Numerics for Partial Differential Equations

a.o.Univ.Prof. Mag.Dr. Stephen Keeling
http://imsc.uni-graz.at/keeling/

Documentation und Literature:
http://imsc.uni-graz.at/keeling/teaching.html

These notes are based especially upon works of:
Vasilios Dougalis, Randall LeVeque and Christian Clason

but also upon works of:
Franz Kappel, Kazufumi Ito, Karl Kunisch and David Gottlieb

Table of Contents I

Table of Contents II

Table of Contents III

Table of Contents IV

Table of Contents V

Types of PDEs

- The standard types of partial differential equations (PDEs) are: *elliptic*, *parabolic* and *hyperbolic*.
- There are standard *algebraic* defintions of these types which one encounters in the continuum study of PDEs.
- For instance, the PDE

$$au_{xx}(x,y)+bu_{xy}(x,y)+cu_{yy}(x,y)+du_x(x,y)+eu_y(x,y)=0$$

  - is elliptic if $b^2-ac<0$,
  - parabolic if $b^2-ac=0$, and
  - hyperbolic if $b^2-ac>0$.
- But what about more complex PDEs? Equation type is typically defined in terms of characteristics, which may be defined for systems as well as for nonlinear problems.
- Especially for the numerical solution of PDEs we may understand these types qualitatively and intuitively in terms of the standard model equations.

Elliptic PDEs

- *Elliptic* PDEs are found typically in the modelling of *stationary fields* such as force-at-a-distance fields.
- For such PDEs every point is coupled with every other point, and there is no notion of evolution in time.
- Consider the displacement field of an unloaded membrane with a curved fixed boundary, modelled by the following *Laplace Equation* with a boundary condition.

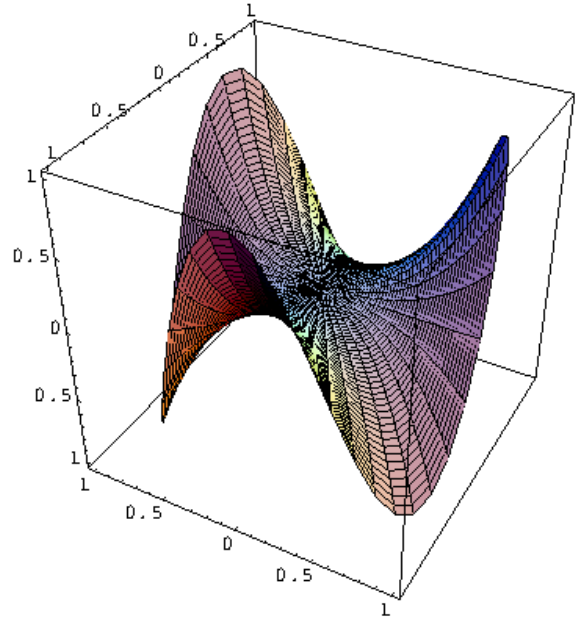
$$\begin{cases} u_{xx}(x,y) + u_{yy}(x,y) = 0, & \text{for } (x,y) \in \Omega = B(0,1) \\ u(x,y) = g(x,y), & \text{for } (x,y) \in \partial\Omega \end{cases}$$

With

$$g(x,y) = \sin(3\tan^{-1}(y/x))$$

the solution is

$$u(x,y) = (x^2+y^2)^{\frac{3}{2}} \sin(3\tan^{-1}(y/x))$$



Parabolic PDEs

- *Parabolic* PDEs are found typically in the modelling of *diffusion* processes.
- For such PDEs every point in space is coupled with every point in space, and there is an evolution in time which is *smoothing* since information travels at an infinite speed.
- Consider the evolution of temperature between suddenly connected hot and cold regions, modelled by the following *Heat Equation* with an initial condition.

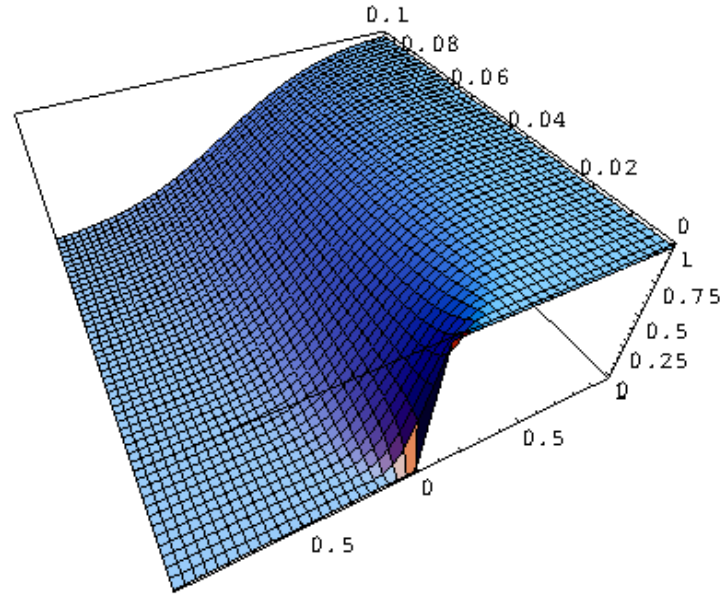
$$\begin{cases} u_t(x,t) &= u_{xx}(x,t), & \text{for } (x,t) \in \mathbb{R} \times (0,\infty) \\ u(x,0) &= u_0(x), & \text{for } (x,t) \in \mathbb{R} \times \{0\} \end{cases}$$

With

$$u_0(x) = \tfrac{1}{2}[1 + \text{sign}(x)]$$

the solution is

$$u(x,t) = \tfrac{1}{2}[1 + \text{erf}(x/\sqrt{4t})]$$



Hyperbolic PDEs

- *Hyperbolic* PDEs are found typically in the modelling of *waves*.
- For such PDEs there is an evolution in time which is *not smoothing* since information travels at a finite speed.
- Consider the evolution of displacement in a string suddenly stretched in a finite region, modelled by the following *Wave Equation* with initial conditions.

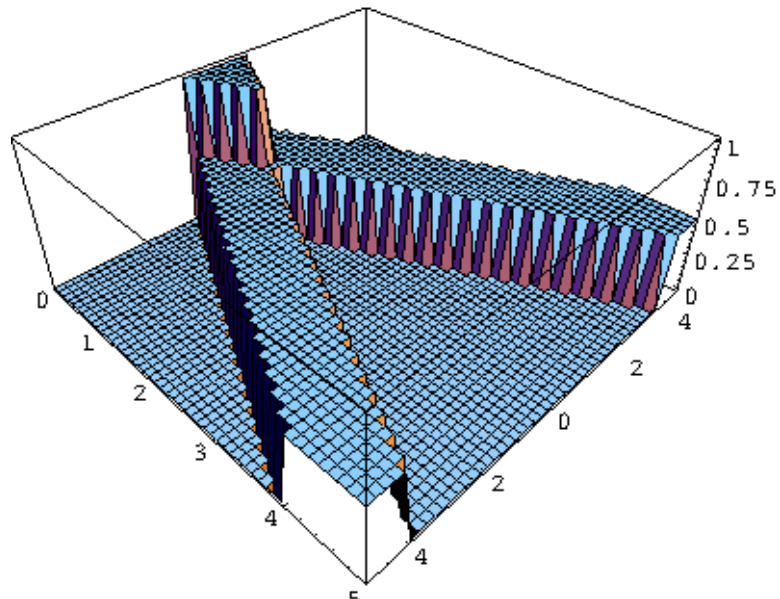
$$\begin{cases} u_{tt}(x,t) &= u_{xx}(x,t), & \text{for } (x,t) \in \mathbb{R} \times (0,\infty) \\ u(x,0) &= u_0(x), & \text{for } (x,t) \in \mathbb{R} \times \{0\} \\ u_t(x,0) &= u_1(x), & \text{for } (x,t) \in \mathbb{R} \times \{0\} \end{cases}$$

With $u_1(x) = 0$ and

$$u_0(x) = \text{sign}(x+1) - \text{sign}(x-1)$$

the solution is

$$\begin{aligned} u(x,t) = &\tfrac{1}{2}[\text{sign}(x+t+1) - \text{sign}(x+t-1)] \\ &+\tfrac{1}{2}[\text{sign}(x-t+1) - \text{sign}(x-t-1)] \end{aligned}$$



Convection Diffusion Equation

- But what about the following convection (velocity $\nu > 0$) and diffusion (diffusivity $\epsilon > 0$) equation?

$$\begin{cases} u_t(x,t) + \nu u_x(x,t) &= \epsilon u_{xx}(x,t), & \text{for } (x,t) \in \mathbb{R} \times (0,\infty) \\ u(x,0) &= g(x), & (x,t) \in \mathbb{R} \times \{0\} \end{cases}$$

With

$$g(x) = \text{sign}(x+1) - \text{sign}(x-1)$$

the solution is

$$u(x,t) = \frac{1}{2}\text{erf}\left[\frac{\nu t - x + 1}{\sqrt{4\epsilon t}}\right] - \frac{1}{2}\text{erf}\left[\frac{\nu t - x - 1}{\sqrt{4\epsilon t}}\right]$$

- This solution manifests more wave character when $\nu \gg \epsilon$ and more diffusion character when $\epsilon \gg \nu$.
- The situation is more complex for non-linear PDEs, but we will use the linear problems presented here as *model problems* upon which methods for non-linear problems can be based.

Hyperbolic Systems

- Note that the wave equation (now with sound speed $c > 0$) can be rewritten in the following form:
$$(\partial_t - c\partial_x)(\partial_t + c\partial_x)u = u_{tt} - c^2 u_{xx} = 0$$
where the factors $(\partial_t - c\partial_x)$ and $(\partial_t + c\partial_x)$ correspond respectively to right and left travelling waves.
- It is then advantageous to understand the numerical approximation for the linear convection equation
$$u_t - cu_x = 0$$
- Setting $v = cu_x$ we can rewrite the scalar wave equation as the following system of PDEs:
$$\begin{pmatrix} v \\ u_t \end{pmatrix}_t = \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix} \begin{pmatrix} v \\ u_t \end{pmatrix}_x$$
- It is also advantageous to understand the numerical approximation for linear hyperbolic systems:
$$\boldsymbol{u}_t = A\boldsymbol{u}_x, \quad \boldsymbol{u} = (u_1, \ldots, u_N)^\top, \quad A \in \mathbb{R}^{N \times N}$$

Classical Solution Procedures

- Consider the following *Poisson Equation* modelling a membrane clamped at the boundary of $\Omega = (0,1)^2$ and loaded internally with a force per unit area $f(x,y)$:

$$\begin{cases} -[u_{xx}(x,y) + u_{yy}(x,y)] &= f(x,y), & \text{for } (x,y) \in \Omega \\ u(x,0) &= g(x,y) = 0, & \text{for } (x,y) \in \partial\Omega \end{cases}$$

- Using separation of variables the solution can be written as:

$$u(x,y) = \sum_{n,m=1}^{\infty} \gamma_{m,n} \sin(n\pi x) \sin(m\pi y)$$

where

$$\gamma_{m,n} = \frac{4}{\pi^2(n^2+m^2)} \int_0^1 \int_0^1 f(x,y) \sin(n\pi x) \sin(n\pi y) dxdy$$

- There are similar *spectral formulas* for solutions to the heat equation.
- But what is the convergence rate? What to do when Ω is not so simple? We can as well use numerical methods!

Well-Posed BVPs

- For a bounded domain, e.g., $\Omega = (0, 1)$, a well-posed initial boundary value problem for the heat equation is given by:

$$\begin{cases} u_t &= u_{xx}, & \text{for } (x,t) \in \Omega \times (0,\infty) \\ u &= g, & \text{for } (x,t) \in \partial\Omega \times (0,\infty) \\ u &= u_0, & \text{for } (x,t) \in \Omega \times \{0\} \end{cases}$$

- For a bounded domain, e.g., $\Omega = (0, 1)$, a well-posed initial boundary value problem for the wave equation is given by:

$$\begin{cases} u_{tt} &= u_{xx}, & \text{for } (x,t) \in \Omega \times (0,\infty) \\ u &= g, & \text{for } (x,t) \in \partial\Omega \times (0,\infty) \\ u &= u_0, & \text{for } (x,t) \in \Omega \times \{0\} \\ u_t &= u_1, & \text{for } (x,t) \in \Omega \times \{0\} \end{cases}$$

- Given sufficient assumptions on the *regularity* of internal forces f, boundary terms g, initial values u_0 and u_1 and the boundary $\partial\Omega$, one can show there exists a unique solution u to our PDE with a certain regularity. We assume this as given.

Dirichlet Problem for the Poisson Equation

- Let $\Omega \subset \mathbb{R}^2$ be bounded, open and connected with sufficiently smooth $\partial\Omega$.
- We seek an approximation of the solution u to the *Dirichlet Problem for the Poisson Equation*,

$$\begin{cases} -\Delta u(x,y) &= f(x,y), \quad \text{for } (x,y) \in \Omega \\ u(x,0) &= g(x,y), \quad \text{for } (x,y) \in \partial\Omega \end{cases} \tag{1}$$

  where $\Delta u = u_{xx} + u_{yy}$ is the Laplace operator and f and g are assumed to be sufficiently regular.
- The *Dirichlet* data are g. The *Neumann* data would be g with a boundary condition $\partial_n u = g$.
- For $h > 0$ cover $\mathbb{R}^2$ with a *grid*

$$G_h = \{(ih, jh) : i, j \in \mathbb{Z}\}.$$

  consisting of *grid points* $(x, y) = (ih, jh)$.
- For every grid point $p = (ih, jh) \in G_h$ define the near neighbors

$$N_h(p) = \{(\alpha h, \beta h) : \alpha, \beta \in \mathbb{Z}, |i - \alpha| + |j - \beta| = 1\}$$

Discrete Laplacian

- Let (figure forthcoming)

$$\begin{aligned} \Omega_h &= \{p \in G_h : p \in \Omega, N_h(p) \subset \bar{\Omega}\} \\ \Omega_h^\star &= \{p \in G_h : p \in \Omega\} \backslash \Omega_h \\ \partial\Omega_h &= \{(x,y) \in \partial\Omega : x = ih \text{ or } y = jh\} \\ \bar{\Omega}_h &= \Omega_h \cup \Omega_h^\star \cup \partial\Omega_h \end{aligned} \tag{2}$$

- For $p = (x,y) \in \Omega_h$ define the discrete Laplacian, i.e., a finite difference approximation to the Laplace operator by

$$\begin{aligned} \Delta_h v(x,y) &= h^{-2}[v(x+h,y) + v(x,y+h) + \\ &\qquad v(x-h,y) + v(x,y-h) - 4v(x,y)] \\ &= h^{-2}[\textstyle\sum_{q \in N(p)} v(q) - 4v(p)] \end{aligned} \tag{3}$$

- For the sequel recall the multi-index notation
$\partial^\alpha u = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_1} u, \quad \alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^n, \quad |\alpha| = \|\alpha\|_{\ell_1}$
and the norms on $\mathcal{C}^k(\bar{\Omega})$,
$$\|u\|_{\mathcal{C}^k(\bar{\Omega})} = \max_{|\alpha| \le 4} \sup_{(x,y) \in \bar{\Omega}} |\partial^\alpha u(x,y)|$$

Discrete Laplacian

Lemma: For $u \in \mathcal{C}^4(\bar{\Omega})$ it holds that

$$\max_{(x,y)\in\Omega_h} |\Delta_h u(x,y) - \Delta u(x,y)| \leq \frac{h^2}{6} \|u\|_{\mathcal{C}^4(\bar{\Omega})}$$

Proof: Exercise with Taylor's Theorem. ■

- This Lemma is a *consistency* result in Ω_h.

Def: A numerical approximation to a differential operator is said to be (locally) *consistent* when, for a sufficiently smooth function, the difference between the discrete and continuous operator applied (locally) to the function converges to zero as discretization is refined infinitely.

- The discrete Laplacian will now be defined for points $(x, y) \in \Omega_h^\star$, whose near neighbors will be written as

  $$N_h^\star(x,y) = \{(x-\alpha h, y), (x, y-\beta h), (x+\gamma h, y), (x, y+\delta h)\}$$

  with
  $$0 < \alpha, \beta, \gamma, \delta \leq 1, \quad \alpha + \beta + \gamma + \delta < 4.$$
- The *Shortley-Weller Formula* gives

Shortley-Weller Formula

the discrete Laplacian for $(x, y) \in \Omega_h^\star$,

$$\Delta_h v(x,y) = \frac{2}{h^2}\left[\frac{1}{\gamma(\alpha+\gamma)}v(x+\gamma h, y) + \frac{1}{\alpha(\alpha+\gamma)}v(x-\alpha h, y) + \frac{1}{\delta(\beta+\delta)}v(x, y+\delta h) + \frac{1}{\beta(\beta+\delta)}v(x, y-\beta h) - \left(\frac{1}{\alpha\gamma}+\frac{1}{\beta\delta}\right)v(x,y)\right] \tag{4}$$

Lemma: For $u \in \mathcal{C}^3(\bar{\Omega})$ it holds that

$$\max_{(x,y)\in\Omega_h^\star} |\Delta_h u(x,y) - \Delta u(x,y)| \leq \frac{2h}{3}\|u\|_{\mathcal{C}^3(\bar{\Omega})}$$

Proof: Exercise with Taylor's Theorem. Hint:

$$\sup_{0<x,y<1}(x^2+y^2)/(x+y) = 1.$$

Note: An $\mathcal{O}(h^2)$ approximation is only possible if $\alpha = \beta = \gamma = \delta = 1$. **Exercise**: prove this.

Finite Difference Scheme

- We now define the finite difference scheme approximating the solution to the Dirichlet Problem for the Poisson equation:

$$\begin{cases} -\Delta_h U(x,y) &= f(x,y), \quad \text{for } (x,y) \in \Omega_h \cup \Omega_h^\star \\ U(x,y) &= g(x,y), \quad \text{for } (x,y) \in \partial\Omega_h \end{cases} \tag{5}$$

- Let $d = \#(\Omega_h \cup \Omega_h^\star)$. The above problem (5) corresponds to a $d \times d$ system of linear equations for the unknown values of U. (The values of U are known and given by g at $\partial\Omega_h$.)

Exercise: Let $\Omega = (0,1)^2$. Write the above system (5) in matrix form and prove (by easier means than used below for the general case) that the system possesses a unique solution.

Theorem (discrete maximum principle): Let v be any grid function satisfying $\Delta_h v(x,y) \geq 0$, $\forall (x,y) \in \Omega_h \cup \Omega_h^\star$. Then

$$\max_{(x,y)\in\bar{\Omega}_h} v(x,y) = \max_{(x,y)\in\partial\Omega_h} v(x,y).$$

Discrete Maximum and Minimum Principles

Proof: Define the clearly non-empty set of grid points

$$\mathcal{P} = \{p \in \bar{\Omega}_h : v(p) = \max_{q \in \bar{\Omega}_h} v(q)\}$$

To avoid the trivial case, suppose $\mathcal{P} \cap (\Omega_h \cup \Omega_h^\star) \neq \emptyset$.

For $p \in \mathcal{P} \cap \Omega_h$,

$$\Delta_h v(p) = \left[\sum_{q \in N_h(p)} v(q) - 4v(p)\right] / h^2 \geq 0.$$

Since $v(p) = \max_{q \in \bar{\Omega}_h} v(q)$, it must be that that $v(q) = v(p)$ for $q \in N_h(p)$; otherwise, if $v(p) > v(q)$ for some $q \in N_h(p)$, then the above inequality is violated.

In this way we continue the argument until we have a $p \in \mathcal{P} \cap \Omega_h$ with $q \in N_h(p) \cap \Omega_h^\star$ and the argument above gives $q \in \mathcal{P} \cap \Omega_h^\star$.

Thus let $p \in \mathcal{P} \cap \Omega_h^\star$. Again it must be that $v(q) = v(p)$ for $q \in N_h^\star(p)$; otherwise, if $v(p) > v(q)$ for some $q \in N_h^\star(p)$, then

$$\Delta_h v(p) < \frac{2}{h^2}\left[\frac{1}{\gamma(\alpha+\gamma)} + \frac{1}{\alpha(\alpha+\gamma)} + \frac{1}{\delta(\beta+\delta)} + \frac{1}{\beta(\beta+\delta)} - \frac{1}{\alpha\gamma} - \frac{1}{\beta\delta}\right] v(p) = 0$$

Discrete Maximum and Minimum Principles

which contradicts the hypothesis that $\Delta_h v \geq 0$ in $\Omega_h \cup \Omega_h^\star$. Since at least one $q \in N_h^\star(p)$ satisfies $q \in \partial\Omega_h$, the claim follows since $q \in \mathcal{P}$. ■

Corollary: (discrete minimum principle): Let v be any grid function satisfying $\Delta_h v(x,y) \leq 0$, $\forall (x,y) \in \Omega_h \cup \Omega_h^\star$. Then

$$\min_{(x,y)\in\bar{\Omega}_h} v(x,y) = \min_{(x,y)\in\partial\Omega_h} v(x,y).$$

Proof: Apply the last theorem to the grid function $-v$. ■

Theorem: The finite difference scheme (5) has a unique solution.

Proof: The scheme (5) is a $d \times d$ system of linear equations $A_h U_h = F_h$, where $U_h \in \mathbb{R}^d$ is a vector of values of U on $\Omega_h \cup \Omega_h^\star$, $A_h \in \mathbb{R}^{d\times d}$ is a matrix independent of f and g and $F_h \in \mathbb{R}^d$ depends upon grid values of f and g. (Recall the Exercise 19.)

Existence of a Discrete Solution

It will be shown that the system $A_h U_h = 0$, corresponding to $f = 0$ and $g = 0$, has only the solution $U_h = 0$. Let U be the grid function with values U_h in $\Omega_h \cup \Omega_h^\star$, corresponding to $f = 0$ and $g = 0$. Then $\Delta_h U(x, y) = 0$, $\forall (x, y) \in \Omega_h \cup \Omega_h^\star$. By the discrete maximum principle,

$$\max_{(x,y)\in\bar{\Omega}_h} U(x, y) = \max_{(x,y)\in\partial\Omega_h} U(x, y) = 0.$$

By the discrete minimum principle,

$$\min_{(x,y)\in\bar{\Omega}_h} U(x, y) = \min_{(x,y)\in\partial\Omega_h} U(x, y) = 0.$$

Hence, $U(x, y) = 0$, $\forall (x, y) \in \bar{\Omega}_h$. It follows that $U_h = 0$ and thus A_h is invertible. ■

- For the proof of convergence $U \to u$ as $h \to 0$, together with an error estimate, we introduce the *discrete Green's function* (analogous to the Green's function used in the continuum setting) as follows.

Discrete Green's Function

Def: For fixed $q \in \bar{\Omega}_h$ define the *discrete Green's function* as the grid function $G_h(p;q)$ for $p \in \bar{\Omega}_h$ as the (unique) solution to

$$\begin{cases} -\Delta_{h,p} G_h(p;q) &= h^{-2}\delta(p;q), & \text{for } p \in \Omega_h \cup \Omega_h^\star \\ G_h(p;q) &= \delta(p;q), & \text{for } p \in \partial\Omega_h \end{cases} \tag{6}$$

where

$$\delta(p;q) = \begin{cases} 1, & p = q \\ 0, & p \neq q \end{cases}$$

Lemma: Let v be a grid function defined on $\bar{\Omega}_h$. Then for any $p \in \bar{\Omega}_h$ there holds

$$v(p) = \sum_{q \in \partial\Omega_h} G_h(p;q) v(q) - h^2 \sum_{q \in \Omega_h \cup \Omega_h^\star} G_h(p;q) \Delta_h v(q)$$

Proof: Define the mesh function for $p \in \bar{\Omega}_h$,

$$w(p) = \sum_{q \in \partial\Omega_h} G_h(p;q) v(q) - h^2 \sum_{q \in \Omega_h \cup \Omega_h^\star} G_h(p;q) \Delta_h v(q)$$

Discrete Green's Function

Then for $p \in \Omega_h \cup \Omega_h^\star$, since $G_h(p;q) = \delta(p;q) = 0$ for $q \in \partial\Omega_h$, we have by (6) that

$$\Delta_h w(p) = -h^2 \sum_{q \in \Omega_h \cup \Omega_h^\star} \Delta_{h,p} G_h(p;q) \Delta_h v(q)$$

$$= -h^2 \sum_{q \in \Omega_h \cup \Omega_h^\star} [-h^{-2}\delta(p;q)] \Delta_h v(q) = \Delta_h v(p)$$

Then for $p \in \partial\Omega_h$,

$$w(p) = \sum_{q \in \partial\Omega_h} \delta(p;q) v(q) = v(p)$$

Hence, w satisfies (5) with $f = \Delta_h v$ and $g = v$, as does v. By the uniqueness of the solution to (5), it follows that $w = v$. ■

- We next summarize properties of the discrete Green's function which are used for the convergence estimate.

Properties of the Discrete Green's Function

Lemma: It holds that

$$G_h(p;q) \geq 0, \quad \forall p, q \in \bar{\Omega}_h.$$

Proof: Fix $q \in \bar{\Omega}_h$. Clearly $G_h(p;q) = \delta(p;q) \geq 0$ if $p \in \partial\Omega_h$. If $p \in \Omega_h \cup \Omega_h^\star$, then by (6), $\Delta_{h,p} G_h(p;q) = -h^{-2}\delta(p;q) \leq 0$. Applying the discrete minimum principle, we obtain

$$G_h(p;q) \geq \min_{s \in \bar{\Omega}_h} G_h(s;q) = \min_{s \in \partial\Omega_h} G_h(s;q) \geq 0. \qquad \blacksquare$$

Lemma: It holds that

$$\sum_{q \in \Omega_h^\star} G_h(p;q) \leq 1, \quad \forall p \in \bar{\Omega}_h.$$

Proof: Define the grid function

$$w(p) = \begin{cases} 1, & p \in \Omega_h \cup \Omega_h^\star \\ 0, & p \in \partial\Omega_h \end{cases}$$

Let $p \in \Omega_h$. Then by definition $\Delta_h w(p) = 0$. Now let $p \in \Omega_h^\star$. Then $-\Delta_h w(p)$ is given by the Shortley-Weller formula (4). Checking

Properties of the Discrete Green's Function

all cases for the number of points in $N_h^\star(p) \cap \partial\Omega_h$ shows that $\Delta_h w(p) \leq -h^{-2}$ (**Exercise**). Thus, w satisfies

$$\Delta_h w(p) \begin{cases} = 0, & p \in \Omega_h \\ \leq -h^{-2}, & p \in \Omega_h^\star \end{cases}$$

Now apply the discrete Green's function to represent w as

$$w(p) = \sum_{q \in \partial\Omega_h} G_h(p;q) w(q) - h^2 \sum_{q \in \Omega_h \cup \Omega_h^\star} G_h(p;q) \Delta_h w(q)$$

$$= -h^2 \sum_{q \in \Omega_h \cup \Omega_h^\star} G_h(p;q) \Delta_h w(q) \geq \sum_{q \in \Omega_h^\star} G_h(p;q).$$

For $p \in \Omega_h \cup \Omega_h^\star$, the claimed estimate follows with $w(p) = 1$ on the left side of the last estimate. For $p \in \partial\Omega_h$, the claimed estimate follows trivially since $G_h(p;q) = 0$, $\forall q \in \Omega_h^\star$. ■

Properties of the Discrete Green's Function

Lemma: Let $\rho = \rho(\Omega)$ denote the diameter of the smallest circumscribed circle containing Ω. Then,

$$h^2 \sum_{q \in \Omega_h \cup \Omega_h^\star} G_h(p;q) \leq \frac{\rho^2}{16}, \quad \forall p \in \bar{\Omega}_h$$

Proof: Let (x_0, y_0) be the center of the smallest circumscribed circle containing Ω. Define the grid function

$$w(x,y) = \tfrac{1}{4}[(x-x_0)^2 + (y-y_0)^2], \quad (x,y) \in \bar{\Omega}_h$$

By a direct calculation (**Exercise**),

$$\Delta_h w(p) = 1, \quad p \in \Omega_h \cup \Omega_h^\star.$$

Also for $p \in \partial\Omega_h$,

$$0 \leq w(p) \leq \tfrac{1}{4}\,(\rho/2)^2 = \rho^2/16.$$

Now define

$$v(p) = h^2 \sum_{q \in \Omega_h \cup \Omega_h^\star} G_h(p;q), \quad p \in \bar{\Omega}_h.$$

Then if $p \in \Omega_h \cup \Omega_h^\star$, using (6), $\Delta_h v(p) = -1$ follows from

Properties of the Discrete Green's Function

$$\begin{aligned}\Delta_h v(p) &= h^2 \sum_{q\in\Omega_h\cup\Omega_h^*} \Delta_{h,p} G_h(p;q) \\ &= h^2 \sum_{q\in\Omega_h\cup\Omega_h^*} [-h^{-2}\delta(p;q)] = -1.\end{aligned}$$

Also if $p \in \partial\Omega_h$, by (6),

$$v(p) = h^2 \sum_{q\in\Omega_h\cup\Omega_h^\star} \delta(p;q) = 0.$$

Hence

$$\begin{cases} \Delta_h[w(p)+v(p)] &= 0, & p\in\Omega_h\cup\Omega_h^\star \\ w(p)+v(p) &\le \rho^2/16, & p\in\partial\Omega_h \end{cases}$$

Applying the discrete maximum principle to $w+v$ gives

$$\max_{p\in\bar\Omega_h} [w(p)+v(p)] \le \rho^2/16.$$

Hence, by the definition of v,

$$w(p) + h^2 \sum_{q\in\Omega_h\cup\Omega_h^\star} G_h(p;q) \le \rho^2/16$$

and since $w \ge 0$, the claimed estimate follows. ■

Convergence of the Discrete Solution

Theorem: Suppose the solution u to (1) satisfies $u \in \mathcal{C}^4(\bar{\Omega})$. Let U be the solution to (5). Then

$$\max_{(x,y)\in\bar{\Omega}_h} |u(x,y) - U(x,y)| \leq \frac{h^2\rho^2}{96}\|u\|_{\mathcal{C}^4(\bar{\Omega})} + \frac{2h^3}{3}\|u\|_{\mathcal{C}^3(\bar{\Omega})}$$

Proof: Let $e(p) = u(p) - U(p)$, $p \in \bar{\Omega}_h$. Since for $p \in \partial\Omega_h$, $e(p) = u(p) - U(p) = g(p) - g(p) = 0$, it follows with Lemma 23 that

$$e(p) = h^2 \sum_{q\in\Omega_h\cup\Omega_h^\star} G_h(p;q)[-\Delta_h e(q)].$$

For $q \in \Omega_h \cup \Omega_h^\star$,

$$\Delta_h e(q) = \Delta_h u(q) - \Delta_h U(q) = \Delta_h u(q) - f(q) = \Delta_h u(q) - \Delta u(q).$$

By Lemmas 17 and 18,

$$|\Delta_h e(q)| = |\Delta_h u(q) - \Delta u(q)| \leq \begin{cases} \frac{1}{6}h^2\|u\|_{\mathcal{C}^4(\bar{\Omega})}, & q \in \Omega_h \\ \frac{2}{3}h\|u\|_{\mathcal{C}^3(\bar{\Omega})}, & q \in \Omega_h^\star \end{cases}$$

Convergence of the Discrete Solution

Hence, by the above Green's function characterization of e,

$$|e(p)| \leq \frac{1}{6}h^2\|u\|_{\mathcal{C}^4(\bar{\Omega})}\left[h^2\sum_{q\in\Omega_h} G_h(p;q)\right] + \frac{2}{3}h\|u\|_{\mathcal{C}^3(\bar{\Omega})}\left[h^2\sum_{q\in\Omega_h^\star} G_h(p;q)\right]$$

where the estimate of $|\Delta_h e|$ has been used together with the property $G_h(p;q) \geq 0$ from Lemma 25. Using the property

$$\sum_{q\in\Omega_h^\star} G_h(p;q) \leq 1$$

from Lemma 25 and the estimate

$$h^2\sum_{q\in\Omega_h} G_h(p;q) \leq h^2\sum_{q\in\Omega_h\cup\Omega_h^\star} G_h(p;q) \leq \rho^2/16$$

of Lemma 27, the claimed convergence estimate follows. ■

Neumann Problem for the Poisson Equation

- We now seek an approximation of the solution u to the *Neumann Problem for the Poisson Equation*,

$$\begin{cases} -\Delta u(x,y) &= f(x,y), \quad \text{for } (x,y) \in \Omega \\ \partial_n u(x,0) &= g(x,y), \quad \text{for } (x,y) \in \partial\Omega \end{cases} \tag{7}$$

  where

$$\partial_n u(x,y) = \nabla u(x,y) \cdot n(x,y), \quad (x,y) \in \partial\Omega$$

  for an outwardly directed unit normal vector $n(x,y)$.

- Also, f and g are assumed to be sufficiently regular and to satisfy the compatibility condition

$$\int_\Omega f(x,y)dxdy = -\int_\partial g(x,y)d\sigma(x,y)$$

  which, according to the *Green's Identity* with $v = 1$,

$$\int_\Omega [u\Delta v - v\Delta u]dxdy = \int_{\partial\Omega} [u\partial_n v - v\partial_n u]d\sigma(x,y)$$

  is necessary for the existence of a solution u to (7).

Neumann Problem for the Poisson Equation

- Note: A solution u to (7) can be unique only up to a constant. One can show there exists a unique solution (with given regularity, depending upon the regularity of the data) under an additional condition such as

$$\int_\Omega u(x,y)dxdy = 0$$

and we will also consider the case that the solution is known at a particular point $(\hat{x},\hat{y}) \in \Omega$,

$$u(\hat{x},\hat{y}) = \hat{u}.$$

- For a finite difference approximation of (7) let Ω_h, $\Omega_h^\star$ and $\partial\Omega_h$ be as in (2) and $\Delta_h U$ as in (3)-(4) for $p \in \Omega_h \cup \Omega_h^\star$.
- Yet the new condition $\partial_n u = g$ must now be discretized.
- For $p \in \partial\Omega_h$ we seek points $p_1, p_2 \in \Omega_h \cup \Omega_h^\star$ for a 3-point approximation to $\partial_n u(p)$,

$$\partial_n u(p) \approx b_1[u(p) - u(p_1)] + b_2[u(p) - u(p_2)]$$

Approximation of the Normal Derivative

- Let (ν, τ) be (inwardly) normal and (counter-clockwise) tangent coordinates, respectively, of a local system with origin at p and $\det[\partial(\nu,\tau)/\partial(x,y)] = 1$.

  (figure forthcoming)

- Let $p_i = (\nu_i, \tau_i)$ and $p = (0,0)$ in the local coordinate system.
- By Taylor's Theorem, for some $q_i, r_i, s_i \in \Omega$,

$$u(p_i) = u(p) + \nu_i \partial_\nu u(p) + \tau_i \partial_\tau u(p) + \tfrac{1}{2}\left[\nu_i^2 \partial_\nu^2 u(q_i) + 2\nu_i\tau_i \partial_{\nu\tau}^2 u(r_i) + \tau_i^2 \partial_\tau^2 u(s_i)\right]$$

- For constant's b_1, b_2,

$$\begin{aligned} b_1 u(p_1) + b_2 u(p_2) - (b_1 + b_2)u(p) = \\ (\nu_1 b_1 + \nu_2 b_2)\partial_\nu u(p) + (\tau_1 b_1 + \tau_2 b_2)\partial_\tau u(p) \\ + \tfrac{1}{2}\Big[\nu_1^2 b_1 \partial_\nu^2 u(q_1) + \nu_2^2 b_2 \partial_\nu^2 u(q_2) \\ + 2\nu_1\tau_1 b_1 \partial_{\nu\tau}^2 u(r_1) + 2\nu_2\tau_2 b_2 \partial_{\nu\tau}^2 u(r_2) \\ + \tau_1^2 b_1 \partial_\tau^2 u(s_1) + \tau_2^2 b_2 \partial_\tau^2 u(s_2)\Big] \end{aligned}$$

Approximation of the Normal Derivative

- The points $p_1 = (\nu_1, \tau_1)$, $p_2 = (\nu_2, \tau_2)$ and the constants b_1, b_2 are now chosen so that

$$(\nu_1 b_1 + \nu_2 b_2)\partial_\nu u(p) + (\tau_1 b_1 + \tau_2 b_2)\partial_\tau u(p) = -\partial_n u(p)$$

- Since $\partial_\nu = -\partial_n$ holds by construction, we require

$$(\nu_1 b_1 + \nu_2 b_2) = 1, \quad (\tau_1 b_1 + \tau_2 b_2) = 0.$$

- An approximation to $\partial_n u(p)$ is thus given by

$$\partial_n u(p) \approx (b_1 + b_2)u(p) - b_1 u(p_1) - b_2 u(p_2)$$

(figure forthcoming)

- For h sufficiently small we may choose $p_1, p_2 \in \Omega_h \cup \Omega_h^\star$ with $\nu_1, \nu_2 > 0$ and $\tau_1 \tau_2 < 0$, and hence,

$$b_1, b_2 > 0.$$

- **Exercise**: A direct calculation shows that

$$\nu_1, \nu_2, \tau_1, \tau_2 = \mathcal{O}(h)$$

and hence

$$b_1, b_2 = \mathcal{O}(1/h).$$

Finite Difference Scheme

- Note: For the case of a horizontal or vertical stretch of $\partial\Omega$ it is natural to take $p_1 = (\nu_1, \tau_1) = (1, 0)$ and $b_1 = 1/h$ with no p_2 or b_2. Nevertheless, for h sufficiently small, it is still possible to carry out the above two-point construction even in this simple case.
- Thus, the following first order estimate follows from the b_1, b_2-weighted Taylor expansion above, $c \neq c(h, u)$, $|\partial_n u(p) - [(b_1 + b_2)u(p) - b_1 u(p_1) - b_2 u(p_2)]| \leq ch\|u\|_{C^2(\bar{\Omega})}$
- We now define the finite difference scheme approximating the solution to the Neumann Problem for the Poisson equation:

$$\begin{cases} -\Delta_h U(x,y) &= f(x,y), & \text{for } (x,y) \in \Omega_h \cup \Omega_h^\star \backslash \{s\} \\ U(x,y) &= u(x,y), & \text{for } (x,y) = s \\ (b_1 + b_2)U(p) & & \\ -b_1 U(p_1) - b_2 U(p_2) &= g(p), & \text{for } (x,y) \in \partial\Omega_h \end{cases} \tag{8}$$

where the value $u(s)$ is assumed to be known in order that there be a unique solution to (7) and (8).

Monotone Matrices

- Matrix methods will now be used to show that (8) has a unique solution.

Def: $A = \{a_{ij}\} \in \mathbb{R}^{N\times N}$ is here *of positive type* iff

a. $a_{ij} \leq 0$, $i, j \in \mathcal{I} = \{1, 2, \ldots, N\}$, $i \neq j$,

b. $\sum_{j=1}^{N} a_{ij} \geq 0$, $i \in \mathcal{I}$,

c. There exists $\mathcal{J}(A) \subset \mathcal{I}$, $\mathcal{J}(A) \neq \emptyset$, such that $\sum_{j=1}^{N} a_{ij} > 0$ for $i \in J(A)$ and

d. for $i \notin \mathcal{J}(A)$ there exists a *connection* in A from i to $\mathcal{J}(A)$, non-zero elements $\{a_{i,k_1}, a_{k_1,k_2}, \ldots, a_{a_{k_m,j}}\}$, $j \in \mathcal{J}(A)$, $\{k_l\}_{l=1}^{m} \subset \mathcal{I}$, $k_{l_1} \neq k_{l_2}$, $k_1 \neq i$, $k_m \neq j$.

Def: $A = \{a_{ij}\} \in \mathbb{R}^{N\times N}$ is here *non-negative*, written $A \geq 0$, iff $a_{ij} \geq 0$, $1 \leq i, j \leq N$, and *non-positive* iff $-A$ is non-negative.

Def: $A = \{a_{ij}\} \in \mathbb{R}^{N\times N}$ is here monotone if $A\boldsymbol{x} \geq 0 \Rightarrow \boldsymbol{x} \geq 0$ (i.e., $\boldsymbol{x} = \{x_i\}$, $x_i \geq 0$) for any $\boldsymbol{x} \in \mathbb{R}^N$.

Monotone Matrices

Lemma: If $A = \{a_{ij}\} \in \mathbb{R}^N$ is monotone, then A is non-singular and A^{-1} is non-negative.

Proof: Let A be monotone. Let $\boldsymbol{x}$ satisfy $A\boldsymbol{x} = \mathbf{0}$. By monotonicity, $\boldsymbol{x} \geq 0$. Also, $A(-\boldsymbol{x}) = 0$ implies $-\boldsymbol{x} \geq 0$. Hence $\boldsymbol{x} = \mathbf{0}$ implies A is invertible. Now let $\boldsymbol{z} \in \mathbb{R}^N$ be the ith column of A^{-1}, so by $AA^{-1} = I$, $A\boldsymbol{z}$ is the ith column of I, and in particular, $A\boldsymbol{z} \geq 0$. By monotonicity, $\boldsymbol{z} \geq 0$, and thus, $A^{-1} \geq 0$.

Lemma: If $A = \{a_{ij}\} \in \mathbb{R}^N$ is of positive type, then A is monotone.

Proof: Since A is of positive type, $a_{ij} \leq 0$ holds for $i \neq j$ by condition (a), and $\sum_{j=1}^{N} a_{ij} \geq 0$ holds for $i \in \mathcal{I}$ by condition (b). In particular, $a_{ii} \geq -\sum_{i \neq j=1}^{N} a_{ij} \geq 0$.

If $a_{ii} = 0$ were to hold for some i, then the last inequality would mean that $a_{ij} = 0$ would hold $\forall j \in \mathcal{I}$. However, this would violate condition (c) if $i \in \mathcal{J}(A)$ or condition (d) if $i \notin \mathcal{J}(A)$. Thus, it

Monotone Matrices

follows that $a_{ii} > 0$, $i \in \mathcal{I}$.

Now suppose that $\boldsymbol{x} \in \mathbb{R}^N$ is such that $A\boldsymbol{x} \geq 0$. It will be shown that $\boldsymbol{x} \geq 0$. Componentwise, $A\boldsymbol{x} \geq 0$ is written as

$$a_{ii}x_i + \sum_{i \neq j = 1}^{N} a_{ij}x_j \geq 0, \quad i \in \mathcal{I}.$$

Since $a_{ij} \leq 0$, $i \neq j$,

$$a_{ii}x_i - \sum_{i \neq j = 1}^{N} |a_{ij}|x_j \geq 0, \quad i \in \mathcal{I}$$

and since $a_{ii} > 0$,

$$x_i \geq \sum_{i \neq j = 1}^{N} |a_{ij}|x_j / a_{ii}, \quad i \in \mathcal{I}.$$

Now let $r \in \mathcal{I}$ be chosen so that $x_r \leq x_i$, $i \in \mathcal{I}$. It will be shown that $x_r \geq 0$, which implies $\boldsymbol{x} \geq 0$.

Assume that $x_r < 0$. Let first $r \in \mathcal{J}(A)$. Then the general

Monotone Matrices

estimate of x_i above gives,

$$x_r \geq \sum_{r\neq j=1}^{N} |a_{rj}| x_j / a_{rr} \geq \sum_{r\neq j=1}^{N} |a_{rj}| x_r / a_{rr}$$

or by dividing by $x_r < 0$,

$$a_{rr} \leq \sum_{r\neq j=1}^{N} |a_{rj}|.$$

However, if $r \in \mathcal{J}(A)$, then condition (c) means

$$\sum_{j=1}^{N} a_{rj} > 0 \quad \text{and thus} \quad \sum_{r\neq j=1}^{N} a_{rj} > -a_{rr}$$

or with condition (a) and the previous estimate of a_{rr},

$$a_{rr} > \sum_{r\neq j=1}^{N} |a_{rj}| \geq a_{rr}$$

a contradiction. So if $x_r < 0$, then $r \notin \mathcal{J}(A)$. Then by condition (d), $\exists a_{r,k_1} \neq 0$. It will be shown that $x_r = x_{k_1}$. By condition (d),

Monotone Matrices

$k_1 \neq r$. If $x_{k_1} > x_r = \min\{x_i\}$, then by the general estimate of x_i,

$$a_{rr}x_r \geq \sum_{r,k_1 \neq j=1}^{N} |a_{rj}|x_j + |a_{r,k_1}|x_{k_1} > \sum_{r,k_1 \neq j=1}^{N} |a_{rj}|x_r + |a_{r,k_1}|x_r$$

or by dividing by $x_r < 0$,

$$a_{rr} < \sum_{r \neq j=1}^{N} |a_{rj}| = -\sum_{r \neq j=1}^{N} a_{rj}$$

which contradicts condition (b). Hence, $x_r = x_{k_1}$. Arguing similarly as if k_1 were r, we find for the connection $\{a_{r,k_1}, a_{k_1,k_2}, \ldots, a_{k_m,j}\}$ in A from r to $j \in \mathcal{J}(A)$ that $x_r = x_{k_1} = x_{k_2} = \cdots = x_{k_m} = x_j$.

However, as it was shown that for $x_r < 0$ to hold it must be that $r \notin \mathcal{J}(A)$, the same argument can be applied to $x_j = x_r < 0$ to conclude the necessity of $j \notin \mathcal{J}$. The contradiction implies that $\min\{x_i\} = x_r \geq 0$ or $\boldsymbol{x} \geq 0$, and hence A is monotone. ■

Existence of a Discrete Solution

Theorem: The finite difference scheme (8) has a unique solution.

Proof: Let $d = \#\bar{\Omega}_h$. The scheme (8) is an $d \times d$ system of linear equations $A_h U_h = F_h$, where $U_h \in \mathbb{R}^d$ is a vector of values of U on $\bar{\Omega}_h$, $A_h \in \mathbb{R}^{d \times d}$ is a matrix independent of f, g and $u(s)$ and $F_h \in \mathbb{R}^d$ depends upon $u(s)$ and grid values of f and g. It will be shown that A_h is of positive type.

To see that condition (a) is satisfied, consider first the rows i of $A_h = \{a_{ij}\}$ corresponding to $p \in \Omega_h \cup \Omega_h^\star \backslash \{s\}$. For these, the equations of (8) are

$$-\Delta_h U(p) = f(p)$$

for which the off-diagonal elements are ≤ 0, the non-trivial ones corresponding to $N_h(p)$ or $N_h^\star(p)$ being negative. For the row i corresponding to $p = s$, $a_{ii} = 1$ and $a_{ij} = 0$, $j \neq i$. For the rows i corresponding to $p \in \partial\Omega_h$, the equations of (8) are

Monotone Matrices

$$(b_1 + b_2)U(p) - b_1 U(p_1) - b_2 U(p_2) = g(p)$$

where $b_1, b_2 > 0$. Thus, condition (a) is satisfied.

To see that condition (b) is satisfied, note that if $p \in \Omega_h \cup \Omega_h^\star \backslash \{s\}$, the corresponding row in A_h has sum of elements zero. Specifically, for $p \in \Omega_h$, checking the sum in (3) gives $(4 - 1 - 1 - 1 - 1)/h^2 = 0$. For $p \in \Omega_h^\star$ a zero-sum is also obtained from the Shortley-Weller formula (4). For $p \in \partial\Omega_h$ a zero-sum is obtained from the above approximation to the normal derivative. For $p = s$, the sum of elements is

$$\sum_{j=1}^{N} a_{ij} = a_{ii} = 1.$$

Now let $\mathcal{J}(A_h)$ consist solely of the index k corresponding to the point s. Then condition (c) is satisfied according to the equation above.

To see that condition (d) is satisfied, let p be a point corresponding to any index $i \neq k$. Then the existence of a

Convergence of the Discrete Solution

connection

$$\{a_{i,k_1}, a_{k_1,k_2}, \ldots, a_{a_{k_m},k}\}$$

in A_h between $i \neq k$ and $k = \mathcal{J}(A_h)$ is equivalent to the existence of a zig-zag path moving horizontally or vertically among grid points in $\bar{\Omega}_h$ from the point p to the point s.
Exercise: The existence of such a path follows with h sufficiently small from the assumption that Ω is connected.
Thus, A_h is of positive type.

By Lemmas 37, A_h is monotone and hence non-singular. ■

- The convergence $U \to u$ as $h \to 0$ can be shown, but the following convergence estimate is stated here without proof.

Theorem: Suppose the solution u to (7) satisfies $u \in \mathcal{C}^3(\bar{\Omega})$.
Let U be the solution to (8). Then

$$\max_{(x,y)\in\bar{\Omega}_h} |u(x,y) - U(x,y)| \leq c(u)h|\log(h)|$$

where the constant $c(u)$ depends upon u but not upon h.

Elliptic BVPs with Variable Coefficients

- We now seek an approximation of the solution u to the *Elliptic BVP with variable coefficients*,

$$\begin{cases} [Lu](x,y) &= f(x,y), \quad \text{for } (x,y) \in \Omega \\ u(x,y) &= g(x,y), \quad \text{for } (x,y) \in \partial\Omega \end{cases} \tag{9}$$

  where

$$[Lu](x,y) = -\nabla \cdot [a_1(x,y)\nabla u(x,y)] + a_0(x,y)u(x,y)$$

- The coefficients are assumed to be sufficiently regular and to satisfy

$$a_1(x,y) \geq \alpha_1 > 0, \quad a_0(x,y) \geq \alpha_0 \geq 0, \quad \forall (x,y) \in \bar{\Omega}$$

- The data f and g are assumed to be sufficiently regular.
- Similarly we could consider the Neumann boundary condition $\partial_n u = g$.
- Another standard boundary condition is given by the *Robin boundary condition* $\sigma_1 \partial_n u + \sigma_0 u = g$ with $\sigma_1, \sigma_0 \geq 0$ and $\sigma_1 + \sigma_0 > 0$.
- In applications, mixed problems also arise in which different boundary conditions are imposed on different parts of $\partial\Omega$, e.g., with σ_1 or σ_0 vanishing at points in $\partial\Omega$.

