Numerical methods for the solution of ordinary differential equations

In this exercise sheet, one considers a generic, scalar initial value problem of the form

$$u'(t) = f(t, u(t)), \text{ for } t > t_0, \text{ and } u(t_0) = \eta,$$

in which the function f satisfies the assumptions of the Cauchy–Lipschitz (or Picard–Lindelöf) theorem and η is a given real number, and its solution by a numerical method, defined by a recurrence relation, or *scheme*, of the form

$$\forall n \in \mathbb{N},\ u^{n+1} = u^n + h\Phi_f(t_n, u^n; h), \text{ and } u^0 = \eta,$$

in which Φ_f is the increment function of the method, depending on the function f, h is the length of the time step, and the terms of sequence $(u^n)_{n\in\mathbb{N}}$ are approximations of the values of the solution u at times $t_n = t_0 + nh$.

Exercise 1. Verify that Heun's method [Heu00], whose scheme is

$$u^{n+1} = u^n + \frac{h}{2}\left(f(t_n, u^n) + f\left(t_n + h, u^n + hf(t_n, u^n)\right)\right)$$

and modified Euler's method, whose scheme is

$$u^{n+1} = u^n + hf\left(t_n + \frac{h}{2}, u^n + \frac{h}{2}f(t_n, u^n)\right)$$

are explicit 2-stage Runge–Kutta methods and give their respective Butcher tableaus.

Exercise 2 (order of some classic Runge–Kutta methods). By studying their respective local truncation errors, find the order of the implicit Euler method, whose scheme is

$$u^{n+1} = u^n + hf(t_{n+1}, u^{n+1}),$$

of the trapezoidal rule method, whose scheme is

$$u^{n+1} = u^n + \frac{h}{2}\left(f(t_n, u^n) + f(t_{n+1}, u^{n+1})\right),$$

of Heun's method, whose scheme is

$$u^{n+1} = u^n + \frac{h}{2}\left(f(t_n, u^n) + f\left(t_n + h, u^n + hf(t_n, u^n)\right)\right).$$

Exercise 3 (order barrier). Consider the initial-value problem

$$x'(t) = \lambda x(t),\ x(0) = 1.$$

1. Apply a single step un pas d'une méthode de Runge–Kutta explicite à s niveaux pour la résolution numérique de ce problème et montrer que u^{n+1} est un polynôme en h de degré au plus égal à s.
2. Infer that the order of an explicit Runge–Kutta method cannot be larger than its number of stages.

Exercise 4 (order conditions without simplifying assumption). Obtain the order conditions on the coefficients a_{21}, b_1, b_2, c_1 and c_2 of an explicit 2-stage Runge–Kutta method of order 2 which does not satisfy the usual simplifying assumptions $c_1 = 0$ and $c_2 = a_{21}$.

Exercise 5 (the ERK methods from [vdHou72]). Construct the Runge–Kutta methods which have a Butcher tableau of the form

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & c_2 & & \\ c_3 & 0 & c_3 & \\ \hline & 0 & 0 & 1 \end{array}.$$

What is the benefit of these methods in terms of storage requirements?

Exercise 6. Show that, for an explicit s-stage Runge–Kutta methods of order[1] s, for which the coefficients a_{ij} et b_i, $1 \leq j < i \leq s$, are nonnegative, the Lipschitz constant Λ of the increment function Φ_f of the method satisfies

$$1 + h\Lambda < e^{hL},$$

where the positive number L is the Lipschitz constant of f.
Hint: use the order conditions for the coefficients of the methods, recalled hereafter:

- $b_1 = 1$ *for* $s = 1$,
- $b_1 + b_2 = 1$ *and* $b_2 c_2 = \frac{1}{2}$ *for* $s = 2$,
- $b_1 + b_2 + b_3 = 1$, $b_2 c_2 + b_3 c_3 = \frac{1}{2}$, $b_2 {c_2}^2 + b_3 {c_3}^2 = \frac{1}{3}$ *and* $b_3 a_{32} c_2 = \frac{1}{6}$ *for* $s = 3$,
- $b_1 + b_2 + b_3 + b_4 = 1$, $b_2 c_2 + b_3 c_3 + b_4 c_4 = \frac{1}{2}$, $b_2 {c_2}^2 + b_3 {c_3}^2 + b_4 {c_4}^2 = \frac{1}{3}$, $b_3 a_{32} c_2 + b_4 (a_{42} c_2 + a_{43} c_3) = \frac{1}{6}$, $b_2 {c_2}^3 + b_3 {c_3}^3 + b_4 {c_4}^3 = \frac{1}{4}$, $b_3 c_3 a_{32} c_2 + b_4 c_4 a_{42} c_2 + b_4 c_4 a_{43} c_3 = \frac{1}{8}$, $b_3 a_{32} {c_2}^2 + b_4 a_{42} {c_2}^2 + b_4 a_{43} {c_2}^3 = \frac{1}{12}$ *and* $b_4 a_{43} a_{32} c_2 = \frac{1}{24}$ *for* $s = 4$.

