Chapter 2

The Discontinuous Galerkin Method for Hyperbolic Problems

In this chapter we shall specify the types of problems we consider, introduce most of our notation, and recall some theory on the DG method. A detailed description of the formulation of the DG method on a hyperbolic PDE is presented. The method presented in this chapter is valid for hyperbolic scalars of conservation laws in multiple space dimensions. The aim of this chapter is to fix some notation and the essential requirements for a DG method to fit in our theory.

2.1 Notations

In this thesis we use the following notations:

The $\mathcal{L}^2$ inner product of two integrable functions u and v is

$$< u, v >_{\mathcal{L}^2} = \int_{-1}^{1} u(x)v(x)dx, \tag{2.1}$$

and the subsequent induced norm is $||u||_{\mathcal{L}^2} = \sqrt{< u, u >}$.

We denote H^s to be the Sobolev space of square integrable functions with all derivatives $\frac{d^k u}{dx^k},\ k = 1, 2, .., s$ being square integrable and equipped with the norm

$$\|u\|_s^2 = \sum_{k=0}^{s} \left\| \frac{d^k u}{dx^k} \right\|_{\mathcal{L}^2}^2,\ s = 1, 2, \ldots. \tag{2.2}$$

The $\mathcal{L}^2$ norm of a function $f(x,y)$ over a region Ω is

$$||f||_{\mathcal{L}^2(\Omega)} = \left(\iint_{\Omega} f^2(x,y) dxdy \right)^{\frac{1}{2}}. \tag{2.3}$$

The local error is denoted by $\epsilon = u - U$, where u and U, respectively, denote the exact and numerical solutions.

If we partition the domain Ω into N triangular elements $\Delta_j, j = 1, ..., N$, the maximum error at the shifted roots of the $(p+1)-$ degree Legendre polynomial over all $outflow$ edges is denoted by

$$\|u - U\|_{\infty}^{*}(t) = \max_{1 \le j \le N} \max_{1 \le i \le p+1} |(u-U)(x_{j,i}, y_{j,i}, t)|, \tag{2.4}$$

where $(x_{j,i}, y_{j,i}), i = 1, 2, \cdots, p+1$, are the coordinates of the shifted roots of Legendre polynomial on the $outflow$ edge of Δ_j.

The divergence of a differentiable vector function $\boldsymbol{a}(x,y) = [a_1(x,y), a_2(x,y))]^T$ is

$$\nabla . \boldsymbol{a} = \frac{\partial a_1}{\partial x} + \frac{\partial a_2}{\partial y}. \tag{2.5}$$

The gradient of a function $u = u(x,y)$ is

$$\nabla u = \frac{\partial u}{\partial x}\boldsymbol{i} + \frac{\partial u}{\partial y}\boldsymbol{j} = [u_x, u_y]^T. \tag{2.6}$$

The Jacobian of a differentiable vector function $\boldsymbol{a}(x,y)$ is

$$\boldsymbol{J}[\boldsymbol{a}(x,y)] = \begin{pmatrix} \frac{\partial a_1}{\partial x} & \frac{\partial a_1}{\partial y} \\ \frac{\partial a_2}{\partial x} & \frac{\partial a_2}{\partial y} \end{pmatrix}. \tag{2.7}$$

The determinant of a matrix function $\boldsymbol{J}$ is denoted $det(\boldsymbol{J})$.

We consider meshes $\Im_h$ that partition the domain Ω into triangles with possible hanging nodes. The parameter h denotes the mesh size of $\Im_h$ given by $h = \max_j(h_j)$, where h_j is the diameter of the element $\Delta_j \in \Im_h$.

The jump of a function v at (x,y) on an edge Γ of Δ_j is defined as

$$[v](x,y) = v^+(x,y) - v^-(x,y), \tag{2.8a}$$

where

$$v^-(x,y) = \lim_{s \to 0^+} v((x,y) + s\boldsymbol{n}), \quad v^+(x,y) = \lim_{s \to 0^-} v((x,y) + s.\boldsymbol{n}) \tag{2.8b}$$

and $\boldsymbol{n}$ denotes the outward unit normal vector to Γ. We will define any other notation as needed.

Next, we provide a review of some aspects of the theory of hyperbolic conservation laws and recall some theory on the method of characteristics.

2.2 Conservation Laws

Hyperbolic PDEs are used to model conservation of physical quantities such as mass, momentum, and energy in a fluid flow. Here we give an overview of the theoretical results and notions for conservation laws where we introduce the solution theory following [28, 38, 61, 62, 39].

We consider a scalar quantity $\rho(t,\boldsymbol{x}) : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ (density of a streaming fluid) which describes the state of the quantity in a point $\boldsymbol{x} \in \mathbb{R}^n$ at time t. The quantity accumulated in a volume Ω at time t is

$$Q(t,\Omega) = \int_{\Omega} \rho(t,\boldsymbol{x}) d\boldsymbol{x}. \tag{2.9}$$

Assuming that the velocity of the gas at the point $\boldsymbol{x}$ at time t is given by $v(t,\boldsymbol{x})$ the flow rate or flux of the gas is given by $\rho(t,\boldsymbol{x})v(t,\boldsymbol{x})$. Since we are interested in the change of the mass in the volume Ω in time, we have to examine the derivative with respect to time for (2.9). Due to the physical principle we like to reveal this is balanced by the flow through the surface of the volume, *i.e.*

$$\frac{d}{dt}\int_{\Omega} \rho(t,\boldsymbol{x}) d\boldsymbol{x} = -\int_{\partial\Omega} \rho(t,\boldsymbol{x}) v(t,\boldsymbol{x}).\boldsymbol{n} d\boldsymbol{s}. \tag{2.10}$$

where $\partial\Omega$ is the boundary of Ω. If the density is sufficiently smooth, interchanging differentiation and integration and applying the divergence Theorem to the right-hand side, we derive

$$\int_{\Omega} [\rho_t(t,\boldsymbol{x}) + \nabla.(\rho(t,\boldsymbol{x})v(t,\boldsymbol{x}))] d\boldsymbol{x} = 0. \tag{2.11}$$

Since this must hold for an arbitrary control volume, (2.11) holds pointwise and the integrand has to be identically zero, *i.e.*

$$\rho_t(t,\boldsymbol{x}) + \nabla.(\rho(t,\boldsymbol{x})v(t,\boldsymbol{x})) = 0. \tag{2.12}$$

This is the divergence form of the conservation of mass, while (2.11) is called the integral form. Hence we have derived the mathematical model for conservation of mass.

