RAIRO
MODÉLISATION MATHÉMATIQUE
ET ANALYSE NUMÉRIQUE

D.-M. CAI
R. S. FALK

Reduced continuity finite element methods for first order scalar hyperbolic equations

RAIRO – Modélisation mathématique et analyse numérique, tome 28, n° 6 (1994), p. 667-698.

<http://www.numdam.org/item?id=M2AN_1994__28_6_667_0>

MATHEMATICAL MODELLING AND NUMERICAL ANALYSIS
MODÉLISATION MATHÉMATIQUE ET ANALYSE NUMÉRIQUE

(Vol 28, n° 6, 1994, p 667 à 698)

REDUCED CONTINUITY FINITE ELEMENT METHODS FOR FIRST ORDER SCALAR HYPERBOLIC EQUATIONS (*)

by D.-M. CAI (1) and R. S FALK (1)

Communicated by J BRAMBLE

Abstract — *Two explicit finite element methods for a first order linear hyperbolic problem in* $\mathbf{R}^2$ *are proposed and analyzed These schemes are designed to produce an approximate solution which has a certain number of continuous moments across element edges* L^2 *error estimates of order* $O(h^{n+1/2})$ *for both schemes are obtained This is the same convergence rate known for the discontinuous Galerkin method, but is achieved with fewer computations Some numerical results for these methods are presented and comparisons are made with other explicit finite element methods for this problem previously studied in the literature*

Résumé — *On analyse deux méthodes explicites d'éléments finis pour un problème hyperbolique linéaire du premier ordre Les schémas sont conçus pour obtenir une solution approchée possédant un certain nombre de moments continus à travers les faces des éléments Des estimations d'erreur* L^2 *d'ordre* $O(h^{n+1/2})$ *sont obtenues pour chacun des deux schémas C'est le même taux de convergence que pour la méthode de Galerkin discontinue, mais il est obtenu avec moins de calculs Quelques résultats numériques sont présentés et des comparaisons sont faites avec d'autres méthodes explicites d'éléments finis de la littérature appliquées au même problème*

1. INTRODUCTION

The finite element approximation of the first order scalar hyperbolic equation

$$\begin{cases} \beta \cdot \nabla u + au = f & \text{in} \quad \Omega \subset \mathbf{R}^2, \\ u = g & \text{on} \quad \Gamma_{\text{in}}(\Omega), \end{cases} \tag{1.1}$$

has been investigated using several different approaches. Previous analysis of this problem was done for two types of explicit approximation schemes :

(*) Received for publication April 1992
AMS(MOS) subject classifications (1985 revision), 65N30, 65M15.
(1) Department of Mathematics, Rutgers University, New Brunswick, NJ 08903
This research supported by NSF Grant DMS-9106051

one which produces a piecewise polynomial approximation which is discontinuous across the triangle edges in the finite element mesh and one which produces a continuous piecewise polynomial approximation. The discontinuous triangular scheme has been analyzed first by Lesaint and Raviart [9], with improved and additional error estimates obtained by Johnson and Pitkäranta [8]. Optimal order error estimates were also derived in the case of semiuniform triangular meshes by Richter [11]. In the case of the continuous scheme, Falk and Richter [3] obtained estimates for a method initiated by Reed and Hill [10] using triangular elements. For rectangular element approximations, Lesaint and Raviart [9] and Winther [12] developed discontinuous and continuous finite element methods, respectively. They both achieved the optimal order of convergence, assuming sufficient regularity.

In this paper, we propose and analyze a class of reduced continuity finite element schemes for this problem. These schemes produce piecewise polynomial approximations which are continuous for a certain number of moments across interelement edges and are devised to retain the advantages of the previous two methods. As in the case of the previous methods, these schemes are explicit, in that the finite element solution may be developed in an explicit manner from element to element, and have the property that the solution in a given layer of elements may be computed in parallel. Hence they can be easily implemented and are economic in practice. This is quite different from the streamline-diffusion method. The latter is an implicit scheme originally introduced by Hughes and Brooks [6] for numerically solving convection dominated convection-diffusion problems and later applied to (1.1) as their corresponding reduced problems by Johnson *et al.* [7]. Since an implicit method must solve a large linear system, its computational cost could be large.

The previous explicit schemes using triangular elements rely on the following unified variational formulation on each triangle T:

$$(\beta \cdot \nabla u_h + a u_h, v)_T - \int_{\Gamma_{\text{in}}(T)} (u_h^+ - u_h^-)\, v\beta \cdot \mathbf{n}\, d\tau = (f, v)_T \quad \text{for} \quad v \in V_{h,T}, \tag{1.2}$$

where the approximate solution $u_h \in \mathbf{P}_n(T)$, the set of polynomials on T of degree $\leqslant n$ and u_h^- and u_h^+ denote the upstream and downstream limits of u_h on $\Gamma_{\text{in}}(T)$. The choice of the test space $V_{h,T}$ and the boundary continuity conditions will then determine each scheme. When $V_{h,T} = \mathbf{P}_n(T)$ and no boundary continuity of u_h is imposed, we get the discontinuous Galerkin method. If $V_{h,T} = \mathbf{P}_{n-l}(T)$, where l denotes the number of inflow sides that T has, and u_h is enforced to be continuous globally, we then obtain a continuous method. Analogously, in our schemes we also make use of (1.2)

and choose $V_{h,T}$ such that it contains all polynomials in the crosswind variable t of degree $\leq n$ and a suitable number of continuous boundary moments in each case. Since the boundary continuity conditions will decrease the degrees of freedom to be determined, thus reducing the number of unknowns to be solved for in the approximate solution, our schemes obviously require fewer computations per triangle than the discontinuous Galerkin method. Employing a test function depending only on t, we obtain L^2 error estimates of order $O(h^{n+1/2})$ and other accuracy properties similar to the discontinuous Galerkin method. This L^2 result is also an improvement on the $O(h^{n+1/4})$ L^2 error estimate previously shown for the continuous method at the cost of a little more computational effort.

We note that it is also possible to develop reduced continuity rectangular elements for equation (1.1) which produce optimal order convergence rates under the assumption of sufficient regularity (*cf.* [1]).

An outline of the paper is as follows. In the next section some basic notation and assumptions are provided. In § 3 we describe two discrete problems and give a characterization of these methods to show more similarities to the continuous and discontinuous methods. The proof of existence and uniqueness of solutions to the discrete problems is given in § 4. In § 5, the main stability results of the proposed methods are established and then used to derive the desired error estimates. Finally, in § 6, we provide some results of numerical experiments for the proposed methods and compare them with the continuous and discontinuous methods.

2. NOTATION AND ASSUMPTIONS

For the sake of simplicity, we consider a model problem of the form

$$\begin{cases} \beta \cdot \nabla u = f & \text{in} \quad \Omega\,, \\ u = g & \text{on} \quad \Gamma_{\text{in}}(\Omega)\,, \end{cases} \tag{2.1}$$

where β is a constant unit vector. For the case with a variable β and a lower order term a, the main results in Theorem 5.3 can still be obtained (*cf.* [1]). In the above, Ω is a bounded polygonal domain in $\mathbf{R}^2$ and $\Gamma_{\text{in}}(\Omega)$ its inflow boundary. By the inflow boundary $\Gamma_{\text{in}}(D)$ of a region D we mean $\{P \in \Gamma(D) : \beta \cdot \mathbf{n}(P) < 0\}$, where $\mathbf{n}(P)$ is the unit outward normal to D at P. Then we set $\Gamma_{\text{out}}(D) = \Gamma(D) - \Gamma_{\text{in}}(D)$.

In what follows, for a region D and a piecewise smooth curve Γ, $(.,.)_D$ and $\|.\|_D$ denote the inner product and the norm on $L^2(D)$, and $\|.\|_{k,D}$ and $|.|_{k,\Gamma}$ denote the norms on $H^k(D)$ and $H^k(\Gamma)$, respectively. Moreover, we shall use $|.|$ to denote the Euclidean norm or its

corresponding matrix norm and define a weighted inner product and its induced norm on $L^2(\Gamma)$ as the following

$$\langle w, v\rangle_\Gamma = \int_\Gamma wv\,|\beta\cdot\mathbf{n}|\,d\tau \quad \text{and} \quad |w|_\Gamma = \langle w, w\rangle_\Gamma^{1/2}.$$

Let $\mathbf{P}_n(D)$ be the space of polynomials of degree $\leqslant n$ on D and $\mathbf{Sp}(v_1, \ldots, v_l)_D$ a vector space spanned by the polynomials v_i, $i = 1, \ldots, l$, over D. We take g_I to be a suitable interpolant of g on $\Gamma_{\text{in}}(\Omega)$ and denote the limit of $w(P \pm \varepsilon\beta)$ as ε decreases to 0 by $w^{\pm}(P)$. Let C stand for a generic constant independent of all major variables u, f, and h, and not necessarily the same at its various occurrences.

To describe the methods we shall analyze, let Δ_h be a quasi-uniform triangulation of Ω such that no maximal diameter of triangular element $T \in \Delta_h$ is bigger than h. More specifically, we assume that Δ_h satisfies the following hypothesis :

$\mathbf{H}_1$ (quasi-uniform)

$$\frac{h_{\max}}{\rho_{\min}} \leqslant M$$

uniformly for all Δ_h when h is sufficiently small, where

$h_{\max} = \max_{T\in\Delta_h} h_T$, h_T = the diameter of T ; and

$\rho_{\min} = \min_{T\in\Delta_h} \rho_T$, ρ_T = the radius of the inscribed circle in T.

Note that each triangle in Δ_h is either of type I (with one inflow side) or of type II (with two inflow sides). We will sometimes consider a partition of Δ_h into certain *layers :*

$$S_1 = \{T \in \Delta_h : \Gamma_{\text{in}}(T) \subset \Gamma_{\text{in}}(\Omega)\},$$
$$S_{i+1} = \{T \in \Delta_h : \Gamma_{\text{in}}(T) \subset \Gamma_{\text{in}}(\Omega - U_{k\leqslant i} S_k)\}, \quad i = 1, 2, \ldots.$$

As will be seen from the construction of the methods proposed in § 3, we can develop the approximate solution layer by layer and simultaneously over all elements within a layer.

We further assume that Δ_h satisfies the following two hypotheses.

$\mathbf{H}_2$ (nonalignment). There exists $\varepsilon_0 > 0$ independent of h such that

$$|\beta\cdot\mathbf{n}| \geqslant \varepsilon_0$$

along all inflow edges of type II triangles.

$\mathbf{H}_3$ The total number of layers in Δ_h is $O(h^{-1})$.

Some remarks about the necessity and validity of these hypotheses are given in the Appendices.

We next list some facts which will be used later in this paper.

(i) The integration by parts formula :

$$\int_T (\beta \cdot \nabla w)\, v = \int_{\Gamma(T)} w\, v\beta \cdot \mathbf{n} - \int_T w(\beta \cdot \nabla v)\,. \tag{2.2}$$

(ii) The inverse inequalities :

$$\|\nabla w\|_T \leq C h^{-1} \|w\|_T \quad \text{for} \quad w \in \mathbf{P}_n(T)\,; \tag{2.3}$$

and

$$|w|_{0,\,\Gamma(T)} \leq C h^{-1/2} \|w\|_T \quad \text{for} \quad w \in \mathbf{P}_n(T)\,. \tag{2.4}$$

For the sake of convenience, when $i = 1, 2, 3$, we denote by Γ_i the sides of $T \in \Delta_h$ numbered counterclockwise, by a_i the opposite vertices of Γ_i, by $\mathbf{n}_i$ the unit outward normals to Γ_i and by τ_i the unit tangential vectors along Γ_i, taken in a counterclockwise direction. We shall always take Γ_3 to be the inflow side of a type I triangle or the outflow side of a type II triangle. On each $T \in \Delta_h$ of type I (II) we establish a local oblique coordinate system (t, s) with the origin at a_1 (a_2) and spanned by the tangent $\tau = \tau_3 (-\tau_3)$ and the characteristic β. Thus every point in a type I triangle has positive s coordinate while that in a type II triangle has negative s coordinate. This notation is illustrated in figure 1.