Exercise 7. Consider the family of methods defined by the scheme

$$u^{n+1} = u^n + h\big((1-\omega) f(t_n, u^n) + \omega f(t_{n+1}, u^{n+1})\big),$$

where ω is a real number chosen in the interval $[0, 1]$. Show that such a method is A-stable, that is, its region of absolute stability includes the entire complex half-plane with negative real part $\mathbb{C}_- = \{z \in \mathbb{C} \,|\, \mathrm{Re}(z) < 0\}$, if and only if $\omega \geq \frac{1}{2}$.

[1] Recall that this is only possible for $1 \leq s \leq 4$.

Finite difference methods

Exercise 1. Assuming that the function u is thrice differentiable at point x in $\mathbb{R}$, find real numbers α, β and γ such that, for any positive real number h,

$$u'(x) = \frac{\alpha\, u(x+2h) + \beta\, u(x) + \gamma\, u(x-h)}{h} + O(h^2).$$

Exercise 2 (analysis of the Lax–Wendroff method for the transport equation). Let u be a smooth solution to the transport equation

$$\forall t \in \mathbb{R}_+^*,\ \forall x \in \mathbb{R},\ \frac{\partial u}{\partial t}(t,x) + c\,\frac{\partial u}{\partial x}(t,x) = 0, \tag{1}$$

c being a given nonzero real number.

1. Show that

$$\forall (n,j) \in \mathbb{N}\times\mathbb{Z},\ u(t_{n+1},x_j) = u(t_n,x_j) - c\,\Delta t\,\frac{\partial u}{\partial x}(t_n,x_j) + \frac{c^2\Delta t^2}{2}\,\frac{\partial^2 u}{\partial x^2}(t_n,x_j) + O(\Delta t^3),$$

   where $t_n = n\Delta t$ and $x_j = j\Delta x$.

For the numerical solution of the equation, one uses a family of finite difference schemes, indexed by the parameter μ and defined by

$$\forall n \in \mathbb{N},\ \forall j \in \mathbb{Z},\ \frac{u_j^{n+1} - u_j^n}{\Delta t} + c\,\frac{u_{j+1}^n - u_{j-1}^n}{2\,\Delta x} - \mu\frac{\Delta t}{\Delta x^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) = 0.$$

2. Determine the value of μ for which the above scheme is consistent and accurate of order 2 at least in both time and space. Show that it corresponds to the Lax–Wendroff method.

3. Compute the amplification factor of the Lax–Wendroff method and infer that it is stable in the ℓ^2-norm under the condition $|c|\,\Delta t \le \Delta x$.

Exercise 3 (the upwind scheme). To approximate the solution to the transport equation (1), one can use the first-order upwind scheme, defined by

$$\forall n \in \mathbb{N},\ \forall j \in \mathbb{Z},\ \begin{aligned} \frac{u_j^{n+1} - u_j^n}{\Delta t} + c\,\frac{u_j^n - u_{j-1}^n}{\Delta x} &= 0 \quad \text{for } c > 0, \\ \frac{u_j^{n+1} - u_j^n}{\Delta t} + c\,\frac{u_{j+1}^n - u_j^n}{\Delta x} &= 0 \quad \text{for } c < 0. \end{aligned}$$

Study the stablity of this difference scheme. Show in particular that it is unstable if the two formulas depending on the sign of c are replaced with one another.

Exercise 4 (analysis of the θ-scheme family in ℓ^2-norm). One is interested in the numerical approximation of the solution to the initial-value problem for the heat equation

$$\forall t \in \mathbb{R}_+^*,\ \forall x \in \mathbb{R},\ \frac{\partial u}{\partial t}(t,x) - \frac{\partial^2 u}{\partial x^2}(t,x) = 0, \tag{2}$$

$$\forall x \in \mathbb{R},\ u(0,x) = u_0(x), \tag{3}$$

in which u_0 is a continuous function in $L^2(\mathbb{R})$. For this purpose, a family of finite difference schemes is considered, which reads

$$\forall n \in \mathbb{N},\ \forall j \in \mathbb{Z},\ \frac{u_j^{n+1} - u_j^n}{\Delta t} - \theta\,\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{(\Delta x)^2} - (1-\theta)\,\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2} = 0,$$

the parameter θ being a real number in $[0,1]$, and complemented by the initialisation

$$\forall j \in \mathbb{Z},\ u_j^0 = u_0(j\,\Delta x).$$

1. Depending on the value of the parameter θ, discuss the order of accuracy and the explicitness or implicitness of the above difference schemes.

2. Using the von Neumann analysis, discuss as well the stability of these schemes in the ℓ^2-norm .

Exercise 5 (analysis of an explicit scheme for the heat equation). To approximate the solution to the initial value problem (2)-(3), one considers the following finite difference scheme

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}-\frac{u_{j+1}^n-2\,u_j^n+u_{j-1}^n}{(\Delta x)^2}=0,\ n\in\mathbb{N},\ j\in\mathbb{Z},$$

completed by the initialisation

$$u_j^0=u_0(j\,\Delta x),\ j\in\mathbb{Z}.$$

1. Give an algorithm for the computation of the numerical approximation. Explain why this approximate solution exhibits a finite propagation speed.