The hyperbolic nature of the equations considered is strongly related to the advection or wave spreading, which is modeled by this class of PDEs. The simplest model for this equation type is the scalar wave equation with constant speed. This equation describes the transport of a scalar quantity u depending on the direction and the velocity. First, we assume the simplest form of this process, which means taking the velocity vector as constant, *i.e.* $v(t,\boldsymbol{x}) = \boldsymbol{a}$, we obtain

$$u_t + \boldsymbol{a}.\nabla u = 0. \tag{2.13}$$

As one can easily see the development in time is balanced by a drift or transport with velocity $\boldsymbol{a}$. So the change of $u(t,\boldsymbol{x})$ depends on the scalar product $< \boldsymbol{a}, \nabla u >$. For simplicity we assume that initial conditions only consist of constant states u^- and u^+ separated by a

discontinuity Σ. The true solution of the Cauchy problem (2.13) with initial condition $u(0, \boldsymbol{x}) = u_0(\boldsymbol{x})$ is simply

$$u(t, \boldsymbol{x}) = u_0(\boldsymbol{x} - \boldsymbol{a}t). \tag{2.14}$$

Here, one immediately sees that the initial data propagate with velocity $\boldsymbol{a}$ in space-time along the rays $\boldsymbol{x} - \boldsymbol{a}t = \boldsymbol{x_0}$. And indeed considering a one-dimensional example one can see, that the initial data $u(0, x) = u_0(x)$ are translated to the right (resp. to the left) for $a > 0$ (resp. $a < 0$).

Looking more closely at the solution (2.14) and drawing it at time t_1 and t_2 into a space-time diagram, we clearly see that the profile of the initial data is transported along these rays. They are called characteristic curves and are defined as the integral curves of the differential equation.

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{a}. \tag{2.15}$$

If we examine the change of the solution along theses curves we see that

$$\frac{d}{dt}u(t, \boldsymbol{x}(t)) = u_t(t, \boldsymbol{x}) + \nabla u(t, \boldsymbol{x})\frac{d\boldsymbol{x}}{dt} = u_t + \boldsymbol{a}.\nabla u = 0. \tag{2.16}$$

This shows that the solution u is constant along these characteristics. Since we have a linear equation and $\boldsymbol{v} = \boldsymbol{a}$ is constant, the characteristic curves are straight lines, which corresponds to linear advection or linear transport. Since u is constant along a characteristic then it must have the same value it had initially.

As stated above, a general conservation law in several space dimensions has the form (1.3). If we restrict ourselves to scalar conservation laws in several space dimensions, we obtain

$$u_t + \nabla.\boldsymbol{F}(u) = 0, \quad u(0, \boldsymbol{x}) = u_0(\boldsymbol{x}). \tag{2.17}$$

with $\boldsymbol{F}(u) = [f_1(u), ..., f_n(u)]^T : \mathbb{R} \to \mathbb{R}^n$. This general form is called conservative formulation. If we assume u is a classical solution of (2.17) we can carry out the differentiation of the flux vector $\boldsymbol{F}(u)$ and derive the nonconservative form of (2.17), *i.e.*

$$u_t + \boldsymbol{F}'(u).\nabla u = 0, \quad u(0, \boldsymbol{x}) = u_0(\boldsymbol{x}). \tag{2.18}$$

With $\boldsymbol{F}'(u) = \boldsymbol{a}(u)$ we find a form similar to (2.13) now with the nonlinear velocity $\boldsymbol{a}(u)$. Thus, the characteristics now have a more general form than (2.15) and we see that the change in time is balanced by the derivative of the flux vector with respect to u.

For simplicity let us restrict ourselves to the one-dimensional case and assume that the flux function is $f(x)$.

$$u_t + f'(u)u_x = u_t + a(u)u_x = 0, \quad u(0, x) = u_0(x). \tag{2.19}$$

Now consider the rate of change of $u(t, x(t))$ as measured by a moving observer, $x = x(t)$. Then the total derivative of u would be

$$\frac{d}{dt}u(t, x(t)) = u_t(t, x) + u_x(t, x)\frac{dx}{dt}. \tag{2.20}$$

If the observer moves with velocity $a(u)$ then $\frac{dx}{dt} = a(u)$. Hence (2.20) leads to

$$\frac{d}{dt}u(t,x(t)) = u_t(t,x) + u_x(t,x)\frac{dx}{dt} = u_t(t,x) + a(u)u_x(t,x) = 0. \tag{2.21}$$

which implies that u is a constant along the curve $\frac{dx}{dt} = a(u)$. This curve is called a characteristic curve, characteristic line, or simply a characteristic of the differential equation (2.19). Having (2.21), we have replaced the PDE (2.19) with two coupled ordinary differential equations,

$$\frac{dx}{dt} = a(u), \qquad \frac{du}{dt} = 0. \tag{2.22}$$

Since u is constant along a characteristic, it must have the same value it had initially. For simplicity, consider a characteristic that passes through the point $(t = 0, x_0)$. From the initial conditions we have $u(0, x_0) = u_0(x_0)$ and thus $u = u_0(x_0)$ along the characteristic. This particular characteristic satisfies the ODE, $\frac{dx}{dt} = a(u_0(x_0)) = a_0$, which states that the slope of the characteristic is a constant, a_0. Thus the characteristic is a straight line given by $x(t) = x_0 + a_0 t$.

If we assume that u is a classical solution of (2.19) then the characteristic curves are straight lines along which the solution is constant. We can also see that the curves are straight lines with constant slopes depending on the initial data.

In fact, each characteristic is a straight line but with different slopes resulting from different initial values. This means that characteristic lines can be parallel, can intersect one another, or diverge from one another. We get parallel characteristics when $a(u)$ is a constant. Figure 2.1 shows characteristic that are parallel and move with constant velocity $a(u)$, characteristics fan out since the slope of the characteristics take on both positive and negative values and two characteristic curves intersecting each other (characteristic lines for a shock wave), respectively.

When characteristics do not intersect, the unique and continuous solution of (2.19) is given implicitly by

$$u(t,x) = u_0(x - a(u)t). \tag{2.23}$$

Accordingly, to find the solution at a point (t, x) we must determine on which characteristic it lies and then trace that characteristic back to the initial value. Then, since u is constant along characteristics, the value of u at (t, x) is the same as u at $(0, x)$.

Here we discuss the solution to the following scalar partial differential equation (two-dimensional one-way wave equation):

$$u_t + \alpha u_x + \beta u_y = 0, \quad u(x, y, 0) = u_0(x, y). \tag{2.24}$$

This PDE is the two-dimensional analogue to the PDE discussed above with the exact solution

$$u(x, y, t) = u_0(x - \alpha t, y - \beta t). \tag{2.25}$$

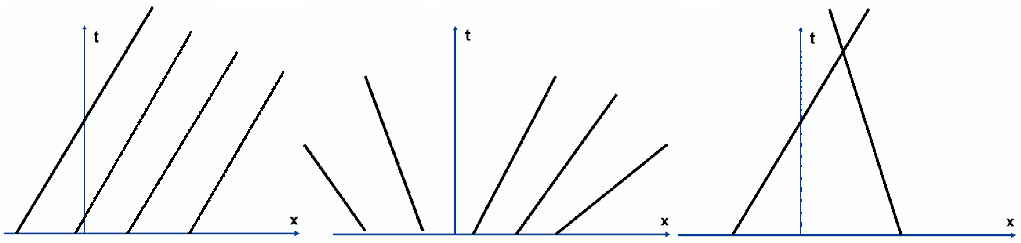

Figure 2.1: Different characteristic lines.

