CHAPTER 2

Numerical approaches to solving PDE's

1. A general framework for numerical solution of PDE's

Now that we have an idea of the differential equations of interest in applications, let us turn to the problem of finding solutions. Analytical techniques such as separation of variables or Fourier analysis are very useful but for a limited class of problems, generally linear PDE's on domains of simple geometry. Most practical interest arises from non-linear PDE's over domains of complicated shape. The fundamental problem facing us is to determine a function $\tilde{q}$ that approximates the solution of the PDE of interest. A large numer of methods have been devised to solve this problem numerically and we shall study individual methods extensively in later chapters. It is instructive to see that basically all methods can be expressed in a general framework that allows comparison of the strengths and weaknesses of individual methods. Say we are faced with the following general problem:

PROBLEM 1. *Find $q : \Omega \to R^n$, $q \in \mathcal{F}$ that satisfies $Lq = 0$ on Ω and $Bq = 0$ on $\partial\Omega$ where L, B are operators defined on the normed linear space $\mathcal{F}$.*

Let us see how this abstract statement maps onto one of the typical PDE problems, the Neumann problem:

$$\begin{cases} \dfrac{\partial^2 q}{\partial x^2} + \dfrac{\partial^2 q}{\partial y^2} = \sigma(x, y, q),\ (x, y) \in \Omega \\ \dfrac{\partial q}{\partial n}(x, y) = F(x, y),\ (x, y) \in \partial\Omega \end{cases} . \tag{1.1}$$

We have

$$L = \frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y} - \sigma(x, y, \cdot) \tag{1.2}$$

$$B = \frac{\partial}{\partial n} - F(x, y) \tag{1.3}$$

with $L : \Omega,\ B : \partial\Omega$ and $\mathcal{F}$ some reasonable space of functions, for instance the space of functions continuous up to second order defined on Ω, $\mathcal{C}^{(2)}(\Omega)$.

The advantage of the abstract formulation is that it allows us to concentrate on the typical steps followed when building a numerical procedure without spending too much time on the individual application at hand. We assume that the exact solution q is difficult to find and therefore concentrate on constructing an approximation $\tilde{q} \cong q$. Applying the operators L, B to the approximation $\tilde{q}$ leads to an error, called a *residual*

$$L\tilde{q} = r,\ B\tilde{q} = s\ . \tag{1.4}$$

It is our objective that the residual be as small as possible. In order to quantify the error made, we need a mapping from the function space to which r belongs to

a real number. Such a mapping is called a *functional* and the typical example is the norm functional, e.g. $\|r\|$ and $\|s\|$. In numerical approximations we often work in the space of p-integrable functions $\mathcal{F} = \mathcal{L}^{(p)}(\Omega)$ for which the norm is defined by

$$\|f\| = \left(\int |f(x)|^p \, dx \right)^{1/p} . \tag{1.5}$$

The norm is a good candidate for evaluating the quality of our approximation $\tilde{q}$ since if $\|r\| = 0$ then we know that $r = 0$ and therefore $\tilde{q} = q$. Unfortunately, the presence of the absolute value operation limits the operations that can be carried out on the norm and it is typical to use other functionals in constructing numerical procedures. Instead of using the norm consider the functionals

$$I(r) = \int_\Omega r \; w \; dx, \; J(s) = \int_{\partial\Omega} s \; w' \; dx \; . \tag{1.6}$$

Here we have introduced weight functions w, w' that assign differing importance to errors made in various parts of the domain of integration. In adopting $I(r), \; J(s)$ as measures of the quality of our approximation $\tilde{q}$ we have abandoned the certainty of knowing that $\tilde{q} = q$ when $I(r) = 0$, since a zero value for $I(r)$ could be obtained by cancellation of positive and negative errors thorughout the domain. Appropriately chosen weight functions alleviate this concern somewhat and this is a tradeoff we accept for now with a view to simplicity of the ensuing algorithm. It is possible to eliminate this drawback using the square of the residual as we shall see later on.