2. Under the hypothesis that $\frac{\Delta t}{(\Delta x)^2}\leq\frac{1}{2}$, prove the following discrete maximum principle

$$\forall j\in\mathbb{Z},\ a\leq u_j^0\leq b,\Rightarrow\forall n\in\mathbb{N},\ \forall j\in\mathbb{Z},\ a\leq u_j^n\leq bj\in\mathbb{Z},$$

   satisfied by the approximation. Deduce that

$$\forall n\in\mathbb{N},\ \|\boldsymbol{u}^n\|_\infty\leq\|\boldsymbol{u}^0\|_\infty,$$

   which proves that the method is stable in the ℓ^∞-norm.

3. One defines the truncation error of the method by

$$\forall n\in\mathbb{N},\ \forall j\in\mathbb{Z},\ \varepsilon_j^{n+1}=u(t_{n+1},x_j)-u(t_n,x_j)-\frac{\Delta t}{(\Delta x)^2}\left(u(t_n,x_{j+1})-2\,u(t_n,x_j)+u(t_n,x_{j-1})\right).$$

   It is assumed that the solution u to the problem is of class $\mathscr{C}^2$ with respect to time and of class $\mathscr{C}^4$ with respect to space and that its partial derivatives are uniformly bounded in $[0,T]\times\mathbb{R}$. Using a Taylor expansion, show that, for any natural integer n such that $(n+1)\,\Delta t\leq T$, one has the estimate

$$\left\|\boldsymbol{\varepsilon}^{n+1}\right\|_\infty\leq C\left((\Delta t)^2\left\|\frac{\partial^2 u}{\partial t^2}\right\|_{L^\infty([0,T]\times\mathbb{R})}+\Delta t\,(\Delta x)^2\left\|\frac{\partial^4 u}{\partial x^4}\right\|_{L^\infty([0,T]\times\mathbb{R})}\right),$$

   in which

$$\|f\|_{L^\infty([0,T]\times\mathbb{R})}=\sup_{(t,x)\in[0,T]\times\mathbb{R}}|f(t,x)|.$$

4. One defines the error of the method by

$$\forall n\in\mathbb{N},\ \forall j\in\mathbb{Z},\ e_j^n=u(t_n,x_j)-u_j^n.$$

   Write down the relations between the sequences $(\boldsymbol{e}^n)_{n\in\mathbb{N}}$ and $(\boldsymbol{\varepsilon}^n)_{n\in\mathbb{N}}$ and prove the following error estimate

$$\sup_{(n+1)\Delta t\leq T}\|\boldsymbol{e}^n\|_\infty\leq CT\left(\Delta t\left\|\frac{\partial^2 u}{\partial t^2}\right\|_{L^\infty([0,T]\times\mathbb{R})}+(\Delta x)^2\left\|\frac{\partial^4 u}{\partial x^4}\right\|_{L^\infty([0,T]\times\mathbb{R})}\right).$$

Exercise 6 (the Richardson and Du Fort–Frankel methods). For the numerical solution of the heat equation (2), one considers the finite difference scheme of the Richardson (or leap-frog) method

$$\frac{u_j^{n+1}-u_j^{n-1}}{2\,\Delta t}-\frac{u_{j+1}^n-2\,u_j^n+u_{j-1}^n}{(\Delta x)^2}=0,\ n\in\mathbb{N}^*,\ j\in\mathbb{Z}.$$

1. What is the practical inconvenience of the above difference scheme?

2. Show that this scheme has second order of accuracy in both time and space, but that it is unconditionally unstable.

The difference scheme of the Du Fort–Frankel method [DF53] is obtained by replacing the term u_j^n in the difference scheme of the Richardson method by the mean value $\frac{u_j^{n+1}+u_j^{n-1}}{2}$.

3. Show that this new method is explicit and unconditionally stable.

4. Analyse the truncation error of the difference scheme and conclude that it is consistent with the equation only if the ratio $\frac{\Delta t}{\Delta x}$ tends to 0 as Δt and Δx tend to 0. Was this result expected?

Exercise 7 (the Lax–Richtmyer equivalence theorem [LR56]). Consider the initial value problem

$$\frac{d}{dt}u(t) = Au(t),\ 0 \le t \le T,$$
$$u(0) = u_0,$$

in which u is an unknown function in the Banach space X, A is a linear operator whose domain $D(A)$ dense in X and which generates a strongly continuous semigroup S on X, and u_0 is the initial datum in $D(A)$.

For the numerical approximation of the solution to this problem, one uses a finite difference method, represented by a strongly continuous function F from $[0,+\infty)$ to $\mathscr{L}(X)$, with $F(0) = id_X$, satisfying the following *consistency* condition: it holds

$$\lim_{\Delta t \to 0} \left\| \frac{F(\Delta t)u(t) - u(t)}{\Delta t} - Au(t) \right\| = 0$$

uniformly for all t in $[0,T]$.