In words, the initial condition is translated with velocity $[\alpha, \beta]^T$, *i.e.*, the solution does not change shape, it simply translates with a constant velocity.

At this point the question naturally arises what happens to a conservation law with nonlinear flux function like (2.19) if characteristic lines intersect and so which characteristic do you follow back to $t = 0$? In the linear case (2.13), this problem does not arise, since the slopes of the characteristic curves are constant, *i.e.* independent of the initial condition. As we have seen before this is not true for the nonlinear case. Since the solution u is constant along a characteristic we may get many solutions if characteristic lines cross each other. For simplicity let us restrict to the one-dimensional case and assume that the flux function f satisfies $f^{'}(u_0(x_1)) > f^{'}(u_0(x_2))$ with $f^{'}(u_0(x_1)) \neq 0$, $f^{'}(u_0(x_2)) \neq 0$ for two points $x_1 < x_2$, *i.e.*

$$m_1 = \frac{1}{f^{'}(u_0(x_1))} < \frac{1}{f^{'}(u_0(x_2))} = m_2. \tag{2.26}$$

Since the slope m_2 corresponding to the characteristic C_2 is steeper than slope m_1 of the characteristic C_1, they necessarily intersect at some point P at time

$$t = \frac{x_2 - x_1}{f^{'}(u_0(x_1)) - f^{'}(u_0(x_2))}. \tag{2.27}$$

In order to show this, let x_s be the position where the shock forms and propagates in time. Since u can not take both values $u_0(x_1)$ and $u_0(x_2)$, a discontinuity has to arise at the point P. The solution breaks and a shock forms. We use the fact that shocks form where characteristic lines intersect. Considering two neighboring characteristics, one derived from x_1, the other from x_2, both at time $t = 0$, we obtain

$$x = x_1 + a(u_0(x_1))t, \quad x = x_2 + a(u_0(x_2))t. \tag{2.28}$$

These two characteristics only intersect at a positive time if $a(u_0(x_1)) > a(u_0(x_2))$. Solving for the intersection point by setting $x_1 + a(u_0(x_1))t = x_2 + a(u_0(x_2))t$ and solving for t we get (2.27).

Taking the limit as $x_2 \to x_1$ we have

$$t = \frac{-1}{\frac{da}{dx_1}}. \tag{2.29}$$

From this we see that characteristics only intersect for $t > 0$ provided $\frac{da(u_0(x_1))}{dx_1} < 0$. To find the first time t that a shock forms we must minimize (2.29) over all times $t > 0$. This can be accomplished by solving

$$\frac{d^2}{dx_1^2}(a(u_0(x_1))) = 0. \tag{2.30}$$

Note that we have not assumed special properties of u and f, so the solution is independent of the smoothness of both functions. This behavior is very special for our class of equations: the possible development of discontinuous solutions from smooth initial data. Here the notion of classical solution fails and a new concept is needed. The above considerations have clearly shown, that classical solutions are not sufficient to resolve (2.17). This dilemma is solved by the notion of weak solutions which means we have to consider solutions in the sense of distributions.

Not surprisingly, a classical solution of the Cauchy problem (2.17) is also a weak solution of the problem. On the other hand, every distributional solution is a classical solution of (2.17) in any domain where u is C^1. For a detailed discussion on weak and measure-valued solutions for conservation laws see [56].

Unfortunately, weak solutions are by no means unique, *i.e.* the propagation velocity at which discontinuities propagate is not necessarily uniquely determined. Furthermore, not every discontinuity is admissible. A necessary condition to the jump discontinuity is given by the Rankine-Hugoniot condition. This condition says the following, let x_s be the position where the shock forms and propagates in time, so that $x_s(t)$ is a smooth function of time with a shock velocity given by $\frac{dx_s}{dt}$ then

$$[u]\frac{dx_s}{dt} = [f]. \tag{2.31}$$

where the jump in a function v at (x, y) is defined in (2.8)

The Rankine-Hugoniot jump condition gives a criterion which solutions are admissible across a discontinuity. In the other word, this condition can be used to give a discontinuous solution u. Notice that the shock has velocity

$$\frac{dx_s}{dt} = \frac{[f]}{[u]}. \tag{2.32}$$

For more information consult [56].

Contact discontinuity: A contact discontinuity occurs if the function f is affine on the interval limited by $u^-(x_0)$ and $u^+(x_0)$, *i.e.* $f'(u^-(x_0)) = f'(u^+(x_0)) = s$ which means that the characteristics run parallel to the discontinuity with speed s. A contact discontinuity is a persistent, discontinuous jump in mass density moving by bulk convection through the system. Since there is negligible mass diffusion, such a jump persists. These jumps usually

appear at the point of contact of different materials, for example, a contact discontinuity separates oil from water. Contacts move at the local bulk convection speed, or more generally the characteristic speed, and can be modeled by using step function initial data in the bulk convection equation. Since contacts are simply a bulk convection effect, they retain any perturbations they receive. Thus we expect contacts to be especially sensitive to numerical methods, *i.e.* any spurious alteration of the contact will tend to persist and accumulate. Note that there is no discontinuity in pressure or velocity across the contact discontinuity, but only in density.

Characteristics theory extends readily to more general nonhomogeneous PDEs and systems of conservation laws, see [35]. We use this theory to gain information about the problem, such as which direction information flows and the possibility of shocks forming. Knowing the direction of flow tells us where in the mesh to begin our computations since we need to progress through the elements in the direction of information flow. When we know a priori that a shock forms, we can employ techniques to improve the quality of the solution, such as doing an *hp*-refinement near the shock or using slope limiters to control spurious oscillations that can form and pollute the approximate solution. Additionally, adaptive methods can incorporate this theory to predict the location of a shock and follow its path, so that the adjustments mentioned earlier can be done automatically.

Next, we recall the DG method formulation for the hyperbolic problem.

2.3 The Discontinuous Galerkin Method

A DG method formulation requires partitioning the domain Ω into a collection of N elements such that $\cup_{j=1}^{N}\overline{\Delta_j}=\overline{\Omega}$ and constructing a Galerkin problem on one element Δ_j by multiplying (1.3a) by a test function $\boldsymbol{v}\in(\mathcal{L}^2(\Delta_j))^m$, integrating the result on Δ_j , and applying the divergence theorem to obtain

$$(\boldsymbol{v},\boldsymbol{u}_t)_{\Delta_j}-(\nabla\boldsymbol{v},\boldsymbol{F}(\boldsymbol{u}))_{\Delta_j}+<\boldsymbol{v},\boldsymbol{F_n}>_{\partial\Delta_j}=(\boldsymbol{v},\boldsymbol{r})_{\Delta_j},\quad\forall\boldsymbol{v}\in(\mathcal{L}^2(\Delta_j))^m. \tag{2.33a}$$

where the normal component of the flux $\boldsymbol{F_n(u)}=\boldsymbol{F(u).n}$, and $\boldsymbol{n}$ is the normal vector to $\partial\Delta_j$. The $\mathcal{L}^2$ volume and surface inner products are

$$(\boldsymbol{v},\boldsymbol{u})_{\Delta_j}=\int_{\Delta_j}\boldsymbol{v}^T\boldsymbol{u}d\tau,\quad<\boldsymbol{v},\boldsymbol{u}>_{\partial\Delta_j}=\int_{\partial\Delta_j}\boldsymbol{v}^T\boldsymbol{u}ds. \tag{2.33b}$$