To simplify the presentation let us concentrate on $I(r)$ only. The addition of boundary conditions is usually a straightforward matter. Now that we have established a means by which to quantify the quality of an approximation we have to decide on how to build the approximation itself. Since $\mathcal{F}$ is a normed linear space we can express any element g of $\mathcal{F}$ as a linear combination over a set of basis functions

$$g = \sum_{i=1}^{\infty} a_i l_i \; . \tag{1.7}$$

For example the set $\{1, \sin x, \cos x, \sin 2x, \cos 2x, \ldots\}$ is a basis for the square-integrable functions defined on $[0, 2\pi]$, $\mathcal{L}^{(2)}([0, 2\pi])$, and the set $\{1, x, x^2, \ldots\}$ is a basis for the infinitely differentiable functions defined over the reals $C^{(\infty)}(\mathbb{R})$. In practical computations we cannot use an infinite sum, so we construct our approximation $\tilde{q}$ using only N terms

$$\tilde{q} = \sum_{i=1}^{N} c_i l_i \; . \tag{1.8}$$

Using (1.8) $I(r)$ becomes

$$I(r) = \int_\Omega L \left(\sum_{i=1}^{N} c_i l_i \right) \; w \; dx \; . \tag{1.9}$$

For arbitrary coefficients c_i we shall have $I(r) \neq 0$. Since we wish $\tilde{q}$ to be a good approximation we can reasonably impose $I(r) = 0$ and thus obtain an equation to

be used in determining the coefficients c_i

$$\int_\Omega L\left(\sum_{i=1}^{N} c_i l_i\right)\ w\ dx = 0\ . \tag{1.10}$$

Now this is just one equation but we have N unknown coefficients. However the weight function is arbitrary so obtaining as many equations as we need is easy: we just choose w from some set of functions $w \in \{w_1, w_2, \ldots, w_N\}$ and obtain a system of N equations for N unknowns

$$\int_\Omega L\left(\sum_{i=1}^{N} c_i l_i\right)\ w_j\ dx = 0,\ j = 1, 2, \ldots, N\ . \tag{1.11}$$

Up to this point we have concentrated on the approximation of q itself. There still remains the question of how to approximate the presumably complex shape of Ω. Generally this is done by approximating Ω by a set of simple-shaped subdomains ω_k such that the measure ρ of the difference between the two sets goes to zero as we increase the number of subdomains

$$\rho\left(\Omega - \bigcup_{m=1}^{M} \omega_m\right) \to 0\ . \tag{1.12}$$

Using this approximation of the domain Ω (1.11) becomes

$$\sum_{m=1}^{M} \int_{\omega_m} L\left(\sum_{i=1}^{N} c_i l_i\right)\ w_j\ dx = 0,\ j = 1, 2, \ldots, N\ . \tag{1.13}$$

This is known as a *weighted residual formulation* and a large number of methods for numerically solving PDE's can be thus expressed. We shall turn to some examples shortly, but let us summarize the basic aspects:

(1) A function space from which we construct approximations is chosen along with a subset of a basis of this space $\{l_1, l_2, \ldots, l_N\}$;
(2) A set of weight functions $\{w_1, w_2, \ldots, w_N\}$ is chosen;
(3) A discretization of the domain $\{\omega_1, \omega_2, \ldots, \omega_M\}$ is chosen.

2. Basic numerical methods

2.1. Finite difference methods.

2.1.1. *Finite difference derivation.* In a finite difference method (FDM) the derivatives appearing in an ODE or PDE are approximated using finite differences. For example the IVP

$$\begin{cases} q' = f(t, q) \\ q(t = 0) = q_0 \end{cases} \tag{2.1}$$

can be solved over the domain $[0, T]$ by finite differences using the following procedure. We construct an approximation $\tilde{q}$ by a set of point values $Q^n = \tilde{q}(t^n)$ with $t^n = nk$, $n = 0, 1, \ldots, N$ and k a step size $k = T/N$ and assume that $\tilde{q}$ varies linearly between the point values. From the point values we can construct myriad approximations of the value of the derivative of $\tilde{q}$, for example

$$\tilde{q}'(t^n) \cong \frac{Q^{n+1} - Q^{n-1}}{2k},$$

FIGURE 1. The piecewise linear form functions.

and obtain practical algorithms to solve (2.1) such as

$$Q^{n+1} = Q^{n-1} + 2k\ f(t^n, Q^n)\ , \tag{2.2}$$

known as the *midpoint rule*.

