



Lecture Notes 3

Finite Volume Discretization of the Heat Equation

We consider finite volume discretizations of the one-dimensional variable coefficient heat equation, with Neumann boundary conditions

$$\begin{aligned} u_t - \partial_x(k(x)\partial_x u) &= S(t, x), & 0 < x < 1, \quad t > 0, \\ u(0, x) &= f(x), & 0 < x < 1, \\ u_x(t, 0) &= u_x(t, 1) = 0, & t \geq 0. \end{aligned} \tag{1}$$

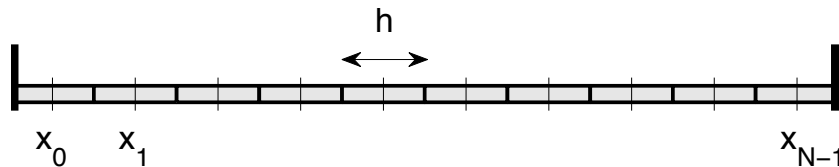
The coefficient $k(x)$ is strictly positive.

1 Semi-discrete approximation

By semi-discretization we mean discretization only in space, not in time. This approach is also called *method of lines*.

Discretization

We discretize space into N equal size grid cells (bins) of size $h = 1/N$, and define $x_j = h/2 + jh$, so that x_j is the center of cell j , see figure. The edges of cell j are then $x_{j-1/2}$ and $x_{j+1/2}$.

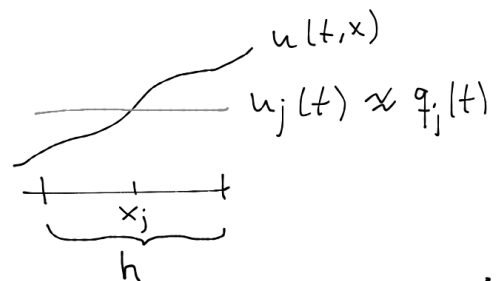


Unknowns

In a finite volume method the unknowns approximate the *average* of the solution over a grid cell. More precisely, we let $q_j(t)$ be the approximation

$$q_j(t) \approx u_j(t) := \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(t, x) dx.$$

Note the contrast with finite difference methods, where pointwise values are approximated, and finite element methods, where basis function coefficients are approximated.



Exact update formula

We derive an exact update formula for $u_j(t)$, the exact local averages. Integrating (1) over cell j and dividing by h we get

$$\begin{aligned} \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u_t(t, x) dx &= \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} \partial_x(k(x)\partial_x u) dx + \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} S(t, x) dx \\ &= \frac{k(x_{j+1/2})u_x(t, x_{j+1/2}) - k(x_{j-1/2})u_x(t, x_{j-1/2})}{h} + \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} S(t, x) dx. \end{aligned}$$

Upon defining the flux

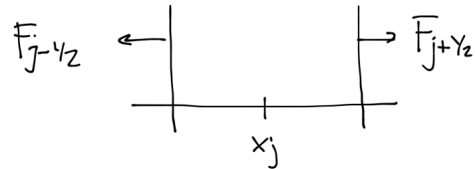
$$F_j(t) = F(t, x_j) = -k(x_j)u_x(t, x_j),$$

and the local average of the source

$$S_j(t) = \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} S(t, x) dx,$$

we get the exact update formula

$$\frac{du_j(t)}{dt} = -\frac{F_{j+1/2}(t) - F_{j-1/2}(t)}{h} + S_j(t).$$



The fluxes $F_{j+1/2}$ and $-F_{j-1/2}$ then represents how much heat flows out through the left and right boundary of the cell.

Note, this is an instance the conservation law in integral form,

$$\frac{d}{dt} \int_V u dV + \int_S \vec{F} \cdot \vec{n} dS = \int_V S dV,$$

where we have picked V as the interval $[x_{j-1/2}, x_{j+1/2}]$, and scaled by $|V| = h$.

