Numerical integration of partial differential equations

1. | Finite difference schemes

Introduction

Definition 1. A linear system of n first order of PDEs for $\mathbf{u}(t,x)$ is a system of the form:

$$\mathbf{A}(t,x)\mathbf{u}_t + \mathbf{B}(t,x)\mathbf{u}_x = \mathbf{C}(t,x)\mathbf{u} + \mathbf{D}(t,x)$$

for certain matrices $\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D} \in \mathcal{M}_q(\mathbb{R})$. The system is called *hyperbolic* if $\mathbf{A}^{-1}\mathbf{B}$ is diagonalizable.

Definition 2. Let $n \in \mathbb{N}$, $m \in \mathbb{Z}$, $h,k > 0$ and $\mathbf{u} : \mathbb{R}^2 \to \mathbb{R}^q$ be a function. We define $\mathbf{u}_m^n := \mathbf{u}(t_n, x_m)$, where $(t_n, x_m) := (nk, x_0 + mh)$, $x_0 \in \mathbb{R}$. We denote by $\mathbf{v}_m^n$ an approximation to $\mathbf{u}_m^n$. The set of points $G := \{(t_n, x_m) : n \in \mathbb{N},\ m \in \mathbb{Z}\}$ is called a *grid*.

Definition 3. Let G be a grid. A *finite difference scheme* is a function

$$\begin{aligned} \mathbf{v} : \quad G \quad &\longrightarrow \quad \mathbb{R} \\ (t_n, x_m) &\longmapsto \mathbf{v}_m^n \end{aligned}$$

that aims to approximate $\mathbf{u}_m^n$, where $\mathbf{u} : \mathbb{R}^2 \to \mathbb{R}^q$ is a function. Here $\mathbf{v}_m^n$ is a function of $\mathbf{v}_m^{n-j}$, $m \in \mathbb{Z}$, $j = 0, \dots, J-1$. The number J is called the *number of steps*. If $J = 1$, we say that the scheme is a *one-step* scheme. Otherwise, we say that the scheme is *multistep*.

Proposition 4. Consider the one dimensional homogeneous traffic equation of constant coefficients

$$u_t + a u_x = f \tag{1}$$

where $a \in \mathbb{R}$ and f is a function. The following are satisfied:

1. *Forward-time forward-space* (*FTFS*):

$$\frac{u_m^{n+1} - u_m^n}{k} + a\frac{u_{m+1}^n - u_m^n}{h} + \mathrm{O}(k) + \mathrm{O}(h) = f_m^n$$

2. *Forward-time backward-space* (*FTBS*):

$$\frac{u_m^{n+1} - u_m^n}{k} + a\frac{u_m^n - u_{m-1}^n}{h} + \mathrm{O}(k) + \mathrm{O}(h) = f_m^n$$

3. *Forward-time central-space* (*FTCS*):

$$\frac{u_m^{n+1} - u_m^n}{k} + a\frac{u_{m+1}^n - u_{m-1}^n}{2h} + \mathrm{O}(k) + \mathrm{O}\left(h^2\right) = f_m^n$$

4. *Backward-time central-space* (*BTCS*):

$$\frac{u_m^{n+1} - u_m^n}{k} + a\frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2h} + \mathrm{O}(k) + \mathrm{O}\left(h^2\right) = f_m^{n+1}$$

5. *Leapfrog scheme*:

$$\frac{u_m^{n+1} - u_m^{n-1}}{2k} + a\frac{u_{m+1}^n - u_{m-1}^n}{2h} + \\ + \mathrm{O}\left(k^2\right) + \mathrm{O}\left(h^2\right) = f_m^n$$

6. *Lax-Friedrichs scheme*:

$$\frac{u_m^{n+1} - \frac{1}{2}(u_{m+1}^n + u_{m-1}^n)}{k} + a\frac{u_{m+1}^n - u_{m-1}^n}{2h} + \\ + \mathrm{O}(k) + \mathrm{O}\left(\frac{h^2}{k}\right) + \mathrm{O}\left(h^2\right) = f_m^n$$

Sketch of the proof. Use the Taylor expansion of $u(t,x)$. □

Corollary 5. Consider the traffic equation of Eq. (1) and let $\lambda := k/h$. Then, we have the following schemes for approximating the solution:

1. Forward-time forward-space

$$v_m^{n+1} = (1 + \lambda a)v_m^n - \lambda a v_{m+1}^n + k f_m^n$$

2. Forward-time backward-space

$$v_m^{n+1} = (1 - \lambda a)v_m^n + \lambda a v_{m-1}^n + k f_m^n$$

3. Forward-time central-space

$$v_m^{n+1} = v_m^n - \frac{\lambda a}{2}v_{m+1}^n + \frac{\lambda a}{2}v_{m-1}^n + k f_m^n$$

4. Backward-time central-space

$$v_m^{n+1} = v_m^n - \frac{\lambda a}{2}v_{m+1}^{n+1} + \frac{\lambda a}{2}v_{m-1}^{n+1} + k f_m^n$$

5. Leapfrog scheme:

$$v_m^{n+1} = v_m^{n-1} - \lambda a v_{m+1}^n + \lambda a v_{m-1}^n + k f_m^n$$

6. Lax-Friedrichs scheme:

$$v_m^{n+1} = \frac{1}{2}\left((1 - \lambda a)v_{m+1}^n + (1 + \lambda a)v_{m-1}^n\right) + k f_m^n$$

Convergence and consistency

Definition 6. A *stability region* is a set $\Lambda \subseteq {\mathbb{R}_{>0}}^2$ such that $(0,0) \in \Lambda'$, that is $(0,0)$ in an accumulation point.

Definition 7. Let (G_j) be a sequence of grids such that the time and space steps $k_j, h_j > 0$ of each one satisfy $\lim_{j\to\infty} k_j = \lim_{j\to\infty} h_j = 0$. We say that a finite difference scheme v approximating a PDE with initial condition $u_0(x)$ is *unconditionally convergent* if for any solution $u(x,t)$ to the PDE we have:

- For all $x \in \operatorname{dom} u_0$ and all increasing sequence $(m_j) \in \mathbb{N}$ such that $(\cdot, x_{m_j}) \in G_j$ and $\lim_{j\to\infty} x_{m_j} = x$, we have $\lim_{j\to\infty} v_{m_j}^0 = u_0(x)$.

- For all $(t,x) \in \operatorname{dom} u$ and all increasing sequences $(m_j),(n_j) \in \mathbb{N}$ such that $(t_{n_j}, x_{m_j}) \in G_j$ and $\lim_{j\to\infty} x_{m_j} = x$, $\lim_{j\to\infty} t_{n_j} = t$, we have $\lim_{j\to\infty} v^{n_j}_{m_j} = u(t,x)$.

The scheme is *conditionally convergent* if $\forall j \in \mathbb{N}$, $(k_j,h_j) \in \Lambda$, for some stability region Λ.