2.1.2. *Weighted residual derivation.* We shall analyze (2.1) and similar algorithms extensively later on, but let us now see how the same method can be obtained via the weighted residual formulation and what insights we can thereby gain. We have chosen $\tilde{q}$ as being a piecewise linear approximation defined at the points $t^n = nk$. In the general language of the weighted residual formulation we have $\Omega = [0, T]$, $\omega_m = [t^{m-1}, t^m]$. A basis for the piecewise linear functions defined on this partition of Ω is given by

$$l_n(t) = \begin{cases} 0 & t < t^{n-1} \\ \frac{1}{k}\left(t - t^{n-1}\right) & t^{n-1} \le t < t^n \\ \frac{1}{k}\left(t^{n+1} - t\right) & t^n \le t < t^{n+1} \\ 0 & t^{n+1} \le t \end{cases} . \tag{2.3}$$

See Fig. (1).

Any piecewise linear function over $\{\omega_m\}$ can be defined as a linear combination of $\{l_n\}$, for instance

$$\tilde{q}(t) = \sum_{i=0}^{N} c_i l_i(t)\ . \tag{2.4}$$

It is apparent from (2.3) that

$$l_m(t^n) = \delta_{mn} = \begin{cases} 1 & \text{if } m = n \\ 0 & \text{if } m \neq n \end{cases} , \tag{2.5}$$

so imposing the conditions $\tilde{q}(t^n) = Q^n$ leads to

$$\tilde{q}(t^n) = \sum_{i=0}^{N} c_i l_i(t^n) = \sum_{i=0}^{N} c_i \delta_{in} = c_n = Q^n\ , \tag{2.6}$$

i.e. the coefficients of the expansion (2.4) are the nodal values Q^n. A similar expansion is made to approximate the values of the r.h.s. term

$$f(t, q) \cong \tilde{f}(t, q) = \sum_{i=0}^{N} F^i l_i(t) \tag{2.7}$$

with $F^n = \tilde{f}(t^n, Q^n)$.

We must now choose appropriate weight functions. Since the approximation we are building depends only on nodal values, a reasonable choice would be a set of weight functions that give importance to the residual obtained at the nodes. Such a set is given by

$$w_j = \delta(t - t^j) \tag{2.8}$$

where $\delta(t-t^j)$ is the Dirac delta functional centered on t^j defined by its integral property

$$\int_0^T g(t)\delta(t-t^j)\ dt = g(t^j) \tag{2.9}$$

for any function $g(t)$.

Having chosen the function space on which to base our aproximation $\tilde{q}$, a basis in this space $\{l_m(t)\}$, the weight functions $\{\delta(t-t^j)\}$ and a discretization of the domain $[0,T]$ we can work through the weighted residual formulation (1.13) to obtain

$$\sum_{m=1}^{M}\int_{t^{m-1}}^{t^m}\left[\frac{d}{dt}\left(\sum_{n=1}^{N}Q^n l_n(t)\right) - \left(\sum_{n=1}^{N}F^n l_n(t)\right)\right]\delta(t-t^j)\ dt = 0\ . \tag{2.10}$$

It is easiest to use the properties of the Dirac-δ function to work through the above expression. We have

$$\int_0^T \frac{d}{dt}\left(\sum_{n=1}^{N}Q^n l_n(t)\right)\ \delta(t-t^j)\ dt = \int_0^T\left(\sum_{n=1}^{N}F^n l_n(t)\right)\ \delta(t-t^j)\ dt \tag{2.11}$$

$$\sum_{n=1}^{N}Q^n\int_0^T l_n'(t)\ \delta(t-t^j)\ dt = \sum_{n=1}^{N}F^n\int_0^T\ l_n(t)\ \delta(t-t^j)\ dt \tag{2.12}$$

$$\sum_{n=1}^{N}Q^n\ l_n'(t_j) = \sum_{n=1}^{N}F^n l_n(t^j) = \sum_{n=1}^{N}F^n\delta_{nj} = F^j \tag{2.13}$$

We should be careful in evaluating $l_n'(t_j)$ since $l_n(t)$ is not differentiable at the nodal points $t=t^n$. If we interpret $l_n'(t_j)$ in principal value as the average of the limits to the left and the right we have

$$l_n'(t_j) = \begin{cases} -\frac{1}{2k} & n = j-1 \\ 0 & n \neq j\pm 1 \\ \frac{1}{2k} & n = j+1 \end{cases}\ . \tag{2.14}$$

Eq. (2.13) becomes

$$\frac{Q^{j+1}-Q^{j-1}}{2k} = F^j \tag{2.15}$$

which is exactly the same expression we had obtained previously, Eq. (2.2).