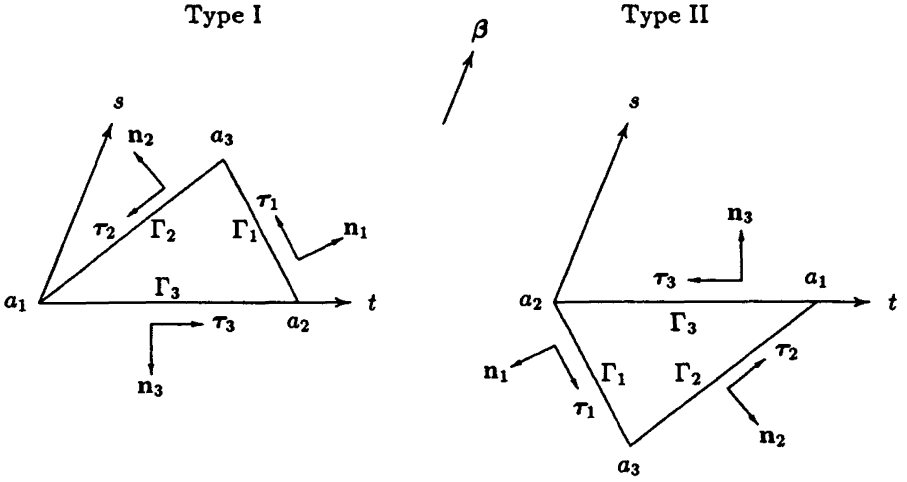

Figure 1.

The relations between this local oblique system (t, s) and the global orthogonal system (x, y) can be easily demonstrated via the following linear transformation

$$\begin{pmatrix} t \\ s \end{pmatrix} = (\tau, \beta)^{-1} \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix},$$

where (x_0, y_0) are the coordinates that the origin of the system (t, s) has under the system (x, y).

We next observe that both $\Gamma_{\text{in}}(T)$ and $\Gamma_{\text{out}}(T)$ can be parameterized in terms of equations : $s = s_{\text{in}}(t)$ and $s = s_{\text{out}}(t)$ for $t \in [0, t_T]$ and T can be described by

$$T = \{(t, s) : t \in [0, t_T],\ s \in [s_{\text{in}}(t), s_{\text{out}}(t)]\}.$$

For any function Φ on T, we denote $\Phi_{\text{out}} = \Phi|_{\Gamma_{\text{out}}(T)}$ and $\Phi_{\text{in}} = \Phi|_{\Gamma_{\text{in}}(T)}$. Equivalently, in terms of (t, s) coordinates, $\Phi_{\text{out}}(t) = \Phi(t, s_{\text{out}}(t))$ and $\Phi_{\text{in}}(t) = \Phi(t, s_{\text{in}}(t))$. We also define, for convenience, a weighted inner product and its induced norm in $L^2([0, t_T])$ as follows

$$\langle w, v \rangle = \int_0^{t_T} w(t)\, v(t) |\beta \cdot \mathbf{n}_3|\, dt \quad \text{and} \quad |w| = \langle w, w \rangle^{1/2}.$$

With this inner product, we introduce a boundary projection $P_t : L^2[0, t_T] \to \mathbf{P}_n[0, t_T]$, which will be used frequently in the sequel. We also denote by P_n, the standard L^2 interior projection into $\mathbf{P}_n(T)$. Moreover, we define the *Extension* $Ev(t, s)$ of $v(t)$ to be a function over $T \in \Delta_h$ such that $\frac{\partial}{\partial_s} Ev(t, s) = 0$ and $Ev(t, 0) = v(t)$, and note that $w_s = \beta \cdot \nabla w$.

Finally, we state a lemma containing two change of variables formulas and an integration identity which we shall frequently use in this paper. The proof is elementary and we omit it here.

LEMMA 2.1 : *Any function w defined on a triangle T of either type satisfies*

$$\int_{\Gamma_{\text{out}}(T)} w |\beta \cdot \mathbf{n}|\, d\tau = \int_0^{t_f} w_{\text{out}} |\beta \cdot \mathbf{n}_3|\, dt, \tag{2.5}$$

$$\int_{\Gamma_{\text{in}}(T)} w |\beta \cdot \mathbf{n}|\, d\tau = \int_0^{t_T} w_{\text{in}} |\beta \cdot \mathbf{n}_3|\, dt; \tag{2.6}$$

and

$$\int_T w\, dx\, dy = \int_0^{t_T} \int_{s_{\text{in}}(t)}^{s_{\text{out}}(t)} w |\beta \cdot \mathbf{n}_3|\, ds\, dt. \tag{2.7}$$

3. FORMULATION OF THE METHODS

Let Δ_h be a triangulation of Ω satisfying hypotheses $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3$ of the previous section. Using the variational equation

$$(\beta \cdot \nabla u_h, v)_T - \int_{\Gamma_{\text{in}}(T)} (u_h^+ - u_h^-) v\beta \cdot \mathbf{n}\, d\tau = (f, v)_T \quad \text{for all} \quad v \in V_{h,T}, \tag{3.1}$$

where $V_{h,T}$ is a test space to be specified, we can formulate our schemes as follows :

Method M_h^1 : For $n \geq 1$, find $u_h \in L^2(\Omega)$ such that $u_h^- = g_I$ on $\Gamma_{\text{in}}(\Omega)$, $u_h|_T \in \mathbf{P}_n(T)$ for any $T \in \Delta_h$ and for triangles of type I, u_h satisfies (3.1) with $V_{h,T} = \mathbf{P}_{n-1}(T) \oplus \mathbf{Sp}(t^n)_T$ and

$$\int_{\Gamma_3} (u_h^+ - u_h^-) \tau^l d\tau = 0 \quad \text{for} \quad l = 0, 1, \ldots, n-1 ; \tag{3.2}$$

while for triangles of type II, u_h satisfies (3.1) with $V_{h,T} = \mathbf{P}_1(T)$ when $n = 1$ or $V_{h,T} = \mathbf{P}_{n-2}(T) \oplus \mathbf{Sp}(t^{n-1}, s^{n-1}, t^n)_T$ and

$$\int_{\Gamma_i} (u_h^+ - u_h^-) \tau^l d\tau = 0 \text{ for } l = 0, 1, \ldots, n-2 \text{ and } i = 1, 2 \tag{3.3}$$

when $n \geq 2$.

Method M_h^2 : For odd $n \geq 1$, find $u_h \in L^2(\Omega)$ such that $u_h^- = g_I$ on $\Gamma_{\text{in}}(\Omega)$ and for a type I triangle T, u_h satisfies the same conditions as in $\mathbf{M}_h^1$; while for a type II triangle T, $u_h|_T \in \mathbf{P}_n(T) \oplus \mathbf{Sp}(st^n)_T$ satisfies (3.1) with $V_{h,T} = \mathbf{P}_{n-2} \oplus \mathbf{Sp}(t^{n-1}, t^n)_T$ and

$$\int_{\Gamma_i} (u_h^+ - u_h^-) \tau^l d\tau = 0 \text{ for } l = 0, 1, \ldots, n-1 \text{ and } i = 1, 2 . \tag{3.4}$$

Remark 3.1 : Note that in $\mathbf{M}_h^1$ the schem › for type II triangles is in fact a discontinuous Galerkin method when $n = 1$. This reflects a common feature of the scheme of order n for two-inflow-side triangles ; they all have $2(n-1)$ continuity conditions on the inflow triangle sides.

Remark 3.2 : There are some difficulties in formulating even-order elements over an arbitrary type II triangle for Method $\mathbf{M}_h^2$. For example, suppose (t, s) is an orthogonal coordinate system and T is a triangle with the vertices $(1, 0)$, $(-1, 0)$, and $(0, -1)$ in (t, s) for simplicity. Then the

polynomial $u_h(t, s) = s(10\,t^2 - 8\,s - 7)$ satisfies all requirements in $\mathbf{M}_h^2$ over a type II triangle when $n = 2$ and $u_h^- = f = 0$. This implies that the second-order element of $\mathbf{M}_h^2$ is not unisolvent over this triangle. In fact, all even-order elements over this triangle are not unisolvent, as can be seen from the proof of Lemma 4.1.

We observe that the approximate solution u_h has a total of $\sigma_n(= (n + 1)(n + 2)/2)$ or $\sigma_n + 1$ degrees of freedom in each triangle. For both methods the number of the continuity conditions on the inflow boundary of a type I triangle is n, leaving a total of $\sigma_{n-1} + 1$ degrees of freedom to be determined. For a type II triangle the degrees of freedom to be determined are $\sigma_{n-2} + 3$ and $\sigma_{n-2} + 2$ for $\mathbf{M}_h^1$ and $\mathbf{M}_h^2$, respectively, which are exactly the dimensions of the test spaces.

4. CHARACTERIZATION AND WELL-POSEDNESS

To help expose the essential features of the discrete problems proposed in the last section, we want to characterize their approximate solutions u_h in a fashion analogous to the continuous and discontinuous methods discussed in [4]. Then we proceed to show that these problems are well-formulated.

Suppose u_h is a solution developed on a type I triangle for either $\mathbf{M}_h^1$ or $\mathbf{M}_h^2$. Then it satisfies (3.1) with $V_{h,T} = \mathbf{P}_{n-1}(T) \oplus \mathbf{Sp}(t^n)_T$. Since $(u_h)_s \in \mathbf{P}_{n-1}(T)$, we have, by making use of the boundary continuity conditions (3.2),

$$(u_h)_s = P_{n-1} f\,.$$

Hence for any $w \in \mathbf{P}_n[0, t_T]$,

$$\langle u_{h,\,\mathrm{in}}^+ - u_{h,\,\mathrm{in}}^-, w\rangle = ((I - P_{n-1}) f, Ew)_T = \left\langle \int_0^{s_{\mathrm{out}}(t)} (I - P_{n-1}) f, w \right\rangle$$

by (3.1) and Lemma 2.1. This implies

$$u_{h,\,\mathrm{in}}^+(t) = u_{h,\,\mathrm{in}}^-(t) + P_t \int_0^{s_{\mathrm{out}}(t)} (I - P_{n-1}) f\, ds\,,$$

in view of $u_{h,\,\mathrm{in}}^+$, $u_{h,\,\mathrm{in}}^- \in \mathbf{P}_n[0, t_T]$. On the other hand, we see that $u_h(t, s) = u_{h,\,\mathrm{in}}^+(t) + \int_0^s (u_h)_s\, ds$. Therefore,

$$u_h(t, s) = u_{h,\,\mathrm{in}}^-(t) + P_t \int_0^{s_{\mathrm{out}}(t)} (I - P_{n-1}) f\, ds + \int_0^s P_{n-1} f\, ds\,. \quad (4.1)$$

To characterize u_h on a type II triangle, we first introduce a function U on T such that $U_{\text{in}} = u_{h,\text{in}}^-$ and $U_s = f$. Then for $v \in V_{h,T}$, we find

$$0 = ((u_h - U)_s, v)_T + \langle u_{h,\text{in}}^+ - U_{\text{in}}, v \rangle$$
$$= -(u_h - U, v_s)_T + \langle u_{h,\text{out}}^- - U_{\text{out}}, v \rangle ,$$

by Lemma 2.1 and after integrating by parts.

When u_h is a solution of $\mathbf{M}_h^1$ for $n \geq 2$ (the case $n = 1$ is the same as the discontinuous scheme; see below), we first take $v = Ew$ in the above identity with $w \in \mathbf{P}_n[0, t_T]$. It follows that

$$u_{h,\text{out}}^- = P_t U_{\text{out}} . \tag{4.2}$$

Observing that $v_s \in \mathbf{P}_{n-3}(T) \oplus \mathbf{Sp}(s^{n-2})_T$ for any $v \in \mathbf{P}_{n-2}(T) \oplus \mathbf{Sp}(t^{n-1}, s^{n-1}, t^n)_T = V_{h,T}$, we conclude that

$$P_{n-3}^* u_h = P_{n-3}^* U , \tag{4.3}$$

where P_{n-3}^* is an L^2 interior projection to $\mathbf{P}_{n-3}(T) \oplus \mathbf{Sp}(s^{n-2})_T$. Moreover, by the continuous moment conditions (3.3), we have

$$\int_{\Gamma_i} u_h^+ \tau^l d\tau = \int_{\Gamma_i} U \tau^l d\tau \text{ for } l = 0, 1, \ldots, n-2 \text{ and } i = 1, 2 . \tag{4.4}$$

Thus we have specified u_h in terms of U.