Definition 8. Let P be a partial differential operator and $\mathbf{f}$ be a function. Given the PDE $P\mathbf{u} = \mathbf{f}$ and a finite difference scheme $P_{k,h}\mathbf{v} = R_{k,h}\mathbf{f}$ with $R_{k,h}\mathbf{1} = \mathbf{1}$, we say that the scheme is *consistent* with the PDE if for any smooth function $\boldsymbol{\phi}(t,x)$ we have:

$$\lim_{k,h\to 0} R_{k,h}P\boldsymbol{\phi} - P_{k,h}\boldsymbol{\phi} = \mathbf{0}$$

where the convergence is pointwise at each point (t,x) in the domain of solutions. We say that the consistency is of order (p,q) in time and space if:

$$\lim_{k,h\to 0} R_{k,h}P\boldsymbol{\phi} - P_{k,h}\boldsymbol{\phi} = \mathrm{O}\left(k^p\right) + \mathrm{O}\left(h^q\right)$$

The consistency is a *conditional consistency* if the limit is for $(k,h) \in \Lambda$, for some stability region Λ. In that case, it makes sense to say that the consistency is of order r in $k = \lambda(h)$ if:

$$\lim_{h\to 0} R_{\lambda(h),h}P\boldsymbol{\phi} - P_{\lambda(h),h}\boldsymbol{\phi} = \mathrm{O}\left(h^r\right)$$

Lemma 9. The Lax-Friedrichs scheme is consistent if and only if $\lim_{h,k\to 0} \dfrac{h^2}{k} = 0$.

Remark. The consistency is not enough to guarantee convergence. For example, consider the PDE $u_t + au_x = 0$, with $a > 0$. The forward-time forward-space scheme is consistent with the PDE, but it is not convergent if we take the initial condition $u_0(x) = \mathbf{1}_{\{x<0\}}$ on the domain $[-1,1]$. Indeed, looking at Fig. 1 we see that from some instant of time, the solution will be 0 everywhere, which cannot be possible. In that case we should use the forward-time backward-space scheme, which is convergent. The usage of this latter method in these cases is called the *upwind condition*.

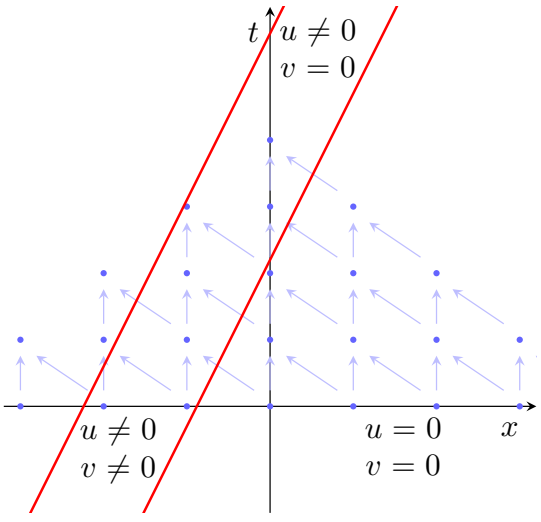

Figure 1: Infringement of the upwind condition. The arrows inward a bullet come from the points from which it depends.

Stability

Definition 10. Let $P_{k,h}\mathbf{v} = 0$ be a finite difference scheme with J steps, that is, a scheme in which we need the last J values of v^n to compute the next one, and Λ be a stability region. We say that it is *stable* if given $T > 0$, there exists $C_T > 0$ such that for any grid with $(k,h) \in \Lambda$ and for any initial values $\mathbf{v}^j_m$, $m \in \mathbb{Z}$, $j = 0,\dots,J-1$ we have

$$\sum_{m\in\mathbb{Z}} \|\mathbf{v}^n_m\|^2 \le C_T \sum_{j=0}^{J-1} \sum_{m\in\mathbb{Z}} \left\|\mathbf{v}^j_m\right\|^2$$

for all $n \in \mathbb{N}$ such that $0 \le nk \le T$.

Lemma 11. If a finite difference scheme of the form of

$$\mathbf{v}^{n+1}_m = \alpha\mathbf{v}^n_m + \beta\mathbf{v}^n_{m+1}$$

satisfies $|\alpha| + |\beta| \le 1$, then it is stable.

Sketch of the proof.

$$\begin{aligned}
\sum_{m\in\mathbb{Z}} \left\|\mathbf{v}^{n+1}_m\right\|^2 &\le \sum_{m\in\mathbb{Z}} \Big(|\alpha|^2 \|\mathbf{v}^n_m\|^2 + 2|\alpha||\beta| \|\mathbf{v}^n_m\| \cdot \\
&\qquad \cdot \left\|\mathbf{v}^n_{m+1}\right\| + |\beta|^2 \left\|\mathbf{v}^n_{m+1}\right\|^2\Big) \\
&\le \sum_{m\in\mathbb{Z}} \Big(|\alpha|^2 \|\mathbf{v}^n_m\|^2 + |\alpha||\beta|(\|\mathbf{v}^n_m\|^2 + \\
&\qquad + \left\|\mathbf{v}^n_{m+1}\right\|^2) + |\beta|^2 \left\|\mathbf{v}^n_{m+1}\right\|^2\Big) \\
&= \sum_{m\in\mathbb{Z}} \left(|\alpha|^2 + 2|\alpha||\beta| + |\beta|^2\right) \|\mathbf{v}^n_m\|^2 \\
&= (|\alpha| + |\beta|)^2 \sum_{m\in\mathbb{Z}} \|\mathbf{v}^n_m\|^2 \\
&\le (|\alpha| + |\beta|)^{2(n+1)} \sum_{m\in\mathbb{Z}} \left\|\mathbf{v}^0_m\right\|^2
\end{aligned}$$

□

Theorem 12 (Courant-Friedrichs-Lewy condition). Consider the traffic equation

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = 0$$

with $\mathbf{A} \in \mathcal{M}_q(\mathbb{R})$ and a finite difference scheme of the form

$$\mathbf{v}^{n+1}_m = \alpha\mathbf{v}^n_{m-1} + \beta\mathbf{v}^n_m + \gamma\mathbf{v}^n_{m+1}$$

with $k/h = \lambda = \text{const}$. Then, if the scheme is convergent, we have $|a_i\lambda| \le 1\ \forall a_i \in \sigma(\mathbf{A})$.

Proof. It suffices to study only the case $q = 1$. Suppose $|a\lambda| > 1$ for some eigenvalue a of $\mathbf{A}$ and let $\mathbf{u}_0(x) = \mathbf{c}\mathbf{1}_{\{|x|>\frac{1}{|\lambda|}\}}$ with $\mathbf{c} = (c_1,\dots,c_q)$ and $c_i \neq 0$. As shown in figure Fig. 2, by the form of the scheme, the numerical solution at $(t,x) = (1,0)$, v^n_0, will only depend on v^0_m with $|m| \le n$. But taking n such that $kn = 1$, we have that $|m|h \le nk/\lambda \le 1/\lambda$. So v^n_0 will depend on x for $|x| \le \frac{1}{\lambda} < |a|$. Thus, in general we will have the numerical solution equal to 0, whereas the exact solution will not be.

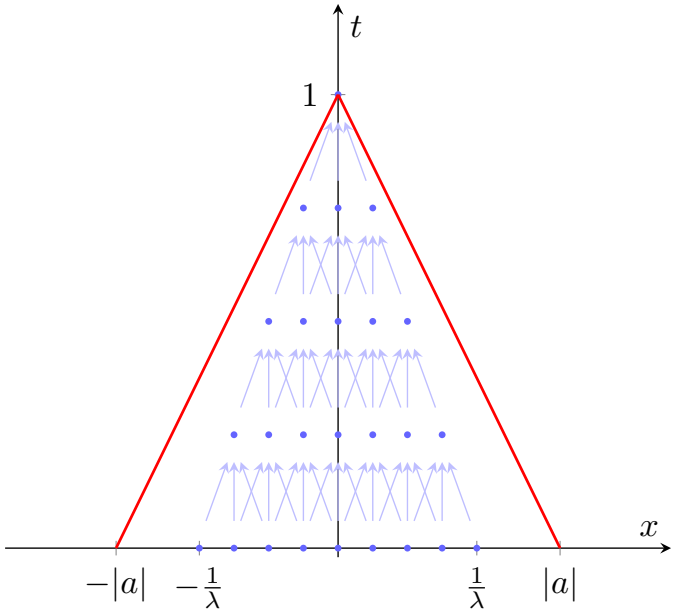

Figure 2: Finite difference scheme (blue) versus the characteristic lines (red). The arrows inward a bullet come from the points from which it depends.