To complete the construction of the numerical method, we must:

1- Select an approximation $\boldsymbol{U_j}\in\mathcal{P}_p(\Delta_j)$ for solution $\boldsymbol{u}$, where, typically, $\mathcal{P}_p$ consists of polynomials of degree p on Δ_j. That means that the space is replaced by a finite dimensional subspace. Let $S^{N,p}$ denote the space of piecewise polynomial functions $\boldsymbol{U_j}$ such that the

restriction of $\boldsymbol{U_j}$ to an element Δ_j is in $\mathcal{P}_p$. A basis for $\mathcal{P}_p(\Delta_j)$ is chosen to be orthogonal in $\mathcal{L}^2$ on Δ_j and this leads to the Dubiner basis commonly used with spectral methods. These basis functions will be defined later.

2- Approximate the normal component of the flux $\boldsymbol{F_n(u)}$ on $\partial\Delta_j$. Due to discontinuous finite element solutions the normal component of the flux in (2.33) is not defined on $\partial\Delta_j$. The usual strategy is to approximate it by a numerical flux $\hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{U_j}, \boldsymbol{U_{nbj}})$ that depends on the solution $\boldsymbol{U_j}$ on Δ_j and $\boldsymbol{U_{nbj}}$ on the neighboring element Δ_{nbj} sharing the portion of the boundary $\Delta_{j,nbj}$ common to both elements. The numerical flux is required to be consistent in the sense that $\hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{u}, \boldsymbol{u}) = \boldsymbol{F(u).n}$.

A good numerical flux should satisfy (i) locally lipschitz and $\hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{u}, \boldsymbol{u}) = \boldsymbol{F_n(u)}$, (ii) a nondecreasing function with respect to the first argument and (iii) a nondecreasing function with respect to the second argument. Examples of popular numerical fluxes: upwind flux, Lax-Friedrichs flux, Godunov flux, Roe flux [23, 26]. A common strategy [44] is to compute the numerical normal flux as the exact or approximate solution of a Riemann problem breaking on $\partial\Delta_{j,nbj}$. Recall that a Riemann problem is an initial value (Cauchy) problem with piecewise constant data. Upon choosing a numerical flux $\hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{U_j}, \boldsymbol{U_{nbj}})$ on the edge $\partial\Delta_{j,nbj}$ separating elements j and nbj, the DG method (2.33) becomes

$$(\boldsymbol{V}, \boldsymbol{U}_{j,t})_{\Delta_j} - (\nabla \boldsymbol{V}, \boldsymbol{F}(\boldsymbol{U_j}))_{\Delta_j} + \sum_{nbj=1}^{n_{\partial\Delta_j}} < \boldsymbol{V}, \hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{U_j}, \boldsymbol{U_{nbj}}) >_{\partial\Delta_j} = (\boldsymbol{V}, \boldsymbol{r})_{\Delta_j}, \quad \forall \boldsymbol{V} \in (\mathcal{P}_p(\Delta_j))^m. \tag{2.34}$$

where $n_{\partial\Delta_j}$ is the number of faces of Δ_j.

3- Evaluate volume and boundary inner products appearing in (2.33) by Gaussian quadratures of orders $2p$ and $2p+1$, respectively.

4- Select a time integration strategy, typically, this is performed by classical Runge-Kutta integration scheme [26] with a time step chosen according to the Courant-Friedrichs-Levy (CFL) condition.

Additional stabilization, known as limiting is needed to suppress spurious oscillations when $p > 0$. The choice of the numerical flux and the limiting strategy are key factors in the success or failure of the approach.

To solve the problem on one element and at fixed time, one needs only to know the $inflow$ boundary data and $\boldsymbol{r}$. Because the data is given on the $inflow$ boundary of the whole domain, one can find an order of elements such that the problem (2.34) can be solved element by element. Instead of solving a big system of linear equations, many small systems of linear equations has to be solved when using DG. Figure 2.2 shows an example of an order of triangles for the DG method applied to scalar-first order hyperbolic PDEs.

For hyperbolic problems, the global problem can be decoupled in a lot of small problems.

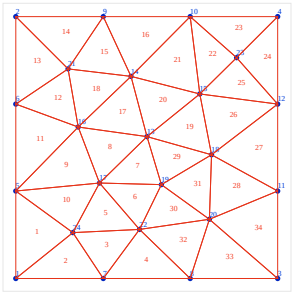

Figure 2.2: Example of an order of triangles for the DG method with $\boldsymbol{a} = [\alpha, \beta]^T$, $\alpha, \beta > 0$

For each element, a linear system has to be solved. Since the approximation does not have to be continuous, one can take an arbitrary basis of polynomials for each element. But the fact that the approximation is discontinuous implies also a larger number of unknowns. Note that the approximation is discontinuous, however; if the exact solution is continuous, the approximation will be discontinuous, but converges to the continuous solution. On the other hand, if the exact solution is discontinuous, with a good choice of the computational mesh, one may catch better the effects of the discontinuous solution than with the continuous Galerkin method.

2.4 Finite Element Spaces

The complete space of polynomials of degree p in 2-D is defined as

$$\mathcal{P}_k = \{q \,|\, q = \sum_{m=0}^{k} \sum_{i=0}^{m} c_i^m x^i y^{m-i}\}, \qquad k = 0, 1, \cdots, p. \tag{2.35}$$

The usual polynomial space in one-dimension is defined as

$$\mathbf{P}_k = \{q | q = \sum_{i=0}^{k} c_i x^i\}, \qquad k = 0, 1, \cdots, p. \tag{2.36}$$

In our error analysis we will also use the following spaces

$$\mathcal{V}_k = \mathcal{P}_k \cup \{x^i y^{k+1-i},\ i = 1, 2, \cdots, k\}, \qquad k = 0, 1, \cdots, p, \tag{2.37}$$

$$\mathcal{U}_k = \mathcal{P}_k \cup \{x^{k+1}, y^{k+1}\}, \qquad k = 0, 1, \cdots, p. \tag{2.38}$$

We note that $\mathcal{V}_0 = \mathcal{P}_0$, $\mathcal{U}_0 = \mathcal{P}_1$ and

$$\mathcal{P}_{p+1} \subset \mathcal{V}_p \cup span(\{x^{p+1}, y^{p+1}\}), \quad p \geq 1. \tag{2.39}$$

These spaces are suboptimal but they lead to a very simple *posteriori* error estimator which we present in chapter 4. The spaces $\mathcal{P}_p$, $\mathcal{V}_p$ and $\mathcal{U}_p$ will be used and they are easily visualized using Pascal's triangle, see Figure 2.3.