Approximation of the flux

To use the exact update formula as the basis for a numerical scheme we must approximate the fluxes $F_{j\pm 1/2}$. Since the value in the midpoint of the cell is a second order approximation of the average, we have for smooth u ,

$$\begin{aligned} F_{j-1/2}(t) &= -k(x_{j-1/2})u_x(t, x_{j-1/2}) = -k(x_{j-1/2})\frac{u(t, x_j) - u(t, x_{j-1})}{h} + O(h^2) \\ &= -k(x_{j-1/2})\frac{u_j(t) - u_{j-1}(t)}{h} + O(h^2). \end{aligned}$$

We therefore use

$$F_{j-1/2}(t) \approx \tilde{F}_{j-1/2}(t) = -k(x_{j-1/2})\frac{q_j(t) - q_{j-1}(t)}{h}.$$

as approximation. This leads to the numerical scheme for inner points $1 \leq j \leq N - 2$,

$$\begin{aligned} \frac{dq_j(t)}{dt} &= -\frac{\tilde{F}_{j+1/2}(t) - \tilde{F}_{j-1/2}(t)}{h} + S_j(t) \\ &= \frac{k(x_{j+1/2})(q_{j+1}(t) - q_j(t)) - k(x_{j-1/2})(q_j(t) - q_{j-1}(t))}{h^2} + S_j(t). \end{aligned} \tag{2}$$

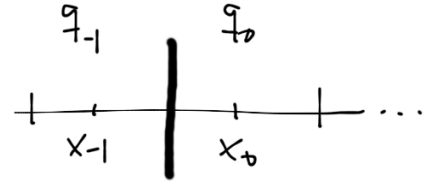
Hence, with $k_j := k(x_j)$,

$$\boxed{\frac{dq_j}{dt} = \frac{k_{j+1/2}q_{j+1} - (k_{j+1/2} + k_{j-1/2})q_j + k_{j-1/2}q_{j-1}}{h^2} + S_j}, \quad (3)$$

for $j = 1, \dots, N - 2$. This is a second order approximation.

Boundary conditions

To complete the scheme (3) we need update formulae also for the boundary points $j = 0$ and $j = N - 1$. These must be derived by taking the boundary conditions into account. We introduce the *ghost cells* $j = -1$ and $j = N$ which are located just outside the domain. The boundary conditions are used to fill these cells with values q_{-1} and q_N , based on the values q_j in the interior cells. The same update formula (3) as before can then be used also for $j = 0$ and $j = N - 1$.



Let us consider our boundary condition $u_x = 0$ at $x = 0$. (We can also think of this as a "no flux" condition, $F = 0$.) We formally extend the definition of the solution u for $x < 0$, i.e. outside the domain, and, as before, approximate

$$\begin{aligned} 0 = u_x(t, 0) &= \frac{u(t, x_0) - u(t, x_{-1})}{h} + O(h^2) \\ &= \frac{u_0(t) - u_{-1}(t)}{h} + O(h^2) \quad \Rightarrow \quad u_{-1}(t) = u_0(t) + O(h^3). \end{aligned}$$

Replacing u_j by our approximation q_j and dropping the $O(h^2)$ term we get an expression for q_{-1} in terms of q_0 as the boundary rule

$$q_{-1}(t) = q_0(t).$$

We now insert this into the update formula (3) for $j = 0$,

$$\frac{dq_0}{dt} = \frac{k_{1/2}q_1 - (k_{1/2} + k_{-1/2})q_0 + k_{-1/2}q_{-1}}{h^2} + S_0 = k_{1/2} \frac{q_1 - q_0}{h^2} + S_0. \quad (4)$$

In exactly the same way we obtain for $j = N - 1$ that $q_N(t) = q_{N-1}(t)$ and therefore

$$\frac{dq_{N-1}}{dt} = k_{N-3/2} \frac{q_{N-2} - q_{N-1}}{h^2} + S_{N-1}. \quad (5)$$

Remark 1 *Dirichlet boundary conditions* $u(t, 0) = u(t, 1) = 0$ can be approximated to second order in two ways.

First, one can use a shifted grid, $x_j = jh$ so that x_0 and x_N , the centers of cells 0 and N , are precisely on the boundary. Then one does not need ghost cells; one just sets $q_0 = q_N = 0$. Note that the number of unknowns are now only $N - 1$, so $A \in \mathbb{R}^{(N-1) \times (N-1)}$ etc.