□

Remark. The idea behind this is that one cannot obtain convergence of the scheme if the numerical domain does not include the analytic domain.

Semidiscrete Fourier transform

Definition 13 (Semidiscrete Fourier transform). The *semidiscrete Fourier transform* of a function $v \in \ell^2(h\mathbb{Z})$, i.e. defined in a mesh of step-size $h > 0$, is the function $\widehat{v} \in L^2\left(\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]\right)$ defined as the Fourier series:

$$\widehat{v}(\xi) = \sum_{m\in\mathbb{Z}} v_m \mathrm{e}^{-\mathrm{i}mh\xi}$$

where

$$v_m = \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} \widehat{v}(\xi)\mathrm{e}^{\mathrm{i}mh\xi}\,\mathrm{d}\xi$$

This latter formula is called *inverse semidiscrete Fourier transform*.

Proposition 14 (Semidiscrete Parseval identity). Let $h > 0$ and $\widehat{v} \in L^2\left(\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]\right)$ be the semidiscrete Fourier transform of $v \in \ell^2(h\mathbb{Z})$. Then

$$\sum_{m\in\mathbb{Z}} |v_m|^2 = \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} |\widehat{v}(\xi)|^2\,\mathrm{d}\xi$$

Proof.

$$\frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} |\widehat{v}(\xi)|^2\,\mathrm{d}\xi = \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} \sum_{m,n\in\mathbb{Z}} v_m \overline{v_n} \mathrm{e}^{-\mathrm{i}(m-n)h\xi}\,\mathrm{d}\xi$$

$$= \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} \sum_{m\in\mathbb{Z}} |v_m|^2\,\mathrm{d}\xi$$

$$= \sum_{m\in\mathbb{Z}} |v_m|^2$$

where in the second step we exchanged the integral and the sum in basis of the Cauchy-Schwarz inequality for sequences and the fact that $v \in \ell^2(h\mathbb{Z})$. □

Von Neumann stability analysis

Definition 15. Let $P_{k,h}v = f$ be a finite difference scheme. For each $n \in \mathbb{N}$, let $\widehat{v}^n \in L^2\left(\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]\right)$ be the function defined as the Fourier series:

$$\widehat{v}^n(\xi) = \sum_{m\in\mathbb{Z}} v_m^n \mathrm{e}^{-\mathrm{i}mh\xi}$$

Hence $v_m^n = \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} \widehat{v}^n(\xi)\mathrm{e}^{\mathrm{i}mh\xi}\,\mathrm{d}\xi$. We denote by $v^n := (v_m^n) \in \ell^2(\mathbb{Z})$ and $\|v^n\|_h{}^2 := h\|v^n\|_2{}^2$. We define the *amplification factor* as the 2π-periodic function in θ, $g(\theta, k, h)$ that satisfies:

$$\widehat{v}^{n+1}(\xi) = g(\xi h, k, h)\widehat{v}^n(\xi)$$

Theorem 16. Let $P_{k,h}v = f$ be a one-step finite difference scheme with constant coefficients whose amplification factor $g(\theta, k, h)$ is continuous on $\mathbb{R} \times \Lambda$, where $(k, h) \in \Lambda$ is a stability region. Then:

1. If $\exists K > 0$ such that $\forall\theta \in \mathbb{R}$ and $\forall(k, h) \in \Lambda$ we have $|g(\theta, k, h)| \le 1 + Kk$, then the scheme is stable in Λ.
2. If $\forall K > 0$ and $\forall\varepsilon > 0$ $\exists\theta \in \mathbb{R}$ and $\exists(k, h) \in \Lambda$ with $k < \varepsilon$ such that $|g(\theta, k, h)| > 1 + Kk$, then the scheme is unstable.

Proof.

1. We have that

$$\widehat{v}^n(\xi) = (g(\xi h, k, h))^n \widehat{v}^0(\xi)$$

Therefore applying twice the 14 Semidiscrete Parseval identity:

$$\sum_{m\in\mathbb{Z}} |v_m^n|^2 = \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} |\widehat{v}^n(\xi)|^2\,\mathrm{d}\xi$$

$$\le (1 + Kk)^{2n} \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} \left|\widehat{v}^0(\xi)\right|^2\,\mathrm{d}\xi$$

$$= (1 + Kk)^{2n} \sum_{m\in\mathbb{Z}} \left|v_m^0\right|^2$$

And note that $\forall T > 0$ with $nk \le T$ we have that:

$$(1 + Kk)^{2n} \le (1 + Kk)^{2\frac{T}{k}} = \left((1 + Kk)^{\frac{1}{Kk}}\right)^{2KT} \le$$

$$\le \mathrm{e}^{2KT} =: C_T$$

because $\sup_{x>0} (1 + x)^{1/x} = \mathrm{e}$.

2. Let $T \geq 2$, $K > 0$, and θ_0, h, k be the ones of the hypothesis with $\varepsilon = \min(1, \frac{1}{K})$. Hence, $k \leq 1$ and $Kk \leq 1$. By the continuity of g, $\exists \theta_1, \theta_2 \in \mathbb{R}$ such that $|g(\theta, k, h)| > 1 + Kk \ \forall \theta \in [\theta_1, \theta_2]$. Let $\widehat{v}^0(\xi) := \sqrt{\frac{h}{2\pi(\theta_2 - \theta_1)}} \mathbf{1}_{\left[\frac{\theta_1}{h}, \frac{\theta_2}{h}\right]}$ and denote $v^0 := (v_m^0) \in \ell^2(\mathbb{Z})$ its inverse transform. An easy check shows that $\left\|v^0\right\| = 1$. Now take $n := \lfloor T/k \rfloor$. Thus:

$$\begin{aligned}
\left\|v^n\right\|_2^2 &= \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} |\widehat{v}^n(\xi)|^2 \, \mathrm{d}\xi \\
&= \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} |g(h\xi, k, h)|^{2n} \left|\widehat{v}^0(\xi)\right|^2 \mathrm{d}\xi \\
&> (1 + Kk)^{2n} \\
&\geq (1 + Kk)^{\frac{2}{k}} \\
&= \left((1 + Kk)^{\frac{1}{Kk}}\right)^{2K} \\
&\geq 2^{2K} \\
&= 2^{2K} \left\|v^0\right\|_2^2
\end{aligned}$$

where in the forth inequality we used that $n \geq T/k - 1 = \frac{T-k}{k} \geq 1$ and in the penultimate step is because $\inf_{x \in [0,1]} (1 + x)^{1/x} = 2$. Hence, the scheme is unstable.

□

Corollary 17. Let $P_{k,h} v = f$ be a one-step finite difference scheme with constant coefficients whose amplification factor $g(\theta, k, h)$ is continuous on $\mathbb{R} \times \Lambda$, where $(k, h) \in \Lambda$ is a stability region. Then:

1. If $|g(\theta, k, h)| \leq 1 \ \forall \theta$ and $\forall (k, h) \in \Lambda$, then the scheme is stable.
2. If $\exists \theta_0 \in \mathbb{R}$ and $\delta > 0$ such that $|g(\theta_0, k, h)| > 1 + \delta$ $\forall (k, h) \in \Lambda$, then the scheme is unstable.