Similarly, we have a characterization for an $\mathbf{M}_h^2$ solution :

$$u_{h,\text{out}}^- = P_t U_{\text{out}} , \tag{4.5}$$
$$P_{n-3} u_h = P_{n-3} U , \tag{4.6}$$

$$\int_{\Gamma_i} u_h^+ \tau^l d\tau = \int_{\Gamma_i} U \tau^l d\tau \text{ for } l = 0, 1, \ldots, n-1 \text{ and } i = 1, 2 . \tag{4.7}$$

Later, from Lemma 4.1, we will see that on a triangle of type II, an approximate solution is completely determined by (4.2), ..., (4.4) for $\mathbf{M}_h^1$ and by (4.5), ..., (4.7) for $\mathbf{M}_h^2$.

Let us now briefly describe the characterizations for the continuous and discontinuous Galerkin methods developed in [4]. For the continuous method, an approximate solution u_h has the representation

$$u_h(t, s) = u_{h,\text{in}}(t) + \int_0^s P_{n-1} f \, ds$$

on a type I triangle and satisfies

$$\frac{d}{dt} u_{h,\,\mathrm{out}} = P_{t,\,n-1} \frac{d}{dt} U_{\mathrm{out}} \quad \text{and} \quad P_{n-3}\, u_h = P_{n-3}\, U$$

on a type II triangle, where U is defined as before with $U_{\mathrm{in}} = u_{h,\,\mathrm{in}}$ and $P_{t,\,n-1}$ denotes the L^2 projection into $\mathbf{P}_{n-1}[0, t_T]$. For the discontinuous method, u_h is characterized as

$$u_h(t, s) = u_{h,\,\mathrm{in}}^{-}(t) + P_t \int_0^{s_{\mathrm{out}}(t)} (I - R_{n-1}) f \, ds + \int_0^{s} R_{n-1} f \, ds$$

on a type I triangle with R_{n-1} denoting the projection into $\mathbf{P}_{n-1}(T)$ with respect to the weighted L^2 inner product $[p, q] = (sp, q)_T$ and

$$u_{h,\,\mathrm{out}}^{-} = P_t\, U_{\mathrm{out}} \quad \text{and} \quad P_{n-1}\, u_h = P_{n-1}\, U$$

on a type II triangle with U as defined for $\mathbf{M}_h^1$ and $\mathbf{M}_h^2$. From the characterizations given above, we can obtain a clearer view of similarities between these four explicit schemes.

We now want to prove an existence and uniqueness result for our methods.

LEMMA 4.1 : *There exist unique solutions to the discrete problems* $\mathbf{M}_h^1$ *and* $\mathbf{M}_h^2$.

Proof : The statement is obvious for a one-inflow-side triangle by the representation (4.1). For a two-inflow-side triangle, we first prove the uniqueness for each method. The existence of the numerical solution then follows since in either case the sum of the dimension of the test space and the number of the continuous moments on the inflow boundary is exactly the same as the number of degrees of freedom of the finite element.

To derive uniqueness of the problem $\mathbf{M}_h^1$ with $n \geq 2$, let us choose the degrees of freedom of the finite element to be the standard ones related to the three vertices a_l, the moments from 0 to $n-2$ on each side Γ_l and the inner product with the polynomials of degree $\leq n-3$ over T (when $n = 2$, this part is void). The corresponding basis functions $\{\phi_l,\ \psi_{lj},\ \eta_k :\ l = 1, 2, 3\ ;\ j = 0, 1, \ldots, n-2\ ;\ k = 1, 2, \ldots, \sigma_{n-3}\} \subset \mathbf{P}_n(T)$ satisfy

$$\phi_l(a_l) = \delta_{ll}\,, \qquad \int_{\Gamma_l} \phi_l\, \tau^m \, d\tau = 0\,, \qquad (\phi_l, w_q)_T = 0\ ;$$

$$\psi_{lj}(a_l) = 0\,, \qquad \int_{\Gamma_l} \psi_{lj}\, \tau^m \, d\tau = \delta_{ll}\, \delta_{jm}\,, \qquad (\psi_{lj}, w_q)_T = 0\ ;$$

$$\eta_k(a_l) = 0\,, \qquad \int_{\Gamma_l} \eta_k\, \tau^m \, d\tau = 0\,, \qquad (\eta_k, w_q)_T = \delta_{kq}$$

for $i, l = 1, 2, 3$; $j, m = 0, 1, \ldots, n-2$; $k, q = 1, 2, \ldots, \sigma_{n-3}$ with $\{w_q\}$ a basis of $\mathbf{P}_{n-3}(T)$ and δ the Kronecker delta. Then we can express u_h as

$$u_h = \sum_{i=1}^{3} \left(c_i \phi_i + \sum_{j=0}^{n-2} d_{ij} \psi_{ij} \right) + \sum_{k=1}^{\sigma_{n-3}} e_k \eta_k$$

where c_i, d_{ij} and e_k are constants.

If $u_{h,\text{in}}^- = f = 0$, then $U \equiv 0$ by its definition. The characterization (4.2), ..., (4.4) imply that $c_1 = c_2 = d_{ij} = e_k = 0$ for all i, j, k. Hence u_h is reduced to $c_3 s\phi(t, s)$ with $\phi \in \mathbf{P}_{n-1}(T)$ and $s\phi = \phi_3$. Note that

$$\int_T s(\phi_t)^2 \, ds \, dt = \int_T (s\phi)_t \, \phi_t \, ds \, dt = \int_T (\phi_3)_t \, \phi_t \, ds \, dt$$

$$= \int_{\Gamma(T)} \phi_3 \, \phi_t \, \tau \cdot \mathbf{n} \, d\tau - \int_T \phi_3 \, \phi_{tt} \, ds \, dt = 0 .$$

Since s does not change sign in T, we have $\phi_t = 0$ in T. This implies that ϕ is a polynomial of s.

On the other hand, $\int_{\Gamma_1} \phi_3 \, \tau^l \, d\tau = 0$, $l = 0, 1, \ldots, n-2$ implies that ϕ_3 is a multiple of $s(s - s_1) \ldots (s - s_{n-1})$ with $0 = s_0 > s_1 > \cdots > s_{n-1} > e$ being n Gauss-Radau quadrature points on the interval $[e, 0]$. It then follows that $u_h = cs(s - s_1) \ldots (s - s_{n-1})$ with c some constant. From (4.3) we see that

$$\begin{aligned} 0 &= c(s(s - s_1) \ldots (s - s_{n-1}), s^{n-2})_T \\ &= c \int_e^0 s(s - s_1)(s - s_2)^2 \ldots (s - s_{n-1})^2 \cdot \left(\int_{m_1 s + b_1}^{m_2 s + b_2} dt \right) |\beta \cdot \mathbf{n}_3| \, ds \\ &= c \, |\beta \cdot \mathbf{n}_3| \, (m_2 - m_1) \int_e^0 s(s - s_1)^2 \ldots (s - s_{n-1})^2 \, ds , \end{aligned}$$

where $t = m_1 s + b_1$ and $t = m_2 s + b_2$ are the parametric equations for Γ_1 and Γ_2, respectively. Here we also used the fact that $\int_e^0 s(s - s_1) \ldots (s - s_{n-1}) \, q_{n-2}(s) \, ds = 0$ for all polynomials q_{n-2} of degree $\leq n - 2$. Since $m_1 \neq m_2$ and $|\beta \cdot \mathbf{n}_3| \neq 0$ we conclude that $c = 0$. Thus $u_h \equiv 0$ in T.

We now turn to the problem $\mathbf{M}_h^2$. Again we have $U \equiv 0$ when $u_{h,\text{in}}^- = f = 0$. Thus by (4.5), $u_{h,\text{out}}^- = 0$. Rewrite u_h as $u_h = s(ct^n + p_{n-1}(t, s))$, where c is a constant and $p_{n-1} \in \mathbf{P}_{n-1}(T)$. We assert that c is

zero. In fact, let $t = m_1 s + b_1$ and $t = m_2 s + b_2$ again be parametric equations for Γ_1 and Γ_2, respectively. Then

$$u_h^+ \big|_{\Gamma_1} = cm_1^n s(s^n + q_{n-1}(s)) \text{ and } u_h^+ \big|_{\Gamma_2} = cm_2^n s(s^n + r_{n-1}(s))$$

for some $n-1$ degree polynomials of s: q_{n-1} and r_{n-1}, since m_1, $m_2 \neq 0$ by the definition of a type II triangle. Note that (4.7) for Γ_1 is now equivalent to

$$\int_e^0 cm_1^n s(s^n + q_{n-1}(s)) s^l \, ds = 0 , \quad l = 0, 1, \ldots, n-1 .$$

If we take $e < s_n < \cdots < s_1 < s_0 = 0$ to be the $n+1$ Gauss-Radau quadrature points on $[e, 0]$, then

$$u_h^+ \big|_{\Gamma_1} = cm_1^n s(s - s_1) \ldots (s - s_n) .$$

Analogously,

$$u_h^+ \big|_{\Gamma_2} = cm_2^n s(s - s_1) \ldots (s - s_n) .$$

Since $u_h^+ \big|_{\Gamma_1}(e) = u_h^+ \big|_{\Gamma_2}(e)$, $cm_1^n = cm_2^n$. The fact that Γ_1 and Γ_2 are not parallel implies $m_1 \neq m_2$. Thus $m_1^n \neq m_2^n$ for odd n. This yields $c = 0$. We now have $u_h = s\, p_{n-1}(t, s) \in \mathbf{P}_n(T)$ and $u_h^+ \big|_{\Gamma_i} = 0$, $i = 1, 2$. When $n = 1$, we can conclude that $u_h \equiv 0$ in T.

When $n \geq 3$, $u_h = \lambda_1 \lambda_2 \lambda_3 p_{n-3}$, where λ_1, λ_2, λ_3 are the barycentric coordinates in T and $p_{n-3} \in \mathbf{P}_{n-3}(T)$. Taking the inner product of u_h with p_{n-3} and applying the positivity of λ_i as well as $P_{n-3} u_h = 0$ from (4.6), we finally obtain $p_{n-3} = 0$. This means $u_h \equiv 0$ in T. The proof of uniqueness is therefore completed. □

Remark 4.1 : In Method $\mathbf{M}_h^1$, the test function s^{n-1} for type II triangles can be replaced by any $s^{n-1-k} t^k$ whenever $0 \leq k < n-1$ is even. This is easy to see from the above proof.

One of the immediate consequences of the above lemma is the following local stability inequality which will be used for deriving the global stability in the next section.

LEMMA 4.2 : *Let u_h be the solution of Method $\mathbf{M}_h^1$ or $\mathbf{M}_h^2$. Then*

$$\|u_h\|_T \leq C \left\{ h^{\frac{1}{2}} \left| u_h^- \right|_{\Gamma_{\text{in}}(T)} + h \|f\|_T \right\} . \tag{4.8}$$

Proof: To use a scaling argument to prove this lemma, we need to introduce a reference triangle $\hat{T}$ with vertices $\hat{a}_1 = (1, 0)$, $\hat{a}_2 = (0, 1)$, $\hat{a}_3 = (0, 0)$. For a generic triangle $T \in \Delta_h$, denote by F_T the invertible affine transformation of $\hat{T}$ to T such that

$$\begin{pmatrix} x \\ y \end{pmatrix} = B \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} + a_3 ,$$

where B is a 2×2 matrix of the form

$$B = (a_1 - a_3, a_2 - a_3) = (|\Gamma_2|\ \tau_2, -|\Gamma_1|\ \tau_1) .$$

Defining $\hat{v}(\hat{x}, \hat{y}) = v \circ F_T(\hat{x}, \hat{y}) = v(x, y)$ for any function v defined on T and

$$\hat{\nabla} = \begin{pmatrix} \partial/\partial\hat{x} \\ \partial/\partial\hat{y} \end{pmatrix} ,$$

we have

$$\beta \cdot \nabla u_h = B^{-1} \beta \cdot \hat{\nabla} \hat{u}_h ,$$

where

$$B^{-1} = \frac{1}{\mathbf{n}_1 \cdot \boldsymbol{\tau}_2} \begin{pmatrix} \mathbf{n}_1^T/|\Gamma_2| \\ \mathbf{n}_2^T/|\Gamma_1| \end{pmatrix} .$$

Observe that the reference transformation F_T preserves the types of triangles. This can be seen by noting that $F^{-1}(\beta) = B^{-1}\beta = \frac{1}{\mathbf{n}_1 \cdot \boldsymbol{\tau}_2} \left(\frac{\beta \cdot \mathbf{n}_1}{|\Gamma_2|}, \frac{\beta \cdot \mathbf{n}_2}{|\Gamma_1|} \right)^I$, $\hat{\mathbf{n}}_1 = (-1, 0)^T$, $\hat{\mathbf{n}}_2 = (0, -1)^T$, and $\mathbf{n}_1 \cdot \boldsymbol{\tau}_2 < 0$ imply that $\beta \cdot \mathbf{n}_1$ and $\beta \cdot \mathbf{n}_2$ have the same signs as $(B^{-1}\beta) \cdot \hat{\mathbf{n}}_1$ and $(B^{-1}\beta) \cdot \hat{\mathbf{n}}_2$, respectively.