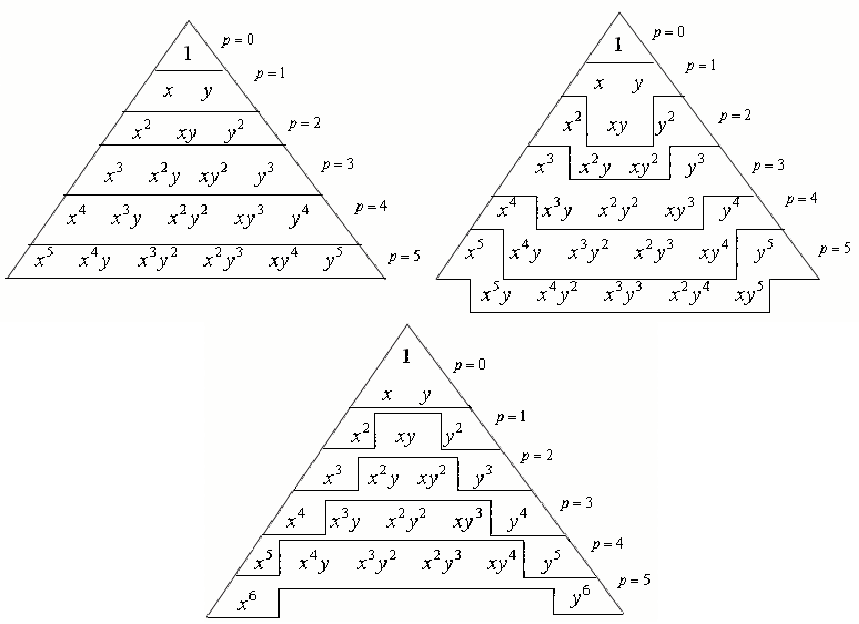

Figure 2.3: The spaces $\mathcal{P}_p$ (upper left), $\mathcal{V}_p$ (upper right) and $\mathcal{U}_p$ (bottom) for $p = 0$ to 5.

These spaces have the following dimensions $\dim(\mathcal{P}_p) = (p+1)(p+2)/2$, $\dim(\mathcal{V}_p) = (p+2)(p+3)/2 - 2$, and $\dim(\mathcal{U}_p) = (p+1)(p+2)/2 + 2$.

These spaces are employed to approximate solutions, we will see that the disadvantage of this approximation scheme is that different elements have different matrices. Moreover, the matrices might not be diagonal even though the basis is orthogonal. In this case, inverses of matrices have to be calculated, which might be ill-conditioned for high-degree polynomial basis functions.

2.5 Mappings for Triangle Elements

General domains can be most easily partitioned into triangular meshes. Any triangle can be linearly transformed into the standard reference triangle T_0.

We map a physical triangle Δ having vertices $(x_i, y_i),\ i = 1, 2, 3$, into the canonical triangle $(0, 0)$, $(1, 0)$ and $(0, 1)$ by a standard affine mapping. We will denote the coordinates in the physical element Δ by (x, y), while the coordinates in the reference element are denoted by (ξ, η). For convenience, we map element Δ onto a canonical right triangle $T_0 = \{(\xi, \eta), 0 \leq \xi, \eta \leq 1, 0 \leq \xi + \eta \leq 1\}$ (Figure 2.4) using the linear transformation

$$\begin{pmatrix} x(\xi,\eta) \\ y(\xi,\eta) \end{pmatrix} = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} + \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}. \tag{2.40}$$

The Jacobian of this linear transformation is the constant,

$$det(J_j) = det \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} = (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1). \tag{2.41}$$

The mapping between Ω_j and T_0 can be generally denoted by $(x, y) = (x(\xi, \eta), y(\xi, \eta))$.

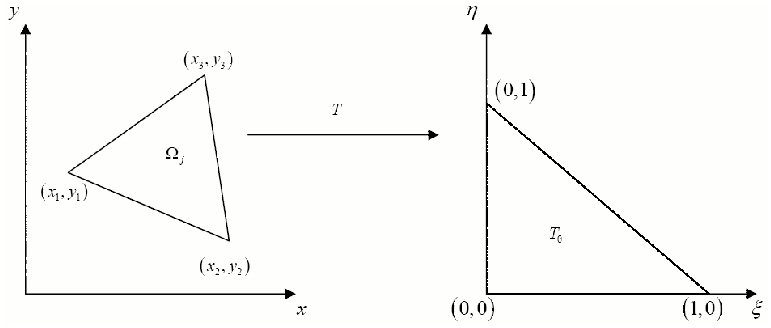

Figure 2.4: Mapping of a triangle Ω_j (left) onto a canonical triangle T_0 (right).

2.6 Orthogonal Polynomial Basis on Triangles

In this thesis, we use the polynomial space $\mathcal{P}_p = span\{\xi^i \eta^j, \quad 0 \leq i, j, i + j \leq p\}$ to approximate solutions. A natural basis for this space is $v_{ij} = \xi^i \eta^j, \quad 0 \leq i, j, i + j \leq p$, which is only used in practice for $p \leq 7$, and for larger p the basis becomes nearly dependent and leads to ill-conditioned problems. For this, we construct an orthogonal basis such as Dubiner basis. The Dubiner basis [30] on triangles is obtained by transforming Jacobi polynomials defined on [-1,1] to polynomials on triangles. The $n^{th}-$degree Jacobi polynomials $P_n^{\alpha,\beta}(x)$ on [-1,1] are orthogonal with respect to the weight function $w(x) = (1 - x)^\alpha (1 + x)^\beta$; *i.e.*,

$$\int_{-1}^{1} (1 - x)^\alpha (1 + x)^\beta P_n^{\alpha,\beta}(x) P_m^{\alpha,\beta}(x) dx = \frac{2^{\alpha+\beta+1} \Gamma(n + \alpha + 1)\Gamma(n + \beta + 1)}{(2n + \alpha + \beta + 1)n!\Gamma(n + \alpha + \beta + 1)} \delta_{nm}. \tag{2.42}$$

where Γ is the Gamma function defined by $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ and δ_{nm} is the Kronecker symbol equal to 1 if $n = m$ and 0, otherwise. In our analysis we will use some formulae for Jacobi polynomials [1] defined by the Rodrigues's Formula:

$$P_n^{\alpha,\beta}(x) = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}[(1-x)^{\alpha+n}(1+x)^{\beta+n}], \qquad \alpha,\beta > -1. \tag{2.43}$$