Second, one can take the average of two cells to approximate the value in between,

$$0 = u(t, 0) = \frac{u(t, x_0) + u(t, x_{-1})}{2} + O(h^2) = \frac{u_0(t) + u_{-1}(t)}{2} + O(h^2),$$

leading to the approximations

$$q_{-1} = -q_0, \quad q_N = -q_{N-1}.$$

Matrix form

We put all the formulae (3), (4), (5) together and write them in matrix form. Introduce

$$\mathbf{q} = \begin{pmatrix} q_0 \\ \vdots \\ q_{N-1} \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} S_0 \\ \vdots \\ S_{N-1} \end{pmatrix}, \quad \mathbf{q}, \mathbf{S} \in \mathbb{R}$$

and

$$A = \frac{1}{h^2} \begin{pmatrix} -k_{1/2} & k_{1/2} & & & & \\ k_{1/2} & -(k_{1/2} + k_{3/2}) & k_{3/2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & k_{N-5/2} & -(k_{N-5/2} + k_{N-3/2}) & k_{N-3/2} & \\ & & & -k_{N-3/2} & k_{N-3/2} & \end{pmatrix} \in \mathbb{R}^{N \times N}. \quad (6)$$

Then we get the linear ODE system

$$\frac{d\mathbf{q}(t)}{dt} = A\mathbf{q}(t) + \mathbf{S}(t). \quad (7)$$

Hence, in this semi-discretization the time-dependent PDE has been approximated by a system of ODEs, where the matrix A is a discrete approximation of the second order differential operator $\partial_x k(x) \partial_x$, including its boundary conditions.

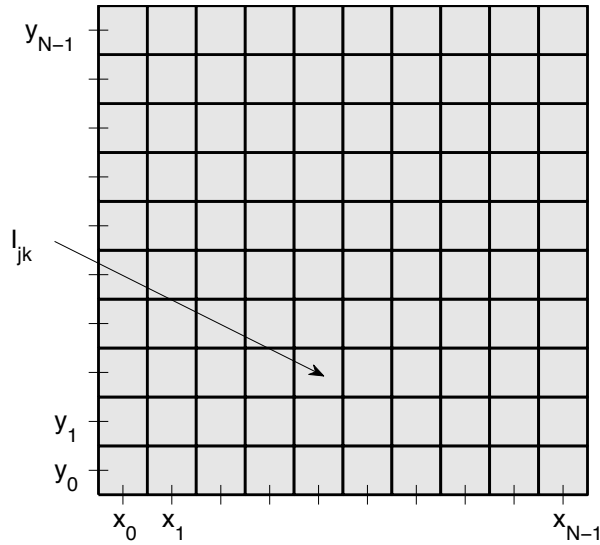
1.1 Brief outline of extensions to 2D

The same strategy can be used in 2D. We give a cursory description of the main steps here. To simplify things we consider the constant coefficient problem,

$$\begin{aligned} u_t - \Delta u &= S(t, x, y), & 0 < x < 1, \quad 0 < y < 1, \quad t > 0, \\ u(0, x, y) &= f(x, y), & 0 < x < 1, \quad 0 < y < 1, \\ u_x(t, 0, y) &= u_x(t, 2\pi, y) = u_y(t, x, 0) = u_y(t, x, 2\pi) = 0, & 0 < x < 1, \quad 0 < y < 1, \quad t \geq 0. \end{aligned}$$

Discretization

We discretize the domain $[0, 1]^2$ into $N \times N$ equal size grid cells of size $h \times h$, where $h = 1/N$. We define $x_j = h/2 + jh$ and $y_k = h/2 + kh$ and denote the cell with center (x_j, y_k) by I_{jk} .