Such a finite difference method is called *stable* if there exists a positive constant M such that

$$\|F(\Delta t)^n\|_{\mathscr{B}(X)} \le M,\ n \in \mathbb{N},\ 0 \le n\Delta t \le T.$$

Finally, such a method is called *convergent* if, for any sequences $(\Delta t_k)_{k\in\mathbb{N}}$ and $(n_k)_{k\in\mathbb{N}}$, such that $\lim_{k\to+\infty} \Delta t_k = 0$, $\lim_{k\to+\infty} n_k = +\infty$, and $\lim_{k\to+\infty} n_k \Delta t_k = t$ with $0 \le t \le T$,

$$\lim_{k\to+\infty} \|F(\Delta t_k)^{n_k} u_0 - S(t)u_0\| = 0.$$

The goal of this exercise is to prove the celebrated result due to Lax and Richtmyer: *for a consistent finite difference method, stability is a necessary and sufficient condition for convergence*.

1. Show that the above consistency condition is equivalent to having

$$\lim_{\Delta t \to 0} \left\| \frac{F(\Delta t)S(t)u_0 - S(t+\Delta t)u_0}{\Delta t} \right\| = 0$$

   uniformly for all t in $[0,T]$.

2. Prove the necessary part of the theorem.
   Hint: argue by contradiction using the ***uniform boundedness principle*** *(also known as the Banach–Steinhaus theorem) recalled hereafter: let X and Y be two Banach spaces and $\mathscr{B}(X,Y)$ be the space of all continuous linear operators from X into Y. Suppose that K is a collection of elements of $\mathscr{B}(X,Y)$. If K is strongly bounded, i.e., for all x in X,*

$$\sup\{\|Lx\|_Y \mid L \in K\} < +\infty,$$

   then it is uniformly bounded, i.e.

$$\sup\{\|L\|_{\mathscr{B}(X,Y)} \mid L \in K\} < +\infty.$$

3. In this question, in order to obtain the sufficient part of the theorem, it is assumed that the finite difference method is stable. Observe that, by the strong continuity of S, it suffices to show that

$$\lim_{k\to+\infty} \|F(\Delta t_k)^{n_k} u_0 - S(n_k\Delta t_k)u_0\| = 0,$$

   where $(\Delta t_k)_{k\in\mathbb{N}}$ and $(n_k)_{k\in\mathbb{N}}$ are sequences such that $\lim_{k\to+\infty} \Delta t_k = 0$, $\lim_{k\to+\infty} n_k = +\infty$, and $\lim_{k\to+\infty} n_k\Delta t_k = t$ with t in $[0,T]$, to prove that the method is convergent.

   (a) Establish that

$$\forall k \in \mathbb{N},\ F(\Delta t_k)^{n_k} - S(n_k\Delta t_k) = \sum_{i=0}^{n_k-1} F(\Delta t_k)^{n_k-1-i}(F(\Delta t_k) - S(\Delta t_k))S(\Delta t_k)^i.$$

   (b) Prove that

$$\forall \varepsilon > 0,\ \exists K \in \mathbb{N},\ \forall s \in [0,t],\ \forall k \in \mathbb{N},\ k \ge K,\ \|F(\Delta t_k)S(s)u_0 - T(\Delta t_k)S(s)u_0\| \le \varepsilon\Delta t_k.$$

   (c) Conclude.

Finite element methods

Weak formulation of an elliptic boundary value problem

Some boundary-value problems for elliptic partial differential equations can be cast into the following form: given a Hilbert space V, a bilinear form $a(\cdot,\cdot)$ on $V \times V$ and a continuous linear form $ell(\cdot)$ on V, find u in V satisfying

$$\forall v_h \in V,\ a(u,v) = \ell(v). \tag{4}$$

According to the Lax–Milgram theorem [LM54], if the form a is both bounded and coercive on V, that is, there exist positive constants C and c such that

$$\forall (u,v) \in V^2,\ |a(u,v)| \leq C \|u\|_V \|V\|_V, \tag{5}$$

$$\forall v \in V,\ a(v,v) \geq c \|V\|_V{}^2, \tag{6}$$

which is assumed from here on, then this problem is well-posed, that is, there exists a unique solution and it holds

$$\|u\|_V \leq \frac{1}{c} \|\ell\|_{V'}.$$

Exercise 1 (Dirichlet, Neumann and Robin problems for an elliptic partial differential equation). Consider a boundary-value problem with a partial differential equation of the form

$$-\sum_{j=1}^{d}\sum_{k=1}^{d} \partial_{x_j}\left(a_{jk}\partial_{x_k} u\right) + \sum_{j=1}^{d} b_j \partial_{x_j} u + cu = f,$$

in a bounded open set Ω in $\mathbb{R}^d$ with regular boundary $\partial\Omega$, where a_{jk}, b_j, c in $L^\infty(\Omega)$ and f in $L^2(\Omega)$ are given functions. It assumed this equation is elliptic, that is, there exists a positive constant α such that

$$\forall \xi \in \mathbb{R}^d,\ \forall x \in \Omega,\ \sum_{j=1}^{d}\sum_{k=1}^{d} a_{jk}(x)\xi_j\xi_k \geq \alpha \sum_{j=1}^{d} \xi_j{}^2.$$