Jacobi polynomials can also be computed using the following recursion formula:

$$\begin{aligned}
&P_0^{\alpha,\beta}(x) = 1,\\
&P_1^{\alpha,\beta}(x) = \tfrac{1}{2}(\alpha-\beta+(\alpha+\beta+2)x),\\
&a_n^1 P_{n+1}^{\alpha,\beta}(x) = (a_n^2 + a_n^3 x)P_n^{\alpha,\beta}(x) - a_n^4 P_{n-1}^{\alpha,\beta}(x),\\
&where\\
&\qquad a_n^1 = 2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta),\\
&\qquad a_n^2 = (2n+\alpha+\beta+1)(\alpha^2-\beta^2),\\
&\qquad a_n^3 = (2n+\alpha+\beta)(2n+\alpha+\beta+1)(2n+\alpha+\beta+2),\\
&\qquad a_n^4 = 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2).
\end{aligned} \tag{2.44}$$

For instance, the $n^{th}-$degree Legendre polynomial is defined in the special case when $\alpha = \beta = 0$, *i.e.*, $L_n(x) = P_n^{0,0}(x)$. In our analysis we also need the right Radau polynomials

$$R_{p+1}(x) = (1-x)P_p^{1,0}(x) = C(L_{p+1}(x) - L_p(x)). \tag{2.45}$$

Let us denote $L_p(\xi)$, $P_p^{\alpha,\beta}(\xi)$ and $R_p(\xi)$ the shifted Jacobi, Legendre and Radau polynomials, respectively, on $[0,1]$. Table 2.1 contains the first seven Legendre polynomials and their roots. We plot $L_n(\xi),\ n = 0, ..., 7$ in Figure 2.5.

Table 2.1: Legendre polynomial on $[0,1]$ and their roots up to degree 6.

n	$L_n(\xi)$	Roots of $L_n(\xi)$
0	1	-
1	-1+2ξ	0.5
2	$1-6\xi+6\xi^2$	0.2113248654051871, 0.7886751345948129
3	$-1+12\xi-30\xi^2+20\xi^3$	0.1127016653792583, 0.5, 0.8872983346207415
4	$1-20\xi+90\xi^2-140\xi^3+70\xi^4$	0.06943184420297371, 0.3300094782075722, 0.6699905217924266, 0.9305681557970275
5	$-1+30\xi-210\xi^2+560\xi^3-630\xi^4+252\xi^5$	0.046910077030668, 0.2307653449471585,0.5, 0.7692346550528415, 0.9530899229693265
6	$1-42\xi+420\xi^2-1680\xi^3+3150\xi^4-$ $2772\xi^5+924\xi^6$	0.03376524289842398, 0.1693953067668677, 0.3806904069584021, 0.6193095930415969, 0.830604693233132, 0.9662347571015774

We plot the first eight right Radau polynomials in Figure 2.6. We also present the right Radau polynomials and their roots up to degree 6, in Table 2.2. The values for these zeros were found using *Mathematica NSolve* function with maximum precision.

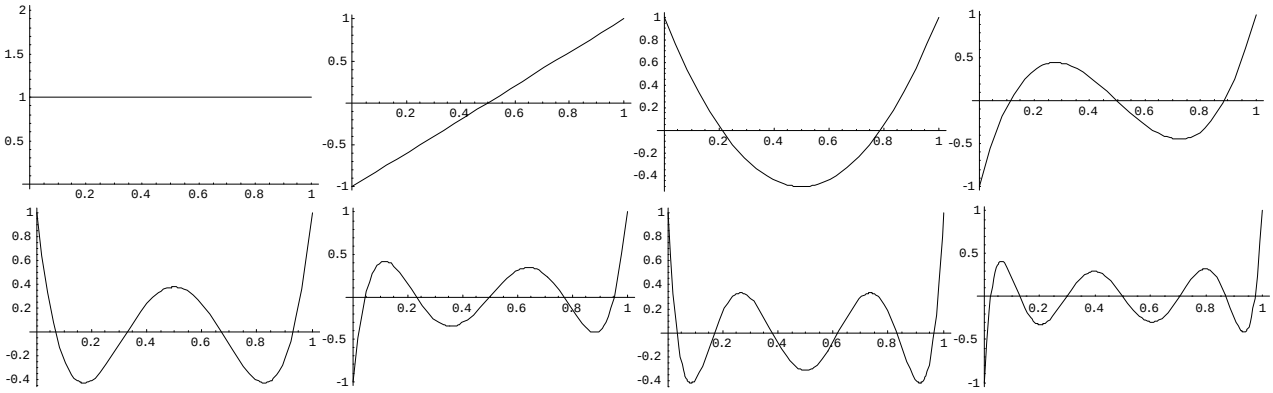

Figure 2.5: Legendre polynomials from degree 0 to 7 (upper left to lower right).

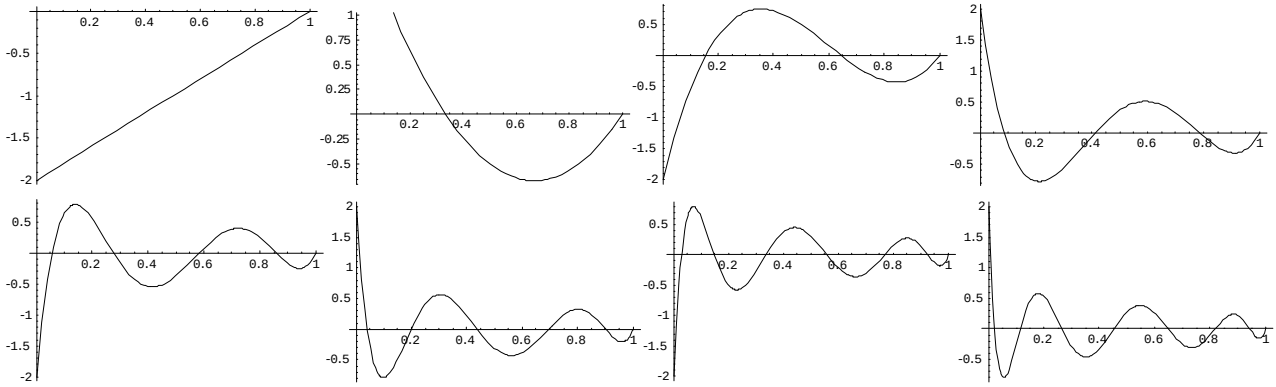

Figure 2.6: Right Radau polynomials from degree 1 to 8 (upper left to lower right).