Unknowns

The unknowns are now $q_{jk}(t)$, which are approximations of the average of the solution over the grid cell I_{jk} ,

$$q_{jk}(t) \approx u_{jk}(t) := \frac{1}{h^2} \int_{I_{jk}} u(t, x, y) dx dy.$$

Exact update formula

The update formula is again an instance of the conservation law in integral form where we pick the volume V as I_{jk} and scale by $|I_{jk}| = h^2$,

$$\frac{d}{dt} \frac{1}{h^2} \int_{I_{jk}} u dx dy + \frac{1}{h^2} \int_{\partial I_{jk}} \vec{F} \cdot \vec{n} dS = \frac{1}{h^2} \int_{I_{jk}} S dx dy,$$

where $F = -\nabla u$. Upon defining

$$S_{jk}(t) := \frac{1}{h^2} \int_{I_{jk}} S(t, x, y) dx dy,$$

this can be written as

$$\frac{du_{jk}(t)}{dt} = -\frac{1}{h^2} \int_{\partial I_{jk}} \vec{F} \cdot \vec{n} dS + S_{jk}(t).$$

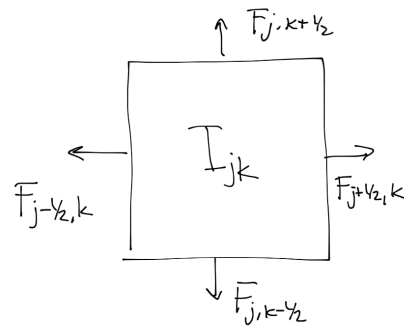
Let $F = (f_1, f_2)$ and define the average flux through each side of the cell,

$$F_{j,k\pm 1/2}(t) := \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} f_2(t, x, y_{k\pm 1/2}) dx, \quad F_{j\pm 1/2,k}(t) := \frac{1}{h} \int_{y_{k-1/2}}^{y_{k+1/2}} f_1(t, x_{j\pm 1/2}, y) dy,$$

we get the exact formula

$$\frac{du_{jk}}{dt} = -\frac{1}{h} (F_{j+1/2,k} - F_{j-1/2,k} + F_{j,k+1/2} - F_{j,k-1/2}) + S_{jk}. \tag{8}$$

Again, the fluxes $F_{j,k\pm 1/2}$ and $F_{j\pm 1/2,k}$ represent the heat flux (upto sign) out through the four sides of the cell.



Approximation of the flux

We use the same type of approximation as in 1D for $F = -\nabla u$. We get for instance

$$\begin{aligned} F_{j,k-1/2} &= -\frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u_y(t, x, y_{k-1/2}) dx = -\frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} \frac{u(t, x, y_k) - u(t, x, y_{k-1})}{h} dx + O(h^2) \\ &= -\frac{1}{h^2} \int_{x_{j-1/2}}^{x_{j+1/2}} u(t, x, y_k) dx + \frac{1}{h^2} \int_{x_{j-1/2}}^{x_{j+1/2}} u(t, x, y_{k-1}) dx + O(h^2) \\ &= -\frac{1}{h^3} \int_{I_{jk}} u(t, x, y) dx dy + \frac{1}{h^3} \int_{I_{j,k-1}} u(t, x, y) dx dy + O(h^2) \\ &= \frac{-u_{jk} + u_{j,k-1}}{h} + O(h^2). \end{aligned}$$

Replacing u_{jk} by q_{jk} , dropping $O(h^2)$ and inserting in the exact formula (8) we get the five-point formula,

$$\boxed{\frac{dq_{jk}}{dt} = \frac{1}{h^2}(q_{j+1,k} + q_{j-1,k} + q_{j,k+1} + q_{j,k-1} - 4q_{jk}) + S_{jk}.}$$

Boundary conditions is done as in 1D. The matrix form is more complicated and the subject of Homework 1.