Elliptic BVPs with Variable Coefficients

- To approximate (9) let $\bar{\Omega}_h$ be defined as in (2).
- For $p = (x, y) \in \Omega_h$ define $L_h \approx L$ by by adapting (3),

$$\begin{aligned}[L_h v](x,y) \;=\; -h^{-2}\big\{&a_1(x+\tfrac{1}{2}h, y)\,[v(x+h,y) - v(x,y)]\\ &-a_1(x-\tfrac{1}{2}h, y)\,[v(x,y) - v(x-h,y)]\\ &+a_1(x, y+\tfrac{1}{2}h)\,[v(x,y+h) - v(x,y)]\\ &-a_1(x, y-\tfrac{1}{2}h)\,[v(x,y) - v(x,y-h)]\big\}\\ &+a_0(x,y)v(x,y)\end{aligned}$$

- For $p = (x, y) \in \Omega_h^\star$ define L_h by adapting (4),

$$\begin{aligned}[L_h v](x,y) \;=\; -\frac{2}{h^2}\Bigg\{&\frac{a_1(x+\frac{1}{2}\gamma h, y)}{\alpha\gamma(\alpha+\gamma)}\,[\alpha v(x+\gamma h, y) - (\alpha+\gamma)v(x,y)]\\ &-\frac{a_1(x-\frac{1}{2}\alpha h, y)}{\alpha\gamma(\alpha+\gamma)}\,[(\alpha+\gamma)v(x,y) - \gamma v(x-\alpha h, y)]\\ &+\frac{a_1(x, y+\frac{1}{2}\delta h)}{\beta\delta(\beta+\delta)}\,[\beta v(x, y+\delta h) - (\beta+\delta)v(x,y)]\\ &-\frac{a_1(x, y-\frac{1}{2}\beta h)}{\beta\delta(\beta+\delta)}\,[(\beta+\delta)v(x,y) - \delta v(x, y-\beta h)]\Bigg\}\\ &+a_0(x,y)v(x,y)\end{aligned}$$

Finite Difference Scheme

- We now define the finite difference scheme approximating the solution to the Elliptic BVP with variable coefficients:

$$\begin{cases} L_h U(x,y) &= f(x,y), \quad \text{for } (x,y) \in \Omega_h \cup \Omega_h^\star \\ U(x,y) &= g(x,y), \quad \text{for } (x,y) \in \partial\Omega_h \end{cases} \tag{10}$$

- Let $d = \#(\Omega_h \cup \Omega_h^\star)$. The above problem (10) corresponds to a $d \times d$ system of linear equations $A_h U_h = F_h$, where $U_h \in \mathbb{R}^d$ contains values of U on $\Omega_h \cup \Omega_h^\star$, $A_h \in \mathbb{R}^{d \times d}$ is a matrix independent of f and g and $F_h \in \mathbb{R}^d$ depends upon grid values of f and g.
- Let $\mathcal{J}(A_h)$ consist of the indices for $p \in \Omega_h^\star$.
- Using the above matrix methods it can be shown that A_h is of positive type, hence monotone, hence invertible, and therefore:

Theorem: The finite difference scheme (10) has a unique solution.

- Convergence $U \to u$ for $h \to 0$ can be shown using methods similar to those shown above.

Non-Linear Elliptic BVP

- Suppose a membrane is fastened to a boundary $\partial\Omega$ and stretched over the interior of Ω with tension $T(x, y)$ (force per unit length). (figure forthcoming)
- Let $u(x, y)$ be the displacement of the membrane resulting from an externally applied force per unit area $f(x, y)$.
- The variational principle used to model the shape of the membrane involves to minimize the following energy with respect to u:

$$J(u) = \int_\Omega T(x, y)\sqrt{1 + |\nabla u(x, y)|^2}dxdy - \int_\Omega f(x, y)u(x, y)dxdy$$

- The first term involves an increase in surface area which represents the elastic energy available for work to oppose the work performed by the external load as represented in the second term.
- Using the approximation $\sqrt{1 + \epsilon^2} - 1 \approx \frac{1}{2}\epsilon^2$, the first integrand above may be approximated by $\frac{T}{2}|\nabla u|^2$ for $|\nabla u| \ll 1$. Then the minimizing u satisfies a Poisson equation ($u|_{\partial\Omega} = 0$).

Non-Linear Elliptic BVP

- Without assuming $|\nabla u| \ll 1$, let $v \in \mathcal{C}_0^\infty(\Omega)$ and

$$\frac{\delta J}{\delta u}(u; v) = \int_\Omega T \frac{\nabla u \cdot \nabla v}{\sqrt{1 + |\nabla u|^2}} - \int_\Omega f v =$$

$$- \int_\Omega v \left[\nabla \cdot \left(T \frac{\nabla u}{\sqrt{1 + |\nabla u|^2}} \right) + f \right] + \int_{\partial\Omega} v \boldsymbol{n} \cdot \left(T \frac{\nabla u}{\sqrt{1 + |\nabla u|^2}} \right)$$

- The necessary optimality condition for a minimizing u is

$$\begin{cases} -\nabla \cdot \left(\dfrac{T}{\sqrt{1 + |\nabla u|^2}} \nabla u \right) &= f, \quad \text{in } \Omega \\ u &= 0, \quad \text{on } \partial\Omega \end{cases} \tag{11}$$

- A common approach to solving such a problem is to use the method of *lagged diffusivity*, where $u_0 = 0$ and for $l \in \mathbb{N}$ the *non-linear coefficient* is replaced by $T/\sqrt{1 + |\nabla u_{l-1}|^2}$ while the displacement is otherwise replaced by u_l.
- For each l the linear PDE can be solved by treating $T/\sqrt{1 + |\nabla u_{l-1}|^2}$ as a variable coefficient.

Dirichlet Poisson BVP on a Square

- The solution to (1) is approximated by solving:

$$\begin{cases} -\Delta_h U(x,y) &= f(x,y), \quad \text{for } (x,y) \in \Omega_h \\ U(x,y) &= g(x,y), \quad \text{for } (x,y) \in \partial\Omega_h \end{cases} \tag{12}$$

  now for $\Omega = (0,1)^2$ and for $N \in \mathbb{N}$, $h = 1/(N+1)$, $\Omega_h^\star = \emptyset$ and $\bar{\Omega}_h = \{(x,y) : x = ih, y = jh, 0 \le i,j \le N+1\}$.
- Let $d = \#\Omega_h = N^2$. The above problem (12) corresponds to a $d \times d$ system of linear equations $A_h U_h = F_h$, where $U_h \in \mathbb{R}^d$ is a vector of values of U on Ω_h, $A_h \in \mathbb{R}^{d\times d}$ is a matrix independent of f and g and $F_h \in \mathbb{R}^d$ depends upon grid values of f and g.
- For simplicity let $g = 0$. **Exercise**: Generalize to $g \neq 0$.
- Let the unknowns be ordered according to the so-called lexicographic ordering: (figure forthcoming)

$$\begin{aligned} U_h = \{&U(x_1,y_1), U(x_2,y_1), \ldots, U(x_N,y_1), \\ &U(x_1,y_2), U(x_2,y_2), \ldots, U(x_N,y_2), \\ &\ldots, U(x_1,y_N), U(x_2,y_N), \ldots, U(x_N,y_N)\}^\top \end{aligned}$$

- The vector F_h is given by the values of f in Ω_h in the lexicographic ordering.

Dirichlet Poisson BVP on a Square

- Then the matrix A_h is given explicitly by

$$A_h = \frac{1}{h^2}\begin{bmatrix} A_1 & A_0 & & & \\ A_0 & A_1 & A_0 & & \\ & \ddots & \ddots & \ddots & \\ & & A_0 & A_1 & A_0 \\ & & & A_0 & A_1 \end{bmatrix}, \quad A_0 = -I, \quad A_1 = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{bmatrix}$$

- (`sparse`!) Matlab commands for solving (12) are

```
B1 = speye(N); h=1/(N+1);
B0 = spdiags(kron([1,1],ones(N,1)),[-1,1],N,N);
A0 = -B1;
A1 = spdiags(kron([-1,4,-1],ones(N,1)),[-1,0,+1],N,N);
Ah = (kron(B0,A0)+kron(B1,A1))/h^2;
Uh = zeros(N+2,N+2);
Uh(2:(N+1),2:(N+1)) = reshape(Ah \ Fh,N,N);
```

Exercise: Implement and estimate the convergence order.

Neumann Poisson BVP on a Square

- The solution to (7) is approximated by solving:

$$\begin{cases} -\Delta_h U(x,y) &= f(x,y), & \text{for } (x,y) \in \Omega_h \\ -\Delta_h U(x,y) + & f(x,y) + & \\ \partial_{n,h} U(x,y)/h &= g(x,y)/h, & \text{for } (x,y) \in \partial\Omega_h \\ \sum_{(x,y)\in\bar{\Omega}_h} U(x,y) &= 0 & \end{cases} \tag{13}$$

  now for $\Omega = (0,1)^2$ and for $N \in \mathbb{N}$, $h = 1/(N+1)$, $\Omega_h^\star = \emptyset$ and $\bar{\Omega}_h = \{(x,y) : x = ih, y = jh, 0 \le i,j \le N+1\}$.

- Also in contrast to (8), $\partial_{n,h}$ can be computed here more simply and then integrated into the discrete Laplacian as seen below.
- **Exercise**: For $(x_i, y_j) \in \partial\Omega_h$ with $x_i = 0$, and otherwise $x_i = ih$, the normal derivative can be approximated as

$$g(x_0,y_j) = \partial_n u(x_0,y_j) = -\partial_x u(x_0,y_j) = -[u(x_0,y_j) - u(x_{-1},y_j)]/h + \mathcal{O}(h)$$

- The normal derivative can also be approximated as

$$g(x_0,y_j) = \partial_n u(x_0,y_j) = -\partial_x u(x_0,y_j) = -[u(x_1,y_j) - u(x_{-1},y_j)]/(2h) + \mathcal{O}(h^2)$$

  which leads to a non-symmetric matrix A_h.

Neumann Poisson BVP on a Square

- Integrating the $\mathcal{O}(h)$ approximation to ∂_n into $\Delta_h U$ gives,

$$f(x_0, y_j) = -\Delta_h U(x_0, y_j) =$$
$$h^{-2}\{[U(x_0, y_j) - U(x_{-1}, y_j)] - [U(x_1, y_j) - U(x_0, y_j)]$$
$$+[U(x_0, y_{j+1}) - U(x_0, y_j)] - [U(x_0, y_j) - U(x_0, y_{j-1})]\}$$
$$= h^{-2}\{-hg(x_0, y_j) - [U(x_1, y_j) - U(x_0, y_j)]$$
$$+[U(x_0, y_{j+1}) - U(x_0, y_j)] - [U(x_0, y_j) - U(x_0, y_{j-1})]\}$$

  and similarly for other points on $\partial\Omega_h$.
- Let $d = \#\bar{\Omega}_h = (N+2)^2$. The above problem (13) corresponds to a $d \times d$ system of linear equations $A_h U_h = F_h$, where $U_h \in \mathbb{R}^d$ is a vector of values of U on $\bar{\Omega}_h$, $A_h \in \mathbb{R}^{d \times d}$ is a matrix independent of f and g and $F_h \in \mathbb{R}^d$ depends upon grid values of f and g.
- For simplicity let $g = 0$. **Exercise**: Generalize to $g \neq 0$.
- Let the unknowns U_h be ordered by the lexicographic ordering.
- The vector F_h is given by the values of f in $\bar{\Omega}_h$ in the lexicographic ordering.

Neumann Poisson BVP on a Square

- Then the matrix A_h is given explicitly by

$$A_h = \frac{1}{h^2}\begin{bmatrix} A_1 & A_0 & & & & & \\ A_0 & A_2 & A_0 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & & A_0 & A_2 & A_0 & \\ & & & & A_0 & A_1 & \end{bmatrix}$$

with $A_0 = -I$ and

$$A_1 = \begin{bmatrix} 2 & -1 & & & \\ -1 & 3 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 3 & -1 \\ & & & -1 & 2 \end{bmatrix} \quad A_2 = \begin{bmatrix} 3 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 3 \end{bmatrix}$$

Neumann Poisson BVP on a Square

- The additional condition in (13) that $\sum_{(x,y)\in\bar{\Omega}} U(x,y) = 0$ corresponds to selecting the solution to (7) with mean value zero. Such a solution (with $g = 0$) is stationary for the Lagrangian functional,

$$L(u) = \left[\int_\Omega \tfrac{1}{2}|\nabla u|^2 - \int_\Omega fu\right] + \lambda \int_\Omega u$$

  where λ is a Lagrange multiplier corresponding to the condition $\int_\Omega u = 0$.
- **Exercise**: The stationarity conditions for L are

$$-\Delta u + \lambda = f \text{ in } \Omega, \quad \partial_n u = 0 \text{ on } \partial\Omega, \quad \text{and} \int_\Omega u = 0.$$

- With $e = (1, 1, \ldots, 1)^\top \in \mathbb{R}^d$ the discrete counterpart to these stationarity conditions is

$$\begin{bmatrix} A_h & e \\ e^\top & 0 \end{bmatrix} \begin{bmatrix} U_h \\ \lambda \end{bmatrix} = \begin{bmatrix} F_h \\ 0 \end{bmatrix}$$

  where the last component of this system is seen as the condition $\sum_{(x,y)\in\bar{\Omega}} U(x,y) = 0$.

Neumann Poisson BVP on a Square

- (`sparse`!) Matlab commands for solving (13) are

```
N1=N+1; N2=N+2; N2N2 = N2*N2; h=1/N1;
B1=sparse(N2,N2); B1(1,1) = 1; B1(N2,N2) = 1;
B2=speye(N2,N2);  B2(1,1) = 0; B2(N2,N2) = 0;
B0=spdiags(kron([1,1],ones(N2,1)),[-1,1],N2,N2);
A0=-speye(N2);
A1=spdiags(kron([-1,3,-1],ones(N2,1)),[-1,0,+1],N2,N2);
   A1(1,1) = 2; A1(N2,N2) = 2;
A2=spdiags(kron([-1,4,-1],ones(N2,1)),[-1,0,+1],N2,N2);
   A2(1,1) = 3; A3(N2,N2) = 3;
Ah=kron(B0,A0)+kron(B1,A1)+kron(B2,A2);
Ah=Ah/h^2;
e =ones(N2N2,1); Uh=[Ah,e;e',0] \ [Fh;0];
Uh=reshape(Uh(1:N2N2),N2,N2);
```

Exercise: Implement and estimate the convergence order.
What happens if `e'*Fh =/= 0` ?

Non-Linear Elliptic BVP on a Square

- To solve the non-linear elliptic BVP (11) on $\Omega = (0,1)^2$ we set $w_0 = 0$ and for a given $l \in \mathbb{N}$ we seek an approximation to the solution w_l of the *now linear* elliptic BVP,

$$\begin{cases} -\nabla \cdot \left(\dfrac{T}{\sqrt{1+|\nabla w_{l-1}|^2}} \nabla w_l \right) &= f, \quad \text{in } \Omega \\ w_l &= 0, \quad \text{on } \partial\Omega \end{cases} \tag{14}$$

- Setting $u = w_l$ and

$$a_l(x,y) = \frac{T(x,y)}{\sqrt{1+|\nabla w_{l-1}(x,y)|^2}} \geq \min_{(x,y)\in\Omega} T(x,y) > 0$$

permits (14) to be formulated like (9), an elliptic BVP with variable coefficients,

$$\begin{cases} [L_l u](x,y) &= f(x,y), \quad \text{for } (x,y) \in \Omega \\ u(x,0) &= 0, \quad \text{for } (x,y) \in \partial\Omega \end{cases} \tag{15}$$

where

$$[L_l u](x,y) = -\nabla \cdot [a_l(x,y) \nabla u(x,y)]$$

Non-Linear Elliptic BVP on a Square

- To approximate (15) let $\bar{\Omega}_h$ and $L_{h,l}$ be defined as for (9) but now with $\Omega_h^\star = \emptyset$, $a_1 = a_l$ and $a_0 = 0$.
- The solution to (15) is approximated with the finite difference scheme

$$\begin{cases} [L_{h,l}U](x,y) &= f(x,y), & \text{for } (x,y) \in \Omega_h \\ U(x,y) &= 0, & \text{for } (x,y) \in \partial\Omega_h \end{cases} \tag{16}$$

  now for $\Omega = (0,1)^2$ and for $N \in \mathbb{N}$, $h = 1/(N+1)$, $\Omega_h^\star = \emptyset$ and
  $$\bar{\Omega}_h = \{(x,y) : x = ih, y = jh, 0 \le i,j \le N+1\}.$$
- Let $d = \#\Omega_h = N^2$. The above problem (16) corresponds to a $d \times d$ system of linear equations $A_{h,l}U_h = F_h$, where $U_h \in \mathbb{R}^d$ is a vector of values of U on Ω_h, $A_{h,l} \in \mathbb{R}^{d\times d}$ is a matrix independent of f and $F_h \in \mathbb{R}^d$ depends upon grid values of f.
- Let the unknowns of U_h and the f-values of F_h be ordered according to the lexicographic ordering.

Non-Linear Elliptic BVP on a Square

- The formulas 45 for $L_{h,l}$ require to evaluate a_l at midpoints $(x + \frac{1}{2}ih, y + \frac{1}{2}jh)$, $|i| + |j| = 1$, between grid points $(x, y) \in \Omega_h$,

$$a_l(x + \tfrac{1}{2}ih, y + \tfrac{1}{2}jh) = \frac{T(x + \frac{1}{2}ih, y + \frac{1}{2}jh)}{\sqrt{1 + |\nabla_h v(x + \frac{1}{2}ih, y + \frac{1}{2}jh)|^2}}, \quad v = w_{l-1}$$

- The gradient $\nabla_h \approx \nabla$ above is approximated compactly with the finite differences (figure forthcoming)

$$\nabla_{h,x} v(x + \tfrac{1}{2}ih, y) = \begin{bmatrix} [D_{h,x} v](x + \frac{1}{2}ih, y) \\ [C_{h,y} v](x + \frac{1}{2}ih, y) \end{bmatrix}$$

for $i = \pm 1$ and

$$\nabla_{h,y} v(x, y + \tfrac{1}{2}jh) = \begin{bmatrix} [C_{h,x} v](x, y + \frac{1}{2}jh) \\ [D_{h,y} v](x, y + \frac{1}{2}jh) \end{bmatrix}$$

for $j = \pm 1$ where (with values understood to be 0 at $\partial\Omega_h$)

$$[D_{h,x} v](x + \tfrac{1}{2}ih, y) = [v(x + ih, y) - v(x, y)]/h$$

$$[D_{h,y} v](x, y + \tfrac{1}{2}jh) = [v(x, y + jh) - v(x, y)]/h$$

Non-Linear Elliptic BVP on a Square

and (with values understood to be 0 at $\partial\Omega_h$)

$$[C_{h,x}v](x, y + \tfrac{1}{2}jh) = \frac{1}{4h}\begin{pmatrix} v(x+h,y) + v(x+h,y+jh) \\ -v(x-h,y) - v(x-h,y+jh) \end{pmatrix}$$

$$[C_{h,y}v](x + \tfrac{1}{2}ih, y) = \frac{1}{4h}\begin{pmatrix} v(x,y+h) + v(x+ih,y+h) \\ -v(x,y-h) - v(x+ih,y-h) \end{pmatrix}$$

- The matrix representations of these operators are

$$D_{h,x} = \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_1 \end{bmatrix}, \quad D_1 = \frac{1}{h}\begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \\ & & & -1 \end{bmatrix}$$

$$D_{h,y} = \frac{1}{h}\begin{bmatrix} D_0 & & & \\ -D_0 & D_0 & & \\ & \ddots & \ddots & \\ & & -D_0 & D_0 \\ & & & -D_0 \end{bmatrix}, \quad D_0 = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$$

Non-Linear Elliptic BVP on a Square

and

$$C_{h,x} = \frac{1}{2}\begin{bmatrix} C_1 & & & & \\ C_1 & C_1 & & & \\ & \ddots & \ddots & & \\ & & C_1 & C_1 & \\ & & & C_1 & \end{bmatrix} \quad C_1 = \frac{1}{2h}\begin{bmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -1 & 0 \end{bmatrix}$$

$$C_{h,y} = \frac{1}{2h}\begin{bmatrix} 0 & C_0 & & & \\ -C_0 & 0 & C_0 & & \\ & \ddots & \ddots & \ddots & \\ & & -C_0 & 0 & C_0 \\ & & & -C_0 & 0 \end{bmatrix} \quad C_0 = \frac{1}{2}\begin{bmatrix} 1 & & & & \\ 1 & 1 & & & \\ & \ddots & \ddots & & \\ & & & 1 & 1 \\ & & & & 1 \end{bmatrix}$$

- Set the tension values at midpoints,

$$T_{h,x} = \{T((i+\tfrac{1}{2})h, jh) : i = 0, \ldots, N, j = 1, \ldots, N\}$$
$$T_{h,y} = \{T(ih, (j+\tfrac{1}{2})h) : i = 1, \ldots, N, j = 0, \ldots, N\}$$

Non-Linear Elliptic BVP on a Square

- Let V_h denote the vector of lexicographically ordered values approximating w_{l-1} on Ω_h.
- Set the values of coefficients at midpoints,

$$a_{l,x} = T_{h,x} ./ \sqrt{1 + (D_{h,x} V_h).\hat{}2 + (C_{h,y} V_h).\hat{}2}$$
$$a_{l,y} = T_{h,y} ./ \sqrt{1 + (C_{h,x} V_h).\hat{}2 + (D_{h,y} V_h).\hat{}2}$$

- Let $\mathcal{D}(V)$ denote a diagonal matrix with the values of the vector V along the diagonal.
- Then the coefficient matrix for the system of linear equations $A_{h,l} U_h = F_h$ for the solution to (16) is given explicitly by

$$A_{h,l} = D_{h,x}^{\top} \mathcal{D}(a_{l,x}) D_{h,x} + D_{h,y}^{\top} \mathcal{D}(a_{l,y}) D_{h,y}$$

- For example, suppose the following data are given:

```
N1 = N+1; NN1 = N*N1; h=1/N1;
Fh = f0*ones(N,N); Fh = Fh(:); % f0 = ?
Tx = T0*ones(N1,N); Tx = Tx(:); % T0 = ?
Ty = T0*ones(N,N1); Ty = Ty(:);
```

Non-Linear Elliptic BVP on a Square

- (`sparse`!) Matlab commands for solving (16) are

```
D0 = speye(N);
D1 = spdiags(kron([-1,1],ones(N,1)),[-1,0],N1,N)/h;
Dx = kron(D0,D1); Dy = kron(D1,D0);
C0 = spdiags(kron([1,1],ones(N,1)),[-1,0],N1,N)/2;
C1 = spdiags(kron([-1,1],ones(N,1)),[-1,1],N,N)/(2*h);
Cx = kron(C0,C1); Cy = kron(C1,C0);
Uh = zeros(N,N); Uh=Uh(:);
% iterate until convergence
    Vh = Uh;
    ax = Tx./sqrt(1 + (Dx*Vh).^2 + (Cy*Vh).^2);
    ay = Ty./sqrt(1 + (Cx*Vh).^2 + (Dy*Vh).^2);
    Ah = Dx'*spdiags(ax,0,NN1,NN1)*Dx ...
       + Dy'*spdiags(ay,0,NN1,NN1)*Dy;
    Uh = Ah \ Fh;
    % if ||Uh-Vh|| small enough, break
```

Exercise: Implement and show convergence w.r.t. l for $f0 \leq \boxed{?} T0$.

Heat Equation

- Let $\Omega = (0,1)$ and $T > 0$ and set $Q = \Omega \times (0,T)$.
- For a given constant diffusivity $\alpha > 0$ we week an approximation of the solution u to the *Heat Equation with Dirichlet Boundary Conditions*,

$$\begin{cases} u_t(x,t) &= \alpha u_{xx}(x,t), & \text{for } (x,t) \in \Omega \times (0,T] \\ u(x,t) &= g(x,t), & \text{for } (x,t) \in \partial\Omega \times [0,T] \\ u(x,0) &= u_0(x), & \text{for } (x,t) \in \Omega \times \{0\} \end{cases} \tag{17}$$

- For simplicity let $g = 0$. **Exercise**: Generalize to $g \neq 0$.
- For compatibility it must hold that $u_0(x) = 0$, $x \in \partial\Omega$.
- Under this compatibility condition and with $u_0 \in \mathcal{C}(\bar{\Omega})$, (17) has a unique solution in $\mathcal{C}^{2,1}(Q)$, where

$$\|u\|_{\mathcal{C}^{m,n}(Q)} = \max_{1\le i\le m} \max_{1\le k\le n} \sup_{(x,t)\in Q} |\partial_x^i \partial_t^k u(x,t)|$$

and, according to a maximum principle,

$$\max_{(x,t)\in\bar{Q}} |u(x,t)| \le \max_{x\in\bar{\Omega}} |u_0(x)|$$

Explicit Finite Difference Scheme

- For the discretization of (17) let $N, K \in \mathbb{N}$ and set
$$h = 1/(N+1), \quad x_i = ih, \quad i = 0, \ldots, N+1$$
$$\tau = T/K, \quad t_k = k\tau, \quad k = 0, \ldots, K$$
and adopt the notation $u_i^k = u(x_i, t_k)$.
- A finite difference scheme for computing $U_i^k \approx u_i^k$ is given first by approximating the PDE according to
$$\frac{U_i^{k+1} - U_i^k}{\tau} = \alpha \frac{U_{i+1}^k - 2U_i^k + U_{i-1}^k}{h^2}$$
- By setting $\lambda = \alpha\tau/h^2$, the IBVP (17) is approximated by

$$\begin{cases} U_i^{k+1} = \lambda U_{i+1}^k + (1-2\lambda)U_i^k + \lambda U_{i-1}^k, & 1 \le i \le N, 1 \le k \le K-1 \\ U_0^k = U_{N+1}^k = 0, & 0 \le k \le K \\ U_i^0 = u_0(x_i), & 0 \le i \le N+1 \end{cases} \tag{18}$$

- Setting $U_h^k = \{U_i^k\}_{i=1}^N$ this *explicit* scheme can be written as
$$U_h^{k+1} - U_h^k = \alpha\tau A_h U_h^k, \quad A_h = \text{tridiag}\{-1, 2, -1\}/h^2$$
or
$$U_h^{k+1} = B_\lambda U_h^k, \quad B_\lambda = \text{tridiag}\{\lambda, 1-2\lambda, \lambda\}$$

Stability of Explicit Scheme

- To prove convergence $U \to u$ as $h, \tau \to 0$, *stability* is first established as follows.

Lemma: Let $\lambda \in (0, 1/2]$. Then the solution $\{U_i^k\}$ to (18) satisfies

$$\max_{0 \le i \le N+1, 0 \le k \le K} |U_i^k| \le \max_{0 \le i \le N+1} |u_0(x_i)|$$

Proof: By (18), for $1 \le i \le N$, it follows with $\lambda \in (0, 1/2]$ that

$$\begin{aligned} |U_i^{k+1}| &\le \lambda |U_{i+1}^k| + (1 - 2\lambda)|U_i^k| + \lambda |U_{i-1}^k| \\ &\le (\lambda + 1 - 2\lambda + \lambda) \max_{0 \le i \le N+1} |U_i^k| \end{aligned}$$

Since $U_0^{k+1} = U_{N+1}^{k+1} = 0$, the max over $0 \le i \le N+1$ can be taken on the left side of the estimate, and the claim follows. ■

- The restriction $\lambda \in (0, 1/2]$ is crucial here for stability.

Exercise: Construct an example for which the scheme (18) gives unbounded values $\{U_i^k\}$ for $\lambda > 1/2$ as $k, h \to 0$.

- The restriction $\lambda \in (0, 1/2]$ is severe here: $\tau = \mathcal{O}(h^2)$!

Consistency of Explicit Scheme

- To prove convergence $U \to u$ as $h, \tau \to 0$, *consistency* is next established as follows.

Lemma: Let $u \in \mathcal{C}^{4,2}(\bar{Q})$ be the solution to (17). Then for

$$b_\lambda u_i^k := \lambda u_{i+1}^k + (1-2\lambda)u_i^k + \lambda u_{i-1}^k$$

$\exists c \neq c(h,\tau)$ such that

$$\max_{1\le i\le N, 0\le k\le K} |u_i^{k+1} - b_\lambda u_i^n| \le c\tau(\tau + h^2)\|u\|_{\mathcal{C}^{4,0}(\bar{Q})}$$

Proof: Expanding each term of $b_\lambda u_i^k$ about (x_i, t_k) gives

$$b_\lambda u_i^k = u_i^k + \lambda h^2 u_{xx}(x_i,t_k) + \frac{\lambda h^4}{24}\left[\frac{\partial^4 u}{\partial x^4}(\theta_{i-\frac{1}{2}}, t_k) + \frac{\partial^4 u}{\partial x^4}(\theta_{i+\frac{1}{2}}, t_k)\right]$$

for constants $\theta_{i-\frac{1}{2}} \in [x_{i-1}, x_i]$ and $\theta_{i+\frac{1}{2}} \in [x_i, x_{i+1}]$. Expanding u_i^{k+1} about (x_i, t_k) gives

$$u_i^{k+1} = u_i^k + \tau u_t(x_i,t_k) + \frac{\tau^2}{2} u_{tt}(x_i, \theta_{k+\frac{1}{2}})$$

for a constant $\theta_{k+\frac{1}{2}} \in [t_k, t_{k+1}]$. Combining these estimates and

Convergence of Explicit Scheme

recalling $\lambda h^2 = \alpha\tau$ and $u_t = \alpha u_{xx}$ gives

$$u_i^{k+1} - b_\lambda u_i^k = \frac{\tau^2}{2} u_{tt}(x_i, \theta_{k+\frac{1}{2}}) - \frac{\lambda h^4}{24}\left[\frac{\partial^4 u}{\partial x^4}(\theta_{i-\frac{1}{2}}, t_k) + \frac{\partial^4 u}{\partial x^4}(\theta_{i+\frac{1}{2}}, t_k)\right]$$

Setting $u_{tt} = \alpha u_{xxt} = \alpha(u_t)_{xx} = \alpha^2 u_{xxxx}$ proves the claim. ■

- Note that the stability restriction $\lambda \in (0, 1/2]$ did not enter into the consistency proof.
- Also consistency is not equivalent to *convergence*, which is proved as follows by *combining consistency with stability*. (cf. Lax Equivalence Theorem.)

Theorem: Let $u \in \mathcal{C}^{4,2}(\bar{Q})$ be the solution to (17). For $\lambda \in (0, 1/2]$ let $\{U_i^k\}$ be the solution to (18). Then $\exists c \neq c(h, \tau)$ such that

$$\max_{0 \le i \le N+1, 0 \le k \le K} |U_i^k - u(x_i, t_k)| \le c(\tau + h^2)\|u\|_{\mathcal{C}^{4,0}(\bar{Q})}$$

Convergence of Explicit Scheme

Proof: Let $e_i^k = U_i^k - u_i^k$, which satisfies $e_0^k = e_{N+1}^0 = 0$ for $k = 0, \ldots, K$ and $e_i^0 = 0$ for $0 \le i \le N+1$. For $1 \le i \le N$ and $0 \le k \le K-1$,

$$\begin{aligned} e_i^{k+1} &= U_i^{k+1} - u_i^{k+1} = b_\lambda U_i^k - u_i^{k+1} \pm b_\lambda u_i^k \\ &= b_\lambda e_i^k + (b_\lambda u_i^k - u_i^{k+1}) \end{aligned}$$

Since $\lambda \in (0, 1/2]$, e_i^{k+1} depends stably upon e_i^k,

$$|e_i^{k+1}| \le \lambda |e_{i+1}^k| + (1-2\lambda)|e_i^k| + \lambda |e_{i-1}^k| + |b_\lambda u_i^k - u_i^{k+1}|$$

and using the consistency estimate,

$$\max_{1 \le i \le N} |e_i^{k+1}| \le \max_{1 \le i \le N} |e_i^k| + \tilde{c}\tau(\tau + h^2)\|u\|_{\mathcal{C}^{4,0}(\bar{Q})}$$

Summing this estimate over $k = 0, \ldots, \kappa - 1$, $\kappa \le K$, gives

$$\max_{1 \le i \le N} |e_i^\kappa| \le \max_{1 \le i \le N} |e_i^0| + \tilde{c}\kappa\tau(\tau + h^2)\|u\|_{\mathcal{C}^{4,0}(\bar{Q})}$$

Setting $c = \tilde{c}T (\ge \tilde{c}\kappa\tau)$ and recalling where e_i^k vanishes gives the claimed convergence estimate. ■

Neumann Heat Equation on a Square

- Let $\Omega = (0,1)^2$ and $T > 0$.
- For a given constant diffusivity $\alpha > 0$ we week an approximation of the solution u to the *Heat Equation with Neumann Boundary Condition*,

$$\begin{cases} u_t(x,y,t) &= \alpha \Delta u(x,y,t), & \text{for } (x,y,t) \in \Omega \times (0,T] \\ \partial_n u(x,y,t) &= g(x,y,t), & \text{for } (x,y,t) \in \partial\Omega \times [0,T] \\ u(x,y,0) &= u_0(x,y), & \text{for } (x,y,t) \in \Omega \times \{0\} \end{cases} \tag{19}$$

- For simplicity let $g = 0$. **Exercise**: Generalize to $g \neq 0$.
- For compatibility it must hold that $\partial_n u_0(x,y) = 0$, $x \in \partial\Omega$.
- Let the Laplacian Δ be approximated with Neumann boundary conditions as was done for the Poisson equation on a square in (13).
- With $\bar{\Omega}_h$ and A_h defined as for 53, (19) can be semi-discretized spatially by the system of ODEs (method of lines),

$$U_h'(t) = -\alpha A_h U_h(t), \quad U_h(0) = U_0 = \{u_0(x_i, y_j)\}, \quad U_h(t) = \{U_{i,j}(t)\}$$

Semi-Discrete Solution

- This semi-discrete solution $U_h(t)$ has the property
$$\tfrac{1}{2} D_t \|U_h(t)\|^2 = U_h(t) \cdot U_h'(t) = -\alpha U_h(t) \cdot [A_h U_h(t)] \leq 0$$
which implies $\|U_h(t)\| \leq \|U_0\|$, analogous to the estimate for the solution u to (19),
$$\tfrac{1}{2} D_t \int_\Omega u^2 = \int_\Omega u u_t = \int_\Omega \alpha u \Delta u = \int_{\partial\Omega} \alpha u \partial_n u - \int_\Omega \alpha |\nabla u|^2 \leq 0$$
- Also $U_h(t)$ satisfies
$$e \cdot U_h'(t) = -\alpha e \cdot [A_h U_h(t)] = -\alpha [A_h e] \cdot U_h(t) = 0$$
which implies $e \cdot U_h'(t) = e \cdot U_0$, analogous to the estimate for the solution u to (19),
$$\tfrac{1}{2} D_t \int_\Omega u = \int_\Omega u_t = \int_\Omega \alpha \Delta u = \int_{\partial\Omega} \alpha \partial_n u = 0$$
- **Exercise**: The fully-discrete explicit Euler scheme,
$$U_h^{k+1} - U_h^k = \tau \alpha A_h U_h^k, \quad U_h^k = \{U_{i,j}^k\}, \quad U_{i,j}^k \approx u(x_i, y_j, t_k)$$
satisfies the property $e \cdot U_h^k = e \cdot U_h^0$ but it satisfies $\|U_h^k\| \leq \|U_h^0\|$ only for $\lambda = \alpha\tau/h^2 \leq 1/4$.

Explicit and Implicit Euler Schemes

- **Exercise**: The implicit Euler scheme

$$U_h^{k+1} - U_h^k = -\tau\alpha A_h U_h^{k+1}, \quad [I + \alpha\tau A_h] U_h^{k+1} = U_h^k$$

satisfies the condition

$$e \cdot U_h^{k+1} = e \cdot [I + \alpha\tau A_h] U_h^{k+1} = e \cdot U_h^k$$

or $e \cdot U_h^k = e \cdot U_0$, and furthermore

$$\|U_h^{k+1}\|^2 \leq U_h^{k+1} \cdot [I + \alpha\tau A_h] U_h^{k+1} = U_h^{k+1} \cdot U_h^k \leq \|U_h^{k+1}\| \|U_h^k\|$$

or, analogous to 65, $\|U_h^k\| \leq \|U_h^0\|$ holds but now without conditions on τ or h.

- **Exercise**: Analogous to 66, show that the implicit Euler scheme is consistent to $\mathcal{O}(\tau(\tau + h^2))$.
- **Exercise**: Analogous to 67, show that the implicit Euler scheme is convergent with convergence rate $\mathcal{O}(\tau + h^2)$.
- (`sparse`!) Matlab commands for solving (19) (`Ah` from 55, `Uh = U0` given) are

```
for k=1:K   Uh = (speye(N2N2)+al*ta*Ah) \ Uh;   end
```

Non-Linear Parabolic IBVP

- Let $f : \Omega \to [0, 1]$ be a measured image defined on $\Omega = (0, 1)^2$ which is to be denoised.
- The steepest descent evolution,

$$\int_\Omega u_t v = -\frac{\delta J}{\delta u}(u; v), \quad u|_{t=0} = f, \quad t \in [0, T], \quad v \in C^\infty(\bar{\Omega})$$

produces a sequence of images u which start at f and become progressively less noisy as time advances.
- The descent direction reduces a regularizer such as

$$J(u) = \alpha \int_\Omega \sqrt{\epsilon + |\nabla u|^2}$$

Here, $J(u) \to \alpha \mathrm{TV}(u)$ for $\epsilon \to 0$ and u sufficiently smooth.
- For the explicit form of the evolution, let $v \in \mathcal{C}^\infty(\bar{\Omega})$ and

$$0 = \int_\Omega u_t v + \frac{\delta J}{\delta u}(u; v) = \int_\Omega \left[v u_t + \alpha \frac{\nabla u \cdot \nabla v}{\sqrt{\epsilon + |\nabla u|^2}} \right] =$$

$$\int_\Omega v \left[u_t - \nabla \cdot \left(\alpha \frac{\nabla u}{\sqrt{\epsilon + |\nabla u|^2}} \right) \right] + \int_{\partial\Omega} v \boldsymbol{n} \cdot \left(\alpha \frac{\nabla u}{\sqrt{\epsilon + |\nabla u|^2}} \right)$$

Non-Linear Anisotropic Diffusion

- The steepest descent evolution is given explicitly by the resulting non-linear anisotropic diffusion equation,

$$\begin{cases} u_t &= \alpha\nabla\cdot\left(\dfrac{\nabla u}{\sqrt{\epsilon+|\nabla u|^2}}\right), & \text{in } \Omega\times(0,T] \\ \partial_n u &= 0, & \text{on } \partial\Omega\times[0,T] \\ u &= f, & \text{in } \Omega\times\{0\} \end{cases} \tag{20}$$

- To approximate the solution, (20) is first discretized temporally with a semi-implicit Euler scheme to obtain

$$\begin{cases} L_k u^k &= u^{k-1}, & \text{in } \Omega \\ \partial_n u^k &= 0, & \text{on } \partial\Omega \end{cases} \qquad \begin{matrix} k=1,\ldots,K \\ u^0 = f \end{matrix} \tag{21}$$

where $T = K\tau$ and

$$L_k u = u - \tau\nabla\cdot(a_k\nabla u), \quad a_k = \frac{\alpha}{\sqrt{\epsilon+|\nabla u^{k-1}|^2}}$$

- To approximate the solution to (21) spatially, let $\bar{\Omega}_h$ and $L_{h,k}$ be defined as for (9) but now with $\Omega_h^\star = \emptyset$, $a_1 = a_k$ and $a_0 = 1$.

Fully Discrete Approximation

- The solution to (21) is approximated spatially with the finite difference scheme

$$\begin{cases} L_{h,k}U^k = U^{k-1}, & \text{in } \Omega_h \\ L_{h,k}U^k + a_k \partial_{n,h} U^k / h = U^{k-1}, & \text{in } \partial\Omega_h \end{cases} \tag{22}$$

  now for $\Omega = (0,1)^2$ and for $N \in \mathbb{N}$, $h = 1/(N+1)$, $\Omega_h^\star = \emptyset$ and $\bar{\Omega}_h = \{(x,y) : x = ih, y = jh, 0 \le i,j \le N+1\}$.

- Let $d = \#\bar{\Omega}_h = (N+2)^2$. The above problem (22) corresponds to a $d \times d$ system of linear equations $A_{h,k}U_h^k = U_h^{k-1}$, where $U_h^k \in \mathbb{R}^d$ is a vector of values of U^k on $\bar{\Omega}_h$ and $A_{h,k} \in \mathbb{R}^{d\times d}$ is a matrix depending upon a_k and hence U^{k-1}.
- Let the unknowns of U_h^k be ordered according to the lexicographic ordering.
- The formulas 45 for $L_{h,k}$ require to evaluate a_k at midpoints $(x + \frac{1}{2}ih, y + \frac{1}{2}jh)$, $|i| + |j| = 1$, between grid points $(x,y) \in \bar{\Omega}_h$,

Approximation of Non-Linear Coefficients

$$a_k(x + \tfrac{1}{2}ih, y + \tfrac{1}{2}jh) = \frac{\alpha}{\sqrt{1 + |\nabla_h U^{k-1}(x + \tfrac{1}{2}ih, y + \tfrac{1}{2}jh)|^2}}$$

- The gradient $\nabla_h \approx \nabla$ above is approximated compactly with the finite differences (figure forthcoming)

$$\nabla_{h,x} v(x + \tfrac{1}{2}ih, y) = \begin{bmatrix} [D_{h,x} v](x + \frac{1}{2}ih, y) \\ [C_{h,y} v](x + \frac{1}{2}ih, y) \end{bmatrix}$$

for $i = \pm 1$ and

$$\nabla_{h,y} v(x, y + \tfrac{1}{2}jh) = \begin{bmatrix} [C_{h,x} v](x, y + \frac{1}{2}jh) \\ [D_{h,y} v](x, y + \frac{1}{2}jh) \end{bmatrix}$$

for $j = \pm 1$ where

$$[D_{h,x} v](x + \tfrac{1}{2}ih, y) = [v(x + ih, y) - v(x, y)]/h$$
$$[D_{h,y} v](x, y + \tfrac{1}{2}jh) = [v(x, y + jh) - v(x, y)]/h$$

Approximation of the Gradient

and (with difference quotients $[v(z+h)-v(z-h)]/(2h)$
replaced by $[v(z+h)-v(z)]/h$ for $z-h<0$
or by $[v(z)-v(z-h)]/h$ for $z+h>1$)

$$[C_{h,x}v](x, y+\tfrac{1}{2}jh) = \frac{1}{4h}\begin{pmatrix} v(x+h,y)+v(x+h,y+jh) \\ -v(x-h,y)-v(x-h,y+jh) \end{pmatrix}$$

$$[C_{h,y}v](x+\tfrac{1}{2}ih, y) = \frac{1}{4h}\begin{pmatrix} v(x,y+h)+v(x+ih,y+h) \\ -v(x,y-h)-v(x+ih,y-h) \end{pmatrix}$$

- The matrix representations of these operators are

$$D_{h,x} = \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_1 \end{bmatrix}, \quad D_1 = \frac{1}{h}\begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}$$

$$D_{h,y} = \frac{1}{h}\begin{bmatrix} -D_0 & D_0 & & \\ & \ddots & \ddots & \\ & & -D_0 & D_0 \end{bmatrix}, \quad D_0 = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$$

Approximation of the Gradient

and

$$C_{h,x} = \frac{1}{2}\begin{bmatrix} C_1 & C_1 & & \\ & \ddots & \ddots & \\ & & C_1 & C_1 \end{bmatrix} \quad C_1 = \frac{1}{2h}\begin{bmatrix} -2 & 2 & & & \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ & & & -2 & 2 \end{bmatrix}$$

$$C_{h,y} = \frac{1}{2h}\begin{bmatrix} -2C_0 & 2C_0 & & & \\ -C_0 & 0 & C_0 & & \\ & \ddots & \ddots & \ddots & \\ & & -C_0 & 0 & C_0 \\ & & & -2C_0 & 2C_0 \end{bmatrix} \quad C_0 = \frac{1}{2}\begin{bmatrix} 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{bmatrix}$$

Exercise: Derive these explicit formulas for $D_{h,x}, D_{h,y}, C_{h,x}, C_{h,y}$ as well as the counterparts for (14) and highlight the differences resulting from the respective boundary conditions.

Approximation of Non-Linear Operator

- Set the values of coefficients at midpoints,

$$a_{k,x} = \alpha./\sqrt{\epsilon + (D_{h,x}U_h^{k-1}).\hat{}2 + (C_{h,y}U_h^{k-1}).\hat{}2}$$
$$a_{k,y} = \alpha./\sqrt{\epsilon + (C_{h,x}U_h^{k-1}).\hat{}2 + (D_{h,y}U_h^{k-1}).\hat{}2}$$

- Let $\mathcal{D}(V)$ denote a diagonal matrix with the values of the vector V along the diagonal.
- With $[D_{h,x}^\top, D_{h,y}^\top] = \nabla_h^\top \approx (\nabla)^* = -\nabla\cdot$, the coefficient matrix $A_{h,k}$ in $A_{h,k}U_h^k = U_h^{k-1}$ for the solution to (22) is given explicitly by

$$A_{h,k} = I + \tau\left[D_{h,x}^\top \mathcal{D}(a_{k,x})D_{h,x} + D_{h,y}^\top \mathcal{D}(a_{k,y})D_{h,y}\right]$$

- For example, suppose the following data are given:

```
N1 = N+1; h=1/N1; ta = T/M;
N2 = N+2; N1N2 = N1*N2; N2N2 = N2*N2; N4 = N2/4;
Fh = [zeros(N4,1);ones(N2-2*N4,1);zeros(N4,1)];
Fh = reshape(Fh*Fh' + 0.5*randn(N2,N2),N2N2,1);
```

Code to Solve the Non-Linear IBVP

- (`sparse`!) Matlab commands for solving (22) are

```
D0 = speye(N2);
D1 = spdiags(kron([-1,1],ones(N2,1)),[0,1],N1,N2)/h;
Dx = kron(D0,D1); Dy = kron(D1,D0);
C0 = spdiags(kron([1,1],ones(N2,1)),[0,1],N1,N2)/2;
C1 = spdiags(kron([-1,1],ones(N2,1)),[-1,1],N2,N2)/(2*h);
C1(1,1) = -1/h; C1(1,2) = 1/h;
C1(N2,N2-1) = -1/h; C1(N2,N2) = 1/h;
Cx = kron(C0,C1); Cy = kron(C1,C0);
Uh = Fh; I = speye(N2N2);
for k=1:K
    ax = al./sqrt(ep + (Dx*Uh).^2 + (Cy*Uh).^2);
    ay = al./sqrt(ep + (Cx*Uh).^2 + (Dy*Uh).^2);
    Ah = I + ta*(Dx'*spdiags(ax,0,N1N2,N1N2)*Dx ...
                 + Dy'*spdiags(ay,0,N1N2,N1N2)*Dy);
    Uh = Ah \ Uh;
end
```

Exercise: Implement and estimate best ϵ, α and T.

Wave Equation

- Let $\Omega = (0,1)$ and $T > 0$ and set $Q = \Omega \times (0,T)$.
- For a given constant wave speed $\omega > 0$ we week an approximation of the solution u to the *Wave Equation with Dirichlet Boundary Conditions*,

$$\begin{cases} u_{tt}(x,t) &= \omega^2 u_{xx}(x,t), & \text{for } (x,t) \in \Omega \times (0,T] \\ u(x,t) &= g(x,t), & \text{for } (x,t) \in \partial\Omega \times [0,T] \\ u(x,0) &= u_0(x), & \text{for } (x,t) \in \Omega \times \{0\} \\ u_t(x,0) &= u_1(x), & \text{for } (x,t) \in \Omega \times \{0\} \end{cases} \tag{23}$$

- For simplicity let $g = 0$. **Exercise**: Generalize to $g \neq 0$.
- For compatibility it must hold that $u_0(x) = u_1(x) = 0$, $x \in \partial\Omega$.
- If the initial data u_0 and u_1 have compact support in Ω, then, for t sufficiently small, waves do not reach $\partial\Omega$, and the solution is given by d'Alembert's formula,

$$u(x,t) = \frac{1}{2}\left[u_0(x-\omega t) + u_0(x+\omega t)\right] + \frac{1}{2\omega}\int_{x-\omega t}^{x+\omega t} u_1(s)ds$$

which shows how u inherits its regularity from that of u_0 and u_1.

Explicit Finite Difference Scheme

- For the discretization of (23) let $N, K \in \mathbb{N}$ and set

$$h = 1/(N+1), \quad x_i = ih, \quad i = 0, \ldots, N+1$$
$$\tau = T/K, \quad t_k = k\tau, \quad k = 0, \ldots, K$$

  and adopt the notation $u_i^k = u(x_i, t_k)$.
- A finite difference scheme for computing $U_i^k \approx u_i^k$ is given first by approximating the PDE according to

$$\frac{U_i^{k+1} - 2U_i^k + U_{i-1}^{k-1}}{\tau^2} = \omega^2 \frac{U_{i+1}^k - 2U_i^k + U_{i-1}^k}{h^2}, \qquad \begin{matrix} 1 \le i \le N \\ k = 1, \ldots, K-1 \end{matrix}$$

  with $U_0^k = U_{N+1}^k = 0$, $k = 0, \ldots, K$, but we need U_j^0 and U_j^1.
- The initial conditions begin naturally with

$$U_i^0 = u_0(x_i), \quad 0 \le i \le N+1.$$

- For U_i^1 note that $U_i^1 \approx u(x_i, t_1)$, which can be expanded as

$$u(x_i, t_1) = u(x_i, 0) + \tau u_t(x_i, 0) + \tfrac{1}{2}\tau^2 u_{tt}(x_i, 0) + \tfrac{1}{6}\tau^3 u_{ttt}(x_i, \theta_{\frac{1}{2}})$$

  for $\theta_{\frac{1}{2}} \in [0, \tau]$.

Initial Conditions of Explicit Scheme

- Using the initial values $u(x,0) = u_0(x)$, $u_t(x,0) = u_1(x)$ and $u_{tt} = \omega^2 u_{xx}$ the expansion gives

$$u(x_i, t_1) = u_0(x_i) + \tau u_1(x_i) + \tfrac{1}{2}\tau^2\omega^2 u_0''(x_i) + \mathcal{O}(\tau^3)$$

- Thus for $1 \leq i \leq N$ we set

$$U_i^1 = u_0(x_i) + \tau u_1(x_i) + \frac{\omega^2\tau^2}{2h^2}\left[u_0(x_{i+1}) - 2u_0(x_i) + u_0(x_{i-1})\right]$$

which has the following order of consistency.

Lemma: For $u_0 \in \mathcal{C}^4(\bar{\Omega})$ and $u \in \mathcal{C}^{0,3}(Q)$, $\exists c \neq c(h,\tau)$ such that

$$\max_{0\leq i\leq N+1} |U_i^1 - u_i^1| \leq c\left[\tau^2 h^2 \|u_0\|_{\mathcal{C}^4(\bar{\Omega})} + \tau^3 \|u\|_{\mathcal{C}^{0,3}(\bar{Q})}\right]$$

Proof: **Exercise** with Taylor expansions. ■

Def: Let $r = \omega\tau/h$ be called the *Courant Number*.