Table 2.2: Right Radau polynomials on $[0, 1]$ and their roots up to degree 6.

n	$R_n^+(\xi)$	Roots of $R_n^+(\xi)$
1	2(-1+ξ)	1
2	$2 - 8\xi + 6\xi^2$	0.33333333333333337, 1
3	$-2 + 18\xi - 36\xi^2 + 20\xi^3$	0.15505102572168217, 0.6449489742783179, 1
4	$2(1 - 16\xi + 60\xi^2 - 80\xi^3 + 35\xi^4)$	0.08858795951270396, 0.4094668644407346, 0.7876594617608479, 1
5	$-2 + 50\xi - 300\xi^2 + 700\xi^3 - 700\xi^4 + 252\xi^5$	0.0571041961145177, 0.27684301363812375, 0.5835904323689155, 0.8602401356562251, 1
6	$2 - 72\xi + 630\xi^2 - 2240\xi^3 + 3780\xi^4 - 3024\xi^5 + 924\xi^6$	0.03980985705146874, 0.19801341787360793, 0.4379748102473867, 0.6954642733536484, 0.9014649142011256, 1

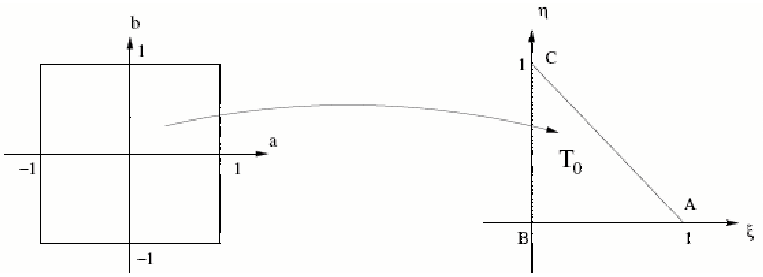

Figure 2.7: Illustration of the mapping between a reference square and a reference triangle.

To construct an orthogonal polynomial basis on the standard reference triangle $T_0 = \{(\xi,\eta), 0 \leq \xi,\eta \leq 1, 0 \leq \xi+\eta \leq 1\}$ we follow [63] and consider the mapping in Figure 2.7 between the reference square S_0 and the reference triangle T_0.

$$(\xi,\eta) = (\frac{(1+a)(1-b)}{4}, \frac{1+b}{2}) \text{ or } (a,b) = (\frac{2\xi}{1-\eta} - 1, 2\eta - 1). \tag{2.46}$$

The mapping (2.46) basically collapse the top edge $b = 1$ of S_0 into the top vertex $(0,1)$ of T_0. The Jacobians of the mapping and its inverse are $J(\xi,\eta) = \frac{\partial(a,b)}{\partial(\xi,\eta)} = \frac{4}{1-\eta}$ and $J^{-1}(\xi,\eta) = \frac{\partial(\xi,\eta)}{\partial(a,b)} = \frac{1-\eta}{4}$. The Dubiner polynomial basis on the canonical element (defined by the vertices $(0,0)$, $(1,0)$ and $(0,1)$) is then defined as

$$\varphi_k^l(\xi,\eta) = L_k(a)(1-b)^k P_l^{2k+1,0}(b) = 2^k L_k(\frac{2\xi}{1-\eta} - 1)(1-\eta)^k P_l^{2k+1,0}(2\eta-1), \; k,l \geq 0, \tag{2.47}$$

where $P_n^{\alpha,\beta}(x)$, $-1 \leq x \leq 1$, is the Jacobi polynomial and $L_n(x) = P_n^{0,0}(x) \in \mathbf{P}_n$, $-1 \leq x \leq 1$, is the Legendre polynomial. We note that Dubiner polynomials satisfy the $\mathcal{L}^2$ orthogonality

$$\int_0^1 \int_0^{1-\xi} \varphi_k^l \varphi_p^q d\eta d\xi = c_{kp}^{lq} \delta_{kp} \delta_{lq}. \tag{2.48}$$

Thus, $\{\varphi_k^l(\xi,\eta), 0 \leq k,l,k+l \leq p\}$ forms an orthogonal basis for the polynomial space $\mathcal{P}_p$. The first fifteen un-normalized Dubiner basis functions for $p = 0$ to 4 are given in Table 2.3 and are displayed in Figure 2.8 for $p = 0$ to 4.

Note that the Dubiner basis functions (2.47) are polynomials in both (ξ,η) and (a,b) spaces. Moreover, it can be observed that in (a,b) space, a Dubiner basis function can be expressed as the product of two polynomials, one in a and the other in b. This property is referred to a warped product by Dubiner to differentiate it from the standard tensor product with quadrilateral domains. For this property, the Dubiner basis can be relatively efficiently manipulated. For example, integrals involving the inner product of $\varphi_k^l(\xi,\eta)$ with an arbitrary function $f(\xi,\eta)$ can be efficiently evaluated using the sum factorization technique. Especially,

Table 2.3: First fifteen un-normalized Dubiner polynomials φ_i^j of degree $0 \leq i+j \leq 4$.

$p=0$	$\varphi_0^0(\xi,\eta) = 1,$
$p=1$	$\varphi_1^0(\xi,\eta) = 4\xi + 2\eta - 2,$ $\varphi_0^1(\xi,\eta) = 3\eta - 1,$
$p=2$	$\varphi_2^0(\xi,\eta) = 24\xi^2 + 24\xi\eta + 4\eta^2 - 24\xi - 8\eta + 4,$ $\varphi_1^1(\xi,\eta) = 20\xi\eta + 10\eta^2 - 4\xi - 12\eta + 2,$ $\varphi_0^2(\xi,\eta) = 10\eta^2 - 8\eta + 1,$
$p=3$	$\varphi_3^0(\xi,\eta) = -8 + 96\xi - 240\xi^2 + 160\xi^3 + 24\eta - 192\xi\eta + 240\xi^2\eta - 24\eta^2 + 96\xi\eta^2 + 8\eta^3,$ $\varphi_2^1(\xi,\eta) = -4 + 24\xi - 24\xi^2 + 36\eta - 192\xi\eta + 168\xi^2\eta - 60\eta^2 + 168\xi\eta^2 + 28\eta^3,$ $\varphi_1^2(\xi,\eta) = -2 + 4\xi + 26\eta - 48\xi\eta - 66\eta^2 + 84\xi\eta^2 + 42\eta^3,$ $\varphi_0^3(\xi,\eta) = -1 + 15\eta - 45\eta^2 + 35\eta^3,$
$p=4$	$\varphi_4^0(\xi,\eta) = 16(70\xi^4 + 140\xi^3(\eta-1) + 90\xi^2(\eta-1)^2 + 20\xi(\eta-1)^3 + (\eta-1)^4),$ $\varphi_3^1(\xi,\eta) = 8(10\xi^2 + 10\xi(\eta-1) + (\eta-1)^2)(-1 + 2\xi + \eta)(9\eta - 1),$ $\varphi_2^2(\xi,\eta) = 4(6\xi^2 + 6\xi(\eta-1) + (\eta-1)^2)(1 - 16\xi + 36\eta^2),$ $\varphi_1^3(\xi,\eta) = 2 - 4\xi - 44\eta + 84\xi\eta + 210\eta^2 - 336\xi\eta^2 - 336\eta^3 + 336\xi\eta^3 + 168\eta^4,$ $\varphi_0^4(\xi,\eta) = 1 - 24\eta + 126\eta^2 - 224\eta^3 + 126\eta^4.$