2 Properties of the semi-discrete approximation

2.1 Conservation

When $S = 0$ in (1) we have seen that the solution has the conservation property

$$\int_0^1 u(t, x) dx = \text{constant} = \int_0^1 f(x) dx.$$

In fact this holds for all conservation laws with "no-flux" boundary conditions $F = 0$, which is easily seen from the integral form of the PDE. An analogue of this holds also for the semi-discrete approximation. More precisely, let us define

$$Q(t) = \sum_{j=0}^{N-1} q_j(t)h.$$

Then

$$Q(t) = \text{constant}. \tag{9}$$

Moreover, if we start the approximation with exact values, $q_j(0) = u_j(0)$, then it follows that $Q(t)$ is exactly the integral of the solution u for all time,

$$Q(t) = \text{constant} = Q(0) = \sum_{j=0}^{N-1} u_j(0)h = \sum_{j=0}^{N-1} \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(0, x) h dx = \int_0^1 u(0, x) dx = \int_0^1 u(t, x) dx.$$

We prove the exact discrete conservation (9) in two ways.

- Recall from (2) that when $S = 0$,

$$\frac{dq_j(t)}{dt} = -\frac{\tilde{F}_{j+1/2}(t) - \tilde{F}_{j-1/2}(t)}{h}. \quad (10)$$

Then

$$\frac{dQ}{dt} = \sum_{j=0}^{N-1} \frac{dq_j}{dt} h = \sum_{j=0}^{N-1} \tilde{F}_{j-1/2}(t) - \tilde{F}_{j+1/2}(t) = \tilde{F}_{-1/2} - \tilde{F}_{N-1/2} = 0,$$

where we used the boundary conditions (4) and (5) which implies that

$$\tilde{F}_{-1/2} = k_{-1/2} \frac{q_0 - q_{-1}}{h} = 0, \quad \tilde{F}_{N-1/2} = k_{N-1/2} \frac{q_N - q_{N-1}}{h} = 0.$$

Note that this discrete conservation property is true for *any* discretization of the type (10) if $\tilde{F}_{-1/2} = \tilde{F}_{N-1/2}$, regardless of how the fluxes $\tilde{F}_{j\pm 1/2}$ are computed.

- Recall from (7) that when $S = 0$,

$$\frac{d\mathbf{q}(t)}{dt} = A\mathbf{q}(t).$$

Moreover,

$$Q(t) = \mathbf{1}^T \mathbf{q} h, \quad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^N.$$

Then

$$\frac{dQ(t)}{dt} = \mathbf{1}^T \frac{d\mathbf{q}(t)}{dt} h = \mathbf{1}^T A\mathbf{q}(t) h = 0,$$

since $\mathbf{1}^T A\mathbf{q}(t) = (A\mathbf{1})^T \mathbf{q}(t)$ by the symmetry of A and $A\mathbf{1}$ are the row sums of A which are zero, by (6).

2.2 Maximum principle

We have seen before that in the continuous case that when $S = 0$ in (1), the maximum value of the solution $u(t, x)$ in $[0, T] \times [0, 1]$ is either attained on the boundary $x \in \{0, 1\}$ or for the initial data at $t = 0$ (regardless of boundary conditions). The corresponding result in the semi-discrete case says that

$$q^* = \max_{j \in \{0, \dots, N-1\}} \sup_{0 \leq t < T} |q_j(t)|$$

is attained either for $j = 0$, $j = N - 1$ or for the initial data $t = 0$. This is easily shown by contradiction. Suppose the maximum is attained at $t = t^* > 0$ and that it is strictly larger than $q_0(t^*)$ and $q_{N-1}(t^*)$. Then, there must be an interior index $j^* \in [1, N - 2]$ such that

$$q_{j^*-1}(t^*) < q_{j^*}(t^*), \quad q_{j^*}(t^*) \geq q_{j^*+1}(t^*).$$

Therefore

$$\frac{dq_{j^*}(t^*)}{dt} = k_{j^*+1/2} \underbrace{(q_{j^*+1} - q_{j^*})}_{\leq 0} - k_{j^*-1/2} \underbrace{(q_{j^*} - q_{j^*-1})}_{> 0} < 0.$$

Hence, there is an ε such that $q_{j^*}(t) > q_{j^*}(t^*)$ for all $t \in (t^* - \varepsilon, t^*)$, which contradicts the assumption that $q_{j^*}(t^*)$ is a maximum.

Furthermore, as in the continuous case local spatial maximum (minimum) of q_j in the interior cannot increase (decrease) in time.