- The finite difference scheme for approximating the solution u to (23) is now given by the explicit method

Consistency of Explicit Scheme

$$\begin{cases} U_i^{k+1} &= r^2(U_{i+1}^k + U_{i-1}^k) + 2(1-r^2)U_i^k - U_i^{k-1} \\ & \qquad 1 \le k \le K-1, 1 \le i \le N \\ U_0^k &= U_{N+1}^k = 0, \qquad 0 \le k \le K \\ U_i^0 &= u_0(x_i), \qquad 0 \le i \le N+1 \\ U_i^1 &= u_0(x_i) + \tau u_1(x_i) + \\ & \quad (r^2/2)\left[u_0(x_{i+1}) - 2u_0(x_i) + u_0(x_{i-1})\right], \\ & \qquad 1 \le i \le N \end{cases} \tag{24}$$

- This scheme is consistent according to the following.

Lemma: Let $u \in \mathcal{C}^{4,2}(\bar{Q})$ be the solution to (23). Then $\exists c \neq c(h,\tau)$ such that

$$\max_{1\le i\le N, 0\le k\le K} |u_i^{k+1} - [r^2(u_{i+1}^k + u_{i-1}^k) + 2(1-r^2)u_i^k - u_i^{k-1}]| \le c\tau^2(\tau^2+h^2)\|u\|_{\mathcal{C}^{4,0}(\bar{Q})}$$

Proof: **Exercise** with Taylor expansions. ■

Forward, Backward, Centered Differences

- Now some standard notation:

$$\begin{aligned} \Delta v_i &= v_{i+1} - v_i, && \text{forward difference} \\ \nabla v_i &= v_i - v_{i-1}, && \text{backward difference} \\ \delta v_i &= v_{i+\frac{1}{2}} - v_{i-\frac{1}{2}}, && \text{central difference} \end{aligned}$$

  and it holds that

$$\Delta\nabla v_i = v_{i+1} - 2v_i + v_{i-1} = \delta\delta v_i.$$

- On the subspace $V_0 = \{\{v_i\}_{i=0}^{N+1} : v_0 = v_{N+1} = 0\} \subset \mathbb{R}^{N+2}$ define the bilinear form

$$a_h(v, w) = -h \sum_{i=1}^{N} (\Delta\nabla v_i) w_i \tag{25}$$

Lemma: It holds that

a. $\forall v, w \in V_0$,

$$a_h(v, w) = a_h(w, v) = h \sum_{i=1}^{N+1} (\nabla v_i)(\nabla w_i)$$

b. $\forall v \in V_0$,

$$a_h(v, v) \geq 0 \quad \text{and} \quad a_h(v, v) = 0 \quad \Rightarrow \quad v = 0.$$

c. By (1) and (2), $a_h(v, w)$ is an inner product on V_0 giving a norm $a_h(v, v)^{\frac{1}{2}}$ so that $\forall v, w \in V_0$,

Bilinear Form

$$|a_h(v,w)| \le a_h(v,v)^{\frac{1}{2}} a_h(w,w)^{\frac{1}{2}}$$

d. $\forall v \in V_0$, $a_h(v,v)^{\frac{1}{2}} \le 2\|v\|_h, \quad \|v\|_h^2 = h\sum_{i=1}^{N} v_i^2$

Proof: **Exercise** ■

Theorem: Let $u \in \mathcal{C}^{4,3}(\bar{Q})$ be the solution to (23). Let $\{U_i^k\}$ be the solution to (24). Then for each $r_0 \in (0,1)$, $\exists c = c(r_0, T)$ such that for $r = \omega\tau/h$, $0 < r \le r_0$,

$$\max_{0\le k\le K} \|U_h^k - u^k\|_h \le c(h^2+\tau^2)\|u\|_{\mathcal{C}^{4,3}(\bar{Q})}$$

Proof: Set $e_i^k = U_i^k - u_i^k$, $0 \le i \le N+1$, and note that $e_0^k = e_{N+1}^k = 0$. Set also $e_h^k = \{e_i^k\}_{i=0}^{N+1}$ and note that $e_h^k \in V_0$. Then for $1 \le i \le N$, $1 \le k \le K-1$,

$$e_i^{k+1} - 2e_i^k + e_i^{k-1} = (U_i^{k+1} - 2U_i^k + U_i^{k-1}) - (u_i^{k+1} - 2u_i^k + u_i^{k-1})$$

$$= r^2(U_{i+1}^k - 2U_i^k + U_{i-1}^k) \pm r^2(u_{i+1}^k - 2u_i^k + u_{i-1}^k) - (u_i^{k+1} - 2u_i^k + u_i^{k-1})$$

Convergence of Explicit Scheme

or

$$e_i^{k+1} - 2e_i^k + e_i^{k-1} = r^2(e_{i+1}^k - 2e_i^k + e_{i-1}^k) + \varepsilon_i^k$$

where

$$\varepsilon_i^k = r^2(u_{i+1}^k - 2u_i^k + u_{i-1}^k) - (u_i^{k+1} - 2u_i^k + u_i^{k-1})$$

Multiplying the last e-equation by $h(e_i^{k+1} - e_i^{k-1})$ and summing over $i = 1, \ldots, N$ gives

$$h\sum_{i=1}^{N}(e_i^{k+1} - 2e_i^k + e_i^{k-1})(e_i^{k+1} - e_i^{k-1}) =: \mathrm{L} = \mathrm{R}_1 + \mathrm{R}_2 :=$$

$$r^2 h\sum_{i=1}^{N}(e_{i+1}^k - 2e_i^k + e_{i-1}^k)(e_i^{k+1} - e_i^{k-1}) + h\sum_{i=1}^{N}\varepsilon_i^k(e_i^{k+1} - e_i^{k-1})$$

Then

$$\mathrm{L} = h\sum_{i=1}^{N}[(e_i^{k+1} - e_i^k) - (e_i^k - e_i^{k-1})][(e_i^{k+1} - e_i^k) + (e_i^k - e_i^{k-1})] =$$

$$h\sum_{i=1}^{N}(e_i^{k+1} - e_i^k)^2 - h\sum_{i=1}^{N}(e_i^k - e_i^{k-1})^2 = \|e_h^{k+1} - e_h^k\|_h^2 - \|e_h^k - e_h^{k-1}\|_h^2$$

Convergence of Explicit Scheme

and with (25) and property (a) in Lemma 84,

$$\mathrm{R}_1 = -r^2 a_h(e_h^k, e_h^{k+1} - e_h^{k-1}) = -r^2 a_h(e_h^k, e_h^{k+1}) + r^2 a_h(e_h^k, e_h^{k-1}) =$$
$$-\tfrac{1}{2} r^2 \left[a_h(e_h^{k+1}, e_h^{k+1}) - a_h(e_h^{k-1}, e_h^{k-1}) \right]$$
$$+\tfrac{1}{2} r^2 \left[a_h(e_h^{k+1} - e_h^k, e_h^{k+1} - e_h^k) - a_h(e_h^k - e_h^{k-1}, e_h^k - e_h^{k-1}) \right]$$

Using these calculations to rewrite $\mathrm{L} = \mathrm{R}_1 + \mathrm{R}_2$ gives

$$\|e_h^{k+1} - e_h^k\|_h^2 - \|e_h^k - e_h^{k-1}\|_h^2$$
$$+\tfrac{1}{2} r^2 \left[a_h(e_h^{k+1}, e_h^{k+1}) \pm a_h(e_h^k, e_h^k) - a_h(e_h^{k-1}, e_h^{k-1}) \right]$$
$$-\tfrac{1}{2} r^2 \left[a_h(e_h^{k+1} - e_h^k, e_h^{k+1} - e_h^k) - a_h(e_h^k - e_h^{k-1}, e_h^k - e_h^{k-1}) \right]$$
$$= h \sum_{i=1}^{N} \varepsilon_i^k (e_i^{k+1} - e_i^{k-1})$$

Summing over $k = 1, \ldots, \kappa - 1$, $\kappa \leq K$, and observing $e_h^0 = 0$

Convergence of Explicit Scheme

gives

$$\left[\|e_h^\kappa - e_h^{\kappa-1}\|_h^2 - \|e_h^1\|_h^2 \pm \tfrac{1}{2}r^2 a_h(e_h^1, e_h^1)\right]$$

$$+\tfrac{1}{2}r^2\left[a_h(e_h^\kappa, e_h^\kappa) + a_h(e_h^{\kappa-1}, e_h^{\kappa-1}) - a_h(e_h^\kappa - e_h^{\kappa-1}, e_h^\kappa - e_h^{\kappa-1})\right]$$

$$=: \mathrm{L}_1 + \mathrm{L}_2 = \mathrm{R} := \sum_{k=1}^{\kappa-1} h \sum_{i=1}^{N} \varepsilon_i^k (e_i^{k+1} - e_i^{k-1})$$

Then,

$$\mathrm{R} \le \sum_{k=1}^{\kappa-1} \|\varepsilon^k\|_h \|e_h^{k+1} - e_h^{k-1}\|_h$$

Also,

$$\mathrm{L}_2 = r^2 a_h(e_h^{\kappa-1}, e_h^\kappa) = r^2\left[a_h(e_h^{\kappa-1}, e_h^\kappa) \pm a_h(e_h^{\kappa-1}, e_h^{\kappa-1})\right]$$
$$= r^2 a_h(e_h^{\kappa-1}, e_h^{\kappa-1}) + r^2 a_h(e_h^{\kappa-1}, e_h^\kappa - e_h^{\kappa-1})$$

Using properties (c) and (d) of Lemma 84 gives,

$$|a_h(e_h^{\kappa-1}, e_h^\kappa - e_h^{\kappa-1})| \le [a_h(e_h^{\kappa-1}, e_h^{\kappa-1})]^{\frac{1}{2}} [a_h(e_h^\kappa - e_h^{\kappa-1}, e_h^\kappa - e_h^{\kappa-1})]^{\frac{1}{2}}$$
$$\le [a_h(e_h^{\kappa-1}, e_h^{\kappa-1})]^{\frac{1}{2}} 2\|e_h^\kappa - e_h^{\kappa-1}\|_h$$

Convergence of Explicit Scheme

Using $2ab \leq a^2 + b^2$, the last inequality becomes

$$|a_h(e_h^{\kappa-1}, e_h^{\kappa} - e_h^{\kappa-1})| \leq a_h(e_h^{\kappa-1}, e_h^{\kappa-1}) + \|e_h^{\kappa} - e_h^{\kappa-1}\|_h^2$$

and hence

$$r^2 a_h(e_h^{\kappa-1}, e_h^{\kappa} - e_h^{\kappa-1}) + r^2 a_h(e_h^{\kappa-1}, e_h^{\kappa-1}) \geq -r^2 \|e_h^{\kappa} - e_h^{\kappa-1}\|_h^2$$

Thus, rewriting $\mathrm{L}_1 + \mathrm{L}_1 = \mathrm{R}$ with these calculations gives

$$\begin{aligned}
&(1 - r^2)\|e_h^{\kappa} - e_h^{\kappa-1}\|_h^2 \\
&\leq \|e_h^{\kappa} - e_h^{\kappa-1}\|_h^2 + r^2 a_h(e_h^{\kappa-1}, e_h^{\kappa-1}) + r^2 a_h(e_h^{\kappa-1}, e_h^{\kappa} - e_h^{\kappa-1}) \\
&\leq \|e_h^1\|_h^2 + \sum_{k=1}^{\kappa-1} \|\varepsilon^k\|_h \|e_h^{k+1} - e_h^{k-1}\|_h
\end{aligned}$$

For $0 < r \leq r_0 < 1$, we have $(1 - r^2) \geq (1 - r_0^2) := 1/c_r$ and

$$\begin{aligned}
\|e_h^{\kappa} - e_h^{\kappa-1}\|_h^2 &\leq c_r \|e_h^1\|_h^2 + c_r \sum_{k=1}^{\kappa-1} \|\varepsilon^k\|_h \|e_h^{k+1} - e_h^{k-1}\|_h \\
&\leq c_r \|e_h^1\|_h^2 + \Bigg[c_r^2 \sum_{k=1}^{\kappa-1} \|\varepsilon^k\|_h^2 \Bigg]^{\frac{1}{2}}_{:=E} \Bigg[\sum_{k=1}^{\kappa-1} \|e_h^{k+1} - e_h^{k-1}\|_h^2 \Bigg]^{\frac{1}{2}}
\end{aligned}$$

Convergence of Explicit Scheme

Using $2ab \leq a^2 + b^2$, the last term can be estimated as

$$\sum_{k=1}^{\kappa-1} \|e_h^{k+1} - e_h^{k-1}\|_h^2 \leq \sum_{k=1}^{\kappa-1} \left[\|e_h^{k+1} - e_h^k\|_h + \|e_h^k - e_h^{k-1}\|_h \right]^2 \leq$$

$$2\sum_{k=1}^{\kappa-1} \|e_h^{k+1} - e_h^k\|_h^2 + 2\sum_{k=1}^{\kappa-1} \|e_h^k - e_h^{k-1}\|_h^2 \leq 4\sum_{k=1}^{\kappa} \|e_h^k - e_h^{k-1}\|_h^2$$

Taking square roots and inserting the result next to E in the previous estimate gives

$$\|e_h^\kappa - e_h^{\kappa-1}\|_h^2 - c_r\|e_h^1\|_h^2 \leq 2E \left[\sum_{k=1}^{\kappa} \|e_h^k - e_h^{k-1}\|_h^2 \right]^{\frac{1}{2}}$$

$$\leq \frac{E^2}{\alpha} + \alpha \sum_{k=1}^{\kappa} \|e_h^k - e_h^{k-1}\|_h^2 \leq \frac{E^2}{\alpha} + \alpha K \max_{1\leq k\leq K} \|e_h^k - e_h^{k-1}\|_h^2$$

where $2ab \leq a^2/\alpha + \alpha b^2$ has been used. Taking $\alpha = 1/(2K)$

Convergence of Explicit Scheme

gives

$$\|e_h^\kappa - e_h^{\kappa-1}\|_h^2 \le c_r\|e_h^1\|_h^2 + 2KE^2 + \tfrac{1}{2}\max_{1\le k\le K}\|e_h^k - e_h^{k-1}\|_h^2$$

or

$$\max_{1\le k\le K}\|e_h^k - e_h^{k-1}\|_h^2 \le 2c_r\|e_h^1\|_h^2 + 4KE^2$$

Using Lemma 83 and $\tau^4 K\kappa \le \tau^4 K^2 = T^2\tau^2$ gives,

$$KE^2 \le K\sum_{k=1}^{\kappa}\left[c\tau^2(\tau^2+h^2)\|u\|_{\mathcal{C}^{4,0}(\bar{Q})}\right]^2 \le c^2T^2\tau^2(\tau^2+h^2)^2\|u\|^2_{\mathcal{C}^{4,0}(\bar{Q})}$$

With $\sqrt{a^2+b^2} \le |a|+|b|$, taking square roots above gives

$$\max_{1\le k\le K}\|e_h^k - e_h^{k-1}\|_h \le c(r_0,T)\left[\|e_h^1\|_h + \tau(\tau^2+h^2)\|u\|_{\mathcal{C}^{4,0}(\bar{Q})}\right] =: \mathrm{F}$$

Since

$$\|e_h^k\|_h \le \|e_h^k - e_h^{k-1}\|_h + \|e_h^{k-1}\|_h, \quad e_h^0 = 0$$

summing over k gives

Convergence of Explicit Scheme

$$\max_{1\le k\le K} \|e_h^k\|_h \le K \max_{1\le k\le K} \|e_h^k - e_h^{k-1}\|_h \le K\mathrm{F}$$

By Lemma 82,

$$\begin{aligned}\|e_h^1\|_h &= \left[\, h\sum_{i=1}^{N}(e_i^1)^2 \,\right]^{\frac{1}{2}} \le \sqrt{hN}\max_{1\le i\le N}|e_i^1| \\ &\le c\left[\tau^2 h^2 \|u_0\|_{\mathcal{C}^4(\bar\Omega)} + \tau^3\|u\|_{\mathcal{C}^{0,3}(\bar Q)}\right] \\ &\le c\tau(\tau^2+h^2)\left[\|u\|_{\mathcal{C}^{4,0}(\bar Q)} + \|u\|_{\mathcal{C}^{0,3}(\bar Q)}\right]\end{aligned}$$

Using this estimate for F gives finally

$$\max_{1\le k\le K} \|e_h^k\|_h \le K\mathrm{F} \le c(K\tau)(\tau^2+h^2)\|u\|_{\mathcal{C}^{4,3}(\bar Q)}$$

Observing $K\tau = T$ gives the claimed convergence estimate. ■

Dirichlet Wave Equation on a Square

- Let $\Omega = (0,1)^2$ and $T > 0$.
- For a given constant wave speed $\omega > 0$ we week an approximation of the solution u to the *Wave Equation with Dirichlet Boundary Condition*,

$$\begin{cases} u_{tt}(x,y,t) &= \omega^2 \Delta u(x,y,t), & \text{for } (x,y,t) \in \Omega \times (0,T] \\ u(x,y,t) &= g(x,y,t), & \text{for } (x,y,t) \in \partial\Omega \times [0,T] \\ u(x,y,0) &= u_0(x,y), & \text{for } (x,y,t) \in \Omega \times \{0\} \\ u_t(x,y,0) &= u_1(x,y), & \text{for } (x,y,t) \in \Omega \times \{0\} \end{cases} \tag{26}$$

- For simplicity let $g = 0$. **Exercise**: Generalize to $g \neq 0$.
- For compatibility it must hold that $u_0(x,y) = u_1(x,y) = 0$, $x \in \partial\Omega$.
- The IBVP (26) can be rewritten in first order form as

$$\begin{pmatrix} \omega\nabla u \\ u_t \end{pmatrix}_t = \begin{pmatrix} 0 & \omega\nabla \\ \omega\nabla\cdot & 0 \end{pmatrix} \begin{pmatrix} \omega\nabla u \\ u_t \end{pmatrix}, \quad \begin{pmatrix} \omega\nabla u \\ u_t \end{pmatrix}_{t=0} = \begin{pmatrix} \omega\nabla u_0 \\ u_1 \end{pmatrix}$$

with the boundary condition $u_t|_{\partial\Omega} = 0$.

Properties of First Order Form

- Note that when the gradient ∇ is seen as an operator equipped with the homogenous boundary condition corresponding to $u_t|_{\partial\Omega} = 0$, then the divergence operator $-\nabla\cdot$ is seen as the adjoint,

$$\int_\Omega \nabla\phi \cdot \Phi = -\int_\Omega \phi \nabla \cdot \Phi, \quad \forall\phi \in C_0^\infty(\Omega), \quad \forall\Phi \in [C^\infty(\bar{\Omega})]^2$$

- Thus, the operator for the first order form of the wave equation shown above can be written as

$$\begin{pmatrix} 0 & \omega\nabla \\ \omega\nabla\cdot & 0 \end{pmatrix} = \begin{pmatrix} 0 & \omega\nabla \\ -\omega(\nabla)^* & 0 \end{pmatrix}$$

  suggesting that the approximation to this operator should be skew symmetric.
- As shown in [58], given $\bar{\Omega}_h$, let the gradient $\nabla \approx \nabla_h = [D_{h,x}; D_{h,y}]$ be approximated with Dirichlet boundary conditions of $D_{h,x}$ and $D_{h,y}$.

Properties of First Order Form

- Then (26) can be semi-discretized spatially by the system of ODEs (method of lines),

$$U_h'(t) = B_h U_h(t), \quad B_h = \begin{pmatrix} 0 & \omega\nabla_h \\ -\omega\nabla_h^\top & 0 \end{pmatrix}$$

where

$$U_h(t) = \begin{pmatrix} \{U^x_{i+\frac{1}{2},j}(t) : 0 \le i \le N, 1 \le j \le N\} \\ \{U^y_{i,j+\frac{1}{2}}(t) : 1 \le i \le N, 0 \le j \le N\} \\ \{U^t_{i,j}(t) : 1 \le i \le N, 1 \le j \le N\} \end{pmatrix}$$

$$\approx \begin{pmatrix} \{\omega u_x(x_{i+\frac{1}{2}}, y_j, t) : 0 \le i \le N, 1 \le j \le N\} \\ \{\omega u_y(x_i, y_{j+\frac{1}{2}}, t) : 1 \le i \le N, 0 \le j \le N\} \\ \{u_t(x_i, y_j, t) : 1 \le i \le N, 1 \le j \le N\} \end{pmatrix}$$

and

$$U_h(0) = \begin{pmatrix} \{\omega \partial_x u_0(x_{i+\frac{1}{2}}, y_j, t) : 0 \le i \le N, 1 \le j \le N\} \\ \{\omega \partial_y u_0(x_i, y_{j+\frac{1}{2}}, t) : 1 \le i \le N, 0 \le j \le N\} \\ \{u_1(x_i, y_j, t) : 1 \le i \le N, 1 \le j \le N\} \end{pmatrix}$$

Semi-Discrete Solution

- This semi-discrete solution $U_h(t)$ has the property

$$\tfrac{1}{2} D_t \|U_h(t)\|^2 = U_h(t) \cdot U_h'(t) = U_h(t) \cdot [B_h U_h(t)] = 0$$

  which implies the conservation $\|U_h(t)\| = \|U_0\|$, analogous to the estimate for the solution u to (26),

$$\tfrac{1}{2} D_t \int_\Omega (u_t^2 + \omega^2 |\nabla u|^2) = \int_\Omega (u_t u_{tt} + \omega^2 \nabla u \cdot \nabla u_t)$$
$$= \int_\Omega u_t u_{tt} + \int_{\partial\Omega} \omega u_t \partial_n u - \int_\Omega \omega^2 u_t \Delta u = 0$$

  corresponding to the conservation of energy, kinetic plus potential.
- Note that the fully-discrete explicit scheme analyzed above does not generally satisfy a conservation property $\|U^k\| = \|U^{k-1}\|$ where $\|\cdot\|$ is an appropriate energy norm.
- **Exercise**: The Crank Nicholson scheme

$$(U_h^{k+1} - U_h^k)/\tau = \tfrac{1}{2} B_h (U_h^{k+1} + U_h^k), \quad [I - \tfrac{1}{2}\tau B_h] U_h^{k+1} = [I + \tfrac{1}{2}\tau B_h] U_h^k$$

  satisfies the conservation condition $\|U_h^{k+1}\| = \|U_h^k\|$.

Crank Nicholson Scheme

- **Exercise**: Analogous to 83, show that the Crank Nicholson scheme is consistent to $\mathcal{O}(\tau(\tau^2 + h^2))$.
- **Exercise**: Analogous to 85 (or more easily using the method of 67) show that the Crank Nicholson scheme is convergent with convergence rate $\mathcal{O}(\tau^2 + h^2)$.
- (`sparse`!) Matlab commands for solving (26) (`Dx` and `Dy` from 59, `U0`, `U1` given) are

```
NN = N*N; N1 = N+1; NN1 = N*N1; h= 1/N1; ta = T/K;
Dh = [Dx;Dy]; Uh = [om*Dh*U0;U1]; uh = U0; vt = U1;
Bh = [sparse(2*NN1,2*NN1),om*Dh;-om*Dh',sparse(NN,NN)];
Ch = speye(2*NN1+NN)+0.5*ta*Bh;
Ah = speye(2*NN1+NN)-0.5*ta*Bh;
for k=1:K
     Uh = Ah \ (Ch*Uh);
     ut = Uh(2*NN1+1:end);
     uh = uh + ta*(ut + vt)/2; vt = ut;
end
```

Non-Linear Hyperbolic IBVP for a Cord

- Let the rest length of a bungee cord be parameterized by $s \in \Omega = (0, 1)$ so the total length L of the cord is 1.
- Let $u(s, t) = (x(s, t), y(s, t), z(s, t)) \in \mathbb{R}^3$ represent the cord at position s and at time t.
- Let the cord be fastened at one end, $u(0, t) = 0$, with known initial position and velocity, $u(s, 0) = u_0(s)$, $u_t(s, 0) = u_1(s)$.
- The cord is loaded externally by $f(s, t) \in \mathbb{R}^3$ (force per unit length) and internally through tension related to the elastic modulus $\kappa(s)$ (force units).
- The density of the cord is $\rho(s)$ (mass per unit length).
- The variational principle used to model the dynamic shape of the cord over the time interval $t \in [0, T]$ is to find a stationary state for the Lagrangian functional (s.t. ICs & BCs)

$$L(u) = \int_0^T \int_0^1 \left[\tfrac{1}{2}\rho |u_t|^2 - \tfrac{1}{2}\kappa (|u_s| - 1)^2 + f \cdot u \right] ds dt$$

to transform kinetic or potential energy most efficiently to the other type of energy. (cf. Newton's Law!)

Stationary Lagrangian for Non-Linear Mechanics

- Without approximating $(|u_s| - 1)^2$, let $Q = \Omega \times (0, T)$, $v \in \{\phi \in \mathcal{C}^\infty(\bar{Q}) : \phi(0, t) = 0, \phi(s, 0) = \phi(s, T) = 0\}$ to obtain

$$\frac{\delta L}{\delta u}(u; v) = \int_Q \left[\rho u_t v_t - \kappa(|u_s| - 1)\frac{u_s \cdot v_s}{|u_s|} + f \cdot v\right]$$

$$= \int_Q v\left[-\rho u_{tt} + \left(\kappa\frac{|u_s| - 1}{|u_s|}u_s\right)_s + f\right]$$

$$+ \int_\Omega \rho u_t v \Big|_{t=0}^{t=T} - \int_0^T \kappa\frac{|u_s| - 1}{|u_s|}u_s \cdot v \Big|_{s=0}^{s=1}$$

- The necessary optimality condition for a stationary u is

$$\begin{cases} \rho u_{tt} = \left(\kappa\dfrac{|u_s| - 1}{|u_s|}u_s\right)_s + f, & \text{for } (s, t) \in Q \\ u(0, t) = 0, \quad (1 - 1/|u_s|)u_s(1, t) = 0, & \text{for } t \in [0, T] \\ u(s, 0) = u_0(s), \quad u_t(s, 0) = u_1(s), & \text{for } s \in \Omega \end{cases}$$

where $u_t(s, 0) = u_1(s)$ is imposed initially instead of a final time condition $u(s, T) = u_T(s)$ corresponding to $v(s, T) = 0$.

Conservation Property for the Nonlinear IBVP

- If the cord is very *taut* and $|u_s| \gg 1$, the non-linear IBVP reduces to a linear IBVP for the wave equation.
- For simplicity we assume that $\omega^2 = \kappa/\rho$ is a constant and that $g = f/\rho$ is a constant vector (e.g., gravitational force).
- Because of the conservation property,

$$\tfrac{1}{2} D_t \int_\Omega \left[\tfrac{1}{2}\rho |u_t|^2 + \tfrac{1}{2}\kappa(|u_s| - 1)^2 - f \cdot u \right]$$

$$= \int_\Omega \left[\rho u_t \cdot u_{tt} + \kappa(|u_s| - 1) \frac{u_s \cdot u_{st}}{|u_s|} - f \cdot u_t \right]$$

$$= \int_\Omega u_t \cdot \left[\rho \cdot u_{tt} - \left(\kappa \frac{|u_s| - 1}{|u_s|} u_s \right)_s - f \right] + \kappa(|u_s| - 1) \frac{u_s \cdot u_t}{|u_s|} \Bigg|_{s=0}^{s=1} = 0$$

  there is tendency to define the state in terms of $\rho^{\frac{1}{2}} u_t$ and $\kappa^{\frac{1}{2}}(1 - 1/|u_s|)u_s$.
- Yet, the non-linear IBVP is written here in first-order form with the state $\vec{u} = (u; u_t) = (x, y, z; x_t, y_t, z_t)$ as follows:

Nonlinear Wave Equation in First Order Form

$$\begin{cases} \begin{pmatrix} u \\ u_t \end{pmatrix}_t = \begin{pmatrix} 0 & 1 \\ \omega^2 \partial_s (1 - 1/|u_s|)\partial_s & -c \end{pmatrix} \begin{pmatrix} u \\ u_t \end{pmatrix} + \begin{pmatrix} 0 \\ g \end{pmatrix} & \text{in } Q \\ \begin{pmatrix} u \\ u_t \end{pmatrix}_{t=0} = \begin{pmatrix} u_0 \\ u_1 \end{pmatrix}, \quad \begin{pmatrix} u \\ u_t \end{pmatrix}_{s=0} = 0, \quad (1 - 1/|u_s|)\partial_s \begin{pmatrix} u \\ u_t \end{pmatrix}_{s=1} = 0 \end{cases} \tag{27}$$

where now damping $c \geq 0$ has also been introduced.

- To approximate the solution, (27) is first discretized temporally with a semi-implicit Euler scheme to obtain

$$\begin{cases} L_k \vec{u}^k = \vec{u}^{k-1} + \tau \vec{g}, & \text{in } \Omega \qquad k = 1, \ldots, K \\ \vec{u}^k|_{s=0} = 0, \quad \partial_s \vec{u}^k|_{s=1} = 0 \qquad \vec{u}^0 = \vec{u}_0 \end{cases} \tag{28}$$

where $\vec{g} = (0; g)$, $\vec{u}^k = (u^k; u_t^k)$, $\vec{u}_0 = (u_0; u_1)$, $T = K\tau$ and

$$L_k \vec{u} = \vec{u} - \tau \begin{pmatrix} 0 & 1 \\ \omega^2 \partial_s a_k \partial_s & -c \end{pmatrix} \vec{u}, \quad a_k = 1 - 1/|u_s^{k-1}|$$

- To approximate the solution to (28) spatially, let Ω be discretized with the grid $\bar{\Omega}_h = \{s_i = ih\}_{i=0}^{N+1}$, $h = 1/(N+1)$.

Fully Discrete Approximation of Nonlinear IBVP

- The time-discretized state
$$\vec{u}^k = (u^k; u_t^k) = (x^k, y^k, z^k; x_t^k, y_t^k, z_t^k)$$
is approximated in a fully discrete scheme by $\vec{U}^k \approx \vec{u}^k$ with grid values (only at s_i, $i > 0$, since $\vec{U}^k(s_0) = 0$)
$\vec{U}_h^k = (U_h^k; U_{h,t}^k) = (X_h^k, Y_h^k, Z_h^k; X_{h,t}^k, Y_{h,t}^k, Z_{h,t}^k) \in \mathbb{R}^{2(N+1)\times 3}$.
- The operator ∂_s is approximated for a grid function v by
$$[D_{h,s}v](s + \tfrac{1}{2}ih) = [v(s + ih) - v(s)]/h$$
where $v(0)$ is understood to be 0 due to the boundary condition at s_0.
- The matrix representation of $D_{h,s}$ is
$$D_{h,s} = \frac{1}{h}\begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}$$
mapping grid values at $\{s_i\}_{i=1}^{N+1}$ to interface values at $\{s_{i+\frac{1}{2}}\}_{i=0}^{N}$.

Fully Discrete Approximation of Nonlinear IBVP

- The coefficient a_k is approximated by

$$a_{h,k} = 1 - 1./|D_{h,s}U_h^{k-1}|$$

where the term $|u_s^{k-1}|$ is approximated by

$$|u_s^{k-1}| \approx |D_{h,s}U_h^{k-1}| = \sqrt{(D_{h,s}X_h^{k-1}).\hat{}2 + (D_{h,s}Y_h^{k-1}).\hat{}2 + (D_{h,s}Z_h^{k-1}).\hat{}2}$$

- Let $\mathcal{D}(V)$ denote a diagonal matrix with the values of the vector V along the diagonal.
- Since in the term $\partial_s^{(l)} a_k \partial_s^{(r)}$ it holds that $\partial_s^{(l)} = -(\partial_s^{(r)})^*$, the term $\partial_s a_k \partial_s$ is approximated with $-D_{h,s}^\top \mathcal{D}(a_{h,k}) D_{h,s}$.
- The spatial discretization of (28) is given by

$$A_{h,k}\vec{U}^k = \vec{U}^k + \tau\vec{g} \tag{29}$$

where the coefficient matrix $A_{h,k}$ is given by

$$A_{h,k} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \tau \begin{pmatrix} 0 & I \\ -\omega^2 D_{h,s}^\top \mathcal{D}(a_{h,k}) D_{h,s} & -cI \end{pmatrix}$$

Code to Solve the Non-Linear IBVP

- (`sparse`!) Matlab commands for solving (29):

```
N1 = N+1; h = 1/N1;
g = [zeros(N1,1),zeros(N1,1),-ones(N1,1)];
D = spdiags(kron([-1,1],ones(N1,1)),[-1,0],N1,N1)/h;
s = linspace(0,1,N2);
u1 = s(2:end)'; u2 = zeros(N1,1); u3 = zeros(N1,1);
u = [u1,u2,u3]; ut = zeros(N1,3); U = [u;ut];
for k=1:K
  u = U(1:N1,:);
  u1 = [0;u(:,1)]; u2 = [0;u(:,2)]; u3 = [0;u(:,3)];
  plot3(u1,u2,u3);
  du = sqrt(sum((D*u).^2,2)); a = 1-1./du;
  B = D'*spdiags(om2*a(:),0,N1,N1)*D;
  A = speye(2*N1) ...
    - ta*[sparse(N1,N1),speye(N1);-B,-c*speye(N1)];
  U = A \ (U+ta*[zeros(N1,3);g]);
end
```

Non-Linear Hyperbolic IBVP for a Membrane

Exercise:

- Let the rest area of a membrane be parameterized by $(\xi, \eta) \in \Omega = (0,1)^2$ so the total area A of the membrane is 1.
- Let $u(\xi, \eta, t) = (x(\xi, \eta, t), y(\xi, \eta, t), z(\xi, \eta, t)) \in \mathbb{R}^3$ represent the membrane at position (ξ, η) and at time t.
- Let the membrane be fastened at one end, $u(\xi, 0, t) = (\xi, 0, 0)$, with known initial position and velocity, $u(\xi, \eta, 0) = u_0(\xi, \eta)$, $u_t(\xi, \eta, 0) = u_1(\xi, \eta)$.
- The membrane is loaded externally by $f(\xi, \eta, t) \in \mathbb{R}^3$ (force per unit area) and internally through tension related to the elastic modulus $\kappa(\xi, \eta)$ (force units).
- The membrane density is $\rho(\xi, \eta)$ (mass per unit area).
- For the Lagrangian functional (s.t. ICs & BCs)

$$L(u) = \int_0^T \int_0^1 \int_0^1 \left[\tfrac{1}{2}\rho |u_t|^2 - \tfrac{1}{2}\kappa(|u_\xi \times u_\eta| - 1)^2 + f \cdot u \right] d\xi d\eta dt$$

show that the necessary optimality condition for a

Non-Linear Hyperbolic IBVP for a Membrane

stationary u in $Q = \Omega \times (0, T)$ (with damping $c > 0$) is

$$\begin{cases} \rho u_{tt} = (A(u)u_\xi)_\xi + (B(u)u_\eta)_\eta + f(u) - cu_t, & \text{for } (\xi, \eta, t) \in Q \\ \alpha(u) = \dfrac{1 - |u_\xi \times u_\eta|}{|u_\xi \times u_\eta|}, \quad A(u) = \alpha(u)[u_\eta]^2, \quad B(u) = \alpha(u)[u_\xi]^2 & \\ u(\xi, 0, t) = (\xi, 0, 0), \quad B(u)u_\eta(\xi, 1, t) = 0, & \text{for } \xi \in [0, 1], t \in [0, T] \\ A(u)[u_\xi(\xi, \eta, t) - (1, 0, 0)] = 0, \quad \xi = 0, 1, & \text{for } \eta \in [0, 1], t \in [0, T] \\ u(\xi, \eta, 0) = u_0(\xi, \eta), \quad u_t(\xi, \eta, 0) = u_1(\xi, \eta), & \text{for } (\xi, \eta) \in \Omega \end{cases}$$

where $u_t(\xi, \eta, 0) = u_1(\xi, \eta)$ is imposed initially instead of a final time condition $u(\xi, \eta, T) = u_T(\xi, \eta)$. Here, for a vector

$$a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad \text{define} \quad [a] = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix} \quad \text{so} \quad u \times v = [u]v$$

- Rewrite this problem in first order form, formulate a spatial discretization in (ξ, η), apply the Crank Nicholson scheme for a discretization in t and implement the fully discrete approximation with Matlab. Does a solution exist for all parameters?

Scalar Convection Equation

- *Finite Difference Methods for Conservation Laws* will first be investigated by considering the IVP for the *scalar convection equation*,

$$\begin{cases} u_t(x,t) - \alpha u_x(x,t) &= 0, & x \in \mathbb{R}, \quad t \in (0,T] \\ u(x,0) &= u_0(x), & x \in \mathbb{R} \end{cases} \tag{30}$$

  where $\alpha > 0$ is constant.
- The solution is given explicitly by the left traveling wave

$$u(x,t) = u_0(x+\alpha t), \quad x \in \mathbb{R}, \quad t \geq 0$$

  so the IVP need not be solved numerically.
- Yet the study of numerical methods for such a simple problem will set the stage for the general theory.
- The solution to (30) is approximated with *forward spatial differences* by $U^k(x) \approx u(x,t_k)$ where $t_k = k\tau$, $\tau = T/K$, and

$$\begin{cases} [U^{k+1}(x) - U^k(x)]/\tau &= \alpha[U^k(x+h) - U^k(x)]/h, & x \in \mathbb{R}, \quad 0 \leq k \leq K-1 \\ U^0(x) &= u_0(x), & x \in \mathbb{R} \end{cases}$$

Forward Difference Scheme

- With the *Courant Number* $r = \alpha\tau/h$, the scheme becomes

$$\begin{cases} U^{k+1}(x) &= rU^k(x+h) + (1-r)U^k(x), & x \in \mathbb{R}, 0 \le k \le K-1 \\ U^0(x) &= u_0(x), & x \in \mathbb{R} \end{cases} \tag{31}$$

- While a spatial grid $\{x_i = ih : i \in \mathbb{Z}\}$ can be introduced, the theory will be carried out using a continuous variable x.
- Although a practical problem would involve a bounded spatial domain and a boundary condition, the analysis of the IVP on an infinite domain can serve as a guide to the local behavior of a more realistic solution.
- A more detailed analysis is necessary when boundary effects would be significant.
- The analysis of (31) (stability, consistency, convergence) will be carried out in $L^2(\mathbb{R})$ with Hilbert space structure.
- The (extended) Fourier Transform will be used for $f, \hat{f} \in L^2(\mathbb{R})$

$$\hat{f}(\xi) = (\mathcal{F}f)(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) e^{-\imath x\xi} dx, \quad \xi \in \mathbb{R} \quad (\imath^2 = -1)$$

Fourier Transforms

with inverse formula, $f(x) = (\mathcal{F}^{-1}\hat{f})(x)$,

$$\check{g}(x) = (\mathcal{F}^{-1}g)(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} g(\xi)e^{\imath x\xi}dx, \quad x \in \mathbb{R}$$

- By Parseval's Identity,

$$(f,g)_{L^2} = \int_{-\infty}^{+\infty} f(x)g^*(x)dx = \int_{-\infty}^{+\infty} \hat{f}(\xi)\hat{g}^*(\xi)d\xi = (\hat{f},\hat{g})_{L^2}$$

  the Fourier and Inverse Fourier Transforms are isometric isomorphisms.
- For $f \in L^2(\mathbb{R})$ absolutely continuous with $f(x) \to 0$, $|x| \to \infty$, it holds that

$$(\mathcal{F}f')(\xi) = \imath\xi\hat{f}(\xi).$$

- Under more general smoothness assumptions it holds that

$$(\mathcal{F}f^{(n)})(\xi) = (\imath\xi)^n\hat{f}(\xi).$$

- The translation operator $(T_a f)(x) = f(x+a)$ satisfies

$$(\mathcal{F}T_a f)(\xi) = e^{\imath\xi a}\hat{f}(\xi).$$

Fourier Transforms

- Under suitable regularity assumptions we take the Fourier Transform of (30) to obtain

$$(\mathcal{F}u_x)(\xi,t) = \imath\xi\hat{u}(\xi,t), \quad (\mathcal{F}u_t)(\xi,t) = \hat{u}_t(\xi,t)$$

and

$$\begin{cases} \hat{u}_t(\xi,t) &= \imath\xi\alpha\hat{u}(\xi,t), \quad \xi\in\mathbb{R}, \quad t\in(0,T] \\ \hat{u}(\xi,0) &= \hat{u}_0(\xi), \quad \xi\in\mathbb{R} \end{cases}$$

- The solution to this ξ-parameterized ODE is

$$\hat{u}(\xi,t) = e^{\alpha\imath\xi t}\hat{u}_0(\xi) = (\mathcal{F}T_{\alpha t}u_0)(\xi) \quad \text{or} \quad u(x,t) = u_0(x+\alpha t)$$

which implies the conservation $\|u(\cdot,t)\|_{L^2} = \|u_0\|_{L^2}$ since

$$\|u(\cdot,t)\|_{L^2}^2 = \|\hat{u}(\cdot,t)\|_{L^2}^2 = \int_{-\infty}^{+\infty} |e^{\alpha\imath\xi t}\hat{u}_0(\xi)|^2 d\xi = \|\hat{u}_0\|_{L^2}^2 = \|u_0\|_{L^2}^2$$

- We now consider convergence of the scheme (31). Taking the Fourier Transform gives

$$\hat{U}^{k+1}(\xi) = G\hat{U}^k(\xi) := \hat{U}^k(\xi) + r[e^{\imath\xi h}\hat{U}^k(\xi) - \hat{U}^k(\xi)]$$

Consistency of the Forward Difference Scheme

or $\hat{U}^k(\xi) = G^k \hat{u}_0(\xi)$ where G is the *amplification factor*,

$$G = G(h, \tau) = 1 + r(e^{\imath \xi h} - 1) \quad (r = \alpha\tau/h)$$

- Consistency of the scheme is readily shown using Taylor expansions,

$$|u(x,(k+1)\tau) - \{u(x,k\tau) + r[u(x+h,k\tau) - u(x,k\tau)]\}| \leq c\tau^2 \|u\|_{\mathcal{C}^{2,0}(\mathbb{R}\times[0,T])}$$

but it is convenient to express consistency in the transformed variable:

Lemma: Let $r = \alpha\tau/h$ be fixed. Then as $\xi h \to 0$ (equivalently as $\xi\tau \to 0$) $\exists c \neq c(\xi, \tau, h)$ s.t.

$$|G - e^{\imath\alpha\xi\tau}| \leq c\tau^2\xi^2 \tag{32}$$

Proof: Follows with

$$G = 1 + re^{\imath\xi h} - r = 1 + r[1 + \imath h\xi + \mathcal{O}(h^2\xi^2)] - r$$

$$= 1 + \imath rh\xi + \mathcal{O}(h^2\xi^2) = 1 + \imath\alpha\tau\xi + \mathcal{O}(\tau^2\xi^2) = e^{\imath\alpha\tau\xi} + \mathcal{O}(\tau^2\xi^2)$$

■

Stability of the Forward Difference Scheme

- Furthermore, (with $[1 + az + \mathcal{O}(z^2)]^c = e^{c[az+\mathcal{O}(z^2)]}$)

$$G^k = (G)^{t_k/\tau} = [1+\imath\alpha\tau\xi+\mathcal{O}(\tau^2\xi^2)]^{t_k/\tau} = e^{\alpha\imath\tau\xi t_k/\tau}\, e^{\mathcal{O}(\tau^2\xi^2 t_k/\tau)}$$

$$= e^{\alpha\imath\xi t_k}\, e^{t_k\tau\xi^2} \quad \text{so} \quad G^k - e^{\alpha\imath\xi k\tau} = e^{\alpha\imath\xi t_k}[e^{\mathcal{O}(t_k\tau\xi^2)} - 1]$$

- Then the pointwise convergence (r, t_k fixed) follows:

$$\lim_{\tau\to 0}[\hat{U}^k(\xi) - \hat{u}(\xi, t_k)] = \lim_{\tau\to 0}[G^k - e^{\alpha\imath\xi k\tau}]\hat{u}_0(\xi) = 0.$$

- However, pointwise convergence does not imply convergence in $L^2(\mathbb{R})$. It must yet be shown that
$$\|U^k - u(\cdot, t_k)\|_{L^2} = \|\hat{U}^k - \hat{u}(\cdot, t_k)\|_{L^2} \to 0,\ \tau \to 0$$
for which the following stability is required.

Lemma: Let $u_0 \in L^2(\mathbb{R})$ and suppose $r \in (0, 1]$. Then

$$\|U^k\|_{L^2} \le \|u_0\|_{L^2}, \quad \forall k \ge 0.$$

Proof: For $r \in (0, 1]$, $|G| \le 1$ follows from

$$|1 + re^{\imath\xi h} - r| = |1 - r + re^{\imath\xi h}| \le |1 - r| + |r| = 1 - r + r = 1.$$

Stability of the Forward Difference Scheme

From $\hat{U}^k(\xi) = G^k \hat{u}_0(\xi)$ it follows

$$|\hat{U}^k(\xi)| = |G|^k |\hat{u}_0(\xi)| \leq |\hat{u}_0(\xi)|, \quad \forall \xi \in \mathbb{R}$$

or

$$\|U^k\|_{L^2}^2 = \|\hat{U}^k\|_{L^2}^2 = \int_{-\infty}^{+\infty} |\hat{U}^k(\xi)|^2 d\xi \leq \int_{-\infty}^{+\infty} |\hat{u}_0(\xi)|^2 d\xi = \|u_0\|_{L^2}^2$$

■

- The condition $|G| \leq 1$ is the *Von Neumann stability condition*.
- The condition $r \in (0, 1]$ is seen above to be sufficient for stability. Yet it is also necessary:

Exercise: Construct an example for which the scheme (31) gives $\|U^k\|_{L^2} \to \infty$ as $k \to \infty$ for fixed $t^k = k\tau$ and fixed $r = \alpha\tau/h > 1$.

- For convergence we introduce the following function spaces (with minimal background).

Def: The *Sobolev Space* $H^m(\mathbb{R})$ consists of those functions whose derivatives up to order m are in $L^2(\mathbb{R})$.

Sobolev Spaces

- The Sobolev spaces have the Hilbert space structure with the scalar product,

$$(u, v)_{H^m} = \sum_{l=0}^{m} \int_{-\infty}^{+\infty} u^{(l)}(x)^* v^{(l)}(x) dx$$

- By Parseval's Identity and the derivative property

$$(u, v)_{H^m} = \sum_{l=0}^{m} \int_{-\infty}^{+\infty} [(\imath\xi)^l \hat{u}(\xi)]^* [(\imath\xi)^l \hat{v}(\xi)] d\xi$$

- The norm on $H^m(\mathbb{R})$ is then given by

$$\|u\|_{H^m} = \left[\sum_{l=0}^{m} \int_{-\infty}^{+\infty} |u^{(l)}(x)|^2 dx \right]^{\frac{1}{2}} = \left[\sum_{l=0}^{m} \int_{-\infty}^{+\infty} |\xi|^{2l} |\hat{u}(\xi)|^2 d\xi \right]^{\frac{1}{2}}$$

Theorem: Let $u_0 \in H^2(\mathbb{R})$. Let $r = \alpha\tau/h \in (0, 1]$ be fixed. Then $\exists c \neq c(\tau, h)$ s.t. for $\tau = T/K$,

$$\max_{0 \leq k \leq K} \|U^k - u(\cdot, t_k)\|_{L^2} \leq c\tau \|u_0\|_{H^2}$$

Convergence of the Forward Difference Scheme

Proof: Using $\hat{u}(\xi,t) = e^{\alpha \imath \xi t}\hat{u}_0(\xi)$ and $\hat{U}^k(\xi) = G^k\hat{u}_0(\xi)$ gives

$$\|U^k - u(\cdot,t_k)\|_{L^2}^2 = \|\hat{U}^k - \hat{u}(\cdot,t_k)\|_{L^2}^2 = \int_{-\infty}^{+\infty} |\hat{u}_0(\xi)|^2 |G^k - e^{\imath\alpha\xi k\tau}|^2 d\xi$$

and this integral I is now partitioned as $I = I_1 + I_2$ with I_1 taken over the set $\{\xi : |\xi| \geq \gamma/h\}$ and I_2 over the set $\{\xi : |\xi| < \gamma/h\}$, where γ is to be determined. First consider I_1. Note from $|G| \leq 1$ that

$$|G^k - e^{\imath\alpha\xi k\tau}|^2 \leq (|G|^k + |e^{\imath\alpha\xi k\tau}|)^2 \leq 4$$

and hence

$$I_1 \leq 4\int_{|\xi|\geq\gamma/h} |\hat{u}_0(\xi)|^2 d\xi \leq \frac{4h^2}{\gamma^2}\int_{|\xi|\geq\gamma/h} |\xi|^2|\hat{u}_0(\xi)|^2 d\xi$$

$$= \frac{4\alpha^2\tau^2}{r^2\gamma^2}\int_{|\xi|\geq\gamma/h} |\xi|^2|\hat{u}_0(\xi)|^2 d\xi \leq c\tau^2\|u_0'\|_{L^2}^2 \leq c\tau^2\|u_0\|_{H^2}^2.$$

Now consider I_2. Let $\gamma > 0$ be chosen so that the consistency estimate (32) holds for $|h\xi| < \gamma$. Also note that for $|a|, |b| \leq 1$,

$$|a^k - b^k| = |(a-b)\textstyle\sum_{m=0}^{k-1} a^{k-m}b^m| \leq k|a-b|.$$

Backward Difference Scheme

Hence for $|h\xi| < \gamma$, (32) gives

$$|G^k - e^{\imath\alpha\xi\tau k}| \leq k|G - e^{\imath\alpha\xi\tau}| \leq ck\tau^2\xi^2 \leq cT\tau\xi^2$$

Hence,

$$I_2 \leq c\tau^2 \int_{|h\xi|<\gamma} |\hat{u}_0(\xi)|^2|\xi|^4 d\xi \leq c\tau^2\|u_0\|^2_{H^2}$$

■

- For $r = 1$ the scheme (31) is exact. (verify!)
- For the IVP (30) with $-\alpha$ replaced by $+\alpha$ (again with $\alpha > 0$) the solution is given by the right traveling wave $u(x,t) = u_0(x - \alpha t)$ and is approximated with *backward spatial differences* by $U^k(x) \approx u(x, t_k)$ where

$$\begin{cases} [U^{k+1}(x) - U^k(x)]/\tau &= \alpha[U^k(x) - U^k(x-h)]/h, & x \in \mathbb{R}, \quad 0 \leq k \leq K-1 \\ U^0(x) &= u_0(x), & x \in \mathbb{R} \end{cases}$$

- **Exercise**: Find the amplification factor G for the backward difference scheme and show that $|G| \leq 1$ holds iff $r \in (0, 1]$. Prove convergence of the scheme and determine the order of convergence.