Transforming (3.1), ..., (3.4) to $\hat{T}$, we have

$$|\det B|\,(B^{-1}\beta \cdot \hat{\nabla}\hat{u}_h, \hat{v})_{\hat{T}} - \sum_{\Gamma_i \subset \Gamma_{\text{in}}(T)} \mu_i |\Gamma_i| \times \int_{\hat{\Gamma}_i} (\hat{u}_h^+ - \hat{u}_h^-)\, \hat{v}\widehat{\beta \cdot \mathbf{n}_i}\, d\hat{\tau} = |\det B|\,(\hat{f}, \hat{v})_{\hat{T}} , \quad \text{for all} \quad \hat{v} \in \hat{V}_{h,\hat{T}} , \tag{4.9}$$

where $\mu_1 = \mu_2 = 1$ and $\mu_3 = \frac{1}{\sqrt{2}}$; and the corresponding boundary continuity conditions

$$\int_{\hat{\Gamma}_i} (\hat{u}_h^+ - \hat{u}_h^-)\, \hat{\tau}^l\, d\hat{\tau} = 0 , \quad \hat{\Gamma}_i \subset \Gamma_{\text{in}}(\hat{T})$$

and

$$l = 0, 1, \ldots, n-2 \,(\text{or } n-1) . \tag{4.10}$$

Let us denote by $\{\hat{\Phi}_J\}$ and $\{\hat{\Psi}_k\}$ the basis functions for the trial and test spaces on $\hat{T}$, respectively, and $\hat{U} = \{\hat{U}_J\}$ the coefficient set. We also set $\hat{\beta}$ to be the direction of $B^{-1}\beta$ and $\hat{\mathbf{n}}_l$ the unit outward normals to $\hat{\Gamma}_l$. Then by the fact that $\dfrac{\mu_l |\Gamma_l|\ \beta \cdot \mathbf{n}_l}{|B^{-1}\beta|\ \det B} = \hat{\beta} \cdot \hat{\mathbf{n}}_l$, (4.9) and (4.10) are equivalent to the following linear algebraic system

$$\sum_J \left\{ (\hat{\beta} \cdot \hat{\nabla}\hat{\Phi}_J,\ \hat{\Psi}_k)_{\hat{T}} - \sum_{\Gamma_l \subset \Gamma_{\text{in}}(T)} \int_{\hat{\Gamma}_l} \hat{\Phi}_J\ \hat{\Psi}_k\, \hat{\beta} \cdot \hat{\mathbf{n}}_l\, d\hat{\tau} \right\} \hat{U}_J$$

$$= \frac{1}{|B^{-1}\beta|} (\hat{f},\ \hat{\Psi}_k)_{\hat{T}} - \sum_{\Gamma_l \subset \Gamma_{\text{in}}(T)} \int_{\hat{\Gamma}_l} \hat{u}_h^-\ \hat{\Psi}_k\, \hat{\beta} \cdot \hat{\mathbf{n}}_l\, d\hat{\tau}\ ,$$

$$k = 1, 2, \ldots, \dim \hat{V}_{h,\hat{T}}\ , \quad (4.11)$$

and

$$\sum_J \left(\int_{\hat{\Gamma}_l} \hat{\Phi}_J\ \hat{\tau}^l\, d\hat{\tau} \right) \hat{U}_J = \int_{\hat{\Gamma}_l} \hat{u}_h^-\ \hat{\tau}^l\, d\hat{\tau}\ , \quad \hat{\Gamma}_l \subset \Gamma_{\text{in}}(\hat{T})$$

and $$l = 0, 1, \ldots, n-2\ (\text{or } n-1)\,, \quad (4.12)$$

or, in matrix form,

$$A\hat{U} = b\,, \quad (4.13)$$

where A is a $\sigma_n \times \sigma_n$ or $(\sigma_n + 1) \times (\sigma_n + 1)$ matrix.

It is obvious that A is uniformly bounded over all triangles by the hypothesis that all angles are bounded away from zero (the minimum angle condition implied in $\mathbf{H}_1$). Together with Lemma 4.1 we can also infer the uniform boundedness of A^{-1} and the bound : $|B^{-1}\beta|^{-1} \leq Ch$. Hence the solution $\hat{U}$ of the system (4.13) satisfies

$$|\hat{U}_l| \leq |A^{-1}|\,|b| \leq C\,|b| \leq C \left\{ |\hat{u}_h^-|_{0,\Gamma_{\text{in}}(\hat{T})} + h\|\hat{f}\|_{\hat{T}} \right\}\ ,$$

where the last inequality can be derived by carefully observing the system (4.11) and (4.12). When T is of type I, $(\hat{\beta} \cdot \hat{\mathbf{n}})_{\text{in}}$ is bounded uniformly away from 0 by the minimum angle condition. When T is of type II, the hypothesis $\mathbf{H}_2$ assures such a property. Therefore, for a triangle of either type,

$$\|\hat{u}_h\|_{\hat{T}} \leq C \sum_l |\hat{U}_l| \leq C \left\{ |\hat{u}_h^-|_{\Gamma_{\text{in}}(\hat{T})} + h\|\hat{f}\|_{\hat{T}} \right\}\ .$$

The desired inequality (4.8) nows follows by transforming from $\hat{T}$ back to T. □

Remark 4.2 : For a type I triangle the a priori estimate (4.8) is in fact an immediate consequence of the representation (4.1). In this case the minimum angle condition is not needed.

Remark 4.3 : If we select the bases for the trial spaces in such a way that all moments appearing in the boundary conditions are included in the degrees of freedom of the finite element, the computational cost can be reduced significantly. More specifically, for a type I element of odd order n in $\mathbf{M}_h^1$ and an element of either type in $\mathbf{M}_h^2$, we can express u_h in the form

$$\hat{u}_h = \sum_{i=1}^{3} \sum_{j=0}^{n-1} \hat{\psi}_{ij}^*(\hat{u}_h)\, \hat{\psi}_{ij} + \sum_{k=1}^{\sigma_{n-3}} \hat{\eta}_k^*(\hat{u}_h)\, \hat{\eta}_k (+ \hat{\xi}^*(\hat{u}_h)\, \hat{\xi}) \quad \text{for a type I (II) } \hat{T},$$

where $\{\hat{\psi}_{ij}^*,\ \hat{\eta}_k^*\}$ is the dual basis of $\{\hat{\psi}_{ij},\ \hat{\eta}_k\}$, a basis of $\mathbf{P}_n(\hat{T})$ satisfying

$$\int_{\hat{r}_l} \hat{\psi}_{ij}\, \hat{\tau}^m\, d\hat{\tau} = \delta_{il}\, \delta_{jm}, \quad (\hat{\psi}_{ij}, \hat{w}_q)_{\hat{T}} = 0 ;$$

$$\int_{\hat{r}_l} \hat{\eta}_k\, \hat{\tau}^m\, d\hat{\tau} = 0, \qquad (\hat{\eta}_k, \hat{w}_q)_{\hat{T}} = \delta_{kq},$$

for $i,\ l = 1,\ 2,\ 3$; $j,\ m = 0,\ 1,\ \ldots,\ n-1$; $k,\ q = 1,\ 2,\ \ldots,\ \sigma_{n-3}$ with $\{\hat{w}_q\}$ a basis of $\mathbf{P}_{n-3}(\hat{T})$; and

$$\hat{\xi} = \hat{s}\hat{t}^n - \sum_{i=1}^{3} \sum_{j=0}^{n-1} \hat{\psi}_{ij}^*(\hat{s}\hat{t}^n)\, \hat{\psi}_{ij} - \sum_{k=1}^{\sigma_{n-3}} \hat{\eta}_k^*(\hat{s}\hat{t}^n)\, \hat{\eta}_k .$$

For a type II triangle in $\mathbf{M}_h^1$ we can select the basis used in the proof of the uniqueness lemma. For an even order element of type I, it is still possible to obtain a basis possessing the desired properties. For example, when $n = 2$, we may take the average values and the first moments on the inflow side and one of the outflow sides and only the average on the other outflow side to be five out of six degrees of freedom required and complete them by a quadratic polynomial which is zero at two Gauss-Legendre points on each triangle side (see [5] or [1, Appendix A.3] for details). Under these special bases the systems to be solved actually have size $\sigma_{n-1} + 1$ for one-inflow-side elements and $\sigma_{n-2} + 3$ or $\sigma_{n-2} + 2$ for two-inflow-side elements.

5. STABILITY AND ERROR ESTIMATES FOR THE TRIANGULAR SCHEMES

In this section our intention is to derive some stability results for methods $\mathbf{M}_h^1$ and $\mathbf{M}_h^2$ and then to obtain error estimates as their consequence. The achievement of this goal is based on the employment of the *a priori* estimate

(4.8) established in the last section and some test functions depending only on t. The analysis framework constituted here will cover not only methods $\mathbf{M}_h^1$ and $\mathbf{M}_h^2$, but also any finite element method which adopts the variational equation (3.1) and satisfies (4.8), where the test space $V_{h,T}$ must contain all polynomials in the crosswind variable t of order $\leq n$. Thus the discontinuous Galerkin method is another typical example. For other possible schemes included in this framework, see [1, Appendix A.6].

We now proceed to establish some basic properties of u_h over T.

LEMMA 5.1 : *The solution u_h of Method $\mathbf{M}_h^1$ or $\mathbf{M}_h^2$ satisfies the following inequality on a triangle T of either type,*

$$\left|u_{h,\,\mathrm{out}}^- - u_{h,\,\mathrm{in}}^-\right|^2 + \left|u_{h,\,\mathrm{in}}^+ - u_{h,\,\mathrm{in}}^-\right|^2 + h\left\|(u_h)_s\right\|_T^2 \leq C\left\{\left|(I-P_t)\,u_{h,\,\mathrm{in}}^-\right|^2 + h\|f\|_T^2\right\}. \quad (5.1)$$

Proof : Set $w_h = u_h - EP_t\,u_{h,\,\mathrm{in}}^-$. Then (3.1), ..., (3.4) still hold with w_h in place of u_h. By Lemma 4.2 and the change of variables formula (2.6), we have

$$\|w_h\|_T \leq C\left\{h^{1/2}\left|w_h^-\right|_{\Gamma_{\mathrm{in}}(T)} + h\|f\|_T\right\} \leq C\left\{h^{1/2}\left|(I-P_t)\,u_{h,\,\mathrm{in}}^-\right| + h\|f\|_T\right\}.$$

Therefore by the inverse estimate (2.3),

$$\left\|(u_h)_s\right\|_T = \left\|(w_h)_s\right\|_T \leq Ch^{-1}\|w_h\|_T \leq C\left\{h^{-1/2}\left|(I-P_t)\,u_{h,\,\mathrm{in}}^-\right| + \|f\|_T\right\},$$

and by (2.4),

$$\begin{aligned}
&\left|u_{h,\,\mathrm{out}}^- - u_{h,\,\mathrm{in}}^-\right| + \left|u_{h,\,\mathrm{in}}^+ - u_{h,\,\mathrm{in}}^-\right| \\
&\leq \left|u_{h,\,\mathrm{out}}^- - P_t\,u_{h,\,\mathrm{in}}^-\right| + \left|u_{h,\,\mathrm{in}}^+ - P_t\,u_{h,\,\mathrm{in}}^-\right| + 2\left|(I-P_t)\,u_{h,\,\mathrm{in}}^-\right| \\
&\leq Ch^{-1/2}\left\|u_h - EP_t\,u_{h,\,\mathrm{in}}^-\right\|_T + 2\left|(I-P_t)\,u_{h,\,\mathrm{in}}^-\right| \\
&\leq C\left\{\left|(I-P_t)\,u_{h,\,\mathrm{in}}^-\right| + h^{1/2}\|f\|_T\right\}.
\end{aligned}$$

The desired inequality then follows from a suitable combination of the above two results. □

The next lemma will be used, together with the previous one, to help establish another local stability result that, unlike (4.8), can be iterated over the entire triangulation to obtain global stability of the methods.