Lemma 18. Let $P_{k,h} v = f$ be a one-step finite difference scheme with constant coefficients. Impose that $v_m^n = g(\theta, k, h)^n \mathrm{e}^{\mathrm{i}m\theta}$ for certain function $g(\cdot, k, h)$. Then, g is the amplification factor of the scheme.

Proof. We have:

$$\begin{aligned}
\widehat{v}^{n+1}(\xi) &= \sum_{m \in \mathbb{Z}} v_m^{n+1} \mathrm{e}^{-\mathrm{i}mh\xi} \\
&= \sum_{m \in \mathbb{Z}} g(\theta, k, h)^{n+1} \mathrm{e}^{\mathrm{i}m\theta} \mathrm{e}^{-\mathrm{i}mh\xi} \\
&= g(\theta, k, h) \widehat{v}^n(\xi)
\end{aligned}$$

□

Proposition 19. Consider the PDE of Eq. (1) with $\lambda = k/h = \text{const}$. Then:

- The FTFS scheme is stable if and only if $a\lambda \in [-1, 0]$.
- The FTBS scheme is stable if and only if $a\lambda \in [0, 1]$.
- The FTCS scheme is always unstable.
- The BTCS scheme is unconditionally stable.
- The Lax-Friedrichs scheme is stable if and only if $|a\lambda| \leq 1$.

Proposition 20 (Lax-Wendroff). Consider the traffic equation of Eq. (1). The *Lax-Wendroff scheme* is:

$$\begin{aligned}
&\frac{u_m^{n+1} - u_m^n}{k} + a \frac{u_{m+1}^n - u_{m-1}^n}{2h} - \frac{a^2 k}{2} \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} \\
&\quad = \frac{f_m^{n+1} + f_m^n}{2} - \frac{ak}{4} \frac{f_{m+1}^n - f_{m-1}^n}{h} + \mathrm{O}\left(k^2\right) + \mathrm{O}\left(h^2\right)
\end{aligned}$$

Sketch of the proof. Expand $u(t + k, x)$ in Taylor series and use that:

$$\begin{aligned}
u_t &= -a u_x + f \\
u_{tt} &= a^2 u_{xx} - a f_x + f_t
\end{aligned}$$

□

Proposition 21. The Lax-Wendroff scheme is a one-step method that has order of consistency 2, and it is stable if and only if $|a\lambda| \leq 1$.

Sketch of the proof. Show that $P_{k,h}\phi - R_{k,h} P\phi = \mathrm{O}\left(h^2\right) + \mathrm{O}\left(k^2\right)$ using a Taylor expansion and is stable if $a\lambda \leq 1$. For the stability, assume $v_m^n = g^n \mathrm{e}^{\mathrm{i}m\theta}$. We need to study the homogeneous part.

$$\begin{aligned}
0 &= \frac{g - 1}{k} + \frac{a}{2h} \left(e^{\mathrm{i}\theta} - e^{-\mathrm{i}\theta}\right) - \frac{a^2 k}{2h^2} \left(e^{\mathrm{i}\theta} - 2 + e^{-\mathrm{i}\theta}\right) \\
g &= 1 - a\lambda \mathrm{i} \sin\theta + a^2\lambda^2 (\cos\theta - 1) \\
g &= 1 - 2a\lambda \mathrm{i} \sin\frac{\theta}{2} \cos\frac{\theta}{2} - 2a^2\lambda^2 \left(\sin\frac{\theta}{2}\right)^2
\end{aligned}$$

Hence:

$$\begin{aligned}
|g|^2 &= 1 - 4a^2\lambda^2 \left(\sin\frac{\theta}{2}\right)^2 + 4a^4\lambda^4 \left(\sin\frac{\theta}{2}\right)^4 + \\
&\qquad + 4a^2\lambda^2 \left(\sin\frac{\theta}{2} \cos\frac{\theta}{2}\right)^2 \\
&= 1 + 4a^2\lambda^2 (1 - a^2\lambda^2) \left(\sin\frac{\theta}{2}\right)^4
\end{aligned}$$

If $|a\lambda| \leq 1$, then $|g|^2 \leq 1$ because $x^2(1 - x^2) \leq 1/4$ for $x \in [-1, 1]$. If $|a\lambda| > 1$, then by taking $\theta = \pi$ we have $|g|^2 > 1$. □

Proposition 22 (Crank-Nicolson). Consider the traffic equation of Eq. (1). The *Crank-Nicolson scheme* is:

$$\begin{aligned}
&\frac{u_m^{n+1} - u_m^n}{k} + a \frac{u_{m+1}^{n+1} - u_{m-1}^{n+1} + u_{m+1}^n - u_{m-1}^n}{4h} = \\
&\qquad = \frac{f_m^{n+1} + f_m^n}{2} + \mathrm{O}\left(k^2\right) + \mathrm{O}\left(h^2\right)
\end{aligned}$$

Note that it is an implicit scheme.

Proposition 23. The Crank-Nicolson scheme is a one-step method that has order of consistency 2, and it is unconditionally stable.

Sketch of the proof. Let $P = \frac{\partial}{\partial t} + a\frac{\partial}{\partial x}$. Let's start with the consistency. Using $\phi = \phi(t,x) = v_m^n$ we can simplify the first term as:

$$\frac{\phi(t+k,x)-\phi}{k} = \phi_t + \frac{k}{2}\phi_{tt} + \mathrm{O}\left(k^2\right)$$

For the second term note that:

$$\begin{aligned}
\phi(t+k,x+h) &= \phi(t+k,x) + h\phi_x(t+k,x) + \\
&\quad + \frac{h^2}{2}\phi_{xx}(t+k,x) + \mathrm{O}\left(h^3\right) \\
-\phi(t+k,x-h) &= -\phi(t+k,x) + h\phi_x(t+k,x) - \\
&\quad - \frac{h^2}{2}\phi_{xx}(t+k,x) + \mathrm{O}\left(h^3\right) \\
\phi(t,x+h) &= \phi + h\phi_x + \frac{h^2}{2}\phi_{xx} + \mathrm{O}\left(h^3\right) \\
-\phi(t,x-h) &= -\phi + h\phi_x - \frac{h^2}{2}\phi_{xx} + \mathrm{O}\left(h^3\right)
\end{aligned}$$

Summing these equations and multiplying by $\frac{a}{4h}$ we get:

$$\frac{a}{2}[\phi_x + \phi_x(t+k,x)] + \mathrm{O}\left(h^2\right) = a\phi_x + \frac{a}{2}k\phi_{xt} + \mathrm{O}\left(h^2\right) + \mathrm{O}\left(k^2\right)$$

Thus:

$$P_{k,h}\phi = \phi_t + a\phi_x + \frac{k}{2}\phi_{tt} + \frac{a}{2}k\phi_{xt} + \mathrm{O}\left(k^2\right) + \mathrm{O}\left(h^2\right)$$

On the other hand:

$$\begin{aligned}
R_{k,h}P\phi &= \frac{\phi_t(t+k,x) + a\phi_x(t+k,x) + \phi_t + a\phi_x}{2} \\
&= \phi_t + a\phi_x + \frac{1}{2}k\phi_{tt} + \frac{a}{2}k\phi_{xt} + \mathrm{O}\left(k^2\right)
\end{aligned}$$