Lax Wendroff Scheme

- **Exercise**: For the solution to $u_t + \alpha u_x = 0$ (with arbitrary sign(α)), the *Lax Wendroff* scheme

$$\begin{cases} \dfrac{U^{k+1}(x) - U^k(x)}{\tau} = -\alpha \dfrac{U^k(x+h) - U^k(x-h)}{2h} \\ \qquad\qquad + \dfrac{\alpha^2 \tau}{2} \dfrac{U^k(x+h) - 2U^k(x) + U^k(x-h)}{h^2} \\ \qquad\qquad\qquad\qquad x \in \mathbb{R}, \quad 0 \le k \le K-1 \\ U^0(x) = u_0(x), \qquad x \in \mathbb{R} \end{cases}$$

  can be seen as a discretization of $u_t + \alpha u_x = \epsilon u_{xx}$ containing the *artificial* or *pseudo-viscosity* term ϵu_{xx} with $\epsilon = \alpha^2 \tau / 2$. Find the amplification factor G for the Lax Wendroff scheme and show that $|G| \le 1$ holds iff $|r| \le 1$. Prove that the scheme converges with the order,

$$\max_{0 \le k \le K} \|U^k - u(\cdot, t_k)\|_{L^2} \le c\tau^2 \|u_0\|_{H^3}$$

- Find the artificial viscosity term for the forward and the backward difference schemes and observe the effect of sign(α) for each.

Linear Hyperbolic Systems

- Now for $A \in \mathbb{R}^m$ consider the IVP (*Cauchy Problem*),

$$\begin{cases} u_t + Au_x = 0, & x \in \mathbb{R}, \quad t > 0 \\ u = u_0, & x \in \mathbb{R}, \quad t = 0 \end{cases} \tag{33}$$

- With the spatial grid $x_i = ih$, $h > 0$, $i \in \mathbb{Z}$, and the temporal grid $t_k = k\tau$, $\tau > 0$, $k \in \mathbb{N}_0$, let U_i^k denote an approximation of the *cell average*,

$$u_i^k = \frac{1}{h} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x, t_k) dx$$

and define the piecewise constant function

$$U_\tau(x, t) = U_i^k, \quad (x, t) \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}) \times [t_k, t_{k+1}).$$

- The ratio τ/h is assumed to be a fixed constant.
- In practice, (33) is solved on a finite spatial domain, e.g., $x \in (0, 1)$, with, e.g., periodic boundary conditions,

$$u(0, t) = u(1, t), \quad t \geq 0$$

Periodic and Inflow Boundary Conditions

or inflow boundary conditions, (figure forthcoming)

$$\begin{cases} v_j(0,t) = l_j(t), & \lambda_j > 0 \\ v_j(1,t) = r_j(t), & \lambda_j < 0 \end{cases} \qquad t \geq 0$$

where $SAS^{-1} = \Lambda = \mathrm{diag}\{\lambda_j\}_{j=1}^m$, $\{v_j\} = v = Su$ and

$$v_t + \Lambda v_x = S(u_t + Au_x) = 0.$$

- An apparently natural discretization for (33) is

$$\frac{U_i^{k+1} - U_i^k}{\tau} + A\frac{U_{i+1}^k - U_{i-1}^k}{2h} = 0$$

but this *explicit Euler* method is unstable. (verify!)

- On the other hand, the *implicit Euler* method

$$\frac{U_i^{k+1} - U_i^k}{\tau} + A\frac{U_{i+1}^{k+1} - U_{i-1}^{k+1}}{2h} = 0$$

is stable but involves an unnatural and expensive coupling among all numerical values at time t_{k+1}.

Upwinding and Other Single-Step Schemes

- Other methods considered earlier are: the forward difference scheme, the backward difference scheme and the Lax Wendroff scheme.
- Other well-known methods include the Lax Friedrichs scheme,

$$\frac{U_i^{k+1} - (U_{i+1}^k + U_{i-1}^k)/2}{\tau} + A\frac{U_{i+1}^k - U_{i-1}^k}{2h} = 0$$

and the Leapfrog scheme (figure forthcoming)

$$\frac{U_i^{k+1} - U_i^{k-1}}{2\tau} + A\frac{U_{i+1}^k - U_{i-1}^k}{2h} = 0$$

but, while all other previously mentioned methods are single-step methods, the Leapfrog scheme is a two-step method.

- The *upwind method* is given by

$$\frac{U_i^{k+1} - U_i^k}{\tau} + A^+\frac{U_i^k - U_{i-1}^k}{h} + A^-\frac{U_{i+1}^k - U_i^k}{h} = 0$$

where $A^+ = S\Lambda^+S^{-1}$, $A^- = S\Lambda^-S^{-1}$
and $\Lambda^+ = \max\{\Lambda, 0\}$, $\Lambda^- = \min\{\Lambda, 0\}$.

Upwinding and Other Single-Step Schemes

Exercise:

- For $\Omega = (0,1)$ write the IBVP

$$\begin{cases} v_{tt} = \omega^2 v_{xx}, \quad (x,t) \in \Omega \times (0,\infty) \\ v(0,t) = v(1,t), \quad v_x(0,t) = v_x(1,t), \quad t \in (0,\infty) \\ v(x,0) = v_0(x), \quad v_t(x,0) = v_1(x), \quad x \in \Omega \end{cases}$$

  as a linear hyperbolic system $u_t + Au_x = 0$ with periodic boundary conditions and with appropriate initial conditions.
- Choose smooth data $v_0(x) = \sin(2k\pi x)$, $k \in \mathbb{N}$, and then rough data $v_0(x) = (\frac{1}{4} < x < \frac{3}{4})$, and then choose v_1 so that there are waves traveling (a) only to the right, (b) only to the left or (c) mirrored in both directions.
- Implement the single-step schemes above and estimate the respective convergence rates. Report on apparent amplitude (dissipation) and phase (dispersion) errors.
- Write new boundary conditions so that left and right traveling waves are purely absorbed at the left and right boundaries respectively. Implement the upwind scheme for this problem.

Consistency of Single-Step Schemes

- All the one-step methods above can be written in the form

$$U_i^{k+1} = H_\tau(U^k; i) \quad \text{or} \quad U_\tau(x, t+\tau) = H_\tau(U_\tau(\cdot, t); x)$$

- Although the $L^2(\mathbb{R})$ norm was useful earlier in the context of Fourier analysis, we will now use the $L^1(\mathbb{R})$ norm,

$$\|u\|_{L^1} = \int_{-\infty}^{+\infty} |u(x)|dx, \quad \|U_\tau(\cdot, t_k)\|_{L^1} = h\sum_i |U_i^k|$$

- A method is called *consistent* if for sufficiently smooth u

$$L_\tau(x,t) = [u(x,t+\tau) - H_\tau(u(\cdot,t);x)]/\tau$$

satisfies $\|L_\tau(\cdot,t)\|_{L^1} \to 0$, $\tau \to 0$.

- For the Lax Friedrichs method, Taylor expansions give

$$\begin{aligned} L_\tau(x,t) &= [u(x,t+\tau) - (u(x-h,t)+u(x+h,t))/2]/\tau \\ &\quad + A[u(x+h,t) - u(x-h,t)]/(2h) \\ &= u_t + Au_x + \tfrac{1}{2}[\tau u_{tt} - (h^2/\tau)u_{xx}] + \mathcal{O}(h^2) \end{aligned}$$

Stability of Single-Step Schemes

since τ/h is constant, so with $u_t + Au_x = 0$,

$$L_\tau(x,t) = \tfrac{1}{2}\tau[A^2 - (h/\tau)^2 I]u_{xx} + \mathcal{O}(\tau^2)$$

and given that u_0 has compact support, so does u_{xx} and

$$\|L_\tau(\cdot,t)\|_{L^1} \le c\tau$$

- A method is called *stable* if $\forall T$, $\exists c, \tau_0 > 0$ such that

$$\|H_\tau^k\|_{L^1} \le c, \quad \forall k,\tau \quad \text{with} \quad k\tau \le T, \quad \text{and} \quad \tau < \tau_0.$$

- For the Lax Friedrichs method with $A = \alpha > 0$ and $\alpha\tau/h \le 1$,

$$\begin{aligned}\|U^{k+1}\|_{L^1} &= h\textstyle\sum_i |U_i^{k+1}| \\ &\le \tfrac{h}{2}\textstyle\sum_i |(1-\alpha\tau/h)U_{i+1}^k| + \tfrac{h}{2}\textstyle\sum_i |(1+\alpha\tau/h)U_{i-1}^k| \\ &= \tfrac{1}{2}(1-\alpha\tau/h)\|U^k\|_{L^1} + \tfrac{1}{2}(1+\alpha\tau/h)\|U^k\|_{L^1} = \|U^k\|_{L^1}\end{aligned}$$

Lax Equivalence Theorem: A consistent, linear method is convergent if and only if it is stable.

Proof: (sufficiency) Set $E_\tau(x,t) = U_\tau(x,t) - u(x,t)$. With the

Lax Equivalence Theorem

initial values consistently approximated, we have that $\|E_\tau(\cdot,0)\|_{L^1} \to 0$ as $\tau \to 0$. Also

$$E_\tau(x, t+\tau) = H_\tau(E_\tau(\cdot,t);x) - \tau L_\tau(x,t)$$

satisfies

$$E_\tau(x,t_k) = H_\tau^k(E_\tau(\cdot,0);x) - \tau \sum_{l=1}^{k} H_\tau^{k-l}(L_\tau(\cdot,t_{l-1});x)$$

and hence with $\|H_\tau^k\|_{L^1} \le c$,

$$\begin{aligned}
\|E_\tau(\cdot,t_k)\|_{L^1} &\le \|H_\tau^k\|_{L^1}\|E_\tau(\cdot,0)\|_{L^1} + \tau\sum_{l=1}^{k}\|H_\tau^{k-l}\|_{L^1}\|L_\tau(\cdot,t_{l-1})\|_{L^1} \\
&\le c\left[\|E_\tau(\cdot,0)\|_{L^1} + T\max_{1\le l\le k}\|L_\tau(\cdot,t_{l-1})\|_{L^1}\right] \to 0
\end{aligned}$$

as $\tau \to 0$. ■

- **Exercise**: Show that the Lax Friedrichs scheme converges if the *CFL condition* (Courant, Friedrichs, Lewy) is satisfied:

$$|\lambda_j \tau/h| \le 1, \quad \forall j = 1,\dots,m.$$

- **Exercise**: Show that the backward difference scheme converges if (and only if) $0 \le \lambda_j\tau/h \le 1$.

Computing Discontinuous Solutions

- While wave equations are often given in differential form, the underlying principle involves conservation of integrated quantities, which are not necessarily pointwise smooth.
- Physical waves often involve shocks, i.e., discontinuities.
- For a sufficiently smooth solution to (30), the Lax Friedrichs scheme converges as $\mathcal{O}(\tau)$, and the Lax Wendroff scheme converges as $\mathcal{O}(\tau^2)$.
- When (30) is solved (*weakly*) by a traveling shock wave,
  - first-order schemes, e.g., Lax Friedrichs, typically give smearing around discontinuities (converging as $\mathcal{O}(\tau^{\frac{1}{2}})$), and
  - second-order schemes, e.g., Lax Wendroff, typically give oscillations around discontinuities (converging as $\mathcal{O}(\tau^{\frac{2}{3}})$).

  (figure forthcoming)
- To explain this behavior, a scheme can be seen to provide a higher order approximation of the solution to a so-called *modified equation*, which depends upon discretization parameters.

Modified Equations

- For example, it was shown above for Lax Friedrichs that

$$L_\tau(x,t) = u_t + Au_x + \tfrac{1}{2}\tau[A^2 - (h/\tau)^2 I]u_{xx} + \mathcal{O}(\tau^2)$$

  and hence the scheme can be viewed as providing a $\mathcal{O}(\tau^2)$ approximation of the solution to the *dissipation* equation

$$u_t + Au_x = Du_{xx}, \quad D = \frac{h^2}{2\tau}\left[I - \frac{\tau^2}{h^2}A^2\right]$$

- For $|\tau\lambda_j/h| < 1$, the matrix D and the term u_{xx} lead to forward diffusion, which smears out shocks. Backward diffusion results from $|\tau\lambda_j/h| > 1$, which is unstable.
- **Exercise**: Determine the modified equation for the explicit Euler method

$$\frac{U_i^{k+1} - U_i^k}{\tau} + A\frac{U_{i+1}^k - U_{i-1}^k}{2h} = 0$$

  and use this to explain why the method is unstable for all τ/h.

Dissipation and Dispersion

- The modified equation for the Lax Wendroff method is the *dispersion* equation (verify!)

$$u_t + Au_x = Du_{xxx}, \quad D = \frac{h^2}{6} A \left[\frac{\tau^2}{h^2} A^2 - I \right]$$

- For the case $A = \alpha > 0$ and $D = \mu$, taking the Fourier Transform of the dispersion equation gives

$$\hat{u}_t(\xi, t) + \imath \xi \alpha \hat{u}(\xi, t) = \mu(\imath\xi)^3 \hat{u}(\xi, t), \quad \hat{u}(\xi, 0) = \hat{u}_0(\xi)$$

which is solved by

$$\hat{u}(\xi, t) = e^{-\imath c(\xi) t} \hat{u}_0(\xi), \quad c(\xi) = \alpha\xi + \mu\xi^3$$

- Considering each frequency component separately, set $u_0(x) = e^{\imath\kappa x}$ to obtain $\hat{u}(\xi) = \sqrt{2\pi}\delta(\xi - \kappa)$ and the traveling wave solution $e^{-\imath c(\kappa)t} \mathcal{F}^{-1}[\sqrt{2\pi}\delta(\xi - \kappa)]$ or

$$u(x, t) = e^{\imath(\kappa x - c(\kappa)t)} = e^{\imath\kappa(x - c_p(\kappa)t)}, \quad c_p(\kappa) = c(\kappa)/\kappa = \alpha + \mu\kappa^2$$

with *phase velocity* $c_p(\kappa)$, the speed at which wave peaks travel, depending upon the wave number κ excited in u_0.

Phase and Group Velocities

- A step function, such as $u_0(x) = \text{sign}(x)$, has a broad Fourier spectrum, $\hat{u}_0(\xi) = \sqrt{2/\pi}/(\imath\xi)$.
- For such broad spectrum data, the wave number κ is mainly visible near $x = c_g(\kappa)t$ instead of $x = \alpha t$, where
$$c_g(\kappa) = c'(\kappa)$$
is the *group velocity*.
- For Lax Wendroff, $c_g(\kappa) = \alpha + 3\mu\kappa^2$, so higher frequency components are dispersed more.
- Numerical errors due to dissipation and dispersion can also be quantified with the amplification factor:
  - Let $G_e(\kappa,\tau) = \rho_e(\kappa,\tau)e^{\imath\phi_e(\kappa,\tau)}$ $(\rho_e, \phi_e \in \mathbb{R})$ denote the single τ step amplification factor for an exact solution with spatial wave number κ.
  - Let $G_c(\kappa,\tau) = \rho_c(\kappa,\tau)e^{\imath\phi_c}(\kappa,\tau)$ $(\rho_c, \phi_c \in \mathbb{R})$ denote the corresponding amplification factor for the computed solution.

  Relative dissipation and dispersion errors are respectively
$$\rho_c(\kappa,\tau)/\rho_e(\kappa,\tau) \quad \text{and} \quad |\phi_c(\kappa,\tau)/\phi_e(\kappa,\tau)|.$$

Relative Dissipation and Dispersion Errors

- Example: With wave number κ in $u_0(x) = e^{\imath\kappa x}$, (30) is solved exactly by $u_e(x,t) = e^{\imath(\kappa x + \omega_e t)}$ iff $\omega_e = -\alpha\kappa$, so $G_e(\kappa,\tau) = e^{-\imath\alpha\kappa\tau}$ and $u(x,t+\tau) = G_e(\kappa,\tau)u(x,t)$.
- For the upwind scheme $U_i^{k+1} = (1-|r|)U_i^k + |r|U_{i-\sigma(a)}^k$ with $r = \alpha\tau/h$, $\sigma(\alpha) = \text{sign}(\alpha)$ and $U_i^k = e^{\imath(\kappa x_i + \omega_e t_k)}$, $G_c(\kappa,\tau) = (1-|r|) + |r|e^{-\imath|\alpha|\kappa\tau/r}$ and $U_i^{k+1} = G_c(\kappa,\tau)U_i^k$.
- In $G_e(\kappa,\tau) = e^{\imath\omega_e\tau}$ and $G_c(\kappa,\tau) = e^{\imath\omega_c\tau}$, $\omega_e(\kappa,\tau)$ and $\omega_c(\kappa,\tau)$ are the so-called exact and computed *dispersion relations*.
- The relative dissipation error is
$$\rho_c(\kappa,\tau)/\rho_e(\kappa,\tau) = \left[1 - 4|r|(1-|r|)\sin^2(\alpha\kappa\tau/(2r))\right]^{\frac{1}{2}}$$
implying some dissipation unless $|r| = 1$.
- The relative dispersion error is
$$|\phi_c(\kappa,\tau)/\phi_e(\kappa,\tau)| = \tan^{-1}\left[\frac{|r|\sin(\kappa\alpha\tau/r)}{(1-|r|+|r|\cos(\kappa\alpha\tau/r))}\right]/(a\kappa\tau)$$
implying no dissipation error for $|r| = \frac{1}{2}, 1$, while there is a phase lag for $|r| < \frac{1}{2}$ and a phase advance for $|r| > \frac{1}{2}$.

$\tan\theta/2 = \frac{\sin\theta}{1+\cos\theta}$

- **Exercise**: Repeat for Lax Friedrichs and Lax Wendroff.

Burger's Equation

- A scalar conservation equation often used to model aspects of fluid flow is *Burger's Equation*

$$u_t + \tfrac{1}{2}(u^2)_x = 0$$

which is deduced from the *weak* (integral) form of solution,

$$\int_{x_1}^{x_2} [u(x,t_2) - u(x,t_1)]dt + \int_{t_1}^{t_2} [f(x_2,t) - f(x_1,t)]dt = 0, \quad \forall [x_1,x_2]\times[t_1,t_2]$$

with flux $f(u) = u^2/2$ of the conserved quantity u.

- When Burger's equation is written in the quasilinear form

$$u_t + uu_x = 0, \quad u(x,0) = u_0(x)$$

the coefficient u can be viewed as the wave speed (cf. α in (30)), and an apparently natural upwind scheme is given by

$$(U_i^{k+1} - U_i^{k+1})/\tau + U_i^k(U_i^k - U_{i-1}^k)/h = 0$$

assuming that $U_i^k \geq 0$. Yet for $u_0(x) = (x < 0)$, we obtain $U_\tau(x) \to u_0(x)$, $\tau \to 0$, which is not a weak solution.

- **Exercise**: Verify these claims.

Conservative Methods for Nonlinear Convection

- To avoid convergence to non-solutions, it is required that numerical methods for solving $u_t + f(u)_x = 0$ can be expressed in *conservative form*,

$$(U_i^{k+1} - U_i^k)/\tau + [F(U^k; i) - F(U^k; i-1)]/h = 0, \quad F(U^k; i) = F(\{U_{i+j}\}_{j=-p}^{q})$$

  where F is the *numerical flux function*.
- The simplest and most frequently used case is ($p = 0$, $q = 1$)

$$(U_i^{k+1} - U_i^k)/\tau + [F(U_i^k, U_{i+1}^k) - F(U_{i-1}^k, U_i^k)]/h = 0$$

- Considering the weak form of the conservation equation on $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [t_k, t_{k+1}]$, conservative form implies naturally that

$$F(U_i^k, U_{i+1}^k) \approx \frac{1}{\tau} \int_{t_k}^{t_{k+1}} f(u(x_{i+\frac{1}{2}}, t))dt$$

- An approach to upwinding for nonlinear fluxes is to use

$$F(U, V) = \begin{cases} f(U), & [f(U) - f(V)]/(U - V) \geq 0 \\ f(V), & [f(U) - f(V)]/(U - V) < 0 \end{cases}$$

  with generlizations for systems to be addressed later.

Consistency for Conservative Methods

- The Lax Friedrichs scheme for nonlinear systems is

  $$[U_i^{k+1} - \tfrac{1}{2}(U_{i-1}^k + U_{i+1}^k)]/\tau + [f(U_{i+1}^k) - f(U_{i-1}^k)]/(2h) = 0$$

  with the corresponding numerical flux function

  $$F(U_i^k, U_{i+1}^k) = \frac{h}{2\tau}[U_i^k - U_{i+1}^k] + \frac{1}{2}[f(U_i^k) + f(U_{i+1}^k)]$$

- A conservative scheme is said to be *consistent* if the numerical flux function is Lipschitz continuous and satisfies $F(u, u, \ldots, u) = f(u)$. (So f is necessarily Lipschitz.)
- **Exercise**: Show that the upwind and Lax Friedrichs schemes are consistent. Also discuss a nonlinear Lax Wendroff scheme.

Theorem: (Lax Wendroff) Choose grid parameters $\tau_l \to 0$ and $h_l \to 0$ as $l \to \infty$. Let $U_l(x,t)$ be given on the l-th grid by a conservative scheme which is consistent with $u_t + f(u)_x = 0$. Suppose U_l converges with bounded variation in L^1_{loc} to u, i.e., $\forall \Omega = (x_1, x_2) \times (t_1, t_2) \subset \mathbb{R} \times [0, \infty)$, $\|U_l - u\|_{L^1(\Omega)} \to 0$, $l \to \infty$, and $\forall T > 0$, $\exists c > 0$ such that $\text{TV}(U_l(\cdot, t)) \leq c$, $\forall t \in [0, T]$, $\forall l$. Then u is a weak solution to the conservation equation.

Lax Wendroff Theorem

Proof: A weak solution u can characterized equivalently by

$$\int_0^{\infty}\int_{-\infty}^{+\infty}[\phi_t u+\phi_x f(u)]dxdt=-\int_{-\infty}^{+\infty}\phi(x,0)u_0(x,0)dx,\quad \forall\phi\in C_0^1(\mathbb{R}\times[0,\infty))$$

For such a ϕ multiply the scheme by $\phi(x_i,t_k)$ to obtain

$$\phi(x_i,t_k)[U_i^{k+1}-U_i^k]/\tau+\phi(x_i,t_k)[F(U^k;i)-F(U^k;i-1)]/h=0$$

and sum over all i and $k\geq 0$,

$$\sum_{k=0}^{\infty}\sum_{i=-\infty}^{+\infty}\phi(x_i,t_k)\frac{U_i^{k+1}-U_i^k}{\tau}+\sum_{k=0}^{\infty}\sum_{i=-\infty}^{+\infty}\phi(x_i,t_k)\frac{F(U^k;i)-F(U^k;i-1)}{h}=0.$$

After summation by parts, i.e.,

$$\sum_{i=1}^{m}a_i(b_i-b_{i-1})=a_mb_m-a_1b_0-\sum_{i=1}^{m-1}(a_{i+1}-a_i)b_i$$

we obtain the following, using the compact support of ϕ, i.e., $\phi(x_i,t_k)=0$ for $|i|$ or k sufficiently large,

Lax Wendroff Theorem

$$-\sum_{i=-\infty}^{+\infty}\phi(x_i,0)U_i^0/\tau-\sum_{k=1}^{\infty}\sum_{i=-\infty}^{+\infty}\frac{\phi(x_i,t_k)-\phi(x_i,t_{k-1})}{\tau}U_i^k$$
$$-\sum_{k=0}^{\infty}\sum_{i=-\infty}^{+\infty}\frac{\phi(x_{i+1},t_k)-\phi(x_i,t_n)}{h}F(U^k;i)=0$$

where each sum is finite, owing to the compact support of ϕ. Multiplying by h and rearranging gives

$$h\tau\sum_{k=1}^{\infty}\sum_{i=-\infty}^{+\infty}\left(\frac{\phi(x_i,t_k)-\phi(x_i,t_{k-1})}{\tau}\right)U_i^k$$
$$+h\tau\sum_{k=0}^{\infty}\sum_{i=-\infty}^{+\infty}\left(\frac{\phi(x_{i+1},t_k)-\phi(x_i,t_k)}{\tau}\right)F(U^k;i)$$
$$=-h\sum_{i=-\infty}^{+\infty}\phi(x_i,0)U_i^0.$$

Due to the convergence in L^1_{loc} of U and the smoothness of ϕ, the first and last terms converge respectively to

Lax Wendroff Theorem

$$\int_0^\infty \int_{-\infty}^{+\infty} \phi_t(x,t)u(x,t)dxdt$$

and

$$-\int_{-\infty}^{+\infty} \phi_t(x,t)u(x,0)dx$$

as $l \to \infty$. For the remaining term in the last equation note

$$|F(U^k;i) - f(U_i^k)| \leq L \max_{-p\leq j\leq q} |U_{i-j}^k - U_i^k| \to 0, \quad \text{a.e.} \quad l \to \infty$$

where the first inequality follows from consistency and the convergence follows from bounded variation. Thus, $F(U^k;i)$ above can be approximated by $f(U_i^k)$ with errors that vanish uniformly a.e. Similarly, due to Lipschitz continuity of f,

$$|f(U_i^k) - f(u_i^k)| \leq L|U_i^k - u_i^k|$$

a difference which converges to 0 in L^1_{loc}. Thus, $F(U^k;i)$ above can actually be approximated by $f(u_i^k)$ with errors that vanish uniformly. The resulting term converges to

$$\int_0^\infty \int_{-\infty}^{+\infty} \phi_x(x,t)f(u(x,t))dxdt$$

as $l \to \infty$.

■

Vanishing Viscosity Solutions

- Weak solutions to conservation equations $u_t + f(u)_x = 0$ are typically not unique.
- A weak solution is regarded as *correct* when it corresponds to a *vanishing viscosity solution*, i.e., $\lim_{\epsilon \to 0} u^\epsilon$ where

$$u_t^\epsilon + f(u^\epsilon)_x = \epsilon u_{xx}$$

- A vanishing viscosity solution can be characterized more conveniently as an *entropy solution*, one which is either smooth or else exhibits shocks with the property that characteristics never emerge from but rather always run into space-time discontinuities.
- Examples: Entropy (vanishing viscosity) solutions to Burger's equation are given by

$$u(x,t) = -\text{sign}(x), \quad u_0(x) = -\text{sign}(x) \quad \text{(shock)}$$

and (figure forthcoming)

$$u(x,t) = \begin{cases} -1, & x \le -t \\ x/t, & -t \le x \le t \\ +1, & t \le x \end{cases} \qquad u_0(x) = \text{sign}(x) \quad \text{(rarefaction)}$$

Entropy Solutions to Conservation Equations

- In the second case above, a weak solution which is not an entropy solution is given by

$$u(x,t) = \text{sign}(x), \quad u_0(x) = \text{sign}(x) \quad \text{(shock)}$$

- **Exercise**: Verify these claims.
- **Exercise**: For initial data $U_i^0 = (i > 0) - (i \leq 0)$, show that the nonlinear upwind scheme converges to the entropy violating shock above.
- **Exercise**: For initial data given by cell averages of $u_0(x) = \text{sign}(x)$, show that with $h = 2\tau$ the nonlinear upwind scheme
  - converges to the correct rarefaction wave for $\tau_l = 1/(2l)$ as $l \to \infty$,
  - converges to the entropy violating shock for $\tau_l = 1/(2l+1)$ as $l \to \infty$,
  - diverges for $\tau_l = 1/l$ as $l \to \infty$.

Roe's Approximate Riemann Solver

- A *Riemann Solver* for a conservation equation $u_t + f(u)_x = 0$ is one in which a weak solution is approximated in the time interval $[t_k, t_{k+1}]$ using the exact solution for piecewise constant data, (Riemann problems)

$$u(x, t_k) = U_i^k, \quad x \in (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$$

- The solution can be computed exactly over a short time interval, and then cell averages may be recomputed.
- This is the essence of *Godunov's Method*, but exact Riemann solvers can be expensive.
- The idea of Roe's Approximate Riemann Solver is to solve a constant coefficient linear hyperbolic system at every cell interface,

$$u_t + A(u_\mathrm{l}, u_\mathrm{r})u_x = 0$$

where the matrix $A(u_\mathrm{l}, u_\mathrm{r})$ depends upon left and right states u_l and u_r at the interface.

Roe's Approximate Riemann Solver

- Roe requires the following conditions on $A(u_l, u_r)$:
  1. $A(u_l, u_r)(u_l - u_r) = f(u_l) - f(u_r)$,
  2. $A(u_l, u_r)$ is diagonalizable with real eigenvalues,
  3. $A(u_l, u_r) \to f'(u)$, $u_l, u_r \to u$.
- The associated numerical flux function is given by

$$F(u_l, u_r) = [f(u_r) + f(u_l)]/2 - |A(u_l, u_r)|(u_r - u_l)/2$$

$$\begin{gathered} \text{where} \quad A^+ = S\Lambda^+ S^{-1}, \quad A^- = S\Lambda^- S^{-1}, \\ |A| = A^+ - A^- = S|\Lambda|S^{-1}, \quad |\Lambda| = \Lambda^+ - \Lambda^- \\ \text{and} \quad \Lambda^+ = \max\{\Lambda, 0\}, \quad \Lambda^- = \min\{\Lambda, 0\}. \end{gathered}$$

- In some cases one may take $A(u_l, u_r) = f'(u_{\text{ave}})$, where u_{ave} is a certain average of u_l and u_r, difficult to implement in practice. Yet, a Roe matrix can be shown to exist for certain applications, e.g., scalar equations and Euler equations.
- For a scalar conservation equation, $u_t + f(u)_x = 0$, $A(u_l, u_l) = \alpha(u_l, u_l)$ is given uniquely by

$$\alpha(u_l, u_l) = [f(u_r) - f(u_l)]/[u_r - u_l]$$

Code to Solve Burger's Equation

- Matlab code to solve Burger's Equation with Roe's scheme:

```
N1 = N+1; N2 = N+2; dx = 1/N1;
dt = dx/10; K = round(1/dt); T = K*dt; r = dt/dx;
x  = linspace(-1,+1,N2); u = 0.5*sign(x); uv = u;
for k=1:K
   f  = u.^2/2;
   ul = u(1:N1); fl = f(1:N1);
   ur = u(2:N2); fr = f(2:N2);
   df = fr-fl; du = ur-ul;
   a  = df.*(du ~= 0)./(du + (du == 0));
   F  = (fl + fr)/2 - abs(a).*du/2;
   dF = [0,F(2:N1)-F(1:N),0];
   u  = u - r*dF;
   u(1)  = u(1)  - r*(a(1)  < 0)*df(1);
   u(N2) = u(N2) - r*(a(N1) > 0)*df(N1);
   plot(x,u); uv = [uv;u];
end
contour(uv,'Fill','on');
```

Roe Matrix for Isothermal Flow

- The Euler equations for inviscid flow in one spatial dimension (e.g., in a pipe) are given by:

$$\begin{aligned} \text{mass conservation:} \quad & \rho_t + (\rho v)_x = 0 \\ \text{momentum conservation:} \quad & (\rho v)_t + (\rho v^2 + p)_x = 0 \\ \text{energy conservation:} \quad & E_t + (v(E+p))_x = 0 \end{aligned}$$

  where the pressure is determined by an equation of state

$$p = p(\rho, v, E).$$

- If the temperature is constant, the energy E can be determined from the density ρ and the velocity v without solving the energy conservation equation. The equation of state reduces to $p = a^2\rho$ where a is the sound speed. The Euler equations reduce to

$$u_t + f(u)_x = 0, \quad u = \begin{bmatrix} \rho \\ \rho v \end{bmatrix}, \quad f(u) = \begin{bmatrix} \rho v \\ \rho v^2 + a^2\rho \end{bmatrix}$$

- **Exercise**: Use the following Roe matrix to solve this system.

$$A(u_l, u_r) = \begin{bmatrix} 0 & 1 \\ a^2 - \bar{v}^2 & 2\bar{v} \end{bmatrix}, \quad \bar{v} = \frac{\rho_l^{1/2} v_l + \rho_r^{1/2} v_r}{\rho_l^{1/2} + \rho_r^{1/2}}$$

Convection Diffusion Equation

- Previous techniques for conservation equations are now applied to solve the convection diffusion PDE in $\Omega = (0,1)^2$,

$$\begin{cases} \partial_t u + \nabla \cdot (\alpha u) &= \nabla \cdot (\kappa \nabla u) + \phi & \text{in} & \quad \Omega \times (0,T] \\ u &= g_1, & \text{on} & \quad \partial\Omega_1 \times [0,T] \\ \partial_n u &= g_2, & \text{on} & \quad \partial\Omega_2 \times [0,T] \\ u &= u_0, & \text{in} & \quad \Omega \times \{0\} \end{cases}$$

  where $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$, $\alpha : \Omega \to \mathbb{R}^2$ and $\kappa : \Omega \to \mathbb{R}_+$.
- For compatibility it must be that $u_0|_{\partial\Omega_1} = g_1$ and $\partial_n u_0|_{\partial\Omega_2} = g_2$ hold.
- Without loss of generality we may assume that $g_2 = 0$. Otherwise, we could shift this boundary condition to the interior as follows:
  - Define a sufficiently smooth function v satisfying $\partial_n v|_{\partial\Omega_2} = g_2$, e.g., $v = u_0$ if g_2 is independent of t.
  - Solve the convection diffusion PDE for $w = u - v$ replacing (a) ϕ with $\phi + \nabla \cdot (\kappa \nabla v) - \nabla \cdot (\alpha v) - \partial_t v$, (b) g_1 with $g_1 - v$ and (c) u_0 with $u_0 - v$.
  - Take $u = w + v$ to solve the original problem.

Convection Diffusion Equation

- For $N \in \mathbb{N}$ and $h = 1/(N+1)$ set

$$\bar{\Omega}_h = \{(x, y) : x = ih, y = jh, 0 \le i, j \le N+1\}$$

and for $K \in \mathbb{N}$ and $\tau = T/K$ set $t_k = k\tau$, $0 \le k \le K$.

- The exact solution $u_{ij}^k = u(x_i, y_j, t_k)$ to the PDE is approximated with $U_{ij}^k \approx u_{ij}^k$ written as $U_h^k = \{U_{ij}^k\}_{i,j=0}^{N+1}$.
- The diffusivity is evaluated at midpoints $(x + \frac{1}{2}ih, y + \frac{1}{2}jh)$, $|i| + |j| = 1$, between grid points $(x, y) \in \Omega_h$,

$$\kappa_{h,x} = \left\{\kappa(x_i + \tfrac{1}{2}ih, y_j)\right\}_{i=0,j=0}^{N,N+1}, \quad \kappa_{h,y} = \left\{\kappa(x_i, y_j + \tfrac{1}{2}jh)\right\}_{i=0,j=0}^{N+1,N}$$

- Let $\mathcal{D}(V)$ denote a diagonal matrix with the values of the vector V along the diagonal.
- Using $D_{h,x}$ and $D_{h,y}$ for Neumann boundary conditions from [76], the diffusion term is approximated by

$$-\nabla \cdot (\kappa \nabla u) \approx A_{h,\kappa} U, \quad A_{h,\kappa} = D_{h,x}^\top \mathcal{D}(\kappa_{h,x}) D_{h,x} + D_{h,y}^\top \mathcal{D}(\kappa_{h,y}) D_{h,y}$$

Convection Diffusion Equation

- For $\alpha = (\alpha_1, \alpha_2)$ set $f(u) = \alpha_1 u$ and $g(u) = \alpha_2 u$ or $\alpha u = (f(u), g(u))$. The convection is approximated using upwinding, $(u_t + f(u)_x + g(u)_y)_{i,j} \approx$

$$\frac{U_{i,j}^{k+1} - U_{i,j}^{k}}{\tau} + \frac{F_{i+\frac{1}{2},j}^{k} - F_{i-\frac{1}{2},j}^{k}}{h} + \frac{G_{i,j+\frac{1}{2}}^{k} - G_{i,j-\frac{1}{2}}^{k}}{h}$$

where the numerical flux functions for f and g are given respectively by

$$F_{i+\frac{1}{2},j}^{k} = \frac{f_{i+1,j}^{k} + f_{i,j}^{k}}{2} - |a_{i+\frac{1}{2},j}^{k}| \frac{U_{i+\frac{1}{2},j}^{k} - U_{i-\frac{1}{2},j}^{k}}{2}, \quad a_{i+\frac{1}{2},j} = \frac{f_{i+1,j}^{k} - f_{i,j}^{k}}{U_{i+\frac{1}{2},j}^{k} - U_{i-\frac{1}{2},j}^{k}}$$

$$G_{i,j+\frac{1}{2}}^{k} = \frac{g_{i,j+1}^{k} + g_{i,j}^{k}}{2} - |a_{i,j+\frac{1}{2}}^{k}| \frac{U_{i,j+\frac{1}{2}}^{k} - U_{i,j-\frac{1}{2}}^{k}}{2}, \quad a_{i,j+\frac{1}{2}} = \frac{g_{i,j+1}^{k} - g_{i,j}^{k}}{U_{i,j+\frac{1}{2}}^{k} - U_{i,j-\frac{1}{2}}^{k}}$$

Convection Diffusion Equation

- Define the numerical flux vectors

$$\delta F_h^k = \begin{cases} F_{i+\frac{1}{2},j} - F_{i-\frac{1}{2},j}, & 0 < i < N+1 \\ (a_{\frac{1}{2},j} < 0)(f_{1,j} - f_{0,j}), & i = 0 \\ (a_{N+\frac{1}{2},j} > 0)(f_{N+1,j} - f_{N,j}), & i = N+1 \end{cases}$$

$$\delta G_h^k = \begin{cases} G_{i,j+\frac{1}{2}} - G_{i,j-\frac{1}{2}}, & 0 < j < N+1 \\ (a_{i,\frac{1}{2}} < 0)(g_{i,1} - g_{i,0}), & j = 0 \\ (a_{i,N+\frac{1}{2}} > 0)(g_{i,N+1} - g_{i,N}), & j = N+1 \end{cases}$$

- Assume $U_h^k = g_1^k$ holds on $\partial\Omega_{1,h}$.
- Let $\tilde{U}_h^{k+1}$ be chosen to satisfy $\tilde{U}_h^{k+1} = g_1^{k+1}$ on $\partial\Omega_{1,h}$.
- Define $\chi_h(x,y) = ((x,y) \notin \partial\Omega_{1,h})$ as the characteristic function for $\bar{\Omega}_h \backslash \partial\Omega_{1,h}$.
- Apply the semi-implicit scheme with $r = \tau/h$,

$$[I + \tau\mathcal{D}(\chi_h)A_{h,\kappa}]U_h^{k+1} = \mathcal{D}(\chi_h)[U_h^k - r\delta F_h^k - r\delta G_h^k] + \mathcal{D}(1-\chi_h)\tilde{U}_h^{k+1}$$

- Note that $U_h^{k+1} = g_1^{k+1}$ necessarily holds on $\partial\Omega_{1,h}$.

Code to Solve Convection Diffusion PDE

```
N1 = N+1; N2 = N+2; N1N2 = N1*N2; N2N2 = N2*N2;
dx = 1/N1; dt = dx/10; K = round(1/dt); T = K*dt;
r  = dt/dx; y = linspace(0,1,N2); x = y'; dy = dx;
ax = 1.0e+0*ones(N2,N2); ay = 1.0e+0*ones(N2,N2);
kx = 1.0e-2*ones(N1,N2); ky = 1.0e-2*ones(N2,N1);
u = [ones(1,N2);zeros(N1,N2)]; u0 = u(:); U = u0;
D0 = speye(N2); bcs = 1:N2:N2N2;
D1 = spdiags(kron([-1,1],ones(N2,1)),[0,1],N1,N2);
Dx = kron(D0,D1)/dx; Dy = kron(D1,D0)/dy;
A  = Dx'*spdiags(kx(:),0,N1N2,N1N2)*Dx ...
   + Dy'*spdiags(ky(:),0,N1N2,N1N2)*Dy; A(bcs,:)=0;
for k=1:K
  ul = u(1:N1,:); ur = u(2:N2,:); du = ur - ul;
  f  = ax.*u;
  fl = f(1:N1,:); fr = f(2:N2,:); df = fr - fl;
  a  = df.*(du ~= 0)./(du + (du == 0));
  F  = (fl + fr)/2 - abs(a).*du/2;
```

Code to Solve Convection Diffusion PDE

```
    dF = [(a(1,:)<0).*df(1,:); ...
           F(2:N1,:)-F(1:N,:); ...
          (a(N1,:)>0).*df(N1,:)]; dF = dF(:);
    ul = u(:,1:N1); ur = u(:,2:N2); du = ur - ul;
    g  = ay.*u;
    gl = g(:,1:N1); gr = g(:,2:N2); dg = gr - gl;
    a  = dg.*(du ~= 0)./(du + (du == 0));
    G  = (gl + gr)/2 - abs(a).*du/2;
    dG = [(a(:,1)<0).*dg(:,1), ...
           G(:,2:N1)-G(:,1:N), ...
          (a(:,N1)>0).*dg(:,N1)]; dG = dG(:);
    dF(bcs) = 0; dG(bcs) = 0; dU = U - r*dF - r*dG;
    U  = (speye(N2N2) + dt*A) \ dU;
    u  = reshape(U,N2,N2); surf(x,y,u);
end
```

Exercise: Implement convection as above and also with central differences and compare.

Function Spaces for Elliptic PDEs

- Let $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$, be open.
- The *Lebesgue Spaces* $L^p(\Omega)$ are Banach spaces consisting of (equivalence classes of) Lebesgue measurable functions (differing at most on a set of measure 0) with finite norm,

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p dx \right)^{\frac{1}{p}}, \quad 1 \leq p < \infty, \quad \|f\|_{L^\infty(\Omega)} = \operatorname*{esssup}_{x \in \Omega} |f(x)|$$

- These satisfy *Hölder's Inequality*,

$$\|fg\|_{L^1(\Omega)} \leq \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)}, \quad p^{-1} + q^{-1} = 1, \quad p \geq q.$$

  So for bounded Ω it follows that $L^p(\Omega) \hookrightarrow L^q(\Omega)$, $p \geq q$.
- Also, $L^1_{\text{loc}}(\Omega) = \cap_{K \subset\subset \Omega} L^1(K)$.
- For $p = 2$, $L^2(\Omega)$ is a Hilbert space with the scalar product,

$$(f, g)_{L^2(\Omega)} = \int_\Omega f(x) g(x) dx$$

  and for $p = q = 2$, Hölder's Inequality gives the Cauchy-Schwarz Inequality.

Lebesgue and Hölder Spaces

- The *Hölder Spaces* $\mathcal{C}^k(\Omega)$ are Banach spaces consisting of functions whose partial derivatives $\partial^\alpha u$, $|\alpha| \leq k$, are continuous on Ω with finite norm,

$$\|u\|_{\mathcal{C}^k(\Omega)} = \sum_{|\alpha| \leq k} \sup_{x \in \Omega} |\partial^\alpha u(x)|$$

- $\mathcal{C}_0^k(\Omega)$ is the subspace of $\mathcal{C}^k(\Omega)$ consisting of functions with support $\bar{K}_u$, $K_u = \{x \in \Omega : u(x) \neq 0\}$, satisfying $\bar{K}_u \subset\subset \Omega$.
- $\mathcal{C}^k(\bar{\Omega})$ is the restriction to Ω of functions in $\mathcal{C}_0^k(\mathbb{R}^n)$.
- Also $\mathcal{C}_0^\infty(\Omega) = \cap_{k \geq 0} \mathcal{C}_0^k(\Omega)$.
- A function $u \in L^1_{\text{loc}}(\Omega)$ has an α-th *weak derivative* $v \in L^1_{\text{loc}}(\Omega)$ in Ω, written $v = \partial^\alpha u$, if

$$\int_\Omega u(x) \partial^\alpha \phi(x) dx = (-1)^{|\alpha|} \int_\Omega v(x) \phi(x) dx, \quad \forall \phi \in \mathcal{C}_0^k(\Omega)$$

- Example: $u(x) = |x|$ has the weak derivative $v(x) = \text{sign}(x)$ in $\Omega = (-1, +1)$, but v does not have a weak derivative in Ω.

Weak Derivatives and Sobolev Spaces

- The *Sobolev Space* $W^{k,p}(\Omega)$, $k \in \mathbb{N}_0$, $1 \leq p \leq \infty$, are Banach Spaces consisting of functions whose weak derivatives $\partial^\alpha u$, $|\alpha| \leq k$, are in $L^p(\Omega)$ with finite norm,

$$\|u\|_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha| \leq k} \|\partial^\alpha u\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}}, \quad \|u\|_{W^{k,\infty}(\Omega)} = \sum_{|\alpha| \leq k} \|\partial^\alpha u\|^\infty_{L^p(\Omega)}$$

- One also uses the semi-norms,

$$|u|_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha| = k} \|\partial^\alpha u\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}}, \quad |u|_{W^{k,\infty}(\Omega)} = \sum_{|\alpha| = k} \|\partial^\alpha u\|^\infty_{L^p(\Omega)}$$

- An alternative definition of the Sobolev Space $H^{k,p}(\Omega)$, $k \in \mathbb{N}_0$, $1 \leq p < \infty$, is given by the completion of $\mathcal{C}^\infty(\bar{\Omega})$ with respect to the $W^{k,p}(\Omega)$ norm above.
- Yet for $\partial\Omega$ sufficiently smooth, $W^{k,p}(\Omega) = H^{k,p}(\Omega)$!
- Here it will always be assumed that $\partial\Omega$ is Lipschitz continuous, which is sufficient to insure that:
- $\mathcal{C}^\infty(\bar{\Omega})$ is dense (in $H^{k,p}(\Omega)$ by definition and hence) in $W^{k,p}(\Omega)$ for $k \in \mathbb{N}_0$, $1 \leq p < \infty$.

Sobolev Embeddings

- For $1 \leq p, q < \infty$ and $\Omega \subset \mathbb{R}^n$ open and bounded with Lipschitz $\partial\Omega$, the following embeddings are continuous

$$W^{k,p}(\Omega) \hookrightarrow \begin{cases} L^q(\Omega), & p < n/k, \quad p \leq q \leq np/(n-p) \\ L^q(\Omega), & p = n/k, \quad p \leq q < np/(n-p) \\ C^0(\bar{\Omega}), & p > n/k \end{cases}$$

- and the following embeddings are compact

$$W^{k,p}(\Omega) \hookrightarrow \begin{cases} L^q(\Omega), & p \leq n/k, \quad 1 \leq q < (n-pk)/(np) \\ C^0(\bar{\Omega}), & p > n/k \end{cases}$$

  In particular, the embedding $W^{k,p}(\Omega) \hookrightarrow W^{k-1,p}(\Omega)$ is compact $\forall k \in \mathbb{N}$, $\forall p \in [1, \infty]$.
- For $\Omega_i \subset \mathbb{R}^n$, $i = 0, \ldots, N$, $N \in \mathbb{N}$, open and bounded with Lipschitz $\partial\Omega_i$ such that $\Omega_i \cap \Omega_j = \emptyset$, $\Omega = \Omega_0 = \cup_{i=1}^{N} \Omega$, the following holds $\forall k \in \mathbb{N}$, $\forall p \in [1, \infty]$,

$$\{v \in C^{k-1}(\bar{\Omega}) : v|_{\Omega_i} \in C^k(\Omega_i), 1 \leq i \leq N\} \subset W^{k,p}(\Omega)$$

Traces

- Let $\Omega \subset \mathbb{R}^n$ be open and bounded with Lipschitz $\partial\Omega$. Set $k \in \mathbb{N}_0$, $1 \leq p < \infty$, $kp < n$ and $1 \leq q \leq (n-1)p/(n-kp)$. Let $\tilde{T}$ be a linear operator satisfying $\tilde{T}\phi = \phi|_{\partial\Omega}$ for $\phi \in C^k(\bar{\Omega})$.
- Then $\tilde{T}$ has a unique extention to a bounded linear operator $T : W^{k,p}(\Omega) \to L^q(\partial\Omega)$, and $\exists c = c(p,\Omega) > 0$ such that

$$\|Tf\|_{L^q(\partial\Omega)} \leq c\|f\|_{W^{k,p}(\Omega)}, \quad \forall f \in W^{k,p}(\Omega)$$

- If $kp = n$ holds, then the estimate holds for any $p \leq q < \infty$.
- Thus, one defines the subspace,

$$W_0^{k,p}(\Omega) = \{f \in W^{k,p}(\Omega) : T\partial^\alpha f = 0 \in L^q(\partial\Omega), |\alpha| \leq k-1\}$$

  or equivalently $H_0^{k,p}(\Omega)$ $(= W_0^{k,p}(\Omega))$ as the completion of $C_0^\infty(\Omega)$ with respect to the $W^{k,p}(\Omega)$ norm.
- Thus $C_0^\infty(\Omega)$ is dense in the subspace.
- For the following let P_k denote the set of polynomials of degree at most k.

Poincaré's Inequality

Theorem (Poincaré): Let $\Omega \subset \mathbb{R}^n$ be open and bounded with Lipschitz $\partial\Omega$. Let $k \in \mathbb{N}$ and $1 \leq p < \infty$. Then $\exists c = c(\Omega) > 0$ such that

$$\|f\|_{W^{k,p}(\Omega)} \leq c|f|_{W^{k,p}(\Omega)}, \quad \forall f \in W_0^{k,p}(\Omega)$$

Proof: Assume for the sake of contradiction, $\exists\{\tilde{f}_i\} \subset W_0^{k,p}(\Omega)$ such that $|\tilde{f}_i|_{W^{k,p}(\Omega)}/\|\tilde{f}_i\|_{W^{k,p}(\Omega)} \overset{i\to\infty}{\longrightarrow} 0$. Define $f_i = \tilde{f}_i/\|\tilde{f}_i\|_{W^{k,p}(\Omega)}$ so that $\|f_i\|_{W^{k,p}(\Omega)} \overset{\forall i}{=} 1$ while $|f_i|_{W^{k,p}(\Omega)} \overset{i\to\infty}{\longrightarrow} 0$. Since $\{f_i\}$ is bounded in $W^{k,p}(\Omega)$ and $W^{k,p}(\Omega)$ is compactly embedded in $W^{k-1,p}(\Omega)$, there is a subsequence $\{f_{i'}\}$ which converges strongly in $W_0^{k-1,p}(\Omega)$ to a limit $f^* \in W_0^{k-1,p}(\Omega)$. Yet since $|f_{i'}|_{W^{k,p}(\Omega)} \overset{i'\to\infty}{\longrightarrow} 0$, the subsequence $\{f_{i'}\}$ is actually Cauchy in $W_0^{k,p}(\Omega)$, meaning that $f^* \in W_0^{k,p}(\Omega)$. Since $|f^*|_{W^{k,p}(\Omega)} = \lim_{i'\to\infty} |f_{i'}|_{W^{k,p}(\Omega)} = 0$, it must be that $\partial^\alpha f^* = 0$, $\forall\alpha$ with $|\alpha| = k$, or that $f^* \in P_{k-1}$. Yet since $T\partial^\alpha f^* = 0$ on $\partial\Omega$ for $|\alpha| \leq k-1$, it must be that $f^* = 0$. However, this contradicts $\|f^*\|_{W^{k,p}(\Omega)} = \lim_{i'\to\infty} \|f_{i'}\|_{W^{k,p}(\Omega)} = 1$. ■

Poincaré's Inequality

- One writes $H^k(\Omega) = W^{k,2}(\Omega)$, $H_0^k(\Omega) = W_0^{k,2}(\Omega)$, $k \in \mathbb{N}$. These are equipped respectively with the scalar products,

$$(f,g)_{H^k(\Omega)} = \sum_{|\alpha|\le k} (\partial^\alpha f, \partial^\alpha g)_{L^2(\Omega)}, \quad (f,g)_{H_0^k(\Omega)} = \sum_{|\alpha|=k} (\partial^\alpha f, \partial^\alpha g)_{L^2(\Omega)}$$

and, based upon Poincaré's inequality, the associated norms

$$\|f\|_{H^k(\Omega)} = (f,f)^{\frac{1}{2}}_{H^k(\Omega)}, \quad \|f\|_{H_0^k(\Omega)} = |f|_{H^k(\Omega)} = (f,f)^{\frac{1}{2}}_{H_0^k(\Omega)}$$

- **Exercise**: Show that the scalar product $(\nabla u, \nabla v)_{L^2(\Omega)^n} + \langle u, v\rangle_{L^2(\partial\Omega)}$ induces a norm on $H^1(\Omega)$.
- The dual of $H_0^k(\Omega)$ (i.e., the bounded linear functionals on $H_0^k(\Omega)$) is denoted by $H^{-k}(\Omega) = (H_0^k(\Omega))^*$ equipped with the (operator) norm,

$$\|f\|_{H^{-k}(\Omega)} = \sup_{\phi \in H_0^k(\Omega), \phi \neq 0} \frac{f(\phi)}{\|\phi\|_{H_0^k(\Omega)}}$$

where $f(\phi) = (f,\phi)_{H^{-k}(\Omega), H_0^k(\Omega)}$ is the *duality pairing*.