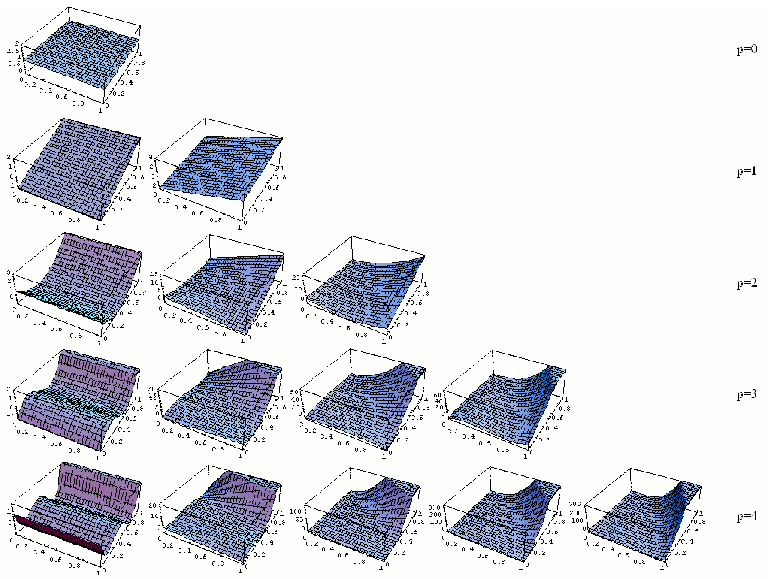

Figure 2.8: Dubiner polynomials for degrees 0 to 4 (upper left to lower right).

we have

$$\iint_{T_0} \varphi_k^l(\xi,\eta) f(\xi,\eta) d\xi d\eta = \frac{1}{8}\int_{-1}^{1}\int_{-1}^{1} L_k(a)(1-b)^k P_l^{2k+1,0}(b) f(a,b)(1-b) da db$$

$$= \frac{1}{8}\int_{-1}^{1} L_k(a)\Big(\int_{-1}^{1}(1-b)^k P_l^{2k+1,0}(b) f(a,b)(1-b) db\Big) da. \tag{2.49}$$

To approximate any integrals, we choose quadrature points in the reference triangle T_0 as

$$(\xi_i,\eta_j) = \Big(\frac{(1+\tau_i)(1-\tau_j)}{4}, \frac{(1+\tau_j)}{2}\Big) \in T_0, \quad 0 \le i,j \le p, \tag{2.50}$$

where τ_i (and ω_i), $i = 0, ..., p$ are the Gauss points (and weights) in the interval $[-1, 1]$. Then any integral over the triangle T_0 can be approximated by the following quadrature

$$\iint_{T_0} f(\xi,\eta) d\xi d\eta \approx \frac{1}{8}\int_{-1}^{1}\int_{-1}^{1} f(a,b)(1-b) da db = \frac{1}{4}\sum_{j=0}^{p}(\omega_j(1-\eta_j)\sum_{i=0}^{p}\omega_i f(\xi_i,\eta_j)). \tag{2.51}$$

It is shown that the extra terms x^{p+1}, y^{p+1} in the finite element space $\mathcal{U}_p$ leads to ill-conditioned problems. These terms in our implementation are replaced by basis functions obtained by orthonormalisation of the polynomial functions x^{p+1}, y^{p+1} on the standard triangle by means of the Gram-Schmidt orthonormalisation procedure. The Gram-Schmidt algorithm is a general procedure that transforms a given set of n independent vectors v_i, $i = 1, ..., n$, into a set of n orthonormal vectors e_i, $i = 1, ..., n$, spanning the same vector space. The procedure is outlined in the following algorithm,

$$\begin{aligned}
&(i) \quad e_1 \leftarrow v_1 / \parallel v_1 \parallel;\\
&(ii) \quad For \quad i = 2, ..., n,\\
&\qquad (a) v_i \leftarrow v_i - \sum_{j=1}^{i-1} < v_i, e_j > e_j;\\
&\qquad (b) e_i \leftarrow v_i / \parallel v_i \parallel .
\end{aligned}$$

In step $(ii)a$ of this algorithm, the i^{th} input vector is replaced by its orthogonal complement with respect to the vector space spanned by the first $(i-1)$ vectors. The orthogonal complement is the difference between a vector and its projection on a sub-space. The vectors e_j, $j < i$ form an orthonormal set, so that the dual vectors in the scalar products for obtaining the projection coefficients, are equal to the primary vectors e_j themselves.

The first six normalized Gram-Schmidt basis functions for $p = 1, 2, 3$ on the canonical triangle T_0 are given in Table 2.4

Table 2.4: First six normalized Gram-Schmidt basis functions

$p=1$	$\varphi_2^0(\xi,\eta)=\sqrt{6}(1-8\xi+10\xi^2)$, $\varphi_1^1(\xi,\eta)=\sqrt{3}(1+4\xi-5\xi^2-12\eta+15\eta^2)$,
$p=2$	$\varphi_3^0(\xi,\eta)=2\sqrt{2}(-1+15\xi-45\xi^2+35\xi^3)$, $\varphi_2^1(\xi,\eta)=2\sqrt{10/3}(-1+3\xi-9\xi^2+7\xi^3+12\eta-36\eta^2+28\eta^3)$,
$p=3$	$\varphi_4^0(\xi,\eta)=\sqrt{10}(1-24\xi+126\xi^2-224\xi^3+126\xi^4)$, $\varphi_3^1(\xi,\eta)=\sqrt{5/3}(2+12\xi-63\xi^2+112\xi^3-63\xi^4-60\eta+315\eta^2-560\eta^3+315\eta^4)$.

2.7 Numerical Integration Rules

We need to compute oriented line integrals arising from the use of Green's theorem and double integrals over general triangular domains. For this, we present a summary of the Gaussian numerical integration rules on the one and two-dimensional canonical elements that we use in our program for the DG method [37].

Approximate numerical integration is performed by evaluating the integrand at a number of well-chosen nodes. The weights ω_i and the node coordinates are carefully chosen so that an as wide as possible class of polynomial functions will be integrated exactly with as few nodes as possible.

The quadrature rules of the Gauss type are based on the summation of weighted function values at non-equidistant integration points. The $n-$point Gauss quadrature rule on the one-dimensional canonical interval $(-1,1)$ reads

$$\int_{-1}^{1} f(\xi)d\xi \approx \sum_{i=1}^{n} \omega_i f(\xi_i). \tag{2.52}$$

The $n-$point Gauss quadrature rules are exact for all polynomials of degree $2n-1$ and lower. The maximum degree of a polynomial that is integrated exactly by an integration rule, is called the degree of precision of the rule. It can be shown that the integration points are roots of Legendre polynomial of degree p.

The integration points and weights for a few selected $n-$point rules are given in [37].

The Gauss quadrature rules in two-dimensions on the canonical triangle $\{(\xi_1,\xi_2), -1\leq\xi_1\leq 1, -1\leq\xi_2\leq-\xi_1\}$ has the form

$$\int_{-1}^{1}\int_{-1}^{-\xi_1} f(\xi_1,\xi_2)d\xi_1 d\xi_2 \approx \sum_{k=1}^{n} \omega_k f(\xi_{1,k},\xi_{2,k}), \tag{2.53}$$

where n denotes the number of integration points. Tabulated Gaussian integration points and weights on the reference triangle T_0 are given in [37].