# Numerical Solution of Partial Differential Equations

Endre Süli

Mathematical Institute
University of Oxford
2024

Lecture 13

# The implicit scheme: stability, consistency and convergence

For $M \geq 2$, we define $\Delta t := T/M$, and for $J \geq 2$ the spatial step is taken to be $\Delta x := (b-a)/J$. We let $x_j := a + j\Delta x$ for $j = 0, 1, \ldots, J$ and $t_m := m\Delta t$ for $m = 0, 1, \ldots, M$.

On the space-time mesh $\{(x_j, t_m) : 0 \leq j \leq J,\ 0 \leq m \leq M\}$ we consider the finite difference scheme

$$
\begin{aligned}
\frac{U_j^{m+1} - 2U_j^m + U_j^{m-1}}{\Delta t^2} - c^2 \frac{U_{j+1}^{m+1} - 2U_j^{m+1} + U_{j-1}^{m+1}}{\Delta x^2} &= f(x_j, t_{m+1}) \quad \text{for} \begin{cases} j = 1, \ldots, J-1, \\ m = 1, \ldots, M-1, \end{cases} \\
U_j^0 &= u_0(x_j) \quad \text{for } j = 0, 1, \ldots, J, \\
U_j^1 &= U_j^0 + \Delta t\, u_1(x_j) \quad \text{for } j = 1, 2, \ldots, J-1, \\
U_0^m = 0 \quad \text{and} \quad U_J^m &= 0 \quad \text{for } m = 1, \ldots, M.
\end{aligned}
\tag{1}
$$

The second numerical initial condition, featuring in equation $(1)_3$, stems from the observation that if $\frac{\partial^2 u}{\partial t^2} \in C([a, b] \times [0, T])$ then

$$\begin{aligned}\frac{u(x_j, \Delta t) - U_j^0}{\Delta t} &= \frac{u(x_j, \Delta t) - u(x_j, 0)}{\Delta t} \\ &= \frac{\partial u}{\partial t}(x_j, 0) + \mathcal{O}(\Delta t) = u_1(x_j) + \mathcal{O}(\Delta t);\end{aligned}$$

thus, by ignoring the $\mathcal{O}(\Delta t)$ term and replacing $u(x_j, \Delta t)$ by its numerical approximation $U_j^1$ we obtain $(1)_3$.

Once the values of $U_j^{m-1}$ and $U_j^m$, for $j = 0, \ldots, J$, have been computed (or have been specified by the initial data, in the case of $m = 1$), the subsequent values $U_j^{m+1}$, $j = 0, \ldots, J$, are computed by solving a system of $J - 1$ linear algebraic equations for the $J - 1$ unknowns $U_j^{m+1}$, $j = 0, \ldots, J - 1$, for $m = 0, \ldots, M - 1$. The finite difference scheme (1) is therefore referred to as the *implicit scheme* for the initial-boundary-value problem.

# Stability of the implicit scheme

Consider the inner products

$$(U, V) := \sum_{j=1}^{J-1} \Delta x \, U_j \, V_j,$$

$$(U, V] := \sum_{j=1}^{J} \Delta x \, U_j \, V_j,$$

and the associated norms, respectively, $\| \cdot \|$ and $\| \cdot ]|$, defined by $\|U\| := (U, U)^{\frac{1}{2}}$ and $\|U]| := (U, U]^{\frac{1}{2}}$.

Note that for two mesh functions $A$ and $B$ defined on the computational mesh $\{x_j : j = 1, \dots, J-1\}$ one has that

$$(A - B, A) = \frac{1}{2}(\|A\|^2 - \|B\|^2) + \frac{1}{2}\|A - B\|^2.$$

Thus, by taking $A = U^{m+1} - U^m$ and $B = U^m - U^{m-1}$, we have

$$\begin{aligned}&(U^{m+1} - 2U^m + U^{m-1}, U^{m+1} - U^m)\\ &= \frac{1}{2}(\|U^{m+1} - U^m\|^2 - \|U^m - U^{m-1}\|^2) + \frac{1}{2}\|U^{m+1} - 2U^m + U^{m-1}\|^2.\end{aligned}$$