- Note $(H_0^k(\Omega))^* = \text{span}\{\partial^\alpha T_f : |\alpha| \le k, T \in C_0^\infty(\Omega)^*, f \in L^2(\Omega)\}$, where f is an actual function. Not so for $(H^k(\Omega))^*$.

Weak Formulation of Elliptic BVPs

- The *strong form* (implicitly requiring strong regularity) of a typical elliptic PDE is

$$\begin{cases} Lu &= f, & \text{in } \Omega \\ Bu &= g, & \text{on } \partial\Omega \end{cases} \tag{34}$$

where $\Omega \subset \mathbb{R}^n$ is open and bounded with Lipschitz $\partial\Omega$ and

$$Lu(x) = -\nabla \cdot [A(x)\nabla u(x)] + c^\top(x)\nabla u(x) + r(x)u(x)$$

for given $A = \{a_{ij}\}_{i,j=1}^n$, $c = \{b_i\}_{i=1}^n$, r, f and g, sufficiently smooth functions on $\bar{\Omega}$.

- Also, typical boundary conditions are given by

$$Bu(x) = \sigma_1(x) n^\top A(x)\nabla u(x) + \sigma_0(x)u(x)$$

for given σ_1 and σ_0, sufficiently smooth functions on $\partial\Omega$.

- Here $\sigma_1 = 0$ gives pure Dirichlet BCs, $\sigma_0 = 0$ gives pure Neumann BCs, and $\sigma_1, \sigma_0 > 0$ gives Robin BCs.

Weak Formulation of Elliptic BVPs

- The problem is *elliptic* if $\exists \alpha > 0$ such that

$$\xi^\top A(x)\xi \geq \alpha \xi^\top \xi, \quad \forall \xi \in \mathbb{R}^n, \quad \forall x \in \Omega.$$

- Assuming all functions are sufficiently smooth, we multiply the PDE by a sufficiently smooth *test function* v and integrate by parts to obtain

$$\begin{aligned}(A\nabla u, \nabla v)_{L^2(\Omega)} + (c^\top \nabla u, v)_{L^2(\Omega)} + (ru, v)_{L^2(\Omega)} \\ = (f, v)_{L^2(\Omega)} + \langle n^\top A \nabla u, v \rangle_{L^2(\partial\Omega)}\end{aligned}$$

  where n is an outwardly directed normal vector.
- Note that this weak formulation requires considerably less regularity of coefficients and data than the strong formulation.
- Define the functional on $H^1(\Omega) \times H^1(\Omega)$ from the weak formulation above,

$$a(u, v) = (A\nabla u, \nabla v)_{L^2(\Omega)} + (c^\top \nabla u, v)_{L^2(\Omega)} + (ru, v)_{L^2(\Omega)}$$

  which can readily be shown to be bilinear.

Weak Formulation of Elliptic BVPs

- Dirichlet BCs: For simplicity, assume first that $g = 0$. The weak formulation of the BVP, with $Bu = u = 0$ in the trace sense, is given as follows. Find $u \in H_0^1(\Omega)$ such that
$$a(u, v) = b(v), \quad \forall v \in H_0^1(\Omega)$$
where $b(v) = (f, v)$. Note that the above term $\langle n^\top A \nabla u, v \rangle$ vanishes since $Tv = 0$.
- Dirichlet BCs: For the case that $g \neq 0$, let $u_g \in H^1(\Omega)$ be chosen to satisfy $Tu_g = g$. (Existence of u_g follows when g and $\partial\Omega$ have enough regularity.) Then seek a function $\tilde{u}$ $(= u - u_g) \in H_0^1(\Omega)$ such that
$$a(\tilde{u}, v) = b(v), \quad \forall v \in H_0^1(\Omega)$$
where $b(v) = (f, v) - a(u_g, v)$ and set $u = \tilde{u} + u_g$.
- Neumann BCs: The weak formulation of the BVP, with $Bu = n^\top A \nabla u = g$, is given as follows. Find $u \in H^1(\Omega)$ such that
$$a(u, v) = b(v), \quad \forall v \in H^1(\Omega)$$
where $b(v) = (f, v) + \langle g, v \rangle$. Note that the term $\langle n^\top A \nabla u, v \rangle$ has been replace by $\langle g, v \rangle$.

Lax Milgram Theorem

- Robin BCs: The weak formulation of the BVP, with $Bu = n^\top A\nabla u + \sigma_0 u = g$, is given as follows. Find $u \in H^1(\Omega)$ such that with $\tilde{a}(u,v) = a(u,v) + \langle \sigma_0 u, v\rangle$ and $b(v) = (f,v) + \langle g, v\rangle$,
$$\tilde{a}(u,v) = b(v), \quad \forall v \in H^1(\Omega)$$
- These problems can be shown with the following theorem to have possess a unique solution.

Theorem (Lax Milgram): Let a Hilbert space H, a bilinear form $a : H \times H \to \mathbb{R}$ and a linear functional $b : H \to \mathbb{R}$ be given satisfying the conditions of cercivity,
$$\exists c_1 > 0 : \quad c_1 \|v\|_H^2 \leq a(v,v), \quad \forall v \in H$$
and continuity
$$\exists c_2, c_3 > 0 : \quad a(u,v) \leq c_2 \|u\|_H \|v\|_H, \quad b(v) \leq c_3 \|v\|_H, \quad \forall u, v \in H.$$
Then $\exists! u \in H$ such that
$$a(u,v) = b(v), \quad \forall v \in H$$
and
$$c_1 \|u\|_H \leq \sup_{v \in H} |b(v)| / \|v\|_H. \qquad \blacksquare$$

Lax Milgram Theorem

- In particular, we have the following results for the previously outlined BVPs.

Theorem: Suppose $a_{ij}, c_i, r \in L^\infty(\Omega)$ and that $A = \{a_{ij}\}$ satisfies the ellipticity condition with $\alpha > 0$. Suppose $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ and set $\beta = \alpha^{-1} \sum_i \|c_i\|^2_{L^\infty(\Omega)}$. Then

1. Dirichlet BCs: For $g = 0$, $\exists! u \in H^1_0(\Omega)$ solving the BVP if $r(x) \geq \beta/2$, a.e. $x \in \Omega$. Also, $\exists C > 0$ s.t. $\|u\|_{H^1(\Omega)} \leq C\|f\|_{L^2(\Omega)}$. If $g \neq 0$, $\exists! u \in H^1(\Omega)$ solving the BVP and $\exists C > 0$ s.t.
$$\|u\|_{H^1(\Omega)} \leq C[\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}].$$
2. Neumann BCs: $\exists! u \in H^1(\Omega)$ solving the BVP if $r(x) \geq \beta/2$, a.e. $x \in \Omega$. Also, $\exists C > 0$ s.t.
$$\|u\|_{H^1(\Omega)} \leq C[\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}].$$
3. Robin BCs: $\exists! u \in H^1(\Omega)$ solving the BVP if $r(x) \geq \beta/2$, a.e. $x \in \Omega$ and if $\sigma_0(x) \geq 0$, a.e. $x \in \partial\Omega$. Also, $\exists C > 0$ s.t.
$$\|u\|_{H^1(\Omega)} \leq C[\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}].$$

Regularity of Weak Solutions to Elliptic BVPs

Proof: **Exercise** with analysis estimates. ■

Theorem (higher regularity): For $k \in \mathbb{N}$ let $\Omega \subset \mathbb{R}^n$, $\partial\Omega \in \mathcal{C}^{k+1}(\mathbb{R}^{n-1})$, $a_{ij} \in \mathcal{C}^k(\bar{\Omega})$, $c_i, r \in W^{k,\infty}(\Omega)$ and $f \in H^k(\Omega)$. Then the solution to the Dirichlet BVP ($g = 0$) satisfies $u \in H^{k+2}(\Omega) \cap H_0^1(\Omega)$ and $\exists C > 0$ s.t.

$$\|u\|_{H^{k+2}(\Omega)} \leq C[\|f\|_{H^k(\Omega)} + \|u\|_{H^1(\Omega)}].$$

Theorem (higher regularity): Let Ω be a convex polygon in $\mathbb{R}^2$ or a parallelpiped in $\mathbb{R}^3$. Suppose $a_{ij} \in \mathcal{C}^1(\bar{\Omega})$ and $c_i, r \in \mathcal{C}^0(\bar{\Omega})$ and $f \in L^2(\Omega)$. Then the solution to the Dirichlet BVP ($g = 0$) satisfies $u \in H^2(\Omega) \cap H_0^1(\Omega)$ and $\exists C > 0$ s.t.

$$\|u\|_{H^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}.$$

Note: The last theorem is particularly applicable for the problems posed earlier on a square.

Function Spaces for Evolution Equations

- Let $T > 0$ and suppose B is a Banach space equipped with the norm $\|\cdot\|_B$. Typically, B will be a Hilbert space such as $H = L^2(\Omega)$ or $V = H_0^1(\Omega)$, where $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$, is open.
- For $k \in \mathbb{N}$ define the *Hölder Space* $\mathcal{C}^k(0, T; B)$ as the space of B-valued functions which are k times continuously differentiable for $t \in [0, T]$. Then $\mathcal{C}^k(0, T; B)$ is a Banach space equipped with the norm,

$$\|u\|_{\mathcal{C}^k(0,t;B)} = \sum_{j=1}^{k} \sup_{t\in[0,T]} \|\partial_t^j u(t)\|_B$$

- For $1 \leq p \leq \infty$ define the *Bochner Space* $L^p(0, t; B)$ as the space of B-valued functions $u(t)$ for which $t \mapsto \|u(t)\|_B$ is in $L^p(0, T)$. Then $L^p(0, t; B)$ is a Banach space equipped with the norm,

$$\|u\|_{L^p(0,T;B)} = \left(\int_0^T \|u(t)\|_B^p dt \right)^{\frac{1}{p}}, \quad \|u\|_{L^\infty(0,T;B)} = \operatorname*{esssup}_{t\in[0,T]} \|u(t)\|_B$$

Function Spaces for Evolution Equations

- For $1 \leq p \leq \infty$ and $k \in \mathbb{N}$ define the *Sobolev Space* $W^{k,p}(0,T;B)$ as the space of B-valued functions $u(t)$ with weak derivatives $\partial_t^j u(t)$, $1 \leq j \leq k$, for which $t \mapsto \|\partial_t^j u(t)\|_B$ is in $L^p(0,T)$. Then $W^{k,p}(0,T;B)$ is a Banach space equipped with the norm,
$$\|u\|_{W^{k,p}(0,T;B)} = \sum_{j=1}^{k} \|\partial_t^j u(t)\|_{L^p(0,T;B)}$$
- More generally, for $1 < p < \infty$ and two reflexive Banach spaces B_0, B_1 with continuous embedding $B_0 \hookrightarrow B$ and $1/p + 1/q = 1$, set
$$W^{1,p}(B_0,B_1) = \{v \in L^p(0,T;B_0) : \partial_t v \in L^q(0,T;B_1)\},$$
which is a Banach spaced equipped with the norm,
$$\|u\|_{W^{1,p}(B_0,B_1)} = \|u\|_{L^p(0,T;B_0)} + \|\partial_t u\|_{L^q(0,T;B_1)}$$
- Let V be a reflexive Banach space, e.g., $V = H_0^1(\Omega)$, which is continuously and densely embedded into a Hilbert space H, e.g., $H = L^2(\Omega)$. Identify H^* with H using the Riesz

Function Spaces for Evolution Equations

representation theorem. Then H^* is continuously and densely embedded into V^*, e.g., $H^{-1}(\Omega)$. Summarized,

$$V \hookrightarrow H \equiv H^* \hookrightarrow V^*$$

and (V, H, V^*) is a *Gelfand triple*.

- Let $1 < p < \infty$ and suppose (V, H, V^*) is a Gelfand triple. Then the embedding

$$W^{1,p}(V, V^*) \hookrightarrow C(0, T; H)$$

is continuous. Thus, $u \in W^{1,p}(V, V^*)$ has well-defined traces $u(0), u(T) \in H$.

- If (V, H, V^*) is an Gelfand triple, then $\forall u, v \in W^{1,p}(V, V^*)$,

$$\partial_t (u(t), v(t))_H = \langle \partial_t u(t), v(t) \rangle_{V^*,V} + \langle \partial_t v(t), u(t) \rangle_{V^*,V} \quad \text{a.e. } t \in (0, T)$$

and hence the partial integration formula holds,

$$\int_0^T \langle \partial_t u(t), v(t) \rangle_{V^*,V} dt = \langle u(T), v(T) \rangle_H - \langle u(0), v(0) \rangle_H - \int_0^T \langle \partial_t v(t), u(t) \rangle_{V^*,V} dt \tag{35}$$

Weak Formulation of Parabolic IBVPs

- The *strong form* (implicitly requiring strong regularity) of a typical parabolic PDE is

$$\begin{cases} u_t &= Lu, & \text{in } \Omega \times (0,T] \\ Bu &= g, & \text{on } \partial\Omega \times [0,T] \\ u &= u_0, & \text{in } \Omega \times \{0\} \end{cases} \tag{36}$$

  where $T \in (0,\infty)$, $\Omega \subset \mathbb{R}^n$ is open and bounded with Lipschitz $\partial\Omega$ and Lu is given as for (34).
- The normed linear space of continous (bounded) linear operators mapping X into Y is $\mathcal{L}(X,Y)$ (or $\mathcal{L}(X)$ for $X = Y$) equipped with the operator norm $\|\cdot\|_{X,Y}$ (or $\|\cdot\|_X$).
- The solution operator $S(t)$ for the IBVP above is roughly $\exp(Lt)$ and it satisfies the following properties.

Def: A *contraction semigroup* on a Hilbert space H is a family of operators $\{S(t)\}_{t\geq 0} \subset \mathcal{L}(H)$ satisfying:

a. $\|S(t)\|_H \leq 1$, b. $S(t+\tau) = S(t)S(\tau), t, \tau \geq 0$,
c. $S(0) = I$, d. $S(t)u \in C([0,\infty), H), \forall u \in H$.

Lumer Philips Theorem

- Differentiation gives formally $D_t S(t) = D_t \exp(Lt) \to L$, $t \to 0$.

Def: The *generator* of a contraction semigroup is an operator L with domain

$$\mathrm{Dom}(L) = \{u \in H : D_t S(t)u|_{t=0} \text{ exists in } H\}$$

and value $Lu = D_t S(t)u|_{t=0}$.

- For $S(t) = \exp(Lt)$ to exist for $t > 0$, the generator L should be a negative operator:

Def: An operator $L \in \mathcal{L}(\mathrm{Dom}(L), H)$ is *dissipative* if $\Re(Lu, u)_H \leq 0$, $\forall u \in \mathrm{Dom}(L)$.

Theorem (Lumer Phillips): Given a Hilbert space H with subspace $\mathrm{Dom}(L)$, an operator $L \in \mathcal{L}(\mathrm{Dom}(L), H)$ generates a contraction semigroup if and only if

- L is dissipative,
- $L - \lambda I$ is surjective $\forall \lambda > 0$.

Existence of Semigroups and Weak Solutions

Def: The *Cauchy problem*

$$u'(t) = Lu(t), \quad u(0) = u_0$$

has a *classical solution* $u \in \mathcal{C}([0,\infty), H)) \cap C^1((0,\infty), H)$ if it satisfies the IVP and $u(t) \in \text{Dom}(L)$, $\forall t > 0$.

Def: A function $u \in L^2([0,\infty), H)$ is a *weak solution* to the Cauchy problem (unique) if

$$-\int_0^\infty (u, \phi_t + L^*\phi)_H dt = (u_0, \phi(0))_H, \quad \forall \phi \in C_0^\infty(\mathbb{R}, \text{Dom}(L^*))$$

Def: A function $u \in \mathcal{C}([0,\infty), H)$ is a *mild solution* to the Cauchy problem (also a weak solution) if

$$\int_0^t u(s)ds \in \text{Dom}(L) \quad \text{and} \quad L\int_0^t u(s)ds = u(t) - u_0.$$

Def: A function $u(t) = S(t)u_0$ is a *strong solution* to the Cauchy problem if $u_0 \in \text{Dom}(L)$.

Theorem: Given a Hilbert space H with subspace $\text{Dom}(L)$, and operator $L \in \mathcal{L}(\text{Dom}(L), H)$ generates a contraction semigroup $S(t)$ if and only if $\forall u_0 \in H$, there exists a mild (and weak) solution $u(t)$ to the Cauchy problem, and $u(t) = S(t)u_0$.

Weak Formulation of Hyperbolic IBVPs

- Consider (36) with $Bu = u$, $g = 0$.
  - Take $H = L^2(\Omega)$.
  - Set $\mathrm{Dom}(L) = H^2(\Omega) \cap H_0^1(\Omega)$.
  - $(Lu, u)_{L^2(\Omega)} = -a(u, u) \leq 0$, $\forall u \in \mathrm{Dom}(L)$, so L is dissipative.
  - $\forall f \in H$, $\forall \lambda > 0$, $\exists! u \in \mathrm{Dom}(L)$ s.t. $\lambda u - Lu = f$, so L satisfies the range condition.
  - By the Lumer Philips Theorem, L generates a contraction semigroup $S(t)$.
  - By the last theorem, $u(t) = S(t)u_0$ is a mild (weak) solution to the Cauchy problem, $u'(t) = Lu(t)$, $u(0) = u_0$.
- Similarly, consider the typical hyperbolic problem,

$$\begin{cases} u_{tt} &= Lu, & \text{in } \Omega \times (0, T] \\ Bu &= g, & \text{on } \partial\Omega \times [0, T] \\ u &= u_0, & \text{in } \Omega \times \{0\} \\ u_t &= u_1, & \text{in } \Omega \times \{0\} \end{cases}$$

  where $T \in (0, \infty)$, $\Omega \subset \mathbb{R}^n$ is open and bounded with Lipschitz $\partial\Omega$ and Lu is given as for (34).

Weak Formulation of Hyperbolic IBVPs

- To establish a semigroup to solve this problem, it is rewritten in first order form,

$$U_t = \mathcal{L}U, \quad \mathcal{B}U_{\partial\Omega} = G, \quad U(0) = U_0$$

$$U_0 = \begin{pmatrix} u_0 \\ u_1 \end{pmatrix}, \quad \mathcal{L} = \begin{pmatrix} 0 & I \\ L & 0 \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix}, \quad G = \begin{pmatrix} g \\ 0 \end{pmatrix}, \quad U_0 = \begin{pmatrix} u_0 \\ u_1 \end{pmatrix}$$

- Consider this problem with $Bu = u$, $g = 0$.
  - Take $H = H_0^1(\Omega) \times L^2(\Omega)$, $(U, V)_H = a(u_1, v_1) + (u_2, v_2)$.
  - Set $\mathrm{Dom}(L) = H^2(\Omega) \cap H_0^1(\Omega)$ and $\mathrm{Dom}(\mathcal{L}) = \mathrm{Dom}(L) \times H_0^1(\Omega)$.
  - $(\mathcal{L}U, U)_{H_0^1(\Omega)\times L^2(\Omega)} = a(u_1, u_2) + (u_2, Lu_1) = 0$, $\forall U \in \mathrm{Dom}(\mathcal{L})$, so $\mathcal{L}$ is dissipative.
  - $\forall F \in H$, $\forall \lambda > 0$, $\exists! U \in \mathrm{Dom}(\mathcal{L})$ s.t. $\lambda u_1 - u_2 = f_1$ and $\lambda u_2 - Lu_1 = f_2$ or $\lambda U - \mathcal{L}U = F$, so $\mathcal{L}$ satisfies the range condition.
  - By the Lumer Philips Theorem, $\mathcal{L}$ generates a contraction semigroup $\mathcal{S}(t)$.
  - By the last theorem, $U(t) = \mathcal{S}(t)U_0$ is a mild solution to the Cauchy problem, $U'(t) = \mathcal{L}U(t)$, $U(0) = U_0$.

Weak Formulation, Non-Autonomous Parabolic PDEs

- Suppose now for the IBVP,

$$\begin{cases} u_t &= Lu + f, & \text{in } \Omega \times (0,T] \\ Bu &= g, & \text{on } \partial\Omega \times [0,T] \\ u &= u_0, & \text{in } \Omega \times \{0\} \end{cases} \tag{37}$$

that

$$Lu(x,t) = -\nabla\cdot[A(x,t)\nabla u(x,t)] + c^\top(x,t)\nabla u(x,t) + r(x,t)u(x,t)$$

for $A = \{a_{ij}(x,t)\}_{i,j=1}^n$, $c = \{c_i(x,t)\}_{i=1}^n$, $r(x,t)$, $f(x,t)$ and $g(x,t)$, sufficiently smooth functions on $\bar{\Omega} \times [0,T]$.

- Suppose that A is uniformly elliptic, i.e., $\exists \alpha > 0$ such that

$$\xi^\top A(x,t)\xi \geq \alpha \xi^\top \xi, \quad \forall \xi \in \mathbb{R}^n, \quad \forall x \in \Omega, \quad \forall t \in [0,T].$$

- Analogous to (34), for each $t \in [0,T]$, define the continuous bilinear form on $H^1(\Omega) \times H^1(\Omega)$,

$$a(t;u,v) = (A(t)\nabla u, \nabla v)_{L^2(\Omega)} + (c(t)^\top \nabla u, v)_{L^2(\Omega)} + (r(t)u, v)_{L^2(\Omega)}$$

for Dirichlet or Neumann BCs, and set $\tilde{a}(t;u,v) = a(t;u,v) + \langle \sigma_0(t)u, v\rangle$ for Robin BCs.

Weak Formulation, Non-Autonomous Parabolic PDEs

- For each $t \in [0, T]$ define the continuous linear form on $H^1(\Omega)$ by $b(t; v) = (f(t), v)$ for Dirichlet BCs, and by $b(t; v) = (f(t), v) + \langle g(t), v \rangle$ for Neumann and Robin BCs.
- As explained in relation to (34), the underlying space V in which solution values are sought depends upon the BCs, and typically $V = H_0^1(\Omega)$ or $V = H^1(\Omega)$.
- For simplicity, assume $u_0 = 0$. Otherwise, (a) let v be a sufficiently smooth function satisfying $v(0) = u_0$, (b) solve (37) for $w = u - v$ replacing f by $f - \partial_t v + Lv$ and (c) take $u = w + v$ to solve the original problem.
- Using the partial integration formula (35), we seek a weak solution to (37) as $u \in \{v \in W^{1,2}(V, V^*) : v(0) = 0\}$ such that

$$\int_0^T \left[\langle \partial_t u(t), v(t) \rangle_{V^*, V} + a(t; u(t), v(t)) \right] dt = \int_0^T b(t; v(t)) dt \quad \forall v \in L^2(0, T; V) \tag{38}$$

Banach-Nečas-Babuška Theorem

- This notion of weak solution is stronger than previously since time derivatives and temporal traces appear explicitly in the formulation. Existence is guaranteed by the following.

Theorem (Banach-Nečas-Babuška): Assume $a(t;\cdot,\cdot) : V \times V \to \mathbb{R}$ satisfies:

- $t \mapsto a(t;u,v)$ is measurable $\forall u, v \in V$.
- $\exists c_1 > 0$ such that

$$a(t;u,v) \geq c_1 \|u\|_V^2, \quad \text{a.e. } t \in (0,T), \quad \forall u \in V.$$

- $\exists c_2 > 0$ such that

$$|a(t;u,v)| \leq c_2 \|u\|_V \|v\|_V, \quad \text{a.e. } t \in (0,T), \quad \forall u, v \in V.$$

Then (38) has a unique solution $u \in \{v \in W^{1,2}(V,V^*) : v(0) = 0\}$ satisfying

$$\|u\|_{W^{1,2}(V,V^*)} \leq \frac{1}{c_1} \sup_{v \in L^2(0,T;V)} \int_0^T b(t;v(t))dt.$$

Finite Element Methods for Elliptic Problems

- We seek an approximation to the solution to (34), more precisely, to a weak solution $u \in V$, where the Hilbert space V depends upon BCs, e.g., $V = H_0^1(\Omega)$ for Dirichlet BCs and $V = H^1(\Omega)$ for Neumann or Robin BCs.
- Given a continuous bilinear functional $a : V \times V \to \mathbb{R}$ and a continuous linear functional $b : V \to \mathbb{R}$, we seek a $u \in V$ such that

$$a(u, v) = b(v), \quad \forall v \in V \tag{39}$$

- Suppose the following inequalities hold $\forall u, v \in V$,

$$c_1 \|u\|_V^2 \leq a(u, u), \quad |a(u, v)| \leq c_2 \|u\|_V \|v\|_V, \quad |b(v)| \leq c_3 \|v\|_V \tag{40}$$

  By the Lax Milgram Theorem, $\exists! u \in V$ satisfying (39).
- For the *conforming Galerkin approach*, choose a finite-dimensional $V_h \subset V$ and seek $u_h \in V_h$ satisfying

$$a(u_h, v_v) = b(v_h), \quad \forall v_h \in V_h \tag{41}$$

Céa's Lemma

- Since $V_h \subset V$ is a closed subspace, V_h is a Hilbert space equipped with $(\cdot,\cdot)_V$. Furthermore, the estimates (40) hold on $V_h \times V_h$ and V_h too. Thus, by the Lax Milgram Theorem, $\exists! u_h \in V_h$ satisfying (41).
- Since V_h is finite-dimensional, it has a basis $\{\Phi_i\}$. The solution $u_h = \sum_i u_i \Phi_i$ to (41) satisfies the system $\sum_j u_i a(\Phi_i, \Phi_j) = b(\Phi_j)$, solved uniquely as $K = \{a(\Phi_i, \Phi_j)\}$ is invertible. (**Exercise**)

Lemma (Céa): Let u_h solve (41) for $V_h \subset V$ and let u solve (39), where a and b satisfy (40). Then

$$\|u - u_h\|_V \leq \frac{c_2}{c_1} \inf_{v \in V_h} \|u - v_h\|_V$$

Proof: Since $V_h \subset V$, $a(u - u_h, v_h) = 0$ holds $\forall v_h \in V_h$. In particular, $a(u - u_h, v_h - u_h) = 0$ for any $v_h \in V_h$. So using (40),

$$c_1 \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h)$$
$$= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h)_{=0} \leq c_2 \|u - u_h\|_V \|u - v_h\|_V$$

Dividing by $\|u - u_h\|_V$ and taking the inf over $v \in V_h$ gives the claimed result. ■

Ritz-Galerkin Approximation

- An estimation of the inf is given by approximation properties to be discussed later.
- If a is symmetric, u satisfies (39) iff u minimizes $J(v) = \frac{1}{2}a(v,v) - b(v)$ over all $v \in V$.
  Exercise: Show this.
- Coercivity, continuity and symmetry imply that $a(u,u)$ is an inner product on V with *energy norm* $\|u\|_a = a(u,u)^{1/2}$.
  Exercise: Show this.
- Under these conditions, the solution $u_h \in V_h$ to (41) is the *Ritz-Galerkin approximation*,

$$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a$$

- When it is necessary to estimate the error $u - u_h$ in a weak norm, a *duality argument* may be used. Let V be continuously embedded in a Hilbert space H, e.g., $V = H_0^1(\Omega)$ and $H = L^2(\Omega)$. Then we have the estimate of the following theorem.

Aubin Nitsche Lemma

- For this, note that when a is symmetric, a unique solution to the *adjoint problem* is given by the Lax Milgram Theorem,

$$a(w,u) = b(w), \quad \forall w \in V.$$

Lemma (Aubin Nitsche): Let u_h be the solution to (41) for $V_h \subset V$ and let u be the solution to (39). Then $\exists c > 0$ such that

$$\|u - u_h\|_H \leq c\|u - u_h\|_V \sup_{g \in H} \left(\frac{1}{\|g\|_H} \inf_{v_h \in V_h} \|\phi_g - v_h\|_V \right)$$

where for a given $g \in H$, ϕ_g is the unique solution to the adjoint problem $a(w, \phi_g) = (g, w)_H$, $\forall w \in V$.

Proof: Let $w = u - u_h$ in the adjoint problem. For any $v_h \in V_h$,

$$\begin{aligned}(g, u - u_h)_H = a(u - u_h, \phi_g) &= a(u - u_h, \phi_g - v_h) \\ &\leq c\|u - u_h\|_V \|\phi_g - u_h\|_V\end{aligned}$$

Then since for $w \in H$, $\|w\|_H = \sup_{g \in H}(g, w)_H / \|g\|_H$,

$$\|u - u_h\|_H = \sup_{g \in H} \frac{(g, u - u_h)_H}{\|g\|_H} \leq c\|u - u_h\|_V \sup_{g \in H} \frac{\|\phi_g - v_h\|_V}{\|g\|_H}$$

Since v_h is arbitrary, taking the inf gives the claimed result. ■

FEM for Dirichlet Poisson BVP on a Square

- Consider the Dirichlet Problem for the Poisson Equation,

$$-\Delta u = f, \quad \text{in } \Omega, \quad u = 0, \quad \text{on } \partial\Omega$$

on a square $\Omega = (0,1)^2$. The weak formulation is

$$a(u,v) = b(v), \quad \forall v \in V, \quad V = H_0^1(\Omega)$$

with $a(u,v) = (\nabla u, \nabla v)_{L^2(\Omega)^2}$ and $b(v) = (f,v)_{L^2(\Omega)}$.

- **Exercise**: Show using Poincaré's inequality that (40) hold.
- For nodes $X = \{x_i\}_{i=0}^{N+1}$, $x_i = ih$, $h = 1/(N+1)$, define

$$S(X) = \{\ell \in C([0,1]) : \ell|_{[x_{i-1},x_i]} \in P_1, v(0) = v(1) = 0\}$$

- Define the *hat* functions, (figure forthcoming)

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h, & x \in [x_{i-1}, x_i] \\ (x_{i+1} - x)/h, & x \in [x_i, x_{i+1}] \end{cases}$$

- Similarly, define $\{y_j\}_{j=0}^{N+1}$, $S(Y)$ and $\phi_j(y)$.
- Finally, define the tensor product space $V_h = S(X) \times S(Y)$.

Error Estimate for Dirichlet Poisson BVP

- **Exercise**: Show that $V_h \subset V$ and that $\{\Phi_{i,j}\}_{i,j=1}^{N}$, $\Phi_{i,j}(x,y) = \phi_i(x)\phi_j(y)$, forms a basis for V_h.
- Since (40) hold, the solution $u_h = \sum_{i,j=1}^{N} u_{i,j}\Phi_{i,j}$ to (41) is given uniquely by solving the system of linear equations

$$\sum_{i,j=1}^{N} u_{i,j} a(\Phi_{i,j}, \Phi_{k,l}) = b(\Phi_{k,l}), \quad k,l = 1,\dots,N \qquad (42)$$

- Since a is symmetric, let V be equipped with the energy norm $\|u\|_a = a(u,u)^{\frac{1}{2}} = \|\nabla u\|_{L^2(\Omega)^2}$.
- **Claim**: $\exists c > 0$ independent of h such that

$$\|u - u_h\|_{L^2(\Omega)} + h\|u - u_h\|_a \leq ch^2\|u\|_{H^2(\Omega)}$$

- According to Céa's Lemma 173 it is sufficient for the energy norm estimate to show that $\|u - \mathcal{I}_h u\|_a = \mathcal{O}(h)$ where $\mathcal{I}_h$ is the bilinear interpolation operator,

$$\mathcal{I}_h u = \sum_{i,j=1}^{N} u_{i,j}\phi_i(x)\phi_j(y) \in V_h, \quad u_{i,j} = \frac{1}{4h^2}\int_{x_{i-1}}^{x_{i+1}}\int_{y_{j-1}}^{y_{j+1}} u(x,y)dxdy$$

understanding $u = 0$ outside Ω.

Error Estimate for Interpolation Operators

- Note that $\mathcal{I}_h = \mathcal{I}_h^x \mathcal{I}_h^y = \mathcal{I}_h^y \mathcal{I}_h^x$ where

$$\mathcal{I}_h^x u(x,y) = \sum_{i=1}^{N} u_i(y)\phi_i(x), \quad u_i(y) = \frac{1}{2h}\int_{x_{i-1}}^{x_{i+1}} u(x,y)dx$$

$$\mathcal{I}_h^y u(x,y) = \sum_{j=1}^{N} u_j(x)\phi_j(y), \quad u_j(x) = \frac{1}{2h}\int_{y_{j-1}}^{y_{j+1}} u(x,y)dy$$

- So the error is estimated by triangulating with

$$(I - \mathcal{I}_h)u = (I - \mathcal{I}_h^x)u + \mathcal{I}_h^x(I - \mathcal{I}_h^y)u$$

or

$$\|u - \mathcal{I}_h u\|_a \leq \|u - \mathcal{I}_h^x u\|_a + \|\mathcal{I}_h^x\|_a \|u - \mathcal{I}_h^y u\|_a$$

given the following bound.

- Estimating $\|\mathcal{I}_h^x\|_a$. Let $U = \mathcal{I}_h^x u$.
  - First consider $\|U_x\|_{L^2(\Omega)}$. For $x \in [x_{i-1}, x_i]$,

$$U_x(x,y) = u_{i-1}(y)\phi'_{i-1}(x) + u_i(y)\phi'_i(x)$$

$$= \frac{1}{2h}\int_{x_{i-2}}^{x_i} u(t,y)dt\left(-\frac{1}{h}\right) + \frac{1}{2h}\int_{x_{i-1}}^{x_{i+1}} u(t,y)dt\left(\frac{1}{h}\right)$$

Error Estimate for Interpolation Operators

$$= \frac{1}{2h^2}\int_{x_{i-1}}^{x_{i+1}} [u(t,y) - u(t-h,y)]\,dt = \frac{1}{2h^2}\int_{x_{i-1}}^{x_{i+1}} \left[\int_{t-h}^{t} u_x(s,y)ds\right] dt$$

- Estimating absolute values

$$|U_x(x,y)| \le \frac{1}{2h^2}\int_{x_{i-1}}^{x_{i+1}} \left[\int_{x_{i-2}}^{x_{i+1}} |u_x(s,y)|ds\right] dt \le \frac{2h}{2h^2}\left[\int_{x_{i-2}}^{x_{i+1}} 1ds\right]^{\frac{1}{2}} \left[\int_{x_{i-2}}^{x_{i+1}} |u_x(s,y)|^2 ds\right]^{\frac{1}{2}}$$

- Squaring and integrating over $x \in [x_{i-1}, x_i]$,

$$\int_{x_{i-1}}^{x_i} |U_x(x,y)|^2 dx \le \frac{1}{h^2}(3h)\int_{x_{i-2}}^{x_{i+1}} |u_x(s,y)|^2 ds \int_{x_{i-1}}^{x_i} 1dx = 3\int_{x_{i-2}}^{x_{i+1}} |u_x(s,y)|^2 ds$$

- Summing over $i = 1, \ldots, N+1$ and integrating over $y \in [0,1]$,

$$\|U_x\|^2_{L^2(\Omega)} \le 9\|u_x\|^2_{L^2(\Omega)}$$

- Now consider $\|U_y\|_{L^2(\Omega)}$. For $x \in [x_{i-1}, x_i]$,

$$U_y(x,y) = \phi_{i-1}(x)u'_{i-1}(y) + \phi_i(x)u'_i(y)$$

$$= \frac{x_i - x}{2h^2}\int_{x_{i-2}}^{x_i} u_y(t,y)dt + \frac{x - x_{i-1}}{2h^2}\int_{x_{i-1}}^{x_{i+1}} u_y(t,y)dt$$

Error Estimate for Interpolation Operators

- Estimating absolute values

$$|U_y(x,y)| \le \frac{h}{2h^2}\int_{x_{i-2}}^{x_i}|u_y(t,y)|dt + \frac{h}{2h^2}\int_{x_{i-1}}^{x_{i+1}}|u_y(t,y)|dt \le \frac{1}{h}\int_{x_{i-2}}^{x_{i+1}}|u_y(t,y)|dt$$
$$\le \frac{1}{h}\left[\int_{x_{i-2}}^{x_{i+1}} 1dt\right]^{\frac{1}{2}}\left[\int_{x_{i-2}}^{x_{i+1}}|u_y(t,y)|^2dt\right]^{\frac{1}{2}}$$

- Squaring and integrating over $x \in [x_{i-1}, x_i]$,

$$\int_{x_{i-1}}^{x_i}|U_y(x,y)|^2dx \le 3\int_{x_{i-2}}^{x_{i+1}}|u_y(x,y)|^2dx$$

- Summing over $i = 1, \ldots, N+1$ and integrating over $y \in [0,1]$,

$$\|U_y\|^2_{L^2(\Omega)} \le 9\|u_y\|^2_{L^2(\Omega)}$$

- Combining the estimates above

$$\|\mathcal{I}_x u\|_a^2 = \|U_x\|^2_{L^2(\Omega)} + \|U_y\|^2_{L^2(\Omega)} \le 9\|u_x\|^2_{L^2(\Omega)} + 9\|u_y\|^2_{L^2(\Omega)} = 9\|u\|_a^2$$

gives the bound

$$\|\mathcal{I}_x\|_a \le 3.$$

Error Estimate for Interpolation Operators

- Estimating the error $e = \mathcal{I}_h^x u - u$.
  - First consider $\|e_x\|_{L^2(\Omega)}$. For $x \in [x_{i-1}, x_i]$,

$$
\begin{aligned}
e_x(x,y) &= u_{i-1}(y)\phi'_{i-1}(x) + u_i(y)\phi'_i(x) - u_x(x,y) \\
&= \frac{1}{2h}\int_{x_{i-2}}^{x_i} u(t,y)dt\left(-\frac{1}{h}\right) + \frac{1}{2h}\int_{x_{i-1}}^{x_{i+1}} u(t,y)dt\left(\frac{1}{h}\right) - u_x(x,y) \\
&= \frac{1}{2h^2}\int_{x_{i-1}}^{x_{i+1}} [u(t,y) - u(t-h,y)]\,dt - u_x(x,y) \\
&= \frac{1}{2h^2}\int_{x_{i-1}}^{x_{i+1}} \left[\int_{t-h}^{t} u_x(s,y)ds\right] dt - u_x(x,y)
\end{aligned}
$$

  - Since $u_x(x,y)$ is independent of s and t,

$$
\begin{aligned}
e_x(x,y) &= \frac{1}{2h^2}\int_{x_{i-1}}^{x_{i+1}}\int_{t-h}^{t} [u_x(s,y) - u_x(x,y)]\,dsdt \\
&= \frac{1}{2h^2}\int_{x_{i-1}}^{x_{i+1}}\int_{t-h}^{t}\left[\int_{x}^{s} u_{xx}(r,y)dr\right] dsdt
\end{aligned}
$$

Error Estimate for Interpolation Operators

- Estimating absolute values,

$$|e_x(x,y)| \leq \frac{1}{2h^2}\int_{x_{i-1}}^{x_{i+1}}\int_{t-h}^{t}\left[\int_{x_{i-2}}^{x_{i+1}}|u_{xx}(r,y)|dr\right]dsdt$$

$$\leq \left[\int_{x_{i-2}}^{x_{i+1}}1dr\right]^{\frac{1}{2}}\left[\int_{x_{i-2}}^{x_{i+1}}|u_{xx}(r,y)|^2dr\right]^{\frac{1}{2}}$$

- Squaring and integrating over $x \in [x_{i-1}, x_i]$,

$$\int_{x_{i-1}}^{x_i}|e_x(x,y)|^2dx \leq 3h^2\int_{x_{i-2}}^{x_{i+1}}|u_{xx}(x,y)|^2dx$$

- Summing over $i = 1, \ldots, N+1$ and integrating over $y \in [0,1]$,

$$\|e_x\|^2_{L^2(\Omega)} \leq 9h^2\|u_{xx}\|^2_{L^2(\Omega)}$$

- Now consider $\|e_y\|_{L^2(\Omega)}$. For $x \in [x_{i-1}, x_i]$,

$$e_y(x,y) = \phi_{i-1}(x)u'_{i-1}(y) + \phi_i(x)u'_i(y) - u_y(x,y)$$

$$= \frac{x_i - x}{2h^2}\int_{x_{i-2}}^{x_i}u_y(t,y)dt + \frac{x - x_{i-1}}{2h^2}\int_{x_{i-1}}^{x_{i+1}}u_y(t,y)dt - u_y(x,y)$$

Error Estimate for Interpolation Operators

- Since $u_y(x,y)$ is independent of t,

$$e_y(x,y) = \frac{x_i - x}{2h^2}\int_{x_{i-2}}^{x_i}[u_y(t,y)-u_y(x,y)]dt + \frac{x - x_{i-1}}{2h^2}\int_{x_{i-1}}^{x_{i+1}}[u_y(t,y)-u_y(x,y)]dt$$

$$= \frac{1}{2h^2}\int_{x_{i-1}}^{x_{i+1}}\{(x_i - x)[u_y(t-h,y) - u_y(x,y)] + (x - x_{i-1})[u_y(t,y) - u_y(x,y)]\}\,dt$$

$$= \frac{1}{2h^2}\int_{x_{i-1}}^{x_{i+1}}\left\{(x_i - x)\left[\int_x^{t-h} u_{xy}(s,y)ds\right] + (x - x_{i-1})\left[\int_x^{t} u_{xy}(s,y)ds\right]\right\}dt$$

- Estimating absolute values,

$$|e_y(x,y)| \le \frac{1}{2h}\int_{x_{i-1}}^{x_{i+1}}\left\{\left[\int_{x_{i-2}}^{x_i}|u_{xy}(s,y)|ds\right] + \left[\int_{x_{i-1}}^{x_{i+1}}|u_{xy}(s,y)|ds\right]\right\}dt$$

$$\le 2\int_{x_{i-2}}^{x_{i+1}}|u_{xy}(s,y)|ds \le 2\left[\int_{x_{i-2}}^{x_{i+1}}1ds\right]^{\frac{1}{2}}\left[\int_{x_{i-2}}^{x_{i+1}}|u_{xy}(s,y)|^2ds\right]^{\frac{1}{2}}$$

- Squaring and integrating over $x \in [x_{i-1}, x_i]$,

$$\int_{x_{i-1}}^{x_i}|e_y(x,y)|^2dx \le 12h^2\int_{x_{i-2}}^{x_{i+1}}|u_{xy}(x,y)|^2dx$$

Energy Estimate for Dirichlet Poisson BVP

- Summing over $i = 1, \ldots, N+1$ and integrating over $y \in [0,1]$,

$$\|e_y\|^2_{L^2(\Omega)} \leq 36h^2 \|u_{xy}\|^2_{L^2(\Omega)}$$

- Combining the estimates for e_x and e_y gives

$$\|u - \mathcal{I}_h^x u\|_a^2 = \|e_x\|^2_{L^2(\Omega)} + \|e_y\|^2_{L^2(\Omega)} \leq 9h^2 \|u_{xx}\|^2_{L^2(\Omega)} + 36h^2 \|u_{xy}\|^2_{L^2(\Omega)}$$

- Estimating the error $E = u - \mathcal{I}_h^y u$ is performed analogously to obtain

$$\|E_x\|^2_{L^2(\Omega)} \leq 36h^2 \|u_{xy}\|^2_{L^2(\Omega)}, \quad \|E_y\|^2_{L^2(\Omega)} \leq 9h^2 \|u_{yy}\|^2_{L^2(\Omega)}$$

$$\|u - \mathcal{I}_h^y u\|_a^2 = \|E_x\|^2_{L^2(\Omega)} + \|E_y\|^2_{L^2(\Omega)} \leq 36h^2 \|u_{xy}\|^2_{L^2(\Omega)} + 9h^2 \|u_{yy}\|^2_{L^2(\Omega)}$$

- Combining all the estimates above gives the interpolation error estimate,

$$\|u - \mathcal{I}_h u\|_a \leq \|u - \mathcal{I}_h^x u\|_a + \|\mathcal{I}_h^x\|_a \|u - \mathcal{I}_h^y u\|_a \leq 144h|u|_{H^2(\Omega)}$$

with which Céa's Lemma 173 gives

$$\|u - u_h\|_a \leq c \inf_{v \in V_h} \|u - v_h\|_a \leq ch\|u\|_{H^2(\Omega)}$$

where $\|u\|_{H^2(\Omega)} \leq c\|f\|_{L^2(\Omega)}$ according to Theorem 160.

L^2 Estimate for Dirichlet Poisson BVP

- The $L^2(\Omega)$ estimate of the error is obtained using the Aubin Nitsche Lemma [175],

$$\|u-u_h\|_{L^2(\Omega)} \leq c\|u-u_h\|_a \sup_{g\in L^2(\Omega)} \left(\frac{1}{\|g\|_{L^2(\Omega)}} \inf_{v_h\in V_h} \|\phi_g - v_h\|_a \right)$$

  where ϕ_g solves the adjoint problem $a(w,\phi_g) = (g,w)_H$, $\forall w \in V$, which is identical to the primal problem, since a is symmetric.

- Exactly as shown for the solution u to the primal problem

$$\inf_{v_h\in V_h} \|\phi_g - v_h\|_a \leq ch\|\phi_g\|_{H^2(\Omega)} \leq ch\|g\|_{L^2(\Omega)}$$

  where Theorem [160] has been used for the last estimate.

- Combining the above estimates gives

$$\|u - u_h\|_{L^2(\Omega)} \leq ch^2\|f\|_{L^2(\Omega)}$$

  which establishes the claim [177].

Implementation of FEM Solver on a Square

- As noted above, the numerical solution $u_h = \sum_{i,j} u_{i,j}\Phi_{i,j}(x,y)$ to (41) is given by solving the linear system of equations `KU=F` in (42), where

$$\mathtt{K} = \{a(\Phi_{i,j}, \Phi_{k,l})\}_{i,j,k,l=1}^{N}, \quad \mathtt{U} = \{u_{i,j}\}_{i,j=1}^{N}, \quad \mathtt{F} = \{b(\Phi_{k,l})\}_{k,l=1}^{N}.$$

- For this one must first *assemble* the *stiffness matrix* `K` and the *load vector* `F`.
- A *node-based assembly* is to perform the computation

```
for k=1:N for l=1:N
  m = (l-1)*N + k
  for i=1:N for j=1:N
    n = (j-1)*N + i
    K(m,n) = a(Phi(i,j),Phi(k,l))
  end
  F(m) = b(Phi(k,l))
end
```

where the lexicographic ordering is used here.

Implementation of FEM Solver on a Square

- A more efficient approach (especially for more complex problems) is *element based assembly* in which integrals are partitioned element-wise,

$$a(u_h, v_h) = \sum_{m,n=1}^{N+1} a_{m,n}(u_h, v_h), \quad b_h(v_h) = \sum_{m,n=1}^{N+1} b_{m,n}(v_h)$$

where with $\Omega_{m,n} = (x_{m-1}, x_m) \times (y_{n-1}, y_n)$,

$$a_{m,n}(u_h, v_h) = (\nabla u_h, \nabla v_h)_{L^2(\Omega_{m,n})}, \quad b_{m,n}(v_h) = (f, \nabla v_h)_{L^2(\Omega_{m,n})}$$

- Then for fixed m, n the following are computed $\forall i, j, k, l$,

$$a_{m,n}(\Phi_{i,j}, \Phi_{k,l}) = (\nabla \Phi_{i,j}, \nabla \Phi_{k,l})_{L^2(\Omega_{m,n})}, \quad b_{m,n}(\Phi_{k,l}) = (f, \nabla \Phi_{k,l})_{L^2(\Omega_{m,n})}$$

simultaneously.

- Yet many of these integrals are known to be zero since very few basis functions have overlapping supports:

$$\Phi_{i,j}(x, y) = 0, \quad (x, y) \notin \cup\{\Omega_{i-k,j-l} : k, l = 0, 1\}$$

- Thus, $\mathbb{K}$ is sparse, a fact which actually motivated the invention of the finite element method by engineers.

