Maximum Principle Analysis

1

Definition 1.1. The θ-method solves the heat equation (11.3) by

$$\begin{aligned}&\frac{U_i^{n+1}-U_i^n}{k}=\theta f(U_i^{n+1},t_{n+1})+(1-\theta)f(U_i^n,t_n)\\&=\frac{\nu}{h^2}[\theta(U_{i-1}^{n+1}-2U_i^{n+1}+U_{i+1}^{n+1})+(1-\theta)(U_{i-1}^n-2U_i^n+U_{i+1}^n)],\end{aligned}\tag{1}$$

or, equivalently

$$\begin{aligned}&-2\theta rU_{i-1}^{n+1}+(1+4\theta r)U_i^{n+1}-2\theta rU_{i+1}^{n+1}\\=&2(1-\theta)rU_{i-1}^n+[1-4(1-\theta)r]U_i^n+2(1-\theta)rU_{i+1}^n,\end{aligned}\tag{2}$$

where $r=\frac{k\nu}{2h^2}$. Here $0\le\theta\le 1$.

Example 1.2. Set $\theta=0$ in θ-method, we get FTCS method and set $\theta=\frac{1}{2}$ we get Crank-Nicolson method.

Lemma 1.3. The LTE of the θ-method is

$$\tau_i^{n+\frac{1}{2}}=O(k+h^2),\tag{3}$$

i.e., this method is second order accurate in space and first order accurate in time.

Proof. To take maximum advantage of cancellations in the Taylor expansions, expand each term about the point $(x_i,t_{n+\frac{1}{2}})$. We obtain

$$\begin{aligned}\tau_i^{n+\frac{1}{2}}:=&\frac{u(x_i,t_n+k)-u(x_i,t_n)}{k}\\&-\frac{\nu}{h^2}\theta(u(x_i-h,t_n+k)-2u(x_i,t_n+k)+u(x_i+h,t_n+k))\\&-\frac{\nu}{h^2}(1-\theta)(u(x_i-h,t_n)-2u(x_i,t_n)+u(x_i+h,t_n))\\=&[u_t-u_{xx}]+\left[\left(\frac{1}{2}-\theta\right)ku_{xxt}-\frac{1}{12}h^2u_{xxxx}\right]\\&+\left[\frac{1}{24}k^2u_{ttt}-\frac{1}{8}k^2u_{xxtt}\right]+\left[\frac{1}{12}\left(\frac{1}{2}-\theta\right)kh^2u_{xxxxt}-\frac{2}{6!}h^4u_{xxxxxx}\right]+\cdots,\end{aligned}\tag{4}$$

Notice the first term always cancels, hence we complete the proof. □

Example 1.4. Set $\theta = \frac{1}{2}$, we see that the Crank–Nicolson method is always second order accurate in both space and time.

Example 1.5. Set $\theta = \frac{1}{2} - \frac{h^2}{12k}$, we get a method whose LTE is $O(k^2 + h^4)$. Note that this requires $h^2 \leq 6k$ to ensure $\theta \geq 0$. In fact, this method is also absolutely stable. You can get it by verifying following (5) is satisfied.

Lemma 1.6. For the θ-method (1) to be absolutely stable, the following conditions must be satisfied:

$$\begin{cases} r \leq \frac{1}{4(1-2\theta)}, & 0 \leq \theta < \frac{1}{2}, \\ r > 0, & \frac{1}{2} \leq \theta \leq 1, \end{cases} \tag{5}$$

which imply the time-step limit $k \leq \frac{h^2}{2(1-\theta)\nu}$ when $0 \leq \theta < \frac{1}{2}$, and unconditionally stability when $\frac{1}{2} \leq \theta \leq 1$.

Exercise 1.7. Prove Lemma 1.6.

Proof. We prove Lemma 1.6 by Von Neumann analysis. Set

$$U_j^n = [g(\xi)]^n e^{ix_j\xi}, \tag{6}$$

then $U_j^{n+1} = g(\xi)U_j^n$ and (1) give

$$\begin{aligned} g(\xi) - 1 &= 2r[\theta g(\xi) + (1-\theta)](e^{-i\xi h} - 2 + e^{i\xi h}) \\ &= 2r[\theta g(\xi) + (1-\theta)]\left(-4\sin^2\left(\frac{\xi h}{2}\right)\right) \end{aligned} \tag{7}$$

i.e.,

$$g(\xi) = \frac{1 - 8(1-\theta)r\sin^2\left(\frac{\xi h}{2}\right)}{1 + 8\theta r\sin^2\left(\frac{\xi h}{2}\right)}. \tag{8}$$

Because $r > 0$, and we are assuming that $0 \leq \theta \leq 1$, it is clear that we can never have $g(\xi) > 1$. Thus instability arises only through the possibility that $g(\xi) < -1$, which implies

$$1 - 8(1-\theta)r\sin^2\left(\frac{\xi h}{2}\right) < -\left[1 + 8\theta r\sin^2\left(\frac{\xi h}{2}\right)\right] \tag{9}$$

i.e.,

$$8r(1-2\theta)\sin^2\left(\frac{\xi h}{2}\right) > 2. \tag{10}$$

The mode most liable to instability is the one for which the left side is largest: $\xi h = \pi$. This is an unstable mode if

$$r(1-2\theta) > \frac{1}{4}. \tag{11}$$

Notice that if $\theta \geq \frac{1}{2}$, (11) will never hold. □

Theorem 1.8 (Maximum principle of the heat equation)**.** If $u(x,t)$ is continuous on rectangle $\bar{\Omega} = [0,1] \times [0,T]$, and satisfies the heat equation (11.3) in $\Omega = (0,1) \times (0,T)$, then the maximum and minimum value of $u(x,t)$ over the rectangle is assumed either initially ($t = 0$), or on the lateral sides ($x = 0$, or $x = 1$).