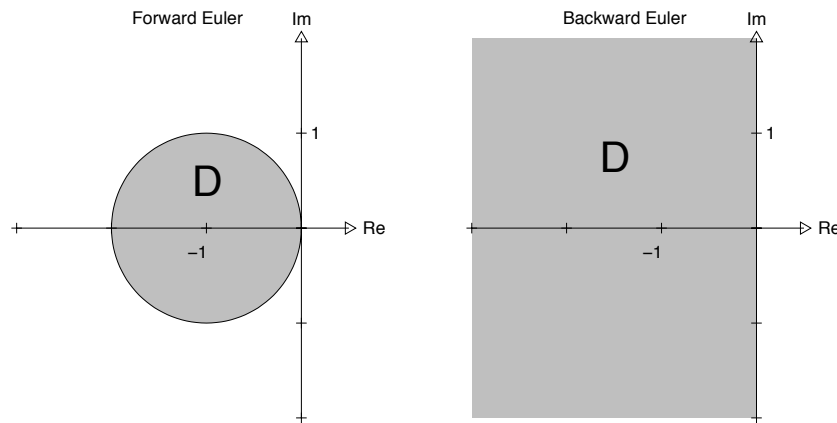


Figure 1. Stability regions \mathcal{D} for the Forward and Backward Euler methods.

3 Fully discrete approximation

The semi-discrete approximation leads to a system of ODEs.

$$\frac{d\mathbf{q}(t)}{dt} = A\mathbf{q}(t) + \mathbf{S}(t). \quad (11)$$

This can be solved by standard numerical methods for ODEs with a time step Δt , e.g. the Forward Euler method

$$u^{n+1} = u^n + \Delta t f(t_n, u^n), \quad t_n = n\Delta t.$$

Applied to (11) this would give the fully discrete scheme for (1),

$$\mathbf{q}^{n+1} = \mathbf{q}^n + \Delta t[A\mathbf{q}^n + \mathbf{S}(t_n)], \quad \mathbf{q}^n \approx \mathbf{q}(t_n).$$

As for any ODE method we must verify its absolute stability: that for our choice of step size Δt

$$\Delta t \lambda_k \in \mathcal{D}, \quad \forall k,$$

where \mathcal{D} is the stability region of the ODE solver and $\{\lambda_k\}$ are the eigenvalues of A . For parabolic problems the real part of the eigenvalues are negative, and the size of them in general grow as $1/h^2$. This is a major difficulty. It means that when the stability region \mathcal{D} is bounded, as in explicit methods, we get a time step restriction of the type

$$\Delta t \leq Ch^2.$$

This is a *severe* restriction, which is seldom warranted from an accuracy point of view. It leads to unnecessarily expensive methods.

Example 1 In the constant coefficient case $k(x) \equiv 1$ the eigenvalues of A are precisely

$$\lambda_k = -\frac{4}{h^2} \sin^2\left(\frac{k\pi h}{2}\right), \quad k = 0, \dots, N-1.$$

This gives the stability condition $\Delta t \leq h^2/2$ for Forward Euler.

The underlying reason for the severe time-step restriction is the fact that parabolic problems include processes on all time-scales: high frequencies decay fast, low frequencies slowly. Their semi-discretization have time-scales spread out all over the interval $[-1/h^2, 0]$, which means that the ODEs are *stiff*. The consequence is that *implicit methods should be used for parabolic problems*. Implicit methods typically have unbounded stability domains \mathcal{D} and have no stability restriction on the time-step — they are *unconditionally stable*. Of course, implicit methods are more expensive per time-step than explicit methods, since a system of equations must be solved, but this is outweighed by the fact that much longer time-steps can be taken. Moreover, when the coefficients do not vary with time, matrices etc. can be constructed once, and re-factored only as changes of time-step make it necessary.