Element Based Assembly

- Since the basis functions are the same on each element, the integrals above can be transformed to a single reference element as follows.
- Recalling definitions, $\Phi_{i,j}(x,y) = \phi_i(x)\phi_j(y)$, note that

$$\phi_i(x) = \hat{\phi}((x - x_i)/h), \quad \hat{\phi}(\xi) = \begin{cases} 1+\xi, & \xi \in [-1,0] \\ 1-\xi, & \xi \in [0,+1] \end{cases}$$

and with $\hat{\Phi}(\xi,\eta) = \hat{\phi}(\xi)\hat{\phi}(\eta)$ the basis functions satisfy

$$\Phi_{i,j}(x,y) = \hat{\Phi}((x - x_i)/h, (y - y_j)/h)$$

- **Exercise**: Translations of $\hat{\Phi}$ lead to the following four *form functions* in the reference element $(\xi,\eta) \in \hat{\Omega} = (0,1)^2$,

$$\hat{\Phi}_4(\xi,\eta) = \xi\eta, \quad \hat{\Phi}_1(\xi,\eta) = (1-\xi)(1-\eta)$$
$$\hat{\Phi}_3(\xi,\eta) = (1-\xi)\eta, \quad \hat{\Phi}_2(\xi,\eta) = \xi(1-\eta)$$

Form Functions

- Thus the element-wise bilinear forms can be transformed according to

$$a_{m,n}(\Phi_{i,j}, \Phi_{k,l}) = \int_{\Omega_{m,n}} \nabla\Phi_{i,j} \cdot \nabla\Phi_{k,l} = \int_{\hat{\Omega}} \nabla\hat{\Phi}_{\tau_{m,n}(i,j)} \cdot \nabla\hat{\Phi}_{\tau_{m,n}(k,l)}$$

  which is non-zero only for $m-1 \leq i,k \leq m$ and $n-1 \leq j,l \leq n$ and otherwise

$$\tau_{m,n}(i,j) = \begin{cases} 1, & i = m-1, \quad j = n-1 \\ 2, & i = m, \quad\quad j = n-1 \\ 3, & i = m-1, \quad j = n \\ 4, & i = m, \quad\quad j = n \end{cases}$$

- Since $1 \leq i,j,k,l \leq N$, there are fewer cases to consider for boundary elements.
- **Exercise**: An explicit calculation shows

$$\left\{ \int_{\hat{\Omega}} \nabla\hat{\Phi}_\tau \cdot \nabla\hat{\Phi}_\sigma \right\}_{\tau,\sigma=1}^{4} = \frac{1}{6} \begin{bmatrix} 4 & -1 & -1 & -2 \\ -1 & 4 & -2 & -1 \\ -1 & -2 & 4 & -1 \\ -2 & -1 & -1 & 4 \end{bmatrix}$$

Explicit Calculation of Stiffness Matrix

- **Exercise**: Element-wise entries of the stiffness matrix are

$$a_{m,n}(\Phi_{i,j},\Phi_{k,l}) = \begin{cases} 2/3, & |i-k| = |j-l| = 0 \\ -1/3, & |i-k| = |j-l| = 1 \\ -1/6, & |i-k| + |j-l| = 1 \end{cases} \qquad \begin{matrix} m-1 \le i,k \le m \\ n-1 \le j,l \le n \end{matrix}$$

- Similarly, if f is approximated by $f_h = \sum_{i,j=1}^{N} f_{i,j}\Phi_{i,j}$, entries of the load vector are given element-wise according to

$$b_{m,n}(\Phi_{k,l}) = \int_{\Omega_{m,n}} f_h \Phi_{k,l} = h^2 \sum_{i,j=1}^{N} f_{i,j} \int_{\hat{\Omega}} \hat{\Phi}_{\tau_{m,n}(i,j)} \hat{\Phi}_{\tau_{m,n}(k,l)}$$

- Recall that $1 \le i,j,k,l \le N$.
- **Exercise**: An explicit calculation shows

$$\left\{ \int_{\hat{\Omega}} \hat{\Phi}_\tau \hat{\Phi}_\sigma \right\}_{\tau,\sigma=1}^{4} = \frac{1}{36} \begin{bmatrix} 4 & 2 & 2 & 1 \\ 2 & 4 & 1 & 1 \\ 2 & 1 & 4 & 2 \\ 1 & 2 & 2 & 4 \end{bmatrix}$$

- So the *mass matrix*

$$\mathbb{M} = \{\mu(\Phi_{i,j},\Phi_{k,l})\}_{i,j,k,l=1}^{N}, \quad \mu(\Phi_{i,j},\Phi_{k,l}) = (\Phi_{i,j},\Phi_{k,l})_{L^2(\Omega)}$$

Explicit Calculation of Mass Matrix

can be similarly partitioned element-wise,

$$\mu(u_h, v_h) = \sum_{m,n=1}^{N+1} \mu_{m,n}(u_h, v_h), \quad \mu_{m,n}(u_h, v_h) = (u_h, v_h)_{L^2(\Omega_{m,n})}$$

- **Exercise**: Element-wise entries of the mass matrix are

$$\mu_{m,n}(\Phi_{i,j}, \Phi_{k,l}) = h^2 \begin{cases} 1/9, & |i-k| = |j-l| = 0 \\ 1/36, & |i-k| = |j-l| = 1 \\ 1/18, & |i-k| + |j-l| = 1 \end{cases} \qquad \begin{array}{r} m-1 \le i, k \le m \\ n-1 \le j, l \le n \end{array}$$

- The load vector is given by $\mathtt{F} = \mathtt{M}\mathtt{f}$ where $\mathtt{f} = \{f_{i,j}\}_{i,j=1}^{N}$.
- With variable coefficients, quadrature must be used to calculate the integrals.
- To accommodate (non-)homogeneous Dirichlet BCs, $u = g$ on $\partial\Omega$, new rows corresponding to boundary indices (k', l') can be added to the stiffness matrix $\mathtt{K}$ and to the load vector $\mathtt{F}$, where $\mathtt{K}$ only has a 1 on the diagonal of the (k', l')'th row and $\mathtt{F}_{k',l'} = g_{k',l'}$.

Implementation of FEM Solver on a Square

- FEM Code with assembly of mass and stiffness matrices to solve the Dirichlet Poisson BVP on a square:

```
Input h, N, f = (f(i,j), i,j=1,...,N)^T
Set K(r,c) = M(r,c) = 0, r,c=1:N*N
for m=1:N+1 for n=1:N+1
  for i=max(1,m-1):min(N,m) for j=max(1,n-1):min(N,n)
    r = (j-1)*N + i
    for k=max(1,m-1):min(N,m) for l=max(1,n-1):min(N,n)
      c = (l-1)*N + k
      K(r,c) = K(r,c) + { 2/3,   |i-k| = |j-l| = 0
                        { -1/3,  |i-k| = |j-l| = 1
                        { -1/6,  |i-k| + |j-l| = 1
      M(r,c) = M(r,c) + h^2 { 1/9,   |i-k| = |j-l| = 0
                            { 1/18,  |i-k| = |j-l| = 1
                            { 1/36,  |i-k| + |j-l| = 1
    end
  end
end
Solve KU=Mf for U=(u(i,j), i,j=1,...,N)^T
```

A Posteriori Error Estimates and Adaptivity

- Suppose $\Omega = (0,1)^2$ is partitioned into rectangular elements,

$$\Omega = \cup_{\mathrm{e}=1}^{\nu} \Omega_{\mathrm{e}}, \quad \Omega_{\mathrm{e}} = (x_{\mathrm{e}} - \delta x_{\mathrm{e}}, x_{\mathrm{e}}) \times (y_{\mathrm{e}} - \delta y_{\mathrm{e}}, y_{\mathrm{e}})$$

- With $h_{\mathrm{e}} = \max\{\delta x_{\mathrm{e}}, \delta y_{\mathrm{e}}\}$ and $h = \max_{\mathrm{e}} h_{\mathrm{e}}$ let the approximation space be

$$V_h = \{v_h \in \mathcal{C}(\Omega) : v_h|_{\partial\Omega} = 0, v_h(x,y)|_{\Omega_{\mathrm{e}}} \in P_1 \times P_1, \mathrm{e} = 1, \ldots, \nu\}$$

- Let $u_h \in V_h$ be the solution to (41).
- To avoid the potentially unnecessary expense of an always uniform grid,
  - begin here with a coarse collection of elements $\{\Omega_{\mathrm{e}}\}_{\mathrm{e}=1}^{\nu}$,
  - and then refine to obtain $\{\Omega_{\mathrm{e}'}\}_{\mathrm{e}'=1}^{\nu'}$ containing new elements, some of which are smaller than previous ones.
- To determine which elements must be refined, an estimate of the local error is necessary.
- As with the previously used duality argument, let $w \in V = H_0^1(\Omega)$ satisfy

$$a(v,w) = (u - u_h, v)_{L^2(\Omega)}, \quad \forall v \in V$$

A Posteriori Error Estimates and Adaptivity

- Then with $v = u - u_h$ and $a(u - u_h, w_h) = 0$ for $w_h \in V_h$,

$$\|u - u_h\|^2_{L^2(\Omega)} = a(u - u_h, w - w_h), \quad \forall w_h \in V_h$$

- In particular, take $w_h \in V_h$ to satisfy

$$a(v_h, w_h) = a(v_h, w), \quad \forall v_h \in V_h$$

so that

$$\|u-u_h\|^2_{L^2(\Omega)} = a(u, w-w_h)-a(u_h, w-w_h) = (f, w-w_h)_{L^2(\Omega)}$$

and u does not appear on the right!

- Writing integrals over elements,

$$\|u - u_h\|^2_{L^2(\Omega)} \leq \sum_{\mathrm{e}=1}^{\nu} \left(\int_{\Omega_\mathrm{e}} |f|^2 \right)^{\frac{1}{2}} \left(\int_{\Omega_\mathrm{e}} |w - w_h|^2 \right)^{\frac{1}{2}}$$

- Using techniques similar to those employed previously,

$$\left(\int_{\Omega_\mathrm{e}} |w - w_h|^2 \right)^{\frac{1}{2}} \leq \gamma h_\mathrm{e}^2 \|w\|_{H^2(\Omega)} \leq \gamma h_\mathrm{e}^2 \|u - u_h\|_{L^2(\Omega)}$$

for a known constant γ.

A Posteriori Error Estimates and Adaptivity

- Combining the above estimates gives

$$\|u - u_h\|_{L^2(\Omega)} \leq \gamma \sum_{e=1}^{\nu} h_e^2 \|f\|_{L^2(\Omega_e)}$$

- Thus the mesh $\{\Omega_e\}_{e=1}^{\nu}$ can be chosen so that

$$\gamma \sum_{e=1}^{\nu} h_e^2 \|f\|_{L^2(\Omega_e)} < \varepsilon$$

  for a given required tolerance ε.
- For more general cases:
  - Suppose Lu is a more general operator than $L = -\Delta$.
  - Then w_h is chosen as an interpolant of w.
  - Local integration by parts

$$(f, w - w_h)_{L^2(\Omega_e)} + a_e(u_h, w - w_h) = (f - Lu_h, w - w_h)_{L^2(\Omega_e)} + (n^\top A \nabla u_h, w - w_h)_{L^2(\partial\Omega_e)}$$

    requires estimating $w - w_h$ in $L^2(\Omega_e)$ and $L^2(\partial\Omega_e)$, and
  - $\|f\|_{L^2(\Omega_e)}$ is replaced by $\|R\|_{L^2(\Omega_e)}$, where $R = f + Lu_h$ is the local residual of the FEM.
  - So a mesh update $\{\Omega_{e'}\}_{e'=1}^{\nu'}$ must adapt to the current u_h.

Finite Element Spaces

- A *finite element method* (FEM) is a Galerkin method based upon approximation with piecewise polynomials.
- One begins with a *reference element* and studies polynomial interpolation on this element.
- Transformations of this reference element are used to cover and thereby partition the domain for a BVP.
- These steps are used to construct a global interpolant built from local ones, and this is used for error estimates.

Def: With the following properties, $(K, \mathcal{P}, \mathcal{N})$ is a finite element:

1. $K \subset \mathbb{R}^n$ (the *element domain* or just *element*) is a simply connected bounded open set with piecewise smooth boundary.
2. $\mathcal{P}$ (the space of *shape functions*) is a finite-dimensional space of functions defined on K.
3. $\mathcal{N} = \{N_i\}_{i=1}^d$ (the *nodal variables* or *degress of freedom*) is a basis of $\mathcal{P}^*$.

Finite Element Spaces

Def: Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element. The basis $\{\psi_i\}_{i=1}^d$ of $\mathcal{P}$ dual to $\mathcal{N}$, i.e., satisfying $N_i(\psi_j) = \delta_{i,j}$, is the *nodal basis* of $\mathcal{P}$.

- Example. For the square domain $\Omega = (0,1)^2$:
  - The reference element is $K = \hat{\Omega} = (0,1)^2$.
  - $\mathcal{P}$ is the space of bilinear functions.
  - $\mathcal{N} = \{N_i\}_{i=1}^4$ are the point evaluations at the corners of $\hat{\Omega}$, i.e., $N_{2i+j+1}(v) = v(i,j)$, $i,j = 0,1$, $\forall v \in \mathcal{P}$.
  - The nodal basis is given by the shape functions $\{\hat{\Phi}_i\}_{i=1}^4$.
- The following permits to verify that a given $\mathcal{N}$ is a basis for $\mathcal{P}^*$.

Lemma: Let $\mathcal{P}$ be a d-dimensional vector space and let $\{N_i\}_{i=1}^d$ be a subset of $\mathcal{P}^*$. Then these are equivalent:

1. $\{N_i\}_{i=1}^d$ is a basis for $\mathcal{P}^*$.
2. If $v \in \mathcal{P}$ satisfies $N_i(v) = 0$, $i = 1, \ldots, d$, then $v = 0$.

Proof: Let $\{\psi_j\}_{j=1}^d$ be a basis of $\mathcal{P}$. Then $\{N_i\}_{i=1}^d$ is a basis for $\mathcal{P}^*$ iff for any $L \in \mathcal{P}^*$, $\exists! \alpha_i$, $1 \leq i \leq d$, such that $L = \sum_{i=1}^d \alpha_i N_i$.

Constructing a Finite Element Space

Equivalently with the basis $\{\psi_j\}_{j=1}^d$ of $\mathcal{P}$, $L(\psi_j) = \sum_{i=1}^d \alpha_i N_i(\psi_j)$, $1 \le j \le d$. Define the matrix $B = \{N_i(\psi_j)\}_{i,j=1}^d$ and the vectors $\ell = \{L(\psi_j)\}_{j=1}^d$ and $a = \{\alpha_i\}_{i=1}^d$. Then (1) is equivalent to B being invertible to solve $Ba = \ell$. On the other hand, given any $v \in \mathcal{P}$, one can write $v = \sum_{j=1}^d \beta_j \psi_j$. Then (2) means that $\sum_{j=1}^d \beta_j N_i(\psi_j) = N_i(v) = 0$, $1 \le i \le d$, implies $v = 0$, or that $B^\top b = 0$ implies $b = \{\beta_j\}_{j=1}^d = 0$. Yet this too is equivalent to B being invertible. ■

- The plan to construct a finite element:
  - Choose an element domain K, e.g., a triangle.
  - Choose a polynomial space $\mathcal{P}$, e.g., linear functions.
  - Choose d degrees of freedom $\mathcal{N} = \{N_i\}_{i=1}^d$ where d is the dimension of $\mathcal{P}$, so the corresponding interpolation problem has a unique solution, e.g., point evaluations at corners of K.
  - Compute the nodal basis of $\mathcal{P}$ with respect to $\mathcal{N}$.
- The following is a useful tool to verify the unique solvability of the interpolation problem.

Solvability of the Interpolation Problem

Lemma: Let $L \neq 0$ be a linear function on $\mathbb{R}^n$ and let P be a polynomial of degree $d \geq 1$ with $P(x) = 0$ for all x with $L(x) = 0$. Then there exists a polynomial Q of degree $d-1$ such that $P = LQ$.

Proof: Note that affine transformations map the space of polynomials of degree d to itself. Without loss of generality, assume P vanishes on the hyperplane orthogonal to the x_n axis, i.e., $L(x) = x_n$ and $P(\hat{x}, 0) = 0$, where $\hat{x} = (x_1, \ldots, x_{n-1})$. Since the degree of P is d,

$$P(\hat{x}, x_n) = \sum_{j=0}^{d} \sum_{|\alpha| \leq d-j} c_{\alpha,j} \hat{x}^\alpha x_n^j.$$

For $x_n = 0$, this implies $0 = P(\hat{x}, 0) = \sum_{|\alpha| \leq d} c_{\alpha,0} \hat{x}^\alpha$ and therefore $c_{\alpha,0} = 0$ for all $|\alpha| \leq d$. Hence,

$$P(\hat{x}, x_n) = \sum_{j=1}^{d} \sum_{|\alpha| \leq d-j} c_{\alpha,j} \hat{x}^\alpha x_n^j = x_n \sum_{j=1}^{d} \sum_{|\alpha| \leq d-j} c_{\alpha,j} \hat{x}^\alpha x_n^{j-1} = x_n Q = LQ$$

where Q has degree $d-1$. ■

Examples of Finite Elements

- Consider first triangular elements in $\mathbb{R}^2$.
- Linear Lagrange elements: (figure forthcoming)
  - The dimension of $\mathcal{P} = P_1$ is 3: $\{1, x, y\}$.
  - Let $\mathcal{N} = \{N_i\}_{i=1}^3$ with $N_i(v) = v(z_i)$ where $\{z_i\}_{i=1}^3$ are the vertices of the triangle K.
  - For $i = 1, 2, 3$, let $L_i \in P_1$ vanish on the edge opposite z_i.
  - Suppose $v \in P_1$ with $v(z_i) = 0$, $i = 1, 2, 3$. Since v is linear, $v = 0$ on each edge of K. By the lemma, $\exists c \in P_0$ with $v = cL_1$. Then $0 = v(z_1) = cL_1(z_1)$, so $c = 0$ and $v = 0$. By the lemma, $\mathcal{N}$ is a basis for $\mathcal{P}^*$.
  - By definition, $(K, \mathcal{P}, \mathcal{N})$ is a finite element.
- Quadratic Lagrange elements: (figure forthcoming)
  - The dimension of $\mathcal{P} = P_2$ is 6: $\{1, x, y, x^2, xy, y^2\}$.
  - Let $\mathcal{N} = \{N_i\}_{i=1}^6$ with $N_i(v) = v(z_i)$ where $\{z_i\}_{i=1}^3$ are the vertices of the triangle K and $\{z_i\}_{i=4}^6$ are the midpoints of the edges of K.
  - For $i = 1, 2, 3$, let $L_i \in P_1$ vanish on the edge opposite v_i, and suppose z_{i+3} is the midpoint of this edge.

Triangular Elements

- Suppose $v \in P_2$ with $v(z_i) = 0$, $i = 1, \ldots, 6$. Since v is quadratic, $v = 0$ on each edge of K. By the lemma, $\exists Q_1 \in P_1$ with $v = Q_1 L_1$. Since $v = Q_1 L_1$ on the edge opposite z_2, $\exists c \in P_0$ with $Q_1 = cL_2$. Then $0 = v(z_6) = cL_1(z_6)L_2(z_6)$, so $c = 0$ and $v = 0$. By the lemma, $\mathcal{N}$ is a basis for $\mathcal{P}^*$.
- By definition, $(K, \mathcal{P}, \mathcal{N})$ is a finite element.

- Cubic Hermite elements: (figure forthcoming)
  - The dimension of $\mathcal{P} = P_3$ is 10: $\{1, x, y, \ldots, x^3, y^3\}$.
  - Let $\mathcal{N} = \{N_i\}_{i=1}^{10}$ with $N_i(v) = v(z_i)$, $i = 1, \ldots, 4$, where $\{z_i\}_{i=1}^{3}$ are the vertices of the triangle K and $z_4 = (z_1 + z_2 + z_3)/3$ is the *barycenter* of K. Then let $N_i(v) = \partial_x v(z_{i-4})$ for $i = 5, 6, 7$, and $N_i(v) = \partial_y v(z_{i-7})$ for $i = 8, 9, 10$.
  - For $i = 1, 2, 3$, let $L_i \in P_1$ vanish on the edge opposite v_i.
  - Suppose $v \in P_3$ with $N_i v = 0$, $i = 1, \ldots, 10$. Since v is cubic, $v = 0$ on each edge of K. Arguing as before, $v = cL_1L_2L_3$, but $0 = v(z_4) = cL_1(z_4)L_2(z_4)L_3(z_4)$, so $c = 0$ and $v = 0$. By the lemma, $\mathcal{N}$ is a basis for $\mathcal{P}^*$.
  - By definition, $(K, \mathcal{P}, \mathcal{N})$ is a finite element.

Rectangular Elements

- Here we understand the space

$$P_k \times P_k = \left\{ \sum_{j=1}^{m} c_j p_j(x) q_j(y) : c_j \in \mathbb{R}, m \in \mathbb{N}, x, y \in \mathbb{R}^n, p_j, q_j \in P_k \right\}$$

- Bilinear Lagrange elements: (figure forthcoming)
  - The dimension of $\mathcal{P} = P_1 \times P_1$ is 4.
  - Let $\mathcal{N} = \{N_i\}_{i=1}^4$ with $N_i(v) = v(z_i)$ where $\{z_i\}_{i=1}^4$ are the vertices of the rectangle K.
  - **Exercise**: Show for $v \in \mathcal{P}$ that $N_i(v) = 0$, $i = 1, \ldots, 4$, implies that $v = 0$, and hence, $\mathcal{N}$ is a basis for $\mathcal{P}^*$.
  - By definition, $(K, \mathcal{P}, \mathcal{N})$ is a finite element.
- Biquadratic Lagrange elements: (figure forthcoming)
  - The dimension of $\mathcal{P} = P_2 \times P_2$ is 9.
  - Let $\mathcal{N} = \{N_i\}_{i=1}^9$ with $N_i(v) = v(z_i)$ where $\{z_i\}_{i=1}^4$ are the vertices of the triangle K, $\{z_i\}_{i=5}^8$ are the midpoints of the edges of K and z_9 is the centroid of K.
  - **Exercise**: Show for $v \in \mathcal{P}$ that $N_i(v) = 0$, $i = 1, \ldots, 9$, implies that $v = 0$, and hence, $\mathcal{N}$ is a basis for $\mathcal{P}^*$.
  - By definition, $(K, \mathcal{P}, \mathcal{N})$ is a finite element.

The Interpolant

Def: Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element and let $\{\psi_i\}_{i=1}^d$ be the corresponding nodal basis of $\mathcal{P}$. For a given function v such that $N_i(v)$ is defined $\forall N_i \in \mathcal{N}$, the *local interpolant* of v is

$$\mathcal{I}_K v = \sum_{i=1}^{d} N_i(v)\psi_i.$$

Lemma: Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element and $\mathcal{I}_K$ the local interpolant operator. Then

a. $v \mapsto \mathcal{I}_K v$ is linear,
b. $N_i(\mathcal{I}_K v) = N_i(v)$, $1 \leq i \leq d$,
c. $\mathcal{I}_K v = v$, $\forall v \in \mathcal{P}$, i.e., $\mathcal{I}_K$ is a projection.

Proof: Claim (a) follows directly from the linearity of each N_i. Claim (b) follows using the definition of a nodal basis in

$$N_i(\mathcal{I}_K v) = N_i\left(\sum_{j=1}^{d} N_j(v)\psi_j\right) = \sum_{j=1}^{d} N_j(v)N_i(\psi_j) = \sum_{j=1}^{d} N_j(v)\delta_{i,j} = N_i(v)$$

$1 \leq i \leq d$, $\forall v$. This implies $N_i(v - \mathcal{I}_K v) = 0$, $\forall 1 \leq i \leq d$, so by Lemma 197, $v - \mathcal{I}_K v = 0$ when $v \in \mathcal{P}$, and claim (c) follows. ■

Triangulations

Def: A subdivision $\mathcal{T}$ of a bounded open $\Omega \subset \mathbb{R}^n$ is a finite collection of open sets $\{K_i\}$ such that

a. $K_i^\circ \cap K_j^\circ = \emptyset$, $i \neq j$,

b. $\cup_i \bar{K}_i = \bar{\Omega}$.

Def: Suppose a subdivision $\mathcal{T}$ of Ω is given such that for each K_i there is a finite element $(K_i, \mathcal{P}_i, \mathcal{N}_i)$ with local interpolant $\mathcal{I}_{K_i}$. Let m be the order of the highest partial derivative appearing in any nodal variable. Then, the *global interplant* $\mathcal{I}_\mathcal{T} v$ of $v \in \mathcal{C}^m(\bar{\Omega})$ on $\mathcal{T}$ is defined by

$$(\mathcal{I}_\mathcal{T} v)|_{K_i} = \mathcal{I}_{K_i} v, \quad K_i \in \mathcal{T}$$

Def: A *triangulation* of a bounded open set $\Omega \subset \mathbb{R}^2$ (necessarily polyhedral) is a subdivision $\mathcal{T}$ of Ω such that

a. every $K_i \in \mathcal{T}$ is a triangle and

b. no vertex of any triangle lies in the interior or on an edge of another triangle.

Remark: A similar definition may be given for $\Omega \subset \mathbb{R}^n$, $n \geq 3$, and the partition of Ω is still called a *triangulation*.

Continuity of Finite Element Space

Def: A global interpolant $\mathcal{I}_\mathcal{T}$ has *continuity order m* (i.e., is $\mathcal{C}^m$) if $\mathcal{I}_\mathcal{T} v \in C^m(\bar{\Omega})$, $\forall v \in \mathcal{C}^m(\bar{\Omega})$, and in this case, the space $V_\mathcal{T} = \{\mathcal{I}_\mathcal{T} v : v \in \mathcal{C}^m(\bar{\Omega})\}$ is called a $\mathcal{C}^m$ *finite element space*.

Theorem: The triangular Lagrange 200 and Hermite 201 elements are $\mathcal{C}^0$ elements.

Proof: For Lagrange elements set $k = 1$ and $m = 0$. For Hermite elements set $k = 3$ and $m = 1$. It need only be shown that for $v \in \mathcal{C}^m(\bar{\Omega})$ the global interpolant $\mathcal{I}_\mathcal{T} v$ is continuous across each edge. Let K_1 and K_2 be two triangles sharing an edge e, and $(K_1, \mathcal{P}_1, \mathcal{N}_1)$ and $(K_2, \mathcal{P}_2, \mathcal{N}_2)$ are the respective finite elements. For Lagrange as well as for Hermite elements, $\mathcal{N}_1$ and $\mathcal{N}_2$ coincide on e. Also, $\mathcal{P}_1 = \mathcal{P}_2 = P_k$. For a fixed $v \in \mathcal{C}^m(\bar{\Omega})$, set $w = \mathcal{I}_{K_1} v - \mathcal{I}_{K_2} v$, where $\mathcal{I}_{K_1} v$ and $\mathcal{I}_{K_2} v$ are extended as global polynomials outside K_1 and K_2, respectively. Hence, $w \in P_k$ and the restriction $w|_e$ is a 1D polynomial with $k + 1$ roots. Hence, $w|_e = 0$, and $\mathcal{I}_\mathcal{T} v$ is continuous across e. ■

Affine Equivalent Finite Elements

Exercise: Show that the bilinear and biquadratic Lagrange elements are also $\mathcal{C}^0$.

- With polynomials of degree 5 and with 21 nodal variables (including values of the function and its first and second derivatives at the vertices in addition to normal derivatives at the midpoints of the sides) the Argyris triangle is $\mathcal{C}^1$.
- With a bicubic Hermite basis of dimension 16 (including values of the function and its first derivatives as well as mixed partial derivatives at the vertices) the Bogner-Fox-Schmit rectangle is $\mathcal{C}^1$. (figures forthcoming)

Def: Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a finite element and $T : \mathbb{R}^n \to \mathbb{R}^n$ be an affine transformation ($T\hat{x} = A\hat{x} + b$, $A, A^{-1} \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$). The finite element $(K, \mathcal{P}, \mathcal{N})$ is *affine equivalent* to $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ if

a. $K = \{A\hat{x} + b : \hat{x} \in \hat{K}\}$,
b. $\mathcal{P} = \{\hat{p} \circ T^{-1} : \hat{p} \in \hat{\mathcal{P}}\}$,
c. $\mathcal{N} = \{N_i : N_i(p) = \hat{N}_i(p \circ T), \forall p \in \mathcal{P}\}$.

A triangulation $\mathcal{T}$ is *affine* if it consists of affine equivalent elements.

Affine Equivalent Finite Elements

Exercise: Show that the nodal bases of $\mathcal{P}$ and $\hat{\mathcal{P}}$ are related by $\hat{\psi}_i = \psi_i \circ T$.

- Hence, if nodal variables on edges are placed symmetrically, triangular Lagrange elements of the same order are affine equivalent, as are triangular Hermite elements.

Lemma: Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ and $(K, \mathcal{P}, \mathcal{N})$ be two affine equivalent finite elements related by the transformation T_K. Then

$$\mathcal{I}_{\hat{K}}(v \circ T_k) = (\mathcal{I}_k v) \circ T_k.$$

Proof: Let $\hat{\psi}_i$ and ψ_i be the nodal basis of $\hat{\mathcal{P}}$ and $\mathcal{P}$. By definition,

$$\mathcal{I}_{\hat{K}}(v \circ T_K) = \sum_{i=1}^{d} \hat{N}_i(v \circ T_k)\hat{\psi}_i = \sum_{i=1}^{d} N_i(v)(\psi_i \circ T_K) = (\mathcal{I}_K v) \circ T_K \quad \blacksquare$$

- Given a reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$, one can thus generate a triangulation $\mathcal{T}$ using affine equivalent elements.

Polynomial Interpolation in Sobolev Spaces

Lemma: If $v \in W^{k,p}(\Omega)$ satisfies $\partial^\alpha v = 0$, $\forall |\alpha| = k$, then $\exists \tilde{v} \in P_{k-1}$ such that $v = \tilde{v}$ a.e.

Proof: If $\partial^\alpha v = 0$ holds $\forall |\alpha| = k$, then $\partial^\beta \partial^\alpha v = 0$, $\forall \beta$. Hence, $v \in \cap_{m=1}^{\infty} W^{m,p}(\Omega)$. By the Sobolev Embedding Theorem, $\exists \tilde{v} \in C^k(\Omega)$ with $v = \tilde{v}$ a.e. Then $\partial^\alpha \tilde{v} = 0$, $\forall |\alpha| = k$, gives $\tilde{v} \in P_{k-1}$. ■

Lemma: $\forall v \in W^{k,p}(\Omega)$, $\exists! q \in P_{k-1}$ such that $\int_\Omega \partial^\alpha (v - q) = 0$, $\forall |\alpha| \leq k - 1$.

Proof: Writing $q = \sum_{|\beta| \leq k-1} c_\beta x^\beta \in P_{k-1}$, the condition amounts to solving the linear system,

$$\sum_{|\beta| \leq k-1} c_\beta \int_\Omega \partial^\alpha x^\beta = \int_\Omega \partial^\alpha v, \quad |\alpha| \leq k - 1.$$

To show that $M = \{\int_\Omega \partial^\alpha x^\beta\}_{|\alpha|,|\beta| \leq k-1}$ is non-singular, let $c = \{c_\beta\}_{|\beta| \leq k-1}$ satisfy $Mc = 0$. Then $q = \sum_{|\beta| \leq k-1} c_\beta x^\beta$ satisfies $\int_\Omega \partial^\alpha q = 0$, $\forall |\alpha| \leq k - 1$. Inserting $\alpha = (k - 1, 0, \ldots, 0)$ then $\alpha = (0, k - 1, \ldots, 0)$ down to $\alpha = (0, 0, \ldots, 1)$ then $\alpha = (0, 0, \ldots, 0)$ shows that $c = 0$. So M is invertible. ■

Bramble Hilbert Lemma

Lemma: $\exists c_0 > 0$ such that $\forall v \in W^{k,p}(\Omega)$ with $\int_\Omega \partial^\alpha v = 0$, $\forall |\alpha| \leq k-1$,

$$\|v\|_{W^{k,p}(\Omega)} \leq c_0 |v|_{W^{k,p}(\Omega)}$$

Proof: As the proof of Poincaré's Inequality 153 (**Exercise**). ■

Theorem (Bramble Hilbert): Let $F : W^{k,p}(\Omega) \to \mathbb{R}$ satisfy

a. $|F(v)| \leq c_1 \|v\|_{W^{k,p}(\Omega)}$ (boundedness)
b. $|F(u+v)| \leq c_2(|F(u)| + |F(v)|)$ (sublinearity)
c. $F(q) = 0, \forall q \in P_{k-1}$ (annihilation)

Then $\exists c > 0$ such that $\forall v \in W^{k,p}(\Omega)$

$$|F(v)| \leq c |v|_{W^{k,p}(\Omega)}$$

Proof: For any $v \in W^{k,p}(\Omega)$ and $q \in P_{k-1}$,

$$|F(v)| = |F(v-q+q)| \leq c_2(|F(v-q)| + |F(q)|) \leq c_1 c_2 \|v-q\|_{W^{k,p}(\Omega)}$$

Given $v \in W^{k,p}(\Omega)$, choose $q \in P_{k-1}$ by Lemma 208 so that $\int_\Omega \partial^\alpha (v-q) = 0, \forall |\alpha| \leq k-1$. Applying the Poincaré Lemma 153 above to $v - q$ gives

Interpolation Error Estimates

$$\|v-q\|_{W^{k,p}(\Omega)} \leq c_0 |v-q|_{W^{k,p}(\Omega)} = c_0 |v|_{W^{k,p}(\Omega)}. \quad \blacksquare$$

Theorem: Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element with $P_{k-1} \subset \mathcal{P}$ for some $k \geq 1$ where each $N \in \mathcal{N}$ is bounded on $W^{k,p}(K)$ for some $1 \leq p \leq \infty$. Then $\exists c = c(n, k, p, l, \mathcal{P}) > 0$ such that

$$|v - \mathcal{I}_K v|_{W^{l,p}(K)} \leq c |v|_{W^{k,p}(K)}, \quad \forall\, 0 \leq l \leq k.$$

Proof: The semi-norm $F(v) = |v - \mathcal{I}_K v|_{W^{l,p}(K)}$ is a sublinear functional on $W^{k,p}(K)$, $\forall\, 0 \leq l \leq k$. Let $\{\psi_i\}_{i=1}^d$ be the nodal basis of $\mathcal{P}$ with respect to $\mathcal{N}$. Since each $N_i \in \mathcal{N}$ is bounded on $W^{k,p}(K)$, F is bounded according to

$$|F(v)| \leq |v|_{W^{l,p}(K)} + |\mathcal{I}_K v|_{W^{l,p}(K)} \leq \|v\|_{W^{k,p}(K)} + \sum_{i=1}^{d} |N_i(v)| |\psi_i|_{W^{l,p}(K)} \leq$$

$$\|v\|_{W^{k,p}(K)} + \sum_{i=1}^{d} c_i \|v\|_{W^{k,p}(K)} |\psi_i|_{W^{l,p}(K)} \leq (1 + c \max_{1 \leq i \leq d} \|\psi_i\|_{W^{l,p}(K)}) \|v\|_{W^{k,p}(K)}$$

Since $\mathcal{I}_K q = q$, $\forall q \in \mathcal{P}$, $F(q) = 0$. The claim then follows from the Bramble Hilbert Lemma 209. $\blacksquare$

Interpolation Error Estimates

- For the interpolation error on an arbitrary finite element $(K, \mathcal{P}, \mathcal{N})$ it is assumed that it is generated by the affine transformation from the reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$, i.e., $T_K : \hat{K} \to K$, $T_K(\hat{x}) = A_K \hat{x} + b_K$ and $\hat{v} = v \circ T_K$ is the function v on K expressed in local coordinates on $\hat{K}$.

Lemma: Let $k \geq 0$ and $1 \leq p \leq \infty$. $\exists c > 0$ such that $\forall K$ and $v \in W^{k,p}(K)$, $\hat{v} = v \circ T_K$ satisfies

$$|\hat{v}|_{W^{k,p}(\hat{K})} \leq c \|A_K\|^k \det(A_K)^{-\frac{1}{p}} |v|_{W^{k,p}(K)}$$

$$|v|_{W^{k,p}(K)} \leq c \|A_K^{-1}\|^k \det(A_K)^{\frac{1}{p}} |\hat{v}|_{W^{k,p}(\hat{K})}$$

Proof: By the chain rule,

$$\partial_{\hat{x}_i} \hat{v} = \sum_{j=1}^{d} \partial_{x_j} v \partial_{\hat{x}_i} x_j = \sum_{j=1}^{d} (A_K)_{i,j} \partial_{x_j} v$$

By the integral transformation rule,

$$\int_{T_K(\hat{K})} v = \det(A_K) \int_{\hat{K}} (v \circ T_K)$$

Interpolation Error Estimates

Let $|\alpha| = k$. For a $c = c(n, k, p) > 0$,

$$\|\partial_{\hat{x}}^{\alpha} \hat{v}\|_{L^p(\hat{K})} \leq c\|A_K\|^k \sum_{|\beta|=k} \|\partial_x^{\beta}(v \circ T_K)\|_{L^p(\hat{K})} \leq c\|A_K\|^k \det(A_K)^{-\frac{1}{p}} |v|_{W^{k,p}(K)}$$

Summing over all $|\alpha| = k$ gives the estimate for $|\hat{v}|_{W^{k,p}(\hat{K})}$.

Arguing similarly using T_K^{-1} gives the estimate for $|v|_{W^{k,p}(K)}$. ■

Def: For a given domain K,

a. the *diameter* is
$$h_K = \max_{x_1, x_2 \in K} \|x_1 - x_2\|$$
b. the *incircle diameter* (i.e., of the largest ball in K) is
$$\rho_K = 2 \operatorname{argmax}_r \{x \in K : B_r(x) \subset K\}$$
c. the *condition number* is
$$\sigma_K = h_K / \rho_K.$$

Lemma: Let T_K be an affine map with $K = T_K(\hat{K})$. Then

$$|\det(A_K)| = \frac{\mathrm{vol}(K)}{\mathrm{vol}(\hat{K})}, \quad \|A_K\| \leq \frac{h_K}{\rho_{\hat{K}}}, \quad \|A_K^{-1}\| \leq \frac{h_{\hat{K}}}{\rho_K}.$$

Local Interpolation Error

Proof: The first property is purely geometrical. For the second,

$$\|A_K\| = \sup_{\|\hat{x}\|=1} \|A_K\hat{x}\| = \frac{1}{\rho_{\hat{K}}} \sup_{\|\hat{x}\|=\rho_{\hat{K}}} \|A_K\hat{x}\|$$

and for any $\hat{x}$ with $\|\hat{x}\| = \rho_{\hat{K}}$, $\exists \hat{x}_1, \hat{x}_2 \in \hat{K}$ with $\hat{x} = \hat{x}_1 - \hat{x}_2$ (e.g., $\hat{x}_0$ at the center of the incircle and $\hat{x}_{1,2} = \hat{x}_0 \pm \frac{1}{2}\hat{x}$) so

$$A_K\hat{x} = T_K\hat{x}_1 - T_K\hat{x}_2 = x_1 - x_2$$

for some $x_1, x_2 \in K$; thus, $\|A_K\hat{x}\| \leq h_K$. The remaining property is obtained by exchanging the roles of K and $\hat{K}$. ■

Theorem (local interpolation error): Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a finite element with $P_{k-1} \subset \hat{\mathcal{P}}$ for some $k \geq 1$ where each $N \in \mathcal{N}$ is bounded on $W^{k,p}(K)$ for some $1 \leq p \leq \infty$. For any finite element $(K, \mathcal{P}, \mathcal{N})$ affine equivalent to $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ by the affine transformation T_K, $\exists c > 0$, $c \neq c(K)$, such that for any $v \in W^{k,p}(K)$,

$$|v - \mathcal{I}_K v|_{W^{l,p}(K)} \leq c h_K^{k-l} \sigma_K^l |v|_{W^{k,p}(K)}, \quad \forall\, 0 \leq l \leq k.$$

Local Interpolation Error

Proof: Let $\hat{v} = v \circ T_K$. By Lemma 207, $(\mathcal{I}_K v) \circ T_K = \mathcal{I}_{\hat{K}} \hat{v}$. Hence, with the estimates of Lemma 211 and using Theorem 210,

$$|v - \mathcal{I}_K v|_{W^{l,p}(K)} \leq c\|A_K^{-1}\|^l |\det(A_K)|^{\frac{1}{p}} |\hat{v} - \mathcal{I}_{\hat{K}} \hat{v}|_{W^{l,p}(\hat{K})}$$

$$\leq c\|A_K^{-1}\|^l |\det(A_K)|^{\frac{1}{p}} |\hat{v}|_{W^{k,p}(\hat{K})} \leq c\|A_K^{-1}\|^l \|A_K\|^k |v|_{W^{k,p}(K)}$$

$$\leq c(\|A_K^{-1}\| \|A_K\|)^l \|A_K\|^{k-l} |v|_{W^{k,p}(K)}$$

Using the estimates from Lemma 212 for fixed $h_{\hat{K}}$ and $\rho_{\hat{K}}$ establishes the claim. ∎

- For a global interpolation error estimate a uniform bound on the condition number σ_K is required.
- This requires a further assumption on the triangulation.

Def: A triangulation $\mathcal{T}$ is *shape regular* if $\exists \kappa > 0$, $\kappa \neq \kappa(h)$, $h = \max_{K \in \mathcal{T}} h_K$, such that $\sigma_K \leq \kappa$, $\forall K \in \mathcal{T}$. (For triangles, all interior angles are bounded from below.)

Global Interpolation Error

Theorem (global interpolation error): Let $\mathcal{T}$ be a shape regular affine triangulation of $\Omega \subset \mathbb{R}^n$ with the reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ satisfying the conditions of Theorem 213 for some $k \geq 1$ and $1 \leq p \leq \infty$. Then $\exists c > 0$, $c \neq c(h)$, such that $\forall v \in W^{k,p}(\Omega)$,

$$\sum_{l=0}^{k} h^l \left(\sum_{K \in \mathcal{T}} |v - \mathcal{I}_K v|^p_{W^{l,p}(K)} \right)^{\frac{1}{p}} \leq c h^k |v|_{W^{k,p}(\Omega)}$$

for the case $1 \leq p < \infty$ and otherwise $\forall v \in W^{k,\infty}(\Omega)$,

$$\sum_{l=0}^{k} h^l \max_{K \in \mathcal{T}} |v - \mathcal{I}_K v|_{W^{l,\infty}(K)} \leq c h^k |v|_{W^{k,\infty}(\Omega)}$$

Proof: Use the upper bound on σ_K and sum over all elements. ■

Exercise: Determine if/how the techniques above can be applied for tensor product polynomial spaces to circumvent the earlier use of average values for the interpolant on the square.

Inverse Error Estimates

Theorem (local inverse estimate): Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a finite element with $\hat{\mathcal{P}} \subset W^{l,p}(\hat{K})$ for an $l \geq 0$ and $1 \leq p \leq \infty$. Let $(K, \mathcal{P}, \mathcal{N})$ be any element affine equivalent to $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ with $h_K \leq 1$ and affine transformation $T_{\hat{K}}$. Then $\exists c > 0$, $c \neq c(K)$, such that $\forall v_h \in \mathcal{P}$

$$\|v_h\|_{W^{l,p}(K)} \leq c h^{k-l} \|v_h\|_{W^{k,p}(K)}, \quad \forall\, 0 \leq k \leq l.$$

Def: A triangulation $\mathcal{T}$ is quasi-uniform if it is shape regular and $\exists \tau > 0$ such that $h_K \geq \tau h$, $\forall K \in \mathcal{T}$.

Theorem (global inverse estimate): Let $\mathcal{T}$ be a quasi-uniform triangulation of $\Omega \subset \mathbb{R}^n$ with reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ having $\hat{\mathcal{P}} \subset W^{l,p}(\hat{K})$ for an $l \geq 0$ and $1 \leq p \leq \infty$. Then $\exists c > 0$, $c \neq c(h)$, such that $\forall v_h \in V_h = \{v \in L^p(\Omega) : v|_K \in \mathcal{P}, \mathcal{P} \in \mathcal{T}\}$ and $\forall\, 0 \leq k \leq l$,

$$\Big[\sum_{K \in \mathcal{T}} \|v_h\|^p_{W^{l,p}(K)}\Big]^{\frac{1}{p}} \leq c h^{k-l} \Big[\sum_{K \in \mathcal{T}} \|v_h\|^p_{W^{k,p}(K)}\Big]^{\frac{1}{p}}$$

for the case $1 \leq p < \infty$ and otherwise

$$\max_{K \in \mathcal{T}} \|v_h\|_{W^{l,\infty}(K)} \leq c h^{k-l} \max_{K \in \mathcal{T}} \|v_h\|_{W^{k,\infty}(K)}$$

Error Estimates for Finite Element Approximations

- Consider only conforming FEM using Lagrange elements.
- Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a reference element for a shape regular triangulation $\mathcal{T}$ of Ω with affine equivalent elements and $P_{k-1} \subset \hat{\mathcal{P}}$ for some $k \geq 1$.
- Denote the affine transformation to the element $(K, \mathcal{P}, \mathcal{N})$ by $T_K(\hat{x}) = A_K \hat{x} + b_K$ and define the $\mathcal{C}^0$ finite element space
$$V_h = \{v_h \in \mathcal{C}(\bar{\Omega}) : (v \circ T_K)|_{\hat{K}} \in \hat{\mathcal{P}}, \forall K \in \mathcal{T}\}.$$
- Recall from Theorem 205 that, $\forall v \in \mathcal{C}^0(\bar{\Omega})$, $\mathcal{I}_{\mathcal{T}} v \in \mathcal{C}^0(\bar{\Omega})$, in fact, $\mathcal{I}_{\mathcal{T}} v \in V_h$, so by Céa's Lemma 173 the discretization error is bounded by the interpolation error.

Theorem: Let $u \in H^1(\Omega)$ be the weak solution to (34) for which the conditions of the Lax Milgram Theorem 158 are satisfied. Under the above assumptions on $\mathcal{T}$ and V_h let $u_h \in V_h$ be the Galerkin approximation to u. If $u \in H^m(\Omega)$ for $n/2 < m \leq k$, then $\exists c > 0$, $c \neq c(h, u)$, such that

$$\|u - u_h\|_{H^1(\Omega)} \leq c h^{m-1} |u|_{H^m(\Omega)}$$

Error Estimates for Finite Element Approximations

Proof: Since $m > n/2$, the Sobolev Embedding Theorem 151 implies $u \in C(\bar{\Omega})$, and hence the pointwise interpolant for u is well defined. Also, $\mathcal{I}_{\mathcal{T}} u \in V_h$, and Céa's Lemma 173 gives

$$\|u - u_h\|_{H^1(\Omega)} \leq c \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq c\|u - \mathcal{I}_{\mathcal{T}} u\|_{H^1(\Omega)}$$

Then setting $p = 2$ and $k = m$ in Theorem 215 gives

$$\|u - \mathcal{I}_{\mathcal{T}} u\|_{H^1(\Omega)} \leq ch^{m-1}|u|_{H^m(\Omega)} \quad \blacksquare$$

Theorem: Assume the primal solution to (34) is $u \in H^m(\Omega)$ for $m > n/2$. Assume the adjoint problem to (34) is well posed and, in particular, for $g \in L^2(\Omega)$ the solution ϕ_g satisfies $\|\phi_g\|_{H^2(\Omega)} \leq c\|g\|_{L^2(\Omega)}$. Under the above assumptions on $\mathcal{T}$ and V_h with $k \geq m$, let $u_h \in V_h$ be the Galerkin approximation to u. Then $\exists c > 0$, $c \neq c(h, u)$, such that

$$\|u - u_h\|_{L^2(\Omega)} \leq ch^m|u|_{H^m(\Omega)}$$

A Posteriori Error Estimates

Proof: The Aubin-Nitsche Theorem 175 gives

$$\|u-u_h\|_{L^2(\Omega)} \leq c\|u-u_h\|_{H^1(\Omega)} \sup_{g\in L^2(\Omega)} \left(\|g\|^{-1}_{L^2(\Omega)} \inf_{v_h\in V_h} \|\phi_g - v_h\|_{H^1(\Omega)} \right)$$

Using Theorem 215 gives

$$\inf_{v_h\in V_h} \|\phi_g - v_h\|_{H^1(\Omega)} \leq \|\phi_g - \mathcal{I}_{\mathcal{T}}\phi_g\|_{H^1(\Omega)} \leq ch|\phi_g|_{H^2(\Omega)} \leq ch\|g\|_{L^2(\Omega)}$$

Combining these estimates with that for $\|u - u_h\|_{H^1(\Omega)}$ from the previous theorem establishes the claim. ■

- For *a posteriori* error estimates, consider the simplified case in (34) that $Lu = -\nabla(\alpha\nabla u)$, $f \in L^2(\Omega)$, $Bu = u$, $g = 0$, $V = H^1_0(\Omega)$ and the weak formulation is to seek $u \in H^1_0(\Omega)$ such that

  $$a(u, v) = (\alpha\nabla u, \nabla v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} = b(v), \quad \forall v \in H^1_0(\Omega)$$

  for $\alpha \in \mathcal{C}^1(\bar{\Omega})$ with $\alpha_1 \geq \alpha(x) \geq \alpha_0 > 0$, $\forall x \in \Omega$.
- Let $V_h \subset V$ and suppose u_h is the Ritz Galerkin approximation.
- The following arguments are analogous for other cases of (34).

Residual Based *A Posteriori* Error Estimates

- Residual based estimates give an $H^1(\Omega)$ error estimate.
- An *a posteriori* error estimate: $\left(\|\cdot\|_{H_0^1(\Omega)} = |\cdot|_{H^1(\Omega)}\right)$

$$
\begin{aligned}
\alpha_0 |u - u_h|_{H^1(\Omega)} &\leq \frac{a(u-u_h, u-u_h)}{|u-u_h|_{H^1(\Omega)}} \leq \sup_{w \in H_0^1(\Omega)} \frac{a(u-u_h, w)}{|w|_{H^1(\Omega)}} \\
&= \sup_{w \in H_0^1(\Omega)} \frac{a(u, w) - (\alpha \nabla u_h, \nabla w)_{L^2(\Omega)}}{|w|_{H^1(\Omega)}} \\
&= \sup_{w \in H_0^1(\Omega)} \frac{(f, w)_{L^2(\Omega)} - (-\nabla \cdot (\alpha \nabla u_h), w)_{H^{-1}(\Omega), H_0^1(\Omega)}}{|w|_{H^1(\Omega)}} \\
&= \sup_{w \in H_0^1(\Omega)} \frac{(f + \nabla \cdot (\alpha \nabla u_h), w)_{H^{-1}(\Omega), H_0^1(\Omega)}}{\|w\|_{H_0^1(\Omega)}} \\
&= \|f + \nabla \cdot (\alpha \nabla u_h)\|_{H^{-1}(\Omega)}
\end{aligned}
$$

but this is inconvenient due to the $H^{-1}(\Omega)$ norm (sup).

- Thus, element-wise integration by parts is used to localize the error.

Residual Based *A Posteriori* Error Estimates

- Let $\mathcal{T}_h$ be the triangulation corresponding to V_h and let $\partial\mathcal{T}_h$ be the set of faces of all $K \in \mathcal{T}_h$.
- The set of all interior faces is $\Gamma_h = \{F \in \partial\mathcal{T}_h : F \cap \partial\Omega = \emptyset\}$.
- For $F \in \Gamma_h$ with $F = \bar{K}_1 \cap \bar{K}_2$, let n_1 and n_2 be unit outward normal to K_1 and K_2, respectively.
- The jump in normal derivative for $w_h \in V_h$ across F is $[\![\nabla w_h \cdot n]\!] = \nabla w_h|_{K_1} \cdot n_1 + \nabla w_h|_{K_2} \cdot n_2$
- Integrating by parts element-wise,

$$a(u-u_h, w) = (f, w)_{L^2(\Omega)} - a(u_h, w) = (f, w)_{L^2(\Omega)} - \sum_{K\in\mathcal{T}_h} \int_K \alpha \nabla u_h \cdot \nabla w$$

$$= \sum_{K\in\mathcal{T}_h} \left(\int_K (f + \nabla\cdot(\alpha\nabla u_h))w - \sum_{F\in\partial K} \int_F \alpha(\nabla u_h \cdot n)w \right)$$

$$= \sum_{K\in\mathcal{T}_h} \int_K (f + \nabla\cdot(\alpha\nabla u_h))w - \sum_{F\in\Gamma_h} \int_F [\![\alpha(\nabla u_h \cdot n)]\!] w$$

where $\int_F \alpha(\nabla u_h \cdot n)w$ is well defined since $u_h|_K \in \mathcal{P} \Rightarrow \alpha(\nabla u_h \cdot n) \in L^2(F)$ and $w \in H_0^1(\Omega) \Rightarrow w \in L^2(F)$.