Finally:

$$P_{k,h}\phi - R_{k,h}P\phi = \mathrm{O}\left(k^2\right) + \mathrm{O}\left(h^2\right)$$

For the stability, substitute $v_m^n = g^n \mathrm{e}^{\mathrm{i}m\theta}$ in the scheme. Simplifying we get:

$$g = \frac{1 + \frac{a\lambda \mathrm{i}}{2}\sin\theta}{1 - \frac{a\lambda \mathrm{i}}{2}\sin\theta}$$

which has always modulus 1. □

Definition 24. Given scheme $P_{k,h}v = f$, usually we cannot use the recurrence to compute the last term of the (finite) grid, with $n \in \{0,\dots,N\}$ and $m \in \{0,\dots,M\}$, v_M^n for each $n \in \mathbb{N}$. Thus, the *numerical boundary condition* is used in these cases. A numerical boundary condition of order p is an extrapolation of order $\mathrm{O}\left(h^p\right)$ of the last term of the grid in terms of the orther ones. Each $u(t,x-\ell h)$ can be expressed as:

$$u(t,x-\ell h) = \sum_{k=0}^{p-1} \frac{(-1)^k \ell^k h^k}{k!} u^{(k)} + \mathrm{O}\left(h^p\right)$$

If we want to get a linear approximation of the form

$$u(t,x) = \sum_{k=1}^{p} \lambda_k u(x-kh)$$

we need to solve the following linear system:

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & (p-1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2^{p-1} & \cdots & (p-1)^{p-1} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_p \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Note that the solution always exists because the matrix is a Vandermonde matrix. For example, numerical boundary conditions of order 1, 2 and 3 are respectively:

$$\begin{aligned}
v_M^n &= v_{M-1}^n \\
v_M^n &= 2v_{M-1}^n - v_{M-2}^n \\
v_M^n &= 3v_{M-1}^n - 3v_{M-2}^n + v_{M-3}^n
\end{aligned}$$

Proposition 25. Consider the following initial value and boundary problem with constant coefficients:

$$\begin{cases} u_t = L(u) \\ u(0,\mathbf{x}) = u_0(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega \subseteq \mathbb{R}^d \\ u(t,\mathbf{x}) = g(t,\mathbf{x}) & \text{if } (t,\mathbf{x}) \in [0,\infty) \times \partial\Omega_1 \subseteq [0,\infty) \times \partial\Omega \end{cases} \tag{2}$$

where L is a differential operator and g is a function. Let $M(\Omega)$ be the set of indices that we compute $\mathbf{v}^n = (v_m^n)_{m \in M(\Omega)}$. Consider a finite difference scheme of the form

$$\mathbf{B}_1 \mathbf{v}^{n+1} = \mathbf{B}_0 \mathbf{v}^n + \mathbf{f}^n \tag{3}$$

where $\mathbf{B}_0$ and $\mathbf{B}_1$ are matrices and $\mathbf{f}^n$ is a vector. Then, the scheme is stable with stability region $\Lambda \subseteq \mathbb{R}_{\geq 0} \times {\mathbb{R}_{\geq 0}}^d$ if and only if $\forall T > 0\ \exists C_T > 0$ such that $\forall (k,h) \in \Lambda$ and $\forall \ell \in \mathbb{N}$ with $0 \leq \ell k \leq T$ we have $\left\| \left({\mathbf{B}_1}^{-1} \mathbf{B}_0 \right)^\ell \right\| \leq C_T$.

Proof. An easy check show that if $\mathbf{v}^0$ and $\mathbf{w}^0$ are such that satisfy the recurrence of Eq. (3), then $\mathbf{v}^\ell - \mathbf{w}^\ell = \mathbf{A}(\mathbf{v}^0 - \mathbf{w}^0)$, where $\mathbf{A} := \left({\mathbf{B}_1}^{-1} \mathbf{B}_0 \right)^\ell$.

$\Longrightarrow$) We will prove by contradiction. Suppose that $\exists T > 0$ such that $\forall C_T > 0$ exist $(k,h) \in \Lambda$ and $\ell \in \mathbb{N}$ with $0 \leq \ell k \leq T$ such that $\left\| \left({\mathbf{B}_1}^{-1} \mathbf{B}_0 \right)^\ell \right\| > C_T$. Then, taking $\mathbf{x}^*$ such that $\|\mathbf{x}^*\| = 1$ and $\|\mathbf{A}\| = \|\mathbf{A}\mathbf{x}^*\|$ we have that for any $\mathbf{v}^0$, defining $\mathbf{w}^0 := \mathbf{v}^0 + \mathbf{x}^*$ we have that:

$$\left\| \mathbf{v}^\ell - \mathbf{w}^\ell \right\| = \|\mathbf{A}\mathbf{x}^*\| = \|\mathbf{A}\| > C_T \left\| \mathbf{v}^0 - \mathbf{w}^0 \right\|$$

where the first equality follows from expanding recursively the norm $\left\| \mathbf{v}^\ell - \mathbf{w}^\ell \right\|$ and using the scheme Eq. (3). Hence, the scheme is not stable.

$\Longleftarrow$) Note that if

$$\left\| \mathbf{v}^\ell - \mathbf{w}^\ell \right\| \leq C_T \left\| \mathbf{v}^0 - \mathbf{w}^0 \right\|$$

then necessarily $\|\mathbf{A}\| \leq C_T$.

□

Theorem 26 (Lax-Richtmyer equivalence theorem). Consider the problem of Eq. (2) and define

$$\mathbf{T}^n := \mathbf{B}_1 \mathbf{u}^{n+1} - \mathbf{B}_0 \mathbf{u}^n - \mathbf{f}^n$$

Suppose that:

1. $\|\mathbf{T}^n\| = \mathrm{O}\left(k^p + \|\mathbf{h}\|^q\right)$ independent of n and $\forall (k,h) \in \Lambda$ and all $(t,x) \in [0,T] \times \Omega$ (consistency condition)
2. $\forall (k,h) \in \Lambda$, $\mathbf{B}_1$ is invertible and $\left\|{\mathbf{B}_1}^{-1}\right\| \leq C_1 k$ for certain $C_1 > 0$ independent of (k,h).
3. The scheme is stable.
4. $\mathbf{v}^0$ is such that $\left\|\mathbf{v}^0 - \mathbf{u}^0\right\| = \mathrm{O}\left(k^p + \|\mathbf{h}\|^q\right)$ uniformly for $(k,h) \in \Lambda$ and $(t,x) \in [0,T] \times \Omega$.

Then, $\forall n \in \mathbb{N}$ with $0 \leq nk \leq T$ we have:

$$\|\mathbf{v}^n - \mathbf{u}^n\| = \mathrm{O}\left(k^p + \|\mathbf{h}\|^q\right)$$

uniformly for $(k,h) \in \Lambda$ and $(t,x) \in [0,T] \times \Omega$.