It is complemented with boundary conditions, which can be of the following types:

- *Dirichlet boundary conditions*: one requires that $u = g$ on $\partial\Omega$ (in the sense of traces) for a given g in $L^2(\partial\Omega)$.
- *Neumann boundary conditions*: one requires that $\sum_{j=1}^{d}\sum_{k=1}^{d} a_{jk}\partial_{x_k} u\, \nu_j = g$ for a given g in $L^2(\partial\Omega)$, where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)$ is the outward unit normal vector to $\partial\Omega$.
- *Robin boundary conditions*: one requires that $e\,u + \sum_{j=1}^{d}\sum_{k=1}^{d} a_{jk}\partial_{x_k} u\, \nu_j = g$ for a given g in $L^2(\partial\Omega)$ and e in $L^\infty(\partial\Omega)$, where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)$ is the outward unit normal vector to $\partial\Omega$.

Set $\beta = \alpha^{-1} \sum_{j=1}^{d} \|b_j\|_{L^\infty(\Omega)}{}^2$.

1. Prove that the Dirichlet problem is well-posed in $H^1(\Omega)$ if g belongs to $H^{\frac{1}{2}}(\partial\Omega)$ and

$$\text{for almost all } x \text{ in } \Omega,\ c(x) - \frac{\beta}{2} \geq 0.$$

2. Prove that the Neumann problem is well-posed in $H^1(\Omega)$ if

$$\text{for almost all } x \text{ in } \Omega,\ c(x) - \frac{\beta}{2} \geq \gamma \geq 0,\ \text{for almost all } x \text{ in } \partial\Omega,\ e(x) \geq \delta \geq 0,$$

and either $\gamma > 0$ or $\delta > 0$.

Hint: use the Lax-Milgram theorem, with the Hölder, Young and Poincaré inequalities to verify the continuity and the coercivity of the forms. Where needed, employ a lifting of the boundary condition.

Conforming Galerkin approximation

A *conforming Galerkin approach* for the solution of this problem consists in choosing a finite-dimensional (hence closed) subspace V_h of V, the subscript h standing for a discretisation parameter meant to tend to 0, and looking for u_h in V_h satisfying

$$\forall v_h \in V_h,\ a(u_h, v_h) = \ell(v_h). \tag{7}$$

Since V_h, equipped with the scalar product on V, is a Hilbert space, the Lax–Milgram theorem immediatly yields the well-posedness of problem (7).

Exercise 2 (Céa's lemma [Céa64]). Prove the following result: *let u be the solution to* (4) *and u_h be the solution to* (7) *for a given $V_h \subset V$. Then, one has*

$$\|u - u_h\|_V \leq \frac{C}{c} \inf_{v_h \in V_h} \|u - v_h\|_V,$$

where C and c are the constants appearing in (5) *and* (6)*, respectively*.

Exercise 3 (the symmetric case). When the bilinear form appearing in the weak formulation of the problem is symmetric, one can obtain a stronger estimate than the one provided by Céa's lemma, by characterising solution to (4) as minimisers of the functional defined by

$$\forall v \in V,\ J(v) := \frac{1}{2} a(v, v) - \ell(v).$$

In what follows, it is assumed that the bilinear form $a(\cdot, \cdot)$ is coercive and symmetric.

1. Prove that a function u in V satisfies (4) if and only if it is the minimiser of J over V.
   Hint: consider $J(u + tv)$ for any function v in V and t in $\mathbb{R}$.

2. Deduce the solution u_h in V_h of (7) satisfies

   $$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a,$$

   that is, u_h is the best approximation of u in V_h in the energy norm defined by

   $$\forall v \in V,\ \|v\|_a = \sqrt{a(v, v)}.$$

Exercise 4 (the Aubin–Nitsche trick). To estimate the error in a weaker norm than the one on the space V, a duality argument is required. The goal of this exercise is to prove the following result, due to Aubin [Aub67] and Nitsche [Nit68]. *Let H be a Hilbert space, with scalar product $(\cdot, \cdot)_H$, and V be a closed subspace of H that becomes a Hilbert space when equipped with a scalar product $(\cdot, \cdot)_V$. Let the embedding $V \hookrightarrow H$ be continuous*[1]*. Let u_h be the solution in V_h of* (7)*. Then, there exists a positive constant C such that*

$$\|u - u_h\|_H \leq C \|u - u_h\|_V \sup_{g \in H \setminus \{0\}} \left(\frac{1}{\|g\|_H} \min_{v_h \in V_h} \|\varphi_g - v_h\|_V \right),$$

where, for any g in H, the function φ_g is the unique solution to the ***adjoint problem***

$$\forall w \in V,\ a(w, \varphi_g) = (g, w)_H.$$

1. Using the above adjoint problem, show that, for any g in H,

   $$\forall v_h \in V_h,\ (g, u - u_h)_H \leq C \|u - u_h\|_V \|\varphi_g - v_h\|_V,$$

   where C is the constant apperaing in (5).

