounds for linear-functional outputs of elliptic partial differential equations. *Comput. Meth. Appl. Mech. Engrg.*, 150:289–312, 1997.

[RH73] W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos National Laboratory, Los Alamos, New Mexico, 1973.

[Roe81] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43:357–372, 1981.

[RS88] P. Rostand and B. Stoufflet. TVD schemes to compute compressible viscous flows on unstructured meshes. In *Proceedings of the Second International Conference on Hyperbolic Problems*. Vieweg:Braunschweig, 1988.

[S̈98] E. Süli. A posteriori error analysis and adaptivity for finite element approximations of hyperbolic problems. In Kröner, Ohlberger, and Rohde, editors, *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws*, volume 5 of *Lecture Notes in Computational Science and Engineering*, pages 122–194. Springer-Verlag, Heidelberg, 1998.

[Shu99] C.-W. Shu. Discontinuous Galerkin methods for convection-dominated problems. In Barth and Deconinck, editors, *High-Order Discretization Methods in Computational Physics*, volume 9 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Heidelberg, 1999.

[SO88] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing scheme. *J. Comp. Phys.*, 77:439–471, 1988.

[Son97] T. Sonar. On the construction of essentially non-oscillatory finite volume approximations to hyperbolic conservation laws on general triangulations: Polynomial recovery, accuracy, and stencil selection. *Comput. Meth. Appl. Mech. Engrg.*, 140:157–181, 1997.

[Son98] T. Sonar. On families of pointwise opitimal finite volume ENO approximations. *SIAM J. Numer. Anal.*, 35(6):2350–2379, 1998.

[Spe87] S.P. Spekreijse. Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws. *Math. Comp.*, 49:135–155, 1987.

[Sül91] E. Süli. Convergence of finite volume schemes for Poisson's equation on nonuniform meshes. *SIAM J. Numer. Anal.*, 28:1419–1430, 1991.

[SW81] J.L. Steger and R.F. Warming. Flux vector splitting of the inviscid gasdynamic equations with application to finite difference methods. *J. Comp. Phys.*, 40:263–293, 1981.

[Swe84] P.K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21(5):995–1011, 1984.

[Van93] P. Vankeirsblick. *Algorithmic Developments for the Solution of Hyperbolic Conservation Laws on Adaptive Unstructured Grids*. PhD thesis, Katholieke Universiteit Leuven, Belgium, 1993.

[VHMD98] C. Viozat, C. Held, K. Mer, and A. Dervieux. On vertex-center unstructured finite-volume methods for stretched anisotropic triangulations. Technical Report 3464, Institut National De Recherche En Informatique Et En Automatique (INRIA), 1998.

[vL79] B. van Leer. Towards the ultimate conservative difference schemes V. A second order sequel to Godunov's method. *J. Comp. Phys.*, 32:101–136, 1979.

[vL82] B. van Leer. Flux-vector splitting for the Euler equations. Technical Report ICASE-82-30, Institite for Computer Applications in Science and Engineering (ICASE), NASA Langley R.C., 1982.

[vL85] B. van Leer. Upwind-difference schemes for aerodynamics problems governed by the Euler equations. volume 22 of *Lectures in Applied Mathematics*. AMS Pub., Providence, Rhode Island, 1985.

[Wie94] M. Wierse. *Higher Order Upwind Scheme on Unstructured Grids for the Compressible Euler Equations in Time Dependent Geometries in 3D*. PhD thesis, University of Freiburg, Germany, 1994.

[ZZ92] O.C. Zienkiewicz and J.Z. Zhu. The superconvergent patch recovery and a posteriori error estimates. Part I: the recovery technique. *Int. J. Numer. Meth. Engrg.*, 33:1331–1364, 1992.