Proof. We have that

$$\mathbf{B}_1 \mathbf{v}^n = \mathbf{B}_0 \mathbf{v}^{n-1} + \mathbf{f}^{n-1}$$
$$\mathbf{B}_1 \mathbf{u}^n = \mathbf{B}_0 \mathbf{u}^{n-1} + \mathbf{f}^{n-1} + \mathbf{T}^{n-1}$$

Then if $\mathbf{A} = {\mathbf{B}_1}^{-1}\mathbf{B}_0$ we have that

$$\mathbf{v}^n - \mathbf{u}^n = \mathbf{A}^n(\mathbf{v}^0 - \mathbf{u}^0) - \sum_{\ell=0}^{n-1} \mathbf{A}^{n-1-\ell} {\mathbf{B_1}}^{-1} \mathbf{T}^\ell$$

And so:

$$\|\mathbf{v}^n - \mathbf{u}^n\| \leq C_T \mathrm{O}\left(k^p + \|\mathbf{h}\|^q\right) + \sum_{\ell=0}^{n-1} C_T C_1 k \mathrm{O}\left(k^p + \|\mathbf{h}\|^q\right)$$

where we have used Theorem 25 for noting that $\forall \ell = 0, \ldots, n-1 \left\|\mathbf{A}^{n-1-\ell}\right\| \leq C_T$. Finally, observe that $kn \leq T$ and the uniformity of the $\mathrm{O}\left(k^p + \|\mathbf{h}\|^q\right)$ allows us to conclude the proof. □

Remark. It can also be shown that the consistency and convergence imply stability.

Theorem 27. Consider a scheme of J steps for a 1st-order-in-time linear PDE of constant coefficients whose amplification factor is g. Let $\Phi(\theta, g)$ be the *amplification polynomial*, that is the polynomial that satisfies g of degree $J-1$. Then, the scheme is stable if and only if:

- for any root $g_j(\theta)$ of Φ we have $|g_j(\theta)| \leq 1$-
- if $\exists \theta_0$ and k such that $|g_k(\theta_0)| = 1$, then this root is simple.

Proposition 28. The Leapfrog scheme for the one-dimensional wave equation of Eq. (1) is stable if and only if $|a\lambda| < 1$.

Proof. An easy check (substituting $v_n^m = g^n \mathrm{e}^{\mathrm{i}m\theta}$ into the scheme) shows that the amplification polynomial is:

$$\Phi(\theta, g) = g^2 + g(2a\lambda \mathrm{i} \sin\theta) - 1$$

The roots are:

$$g_\pm = -a\lambda \mathrm{i} \sin\theta \pm \sqrt{1 - a^2\lambda^2 (\sin\theta)^2}$$

If $|a\lambda| < 1$, then $|g_\pm|^2 = 1$ and the two roots are simple $\forall \theta \in \mathbb{R}$. If $|a\lambda| > 1$ and $\theta = \frac{\pi}{2}$, then either $|g_+| > 1$ or $|g_-| > 1$ and the scheme is unstable. Finally, if $|a\lambda| = 1$ and $\theta = \frac{\pi}{2}$, then the scheme is unstable because there is a double root. □

Second order PDEs

Definition 29. Consider a second order PDE of the form:

$$Au_{tt} + 2Bu_{tx} + Cu_{xx} + Du_t + Eu_x + Fu = G \qquad (4)$$

where $A, B, C, D, E, F, G : \mathbb{R}^2 \to \mathbb{R}$ are smooth functions. The ivp defined in a curve $\gamma(s) = (t,x) = (f(s), g(s))$ is given by the extra conditions:

$$\begin{cases} u(f(s), g(s)) = h(s) \\ u_t(f(s), g(s)) = \phi(s) \\ u_x(f(s), g(s)) = \psi(s) \end{cases}$$

which are tied to the *compatibility condition* $h' = \phi f' + \psi g'$ that follows from the chain rule. The characteristic curves are the curves from which we cannot find the highest order derivatives of u from the initial conditions and the PDE. Differentiating $u_t(s)$ and $u_x(s)$ we get the system of equations for u_{tt}, u_{tx} and u_{xx}:

$$\begin{cases} Au_{tt} + 2Bu_{tx} + Cu_{xx} = G - D\phi - E\psi - Fh \\ f' u_{tt} + g' u_{tx} = \phi' \\ f' u_{tx} + g' u_{xx} = \psi' \end{cases}$$

The determinant of the matrix associated of the system is $\Delta = A(g')^2 - 2Bf'g' + C(f')^2$. Equating this determinant to zero and using the chain rule we get:

$$A\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 - 2B\frac{\mathrm{d}x}{\mathrm{d}t} + C = 0$$

The PDE is called *elliptic* if $AC - B^2 > 0$, *hyperbolic* if $AC - B^2 < 0$ and *parabolic* if $AC - B^2 = 0$.

Definition 30. Consider a finite difference scheme with J steps for a 2n order homogeneous PDE and Λ be a stability region. We say that it is *stable* is given $T > 0$, there exists $C_T > 0$ such that for any grid with $(k,h) \in \Lambda$ and for any initial values $\mathbf{v}_m^j$, $m \in \mathbb{Z}$, $j = 0, \ldots, J-1$ we have

$$\sum_{m \in \mathbb{Z}} \|\mathbf{v}_m^n\|^2 \leq (1 + n^2) C_T \sum_{j=0}^{J-1} \sum_{m \in \mathbb{Z}} \left\|\mathbf{v}_m^j\right\|^2$$

for all $n \in \mathbb{N}$ such that $0 \leq nk \leq T$.

Theorem 31. Consider a finite difference scheme with J steps for a 2n order homogeneous PDE whose amplification factor is g and $\Phi(\theta, g)$ is the amplification polynomial. Then, the scheme is stable if and only if:

- for any root $g_j(\theta)$ of Φ we have $|g_j(\theta)| \leq 1$.
- if $\exists \theta_0$ and k such that $|g_k(\theta_0)| = 1$ then this root is at most double.

Parabolic equations

Proposition 32. Consider the heat equation $u_t = \alpha u_{xx} + f$. We have the following schemes for approximating the solution:

1. *Forward-time central-space*:
$$\frac{v_m^{n+1} - v_m^n}{k} = \alpha \frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + f_m^n$$
2. *Backward-time central-space*:
$$\frac{v_m^{n+1} - v_m^n}{k} = \alpha \frac{v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}}{h^2} + f_m^{n+1}$$
3. *Crank-Nicolson scheme*:
$$\frac{v_m^{n+1} - v_m^n}{k} = \frac{\alpha}{2} \frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + \\ + \frac{\alpha}{2} \frac{v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}}{h^2} + \frac{1}{2}(f_m^{n+1} + f_m^n)$$
4. *Leapfrog scheme*:
$$\frac{u_m^{n+1} - u_m^{n-1}}{2k} = \alpha \frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + f_m^n$$
5. *Du-Fort-Frankel scheme*:
$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} = \alpha \frac{v_{m+1}^n - [v_m^{n+1} + v_m^{n-1}] + v_{m-1}^n}{h^2} + f_m^n$$

Elliptic equations

Definition 33. Let $Pu = f$ be an elliptic PDE on Ω. We define the following boundary conditions on $\partial\Omega$:

1. *Dirichlet*: $u = f$
2. *Neumann*: $\frac{\partial u}{\partial n} = g$
3. *Robin*: $\alpha u + \frac{\partial u}{\partial n} = h$

Definition 34. Consider the following scheme for the Poisson equation $u_{xx} + u_{yy} = f$:
$$\frac{v_{\ell+1,m} - 2v_{\ell,m} + v_{\ell-1,m}}{h^2} + \frac{v_{\ell,m+1} - 2v_{\ell,m} + v_{\ell,m-1}}{h^2} = f_{\ell,m} \tag{5}$$
where we have chosen the same step size h for both x and y directions. We define the *discrete laplacian* as:
$$(\Delta_h v)_{\ell,m} := \frac{v_{\ell+1,m} + v_{\ell-1,m} + v_{\ell,m+1} + v_{\ell,m-1} - 4v_{\ell,m}}{h^2}$$