2. Deduce that

   $$\forall v_h \in V_h,\ \|u - u_h\|_H \leq C \|u - u_h\|_V \sup_{g \in H \setminus \{0\}} \frac{\|\varphi_g - v_h\|_V}{\|g\|_H}.$$

   Hint: use the Riesz representation theorem in H.

3. Conclude.

[1] One may think of $H = L^2(\Omega)$ and $V = H^1(\Omega)$, for instance.

Finite elements

Exercise 5 (simplicial Lagrange finite element). Let $\{a_0,\dots,a_d\}$ be a family of points in $\mathbb{R}^d$ and assume that the sequence of vectors $a_1-a_0,\dots,a_d-a_0$ is linearly independent. Then, the convex hull of $\{a_0,\dots,a_d\}$ is called a *(d-)simplex*.
Given a simplex K in $\mathbb{R}^d$, one can consider the associated *barycentric coordinates* $\{\lambda_0,\dots,\lambda_d\}$ defined as follows:

$$\forall i\in\{0,\dots,d\},\ \forall x\in\mathbb{R}^d,\ \lambda_i(x)=1-\frac{(x-a_i)\cdot\boldsymbol{\nu_i}}{(a_j-a_i)\cdot\boldsymbol{\nu_i}},$$

where $\boldsymbol{\nu_i}$ is the outward unit normal vector to the face F_i of K opposite to a_i, and a_j is an arbitrary vertex in F_i (this definition is independent of the choice of this vextex). The barycentric coordinate λ_i is an afine function; it is equal to 1 at a_i and vanishes on F_i. Furthermore, its level-sets are hyperplanes parallel to F_i and the barycenter of K has barycentric coordinates $\left(\frac{1}{d+1},\dots,\frac{1}{d+1}\right)$. The barycentric coordinates satisfy the following properties

$$\forall i\in\{0,\dots,d\},\ \forall x\in K,\ 0\le\lambda_i(x)\le 1,$$

$$\forall x\in\mathbb{R}^d,\ \sum_{i=0}^{d}\lambda_i(x)=1 \text{ and } \sum_{i=0}^{d}\lambda_i(x)(x-a_i)=\mathbf{0}.$$

Let K be a simplex in $\mathbb{R}^d$, k be a nonzero natural integer, and let $P=P_k$ be the space of polynomial functions in the variables $x_1,\dots,x_d$ with real coefficients and global degree at most k. Consider the set of nodes $\{a_1,\dots,a_{\dim(P)}\}$ with barycentric coordinates $\left(\frac{i_0}{k},\dots,\frac{i_d}{k}\right)$, $0\le i_0,\dots,i_d\le k$, $i_0+\dots,+i_d=k$. Finally, let $\Sigma=\{\sigma_1,\dots,\sigma_{\dim(P)}\}$ be the set of linear forms on P such that

$$\forall i\in\{1,\dots,\dim(P)\},\ \forall p\in P,\ \sigma_i(p)=p(a_i).$$

The goal of this exercise is to prove that the triplet $\{K,P,\Sigma\}$ is a finite element.

1. Let p belong to P and assume that p vanishes on the $\mathbb{R}^d$-hyperplane of equation $\lambda=0$, with λ a nonzero affine function in the variables $x_1,\dots,x_d$. Prove that there exists a polynomial function q in the variables $x_1,\dots,x_d$ with global degree at most $k-1$ such that $p=\lambda q$.
   Hint: since affine transformations map the space of polynomial functions of degree at most k to itself, one may assume that p vanishes on the hyperplace orthogonal to one of the coordinate axis.
2. Conclude.

Exercise 6 (geometrical estimates). For an element domain K of a given finite element (K,P,Σ), one defines

- the diameter $h_K:=\max_{(x,y)\in K^2}\|x-y\|$,
- the insphere diameter $\rho_K:=2\max\{\rho>0\,|\,B(x,\rho)\subset K \text{ for some } x \text{ in } K\}$.

Let T_K be the affine mapping generating K of a reference element domain $\hat{K}$, that is

$$T_K:\hat{K}\to K,\quad \hat{x}\mapsto A_K\hat{x}+b_K.$$

1. Establish, for any sufficiently smooth function v defined on K, the transformation rule

$$\int_{T_K(\hat{K})}v(x)\,\mathrm{d}x=\int_{\hat{K}}(v\circ T_K)(\hat{x})\,|\det(A_K)|\,\mathrm{d}\hat{x}.$$

2. Deduce that $|\det(A_K)|=\frac{\mathrm{vol}(K)}{\mathrm{vol}(\hat{K})}$.
3. Show that $\|A_K\|\le\frac{h_K}{\rho_{\hat{K}}}$, where $\|A_K\|=\sup_{\|\hat{x}\|=1}\|A_K\hat{x}\|$, and infer that $\|{A_K}^{-1}\|\le\frac{h_{\hat{K}}}{\rho_K}$.