LEMMA 5.2 : *There holds, for a triangle T of either type,*

$$\left|u_{h,\,\mathrm{out}}^-\right|^2 + \left|(I-P_t)\,u_{h,\,\mathrm{in}}^-\right|^2 \leq \left|u_{h,\,\mathrm{in}}^-\right|^2 + (f,\,EP_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-))_T + Ch\|f\|_T^2.$$

Proof: Take $v(t, s) = E[w(t)] \in V_{h,T}$ in (3.1). Then application of the integration by parts formula (2.2) and identities (2.5) and (2.6) yields

$$\langle u_{h,\,\mathrm{out}}^- - u_{h,\,\mathrm{in}}^-,\, w\rangle = (f, Ew)_T . \tag{5.2}$$

Selecting $w(t) = P_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-)(t)$, we have

$$|P_t\, u_{h,\,\mathrm{out}}^-|^2 = |P_t\, u_{h,\,\mathrm{in}}^-|^2 + (f, EP_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-))_T . \tag{5.3}$$

When T is of type I, note that $|u_{h,\,\mathrm{out}}^-|^2 = |P_t\, u_{h,\,\mathrm{out}}^-|^2 + |(I - P_t)\, u_{h,\,\mathrm{out}}^-|^2$ and

$$\begin{aligned} |(I - P_t)\, u_{h,\,\mathrm{out}}^-|^2 &= |(I - P_t)(u_{h,\,\mathrm{out}}^- - u_{h,\,\mathrm{in}}^-)|^2 \\ &\leqslant |u_{h,\,\mathrm{out}}^- - u_{h,\,\mathrm{in}}^-|^2 \leqslant Ch\,\|f\|_T^2 \end{aligned}$$

by the fact that $u_{h,\,\mathrm{in}}^- = P_t\, u_{h,\,\mathrm{in}}^-$ and Lemma 5.1. It then follows from (5.3) that

$$|u_{h,\,\mathrm{out}}^-|^2 \leqslant |u_{h,\,\mathrm{in}}^-|^2 + (f, EP_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-))_T + Ch\,\|f\|_T^2 .$$

When T is of type II, noting that $|u_{h,\,\mathrm{in}}^-|^2 = |P_t\, u_{h,\,\mathrm{in}}^-|^2 + |(I - P_t)\, u_{h,\,\mathrm{in}}^-|^2$ and $P_t\, u_{h,\,\mathrm{out}}^- = u_{h,\,\mathrm{out}}^-$, we then have

$$|u_{h,\,\mathrm{out}}^-|^2 + |(I - P_t)\, u_{h,\,\mathrm{in}}^-|^2 = |u_{h,\,\mathrm{in}}^-|^2 + (f, EP_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-))_T . \quad \square$$

Before combining Lemma 5.1 and Lemma 5.2 to get the local stability desired, we need the following identities to simplify its proof.

LEMMA 5.3 : *When T is of type I, then*

$$\begin{aligned} &P_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-)(t) \\ &\quad = 2\, u_{h,\,\mathrm{out}}^-(t) - 2(I - P_t) \int_0^{s_{\mathrm{out}}(t)} (u_h)_s\, ds - Pt \int_0^{s_{\mathrm{out}}(t)} f\, ds \qquad (5.4) \\ &\quad = 2\, u_{h,\,\mathrm{in}}^-(t) + P_t \int_0^{s_{\mathrm{out}}(t)} f\, ds ; \qquad (5.5) \end{aligned}$$

while when T is of type II, then

$$\begin{aligned} &P_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-)(t) \\ &\quad = 2\, u_{h,\,\mathrm{out}}^-(t) - P_t \int_{s_{\mathrm{in}}(t)}^0 f\, ds \qquad (5.6) \\ &\quad = 2\, u_{h,\,\mathrm{in}}^-(t) - 2(I - P_t)\, u_{h,\,\mathrm{in}}^-(t) + P_t \int_{s_{\mathrm{in}}(t)}^0 f\, ds . \qquad (5.7) \end{aligned}$$

Proof. For a triangle T of type I, by (5.2) and (2.7), we have

$$P_t\, u^-_{h,\,\mathrm{out}}(t) = u^-_{h,\,\mathrm{in}}(t) + P_t \int_0^{s_{\mathrm{out}}(t)} f\, ds\,. \tag{5.8}$$

Also it is easy to see that

$$u^-_{h,\,\mathrm{out}}(t) = u^+_{h,\,\mathrm{in}}(t) + \int_0^{s_{\mathrm{out}}(t)} (u_h)_s\, ds\,.$$

Noting that $(I - P_t)\, u^+_{h,\,\mathrm{in}} = 0$, the application of $(I - P_t)$ on both sides of the above identity yields

$$(I - P_t)\, u^-_{h,\,\mathrm{out}}(t) = (I - P_t) \int_0^{s_{\mathrm{out}}(t)} (u_h)_s\, ds\,.$$

This implies

$$P_t\, u^-_{h,\,\mathrm{out}}(t) = u^-_{h,\,\mathrm{out}}(t) - (I - P_t) \int_0^{s_{\mathrm{out}}(t)} (u_h)_s\, ds\,.$$

Hence,

$$\begin{aligned} P_t(u^-_{h,\,\mathrm{out}} + u^-_{h,\,\mathrm{in}})(t) &= 2\, P_t\, u^-_{h,\,\mathrm{out}}(t) - P_t \int_0^{s_{\mathrm{out}}(t)} f\, ds \\ &= 2\, u^-_{h,\,\mathrm{out}}(t) - 2(I - P_t) \int_0^{s_{\mathrm{out}}(t)} (u_h)_s\, ds \\ &\quad - P_t \int_0^{s_{\mathrm{out}}(t)} f\, ds\,. \end{aligned}$$

Also from (5.8),

$$P_t(u^-_{h,\,\mathrm{out}} + u^-_{h,\,\mathrm{in}})(t) = 2\, u^-_{h,\,\mathrm{in}}(t) + P_t \int_0^{s_{\mathrm{out}}(t)} f\, ds\,.$$

These are the desired identities (5.4) and (5.5).

The identities (5.6) and (5.7) for a triangle T of type II can be obtained directly from (5.2) upon noting that $P_t\, u^-_{h,\,\mathrm{out}} = u^-_{h,\,\mathrm{out}}$. □

THEOREM 5.1 : *For the solution u_h of Method $\mathbf{M}_h^1$ or $\mathbf{M}_h^2$, there exists a positive constant M independent of h, f and u such that for a triangle $T \in \Delta_h$*

$$\begin{aligned} &\left|u^-_{h,\,\mathrm{out}}\right|^2 - 2\left\langle u_{\mathrm{out}},\, u^-_{h,\,\mathrm{out}}\right\rangle + M\left\{h\left\|(u_h)_s\right\|^2_T + \left|u^+_{h,\,\mathrm{in}} - u^-_{h,\,\mathrm{in}}\right|^2\right\} \\ &\qquad \leqslant \left|u^-_{h,\,\mathrm{in}}\right|^2 - 2\left\langle u_{\mathrm{in}},\, u^-_{h,\,\mathrm{in}}\right\rangle + C\left\{h\|f\|^2_T + |u|^2_{T(T)}\right\}, \end{aligned}$$

where u is the exact solution of the model problem (2.1).

Proof · (i) When T is of type I, by (5.4) and (5.5),

$$
\begin{aligned}
&(f,\ EP_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-))_T \\
&\quad = (u_s,\ EP_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-))_T = \int_{\Gamma(T)} uEP_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-)\,\boldsymbol{\beta}\cdot\mathbf{n}\,d\tau \\
&\quad = \left\langle u_{\mathrm{out}},\ P_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-)\right\rangle - \left\langle u_{\mathrm{in}},\ P_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-)\right\rangle \\
&\quad = \left\langle u_{\mathrm{out}},\ 2\,u_{h,\,\mathrm{out}}^- - 2(I - P_t)\int_0^{s_{\mathrm{out}}(t)} (u_h)_s\,ds - P_t\int_0^{s_{\mathrm{out}}(t)} f\,ds\right\rangle \\
&\qquad - \left\langle u_{\mathrm{in}},\ 2\,u_{h,\,\mathrm{in}}^- + P_t\int_0^{s_{\mathrm{out}}(t)} f\,ds\right\rangle \\
&\quad \leqslant 2\left\langle u_{\mathrm{out}},\ u_{h,\,\mathrm{out}}^-\right\rangle - 2\left\langle u_{\mathrm{in}},\ u_{h,\,\mathrm{in}}^-\right\rangle + \varepsilon h\left\|(u_h)_s\right\|_T^2 \\
&\qquad + C(\varepsilon)|u|_{\Gamma(T)}^2 + Ch\|f\|_T^2 .
\end{aligned}
$$

Here ε is a positive constant to be determined and for the last inequality we have used the following estimation based on the Schwarz inequality, Lemma 2.1 and the arithmetic-geometric mean inequality :

$$
\begin{aligned}
&\left|\left\langle u_{\mathrm{out}},\ (I - P_t)\int_0^{s_{\mathrm{out}}(t)} (u_h)_s\,ds\right\rangle\right| \\
&\qquad \leqslant |u_{\mathrm{out}}|\left|(I - P_t)\int_0^{s_{\mathrm{out}}(t)} (u_h)_s\,ds\right| \\
&\qquad \leqslant |u|_{\Gamma_{\mathrm{out}}(T)}\cdot Ch^{1/2}\left\|(u_h)_s\right\|_T \leqslant \varepsilon h\left\|(u_h)_s\right\|_T^2 + C(\varepsilon)|u|_{\Gamma_{\mathrm{out}}(T)}^2 ,
\end{aligned}
$$

and analogously,

$$
\left|\left\langle u_{\mathrm{out}} + u_{\mathrm{in}},\ P_t\int_0^{s_{\mathrm{out}}(t)} f\,ds\right\rangle\right| \leqslant C\left\{h\|f\|_T^2 + |u|_{\Gamma_{\mathrm{out}}(T)}^2 + |u|_{\Gamma_{\mathrm{in}}(T)}^2\right\} .
$$

Combining the above results with Lemma 5.2 yields

$$
\begin{aligned}
\left|u_{h,\,\mathrm{out}}^-\right|^2 \leqslant \left|u_{h,\,\mathrm{in}}^-\right|^2 &+ 2\left\langle u_{\mathrm{out}},\ u_{h,\,\mathrm{out}}^-\right\rangle - 2\left\langle u_{\mathrm{in}},\ u_{h,\,\mathrm{in}}^-\right\rangle \\
&+ \varepsilon h\left\|(u_h)_s\right\|_T^2 + C(\varepsilon)|u|_{\Gamma(T)}^2 + Ch\|f\|_T^2 .
\end{aligned}
$$

Finally the desired inequality for this case is obtained by adding (5.1) to the above inequality and taking $\varepsilon = M = 1/2$.