Theorem 35 (Discrete maximum principle). Consider the Poisson equation $u_{xx} + u_{yy} = f$, let $v = (v_{\ell,m})$ be the scheme of Eq. (5) and suppose that $\Delta_h v \geq 0$ on a region Ω. Then:
$$\max_{\overline{\Omega}} v = \max_{\partial\Omega} v$$

Sketch of the proof. The condition $\Delta_h v \geq 0$ is equivalent to:
$$v_{\ell,m} \leq \frac{1}{4}(v_{\ell+1,m} + v_{\ell-1,m} + v_{\ell,m+1} + v_{\ell,m-1})$$
Now note that if there is a maximum in the interior of the region, its four neighbours must be equal to it. □

Corollary 36 (Discrete minimum principle). Consider the Poisson equation $u_{xx} + u_{yy} = f$, let $v = (v_{\ell,m})$ be the scheme of Eq. (5) and suppose that $\Delta_h v \leq 0$ on a region Ω. Then:
$$\min_{\overline{\Omega}} v = \min_{\partial\Omega} v$$

Proof. Apply 35 Discrete maximum principle to $-v$. □

Theorem 37. Consider the Poisson equation $u_{xx} + u_{yy} = f$, let $v = (v_{\ell,m})$ be the scheme of Eq. (5) defined on $\Omega = [0,1]^2$. If $v = 0$ on $\partial\Omega$ then:
$$\|v\|_\infty \leq \frac{1}{8} \|\Delta_h v\|_\infty$$

Proof. From $(\Delta_h v)_{\ell,m} = f_{\ell,m}$ in the internal nodes of the grid, we have that $|\Delta_h v| \leq \|f\|_\infty$. Now consider the non-negative function $w_{\ell,m}$ defined as:
$$w_{\ell,m} := \frac{1}{4}\left[(x_\ell - 1/2)^2 + (y_m - 1/2)^2\right]$$
An easy check shows that $(\Delta_h w)_{\ell,m} = 1$ and $\|w\|_{L^\infty(\partial\Omega)} = \frac{1}{8}$. So on the one hand, $\Delta_h(v - \|f\|_\infty w) \leq 0$ and by the 36 Discrete minimum principle:
$$-\|f\|_\infty \|w\|_{L^\infty(\partial\Omega)} \leq v_{\ell,m} - \|f\|_\infty w_{\ell,m}$$
And on the other hand, $\Delta_h(v + \|f\|_\infty w) \geq 0$ and by the 35 Discrete maximum principle:
$$\|f\|_\infty \|w\|_{L^\infty(\partial\Omega)} \geq v_{\ell,m}$$
Thus:
$$\|v\|_\infty \leq \|w\|_{L^\infty(\partial\Omega)} \|f\|_\infty = \frac{1}{8} \|\Delta_h v\|_\infty$$
□

Theorem 38. Let u be the solution to $\Delta u = f$ with Dirichlet boundary condition on the unit square and let $v_{\ell,m}$ be the solution to $\Delta_h v = f_{\ell,m}$ with $v_{\ell,m} = u(x_\ell, y_m)$ on the boundary. Then:
$$\|u - v\|_\infty \leq Ch^2 \left\|\partial^4 u\right\|_\infty$$
for some constant $C \in \mathbb{R}$, where $\left\|\partial^4 u\right\|_\infty := \max\left\{\left\|{\partial_x}^4 u\right\|_\infty, \left\|{\partial_y}^4 u\right\|_\infty\right\}$

Proof. Note that $\Delta_h u = f + \varepsilon$, with $|\varepsilon| \leq \tilde{C}h^2 \left\|\partial^4 u\right\|_\infty$ for some constant $\tilde{C} \in \mathbb{R}$. Since $u - v = 0$ on the boundary, by Theorem 37 we have:
$$\|u - v\|_\infty \leq \frac{1}{8} \|f + \varepsilon - f\|_\infty \leq Ch^2 \left\|\partial^4 u\right\|_\infty$$
□

2. | Introduction to finite element methods

The *finite element method* is one of the most popular, general, powerful and elegant approaches for approximating the solutions of PDEs. Unlike finite difference methods, it naturally handles complicated domains (useful for engines and aeroplanes) and minimally regular data (such as discontinuous forcing terms).
There are four basic ingredients in the finite element method:

1. A variational formulation of the problem in an infinite-dimensional space V.
2. A variational formulation of the problem in a finite-dimensional space $V_h \subset V$.
3. The construction of a basis for V_h.
4. The assembly and solution of the resulting linear system of equations.

The variational formulation

Definition 39. Let $\Omega \subseteq \mathbb{R}^n$ be an open bounded connected set such that ∂U is of class $\mathcal{C}^1$, $f \in \mathcal{C}(\Omega)$ and $g \in \mathcal{C}(\partial\Omega)$. Consider the following Dirichlet problem of finding $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}(\overline{\Omega})$ such that:

$$\begin{cases} -\Delta u = f & \text{in } U \\ u = 0 & \text{on } \partial U \end{cases} \tag{6}$$

Let

$$V := \{v : \Omega \to \mathbb{R} : \|v\|_{L^2(\Omega)} + \|\nabla v\|_{L^2(\Omega)} < \infty, v|_{\partial\Omega} = 0\}$$

The *variational formulation* (or *weak formulation*) of the problem is to find $u \in V$ such that:

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}\mathbf{x} = \int_\Omega f v \,\mathrm{d}\mathbf{x} \quad \forall v \in V \tag{7}$$

Remark. The variational formulation can be obtained by multiplying Eq. (6) by v and using the **?? ??**.

Theorem 40. If $f \in \mathcal{C}(\Omega)$, then the solutions to Eq. (7) are $\mathcal{C}^2(\Omega)$.

Lemma 41. If $u \in V$ is a solution to Eq. (7), then u is a solution to Eq. (6).

Proof. Note that $\nabla u \cdot \nabla v = \mathbf{div}(v\nabla u) - v\Delta u$. Thus, using the **?? ??** we have:

$$\begin{aligned} 0 &= \int_\Omega \nabla u \cdot \nabla v - f v \,\mathrm{d}\mathbf{x} \\ &= \int_\Omega v(-\Delta u - f)\,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} v \nabla u \cdot \mathbf{n}\,\mathrm{ds} \\ &= \int_\Omega v(-\Delta u - f)\,\mathrm{d}\mathbf{x} \end{aligned}$$

because $v = 0$ on $\partial\Omega$. Now using the **?? ??**, we conclude that we must have $-\Delta u = f$ in Ω. □

Definition 42 (Galerkin approximation). Let $V_h \subset V$ be a finite-dimensional subspace of V. The *Galerkin approximation* of Eq. (7) is to find $u_h \in V_h$ such that:

$$\int_\Omega \nabla u_h \cdot \nabla v_h \,\mathrm{d}\mathbf{x} = \int_\Omega f v_h \,\mathrm{d}\mathbf{x} \quad \forall v_h \in V_h \tag{8}$$

Construction of function spaces

Definition 43 (Mesh). A *mesh* is a geometric decomposition of a domain Ω into a finite collection of *cells* $\{K_i\}_{i=1}^N$ such that:

1. $\mathrm{Int}(K_i) \cap \mathrm{Int}(K_j) = \varnothing$ for all $i \neq j$.
2. $\bigcup_{i=1}^N K_i = \overline{\Omega}$.

The cells are usually chosen to be n-simplexes or n-parallelepipeds.