2.1.3. *Comparison of the two derivations.* It is satisfying to see a method derived in two ways, but an immediate question is what is to be gained by more complicated weighted residual procedure in comparison to the straightforward finite difference derivation. Generally the benefit arises in theoretical considerations of the behavior of the method. For instance let us consider the following theorem from approximation theory.

THEOREM 1. *Let $\mathcal{V}$ be a metric linear space with a metric induced by the scalar product $(\cdot,\cdot)$ on $\mathcal{V}$ and let $\mathcal{S}$ be a subspace of $\mathcal{V}$. Let $v \in \mathcal{V}$ and $u \in \mathcal{S}$. If $v-u$ is orthognal to any $w \in \mathcal{S}$, then u is the best approximation of v within $\mathcal{S}$.*

PROOF. Let $d(u,v)$ be the distance induced by the scalar product between u and v. Ask whether any other $w \in \mathcal{S}$ gives a smaller distance to v

$$[d(v,w)]^2 = (v-w, v-w) = (v-u+u-w, v-u+u-w) \tag{2.16}$$

$$= (v-u, v-u) + 2(u-w, v-u) + (u-w, u-w) \tag{2.17}$$

$$= \|v-u\|^2 + \|u-w\|^2 + 2(u-w, v-u)\ . \tag{2.18}$$

Since $v-u$ is orthogonal to any element in $\mathcal{S}$, it is orthogonal to $u-w$ so $(u-w, v-u) = 0$ and

$$[d(v,w)]^2 = \|v-u\|^2 + \|u-w\|^2 \geq \|v-u\|^2 = [d(v,u)]^2 \tag{2.19}$$

and we conclude that u is the best approximation of v within $\mathcal{S}$. □

We can apply such theorems to weighted residual derivations to infer the behavior of the numerical approximation. Applied to the above example, $\mathcal{V}$ would be the space of differentiable functions to which q belongs. $\mathcal{S}$ would be the subspace of piecewise continuous functions where we defined our approximation $\tilde{q}$. The best approximation we could obtain would be orthogonal to the complement $\mathcal{S}$ of within $\mathcal{V}$. This subspace would contain functions that are not expressible as an expansion along the set (2.3), for instance functions that vary more rapidly than the time step chosen k. Predictions such as this are typically more difficult to obtain from the simpler finite difference derivation.

Let us consider an application of the above theorem. Let $\mathcal{V}$ be the space of differentiable functions defined on the interval $[0,T]$. We introduce the scalar product

$$(u,v) = \int_0^T u(t)v(t)\ dt\ ,$$

with $u, v \in \mathcal{V}$. The metric induced by the scalar product is

$$d(u,v) = (u-v, u-v)^{1/2} = \left(\int_0^T [u(t)-v(t)]^2\ dt \right)^{1/2}\ .$$

Now let us consider the problem of approximating elements of $\mathcal{V}$ by piecewise linear functions defined by their point values on a partition of the interval $[0,T]$. Let $\{0 = t^0, t^1, \ldots, t^{N-1}, t^N = T\}$ be the partition of this interval. To keep things simple we'll use an uniform partition with $t^n = t^0 + nk$, $k = T/N$. The subspace of piecewise linear functions is

$$\mathcal{S} = \left\{ (Q^n)\ \mid n = 0, 1, \ldots, N,\ \tilde{q}(t) = \frac{\left(Q^n - Q^{n-1}\right)}{k}(t - t^{n-1}) + Q^{n-1} \text{ for } t^{n-1} \leq t \leq t^n \right\}\ .$$