Polynomial interpolation in Sobolev spaces

Exercise 7 (the Bramble–Hilbert lemma [BH70]). The Bramble–Hilbert lemma is an essential tool in proving bounds for the interpolation error in the finite element method, which amount to consistency error estimates. Let us state it. *Let Ω be an open subset of $\mathbb{R}^d$ with a Lipschitz-continuous boundary. For some natural integer k and some real number p in $[0,+\infty]$, let f be a continuous linear form on the space $W^{k+1,p}(\Omega)$ with the annihilation property that*

$$\forall q\in P_k(\Omega),\ f(q)=0.$$

Then, there exists a constant $C(\Omega)$ such that

$$\forall v\in W^{k+1,p}(\Omega),\ |f(v)|\le C(\Omega)\|f\|_{W^{k+1,p}(\Omega)'}\,|v|_{W^{k+1,p}(\Omega)}$$

The goal of this exercise is to prove this result.

1. We first establish a preliminary result, due to Deny and Lions.

   (a) Let $N = \dim(P_k(\Omega))$. Show that there exist continuous linear forms $f_1, \dots, f_N$ over $W^{k+1,p}(\Omega)$ such that, for any q in $P_k(\Omega)$, $f_1(q) = \cdots = f_N(q) = 0$ if and only if $q = 0$.
   Hint: consider the dual basis of $P_k(\Omega)$ and use the Hahn–Banach extension theorem.

   (b) Arguing by contradiction, we will now prove that there exists a positive constant $C(\Omega)$ such that

   $$\forall v \in W^{k+1,p}(\Omega),\ \|v\|_{W^{k+1,p}(\Omega)} \leq C(\Omega)\left(|v|_{W^{k+1,p}(\Omega)} + \sum_{i=1}^{N} |f_i(v)|\right).$$

   We thus assume assume that there exists a sequence $(v_l)_{l\in\mathbb{N}}$ in $W^{k+1,p}(\Omega)$ such that

   $$\forall l \in \mathbb{N},\ \|v_l\|_{W^{k+1,p}(\Omega)} = 1,\ \text{and}\ \lim_{l\to+\infty}\left(|v_l|_{W^{k+1,p}(\Omega)} + \sum_{i=1}^{N} |f_i(v_l)|\right) = 0.$$

   i. Show that there exist a subquence, again denoted $(v_l)_{l\in\mathbb{N}}$, and a function v in $W^{k,p}(\Omega)$ such that

   $$\lim_{l\to+\infty} \|v_l - v\|_{W^{k,p}(\Omega)} = 0.$$

   ii. Deduce that the subsequence converges in $W^{k+1,p}(\Omega)$ and that its limit v is such that, for any multi-index α with $|\alpha| = k+1$, $|D^\alpha v| = 0$.

   iii. Infer that v is almost evywhere equal to a polynomial function of degree lower of equal to k and conclude that $v = 0$.
   Hint: use the Sobolev imbedding theorem.

   iv. Conclude.

   (c) Use the preceding result to prove that there exists a positive constant $C(\Omega)$ such that

   $$\forall v \in W^{k+1,p}(\Omega),\ \inf_{q\in P_k(\Omega)} \|v+q\|_{W^{k+1,p}} \leq C(\Omega)\,|v|_{W^{k+1,p}(\Omega)}.$$

2. Show that

   $$\forall v \in W^{k+1,p}(\Omega),\ |f(v)| \leq \|f\|_{W^{k+1,p}(\Omega)'} \inf_{q\in P_k(\Omega)} \|v+q\|_{W^{k+1,p}(\Omega)}.$$

3. Conclude.

Exercise 8 (interpolation error estimates on a reference element). Let (K, P, Σ) be a finite element with $P_k \subset P$ for some natural integer k and all degrees of liberty in Σ bounded on $W^{k+1,p}(K)$, $1 \leq p \leq +\infty$. Prove that there exists a positive constant c, depending only on d, k, p, l and (K, P, Σ), such that

$$\forall v \in W^{k+1,p}(K),\ \forall l \in \{0, \dots, k+1\},\ |v - I_K v|_{W^{l,p}(K)} \leq c\ |v|_{W^{k+1,p}(K)},$$

where $I_K v$ denotes the local interpolant of v.
Hint: use the Bramble–Hilbert lemma.

Spectral methods

Continuous Fourier series

Exercise 1 (the Dirichlet kernel). A truncated Fourier series can also be expressed in the convolution form. Let u be a continuous and periodic function of bounded variation on $[0, 2\pi)$ and N be a natural integer. One has

$$\sum_{k=-N}^{N} \hat{u}_k e^{ikx} = \frac{1}{2\pi} \int_0^{2\pi} \mathscr{D}_N(x-t)u(t)\,\mathrm{d}t,$$

where $\hat{u}_k$ is the kth Fourier coefficient of u and $\mathscr{D}_N$ is the *Dirichlet kernel* given by

$$\mathscr{D}_N(x) = \sum_{k=-N}^{N} e^{ikx} = 1 + 2\sum_{k=1}^{N} \cos(kx) = \frac{\sin\left(\left(N+\frac{1}{2}\right)x\right)}{\sin\left(\frac{x}{2}\right)}.$$