Definition 44. The *finite element method* (*FEM*) is a particular choice of Galerkin approximation, where the discrete function space V_h is:

$$V_h := \{v \in \mathcal{C}(\Omega) : v \text{ is piecewise linear when restricted to a cell}\}$$

Note that the functions in V_h are uniquely determined by its values at the vertices of the cell because of the unicity of the interpolating polynomial. The vertices of the cells are called *nodes*.

Definition 45. Given the locations $\mathbf{x}_i$ of the M *nodes* in $\mathrm{Int}\,\Omega$, we define the *nodal basis* $(\phi_1, \ldots, \phi_M)$ as the functions ϕ_i such that:

$$\phi_i(\mathbf{x}_j) = \delta_{ij}$$

Lemma 46. The nodal basis is indeed a basis of V_h.

Proof. Let $v \in V_h$. Then, v can be written as:

$$v = \sum_{i=1}^M v(\mathbf{x}_i)\phi_i$$

Since it is uniquely determined by its values at the nodes, the equality holds. So, $\langle \phi_1, \ldots, \phi_M \rangle = V_h$. Furthermore, if we have $\sum_{i=1}^M c_i\phi_i = 0$, then evaluating at $\mathbf{x}_j$ we have $c_j = 0\ \forall j = 1, \ldots, M$. □

Linear algebraic formulation

Proposition 47. Given a mesh of Ω, consider the space $V_h \subset V$ and its associate nodal basis. Suppose:

$$u_h = \sum_{i=1}^M u_i\phi_i \qquad v_h = \sum_{i=1}^M v_i\phi_i$$

Then, if $\mathbf{u} = (u_1, \ldots, u_M)^{\mathrm{T}}$ we have:

$$\mathbf{Au} = \mathbf{b}$$

where $\mathbf{A} = (a_{ij})$ and $\mathbf{b} = (b_i)$ are defined as:

$$a_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \,\mathrm{d}\mathbf{x} \qquad b_i = \int_\Omega f\phi_i \,\mathrm{d}\mathbf{x}$$

The matrix $\mathbf{A}$ is usually called the *stiffness matrix* and $\mathbf{b}$ the *load vector*.

Proof. Since, $u_h \in V_h$, and using the linearity of the integral we have:

$$\int_\Omega \nabla u_h \cdot \nabla v_h \, \mathrm{d}\mathbf{x} = \int_\Omega f v_h \, \mathrm{d}\mathbf{x}$$

$$\sum_{i=1}^{M} v_i \int_\Omega \nabla u_h \cdot \nabla \phi_i \, \mathrm{d}\mathbf{x} = \sum_{i=1}^{M} v_i \int_\Omega f \phi_i \, \mathrm{d}\mathbf{x}$$

As this holds for all $v_h \in V_h$, we have that this is equivalent to

$$\int_\Omega \nabla u_h \cdot \nabla \phi_i \, \mathrm{d}\mathbf{x} = \int_\Omega f \phi_i \, \mathrm{d}\mathbf{x}$$

for $i = 1, \dots, M$, which implies:

$$\sum_{j=1}^{M} u_j \int_\Omega \nabla \phi_j \cdot \nabla \phi_i \, \mathrm{d}\mathbf{x} = \int_\Omega f \phi_i \, \mathrm{d}\mathbf{x}$$

□

Remark. Solving this system of linear equations we obtain the approximation by finite elements of the Dirichlet problem for the Poisson equation (Eq. (6)). Note that the approximate solution is a piecewise linear function which may not be differentiable at the vertices of the cells. Even so, the approximate solution converges to the exact solution as the mesh is refined.

Remark. On the computation of the coefficients a_{ij} we should proceed as follows:

$$a_{ij} = \sum_{m=1}^{N} \int_{K_m} \nabla \phi_i \cdot \nabla \phi_j \, \mathrm{d}\mathbf{x}$$

Note, however, that many of these integrals will be zero. Indeed, if $\{P_i\}_{i=1,\dots,M}$ are the nodes of the mesh and $P_i \notin K_m$ for some i, then $\varphi_i = 0$ on the nodes of K_m, and therefore $\varphi_i = 0$ and $\nabla \varphi_i = 0$ on K_m. Thus, we only need to compute the integrals for K_m such that $P_i, P_j \in K_m$. For these (a priori) non-zero integrals, we use a reference n-simplex to compute them.

Proposition 48. Let S be an n-simplex with vertices at $Q_0 = \mathbf{0}$, $Q_i = \mathbf{e}_i$ (thought as a point), $i = 1, \dots, n$, where $\mathbf{e}_i$ is the i-th vector of the canonical basis of $\mathbb{R}^n$. Consider the FEM method for the Eq. (6). Then:

$$\int_{K_m} \nabla \varphi_{K_m,\ell} \cdot \nabla \varphi_{K_m,k} \, \mathrm{d}\mathbf{x} = \frac{d_m}{n!} \nabla \psi_\ell \Big(\mathbf{D}\boldsymbol{\sigma}_m{}^{\mathrm{T}} \mathbf{D}\boldsymbol{\sigma}_m \Big)^{-1} \nabla \psi_k{}^{\mathrm{T}}$$

where $\boldsymbol{\sigma}_m$ is the affine transformation that carries the reference simplex S onto K_m, $d_m = |\det \mathbf{D}\boldsymbol{\sigma}_m|$, $\phi_{K_m,\ell}$ denote that basis function such that evaluates to 1 at the ℓ-th vertex of K_m (with an ordering fixed), $\ell = 0, \dots, n$, and:

$$\psi_k(\mathbf{x}) = \begin{cases} 1 - \sum_{i=1}^{n} x_i & k = 0 \\ x_k & k = 1, \dots, n \end{cases}$$

Proof. Note $\psi_k(Q_k) = \delta_{ij}$ and so by the unicity of the interpolation we have $\varphi_{K_m,\ell} \circ \sigma_m = \psi_\ell$, $\ell = 0, \dots, n$. Thus, by the chain rule, $\nabla \psi_\ell = \nabla \varphi_{K_m,\ell} \mathbf{D}\boldsymbol{\sigma}_m$, and so:

$$\begin{aligned} \int_{K_m} \nabla \varphi_{K_m,\ell} \cdot \nabla \varphi_{K_m,k} \, \mathrm{d}\mathbf{y} &= \int_S \nabla \varphi_{K_m,\ell} \cdot (\nabla \varphi_{K_m,k})^{\mathrm{T}} d_m \, \mathrm{d}\mathbf{x} \\ &= \int_S \nabla \psi_\ell (\mathbf{D}\boldsymbol{\sigma}_m)^{-1} \Big[(\mathbf{D}\boldsymbol{\sigma}_m)^{-1} \Big]^{\mathrm{T}} \nabla \psi_k{}^{\mathrm{T}} d_m \, \mathrm{d}\mathbf{x} \\ &= \frac{d_m}{n!} \nabla \psi_\ell \Big(\mathbf{D}\boldsymbol{\sigma}_m{}^{\mathrm{T}} \mathbf{D}\boldsymbol{\sigma}_m \Big)^{-1} \nabla \psi_k{}^{\mathrm{T}} \end{aligned}$$

where we used that the volume of the n-simplex S is $1/n!$ and all the terms inside the integral is constant. □

Remark. With the same idea, the integrals b_i can be computed as:

$$\int_{K_m} f \varphi_{K_m,\ell} = d_m \int_S f \circ \sigma_m \psi_\ell$$

and we use a quadrature formula to approximate over a triangle.