(ii) When T is of type II, by (5.6) and (5.7)

$$
\begin{aligned}
(f, EP_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-))_T &= (u_s, EP_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-))_T \\
&= \langle u_{\mathrm{out}}, P_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-)\rangle - \langle u_{\mathrm{in}}, P_t(u_{h,\,\mathrm{out}}^- + u_{h,\,\mathrm{in}}^-)\rangle \\
&= \left\langle u_{\mathrm{out}}, 2u_{h,\,\mathrm{out}}^- - P_t\int_{s_{\mathrm{in}}(t)}^0 f\,ds\right\rangle - \left\langle u_{\mathrm{in}}, 2P_t u_{h,\,\mathrm{in}}^- + P_t\int_{s_{\mathrm{in}}(t)}^0 f\,ds\right\rangle \\
&= 2\langle u_{\mathrm{out}}, u_{h,\,\mathrm{out}}^-\rangle - 2\langle u_{\mathrm{in}}, u_{h,\,\mathrm{in}}^-\rangle + 2\langle u_{\mathrm{in}}, (I-P_t)u_{h,\,\mathrm{in}}^-\rangle \\
&\quad - \left\langle u_{\mathrm{out}} + u_{\mathrm{in}}, P_t\int_{s_{\mathrm{in}}(t)}^0 f\,ds\right\rangle \\
&\leqslant 2\langle u_{\mathrm{out}}, u_{h,\,\mathrm{out}}^-\rangle - 2\langle u_{\mathrm{in}}, u_{h,\,\mathrm{in}}^-\rangle + \frac{1}{2}\left|(I-P_t)u_{h,\,\mathrm{in}}^-\right|^2 \\
&\quad + C\left\{h\|f\|_T^2 + |u|_{\Gamma(T)}^2\right\}.
\end{aligned}
$$

Here we have used a similar estimation to part (i). By Lemma 5.2, we then have

$$
\begin{aligned}
&\left|u_{h,\,\mathrm{out}}^-\right|^2 - 2\langle u_{\mathrm{out}}, u_{h,\,\mathrm{out}}^-\rangle + \frac{1}{2}\left|(I-P_t)u_{h,\,\mathrm{in}}^-\right|^2 \\
&\qquad \leqslant \left|u_{h,\,\mathrm{in}}^-\right|^2 - 2\langle u_{\mathrm{in}}, u_{h,\,\mathrm{in}}^-\rangle + C\left\{h\|f\|_T^2 + |u|_{\Gamma(T)}^2\right\}. \qquad (5.9)
\end{aligned}
$$

From (5.1), there exists a positive constant M such that

$$
\begin{aligned}
&M\left\{h\|(u_h)_s\|_T^2 + \left|u_{h,\,\mathrm{in}}^+ - u_{h,\,\mathrm{in}}^-\right|^2\right\} \\
&\qquad \leqslant \frac{1}{2}\left\{\left|(I-P_t)u_{h,\,\mathrm{in}}^-\right|^2 + h\|f\|_T^2\right\}. \qquad (5.10)
\end{aligned}
$$

Adding (5.9) and (5.10) establishes the second part of the theorem. □

In order to state the global stability results, we need some additional notation. First let us recall that $\{S_j\}$ are the layers defined in Section 2. Then we define *Fronts* F_j as follows :

$$
\begin{aligned}
F_0 &= \Gamma_{\mathrm{in}}(\Omega), \\
F_j &= F_{j-1} \cup \Gamma_{\mathrm{out}}(S_j) - \Gamma_{\mathrm{in}}(S_j), \quad j = 1, 2, \ldots .
\end{aligned}
$$

Also we represent $\Omega_j = \bigcup_{k=1}^{j} S_k$.

THEOREM 5.2 : *If u_h is the solution of Method $\mathbf{M}_h^1$ or $\mathbf{M}_h^2$, then*

$$
\begin{aligned}
&\left|u_h^-\right|_{F_j}^2 + M\left\{h\|(u_h)_s\|_{\Omega_j}^2 + \sum_{T\subset\Omega_j}\left|u_{h,\,\mathrm{in}}^+ - u_{h,\,\mathrm{in}}^-\right|_{\Gamma_3(T)}^2\right\} \\
&\qquad \leqslant C\left\{\left|u_h^-\right|_{F_0}^2 + h\|f\|_{\Omega_j}^2 + \sum_{T\subset\Omega_j}|u|_{\Gamma(T)}^2\right\} \qquad (5.11)
\end{aligned}
$$

with M a positive constant. Furthermore

$$|u_h^-|^2_{\Gamma_{\text{out}}(\Omega)} + \|u_h\|^2_\Omega + h\|(u_h)_s\|^2_\Omega + \sum_{T\in\Delta_h} |u^+_{h,\text{in}} - u^-_{h,\text{in}}|^2_{\Gamma_3(T)}$$

$$\leq C\left\{|u_h^-|^2_{\Gamma_{\text{in}}(\Omega)} + h\|f\|^2_\Omega + \sum_{T\in\Delta_h} |u|^2_{\Gamma(T)}\right\}. \quad (5.12)$$

Proof: Recalling the inner product and norm notation we defined before, we see by (2.5) and (2.6),

$$|u^-_{h,\text{out}}|^2 = |u_h^-|^2_{\Gamma_{\text{out}}(T)}, \qquad |u^-_{h,\text{in}}|^2 = |u_h^-|^2_{\Gamma_{\text{in}}(T)};$$

and

$$\langle u_{\text{out}}, u^-_{h,\text{out}}\rangle = \langle u, u_h^-\rangle_{\Gamma_{\text{out}}(T)}, \quad \langle u_{\text{in}}, u^-_{h,\text{in}}\rangle = \langle u, u_h^-\rangle_{\Gamma_{\text{in}}(T)}.$$

By summing the inequality in Theorem 5.1 over all layers S_k, $1 \leq k \leq j$, we obtain

$$|u_h^-|^2_{F_j} - 2\langle u, u_h^-\rangle_{F_j} + M\left\{h\|(u_h)_s\|^2_{\Omega_j} + \sum_{T\subset\Omega_j} |u^+_{h,\text{in}} - u^-_{h,\text{in}}|^2_{\Gamma_3(T)}\right\}$$

$$\leq C\left\{|u_h^-|^2_{F_0} - 2\langle u, u_h^-\rangle_{F_0} + h\|f\|^2_{\Omega_j} + \sum_{T\subset\Omega_j} |u|^2_{\Gamma(T)}\right\}.$$

Thus

$$|u_h^-|^2_{F_j} + M\left\{h\|(u_h)_s\|^2_{\Omega_j} + \sum_{T\subset\Omega_j} |u_h^+ - u_h^-|^2_{\Gamma_{\text{in}}(T)}\right\}$$

$$\leq \frac{1}{2}|u_h^-|^2_{F_j} + 2|u|^2_{F_j} + C\left\{2|u_h^-|_{F_0} + |u|^2_{F_0} + h\|f\|^2_{\Omega_j} + \sum_{T\subset\Omega_j} |u|^2_{\Gamma(T)}\right\}$$

$$\leq \frac{1}{2}|u_h^-|^2_{F_j} + C\left\{|u_h^-|_{F_0} + h\|f\|^2_{\Omega_j} + \sum_{T\subset\Omega_j} |u|^2_{\Gamma(T)}\right\}.$$

Subtracting $(1/2)|u_h^-|^2_{F_j}$ and multiplying by 2 on both sides, we then establish the inequality (5.11).

From Lemma 4.2, we have

$$\|u_h\|^2_{S_j} \leq C\left\{h|u_h^-|^2_{F_{j-1}} + h^2\|f\|^2_{S_j}\right\}.$$

Applying (5.11), we obtain

$$\|u_h\|_{S_j}^2 \leq Ch\left\{|u_h^-|_{F_0}^2 + h\|f\|_{\Omega_j}^2 + \sum_{T\subset\Omega_j}|u|_{\Gamma(T)}^2 + h\|f\|_{S_j}^2\right\}$$

$$\leq Ch\left\{|u_h^-|_{F_0}^2 + h\|f\|_{\Omega_j}^2 + \sum_{T\subset\Omega_j}|u|_{\Gamma(T)}^2\right\}.$$

Summing over all layers S_j, $1 \leq j \leq N$ (N is the total number of the layers in Δ_h), and noting that $N = O(h^{-1})$ by Hypothesis $\mathbf{H}_3$, we infer

$$\|u_h\|_{\Omega}^2 \leq C\left\{|u_h^-|_{F_0}^2 + h\|f\|_{\Omega}^2 + \sum_{T\in\Delta_h}|u|_{\Gamma(T)}^2\right\}. \tag{5.13}$$

On the other hand, the application of (5.11) with $F_j = \Gamma_{\text{out}}(\Omega)$ yields

$$|u_h^-|_{\Gamma_{\text{out}}(\Omega)}^2 + h\|(u_h)_s\|_{\Omega}^2 + \sum_{T\in\Delta_h}|u_{h,\text{in}}^+ - u_{h,\text{in}}^-|_{\Gamma_3(T)}^2$$

$$\leq C\left\{|u_h^-|_{F_0}^2 + h\|f\|_{\Omega}^2 + \sum_{T\in\Delta_h}|u|_{\Gamma(T)}^2\right\}. \tag{5.14}$$

The result (5.12) is the sum of (5.13) and (5.14). □

Our final concern in this section is to derive error estimates. In fact, they are simply a corollary of Theorem 5.2.

Let u_I be any continuous interpolant of the exact solution u such that $u_I|_T \in \mathbf{P}_n(T)$ for any $T \in \Delta_h$ and satisfies

$$\|u - u_I\|_{j,T} \leq Ch^{n+1-j}\|u\|_{n+1,T}, \quad j = 0, 1; \tag{5.15}$$

$$|u - u_I|_{0,\Gamma(T)} \leq Ch^{n+1/2}\|u\|_{n+1,T}. \tag{5.16}$$

One example is that u_I interpolates u at σ_n equispaced points on T. It is well known that this interpolant satisfies the approximation properties given above (refer to [2, Chap. 3] for details).

We now set $e_h = u_h - u_I$. Then (3.1), ..., (3.4) remain valid when u_h and f are replaced by e_h and $(u - u_I)_s$, respectively. Hence we may apply the previous results with u replaced by $u - u_I$. By taking $u_h^- = u_I^- = g_I$ on $\Gamma_{\text{in}}(\Omega)$ and inserting (5.15) and (5.16) into Theorem 5.2, we now conclude

THEOREM 5.3 : *Let $u \in H^{n+1}(\Omega)$ be the solution of equation* (2.1) *and u_h the solution of the approximation* **Method** $\mathbf{M}_h^1$ *or* $\mathbf{M}_h^2$. *Then there exists a constant C independent of u and h such that*

$$\|u - u_h\|_{\Omega} \leq Ch^{n+1/2}\|u\|_{n+1,\Omega},$$

$$\|\beta\cdot\nabla(u - u_h)\|_{\Omega} \leq Ch^{n}\|u\|_{n+1,\Omega},$$

$$\left\{\int_{F_j}(u - u_h)^2\,d\tau\right\}^{1/2} \leq Ch^{n+1/2}\|u\|_{n+1,\Omega}, \quad \textit{for} \quad j = 1, 2, \ldots,$$

and

$$\left\{ \sum_{T \in \Delta_h} |u_h^+ - u_h^-|^2_{\Gamma_{\text{in}}(T)} \right\}^{1/2} \leq C h^{n+1/2} \|u\|_{n+1,\Omega} .$$

COROLLARY 5.1.

$$\|\nabla(u - u_h)\|_\Omega \leq C h^{n-1/2} \|u\|_{n+1,\Omega} .$$

This corollary can be easily derived by the inverse property. Its error order, however, may not be the best possible we can get in these two methods. The actual situations could be better (see Tables 6.1 and 6.2 in § 6).

6. NUMERICAL RESULTS

We now present some numerical results for our proposed schemes and compare them with the corresponding results for the continuous and discontinuous methods.

To generate a triangulation, we first divide the region Ω (for simplicity, we always select Ω to be a unit square in our experiments) uniformly into N^2 squares and then divide each square into four triangles. This is done by randomly selecting a common vertex in the neighborhood of the centroid with the property that all inflow sides of type II triangles are uniformly away from the characteristic direction. The resulting mesh is then nonuniform (*cf. fig.* 2). We shall approximate the solution of each test problem by both quadratic and linear elements.