1. Show that $\mathscr{D}_N$ is an even function, symmetric about $x = \frac{1}{2}$.
2. Show that
$$\frac{1}{2\pi} \int_0^{2\pi} \mathscr{D}_N(t)\,\mathrm{d}t = 1,$$
and
$$\frac{1}{2\pi} \int_0^{2\pi} (\mathscr{D}_N(t))^2\,\mathrm{d}t = 2N+1.$$
3. Show that
$$\forall N \in \mathbb{N} \setminus \{0,1\},\ \int_0^{2\pi} |\mathscr{D}_N(t)|\,\mathrm{d}t \leq c\,\ln(N),$$
where c is a positive constant independent of N.
4. Prove that, for any u in $X_N = \mathrm{Span}\{e^{ikx} \mid k \in \{-N, \dots, N\}\}$,
$$\forall x \in [0, 2\pi),\ u(x) = \frac{1}{2\pi} \int_0^{2\pi} \mathscr{D}_N(x-t)u(t)\,\mathrm{d}t,$$
and
$$\|u\|_{L^\infty([0,2\pi))} \leq \sqrt{2N+1}\|u\|_{L^2([0,2\pi))}.$$

Discrete Fourier series

Exercise 2 (Fourier interpolant of a function). Let N be an even natural integer, u be a continuous and periodic function u on $[0, 2\pi)$ and

$$\forall j \in \{0, \dots, N-1\},\ x_j = \frac{2\pi j}{N}.$$

The *Fourier interpolant of u at the node* $x_0, \dots, x_{N-1}$ is

$$I_N(u)(x) = \frac{1}{N} \sum_{k=-N/2+1}^{N/2} \tilde{u}_k e^{ikx},$$

where $\tilde{u}_k$ is the kth discrete Fourier coefficient of u, defined by

$$\tilde{u}_k = \sum_{j=0}^{N-1} u(x_j) e^{-ikx_j}.$$

1. Write $I_N u$ in the Lagrange basis, that is

$$I_N u(x) = \sum_{j=0}^{N-1} u(x_j)\ell_j(x),$$

where

$$\forall (j,k) \in \{0,\dots,N-1\}^2,\ \ell_j(x_k) = \delta_{jk},$$

and show that

$$\forall j \in \{0,\dots,N-1\},\ \ell_j(x) = \frac{1}{N}\sin\left(\frac{x-x_j}{2}N\right)\cot\left(\frac{x-x_j}{2}\right).$$

2. The differentiation process in the physical space of the Fourier interpolant can be formulated as a matrix-vector multiplication, that is, there exists a so-called *first-order differentiation matrix* D_N such that the values of of the derivative of the interpolant at the nodes are given by

$$\forall j \in \{0,\dots,N-1\},\ (I_N u)'(x_j) = \sum_{l=0}^{N-1} (D_N)_{j+1\,l+1} u(x_l).$$

Show that

$$\forall (j,l) \in \{1,\dots,N\}^2,\ (D_N)_{jl} = \begin{cases} \frac{1}{2}(-1)^{j+l}\cot\left(\frac{(j-l)\pi}{N}\right) & \text{if } j \neq l, \\ 0 & \text{if } j = l. \end{cases}$$

Hint: use the result of the preceding question.

References

[Aub67] J.-P. AUBIN. Behavior of the error of the approximate solutions of boundary value problems for linear elliptic operators by Galerkin's and finite difference methods. *Ann. Scuola Norm. Sup. Pisa Sci. Fis. Mat. (3)*, 21(4):599–637, 1967.

[BH70] J. H. BRAMBLE and S. R. HILBERT. Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation. *SIAM J. Numer. Anal.*, 7(1):112–124, 1970. DOI: 10.1137/0707006.

[Céa64] J. CÉA. Approximation variationnelle des problèmes aux limites. *Ann. Inst. Fourier (Grenoble)*, 14(2):345–444, 1964. DOI: 10.5802/aif.181.

[DF53] E. C. DU FORT and S. P. FRANKEL. Stability conditions in the numerical treatment of parabolic differential equations. *Math. Tables Aids Comp.*, 7(43):135–152, 1953. DOI: 10.1090/S0025-5718-1953-0059077-7.

[Heu00] K. HEUN. Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen. *Z. Math. Phys.*, 45:23–38, 1900.

[LM54] P. D. LAX and A. N. MILGRAM. Parabolic equations. In L. BERS, S. BOCHNER, and F. JOHN, editors, *Contributions to the theory of partial differential equations*. Volume 33, Annals of mathematics studies, pages 167–190. Princeton University Press, 1954. DOI: 10.1515/9781400882182-010.

[LR56] P. D. LAX and R. D. RICHTMYER. Survey of the stability of linear finite difference equations. *Comm. Pure Appl. Math.*, 9(2):267–293, 1956. DOI: 10.1002/cpa.3160090206.

[Nit68] J. NITSCHE. Ein kriterium für die quasi-optimalität des Ritzschen Verfahrens. *Numer. Math.*, 11(4):346–348, 1968. DOI: 10.1007/BF02166687.

[vdHou72] P. J. van der HOUWEN. Explicit Runge–Kutta formulas with increased stability boundaries. *Numer. Math.*, 20(2):146–164, 1972. DOI: 10.1007/BF01404404.