Similarly as above, for two mesh functions $A$ and $B$ defined on the computational mesh $\{x_j : j = 1, \dots, J\}$ we have that

$$(A - B, A] = \frac{1}{2}(\|A]|^2 - \|B]|^2) + \frac{1}{2}\|A - B]|^2.$$

Hence, by summation by parts and taking $A = D_x^- U^{m+1}$ and $B = D_x^- U^m$:

$$\begin{aligned}\left(-D_x^+ D_x^- U^{m+1}, U^{m+1} - U^m\right) &= (D_x^- U^{m+1}, D_x^- (U^{m+1} - U^m)] \\ &= (D_x^- U^{m+1} - D_x^- U^m, D_x^- U^{m+1}] \\ &= \frac{1}{2}(\| D_x^- U^{m+1}]|^2 - \| D_x^- U^m]|^2) \\ &\quad + \frac{1}{2}\| D_x^- (U^{m+1} - U^m)]|^2.\end{aligned}$$

By taking the $(\cdot, \cdot)$ inner product of $(1)_1$ with $U^{m+1} - U^m$ and using the identities stated above we therefore obtain:

$$\begin{aligned}&\frac{1}{2}\left(\left\|\frac{U^{m+1} - U^m}{\Delta t}\right\|^2 - \left\|\frac{U^m - U^{m-1}}{\Delta t}\right\|^2\right) + \frac{1}{2}\Delta t^2 \left\|\frac{U^{m+1} - 2U^m + U^{m-1}}{\Delta t^2}\right\|^2 \\ &\quad + \frac{c^2}{2}(\| D_x^- U^{m+1}]|^2 - \| D_x^- U^m]|^2) + \frac{c^2}{2}\Delta t^2 \left\| D_x^- \left(\frac{U^{m+1} - U^m}{\Delta t}\right)\right]|^2 \\ &\quad = (f(\cdot, t_{m+1}), U^{m+1} - U^m).\end{aligned} \tag{2}$$

In the special case when $f$ is identically zero the equality (2) gives

$$\left\| \frac{U^{m+1} - U^m}{\Delta t} \right\|^2 + c^2 \| D_x^- U^{m+1} ]|^2 \leq \left\| \frac{U^m - U^{m-1}}{\Delta t} \right\|^2 + c^2 \| D_x^- U^m ]|^2. \quad (3)$$

Let us define:

$$\mathcal{M}^2(U^m) := \left\| \frac{U^{m+1} - U^m}{\Delta t} \right\|^2 + c^2 \| D_x^- U^{m+1} ]|^2.$$

With this notation (3) becomes

$$\mathcal{M}^2(U^m) \leq \mathcal{M}^2(U^{m-1}), \qquad \text{for all } m = 1, \ldots, M-1,$$

and therefore

$$\mathcal{M}^2(U^m) \leq \mathcal{M}^2(U^0), \qquad \text{for all } m = 1, \ldots, M-1.$$

The mapping

$$U \mapsto \max_{m \in \{0,\dots,M-1\}} [\mathcal{M}^2(U^m)]^{1/2}$$

is a norm on the linear space of mesh functions $U$ defined on the space-time mesh $\{(x_j, t_m) : j = 0, 1, \dots, J,\ m = 0, 1, \dots, M\}$ such that $U_0^m = U_J^m = 0$ for all $m = 0, 1, \dots, M$, called the discrete energy norm.

Thus we have shown that when $f$ is identically zero the implicit scheme (1) is (unconditionally) stable in this norm.

We now return to the general case when $f$ is not identically zero. Our starting point is the equality (2). By the Cauchy–Schwarz inequality,

$$\begin{aligned}(f(\cdot, t_{m+1}), U^{m+1} - U^m) &\leq \|f(\cdot, t_{m+1})\| \, \|U^{m+1} - U^m\| \\ &= \sqrt{\Delta t \, T} \, \|f(\cdot, t_{m+1})\| \sqrt{\frac{\Delta t}{T}} \left\| \frac{U^{m+1} - U^m}{\Delta t} \right\| \\ &\leq \frac{\Delta t \, T}{2} \|f(\cdot, t_{m+1})\|^2 + \frac{\Delta t}{2T} \left\| \frac{U^{m+1} - U^m}{\Delta t} \right\|^2 ,\end{aligned} \tag{4}$$

where in the transition to the last line we used the elementary inequality

$$\alpha\beta \leq \frac{1}{2}\alpha^2 + \frac{1}{2}\beta^2, \qquad \text{for } \alpha, \beta \in \mathbb{R}.$$