Residual Based *A Posteriori* Error Estimates

- To avoid $w \in H_0^1(\Omega)$ in the last estimate, as well as in the denominator of the previous $H^{-1}(\Omega)$ estimate, we would like to use $\mathcal{I}_{\mathcal{T}_h} w$ and apply interpolation estimates.
- Yet $w \in H_0^1(\Omega)$ does not in general give $w \in \mathcal{C}(\bar{\Omega})$ and so pointwise evaluation is not necessarily defined.
- Thus, interpolation is combined with projection.
- For $K \in \mathcal{T}_h$ define the elements touching K,
$$\omega_K = \cup\{\bar{K}' \in \mathcal{T}_h : \bar{K}' \cap \bar{K} \neq \emptyset\}.$$
- For every face F of K define the elements sharing F,
$$\omega_F = \cup\{\bar{K}' \in \mathcal{T}_h : F \in \bar{K}'\} \subset \omega_K.$$
- For every node z of K define the elements sharing z,
$$\omega_z = \cup\{\bar{K}' \in \mathcal{T}_h : z \in \bar{K}'\} \subset \omega_K.$$
- The $L^2(\omega_z)$ projection of $v \in H^1(\Omega)$ onto P_k is π_z satisfying
$$\int_{\omega_z} [\pi_z v - v] q = 0, \quad \forall q \in P_k.$$
- For $z \in \partial\Omega$, $\pi_z v = 0$ for the homogeneous Dirichlet BCs.
- The local *Clément Interpolant* of $v \in H_0^1(\Omega)$ is
$$\mathcal{I}_C v = \sum_{i=1}^{d} N_i(\pi_z(v))\psi_i$$

Residual Based *A Posteriori* Error Estimates

- With the Bramble Hilbert Lemma [209] it can be shown, $\exists c > 0$ (known!), s.t. $\forall w \in H_0^1(\Omega)$, $\forall K \in \mathcal{T}_h$, $\forall F \in \partial K$,

$$\|w - \mathcal{I}_C w\|_{L^2(K)} + h_K^{\frac{1}{2}} \|w - \mathcal{I}_C w\|_{L^2(F)} \leq c h_K |w|_{H^1(\omega_K)}$$

- It is assumed that $\mathcal{I}_C w \in V_h$, so $a(u - u_h, \mathcal{I}_C w) = 0$ holds. Hence, using the previous estimates, $(\|\cdot\|_{H_0^1(\Omega)} = |\cdot|_{H^1(\Omega)})$

$$|u - u_h|_{H^1(\Omega)} = \frac{1}{\alpha_0} \sup_{w \in H_0^1(\Omega)} \frac{a(u - u_h, w - \mathcal{I}_C w)}{|w|_{H^1(\Omega)}}$$

$$\leq \frac{1}{\alpha_0} \sup_{w \in H_0^1(\Omega)} \frac{1}{|w|_{H^1(\Omega)}} \left(\sum_{K \in \mathcal{T}_h} \|f + \nabla \cdot (\alpha \nabla u_h)\|_{L^2(K)} \|w - \mathcal{I}_C w\|_{L^2(K)} \right.$$

$$\left. + \sum_{F \in \Gamma_h} \|[\![\alpha(\nabla u_h \cdot n)]\!]\|_{L^2(F)} \|w - \mathcal{I}_C w\|_{L^2(F)} \right)$$

$$\leq c \sup_{w \in H_0^1(\Omega)} \frac{1}{|w|_{H^1(\Omega)}} \left(\sum_{K \in \mathcal{T}_h} h_K |w|_{H^1(\omega_K)} \|f + \nabla \cdot (\alpha \nabla u_h)\|_{L^2(K)} \right.$$

$$\left. + \sum_{F \in \Gamma_h} h_K^{\frac{1}{2}} |w|_{H^1(\omega_F)} \|[\![\alpha(\nabla u_h \cdot n)]\!]\|_{L^2(F)} \right)$$

Duality Based *A Posteriori* Error Estimates

- Since $|w|_{H^1(\omega_K)}, |w|_{H^1(\omega_F)} \leq |w|_{H^1(\Omega)}$, $\forall K \in \mathcal{T}_h$, $\forall F \in \partial\mathcal{T}_h$,

$$\|u-u_h\|_{H^1(\Omega)} \leq c \sum_{K\in\mathcal{T}_h} h_K \|f+\nabla\cdot(\alpha\nabla u_h)\|_{L^2(K)} + c \sum_{F\in\Gamma_h} h_K^{\frac{1}{2}} \|[\![\alpha(\nabla u_h \cdot n)]\!]\|_{L^2(F)}$$

  where c depends upon the Clément Interpolant.
- This estimate contains known quantities on the right side which can be used to refine a mesh.
- If $\alpha \in \mathcal{C}^1(\bar{\Omega})$, the Aubin Nitsche trick can be used to obtain an *a posteriori* error estimate in $L^2(\Omega)$ and thereby avoid the Clément Interpolant.
- Let $w \in V = H_0^1(\Omega)$ be the solution to

$$a(v, w) = (u - u_h, v), \quad \forall v \in H_0^1(\Omega)$$

- By the symmetry of a, the dual problem is well posed. In fact, $w \in H^2(\Omega) \cap H_0^1(\Omega)$ and hence for $n = 2$, $w \in \mathcal{C}(\bar{\Omega})$, and pointwise evaluation is defined for the interpolant $\mathcal{I}_{\mathcal{T}_h}$.
- With $v = u - u_h \in V$, $w_h = \mathcal{I}_{\mathcal{T}_h} w \in V_h$ and $a(u - u_h, w_h) = 0$,

$$\begin{aligned} \|u - u_h\|^2_{L^2(\Omega)} &= (u - u_h, u - u_h)_{L^2(\Omega)} = a(u - u_h, w - w_h) \\ &= (f, w - w_h)_{L^2(\Omega)} - a(u_h, w - w_h) \end{aligned}$$

Duality Based *A Posteriori* Error Estimates

- Integrating by parts element-wise gives

$$\begin{aligned}\|u-u_h\|^2_{L^2(\Omega)} &\leq \sum_{K\in\mathcal{T}_h} \|f+\nabla\cdot(\alpha\nabla u_h)\|_{L^2(K)}\|w-w_h\|_{L^2(K)} \\ &+ \sum_{F\in\Gamma_h} \|[\![\alpha(\nabla u_h\cdot n)]\!]\|_{L^2(F)}\|w-w_h\|_{L^2(F)}\end{aligned}$$

- Using Theorem 213 with $k=2$, $l=0$ and $p=2$ gives

$$\|w-w_h\|_{L^2(K)} \leq ch_K^2\|w\|_{H^2(K)}$$

- Using the Bramble Hilbert Lemma 209 and transformations between F, K and $\hat{F}, \hat{K}$ (**Exercise**: See, e.g., 242),

$$\|w-w_h\|_{L^2(F)} \leq ch_K^{3/2}\|w\|_{H^2(K)}$$

- From Theorem 160,

$$\|w\|_{H^2(\Omega)} \leq c\|u-u_h\|_{L^2(\Omega)}$$

- Combining the estimates above gives

$$\|u-u_h\|_{L^2(\Omega)} \leq c\sum_{K\in\mathcal{T}_h} h_K^2\|f+\nabla\cdot(\alpha\nabla u_h)\|_{L^2(K)} + c\sum_{F\in\Gamma_h} h_K^{3/2}\|[\![\alpha(\nabla u_h\cdot n)]\!]\|_{L^2(F)}$$

- This estimate contains known quantities on the right side which can be used to refine a mesh.

Implementation

- The goal here is only to use model problems to achieve an understanding of the structure of professional software packages, whose existence circumvents the need to write a finite element solver from scratch.
- The focus here is on triangular Lagrange and Hermite elements on polygonal domains.
- Constructions can be extended to higher-dimensional and quadrilateral elements.
- The geometric information about the triangulation is stored in arrays such as the following:
  - *nodes* contains coordinates of vertices

    ```
    nodes(i) = (x(i),y(i))
    ```
  - *elements* contains references to element nodes

    ```
    elements(i,:) = (i1(i),i2(i),i3(i))
    ```
  - Here `i1(:)`, `i2(:)` and `i3(:)` are pointer arrays, so `elements(i,1)` refers to `nodes(i1(i))`.
  - An entry `nodes(.)` appears twice when both the function and its gradient are evaluated.

Assembly

- For Dirichlet BCs, *boundary points* are stored, e.g., in `bndry_nodes`
- For Neumann BCs, *boundary faces* are stored, e.g., in `bndry_faces`

- Mesh Generation is an active research area, but for a uniform mesh on a simple geometry,
  - a mesh can be generated by hand, or
  - by using `delaunay` in Matlab (given `nodes`), or
  - by using `distmesh` in Matlab to create a mesh from a geometric description of the boundary.
- *Assembly* of the stiffness matrix $\{a(\Phi_i, \Phi_j)\}$ is achieved most efficiently element-wise by transforming to a reference element.
- Consider the *reference element*
$$\hat{K} = \{(\xi_1, \xi_2) \in \mathbb{R}^2 : 0 \leq \xi_1, \xi_2 \leq 1, \xi_1 + \xi_2 \leq 1\}$$
with vertices in the following order
$$z_1 = (0,0),\ z_2 = (1,0) \text{ and } z_3 = (0,1)$$

Assembly

- For a triangle K with the ordered set of vertices
$$(x_1, y_1), \quad (x_2, y_2), \quad (x_3, y_3)$$
the affine transformation from $\hat{K}$ to K is
$$T_K(\xi) = A_K \xi + b_K, \quad A_K = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}, \quad b_K = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$$
- Given the nodal variables $\hat{\mathcal{N}} = \{\hat{N}_i\}$ the nodal basis functions $\{\hat{\psi}_i\}$ are computed from the conditions $\hat{N}_i(\hat{\psi}_j) = \delta_{ij}$. For the linear Lagrange element, these are
$$1 - \xi_1 - \xi_2, \quad \xi_1, \quad \xi_2$$
- If the coefficients of the bilinear form a are constant, the integrals on the reference element can be computed exactly, noting $\hat{\psi}_i = \psi_i \circ T_K$ and
$$\nabla \psi|_K(x) = A_K^{-\top} \nabla \hat{\psi}(\xi)$$
- If the coefficients do not permit exact integration, numerical *quadrature* must be used,
$$\int_K v(x) = \sum_{k=1}^{r} w_k v(x_k)$$
where x_k are *quadrature nodes* and w_k are *quadrature weights*.

Quadrature

- Performing numerical quadrature replaces the bilinear form a with an approximation a_h, so
  - the quadrature must be performed in such a way that the approximate formulation is well posed and
  - that the quadrature error is negligible compared to other approximation errors.

Theorem (effect of quadrature): Let $\mathcal{T}_h$ be a shape regular affine triangulation with $P_1 \subset \hat{\mathcal{P}} \subset P_k$ for $k \geq 1$. Suppose the quadrature on $\hat{K}$ is of order $2k-2$, that all weights are positive and that h is sufficiently small. Then the discrete problem is well posed.

Also if surface integrals are approximated by a quadrature rule of order $2k-1$ and the conditions of Theorem 213 hold, then $\exists c > 0$ such that for $f \in H^{k-1}(\Omega)$, $g \in H^k(\partial\Omega)$ and h sufficiently small,

$$\|u - u_h\|_{H^1(\Omega)} \leq ch^{k-1}(\|u\|_{H^k(\Omega)}) + \|f\|_{H^{k-1}(\Omega)} + \|g\|_{H^k(\partial\Omega)}) \quad \blacksquare$$

Quadrature

- Rule of thumb:
  - quadrature should be exact for second order derivatives if coefficients are constant,
  - for linear elements (constant gradients) quadrature of order 0 (i.e., the midpoint rule) is sufficient to obtain an order h error estimate.
- For higher order elements, Gauß quadrature is usually used, simplified using *barycentric coordinates*:
  - If the vertices of K are $\{(x_i, y_i)\}_{i=1}^3$, the barycentric coordinates $(\zeta_1, \zeta_2, \zeta_3)$ of $(x, y) \in K$ are determined by
$$\zeta_1, \zeta_2, \zeta_3 \in [0, 1], \quad \zeta_1 + \zeta_2 + \zeta_3 = 1$$
$$(x, y) = \zeta_1(x_1, y_1) + \zeta_2(x_2, y_2) + \zeta_3(x_3, y_3)$$
  - These are invariant under affine transformation: if $\xi \in \hat{K}$ has barycentric coordinates $(\zeta_1, \zeta_2, \zeta_3)$ with respect to vertices of $\hat{K}$, then $x = T_K\xi$ has the same coordinates with respect to the vertices of K.
- The element contributions of the local basis functions are

$$\int_K \nabla_x \psi_i(x)^\top A(x) \nabla_x \psi_j(x) \approx \det(A_K) \sum_{k=1}^{n_l} w_k (A_K^{-\top} \nabla_\xi \hat{\psi}_i(\xi_k))^\top A(x_k) (A_K^{-\top} \nabla_\xi \hat{\psi}_j(\xi_k))$$

Quadrature

where

- $A(x)$ is the matrix of the highest order term $\nabla \cdot (A \nabla u)$,
- n_l is the number of Gauß nodes,
- x_k and ξ_k are the Gauß nodes in K and $\hat{K}$ respectively, and
- ψ_i and $\hat{\psi}_i = \psi_i \circ T_K$ are the basis functions on K and $\hat{K}$ respectively.

and other integrals in the bilinear form a and linear form b are computed similarly.

- The most flexible and efficient method to implement (non-)homogeneous Dirichlet BCs, $u = g$ on $\partial\Omega$, is to construct the stiffness matrix `K` and load vector `F` as above, and then replace each row in `K` and entry in `F` corresponding to a node in `bndry_nodes` as follows:

```
for i=1:length(bndry_nodes)
  k = bndry_nodes(i)
  K(k,j) = 0 for all j
  F(k) = g(nodes(k))
end
```

Quadrature

- For other BCs with boundary integrals, contributions are assembled for each face, where the loop below over elements is replaced by a loop over `bndry_faces` and 1D Gauß quadrature is used.
- Algorithm for the FEM with Lagrange triangles:

```
Input:  nodes, elements, data a(i,j), b(j), c, f
Compute Gauß nodes ξ_l and weights w_l on K̂
Compute ψ̂_i(ξ_l) and ∇_ξ ψ̂_i(ξ_l)
Set K(i,j) = F(j) = 0 for all i,j
for k = 1:length(elements)
   Compute T_K, det(A_K) for K = elements(k)
   Evaluate data at transformed Gauß nodes T_K(ξ_l)
   Compute a(ψ_i,ψ_j),(f,ψ_j),∀i,j on K with Gauß quadrature
   for i,j=1:d
      Set r = elements(k,i), s = elements(k,j)
      Set K(r,s) ← K(r,s) + a(ψ_i,ψ_j), F(s) ← F(s) + (f,ψ_j)
        Output:  K, F
```

Generalized Galerkin Approach

- Cases where preceeding constructions do not apply:
  - *Petrov-Galerkin*: The weak formulation is to seek $u \in U$ satisfying $a(u, v) = b(v)$, $\forall v \in V$, where $V \neq U$.
  - *Non-conforming*: The discrete problem is to find $u_h \in U_h$ satisfying $a(u_h, v_h) = b(v_h)$, $\forall v_h \in V_h$, and $U_h \not\subset U$, $V_h \not\subset V$.
  - *Non-consistent*: The discrete problem is to find $u_h \in U_h$ satisfying $a_h(u_h, v_n) = b_h(v)$, $\forall v_h \in V_h$, and a_h, b_h are not defined on $U \times V$ or V.
- For a more general formulation let U, V be Banach spaces with duals U^*, V^* and V is reflexive. Given $a : U \times V \to \mathbb{R}$ and $b \in V^*$, the weak formulation is to seek $u \in U$ satisfying

$$a(u, v) = b(v), \quad \forall v \in V \tag{43}$$

and existence of a solution is given by the following generalization of the Lax Milgram Theorem.

Theorem (Banach Nečas Babuška): Under the conditions

a. (inf-sup) $\exists c_1 > 0$ such that

$$\inf_{u \in U} \sup_{v \in V} \frac{a(u, v)}{\|u\|_U \|v\|_V} \geq c_1$$

Generalized Galerkin Approach

b. (continuity) $\exists c_2, c_3 > 0$ such that

$$a(u,v) \le c_2 \|u\|_U \|v\|_V, \quad b(v) \le c_3 \|v\|_V, \quad \forall u \in U, \quad \forall v \in V$$

c. (injectivity) For any $v \in V$,

$$a(u,v) = 0, \quad \forall u \in U \quad \Rightarrow \quad v = 0$$

there exists a unique solution $u \in U$ to (43) satisfying $\|u\|_U \le \|b\|_{V^*}/c_1$. ■

- Note that if $U = V$, then coercivity of a implies the inf-sup condition as well as injectivity, giving the Lax Milgram Theorem.
- For the *non-conforming* Galerkin approach, set finite dimensional approximation spaces $U_h \approx U$ and $V_h \approx V$, introduce a bilinear form $a_h : U_h \times V_h \to \mathbb{R}$, a linear form $b_h : V_h \to \mathbb{R}$ and seek a numerical solution $u_h \in U_h$ satisfying

$$a_h(u_h, v_h) = b_h(v_h), \quad \forall v_h \in V_h \tag{44}$$

Generalized Galerkin Approach

- Although $U_h \subset U$ and $V_h \subset V$ are not required, an error estimate requires a way to compare elements in U and U_h.
- For this, suppose $\exists U_*$, a subspace of U containing the exact solution $u \in U_*$.
- Also, define the space
$$U(h) = U_* + U_h = \{w + w_h : w \in U_*, w_h \in U_h\}$$
endowed with a norm $\|\cdot\|_{U(h)}$ satisfying

  a. $\|u_h\|_{U(h)} = \|u_h\|_{U_h}, \forall u_h \in U_h,$
  b. $\|u\|_{U(h)} \leq c\|u\|_U, \forall u \in U_*.$

  e.g.,
$$\|u\|_{U(h)} = \inf_{w \in U_*, w_h \in U_h, u = w + w_h} (\|w\|_U + \|w_h\|_{U_h})$$

- Since $U_h \subset U$ and $V_h \subset V$ are not assumed, existence of a solution to the discrete problem must be established separately:

Generalized Galerkin Approach

Theorem Let U_h and V_h be finite dimensional with $\dim(U_h) = \dim(V_h)$. Under the conditions

a. (inf-sup) $\exists c_1 > 0$ such that

$$\inf_{u_h \in U_h} \sup_{v_h \in V_h} \frac{a_h(u_h, v_h)}{\|u_h\|_{U_h} \|v_h\|_{V_h}} \geq c_1$$

b. (continuity) $\exists c_2, c_3 > 0$ such that $\forall u_h \in U_h$, $\forall v_h \in V_h$,

$$a_h(u_h, v_h) \leq c_2 \|u_h\|_{U_h} \|v_h\|_{V_h}, \quad b_h(v_h) \leq c_3 \|v_h\|_{V_h}$$

there exists a unique solution $u_h \in U_h$ to (44) satisfying $\|u_h\|_{U_h} \leq \|b_h\|_{V_h^*} / c_1$. ■

- Note that the inf-sup condition implies the invertibility of the stiffness matrix, and hence injectivity follows.
- The counterpart condition of coercivity in the Lax Milgram Theorem when applied to the discrete problem implies that the stiffness matrix is SPD.
- Error estimates for non-conforming methods are based upon the following generalizations of Céa's Lemma [173].

Generalized Galerkin Approach

Theorem (First Strang Lemma): Let the conditions of Theorem 236 be satisfied. Assume also that

a. $U_h \subset U = U(h)$ and $V_h \subset V$, $\quad (\|\cdot\|_{U_h} = \|\cdot\|_{U(h)} = \|\cdot\|_U, \|\cdot\|_{V_h} = \|\cdot\|_V)$

b. $\exists c_4 > 0$, $c \neq c(h)$, such that

$$a(u, v_h) \leq c_4 \|u\|_{U(h)} \|v_h\|_{V_h}, \quad \forall u \in U, \quad \forall v_h \in V_h$$

Then the solutions u and u_h to (43) and (44) satisfy

$$\|u - u_h\|_{U(h)} \leq \frac{1}{c_1} \sup_{v_h \in V_h} \frac{|b(v_h) - b_h(v_h)|}{\|v_h\|_{V_h}}$$

$$+ \inf_{w_h \in U_h} \left[\left(1 + \frac{c_4}{c_1}\right) \|u - w_h\|_{U(h)} + \frac{1}{c_1} \sup_{v_h \in V_h} \frac{|a(w_h, v_h) - a_h(w_h, v_h)|}{\|v_h\|_{V_h}} \right]$$

Proof: Let $w_h \in U_h$ be given. By the discrete inf-sup condition,

$$c_1 \|u_h - w_h\|_{U(h)} \leq \sup_{v_h \in V_h} \frac{a_h(u_h - w_h, v_h)}{\|v_h\|_{V_h}}$$

Generalized Galerkin Approach

Using (43) and (44),

$$a_h(u_h - w_h, v_h) = a(u - w_h, v_h) + a(w_h, v_h) - a_h(w_h, v_h) + b_h(v_h) - b(v_h).$$

Using this in the last estimate and applying assumption (b),

$$c_1 \|u_h - w_h\|_{U(h)} \leq c_4 \|u - w_h\|_{U(h)} + \sup_{v_h \in V_h} \frac{|a(w_h, v_h) - a_h(w_h, v_h)|}{\|v_h\|_{V_h}}$$

$$+ \sup_{v_h \in V_h} \frac{|b(v_h) - b_h(v_h)|}{\|v_h\|_{V_h}}$$

Using the triangle inequality,

$$\|u - u_h\|_{U(h)} \leq \|u - w_h\|_{U(h)} + \|u_h - w_h\|_{U(h)}$$

the claim follows after taking the inf over $w_h \in U_h$. ■

Theorem (second Strang Lemma): Let the conditions of Theorem 236 be satisfied. Assume also that a_h can be extended to $U(h) \times V_h$. Assume further, $\exists c_4 > 0$, $c \neq c(h)$, such that

$$|a_h(u, v_h)| \leq c_4 \|u\|_{U(h)} \|v_h\|_{V_h}, \quad \forall u \in U(h), \quad \forall v_h \in V_h.$$

Generalized Galerkin Approach

Then the solutions u and u_h to (43) and (44) satisfy

$$\|u-u_h\|_{U(h)} \leq \left(1+\frac{c_4}{c_1}\right) \inf_{w_h \in U_h} \|u-w_h\|_{U(h)} + \frac{1}{c_1} \sup_{v_h \in V_h} \frac{|b(v_h)-a_h(u,v_h)|}{\|v_h\|_{V_h}}$$

Proof: Let $w_h \in U_h$ be given. By the discrete inf-sup condition,

$$c_1 \|u_h - w_h\|_{U(h)} \leq \sup_{v_h \in V_h} \frac{a_h(u_h - w_h, v_h)}{\|v_h\|_{V_h}}$$

Using (44) with $v_h \in V_h$,

$$\begin{aligned} a_h(u_h - w_h, v_h) &= a_h(u_h - u, v_h) + a_h(u - w_h, v_h) \\ &= b_h(v_h) - a_h(u, v_h) + a_h(u - w_h, v_h). \end{aligned}$$

The assumption on a_h implies

$$c_1 \|u_h - w_h\|_{U(h)} \leq \sup_{v_h \in V_h} \frac{|b_h(v_h) - a_h(u, v_h)|}{\|v_h\|_{V_h}} + c_4 \|u - w_h\|_{U(h)}$$

Using the triangle inequality,

$$\|u - u_h\|_{U(h)} \leq \|u - w_h\|_{U(h)} + \|u_h - w_h\|_{U(h)}$$

the claim follows after taking the inf over $w_h \in U_h$. ■

Generalized Galerkin Approach

- The first Strang Lemma [237] can be applied to show the effect of quadrature on the Galerkin approximation.
- For simplicity, consider to find $u \in H_0^1(\Omega) = U$ s.t.

$$a(u, v) = (\alpha \nabla u, \nabla v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} = b(v), \quad \forall v \in H_0^1(\Omega)$$

with $f, \alpha \in W^{1,\infty}(\Omega) \hookrightarrow C(\bar{\Omega})$, $\alpha_1 \geq \alpha(x) \geq \alpha_0 > 0$.

- Let $V_h \subset V = H_0^1(\Omega)$ be constructed from triangular Lagrange elements of degree m on an affine equivalent triangulation $\mathcal{T}_h$.
- The discrete bilinear form $a_h : V_h \times V_h \to \mathbb{R}$ is

$$a_h(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \sum_{k=1}^{m} w_k \alpha(x_k) \nabla u_h(x_k) \cdot \nabla v_h(x_k)$$

where w_k and x_k are the Gauß quadrature weights and nodes on the element K.

- Recall that this formula is exact for polynomials of degree up to $2m - 1$, and all weights w_k are positive.

Generalized Galerkin Approach

- Estimating the quadrature on K for a_h,

$$\left(\sum_{k=1}^{m} w_k \alpha(x_k) \nabla u_h(x_k) \cdot \nabla v_h(x_k)\right)^2 \leq$$

$$\alpha_1^2 \left(\sum_{k=1}^{m} w_k |\nabla u_h(x_k)|^2\right) \left(\sum_{k=1}^{m} w_k |\nabla v_h(x_k)|^2\right) = \alpha_1^2 |u_h|^2_{H^1(K)} |v_h|^2_{H^1(K)}$$

  where the last equation follows since the components of ∇u_h and ∇v_h are in P_{m-1} and hence $|\nabla u_h|^2, |\nabla v_h|^2 \in P_{2m-2}$.
- Combining the above estimates shows that a_h is continuous on $V_h \times V_h$,

$$|a_h(u_h, v_h)| \leq c \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}$$

- Similarly, a_h is coercive,

$$a_h(u_h, u_h) \geq \alpha_0 \sum_{K \in \mathcal{T}_h} \sum_{k=1}^{m} w_k |\nabla u_h(x_k)|^2 = \alpha_0 |u_h|^2_{H^1(\Omega)} \geq c \|u_h\|^2_{H^1(\Omega)}$$

  where Poincaré's Lemma 153 is used in the last inequality.

Generalized Galerkin Approach

- To show that the linear form

$$b_h(v_h) = \sum_{K \in \mathcal{T}_h} \sum_{k=1}^{m} w_k f(x_k) v_h(x_k)$$

is bounded in terms of $\|v_h\|_{H^1(\Omega)}$, note first that,

$$b(v_h) - b_h(v_h) = \sum_{K \in \mathcal{T}_h} E_K(fv_h), \quad E_K(v) = \int_K v - \sum_{k=1}^{m} w_k v(x_k)$$

where E_K is a bounded, sublinear functional on $W^{1,\infty}(K)$ which vanishes $\forall v \in P_0 \subset P_{2m-1}$.

- So the Bramble Hilbert Lemma [209] can be applied on the reference element $\hat{K}$ to obtain

$$|E_{\hat{K}}(\hat{v})| \leq c|\hat{v}|_{W^{1,\infty}(\hat{K})}, \quad \forall \hat{v} \in W^{1,\infty}(\hat{K}).$$

- On the reference element $\hat{K}$ define $\hat{f} = f \circ T_K$ and $\hat{v}_h = v_h \circ T_K \in \hat{\mathcal{P}}$ to obtain

$$|E_{\hat{K}}(\hat{f}\hat{v}_h)| \leq c\|\hat{f}\|_{W^{1,\infty}(\hat{K})}\|\hat{v}_h\|_{W^{1,\infty}(\hat{K})} \leq c\|\hat{f}\|_{W^{1,\infty}(\hat{K})}\|\hat{v}_h\|_{L^2(\hat{K})}$$

since all norms are equivalent on finite dimensional $\hat{\mathcal{P}}$.

Generalized Galerkin Approach

- Using the integral transformation rule [211],

$$|E_K(fv)| = |E_{\hat{K}}(\hat{f}\hat{v})|\det(A_K) \leq c\|\hat{f}\|_{W^{1,\infty}(\hat{K})}\|\hat{v}_h\|_{L^2(\hat{K})}\det(A_K)$$

- Using Lemma [211],

$$\|\hat{f}\|_{W^{1,\infty}(\hat{K})} \leq c\|A_K\|\|f\|_{W^{1,\infty}(K)}$$
$$\|\hat{v}_h\|_{L^2(\hat{K})} \leq c\|v_h\|_{L^2(K)}\det(A_K)^{-\frac{1}{2}}$$

- Combining these estimates gives

$$\begin{aligned}|E_K(fv)| &\leq c\|A_K\|\det(A_K)^{\frac{1}{2}}\|f\|_{W^{1,\infty}(K)}\|v_h\|_{L^2(K)} \\ &\leq ch_K\det(A_K)^{\frac{1}{2}}\|f\|_{W^{1,\infty}(K)}\|v_h\|_{L^2(K)}\end{aligned}$$

- Summing over elements,

$$|b(v_h) - b_h(v_h)| \leq ch\|f\|_{W^{1,\infty}(\Omega)}\sum_{K\in\mathcal{T}_h}\mathrm{vol}(K)^{\frac{1}{2}}\|v_h\|_{L^2(K)} \leq$$

$$ch\|f\|_{W^{1,\infty}(\Omega)}\left[\sum_{K\in\mathcal{T}_h}\mathrm{vol}(K)\right]^{\frac{1}{2}}\left[\sum_{K\in\mathcal{T}_h}\|v_h\|^2_{L^2(K)}\right]^{\frac{1}{2}} = ch\|f\|_{W^{1,\infty}(\Omega)}\|v_h\|_{L^2(\Omega)}$$

Generalized Galerkin Approach

- Then boundedness of b_h is obtained as follows

$$|b_h(v_h)| \leq |b_h(v_h) - b(v_h)| + |b(v_h)| \leq$$

$$ch\|f\|_{W^{1,\infty}(\Omega)}\|v_h\|_{L^2(\Omega)} + \|f\|_{L^2(\Omega)}\|v_h\|_{L^2(\Omega)}$$

$$\leq (ch\|f\|_{W^{1,\infty}(\Omega)} + \|f\|_{L^2(\Omega)})\|v_h\|_{L^2(\Omega)}$$

$$\leq c\|f\|_{W^{1,\infty}(\Omega)}\|v_h\|_{H^1(\Omega)}$$

- Thus, the discrete problem is well-posed by Theorem 236.
- For error estimates, assume in the following that linear Lagrange elements are used.
- By the first Strang Lemma 237, the discretization error is bounded by the approximation error and the quadrature error.
- Theorem 215 gives for the second term in 237,

$$\inf_{w_h \in V_h} \|u - w_h\|_{H^1(\Omega)} \leq ch|u|_{H^2(\Omega)}$$

Generalized Galerkin Approach

- For the quadrature error in the bilinear form, note that for $w_h, v_h \in V_h$, the gradients ∇w_h and ∇v_h are constant on each element K and hence

$$a(w_h, v_h) - a_h(w_h, v_h) =$$

$$\sum_{K \in \mathcal{T}_h} \left(\int_K \alpha \nabla w_h \cdot \nabla v_h - \sum_{k=1}^{m} w_k \alpha(x_k) \nabla w_h(x_k) \cdot \nabla v_h(x_k) \right)$$

$$= \sum_{K \in \mathcal{T}_h} \nabla w_h \cdot \nabla v_h \left(\int_K \alpha - \sum_{k=1}^{m} w_k \alpha(x_k) \right)$$

- Recall the earlier estimate based upon the Bramble Hilbert Lemma 209 and Theorem 213, i.e., take $fv_h = \alpha \cdot 1$ and

$$E_K(\alpha) = \int_K \alpha - \sum_{k=1}^{m} w_k \alpha(x_k)$$

to obtain

$$E_K(\alpha) \leq c h_K \det(A_K)^{\frac{1}{2}} \|\alpha\|_{W^{1,\infty}(K)} \|1\|_{L^2(K)} \leq c h_K \mathrm{vol}(K) \|\alpha\|_{W^{1,\infty}(K)}.$$

Generalized Galerkin Approach

- Using this estimate in the previous calculation gives

$$|a(w_h, v_h) - a_h(w_h, v_h)| \leq \sum_{K \in \mathcal{T}_h} |\nabla w_h \cdot \nabla v_h| |E_K(\alpha)|$$
$$\leq c \sum_{K \in \mathcal{T}_h} h_K |\alpha|_{W^{1,\infty}(K)} |\nabla w_h \cdot \nabla v_h| \mathrm{vol}(K)$$
$$\leq ch |\alpha|_{W^{1,\infty}(K)} \|w_h\|_{H^1(\Omega)} \|v_h\|_{H^1(\Omega)}.$$

  where $\int_K |\nabla w_h \cdot \nabla w_h| = |\nabla w_h \cdot \nabla w_h| \mathrm{vol}(K)$ since linear Lagrange elements are used.
- It was shown earlier that

$$|b(v_h) - b_h(v_h)| \leq ch \|f\|_{W^{1,\infty}(\Omega)} \|v_h\|_{H^1(\Omega)}$$

- Combining these estimates with the first Strang Lemma 237 yields

$$\|u - u_h\|_{H^1(\Omega)} \leq ch(\|f\|_{W^{1,\infty}(\Omega)} + |u|_{H^2(\Omega)})$$

  where $\inf_{w_h \in V_h} |\alpha|_{W^{1,\infty}(\Omega)} \|w_h\|_{V_h} = 0$ has been applied to the third term in 237.

Discontinuous Galerkin Approach

- Discontinuous Galerkin methods are based upon nonconforming finite element spaces with piecewise polynomials not necessarily continuous across elements.
- Such an approach is flexible (different polynomials on adjacent elements need not to match up) and it is natural for first order equations with discontinuities.
- For $\Omega \subset \mathbb{R}^n$ (polygonal), $\beta \in W^{1,\infty}(\Omega)^n$ and $\mu \in L^\infty(\Omega)$, consider the (steady-state, i.e., $u_t = 0$ on the left) convection-reaction equation

$$\beta \cdot \nabla u + \mu u = f, \quad \Omega \subset \mathbb{R}^n$$

with $u = 0$ on the inflow boundary

$$\Omega^- = \{s \in \partial\Omega : \beta(s) \cdot n(s) < 0\}$$

which is well separated from the outflow boundary

$$\Omega^+ = \{s \in \partial\Omega : \beta(s) \cdot n(s) > 0\}$$

in the sense that $\min_{s\in\Omega^-, t\in\Omega^+} |s - t| > 0$.

Discontinuous Galerkin Approach

- For existence note that the *graph space*

$$W = \{v \in L^2(\Omega) : \beta \cdot \nabla w \in L^2(\Omega)\}$$

is a Hilbert space equipped with the scalar product,

$$(v, w)_W = (v, w)_{L^2(\Omega)} + (\beta \cdot \nabla v, \beta \cdot \nabla w)_{L^2(\Omega)^n}$$

and W functions have traces in

$$L^2_\beta(\partial\Omega) = \{v \text{ meas on } \partial\Omega : \int_{\partial\Omega} |\beta \cdot n|^2 v < \infty\}$$

giving the integration by parts formula

$$\int_\Omega (\beta\cdot\nabla v)w+(\beta\cdot\nabla w)v+(\nabla\cdot\beta)vw = \int_{\partial\Omega} \beta\cdot nvw, \quad \forall v, w \in W$$

- The BCs are satisfied in

$$U = \{v \in W : v|_{\partial\Omega^-} = 0\}$$

- The weak formulation of the convection-reaction equation is to seek $u \in U$ satisfying

$$a(u, v) = (\beta\cdot\nabla u, v)_{L^2(\Omega)}+(\mu u, v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)}, \quad \forall v \in W \tag{45}$$

Discontinuous Galerkin Approach

Theorem: If $\mu(x) - \frac{1}{2}\|\nabla \cdot \beta(x)\|^2 \geq \mu_0 > 0$, a.e. $x \in \Omega$, then $\exists! u \in U$ satisfying (45). Also, $\exists c > 0$, $c \neq c(u, f)$, such that $\|u\|_W \leq c\|f\|_{L^2(\Omega)}$.

Proof: is an **Exercise** in the application of the Banach-Nečas-Babuška Theorem 233, but to see the significance of μ_0, note that for the inf-sup condition, ones uses

$$\int_\Omega u^2 \nabla \cdot \beta + 2u(\nabla u \cdot \beta) = \int_\Omega \nabla \cdot (\beta u^2) = \int_{\partial\Omega} u^2 \beta \cdot n$$

to obtain

$$a(u,u) = \int_\Omega (\beta \cdot \nabla u)u + \mu u^2 = \int_\Omega (\mu - \tfrac{1}{2}\nabla \cdot \beta)u^2 + \int_{\partial\Omega} \tfrac{1}{2}(\beta \cdot n)u^2 \geq \mu_0 \|u\|^2_{L^2(\Omega)}$$

where the inequality depends the assumption with μ_0 and upon $u|_{\partial\Omega^-} = 0$ and $\beta \cdot n > 0$ on $\partial\Omega^+$. It follows

$$\|u\|_{L^2(\Omega)} \leq \mu_0^{-1} \frac{a(u,u)}{\|u\|_{L^2(\Omega)}} \leq \mu_0^{-1} \sup_{v \in L^2(\Omega)} \frac{a(u,v)}{\|v\|_{L^2(\Omega)}}$$

Similarly,

Discontinuous Galerkin Approach

$$\|\beta\cdot\nabla u\|_{L^2(\Omega)} \le \sup_{v\in L^2(\Omega)} \frac{(\beta\cdot\nabla u, v)_{L^2(\Omega)}}{\|v\|_{L^2(\Omega)}} = \sup_{v\in L^2(\Omega)} \frac{a(u,v)-(\mu u, v)_{L^2(\Omega)}}{\|v\|_{L^2(\Omega)}} \le$$

$$\sup_{v\in L^2(\Omega)} \frac{a(u,v)}{\|v\|_{L^2(\Omega)}} + \|\mu\|_{L^\infty(\Omega)}\|u\|_{L^2(\Omega)} \le (1+\mu_0^{-1}\|\mu\|_{L^\infty(\Omega)}) \sup_{v\in L^2(\Omega)} \frac{a(u,v)}{\|v\|_{L^2(\Omega)}}$$

Taking the inf over $u \in U$ of the sum gives the inf-sup condition.

Continuity of a over $U \times L^2(\Omega)$ is established by direct estimation. For injectivity, one assumes that $v \in L^2(\Omega)$ satisfies $a(u,v) = 0$, $\forall u \in U$, and argues v must be sufficiently regular to obtain $\mu v = \nabla\cdot(\beta v)$ and $v \in W$. Then with integration by parts and existence of traces for v, direct calculations give $\int_{\partial\Omega}(\beta\cdot n)uv = a(u,v) - (u, \mu v - \nabla\cdot(\beta v))_{L^2(\Omega)} = 0$ and $\mu_0\|v\|_{L^2(\Omega)} \le (v, \mu v - \nabla\cdot(\beta v))_{L^2(\Omega)} = 0$. ■

- For the discontinuous Galerkin approach let $k \in \mathbb{N}_0$, suppose $\mathcal{T}_h$ is a triangulation of Ω and set
$$V_h = \{v \in L^2(\Omega) : v|_K \in P_k, K \in \mathcal{T}_h\}$$
where no continuity is required across faces.

Discontinuous Galerkin Approach

- The discrete counterpart to (45) is to seek $u_h \in V_h$ satisfying

$$a_h(u_h, v_h) = (f, v_h)_{L^2(\Omega)}, \quad \forall v_h \in V_h \tag{46}$$

where

$$a_h(u_h, v_h) = (\mu u_h + \beta \cdot \nabla_h u_h, v_h)_{L^2(\Omega)} - \int_{\partial\Omega^-} (\beta \cdot n) u_h v_h - \sum_{F \in \Gamma_h} \int_F \beta \cdot [\![u_h]\!] \{\!\{v_h\}\!\}$$

and

$$\{\!\{v\}\!\}_F = \tfrac{1}{2}(v|_{K_1} + v|_{K_2}), \quad [\![v]\!]_F = n v|_{K_1} + n v|_{K_2}, \quad \nabla_h v_h|_K = \nabla(v_h|_K).$$

- To see that a_h is coercive with respect to the norm on V_h

$$|\!|\!| u_h |\!|\!|^2 = \mu_0 \|u_h\|^2_{L^2(\Omega)} + \tfrac{1}{2} \int_{\partial\Omega} \beta \cdot n u_h^2$$

integrate by parts on each element of $a_h(u_h, u_h)$, $u_h \in V_h$, to obtain

$$(\mu u_h + \beta \cdot \nabla_h u_h, u_h)_{L^2(\Omega)} = \sum_{K \in \mathcal{T}_h} \int_K \mu u_h^2 + (\beta \cdot \nabla u_h) u_h$$

$$= \sum_{K \in \mathcal{T}_h} \int_K \mu u_h^2 - \tfrac{1}{2}(\nabla \cdot \beta) u_h^2 + \sum_{K \in \mathcal{T}_h} \int_{\partial K} \tfrac{1}{2}(n \cdot \beta) u_h^2$$

Discontinuous Galerkin Approach

- Since $\beta \in W^{1,\infty}(\Omega)^2$ is continuous,

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}\tfrac{1}{2}(\beta\cdot n)u_h^2 = \sum_{F\in\Gamma_h}\int_F \tfrac{1}{2}\beta\cdot[\![u_h^2]\!] + \sum_{F\in\partial\mathcal{T}_h\setminus\Gamma_h}\int_F \tfrac{1}{2}(\beta\cdot n)u_h^2$$

- Since $n_1 = -n_2 = n$ between elements K_1 and K_2,

$$\tfrac{1}{2}[\![w^2]\!] = \tfrac{1}{2}(w_{K_1}^2 - w_{K_1}^2)n = \tfrac{1}{2}(w_{K_1}+w_{K_1})(w_{K_1}-w_{K_1})n = \{\!\{w\}\!\}_F[\![w]\!]_F$$

- Combining terms over $\partial\Omega$ gives

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}\tfrac{1}{2}(\beta\cdot n)u_h^2 - \int_{\partial\Omega^-}(\beta\cdot n)u_h^2 = \sum_{F\in\Gamma_h}\int_F \beta\cdot[\![w]\!]_F\{\!\{w\}\!\}_F + \int_{\partial\Omega}\tfrac{1}{2}|\beta\cdot n|u_h^2$$

- Thus a_h satisfies

$$\begin{aligned} a_h(u_h,u_h) &= \sum_{K\in\mathcal{T}_h}\int_K(\mu - \tfrac{1}{2}(\nabla\cdot\beta))u_h^2 + \int_{\partial\Omega}\tfrac{1}{2}|\beta\cdot n|u_h^2 \\ &\geq \mu_0\|u_h\|_{L^2(\Omega)}^2 + \int_{\partial\Omega}\tfrac{1}{2}|\beta\cdot n|u_h^2 = |\!|\!|u_h|\!|\!|^2 \end{aligned}$$

- Similarly, one can show (**Exercise**), $\forall u_h, v_h \in V_h$,

$$a_h(u_h,v_h) \leq |\!|\!|u_h|\!|\!|\cdot|\!|\!|v_h|\!|\!|, \quad (f,v_h)_{L^2(\Omega)} \leq \|f\|_{L^2}\cdot|\!|\!|v_h|\!|\!|$$

Mixed Finite Element Methods

- Thus by Theorem 236, $\exists! u_h \in V_h$ satisfying (46).
- For an error estimate, assume that the solution to (45) satisfies $u \in U_* = U \cap H^1(\Omega)$. Then traces $u|_F$ are well defined in $L^2(F)$ for $F \in \Gamma_h$ and $a_h(u, v_h)$, $v \in V_h$, is well defined.
- With the norm on $U(h) = U_* + V_h$,
$$|||w|||_*^2 = \|w\|_*^2 + \sum_{K \in \mathcal{T}_h} (\|\beta \cdot \nabla w\|_{L^2(K)}^2 + h_K^{-1} \|w\|_{L^2(\partial K)}^2)$$
boundedness of a_h follows:
$$a_h(u, v_h) \leq c |||u|||_* |||v_h|||, \quad \forall u \in U(h), \quad \forall v_h \in V_h.$$
- Also, the solution u to (45) satisfies
$$a_h(u, v_h) = (f, v_h)_{L^2(\Omega)}, \quad \forall v_h \in V_h$$
- Thus, the Second Strang Lemma 238 can be used to obtain:

Theorem: Assume the solution to (45) satisfies $u \in U(h) \cap H^{k+1}(\Omega)$. Then $\exists c > 0$, $c \neq c(h)$, such that the solution u_h to (46) satisfies
$$|||u - u_h|||_* \leq c h^k |u|_{H^{k+1}(\Omega)}.$$

Mixed Finite Element Methods

- Mixed FEM are natural to solve variational problems with constraints.
- Let V and M be reflexive Banach spaces.
- Let the bilinear form $a : V \times V \to \mathbb{R}$ be symmetric, coercive and bounded.
- The solution to $a(u, v) = \langle f, v \rangle_{V^*,V}$, $\forall v \in V$, is the unique minimizer for $J(u) = \frac{1}{2}a(u, u) - \langle f, u \rangle_{V^*,V}$.
- Under the constraint $b(u, \mu) = \langle g, \mu \rangle_{M^*,M}$, $\forall \mu \in M$, for the bilinear form $b : V \times M \to \mathbb{R}$, one introduces for $u \in V$, $\lambda \in M$,

$$L(u, \lambda) = \frac{1}{2}a(u, u) - \langle f, u \rangle_{V^*,V} - b(u, \lambda) - \langle g, \lambda \rangle_{M^*,M}$$

and seeks a saddle point,

$$L(u, \lambda) = \inf_{v \in V} \sup_{\mu \in M} L(v, \mu)$$

whose weak first order optimality condition is

$$\begin{cases} a(u, v) + b(v, \lambda) &= \langle f, v \rangle_{V^*,V}, & \forall v \in V \\ b(u, \mu) &= \langle g, \mu \rangle_{M^*,M}, & \forall \mu \in M \end{cases} \tag{47}$$

Mixed Finite Element Methods

- Existence of a solution is given as follows.

Theorem (Brezzi Splitting): Assume that

b. $a : V \times V \to \mathbb{R}$ satisfies the conditions of Theorem 233,

c. $b : V \times M \to \mathbb{R}$ satisfies the LBB condition, i.e., for $\beta > 0$,

$$\inf_{v \in V} \sup_{\mu \in M} \frac{b(v, \mu)}{\|v\|_V \|\mu\|_M} \geq \beta$$

Then $\exists!(u, \lambda) \in V \times M$ satisfying (47) and

$$\|u\|_V + \|\lambda\|_M \leq c(\|f\|_{V^*} + \|g\|_{M^*}).$$

- For a Galerkin approximation choose subspaces $V_h \subset V$ and $M_h \subset M$ and seek $u_h \in V_h$, $\mu_h \in M_h$ such that

$$\begin{cases} a(u_h, v_h) + b(v_h, \lambda_h) &= \langle f, v_h \rangle_{V^*, V}, & \forall v \in V_h \\ b(u_h, \mu_h) &= \langle g, \mu_h \rangle_{M^*, M}, & \forall \mu_h \in M_h \end{cases} \tag{48}$$

- Existence of a solution is given analogously as above.

Mixed Finite Element Methods

Theorem: Assume $\exists \alpha_h, \beta_h > 0$ such that with $K_h = \{v_h \in V_h : b(v_h, \mu_h) = 0, \forall \mu_h \in M_h\}$,

$$\inf_{u_h \in K_h} \sup_{v_h \in K_h} \frac{a(v,\mu)}{\|u\|_V \|v\|_V} \geq \alpha_h \qquad (\text{LBB}_h:) \inf_{\mu_h \in M_h} \sup_{v_h \in V_h} \frac{b(v,\mu)}{\|v\|_V \|\mu\|_M} \geq \beta_h$$

Then $\exists!(u_h, \lambda_h) \in V_h \times M_h$ satisfying (48) and

$$\|u_h\|_{V_h} + \|\lambda_h\|_{M_h} \leq c(\|f\|_{V^*} + \|g\|_{M^*}).$$

Theorem (Fortin Criterion): Assume the LBB condition holds. Then the LBB_h condition holds iff $\exists \Pi_h : V \to V_h$ such that

$$b(\Pi_h v, \mu_h) = b(v, \mu_h), \quad \forall \mu_h \in M_h$$

and $\exists \gamma_h > 0$ such that $\|\Pi_h v\|_V \leq \gamma_h \|v\|_V$, $\forall v \in V$.

Theorem: Let the conditions of the previous two theorems be satisfied. Let (u, λ) be the solution to (47) and let (u_h, λ_h) be the solution to (48). Then $\exists c > 0$, $c \neq c(u, \lambda)$ such that

$$\|u - u_h\|_V + \|\lambda - \lambda_h\|_M \leq c(\inf_{v_h \in V_h} \|u - v_h\|_V + \inf_{\mu \in M_h} \|\lambda - \mu_h\|_M).$$

Mixed FEM for Piecewise Constant Data

- Suppose measurements of $u^\star$ on $\Omega = (0,1)^2$ include the simultaneous effect of noise and a smooth modulation.
- Let $\Omega_{\rm d}$ be a compact subset of Ω where measurements $\tilde{u}$ and κ satisfy $\tilde{u}, \kappa \geq c > 0$ and $u^\star \approx \tilde{u}/\kappa$ in $\Omega_{\rm d}$.
- Let $\chi_{\rm d}$ be the characteristic function for $\Omega_{\rm d}$.
- Then $u^\star$ is estimated as a minimizer for

$$J_\epsilon(u) = \int_\Omega [|\kappa u - \tilde{u}|^2 + \epsilon |\nabla^2 u|^2]$$

- The necessary optimality condition for a minimizer $u^\epsilon \in H^2(\Omega)$ is

$$\epsilon a(u^\epsilon, v) + b(u^\epsilon, v) = d(v), \quad \forall v \in H^2(\Omega) \qquad (49)$$

where

$$a(u,v) = \int_\Omega \nabla^2 u : \nabla^2 v, \quad b(u,v) = \int_\Omega \kappa^2 uv, \quad d(v) = \int_\Omega \kappa \tilde{u} v$$

- (**Exercise**) $\exists! u^\epsilon \in H^2(\Omega)$ satisfing (49), and when $\tilde{u}/\kappa$ can be extended to $H^2(\Omega)$, u^ϵ converges weakly in $H^2(\Omega)$ to a unique limit $u^* \in H^2(\Omega)$ as $\epsilon \to 0$.