Example 6.1 : Let us first consider the equation

$$\frac{1}{\sqrt{5}} \frac{\partial u}{\partial x} + \frac{2}{\sqrt{5}} \frac{\partial u}{\partial y} = 0 \quad \text{in} \quad \Omega = (0, 1) \times (0, 1) ,$$

with the initial data chosen to make the exact solution be $u = |z|^\alpha$, where $z = (2x - y)/\sqrt{5}$ is a coordinate orthogonal to the characteristic direction and α is a positive number to be selected later. Note that $|z|^\alpha \in H^{\alpha + 1/2 - \varepsilon}(\Omega)$ for any $\varepsilon > 0$. We shall estimate errors in the L^2 norm $\|u - u_h\|_\Omega$, the L^2 norm of the gradient $\|\nabla(u - u_h)\|_\Omega$ as well as the L^2 norm of the characteristic derivative $\|\beta \cdot \nabla(u - u_h)\|_\Omega$.

To see the rate of convergence under the regularity condition required in the theory, we select $\alpha = 2.5$ for the quadratic approximation and $\alpha = 1.5$ for the linear. Table 6.1 illustrates some numerical results for our reduced continuity (for brevity, RC) method $\mathbf{M}_h^1$ as well as for the continuous and discontinuous methods using quadratic approximations. We observe that

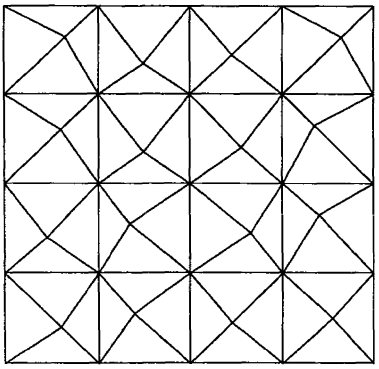

Figure 2. — Triangular mesh for $N = 4$

the order of L^2 error in u_h for $\mathbf{M}_h^1$ tends to be 2 5, while that in $\beta \nabla u_h$ is approaching 1 0 These agree well with our theoretical predictions and also show that our theoretical results for these two errors are best

TABLE 6 1

Numerical results for Example 6 1 Quadratic approximation $\alpha = 2\ 5$

$E_1 = \|u - u_h\|_\Omega$, $E_2 = \|\nabla(u - u_h)\|_\Omega$ and $E_3 = \|\beta \nabla(u - u_h)\|_\Omega$

	Continuous Method		RC Method $\mathbf{M}_h^1$		Discontinuous Method	
N	E_1	Rate	E_1	Rate	E_1	Rate
16	167 (-4)	2 18	354 (-5)	2 87	361 (-5)	2 87
32	368 (-5)	2 18	515 (-6)	2 78	523 (-6)	2 79
64	803 (-6)	2 19	791 (-7)	2 70	798 (-7)	2 71
128	174 (-6)	2 21	127 (-7)	2 63	128 (-7)	2 64
256	372 (-7)	2 22	214 (-8)	2 58	214 (-8)	2 58
N	E_2	Rate	E_2	Rate	E_2	Rate
16	744 (-3)	1 56	438 (-3)	1 87	365 (-3)	1 89
32	257 (-3)	1 53	121 (-3)	1 85	101 (-3)	1 85
64	900 (-4)	1 51	337 (-4)	1 85	284 (-4)	1 83
128	318 (-4)	1 50	948 (-5)	1 83	812 (-5)	1 80
256	113 (-4)	1 49	271 (-5)	1 81	237 (-5)	1 78
N	E_3	Rate	E_3	Rate	E_3	Rate
16	189 (-3)	1 83	260 (-3)	1 86	160 (-3)	1 87
32	521 (-4)	1 86	708 (-4)	1 88	435 (4)	1 88
64	141 (-4)	1 88	189 (-4)	1 90	116 (-4)	1 90
128	378 (-5)	1 90	504 (-5)	1 91	311 (-5)	1 91
256	100 (-5)	1 92	134 (-5)	1 92	824 (6)	1 92

possible under their regularity assumptions. We also see an interesting fact : the rate of the L^2 error in the gradient of u_h is about 1.8, much better than our theoretical result 1.5 in this case, which is derived from L^2 error estimates by the inverse property. Similar phenomena can be observed for the linear approximation from Table 6.2, where the RC method $\mathbf{M}_h^2$ is compared with the discontinuous Galerkin method which coincides with $\mathbf{M}_h^1$ in this case : linear, $a = f = 0$.

Comparing the data for the RC methods with those for the continuous and discontinuous methods in Tables 6.1 and 6.2, we see that the rates of convergence of the RC methods are close to their counterparts in the discontinuous ones while the convergence rates for the continuous schemes are slightly lower. All the experimental convergence rates match with their corresponding theoretical results. The fact that the number of unknowns increases as we go from the continuous to RC to discontinuous method is generally reflected in a corresponding decrease in absolute errors.

TABLE 6.2.

Numerical results for Example 6.1. Linear approximation. $\alpha = 1.5$.

$E_1 = \|u - u_h\|_\Omega$, $E_2 = \|\nabla(u - u_h)\|_\Omega$ and $E_3 = \|\beta \cdot \nabla(u - u_h)\|_\Omega$.

	RC Method $\mathbf{M}_h^2$		Discontinuous Method	
N	E_1	Rate	E_1	Rate
16	.426 (-3)	1.88	.385 (-3)	1.88
32	.124 (-3)	1.78	.113 (-3)	1.77
64	.385 (-4)	1.69	.356 (-4)	1.67
128	.127 (-4)	1.60	.120 (-4)	1.57
256	.440 (-5)	1.54	.422 (-5)	1.51
N	E_2	Rate	E_2	Rate
16	.356 (-1)	.92	.208 (-1)	.86
32	.192 (-1)	.89	.119 (-1)	.80
64	.103 (-1)	.89	.681 (-2)	.81
128	.561 (-2)	.88	.392 (-2)	.80
256	.308 (-2)	.87	.227 (-2)	.79
N	E_3	Rate	E_3	Rate
16	.232 (-1)	.92	.139 (-1)	.92
32	.123 (-1)	.92	.744 (-2)	.90
64	.640 (-2)	.94	.391 (-2)	.93
128	.334 (-2)	.94	.205 (-2)	.93
256	.174 (-2)	.94	.107 (-2)	.94

To see the effect of the additional differentiability in the exact solution, we also performed experiments in the cases $\alpha = 3$ for the quadratic and $\alpha = 2$ for the linear An improvement of the rate of convergence is observed. Except for the rate of L^2 error for the continuous method, which slightly lagged behind, the convergence rate for the other methods approached the optimal.

Example 6 2 The following equation is considered

$$\frac{1}{\sqrt{5}}\frac{\partial u}{\partial x} + \frac{2}{\sqrt{5}}\frac{\partial u}{\partial y} + u = \left(1 + \frac{3}{\sqrt{5}}\right)\exp(x+y) \quad \text{in} \quad \Omega = (0,1)\times(0,1)$$

Here we note that the lower order term a is nonzero and the exact solution $u = \exp(x + y)$ is a smooth function The computations (omitted here again) show that all methods discussed in Example 6.1 achieve their corresponding optimal order of convergence in this case

APPENDIX A

We shall give a counterexample to show that if Hypothesis $\mathbf{H}_2$ is violated, then the local stability inequality (4.8), which plays an important role in our analysis, will no longer be true. For simplicity, we only consider $\mathbf{M}_h^1$ with $n \geqslant 2$ in our example This approach can be applied to $\mathbf{M}_h^2$ upon slightly modifying the proof of Lemma 4.1 for this scheme.

Let $\{T_k\}$ be a sequence of unit isoceles right triangles of type II with respect to the characteristic direction $\beta_0 = (0, 1)^T$ such that the outflow side of T_k is its hypotenuse and the angle between β_0 and the left inflow side of T_k tends to zero Equivalently, we can consider the triangle $T = \Delta a_1 a_2 a_3$ with the various characteristic directions $\beta_k = (\cos\theta_k, \sin\theta_k)^T$, $k = 0, 1, 2, \ldots$, where $\lim_{k\to\infty} \theta_k = \theta_0 = \pi/2$, $a_1 = (1, 0)$, $a_2 = (0, 1)$, and $a_3 = (0, 0)$. As before, we denote $\Gamma_1 = \overline{a_2 a_3}$ and $\Gamma_2 = \overline{a_3 a_1}$.

For each $k = 0, 1, 2, \ldots$, let $u_{h\,k}$ be the discrete solution of $\mathbf{M}_h^1$ on T satisfying (3.1) and (3.3) with $\beta = \beta_k$, $u_{h\,k}^-\big|_{\Gamma_1} = 1$, and $u_{h\,k}^-\big|_{\Gamma_2} = f = 0$ Note that T has a side Γ_1 parallel to β_0 and is therefore a type I triangle with respect to β_0 by the original definition of the *type*. We can, however, assume it has a type II structure and the proof of Lemma 4 1 is still valid. Hence $u_{h,0}$ is well defined and $\|u_{h\,0}\|_T \neq 0$ since $u_{h,0}^-\big|_{\Gamma_1} = 1$.

In terms of a matrix formulation, each $u_{h\,k}$ corresponds to a coefficient vector U_k which is the solution of the linear system

$$A_k U_k = b_k$$

in the form of (4.11) and (4.12). It is not difficult to see that $A_k \to A_0$ and $b_k \to b_0$ as $k \to \infty$ in the standard Euclidean norms. The unique existence of $u_{h,0}$ implies the invertibility of A_0. Hence

$$U_k = A_k^{-1} b_k \to A_0^{-1} b_0 = U_0 .$$

We then have $\lim_{k \to \infty} u_{h,k} = u_{h,0}$ uniformly on T. Thus, $u_{h,k} \to u_{h,0}$ in $L^2(T)$. If now (4.8) remains valid for all $u_{h,k}$, we have

$$\begin{aligned} \| u_{h,k} \|_T^2 &\leq Ch \int_{\Gamma_{\text{in}}(T)} | u_{h,k}^- |^2 \, | \beta_k \cdot \mathbf{n} | \, d\tau \\ &= Ch \int_{\Gamma_1} | \beta_k \cdot \mathbf{n}_1 | \, d\tau = Ch \cos \theta_k . \end{aligned}$$

This leads to

$$0 \neq \| u_{h,0} \|_T^2 = \lim_{k \to \infty} \| u_{h,k} \|_T^2 = \lim_{k \to \infty} Ch \cos \theta_k = 0 ,$$

a contradiction. Therefore, (4.8) may not be true if $\mathbf{H}_2$ is violated.

APPENDIX B

In general, Hypothesis $\mathbf{H}_3$ is not true even though Hypotheses $\mathbf{H}_1$ and $\mathbf{H}_2$ are satisfied. The left picture in figure B.1 illustrates such a fact, where the number in each triangle indicates the layer to which it belongs. We see that there are about $O(h^{-2})$ many layers in this mesh. Moreover, we can easily see that most of the triangles in this mesh are obtuse triangles. This suggests that to obtain a triangulation with only $O(h^{-1})$ many layers we may need to pose some restrictions on these sort of triangles. In the following lemma, we give a sufficient condition for producing a mesh satisfying $\mathbf{H}_3$.

LEMMA B.1 : *Suppose that Δ_h is a triangulation satisfying Hypotheses $\mathbf{H}_1$ and $\mathbf{H}_2$ and that all inner angles of type II triangles are at most $\frac{\pi}{2}$. Then Δ_h satisfies $\mathbf{H}_3$.*

An example of Lemma B.1 is depicted in the right picture of figure B.1.

To prove Lemma B.1 we need some notation. Take $d = 2 \rho_{\min}$, where $\rho_{\min}$ is as defined in $\mathbf{H}_1$. Denote $\theta_{\min}$ as the minimal inner angle in Δ_h implied in $\mathbf{H}_1$, i.e., if θ is an inner angle of a triangle in Δ_h, then

$$\theta_{\min} \leq \theta .$$

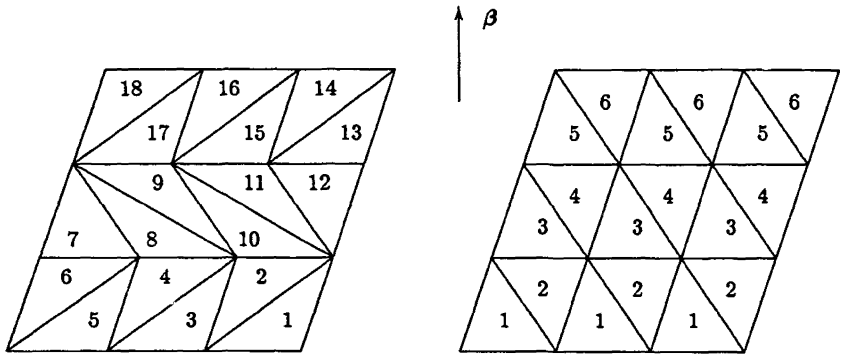

Figure B.1.