Substituting (4) into (2) we deduce that

$$\begin{aligned}
&\left(1-\frac{\Delta t}{T}\right)\left(\left\|\frac{U^{m+1}-U^{m}}{\Delta t}\right\|^2 + c^2\|D_x^- U^{m+1}]|^2\right) \\
&\quad \le \left\|\frac{U^{m}-U^{m-1}}{\Delta t}\right\|^2 + c^2\|D_x^- U^{m}]|^2 + \Delta t\, T\,\|f(\cdot, t_{m+1})\|^2.
\end{aligned} \tag{5}$$

By recalling the definition of $\mathcal{M}^2(U^m)$ we can rewrite (5) in the following compact form:

$$\left(1-\frac{\Delta t}{T}\right)\mathcal{M}^2(U^m) \le \mathcal{M}^2(U^{m-1}) + \Delta t\, T\,\|f(\cdot, t_{m+1})\|^2.$$

As, by assumption, $M \geq 2$, it follows that $\Delta t := T/M \leq T/2$, whereby $\Delta t/T \leq 1/2$. By noting that

$$1 - x \geq \frac{1}{1+2x} \qquad \forall\, x \in \left[0, \tfrac{1}{2}\right],$$

it follows with $x = \Delta t/T$ that

$$\begin{aligned} \mathcal{M}^2(U^m) &\leq \left(1 + \frac{2\,\Delta t}{T}\right) \mathcal{M}^2(U^{m-1}) + \Delta t\; T \left(1 + \frac{2\,\Delta t}{T}\right) \|f(\cdot, t_{m+1})\|^2 \\ &\leq \left(1 + \frac{2\,\Delta t}{T}\right) \mathcal{M}^2(U^{m-1}) + 2\,\Delta t\; T\, \|f(\cdot, t_{m+1})\|^2. \end{aligned}$$

We need the following result, which is easily proved by induction.

**Lemma**

*Suppose that $M \geq 2$ is an integer, $\{a_m\}_{m=0}^{M-1}$ and $\{b_m\}_{m=1}^{M-1}$ are nonnegative real numbers, $\alpha > 0$, and*

$$a_m \leq \alpha\, a_{m-1} + b_m \qquad \text{for } m = 1, 2, \ldots, M-1.$$

*Then,*

$$a_m \leq \alpha^m a_0 + \sum_{k=1}^{m} \alpha^{m-k} b_k \qquad \text{for } m = 1, 2, \ldots, M-1.$$

We shall apply Lemma 1 with

$$a_m = \mathcal{M}^2(U^m), \quad b_m = 2\,\Delta t\; T\,\|f(\cdot, t_{m+1})\|^2, \quad \alpha = 1 + \frac{2\,\Delta t}{T}$$

to deduce that, for $m = 1, 2, \ldots, M-1$,

$$\mathcal{M}^2(U^m) \leq \left(1 + \frac{2\,\Delta t}{T}\right)^m \mathcal{M}(U^0) + 2\,\Delta t\; T \sum_{k=1}^{m} \left(1 + \frac{2\,\Delta t}{T}\right)^{m-k} \|f(\cdot, t^{k+1})\|^2.$$

We note that

$$\left(1 + \frac{2\,\Delta t}{T}\right)^m \leq \left(1 + \frac{2\,\Delta t}{T}\right)^M = \left(1 + \frac{2\,\Delta t}{T}\right)^{\frac{T}{\Delta t}} \leq \mathrm{e}^2,$$

where the last inequality follows from the inequality

$$(1 + 2x)^{\frac{1}{x}} \leq e^2 \quad \forall\, x \in \left(0, \tfrac{1}{2}\right],$$

with $x = \Delta t / T$.

Thus we deduce the following stability result for the implicit scheme (1).