A piecewise linear function is obviously continuous but is not differentiable at the partition points t^n so at present we cannot affirm that $\mathcal{S}$ is a subspace of $\mathcal{V}$. This drawback is easily eliminated by imposing a principal value definition of the derivative at the partition points, i.e. for $\tilde{q} \in \mathcal{S}$ we define

$$\tilde{q}'(t^n) = \frac{1}{2}\left[\lim_{\substack{t\to t^n \\ t<t^n}} \tilde{q}'(t) + \lim_{\substack{t\to t^n \\ t>t^n}} \tilde{q}'(t)\right] = \frac{Q^{n+1} - Q^{n-1}}{2k} \ .$$

With this definition we can now state that $\mathcal{S}$ is a subspace of $\mathcal{V}$. Let's determine the scalar product in $\mathcal{S}$ induced by that defined for $\mathcal{V}$. With $\tilde{q}, \tilde{r} \in \mathcal{S}$ we have

$$(\tilde{q}, \tilde{r}) = \int_a^b \tilde{q}(t)\tilde{r}(t)\ dt = \sum_{n=1}^{N} \int_{t^{n-1}}^{t^n} \left[\frac{\left(Q^n - Q^{n-1}\right)}{k}(t - t^{n-1}) + Q^{n-1}\right]\left[\frac{\left(R^n - R^{n-1}\right)}{k}(t - t^{n-1}) + R^{n-1}\right] dt$$
$$= \frac{k}{6}\left(2Q^nR^n + 2Q^{n-1}R^{n-1} + Q^nR^{n-1} + Q^{n-1}R^n\right) \ .$$

Let us now look for the best approximation of an element $q \in \mathcal{V}$ by an element $\tilde{q} \in \mathcal{S}$. By the above theorem, the best approximation possible satisfies the condition

$$(q - \tilde{q}, \tilde{r}) = 0$$

for all $\tilde{r} \in \mathcal{S}$. This leads to

$$\int_0^T [q(t) - \tilde{q}(t)]\ \tilde{r}(t)\ dt = 0$$

or

$$\sum_{n=1}^{N} \int_{t^{n-1}}^{t^n} \left[q(t) - \frac{\left(Q^n - Q^{n-1}\right)}{k}(t - t^{n-1}) - Q^{n-1}\right]\ \tilde{r}(t) = 0 \ .$$

The unknowns of the above problem are the nodal values $\{Q^n\}$ and we can obtain the relation

(2.20)

$$\sum_{n=1}^{N}\left\{\frac{\left(Q^n - Q^{n-1}\right)}{k}\int_{t^{n-1}}^{t^n}(t - t^{n-1})\ \tilde{r}(t)\ dt + Q^{n-1}\int_{t^{n-1}}^{t^n}\tilde{r}(t)\ dt\right\} = \int_0^T q(t)\,\tilde{r}(t)\,dt \ .$$

This is a linear relation in the unknowns. If this relation is valid for *all* $\tilde{r} \in \mathcal{S}$ we have determined the best approximation of q by elements within $\mathcal{S}$. Note that $\mathcal{S}$ is of finite dimension $N+1$ so we need only choose $N+1$ functions $\tilde{r}_j \in \mathcal{S}$, $j = 0, 1, \ldots, N$ that form a basis of $\mathcal{S}$. If (2.20) is verified for the elements of the basis then it is verified for all elements within $\mathcal{S}$ (why?). We obtain the system of equations

$$\sum_{n=1}^{N}\left\{\frac{\left(Q^n - Q^{n-1}\right)}{k}\int_{t^{n-1}}^{t^n}(t - t^{n-1})\ \tilde{r}_j(t)\ dt + Q^{n-1}\int_{t^{n-1}}^{t^n}\tilde{r}_j(t)\ dt\right\} = \int_a^b q(t)\,\tilde{r}_j(t)\,dt\ ,\ j = 0, 1, \ldots, N,$$

which is linear system with $N+1$ equations for the $N+1$ unknowns Q^n.