Also from $\mathbf{H}_2$, there exists a $\theta_0 > 0$ such that any acute angle between β and an inflow side of a type II triangle is at least θ_0. We then set

$$\theta_* = \min \{\theta_{\min}, \theta_0\} .$$

By a *path* from $T_b \in \Delta_h$ to $T_e \in \Delta_h$ we mean an ordered set of 2 or more triangles

$$T_1 \prec T_2 \prec \cdots \prec T_l$$

such that $T_1 = T_b$, $T_l = T_e$ and $\Gamma_{\text{out}}(T_{i-1}) \cap \Gamma_{\text{in}}(T_i)$ is the common side of T_{i-1} and T_i, $i = 2, \ldots, l$. The *size* of a path is the number of triangles in that path.

Place Ω in an orthogonal coordinate system (z, s) where the s direction is β. Then to each $T \in \Delta_h$ we associate a quantity s_* such that

$$s_*(T) = \min \{s(P) : P \in \Gamma_{\text{out}}(T)\} .$$

We can now state and prove a lemma which will be used in the proof of Lemma B.1.

LEMMA B.2 : *Under the assumptions of Lemma* B.1,

(i) if $T_1 \prec T_2$, *then*

$$s_*(T_1) \leqslant \min_{P \in T_2} s(P) \leqslant s_*(T_2) ;$$

(ii) *if* $T_1 \prec T_2$ *and* T_2 *is of type II, then*

$$s_*(T_2) \geqslant s_*(T_1) + d \sin \theta_0 ;$$

(iii) *if* $T_1 \prec T_2 \prec \cdots \prec T_l$ *is a path consisting of type I triangles and*

$$l > \left[\frac{\pi}{\theta_{\min}} \right],$$

then there exists a $k(\leq l)$ *such that*

$$s_*(T_k) \geq s_*(T_1) + d \sin \theta_{\min}.$$

Proof: (i) If T_2 is of type I, then $s_*(T_2) = \min_{P \in T_2} s(P)$. If T_2 is of type II, by the assumption on type II triangles in Lemma B.1 and $\mathbf{H}_2$, we have $s_*(T_2) > \min_{P \in T_2} s(P)$. Hence, for either case the following inequalities hold

$$s_*(T_2) \geq \min_{P \in T_2} s(P) = \min_{P \in \Gamma_{\text{in}}(T_2)} s(P) \geq \min_{P \in \Gamma_{\text{out}}(T_1)} s(P) = s_*(T_1).$$

(ii) Let $T_2 = \Delta a_1 a_2 a_3$, labeled counterclockwise, with the outflow side $\overline{a_1 a_2}$. Suppose $\overline{a_2 a_3}$ is the common side shared by T_1 and T_2. For a type II triangle, since all of its inner angles are no bigger than $\pi/2$ by the assumptions in Lemma B.1 and an acute angle between an inflow side and β is at least θ_0 by $\mathbf{H}_2$, this angle is also at most $\pi/2 - \theta_0$. If $s(a_1) \geq s(a_2)$, noting that $|\overline{a_2 a_3}| \geq d$, we then find

$$s(a_2) \geq s(a_3) + d \cos\left(\frac{\pi}{2} - \theta_0\right) = s(a_3) + d \sin \theta_0.$$

Thus,

$$s_*(T_2) = s(a_2) \geq s(a_3) + d \sin \theta_0 \geq s_*(T_1) + d \sin \theta_0.$$

For the case $s(a_1) < s(a_2)$, a similar argument on $\overline{a_3 a_1}$ will lead to the same conclusion.

(iii) Let $T_1 = \Delta a_1 a_2 a_3$ with the inflow side $\overline{a_1 a_2}$ and let $\overline{a_2 a_3}$ be the common side of T_1 and T_2. For the case when the common side of T_1 and T_2 is $\overline{a_1 a_3}$, the following argument remains valid with a_2 replaced by a_1.

If one of a_2 and a_3, say a_2, is the common vertex for all T_i, $1 \leq i \leq l$, then all T_i must lie on the same side of the line passing a_2 and parallel to β (see *fig.* B.2(*a*)). Since $< a_3 a_2 a_{l+2} \leq \pi - \theta_{\min}$ and $< a_{i+1} a_2 a_{i+2} \geq \theta_{\min}$, $2 \leq i \leq l$, by $\mathbf{H}_1$, we have

$$l \leq \left[\frac{\pi}{\theta_{\min}} \right].$$

This violates the assumption on l. Therefore, there must exist a $k(\leqslant l)$ such that all T_i, $1 \leqslant i \leqslant k-1$, except T_k have a common vertex. k cannot be $\leqslant 3$. This is consistent with $k \leqslant l$ since $l > \left[\frac{\pi}{\theta_{\min}} \right] \geqslant 3$. Again assume that the common vertex is a_2.

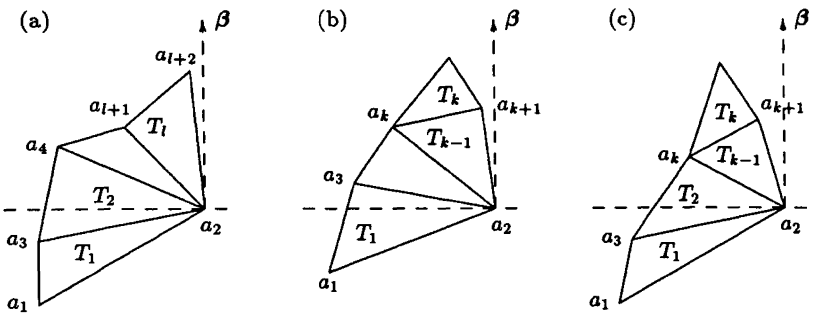

Figure B.2.

If $s(a_3) \geqslant s(a_2)$ (see *fig.* B.2(*b*)), since all acute angles between $\overline{a_i a_2}$ and β, $4 \leqslant i \leqslant k+1$, are at most $\pi/2 - \theta_{\min}$, we have

$$s_*(T_k) = \min \{s(a_k), s(a_{k+1})\}$$
$$\geqslant s(a_2) + d \sin \theta_{\min} \geqslant s_*(T_1) + d \sin \theta_{\min}.$$

If $s(a_3) < s(a_2)$ (see *fig.* B.2(*c*)), since s_* is nondecreasing along a path, we obtain

$$s_*(T_k) \geqslant s_*(T_2) = s(a_3)$$
$$\geqslant s(a_1) + d \sin \theta_{\min} = s_*(T_1) + d \sin \theta_{\min}.$$

Analogously, we can obtain these results when that common vertex is a_3. The conclusion in this part is then proved. □

Proof of Lemma B.1 : Consider any path $T_1 \prec T_2 \prec \cdots \prec T_k$ such that $k = \left[\frac{\pi}{\theta_{\min}} \right] + 2$. If T_i is of type II for some $2 \leqslant i \leqslant k$, then $s_*(T_i) \geqslant s_*(T_{i-1}) + d \sin \theta_0$ by (ii). Otherwise, by (iii) we have $s_*(T_j) \geqslant s_*(T_2) + d \sin \theta_{\min}$ for some $2 \leqslant j \leqslant k$. Since s_* is nondecreasing along a path by (i), we conclude that

$$s_*(T_k) \geqslant s_*(T_1) + d \sin \theta_*.$$

For any path from T_1 to T_L of size

$$L = \left(\left[\frac{\pi}{\theta_{\min}} \right] + 2 \right) \left(\left[\frac{1}{\sin \theta_*} \right] + 1 \right) (M + 1),$$

where M is the uniform upper bound for h/d as indicated in $\mathbf{H}_1$, consider it to be a union of disjoint subpaths of size $\left[\frac{\pi}{\theta_{\min}} \right] + 2$. Then the application of the above discussion and (i) yields

$$\begin{aligned} s_*(T_L) &\geqslant s_*(T_1) + d \sin \theta_* \left(\left[\frac{1}{\sin \theta_*} \right] + 1 \right) (M + 1) \\ &\geqslant s_*(T_1) + d(M + 1) \geqslant s_*(T_1) + h + d . \end{aligned}$$

Since the diameter of a triangle is no bigger than h, we have

$$\min_{P \in T_L} s(P) > s_*(T_L) - h \geqslant s_*(T_1) + d . \tag{B.1}$$

Now consider a partition of the interval $\left[\inf_{P \in \Omega} s(P), \sup_{P \in \Omega} s(P) \right]$:

$$\inf_{P \in \Omega} s(P) = s_0 < s_1 < \cdots < s_N = \sup_{P \in \Omega} s(P)$$

such that $s_i - s_{i-1} = d$, $i = 1, 2, \ldots, N - 1$, and $s_N - s_{N-1} \leqslant d$. Then $N \leqslant C/h$ by $\mathbf{H}_1$. To each subinterval $[s_{i-1}, s_i)$ we associate a strip subset of Ω such that a point of Ω lies in that strip if and only if its s coordinate belongs to $[s_{i-1}, s_i)$. Then Ω is decomposed into N strips.

Let us start with the first strip of Ω (corresponding to $[s_0, s_1)$). It can only overlap with at most $L - 1$ layers. Otherwise, there exists a path from T_b to T_e of size at least L such that

$$s_0 \leqslant \min_{P \in T_b} s(P) \leqslant s_*(T_b) \leqslant \min_{P \in T_e} s(P) \leqslant s_1 .$$

Hence $\min_{P \in T_e} s(P) - s_*(T_b) \leqslant s_1 - s_0 = d$. This contradicts (B.1). Therefore, the total number of layers is no bigger than $LN \leqslant C/h$, where C depends on $\theta_{\min}$ and θ_0 but not h. □

ACKNOWLEDGMENT

We would like to thank Gerard Richter for several helpful discussions and the referee for a careful reading of this manuscript and many helpful comments that led to this revised version.

REFERENCES

[1] D. CAI, Reduced continuity finite element methods for hyperbolic equations, *Ph. D. Dissertation,* Rutgers University, 1991.

[2] P. G. CIARLET, The Finite Element Method for Elliptic Equations, *North-Holland,* Amsterdam, 1978.

[3] R. S. FALK and G. R. RICHTER, Analysis of a continuous finite element method for hyperbolic equations, *SIAM J. Numer. Anal.,* 24 (1987), pp. 257-278.

[4] R. S. FALK and G. R. RICHTER, Local estimates for a finite element method for hyperbolic and convection-diffusion equations, SIAM J. Numer. Anal., 29 (1992), pp. 730-754.

[5] M. FORTIN and M. SOULIE, A non-conforming piecewise quadratic finite element on triangles, *Internat. J. Numer. Methods Engrg,* 19 (1983), pp. 505-520.

[6] T. J. R. HUGHES and A. BROOKS, A multidimensional upwind scheme with no crosswind diffusion, in Finite Element Methods for Convection Dominated Flows (T. J. R. Hughes, ed.), *AMD (ASME),* 1979, pp. 19-35.

[7] C. JOHNSON, U. NÄVERT and J. PITKÄRANTA, Finite element methods for linear hyperbolic problems, *Comput. Methods Appl. Mech. Engrg,* 45 (1984), pp. 285-312.

[8] C. JOHNSON and J. PITKÄRANTA, An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation, *Math. Comp.,* 46 (1986), pp. 1-26.

[9] P. LESAINT and P. A. RAVIART, On a finite element method for solving the neutron transport equation, in Mathematical Aspects of Finite Elements in Partial Differential Equations (C. de Boor, ed.), *Academic Press,* New York, 1974, pp. 89-123.

[10] W. H. REED and T. R. HILL, Triangular mesh methods for the neutron transport equation, Los Alamos Scientific Laboratory, *Report LA-UR-73-479* (1973).

[11] G. R. RICHTER, An optimal-order error estimate for the discontinuous Galerkin method, *Math. Comp.,* 50 (1988), pp. 75-88.

[12] R. WINTHER, A stable finite element method for first-order hyperbolic systems, *Math. Comp.,* 36 (1981), pp. 65-86.