### Theorem

*The implicit finite difference approximation* (1) *of the initial-boundary-value problem, on a finite difference mesh of spacing* $\Delta x = (b-a)/J$ *with* $J \geq 2$ *in the* $x$*-direction and* $\Delta t = T/M$ *with* $M \geq 2$ *in the* $t$*-direction, is (unconditionally) stable in the sense that, for* $m = 1, \ldots, M-1$,

$$\mathcal{M}^2(U^m) \leq \mathrm{e}^2\, \mathcal{M}^2(U^0) + 2\,\mathrm{e}^2\, T \sum_{k=1}^{m} \Delta t\, \|f(\cdot, t_{k+1})\|^2 \,, ,$$

*independently of the choice of* $\Delta x$ *and* $\Delta t$.

# Consistency of the implicit scheme

We define the consistency error of the scheme by

$$T_j^{m+1} := \frac{u_j^{m+1} - 2u_j^m + u_j^{m-1}}{\Delta t^2} - c^2 \frac{u_{j+1}^{m+1} - 2u_j^{m+1} + 2u_{j-1}^{m+1}}{\Delta x^2} - f(x_j, t_{m+1}),$$

and

$$T_j^1 := \frac{u_j^1 - u_j^0}{\Delta t} - u_1(x_j), \qquad j = 1, \ldots, J-1,$$

where $u_j^m := u(x_j, t_m)$.

By Taylor series expanions with remainder terms:

$$|T_j^{m+1}| \leq \frac{1}{12}c^2\Delta x^2 M_{4x} + \frac{5}{3}\Delta t M_{3t}, \qquad \begin{cases} j = 1, \dots, J-1, \\ m = 1, \dots, M-1, \end{cases} \tag{6}$$

where

$$M_{4x} := \max_{(x,t)\in[a,b]\times[0,T]} \left|\frac{\partial^4 u}{\partial x^4}(x,t)\right| \quad \text{and} \quad M_{3t} := \max_{(x,t)\in[a,b]\times[0,T]} \left|\frac{\partial^3 u}{\partial t^3}(x,t)\right|.$$

Furthermore, again by Taylor series expansion with a remainder term:

$$|T_j^1| \leq \frac{1}{2}\Delta t\, M_{2t}, \quad j = 1, \dots, J-1,$$

where

$$M_{2t} := \max_{(x,t)\in[a,b]\times[0,T]} \left|\frac{\partial^2 u}{\partial t^2}(x,t)\right|.$$

# Convergence of the implicit scheme

We define the *global error*

$$e_j^m := u(x_j, t_m) - U_j^m, \qquad \begin{cases} j = 0, \ldots, J, \\ m = 0, \ldots, M. \end{cases}$$

It follows from the definitions of $T_j^{m+1}$ and $T_j^1$ that

$$\frac{e_j^{m+1} - 2e_j^m + e_j^{m-1}}{\Delta t^2} - c^2 \frac{e_{j+1}^{m+1} - 2e_j^{m+1} + 2e_{j-1}^{m+1}}{\Delta x^2} = T_j^{m+1},$$

for $j = 1, \ldots, J-1$ and $m = 1, \ldots, M-1$, and

$$e_j^1 = e_j^0 + \Delta t \, T_j^1, \qquad j = 1, \ldots, J-1.$$

Furthermore, $e_j^0 = 0$ for $j = 0, 1, \ldots, J$, and $e_0^m = e_J^m = 0$ for $m = 1, \ldots, M$.

Hence, the global error $e$ satisfies an identical finite difference scheme as $U$, but with $f(x_j, t_{m+1})$ replaced by $T_j^{m+1}$, $U_j^0 = u_0(x_j)$ replaced by $e_j^0 = 0$, and $u_1(x_j)$ replaced by $T_j^1$.

Theorem 2 with $U^m$ replaced by $e^m$, $U^0$ replaced by $e^0$ and $f(x_j, t_{k+1})$ replaced by $T_j^{k+1}$ for $j = 1, \ldots, J-1$ and $k = 1, \ldots, M-1$, gives that

$$\mathcal{M}^2(e^m) \leq \mathrm{e}^2\,\mathcal{M}^2(e^0) + 2\,\mathrm{e}^2\,T \sum_{k=1}^{m} \Delta t\,\left\| T^{k+1} \right\|^2, \quad \text{for } m = 1, \ldots, M-1.$$

It remains to bound the terms on the r.h.s. of this inequality.