The same technique can be applied to determine best approximations to elements $v \in \mathcal{V}$ which are not specified directly. In the above initial value problem (2.1) we are not given q but rather its derivative $q'(t) = Lq(t) = f(t, q)$. We can impose a condition

$$(Lq - L\tilde{q}, \tilde{r}) = (f - L\tilde{q}, \tilde{r}) = 0$$

namely that the residual be orthogonal to the subspace $\mathcal{S}$. This leads to the system

$$\sum_{n=1}^{N} \left\{ \frac{\left(Q^n - Q^{n-1}\right)}{k} \int_{t^{n-1}}^{t^n} \tilde{r}_j(t)\ dt \right\} = \int_0^T f(t,q)\ \tilde{r}_j(t)\ dt\ ,\ j = 0, 1, \ldots, N,$$

again with $\{r_j\}$, $j = 0, 1, \ldots, N$ a basis of $\mathcal{S}$. Note that if we choose $r_j(t) = l_j(t)$, the linear form functions from (2.3), we obtain

$$\sum_{n=1}^{N} \left\{ \frac{\left(Q^n - Q^{n-1}\right)}{k} \int_{t^{n-1}}^{t^n} l_j(t)\ dt \right\} = \frac{Q^{j+1} - Q^{j-1}}{2} = \int_{t^{j-1}}^{t^{j+1}} f(t,q)\, l_j(t)\, dt\ ,\ j = 0, 1, \ldots, N\ .$$

If we additionally introduce a piecewise linear approximation for $f(t,q)$ by

$$f(t,q) = \sum_{n=0}^{N} F^n l_n(t)$$

with $F^n = f(t^n, Q^n)$ we obtain

$$\frac{Q^{j+1} - Q^{j-1}}{2} = \sum_{n=0}^{N} F^n \int_{t^{j-1}}^{t^{j+1}} l_n(t)\ l_j(t)\ dt\ ,\ j = 0, 1, \ldots, N\ .$$

Only three of the integrals give non-zero values,

$$\int_{t^{j-1}}^{t^{j+1}} l_{j-1}(t)\ l_j(t)\ dt = \frac{k}{6}\ ,$$
$$\int_{t^{j-1}}^{t^{j+1}} l_j(t)\ l_j(t)\ dt = \frac{2k}{3}\ ,$$
$$\int_{t^{j-1}}^{t^{j+1}} l_{j+1}(t)\ l_j(t)\ dt = \frac{k}{6}\ ,$$

leading to the final formula

$$\frac{Q^{j+1} - Q^{j-1}}{2} = kF^j\ ,$$

the same as that obtained above, (2.2). We can now however state that the numerical solution obtained by this formula is the *best* possible approximation of the initial value problem (2.1) by piecewise linear functions when the right-hand-side term $f(t,q)$ is also approximated by piecewise linear functions. This statement could not have been made from the simple finite difference derivation.

2.2. Finite volume methods.

2.2.1. *Derivation by integrating over a finite volume.* Finite volume methods for PDE's are derived by integrating the equation of interest over a finite region of the solution domain and introducing an average value for the unknown function over each finite region. Let us exemplify for an elementary PDE, the one-dimensional, constant-velocity advection equation

$$q_t + uq_x = 0\ .$$

Say we wish to determine the solution for $(x,t) \in [0,1] \times [0,T]$ starting from the initial condition $q(x, t=0) = \sin(\pi x)$. In a finite volume method we first introduce a partition of the definition domain into finite volumes $V_i^n = [x_{i-1}, x_i] \times [t^{n-1}, t^n]$

with $x_i = ih$, $h = 1/M$, $t^n = nk$, $k = T/N$. We then integrate the PDE over a finite volume

$$\int_{t^{n-1}}^{t^n} \int_{x_{i-1}}^{x_i} (q_t + uq_x)\ dx\ dt = 0\ ,$$

and obtain

$$\int_{x_{i-1}}^{x_i} \left[q(x,t^n) - q(x,t^{n-1})\right]\ dx + \int_{t^{n-1}}^{t^n} \left[uq(x_i,t) - uq(x_{i-1},t)\right] dt = 0\ .$$