The " θ -method" is a class of ODE methods defined as

$$u^{n+1} = u^n + \Delta t [\theta f(t_{n+1}, u^{n+1}) + (1 - \theta)f(t_n, u^n)], \quad 0 \leq \theta \leq 1.$$

This includes some common methods:

$$\begin{aligned} \theta = 0 & \Rightarrow \text{Forward Euler (explicit, 1st order),} \\ \theta = 1/2 & \Rightarrow \text{Crank-Nicolson (implicit, 2nd order),} \\ \theta = 1 & \Rightarrow \text{Backward Euler (implicit, 1st order),} \end{aligned}$$

Applied to (11) we have

$$\mathbf{q}^{n+1} = \mathbf{q}^n + \Delta t A[\theta \mathbf{q}^{n+1} + (1 - \theta)\mathbf{q}^n] + \Delta t \underbrace{[\theta \mathbf{S}(t_{n+1}) + (1 - \theta)\mathbf{S}(t_n)]}_{\equiv \mathbf{S}_\theta^n}, \quad (12)$$

or

$$(1 - \theta \Delta t A)\mathbf{q}^{n+1} = (1 + (1 - \theta)\Delta t A)\mathbf{q}^n + \Delta t \mathbf{S}_\theta^n.$$

For the constant coefficient problem $k \equiv 1$ one can show the time-step restriction

$$\Delta t \leq h^2 \begin{cases} \frac{1}{2(1-2\theta)}, & \theta < 1/2, \\ \infty, & 1/2 \leq \theta \leq 1, \quad (\text{unconditionally stable}). \end{cases}$$

Remark 2 *The Crank-Nicolson scheme is second order accurate but gives slowly decaying oscillations for large eigenvalues. It is unsuitable for parabolic problems with rapidly decaying transients. The $\theta = 1$ scheme damps all components, and should be used in the initial steps.*

Remark 3 *The most used family of time-stepping schemes for parabolic problems are the Backward Differentiation Formulas (BDF), of order 1 through 5 which are $A(\alpha)$ -stable. They are multistep methods generalizing Backward Euler to higher order. For instance, the second order BDF method is*

$$u^{n+1} = \frac{4}{3}u^n - \frac{1}{3}u^{n-1} + \frac{2}{3}\Delta t f(t_{n+1}, u^{n+1}).$$

BDF methods are also known as Gear's methods and available in MATLAB as ODE15S.

3.1 Fully discrete conservation

For the θ -method we also have discrete conservation when $S \equiv 0$. Let

$$Q^n \equiv \sum_{j=0}^{N-1} q_j^n h = \mathbf{1}^T \mathbf{q}^n h.$$

Then upon multiplying by $\mathbf{1}$ from the left in (12) we get

$$Q^{n+1} = \mathbf{1}^T \mathbf{q}^{n+1} h = \mathbf{1}^T \mathbf{q}^n h + \Delta t \mathbf{1}^T A [\theta \mathbf{q}^{n+1} + (1-\theta) \mathbf{q}^n] h = Q^n + \Delta t (A \mathbf{1})^T [\theta \mathbf{q}^{n+1} + (1-\theta) \mathbf{q}^n] h = Q^n,$$

since as before $A \mathbf{1} = 0$. In particular, if initial data is exact, $q_j^0 = u_j(0)$, then

$$Q^n = \int_0^1 u(t_n, x) dx,$$

for all $n \geq 0$. The same is true for the second order BDF method if the initialization of the first step Q^1 is conservative so that $Q^1 = Q^0$. Then

$$\mathbf{q}^{n+1} = \frac{4}{3} \mathbf{q}^n - \frac{1}{3} \mathbf{q}^{n-1} + \frac{2}{3} \Delta t A \mathbf{q}^{n+1}$$

implies

$$Q^{n+1} = \mathbf{1}^T \mathbf{q}^{n+1} h = \frac{4}{3} \mathbf{1}^T \mathbf{q}^n h - \frac{1}{3} \mathbf{1}^T \mathbf{q}^{n-1} h + \frac{2}{3} \mathbf{1}^T \Delta t A \mathbf{q}^{n+1} h = \frac{4}{3} Q^n - \frac{1}{3} Q^{n-1}.$$

With the stipulated initial data this difference equation has the solution $Q^n = Q^0$ for all $n \geq 0$.

4 Acknowledgement

Part of these notes are based on earlier notes by Prof. Jesper Ooppelstrup.