Because $(J-1)\Delta x \leq b-a$, it follows from (6) that

$$\max_{1\leq k\leq m} \left\| T^{k+1} \right\|^2 = \max_{1\leq k\leq m} \sum_{j=1}^{J-1} \Delta x \, | T_j^{k+1} |^2$$
$$\leq (b-a) \left[ \frac{1}{12} c^2 \Delta x^2 M_{4x} + \frac{5}{3} \Delta t M_{3t} \right]^2 .$$

On the other hand,

$$\mathcal{M}^2(e^0) = \left\| \frac{e^1 - e^0}{\Delta t} \right\|^2 + \|D_x^- e^1]|^2 = \|T^1\|^2 + \|D_x^- e^1]|^2$$
$$\leq (b-a)\left[\frac{1}{2}\Delta t\, M_{2t}\right]^2 + \|D_x^- e^1]|^2.$$

Since

$$D_x^- e_j^1 = D_x^- e_j^0 + \Delta t\, D_x^- T_j^1 = \Delta t\, D_x^- T_j^1 = \int_0^{\Delta t} (\Delta t - t)\, D_x^- \frac{\partial^2 u}{\partial t^2}(x_j, t)\,\mathrm{d}t$$
$$= \frac{1}{\Delta x}\int_0^{\Delta t}(\Delta t - t)\int_{x_{j-1}}^{x_j} \frac{\partial^3 u}{\partial x\, \partial t^2}(x,t)\,\mathrm{d}x\,\mathrm{d}t,$$

we have that

$$|D_x^- e_j^1| \leq \frac{1}{2}\Delta t^2\, M_{1x2t}, \qquad \text{where} \quad M_{1x2t} := \max_{(x,t)\in[a,b]\times[0,T]} \left|\frac{\partial^3 u}{\partial x \partial t^2}\right|,$$

whereby

$$\|D_x^- e^1]|^2 \leq (b-a)\left[\frac{1}{2}\Delta t^2 M_{1x2t}\right]^2.$$

Therefore,

$$\mathcal{M}^2(e^0) \leq (b-a)\left[\frac{1}{2}\Delta t\, M_{2t}\right]^2 + (b-a)\left[\frac{1}{2}\,\Delta t^2 M_{1x2t}\right]^2.$$

Hence, finally,

$$\begin{aligned}\mathcal{M}^2(e^m) \leq \mathrm{e}^2(b-a)\left[\frac{1}{2}\Delta t\, M_{2t}\right]^2 + \mathrm{e}^2(b-a)\left[\frac{1}{2}\,\Delta t^2 M_{1x2t}\right]^2 \\ + 2\mathrm{e}^2\, T^2(b-a)\left[\frac{1}{12}c^2\Delta x^2 M_{4x} + \frac{5}{3}\Delta t M_{3t}\right]^2\end{aligned}$$

for $m = 1, \ldots, M-1$. Thus, provided that $M_{2t}$, $M_{1x2t}$, $M_{4x}$ and $M_{3t}$ are all finite, we have that

$$\max_{m \in \{1,\ldots,M-1\}} [\mathcal{M}^2(u^m - U^m)]^{\frac{1}{2}} = \mathcal{O}(\Delta x^2 + \Delta t).$$

Summary:

The implicit scheme exhibits second order convergence with respect to the spatial discretization step $\Delta x$ and first-order convergence with respect to the temporal discretization step $\Delta t$ in the norm $\max_{m\in\{1,\dots,M-1\}}[\mathcal{M}^2(\cdot)]^{\frac{1}{2}}$.

Thanks to the unconditional stability of the implicit scheme, its convergence is also *unconditional* in the sense that there is no limitation on the size of the time step $\Delta t$ in terms of the spatial mesh-size $\Delta x$ for convergence of the sequence of numerical approximations to the solution of the wave equation to occur as $\Delta x$ and $\Delta t$ tend to 0.