An average value for the field variable q over the interval $[x_{i-1}, x_i]$ is now introduced at each time

$$Q_i^n = \frac{1}{h} \int_{x_{i-1}}^{x_i} q(x,t^n)\ dx\ .$$

Similarily an average value for the flux $f(q) = uq$ is introduced over the time interval $[t^{n-1}, t^n]$ for each x_i

$$F_i^{n-1} = \frac{1}{k} \int_{t^{n-1}}^{t^n} uq(x_i,t)\ dt\ .$$

Such averages are known to exist by virtue of the mean value theorem. We obtain an update formula

$$Q_i^n = Q_i^{n-1} - \frac{k}{h}\left(F_i^{n-1} - F_{i-1}^{n-1}\right)\ .$$

At first glance this looks just like a backward in time, backward in space finite difference approximation. The interpretation is a bit different though; whereas in a finite difference approximation we would have taken Q_i^n to be the value at $x = x_i$, $t = t^n$ in a finite volume method it is understood as the average value at t^n over the cell $[x_{i-1}, x_i]$. This difference is slight though and indeed we shall see that most finite volume methods have equivalent finite difference formulations.

2.2.2. *Weighted residual derivation.* We can easily recognize the appropriate weights and approximation spaces that lead to a finite volume method. Since we're using average spatial values over a computational cell the appropriate approximation space is that of piecewise constant functions

$$c_i(x) = \begin{cases} 0 & x \le x_{i-1} \\ 1 & x_{i-1} < x \le x_i \\ 0 & x > x_i \end{cases}\ .$$

The time dependence of $\tilde{q}$ is not specified in the finite volume derivation. We can assume a linear variation given by (2.3) such that

$$\tilde{q}(x,t) = \sum_{i=1}^{M}\sum_{n=0}^{N} Q_i^n c_i(x) l_n(t)$$

is our approximation of $q(x,t)$. The weight functions are unity within a cell and zero outside

$$w_j^n(x,t) = \begin{array}{ll} 1 & x_{i-1} \le x \le x_i \text{ and } t^{n-1} \le t \le t^n \\ 0 & \text{otherwise} \end{array}$$

The weighted residual formulation (1.13) becomes

$$\int_0^T \int_0^1 L\left(\sum_{i=1}^{M}\sum_{n=0}^{N} Q_i^n c_i(x) l_n(t)\right) w_j^p(x,t)\ dx\ dt = 0$$

with $L = \partial_t + u\partial_x$. The properties of the weight functions leads to

$$\int_{t^{p-1}}^{t^p}\int_{x_{j-1}}^{x_j} L\left(\sum_{i=1}^{M}\sum_{n=0}^{N} Q_i^n c_i(x) l_n(t)\right)\, dx\, dt = 0\ .$$

We can carry out some of the integrations analytically to obtain

$$\int_{x_{j-1}}^{x_j}\sum_{i=1}^{M}\sum_{n=0}^{N} Q_i^n c_i(x) l_n(t^p)\, dx - \int_{x_{i-1}}^{x_i}\sum_{i=1}^{M}\sum_{n=0}^{N} Q_i^n c_i(x) l_n(t^{p-1})\, dx +$$
$$\int_{t^{p-1}}^{t^p}\sum_{i=1}^{M}\sum_{n=0}^{N} u Q_i^n c_i(x_j) l_n(t)\, dt - \int_{t^{p-1}}^{t^p}\sum_{i=1}^{M}\sum_{n=0}^{N} u Q_i^n c_i(x_{j-1}) l_n(t)\, dt = 0$$

and we can use the identities $c_i(x_j) = \delta_{ij}$, $l_n(t^p) = \delta_{np}$ to obtain

$$\int_{x_{j-1}}^{x_j}\sum_{i=1}^{M} Q_i^p c_i(x)\, dx - \int_{x_{i-1}}^{x_i}\sum_{i=1}^{M} Q_i^{p-1} c_i(x)\, dx +$$
$$\int_{t^{p-1}}^{t^p}\sum_{n=0}^{N} u Q_j^n l_n(t)\, dt - \int_{t^{p-1}}^{t^p}\sum_{n=0}^{N} u Q_{j-1}^n l_n(t)\, dt = 0\ .$$