Mixed FEM for Piecewise Constant Data

- Yet, for typically discontinuous data, a direct FEM approach to approximating the solution u^ϵ to (49) fails (miserably!) to approximate u^* for a fixed h as $\epsilon \to 0$. A discontinuous Galerkin method is only conditionally accurate.
- To obtain a saddle point formulation of u^*, define

$$L^2_{\mathrm{d}}(\Omega) = \{v \in L^2(\Omega) : (1 - \chi_{\mathrm{d}})v = 0\}$$

$$H_0(\Delta^2) = \left\{ \begin{array}{c} w \in H^2(\Omega) : \Delta^2 w \in L^2(\Omega), \\ a(w, \phi) = (\Delta^2 w, \phi)_{L^2(\Omega)}, \forall \phi \in H^2(\Omega) \end{array} \right\}$$

- **Theorem**: $\exists!(\lambda^\epsilon, u^\epsilon) \in L^2_{\mathrm{d}}(\Omega) \times H_0(\Delta^2)$ satisfying

$$\left\{ \begin{array}{rcll} -\epsilon b(\mu, \lambda^\epsilon) + b(\mu, u^\epsilon) & = & d(\mu), & \forall \mu \in L^2_{\mathrm{d}}(\Omega) \\ b(\lambda^\epsilon, v) + a(u^\epsilon, v) & = & 0, & \forall v \in H^2(\Omega) \end{array} \right. \tag{50}$$

  where u^ϵ satisfies (49). (Formally, set $\mu = v$ in I, multiply II by ϵ, and sum I and II.)
- For the following, define $H^{-2}_{\mathrm{d}}(\Omega)$ as the subspace of $H^{-2}(\Omega)$ given by the completion of $L^2_{\mathrm{d}}(\Omega)$ with respect to the norm

$$\|\mu\|_{H^{-2}_{\mathrm{d}}(\Omega)} = \sup_{v \in H^2(\Omega)} \frac{\int_\Omega \chi_{\mathrm{d}} \mu v}{\|v\|_{H^2(\Omega)}}$$

Mixed FEM for Piecewise Constant Data

- **Theorem**: $(\lambda^\epsilon, u^\epsilon)$ satisfying (50) converges weakly in $H_{\mathrm{d}}^{-2}(\Omega) \times H^2(\Omega)$ as $\epsilon \to 0$ to a unique limit $(\lambda^*, u^*) \in H_{\mathrm{d}}^{-2}(\Omega) \times H^2(\Omega)$ satisfying

$$\begin{cases} b(\mu, u^*) &= d(\mu), \quad \forall \mu \in H_{\mathrm{d}}^{-2}(\Omega) \\ b(\lambda^*, v) + a(u^*, v) &= 0, \quad \forall v \in H^2(\Omega) \end{cases} \tag{51}$$

  where u^* is the weak limit in $H^2(\Omega)$ of u^ϵ satisfying (49).
- For a mixed FEM approximation, set $x_i = ih$, $h = 1/(N+1)$, $i = 0, \ldots, N+1$, and define the maximally smooth splines

$$S_h^{(k)}(0,1) = \{s \in C^{k-1}(0,1) : s|_{[x_{i-1}, x_i]} \in P_k, i = 1, \ldots, N+1\}$$

  with $S_h^{(k)}(\Omega)$ being tensor products of such splines.
- Assume the grid conforms to the data support so $\chi_{\mathrm{d}} \in S_h^{(0)}(\Omega)$.
- Let $S_{h,\mathrm{d}}^{(k)}(\Omega)$ be the subspace of $S_h^{(k)}(\Omega)$ supported only on Ω_{d}.
- Assume the data are approximated respectively with $\kappa_h, \tilde{u}_h \in S_{h,\mathrm{d}}^{(0)}(\Omega)$ satisfying

$$\|\kappa - \kappa_h\|_{L^\infty(\Omega)} \to 0, \quad \|\tilde{u} - \tilde{u}_h\|_{L^\infty(\Omega)} \to 0, \quad h \to 0.$$

Mixed FEM for Piecewise Constant Data

- Define the approximate forms

$$b_h(u,v) = \int_\Omega \kappa_h^2 uv, \quad d_h(v) = \int_\Omega \kappa_h \tilde{u}_h v$$

- For a mixed FEM approximation to the solution to (50) seek $(\lambda_h^\epsilon, u_h^\epsilon) \in S_{h,\mathrm{d}}^{(0)}(\Omega) \times S_h^{(2)}(\Omega)$ satisfying

$$\begin{cases} -\epsilon b_h(\mu_h, \lambda_h^\epsilon) + b_h(\mu_h, u_h^\epsilon) &= d_h(\mu_h), \quad \forall \mu_h \in S_{h,\mathrm{d}}^{(0)}(\Omega) \\ b_h(\lambda_h^\epsilon, v_h) + a(u_h^\epsilon, v_h) &= 0, \quad \forall v_h \in S_h^{(2)}(\Omega) \end{cases} \tag{52}$$

- **Theorem**: $\exists!(\lambda_h^\epsilon, u_h^\epsilon) \in S_{h,\mathrm{d}}^{(0)}(\Omega) \times S_h^{(2)}(\Omega)$ satisfying (52), and as $\epsilon \to 0$, $(\lambda_h^\epsilon, u_h^\epsilon)$ converges to a unique limit $(\lambda_h^*, u_h^*) \in S_{h,\mathrm{d}}^{(0)}(\Omega) \times S_h^{(2)}(\Omega)$ satisfying

$$\begin{cases} b_h(\mu_h, u_h^*) &= d_h(\mu_h), \quad \forall \mu_h \in S_{h,\mathrm{d}}^{(0)}(\Omega) \\ b_h(\lambda_h^*, v_h) + a(u_h^*, v_h) &= 0, \quad \forall v_h \in S_h^{(2)}(\Omega) \end{cases} \tag{53}$$

- Under provable conditions of Theorems 256, it can be shown that $\|u^* - u_h^*\|_{H^2(\Omega)} + \|\lambda^* - \lambda_h^*\|_{H_{\mathrm{d}}^{-2}(\Omega)} \to 0$, $h \to 0$.

Trotter Kato Theorem for Evolution Equations

- Let $\{S(t)\}_{t\geq 0} \subset \mathcal{L}(H)$ be a $\mathcal{C}^0$ semigroup on a Hilbert space H with generator L.
- The goal is to construct approximations $L_h \approx L$ on spaces H_h which generate $\mathcal{C}^0$ semigroups $\{S_h(t)\}_{t\geq 0} \subset \mathcal{L}(H_h)$ satisfying $S_h(t) \approx S(t)$.
- Assume there are restriction operators $P_h \in \mathcal{L}(H, H_h)$ and expansion operators $E_h \in \mathcal{L}(H_h, H)$ satisfying
  i. $\|P_h\|_{H,H_h} \leq M_1$, $\|E_h\|_{H_h,H} \leq M_2$, where $M_1 \neq M_1(h)$, $M_2 \neq M_2(h)$.
  ii. $P_h E_h = I_h$ on H_h.

Theorem: Let $\{S(t)\}_{t\geq 0} \subset \mathcal{L}(H)$ be a $\mathcal{C}^0$ semigroup on a Hilbert space H. Then $\exists M \geq 1$ and $\omega \in \mathbb{R}$ such that $\|S(t)\|_H \leq Me^{\omega t}$, $\forall t \geq 0$.

Def: If a $\mathcal{C}^0$ semigroup $\{S(t)\}_{t\geq 0} \subset \mathcal{L}(H)$ satisfies $\|S(t)\|_H \leq Me^{\omega t}$, $\forall t \geq 0$, one writes $L \in G(M, \omega, H)$ for the generator L.

Stability, Consistency, Stability

Theorem (Trotter Kato): Let Hilbert spaces H and H_h be given and assume the restriction operators $P_h \in \mathcal{L}(H, H_h)$ and expansion operators $E_h \in \mathcal{L}(H_h, H)$ satisfy the assumptions above. Let $L \in G(M, \omega, H)$ and $L_h \in G(M_h, \omega_h, H_h)$ be generators of $\mathcal{C}^0$ semigroups $\{S(t)\}_{t\geq 0} \subset \mathcal{L}(H)$ and $\{S_h(t)\}_{t\geq 0} \subset \mathcal{L}(H_h)$, respectively. Then (a) and (b) are equivalent to (c) in the following.

a. (stability) $\exists \tilde{M} \geq 1$, $\tilde{\omega} \in \mathbb{R}$, such that $M, M_h \leq \tilde{M}$ and $\omega, \omega_h \leq \tilde{\omega}$, $\forall h > 0$, where $\tilde{M} \neq \tilde{M}(h)$ and $\tilde{\omega} \neq \tilde{\omega}(h)$.

b. (consistency) $\exists \lambda_0 \in \rho(L) \cap \cap_{h>0} \rho(L_h)$ such that $\forall u \in H$,

$$\|E_h(\lambda_0 I_h - L_h)^{-1} P_h u - (\lambda_0 I - L)^{-1} u\|_H \to 0, \quad h \to 0.$$

c. (convergence) $\forall u \in H$, $\forall t \geq 0$,

$$\|E_h S_h(t) P_h u - S(t)u\|_H \to 0, \quad h \to 0$$

uniformly on bounded t intervals.

If (c) holds, then (b) holds $\forall \lambda$ with $\Re\lambda > \tilde{\omega}$.

Alternative Consistency Condition

- Unfortunately, the consistency condition can be very difficult to verify, so the following theorem gives alternative conditions.

Theorem: Let the assumptions of the previous theorem be satisfied. Then (a) and (b′) are equivalent to (c), where alternative conditions for consistency are given as

b′. (1) $\forall u \in H$, $\|E_h P_h u - u\|_H \to 0$, $h \to 0$,

(2) $\exists D \subset \mathrm{dom}(L)$ such that $\bar{D} = H$ and $\overline{(\lambda_0 I - L)D} = H$ for some $\lambda_0 > \tilde{\omega}$, and

(3) $\forall u \in D$, $\exists \{\bar{u}_h\}_{h>0}$ with $\bar{u}_h \in \mathrm{dom}(L_h)$ such that

$$E_h \bar{u}_h \to u \quad \text{and} \quad E_h L_h \bar{u}_h \to Lu, \quad h \to 0.$$

- Example: For the convection equation on $\Omega = (0, 1)$, $t \geq 0$,

$$u_t + u_x = 0, \quad x \in \Omega, t > 0, \quad u(t, 0) = 0, \quad u(0, x) = u_0(x)$$

choose $H = L^2(\Omega)$.

Application to the Convection Equation

- Equip the generator $Lu = -u_x$ with

$$\mathrm{dom}(L) = \{u \in H^1(\Omega) : u(0) = 0\}$$

- L is dissipative since

$$(Lu, u)_{L^2(\Omega)} = -\tfrac{1}{2}\int_0^1 (u^2)_x = -\tfrac{1}{2}u^2|_0^1 = -\tfrac{1}{2}u(1)^2 \leq 0, \quad \forall u \in \mathrm{dom}(L)$$

- The range condition is satisfied since $\forall f \in L^2(\Omega)$, $\forall \lambda > 0$,

$$(L - \lambda)u = f \quad \Leftrightarrow \quad u(x) = -\int_0^x e^{\lambda(y-x)} f(y)dy \in \mathrm{dom}(L)$$

- By the Lumer Philips Theorem [165], $\exists\{S(t)\}_{t\geq 0} \subset \mathcal{L}(H)$, a contraction semigroup, i.e., $L \in G(1, 0, H)$.
- Now consider the semi-discrete approximation,

$$u_i'(t) = [u_{i-1}(t) - u_i(t)]/h, \quad i = 1, \ldots, n, \quad u_0(t) = 0.$$

where $H_h = \mathbb{R}^n$ and $u_i(t) \approx u(x_i, t)$, $x_i = ih$, $h = 1/n$.

Application to the Convection Equation

- Define the discrete generator by

$$(L_h u)_1 = -u_1/h, \quad (L_h u)_i = [u_{i-1} - u_i]/h, \quad i = 2, \ldots, n, \quad u = \{u_i\}_{i=1}^n$$

  with $\mathrm{dom}(L_h) = \mathbb{R}^n$. ($u_0 = 0$ is not included in the state.)
- Define the restriction operator $P_h \in \mathcal{L}(H, H_h)$ by

$$(P_h u)_i = \frac{1}{h} \int_{x_{i-1}}^{x_i} u(x)dx, \quad 1 \le i \le n, \quad u \in H$$

- Define the expansion operator $E_h \in \mathcal{L}(H_h, H)$ by

$$E_h u = \sum_{i=1}^{n} u_i \chi_{(x_{i-1}, x_i]}, \quad u \in H_h$$

  where χ_S denotes the characteristic function for the set S.
- Define the scalar product and norm on $H_h = \mathbb{R}^n$ by

$$(u, v)_h = h \sum_{i=1}^{n} u_i v_i, \quad \|u\|_h = (u, u)_h^{\frac{1}{2}}, \quad u, v \in H_h$$

- P_h and E_h clearly satisfy $P_h E_h = I_h$ on H_h, so condition (ii) is satisfied.

Application to the Convection Equation

- To show condition (i), note that $\forall u \in H$,

$$\|P_h u\|_{H_h}^2 = h \sum_{i=1}^{n} \left| \frac{1}{h} \int_{x_{i-1}}^{x_i} u(x)dx \right|^2 \leq \frac{1}{h} \sum_{i=1}^{n} h \int_{x_{i-1}}^{x_i} u^2(x)dx = \|u\|_{L^2(\Omega)}^2$$

  and $\forall u \in H_h$, using $\chi_i(x) = \chi_{(x_{i-1},x_i]}(x)$ and $\chi_i \chi_j = \delta_{ij}$,

$$\|E_h u\|_{L^2(\Omega)}^2 = \int_0^1 \left| \sum_{i=1}^{n} u_i \chi_i(x) \right|^2 dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} u_i^2 dx = \|u\|_h^2$$

- L_h is dissipative since

$$\begin{aligned} (L_h u, u)_h &= h\{u_1(0 - u_1)/h + \textstyle\sum_{i=2}^{n} u_i[u_{i-1} - u_i]/h\} \\ &= -\textstyle\sum_{i=1}^{n} u_i^2 + \sum_{i=2}^{n} u_i u_{i-1} \\ \leq -\textstyle\sum_{i=1}^{n} u_i^2 + \frac{1}{2}\sum_{i=2}^{n} u_i^2 + \frac{1}{2}\sum_{i=2}^{n} u_{i-1}^2 &\leq 0, \quad \forall u \in \mathrm{dom}(L_h) \end{aligned}$$

- The range condition is satisfied since $\forall f \in H_h = \mathbb{R}^n$, $\forall \lambda > 0$, $(L_h - \lambda)$ is diagonally dominant and

$$(L_h - \lambda)u = f \quad \Leftrightarrow \quad u = (L_h - \lambda)^{-1} f \in \mathrm{dom}(L_h)$$

- By the Lumer Philips Theorem [165], $\exists \{S_h(t)\}_{t \geq 0} \subset \mathcal{L}(H_h)$, a contraction semigroup, i.e., $L_h \in G(1, 0, H_h)$.

Application to the Convection Equation

- Thus, the stability condition (a) has been established, and the consistency condition (b′) will now be established.
- For (b′ 1), let $u \in H$ and set $\tilde{u} \in \mathcal{C}^1(\bar{\Omega})$ to satisfy $\|u - \tilde{u}\|_{L^2(\Omega)} \leq \epsilon$, so that using (i),

$$\begin{aligned} \|E_h P_h u - u\|_H &\leq \|E_h P_h \tilde{u} - \tilde{u}\|_H + \|(E_h P_h - I)(u - \tilde{u})\|_H \\ &\leq \|E_h P_h \tilde{u} - \tilde{u}\|_H + 2\epsilon \end{aligned}$$

and using $\chi_i(x) = \chi_{(x_{i-1}, x_i]}$ and $\chi_i \chi_j = \delta_{ij}$,

$$\|E_h P_h \tilde{u} - \tilde{u}\|^2_{L^2(\Omega)} = \int_0^1 \Big| \sum_{i=1}^n [\tilde{u}(x_i) - \tilde{u}(x)] \chi_i(x) \Big|^2 dx$$

$$\leq \sum_{i=1}^n \int_{x_{i-1}}^{x_i} |\tilde{u}(x_i) - \tilde{u}(x)|^2 dx \leq \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \Big| \int_x^{x_i} \tilde{u}'(y) dy \Big|^2 dx$$

$$\leq \sum_{i=1}^n \int_{x_{i-1}}^{x_i} h dx \int_{x_{i-1}}^{x_i} |\tilde{u}'(y)|^2 dy = h^2 \|u'\|^2_{L^2(\Omega)} \to 0, \quad h \to 0$$

- For (b′ 2), let $D = \{u \in \mathcal{C}^2(\bar{\Omega}) : u(0) = 0\}$ and note first that $\bar{D} = H$. To show that $\overline{(\lambda - L)D} = H$ holds $\forall \lambda > \tilde{\omega} = 0$, let

Application to the Convection Equation

$f \in H = L^2(\Omega)$ be arbitrary. Let $\{f_k\}_{k\geq 1} \subset \mathcal{C}^1(\bar{\Omega})$ be chosen so that $\|f - f_k\|_H \to 0$, $k \to \infty$. Then

$$u_k(x) = -\int_0^x e^{\lambda(y-x)} f_k(y)dy$$

satisfies $u_k \in D$, $(\lambda - L)u_k = f_k$ and

$$\|(\lambda - L)u_k - f\|_H = \|f_k - f\|_H \to 0,\ k \to \infty.$$

- For (b′ 3), set $\bar{u} = \{u(x_i)\}_{i=1}^n$ for $u \in D$. To show that $\|E_h\bar{u} - u\|_H \to 0$, $h \to 0$,

$$\|E_h\bar{u} - u\|_H^2 = \int_0^1 \Big| \sum_{i=1}^n [u(x_i) - u(x)]\chi_i(x) \Big|^2 dx$$

$$= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} |u(x_i) - u(x)|^2 dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \Big| \int_x^{x_i} |u'(y)dy \Big|^2 dx$$

$$\leq \sum_{i=1}^n \int_{x_{i-1}}^{x_i} (x_i - x) \int_x^{x_i} |u'(y)|^2 dydx \leq \sum_{i=1}^n \int_{x_{i-1}}^{x_i} (x_i - x)dx \int_{x_{i-1}}^{x_i} |u'(y)|^2 dy$$

$$\leq \sum_{i=1}^k \int_{x_{i-1}}^{x_i} |u'(y)|^2 \frac{h^2}{2} dy = \frac{h^2}{2}\|u'\|_{L^2}^2 \to 0, \quad h \to 0.$$

Application to the Convection Equation

and similarly to show that $\|E_h L_h \bar{u} - Lu\|_H \to 0$, $h \to 0$,

$$\|E_h L_h \bar{u} - Lu\|_H^2 = \int_0^1 \Big| \sum_{i=1}^n \Big[\frac{u(x_{i-1}) - u(x_i)}{h} + u'(x) \Big] \chi_i(x) \Big|^2 dx$$

$$= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \Big| -\frac{1}{h} \int_{x_{i-1}}^{x_i} u'(y) dy + u'(x) \Big|^2 dx$$

$$= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \Big| \frac{1}{h} \int_{x_{i-1}}^{x_i} [u'(x) - u'(y)] dy \Big|^2 dx$$

$$\le \frac{1}{h^2} \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \Big[h \int_{x_{i-1}}^{x_i} |u'(x) - u'(y)|^2 dy \Big] dx$$

$$\le \frac{1}{h} \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \Big[\int_{x_{i-1}}^{x_i} \Big| \int_y^x u''(z) dz \Big|^2 dy \Big] dx$$

$$\le \frac{1}{h} \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \Big[\int_{x_{i-1}}^{x_i} \Big[h \int_{x_{i-1}}^{x_i} |u''(z)|^2 dz \Big] dy \Big] dx = h^2 \|u''\|_{L^2(\Omega)}^2 \overset{h \to 0}{\longrightarrow} 0$$

- Hence, the conditions of stability and consistency are satisfied and the convergence property follows.

Application to the Heat Equation

- For the heat equation on $\Omega = (0,1)$, $t \geq 0$,

$$u_t = u_{xx}, \quad x \in \Omega, t > 0, \quad u(t,0) = u(t,1) = 0, \quad u(0,x) = u_0(x)$$

  choose $H = L^2(\Omega)$. **Exercise**: Repeat for Neumann BCs.
- Equip the generator $Lu = u_{xx}$ with

$$\mathrm{dom}(L) = H^2(\Omega) \cap H_0^1(\Omega)$$

- L is dissipative since

$$(Lu, u)_{L^2(\Omega)} = -(u_x, u_x)^2_{L^2(\Omega)} \leq 0, \quad \forall u \in \mathrm{dom}(L)$$

- The range condition is satisfied since $\forall f \in L^2(\Omega)$, $\forall \lambda > 0$, $\exists! u \in H_0^1(\Omega)$ such that with $b(v) = (f, v)_{L^2(\Omega)}$ and $a_\lambda(u,v) = (u_x, v_x)_{L^2(\Omega)} + \lambda(u,v)_{L^2(\Omega)}$,

$$a_\lambda(u,v) = b(v), \quad \forall v \in H_0^1(\Omega)$$

  and $f \in L^2(\Omega)$ implies by Theorem 160 that $u \in H^2(\Omega)$.
- By the Lumer Philips Theorem 165, $\exists \{S(t)\}_{t \geq 0} \subset \mathcal{L}(H)$, a contraction semigroup, i.e., $L \in G(1,0,H)$.

Application to the Heat Equation

- Now consider the semi-discrete approximation,

$$u_i'(t) = [u_{i+1}(t) - 2u_i(t) + u_{i-1}(t)]/h^2, \quad i = 1, \ldots, n$$
$$u_0(t) = u_{n+1}(t) = 0$$

  where $H_h = \mathbb{R}^n$ and $u_i(t) \approx u(x_i, t)$, $x_i = ih$, $h = 1/(n+1)$.
- With $u = \{u_i\}_{i=1}^n$ and understanding $u_0 = u_{n+1} = 0$ (not included in the state), define the discrete generator by

$$(L_h u)_i = [u_{i+1} - 2u_i + u_{i-1}]/h^2, \quad i = 1, \ldots, n$$

  with $\mathrm{dom}(L_h) = \mathbb{R}^n$.
- Define the restriction operator $P_h \in \mathcal{L}(H, H_h)$, the expansion operator $E_h \in \mathcal{L}(H_h, H)$ and the scalar product $(\cdot, \cdot)_h$ as before, and recall that conditions (i) and (ii) hold.
- L_h is dissipative since

$$h(L_h u, u)_h = \sum_{i=1}^{n-1} u_i u_{i+1} - 2\sum_{i=1}^{n} u_i^2 + \sum_{i=2}^{n} u_i u_{i-1}$$
$$\leq \sum_{i=1}^{n-1} \tfrac{1}{2}(u_i^2 + u_{i+1}^2) - 2\sum_{i=1}^{n} u_i^2 + \sum_{i=2}^{n} \tfrac{1}{2}(u_i^2 + u_{i-1}^2) \leq 0$$

Application to the Heat Equation

- The range condition is satisfied since $\forall f \in H_h = \mathbb{R}^n$, $\forall \lambda > 0$, $(L_h - \lambda)$ is SPD and

$$(L_h - \lambda)u = f \quad \Leftrightarrow \quad u = (L_h - \lambda)^{-1} f \in \mathrm{dom}(L_h)$$

- By the Lumer Philips Theorem 165, $\exists \{S_h(t)\}_{t \geq 0} \subset \mathcal{L}(H_h)$, a contraction semigroup, i.e., $L_h \in G(1, 0, H_h)$.
- Thus, the stability condition (a) has been established, and the consistency condition (b′) will now be established.
- Condition (b′ 1) holds as before.
- For (b′ 2), let $D = \mathcal{C}^3(\bar{\Omega}) \cap \mathcal{C}_0(\bar{\Omega}) \subset$ dom(L) and note first that $\bar{D} = H$. To show that $\overline{(\lambda - L)D} = H$ holds $\forall \lambda > \tilde{\omega} = 0$, let $f \in H = L^2(\Omega)$ be arbitrary and choose $\{f_k\}_{k \geq 1} \subset \mathcal{C}^1(\bar{\Omega})$ with $\|f - f_k\|_H \to 0$, $k \to \infty$. Then $\exists! u_k \in H_0^1(\Omega)$ such that

$$a_\lambda(u, v) = (f_k, v)_{L^2(\Omega)}, \quad \forall v \in H_0^1(\Omega).$$

Since $f_k \in L^2(\Omega)$, Theorem 160 implies that $u_k \in H^2(\Omega)$ and $(\lambda - L)u_k = f_k$ or $u_k'' = \lambda u_k - f_k \in \mathcal{C}^1(\bar{\Omega})$. Thus, $u_k \in D$ and

$$\|(\lambda - L)u_k - f\|_H = \|f_k - f\|_H \to 0, \; k \to \infty.$$

Application to the Heat Equation

- For (b′ 3), set $\bar{u} = \{u(x_i)\}_{i=1}^n$ for $u \in D$. Then $\|E_h\bar{u} - u\|_H \to 0$ holds for $h \to 0$ as before.

  To show that $\|E_h L_h \bar{u} - Lu\|_H \to 0$, $h \to 0$,

$$\|E_h L_h \bar{u} - Lu\|_H^2 =$$

$$\int_0^1 \Big| \sum_{i=1}^n \frac{[u(x_{i-1}) - u(x_i)] - [u(x_i) - u(x_{i-1})]}{h^2} \chi_{(x_{i-1},x_i]} - u''(x) \Big|^2 dx$$

$$= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \Big| \frac{1}{h^2} \int_{x_{i-1}}^{x_i} [u'(y+h) - u'(y)] dy - u''(x) \Big|^2 dx$$

$$= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \Big| \frac{1}{h^2} \int_{x_{i-1}}^{x_i} \Big[\int_y^{y+h} u''(t) dt \Big] dy - u''(x) \Big|^2 dx$$

$$= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \Big| \frac{1}{h^2} \int_{x_{i-1}}^{x_i} \Big[\int_y^{y+h} [u''(t) - u''(x)] dt \Big] dy \Big|^2 dx$$

Application to the Heat Equation

$$\leq \frac{1}{h^4} \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left| \int_{x_{i-1}}^{x_i} \left[h \int_{y}^{y+h} |u''(t) - u''(x)|^2 dt \right]^{\frac{1}{2}} dy \right|^2 dx$$

$$\leq \frac{1}{h^3} \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left[h \int_{x_{i-1}}^{x_i} \left[\int_{y}^{y+h} |u''(t) - u''(x)|^2 dt \right] dy \right] dx$$

$$\leq \frac{1}{h^2} \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left[\int_{x_{i-1}}^{x_i} dy \right] \left[\int_{x_{i-1}}^{x_{i+1}} |u''(t) - u''(x)|^2 dt \right] dx$$

$$= \frac{1}{h} \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left[\int_{x_{i-1}}^{x_{i+1}} \left| \int_{x}^{t} u'''(s) ds \right|^2 dt \right] dx$$

$$\leq \frac{1}{h} \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left[\int_{x_{i-1}}^{x_{i+1}} \left| 2h \int_{x_{i-1}}^{x_{i+1}} |u'''(s)|^2 ds \right] dt \right] dx$$

$$= 2h^2 \|u'''\|^2_{L^2(\Omega)} \overset{h \to 0}{\longrightarrow} 0$$

- Hence, the conditions of stability and consistency are satisfied by the semi-discrete scheme, and the convergence property follows.

Application to the Wave Equation

- For the wave equation on $\Omega = (0,1)$, $t \geq 0$,

$$u_{tt} = u_{xx}, \quad x \in \Omega, t > 0, \quad u(t,0) = u(t,1) = 0$$
$$u(0,x) = u_0(x), \quad u_t(0,x) = u_1(x)$$

written in first-order form,

$$U_t = LU, \quad U = \begin{pmatrix} u \\ u_t \end{pmatrix}, \quad L = \begin{pmatrix} 0 & I \\ \partial_{xx} & 0 \end{pmatrix}$$

choose $H = H_0^1(\Omega) \times L^2(\Omega)$ with

$$(U,V)_H = (\partial_x u_1, \partial_x v_1)_{L^2(\Omega)} + (u_2, v_2)_{L^2(\Omega)}$$
$$U = (u_1, u_2), \quad V = (v_1, v_2)$$

i.e., $H_0^1(\Omega)$ is equipped with the norm $\|u\|_{H_0^1(\Omega)} = |u|_{H^1(\Omega)}$.

- Equip the generator L with

$$\mathrm{dom}(L) = [H^2(\Omega) \cap H_0^1(\Omega)] \times H_0^1(\Omega)$$

- L is dissipative since $\forall U \in \mathrm{dom}(L)$,

$$(LU,U)_H = (\partial_x u_2, \partial_x u_1)_{L^2(\Omega)} + (\partial_{xx} u_1, u_2)_{L^2(\Omega)} = \partial_x u_1 u_2 |_{x=0}^{x=1} = 0.$$

Application to the Wave Equation

- The range condition is satisfied since $\forall \lambda \geq 0$,
  $\forall (f_1, f_2) = F \in H = H_0^1(\Omega) \times L^2(\Omega)$,
  $\exists! (u_1, u_2) = U \in H_0^1(\Omega)^2$ such that with $b_f(v) = (f, v)_{L^2(\Omega)}$
  and $a_\mu(u, v) = (u_x, v_x)_{L^2(\Omega)} + \mu(u, v)_{L^2(\Omega)}$,

  $$a_{\lambda^2}(u_1, v) = b_{f_2 + \lambda f_1}(v), \quad \forall v \in H_0^1(\Omega), \quad u_2 = \lambda u_1 - f_1 \in H_0^1(\Omega).$$

  Since $f_2 + \lambda f_1 \in L^2(\Omega)$, Theorem 160 implies $u_1 \in H^2(\Omega)$.
  Hence, $U \in \mathrm{dom}(L)$.
- By the Lumer Philips Theorem 165, $\exists \{S(t)\}_{t \geq 0} \subset \mathcal{L}(H)$, a contraction semigroup, i.e., $L \in G(1, 0, H)$.
- Now consider the semi-discrete approximation,

$$U = (u_1, u_2), \quad u_1(x, t) = \sum_{i=1}^{n} \alpha_i(t) \phi_i(x), \quad u_2(x, t) = \sum_{i=1}^{n} \beta_i(t) \chi_i(x)$$

where for $x_i = ih$, $h = 1/(n+1)$,

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h, & x \in [x_{i-1}, x_i] \\ (x_{i+1} - x)/h, & x \in [x_i, x_{i+1}] \\ 0, & \text{otherwise} \end{cases} \qquad \chi_i(x) = \begin{cases} 1/2, & x \in [x_{i-1}, x_{i+1}] \\ 0, & \text{otherwise} \end{cases}$$

Application to the Wave Equation

- Define the finite dimensional subspaces

$$H_h = \Phi_h \times X_h = \text{span}\{\phi_i\}_{i=1}^n \times \text{span}\{\chi_i\}_{i=1}^n$$

  equipped $\forall h > 0$ with the inner product on H.
- Define the restriction operator $P_h \in \mathcal{L}(H, H_h)$, $P_h = (P_h^{(1)}, P_h^{(2)})$, with the orthogonal projections,

$$(P_h^{(1)} v, \phi)_{H_0^1(\Omega)} = (v, \phi)_{H_0^1(\Omega)}, \quad \forall v \in H_0^1(\Omega), \quad \forall \phi \in \Phi_h$$

$$(P_h^{(2)} f, \chi)_{L^2(\Omega)} = (f, \chi)_{L^2(\Omega)}, \quad \forall f \in L^2(\Omega), \quad \forall \chi \in X_h$$

- Define the expansion operator $E_h \in \mathcal{L}(H_h, H)$ by $E_h = P_h^*$, i.e., the injection $H_h \to H$.
- (**Exercise**): P_h and E_h satisfy (i), (ii) and (b′ 1), and furthermore, $(P_h^{(1)} v)(x_i) = v(x_i)$, $i = 1, \ldots, n$.
- To satisfy $a(U, V) = (LU, V)_H$, $\forall U \in \text{dom}(L)$, $\forall V \in H$, define a on $H_0^1(\Omega) \times H_0^1(\Omega)$ for $U = (u_1, u_2)$, $V = (v_1, v_2)$, by

$$a(U, V) = (\partial_x u_2, \partial_x v_1)_{L^2(\Omega)} - (\partial_x u_1, \partial_x v_2)_{L^2(\Omega)}$$

Application to the Wave Equation

- Since $H_h \not\subset H_0^1(\Omega) \times H_0^1(\Omega)$, a bilinear form a_h cannot be defined by the restriction of a to H_h.
- However, $\Phi_h \times \Phi_h \subset H_0^1(\Omega) \times H_0^1(\Omega)$ so take a_h as the restriction of a to $\Phi_h \times \Phi_h$.
- (**Exercise**): The orthogonal projection of X_h onto Φ_h satisfies

$$P_h^{(2)}\phi = \sum_{i=1}^{n} \alpha_i \chi_i \in X_h \quad \text{for} \quad \phi = \sum_{i=1}^{n} \alpha_i \phi_i \in \Phi_h$$

  and hence the mapping $\iota_h(W) = (w_1, P_h^{(2)} w_2) \in H_h$ for $(w_1, w_2) = W \in \Phi_h \times \Phi_h$, is an isomorphism (invertible).
- Define a_h on H_h by

$$a_h(U, V) = a_h(\iota_h^{-1} U, \iota_h^{-1} V) = a(\iota_h^{-1} U, \iota_h^{-1} V), \quad U, V \in H_h$$

  and the approximate generators by

$$(L_h U, V)_H = a_h(U, V), \quad U, V \in H_h$$

  with $\mathrm{dom}(L_h) = H_h$.

Application to the Wave Equation

- (**Exercise**): For $U \in H_h$,

$$U = \left(\sum_{i=1}^{n} \alpha_i \phi_i, \sum_{i=1}^{n} \beta_i \chi_i, \right), \quad L_h U = \left(\sum_{i=1}^{n} \gamma_i \phi_i, \sum_{i=1}^{n} \delta_i \chi_i, \right)$$

$\alpha = \{\alpha_i\}_{i=1}^n$, $\beta = \{\beta_i\}_{i=1}^n$, $\gamma = \{\gamma_i\}_{i=1}^n$, $\delta = \{\delta_i\}_{i=1}^n$ satisfy

$$A_h \alpha = -B_h \delta, \quad \gamma = \beta$$

where

$$A_h = \frac{1}{h} \text{ tridiag}\{-1, 2, -1\}, \quad B_h = \frac{h}{4} \text{ tridiag}\{1, 2, 1\}.$$

and the following matrix representation of L_h is invertible

$$\begin{pmatrix} 0 & I_h \\ -B_h^{-1} A_h & 0 \end{pmatrix}$$

- The range condition is satisfied since $\forall F \in H_h$, $\forall \lambda \geq 0$,

$$(L_h - \lambda) U = F \quad \Leftrightarrow \quad U = (L_h - \lambda)^{-1} F \in \text{dom}(L_h)$$

- L_h is dissipative since

$$(L_h U, U)_H = a_h(\iota_h^{-1} U, \iota_h^{-1} U) = a(\iota_h^{-1} U, \iota_h^{-1} U) = 0, \quad \forall U \in H_h.$$

Application to the Wave Equation

- By the Lumer Philips Theorem [165], $\exists\{S_h(t)\}_{t\geq 0} \subset \mathcal{L}(H_h)$, a contraction semigroup, i.e., $L_h \in G(1,0,H_h)$.
- Thus, the stability condition (a) has been established, and the consistency condition (b) will now be established.
- It has been shown that $0 \in \rho(L)$ and $0 \in \rho(L_h)$, $\forall h > 0$.
- For $(f,g) = F \in H$, set $(u,v) = U = L^{-1}F$ so $v = f$ and

$$-(\partial_x u, \partial_x \psi)_{L^2(\Omega)} = (\partial_x^2 u, \psi)_{L^2(\Omega)} = (g,\psi)_{L^2(\Omega)}, \quad \forall \psi \in H_0^1(\Omega)$$

- Set $(u_h, v_h) = U_h = L_h^{-1} P_h F$ so that $\forall (\phi_h, \psi_h) = \Psi_h \in H_h$,

$$(P_h F, \Psi_h)_H = (L_h U_h, \Psi_h)_H = a(\iota_h^{-1} U_h, \iota_h^{-1} \Psi_h)$$

- Define $\tilde{U}_h = (u_h, \tilde{v}_h)$ with $v_h = P_h^{(2)} \tilde{v}_h$ and $\tilde{\Psi}_h = (\phi_h, \tilde{\psi}_h)$ with $\psi_h = P_h^{(2)} \tilde{\psi}_h$ so that $\tilde{U}_h = \iota_h^{-1} U_h$ and $\tilde{\Psi}_h = \iota_h^{-1} \Psi_h$ and

$$\begin{aligned}(\partial_x P_h^{(1)} f, \partial_x \phi_h)_{L^2(\Omega)} + (P_h^{(2)} g, \psi_h)_{L^2(\Omega)} &= (P_h F, \Psi_h)_H \\ = a(\tilde{U}_h, \tilde{\Psi}_h) &= (\partial_x \tilde{v}_h, \partial_x \phi_h)_{L^2(\Omega)} - (\partial_x u_h, \partial_x \tilde{\psi}_h)_{L^2(\Omega)}\end{aligned}$$

Application to the Wave Equation

- In particular, with $\psi_h = 0$ and so $\tilde{\psi}_h = 0$,

$$(\partial_x P_h^{(1)} f, \partial_x \phi_h)_{L^2(\Omega)} = (\partial_x \tilde{v}_h, \partial_x \phi_h)_{L^2(\Omega)}$$

  and hence $\tilde{v}_h = P_h^{(1)} f$ and

$$v_h = P_h^{(2)} \tilde{v}_h = P_h^{(2)} P_h^{(1)} f \quad (= \textstyle\sum_{i=1}^n f(x_i)\chi_i).$$

- The remainder of $(P_h F, \Psi_h)_H = a(\tilde{U}_h, \tilde{\Psi}_h)$ gives

$$(P_h^{(2)} g, P_h^{(2)} \tilde{\psi}_h)_{L^2(\Omega)} = -(\partial_x u_h, \partial_x \tilde{\psi}_h)_{L^2(\Omega)}$$

- Set $\bar{u}_h = P_h^{(1)} u \in \Phi_h$ so that

$$-(\partial_x \bar{u}, \partial_x \psi_h)_{L^2(\Omega)} = -(\partial_x u, \partial_x \psi_h)_{L^2(\Omega)} = (g, \psi_h)_{L^2(\Omega)}, \quad \forall \psi \in \Phi_h$$

- Combining equations for u_h and $\bar{u}_h$

$$\begin{aligned}(\partial_x(\bar{u}_h - u_h), \partial_x \tilde{\psi}_h)_{L^2(\Omega)} &= (P_h^{(2)} g, P_h^{(2)} \tilde{\psi}_h)_{L^2(\Omega)} - (g, \tilde{\psi}_h)_{L^2(\Omega)} \\ &= (P_h^{(2)} g - g, P_h^{(2)} \tilde{\psi}_h)_{L^2(\Omega)} + (g, P_h^{(2)} \tilde{\psi}_h - \tilde{\psi}_h)_{L^2(\Omega)} \\ &= (g, P_h^{(2)} \tilde{\psi}_h - \tilde{\psi}_h)_{L^2(\Omega)}, \quad \forall \tilde{\psi}_h \in \Phi_h\end{aligned}$$

Application to the Wave Equation

- Taking the sup of both sides over $\tilde{\psi}_h \in \Phi_h^{(1)} = \{\phi_h \in \Phi_h : \|\phi_h\|_{H_0^1(\Omega)} \leq 1\}$ gives

$$\|\bar{u}_h - u_h\|_{H_0^1(\Omega)} \leq \|g\|_{L^2(\Omega)} \sup_{\tilde{\psi}_h \in \Phi_h^{(1)}} \|P_h^{(2)} \tilde{\psi}_h - \tilde{\psi}_h\|_{L^2(\Omega)}$$

- (**Exercise**) From an estimate $\|P_h^{(2)} v - v\|_{L^2(\Omega)} \leq h\|v\|_{H_0^1(\Omega)}$,

$$\|\bar{u}_h - u_h\|_{H_0^1(\Omega)} \to 0, \quad h \to 0.$$

- Since $u \in H^2(\Omega)$, it follows from an estimate for the interpolating splines $\|P_h^{(1)} v - v\|_{H_0^1(\Omega)} \leq h\|v\|_{H^2(\Omega)}$,

$$\|\bar{u}_h - u\|_{H_0^1(\Omega)} \to 0, \quad h \to 0.$$

- Thus, consistency and hence convergence follows from

$$\begin{aligned}
\|E_h L_h^{-1} P_h F - L^{-1} F\|_H^2 &= \|U_h - U\|_H^2 = \|(u_h, P_h^{(2)} P_h^{(1)} f) - (u, f)\|_H^2 \\
&= \|u_h - u\|_{H_0^1(\Omega)}^2 + \|P_h^{(2)} P_h^{(1)} f - f\|_{L^2(\Omega)}^2 \\
&\leq 2\|u - \bar{u}_h\|_{H_0^1(\Omega)} + 2\|u_h - \bar{u}_h\|_{H_0^1(\Omega)} \\
+2\|P_h^{(2)}[P_h^{(1)} f - f]\|_{L^2(\Omega)}^2 &+ 2\|P_h^{(2)} f - f\|_{L^2(\Omega)}^2 \to 0, \quad h \to 0.
\end{aligned}$$

Spectral Methods for Evolution Equations

- The approach here is to solve an evolution equation

$$u_t = Lu, \quad x \in \Omega, t > 0, \quad u(0, x) = u_0(x)$$

by expressing the approximate solution in the form

$$u_N(x, t) = \sum_{n=1}^{N} a_n(t)\phi_n(x)$$

where $\phi_n \in \text{dom}(L)$, $\forall n$, and these basis functions typically do not have compact support.

- The coefficients a_n are determined by Galerkin equations

$$D_t(\phi_n, u_N)_{L^2(\Omega)} = (\phi_n, Lu_n)_{L^2(\Omega)}$$

or
$$\sum_{m=1}^{N}(\phi_n, \phi_m)a'_m(t) = \sum_{m=1}^{N}(\phi_n, L\phi_m)a_m(t)$$

- For instance, if $\{\phi_n\}$ are chosen as orthonormal eigenfunctions of L, then the mass and stiffness matrices are diagonal and the error $u(x, t) - u_N(x, t)$ can converge to zero more rapidly than $e^{-N^2 t}$ as $N \to \infty$ for any $t > 0$.

Spectral Methods for Evolution Equations

- A method is said to provide *spectral accuracy* when the error converges to zero faster than any fixed power of N, restricted only by the smoothness of the exact solution.
- For instance, solving the heat equation on a simple domain by separation of variables is a spectral method using the eigenfunctions of the Laplacian as a basis for the approximation spaces.
- Also, basis functions may be constructed from Chebyshev polynomials T_n satisfying $T_n(\cos\theta) = \cos(n\theta)$ for degree n.
- Consider the evolution equation,

$$u_t + u_x = 0, \quad x \in [-1, 1], t > 0, \quad u(-1, t) = 0, \quad u(x, 0) = g(x)$$

- Let $\phi_n(x) = T_n(x) - (-1)^n T_0(x)$, and since $T_n(-1) = (-1)^n$, $\forall n$, the BCs are satisfied with $\phi_n(-1) = 0$.
- With $(f, g) = \int_{-1}^{+1} f(x)g(x)/\sqrt{1 - x^2}dx$, the matrix (ϕ_n, ϕ_m) is given by

$$(\phi_n, \phi_m) = \delta_{n,m}\pi/2 + (-1)^{n+m}\pi$$

since

Spectral Methods for Evolution Equations

$$(T_n, T_m) = \int_0^\pi \cos n\theta \cos m\theta d\theta = c_n \delta_{n,m} \pi/2, \quad c_n = 1 + \delta_{n,0}$$

- Using

$$2T_n(x) = \frac{T'_{n+1}(x)}{n+1} - \frac{T'_{n-1}(x)}{n-1}$$

the matrix $(\phi_n, L\phi_m)$ is given by

$$(\phi_n, \phi'_m) = \begin{cases} \pi m, & n \text{ odd and either } m \text{ odd or } m > n \\ -\pi m, & m \text{ odd, } n \text{ even and } n > m \\ 0, & \text{otherwise} \end{cases}$$

- Using these results gives the coefficients according to

$$a'_n(t) + 2(-1)^n \sum_{m=1}^{N} (-1)^m a'_m(t) = -2 \sum_{\text{odd } p+n=2n+1}^{N} p a_p(t)$$

$$+2(-1)^n \sum_{\text{odd } p=1}^{N} p a_p(t), \quad n = 1, \ldots, N$$

Non-Autonomous Evolution Equations

- The solution to the non-autonomous problem

$$u_t = L(t)u + f(t), \quad L(t) = \nabla \cdot [A(t)\nabla u] + r(t)u, \quad u(0) = u_0$$

with $u(t) \in H_0^1(\Omega)$, $t \geq 0$, is approximated using a Galerkin formulation applied to the weak formulation in (38).

- Define the bilinear form $a(t; \cdot, \cdot) : V \times V \to \mathbb{R}$ by

$$a(t; u, v) = (A(t)\nabla u, \nabla v)_{L^2(\Omega)} + (r(t)u, v)_{L^2(\Omega)}$$

and the linear form $b(t; \cdot) : V \to \mathbb{R}$ by

$$b(t; v) = (f(t), v)_{L^2(\Omega)}$$

- Suppose that spatial approximation spaces $V_h \subset V = H_0^1(\Omega)$ are chosen with a basis $\{\phi_i\}_{i=1}^{N_h}$, $N_h = \dim(V_h)$,
- The L^2-projected data $P_h f$, satisfying

$$(P_h f, \phi)_{L^2(\Omega)} = (f, \phi)_{L^2(\Omega)}, \quad \forall \phi \in V_h,$$

and the semi-discrete numerical solution u_h, satisying

$$(\partial_t u_h, \phi) + a(t; u_h, \phi) = b(t; \phi), \quad \forall \phi \in V_h,$$

take the form,

$$P_h f(x, t) = \sum_{i=1}^{N_h} F_i(t)\phi_i(x) \quad \text{and} \quad u_h(x, t) = \sum_{i=1}^{N_h} U_i(t)\phi_i(x).$$

Non-Autonomous Evolution Equations

- Define the mass matrix M and the time-dependent stiffness matrix K, respectively, by

$$M = (\phi_i, \phi_j)_{L^2(\Omega)}, \quad K(t) = \{a(t; \phi_i, \phi_j)\}$$

- Then the semi-discrete approximation

$$(\partial_t u_h, \phi) + a(t; u_h, \phi) = b(t; \phi), \quad \forall \phi \in V_h,$$

can be written as

$$MU'(t) + K(t)U(t) = MF(t)$$

where $U(t) = \{U_i(t)\}_{i=1}^{N_h}$ and $F(t) = \{F_i(t)\}_{i=1}^{N_h}$.

- This system of ODEs can then be solved by time-stepping schemes such as backward Euler

$$M[U^{m+1} - U^m]/\tau + K(t^{m+1})U^{m+1} = MF(t^{m+1})$$

for $m = 0, \ldots, \mu - 1$, $T = \mu\tau$, or Crank Nicholson

$$M[U^{m+1} - U^m]/\tau + [K(t^m)U^m + K(t^{m+1})U^{m+1}]/2 = M[F(t^m) + F(t^{m+1})]/2$$

as presented earlier.

Space-Time Galerkin Schemes

- On the other hand, a Galerkin approach can be formulated in space-time.
- Recall the spaces $X = \{v \in W^{1,2}(V, V^*) : v(0) = 0\}$ and $Y = L^2(0, T; V)$ and choose finite-dimensional subspaces $X_h \subset X$ and $Y_h \subset Y$.
- For instance, let $t^m = m\tau$, $1 \leq m \leq \mu$, $t^\mu = T$, and for each t^m, choose a (possibly different) $V_m \subset V$.
- Let $P_k(t^{m-1}, t^m; V_m)$ denote polynomials on $[t^{m-1}, t^m]$ with degree at most k with values in V_m.
- Define

$$X_h = \{v_h \in \mathcal{C}(0, T; V) : v_h|_{[t^{m-1}, t^m]} \in P_k(t^{m-1}, t^m; V_m)|_{1 \leq m \leq \mu}, v_h(0) = 0\}$$

$$Y_h = \{v_h \in L^2(0, T; V) : v_h|_{[t^{m-1}, t^m]} \in P_{k-1}(t^{m-1}, t^m; V_m)|_{1 \leq m \leq \mu}\}$$

- To approximate the solution u to (38), seek $u_h \in X_h$ such that $\forall v_h \in Y_h$,

$$\int_0^T \left[\langle \partial_t u_h(t), v_h(t)\rangle_{V^*,V} + a(t; u_h(t), v_h(t))\right] dt = \int_0^T b(t; v_h(t)) dt \tag{54}$$

Space-Time Galerkin Schemes

- Take $k = 1$, e.g., so u_h is piecewise linear in time:

$$u_h(t) = \frac{t^m - t}{t^m - t^{m-1}} u_h^{m-1} + \frac{t - t^{m-1}}{t^m - t^{m-1}} u_h^m, \quad t \in [t^{m-1}, t^m]$$

  where $u_h^m = u_h(t^m)$, and functions $v_h \in Y_h$ are constant,

$$v_h(t) = v_h(t^{m-1}) \ = v_h, \quad t \in [t^{m-1}, t^m]$$

- Inserting these into (54) gives for $m = 1, \ldots, \mu$

$$\forall v_h \in V_m, \quad \langle u_h^m - u_h^{m-1}, v_h) \rangle_{V^*,V} =$$

$$\int_{t^{m-1}}^{t^m} b(t; v_h) dt - \int_{t^{m-1}}^{t^m} a(t; u_h(t), v_h) dt$$

$$\approx b((t^{m-1} + t^m)/2; v_h)(t^m - t^{m-1})$$

$$-a((t^{m-1} + t^m)/2; u_h^{m-1} + u_h^m, v_h)(t^m - t^{m-1})/2$$

  where a midpoint rule is used partially for the last step.

- Discontinuous Galerkin methods can also be applied by taking $X_h = Y_h$, defined as above, but $v_h|_{(t^{m-1},t^m]} \in P_k(t^{m-1}, t^m; V_m)$, so functions are temporally continuous from the left, not necessarily agreeing with limits from the right.