Another application of the properties of the form functions easily leads to

$$hQ_j^p - hQ_j^{p-1} + kF_j^{p-1} - kF_{j-1}^{p-1} = 0 \tag{2.21}$$

the same formula as was obtained previously. Note that because we have assumed a certain variation in time of $\tilde{q}(x,t)$ we actually can establish a formula for the time-averaged fluxes

$$F_j^{p-1} = \frac{u}{k}\int_{t^{p-1}}^{t^p}\left[Q_j^{p-1} l_{p-1}(t) + Q_j^p l_p(t)\right]\, dt \tag{2.22}$$
$$= \frac{u}{2}\left(Q_j^{p-1} + Q_j^p\right)\ . \tag{2.23}$$

2.3. Finite element methods. Finite element methods have traditionally been developed directly from the weighted residual formulation. There various ways in which weights and approximating functions are chosen that shall be studied in some detail later on. As an initial example let us consider the popular Galerkin procedure in which the weight functions are chosen identical to the approximating form functions. A simple example is to build a finite element method to solve the Poisson equation with Dirichlet boundary conditions

$$q_{xx} + q_{yy} = 0 \text{ on } \Omega \tag{2.24}$$
$$q(\Sigma) = f \text{ on } \Sigma = \partial\Omega. \tag{2.25}$$

Let us assume a simple rectangular domain $\Omega = [0,1]\times[0,1]$ and introduce an uniform partition $[x_{i-1}, x_i]\times[y_{j-1}, y_j]$, $x_i = ih$, $y_j = jh$, $h = 1/M$. There are many variants of the finite element method. The basic common thread is that a local approximation valid over a single finite element is used. Consider the so-called

form functions

$$N_1(\xi,\eta) = \frac{1}{4}(\xi-1)(\eta-1) \tag{2.26}$$

$$N_2(\xi,\eta) = \frac{1}{4}(\xi-1)(\eta+1) \tag{2.27}$$

$$N_3(\xi,\eta) = \frac{1}{4}(\xi+1)(\eta-1) \tag{2.28}$$

$$N_4(\xi,\eta) = \frac{1}{4}(\xi+1)(\eta+1) \tag{2.29}$$

for $-1 \le \xi, \eta \le 1$ and $N_1 = N_2 = N_3 = N_4 = 0$ otherwise. A local approximation valid over the element (i,j) which has its lower left corner at (x_i, y_j) is given by

$$\tilde{q}(\xi,\eta) = \sum_{k=1}^{4} Q_k^{(i,j)} N_k(\xi,\eta) \tag{2.30}$$

Here k is the local numbering of the unknowns associated with element (i,j). The weighted residual formulation leads to

$$\int_{-1}^{+1}\int_{-1}^{+1} L\left(\sum_{k=1}^{4} Q_k^{(i,j)} N_k(\xi,\eta)\right) N_k(\xi,\eta)\; d\xi d\eta = 0,\; 1 \le k \le 4,\; 1 \le i,j \le M\ . \tag{2.31}$$

This leads to a linear system for the unknown values $Q_k^{(i,j)}$. There intervene various practical complications involving numbering of unknowns and the fact that unknowns are based at the nodes and hence shared by more than one element. The numbering problem is readily solved while the issue of shared unknown values leads to the process of matrix assembly which shall be studied in detail when we analyze finite element methods.

2.4. Spectral methods. Up to now we have predominantly used *local*, *polynomial* approximating functions. There are good reasons to make such a choice for very many problems, especially those that contain jumps (discontinuities) in the unknown function or its derivatives. However if it is known that the function is smooth other basis functions can be used, for example global basis functions defined over an entire domain. When these basis functions have a special relation with the operator of the PDE being solved the resulting method is known as a spectral method. For instance for the operator

$$L = -\frac{d^2}{dx^2} \tag{2.32}$$

the functions $\sin(rx), \cos(rx)$ play a special role since they satisfy the eigenfunction relation

$$L\sin(rx) = r^2\sin(rx),\; L\cos(rx) = r^2\cos(rx)\ . \tag{2.33}$$

Such eigenfunctions have remarkable advantages as basis functions; in a very precise sense they offer the most compact representation of any given function in the space which they span. When they are used as basis functions in a numerical approximation the resulting method is known as a *spectral method*. Weight functions must also be chosen and there exist various procedures to do this. Again the Galerkin procedure is widely used in which the basis functions are also used as weight functions.