Example 1.9. If we denote the set of points comprising the three sides by $\Gamma = \{(x,t) \in \bar{\Omega} \mid t = 0 \text{ or } x = 0 \text{ or } x = 1\}$, then the maximum principle can be written as

$$\begin{aligned} \max_{(x,t)\in\bar{\Omega}} \{u(x,t)\} &= \max_{(x,t)\in\Gamma} \{u(x,t)\}, \\ \min_{(x,t)\in\bar{\Omega}} \{u(x,t)\} &= \min_{(x,t)\in\Gamma} \{u(x,t)\}. \end{aligned} \tag{12}$$

Theorem 1.10 (Discrete maximum principle)**.** The θ-method of (1) with $0 \leq \theta \leq 1$ and $r \leq \frac{1}{4(1-\theta)}$ yields U_j^n satisfying

$$U_{min} \leq U_j^n \leq U_{max}, \tag{13}$$

where

$$U_{min} := \min \{U_0^s,\ 0 \leq s \leq n;\ U_j^0,\ 0 \leq j \leq m+1;\ U_{m+1}^s,\ 0 \leq s \leq n\}, \tag{14}$$

and

$$U_{max} := \max \{U_0^s,\ 0 \leq s \leq n;\ U_j^0,\ 0 \leq j \leq m+1;\ U_{m+1}^s,\ 0 \leq s \leq n\}. \tag{15}$$

Proof. We write (2) in the form

$$\begin{aligned} (1+4\theta r)U_j^{n+1} = {} & 2\theta r(U_{j-1}^{n+1} + U_{j+1}^{n+1}) + 2(1-\theta)r(U_{j-1}^n + U_{j+1}^n) \\ & + [1 - 4(1-\theta)r]U_j^n, \end{aligned} \tag{16}$$

Then under the hypotheses of the theorem all the coefficients on the right are nonnegative and sum to $(1+4\theta r)$. Now suppose that U attains its maximum at an internal point, and

this maximum is U_j^{n+1}, and let U^* be the greatest of the five values of U appearing on the right-hand side of (16). Then since the coefficients are nonnegative,

$$\begin{aligned}(1+4\theta r)U_j^{n+1} \le\ & 2\theta r(U^*+U^*)+2(1-\theta)r(U^*+U^*)\\ &+[1-4(1-\theta)r]U^*=(1+4\theta r)U^*,\end{aligned} \tag{17}$$

i.e., $U_j^{n+1} \le U^*$. But since U_j^{n+1} is assumed to be the maximum value, we also have $U_j^{n+1} \ge U^*$, so $U_j^{n+1} = U^*$. Indeed, the maximum value must also be attained at each neighbouring point which has a non-zero coefficient in (16). The same argument can then be applied at each of these points, showing that the maximum is attained at a sequence of points, until a boundary point is reached. The maximum is therefore attained at a boundary point. An identical argument shows that the minimum is also attained at a boundary point, and the proof is complete. □

Theorem 1.11. Asuume the stable condition (5) holds, then the θ-method (1) is convergent with consistent initial and Dirichlet boundary data.

Proof. Firstly we assume that numerical errors arise from the truncation errors of the finite difference approximations, but that the boundary values are used exactly. This part is the same as sufficiency of theorem 11.23.
Apply the θ-method (2) to the exact solution $u(x_i, t_n)$ and we obtain

$$\begin{aligned}(1+4\theta r)u(x_i,t_{n+1}) =\ & 2\theta r(u(x_{i-1},t_{n+1})+u(x_{i+1},t_{n+1}))+2(1-\theta)r(u(x_{i-1},t_n)+u(x_{i+1},t_n))\\ &+[1-4(1-\theta)r]u(x_i,t_n)+k\tau_i^{n+\frac{1}{2}},\end{aligned} \tag{18}$$

Subtracting (18) from (16), the global error $E_i^n = U_i^n - u(x_i, t_n)$ is determined from the relations

$$\begin{aligned}(1+4\theta r)E_i^{n+1} =\ & 2\theta r(E_{i-1}^{n+1}+E_{i+1}^{n+1})+2(1-\theta)r(E_{i-1}^n+E_{i+1}^n)\\ &+[1-4(1-\theta)r]E_i^n-k\tau_i^{n+\frac{1}{2}}\end{aligned} \tag{19}$$

for $i = 1, 2, \cdots, m$ and $n = 0, 1, \cdots$ together with initial and boundary conditions. By assumption, E_i^0, E_0^n and E_{m+1}^n are all zero with $i = 0, 1, \cdots, m+1$ and $n = 0, 1, \cdots$. Then we define

$$E^n := \max_{0\le i\le m+1}|E_i^n|, \quad \tau^{n+\frac{1}{2}} := \max_{1\le i\le m}\left|\tau_i^{n+\frac{1}{2}}\right|. \tag{20}$$

Because of the nonnegative coefficients, it follows that

$$(1+4\theta r)E^{n+1} \le\ 4\theta r E^{n+1}+E^n+k\tau^{n+\frac{1}{2}} \tag{21}$$

and hence that

$$E^{n+1} \leq \ E^n + k\tau^{n+\frac{1}{2}}. \tag{22}$$

Since $E^0 = 0$,

$$E^n \leq k \sum_{0}^{n-1} \tau^{s+\frac{1}{2}} \leq nk \max_s \tau^{s+\frac{1}{2}}, \tag{23}$$

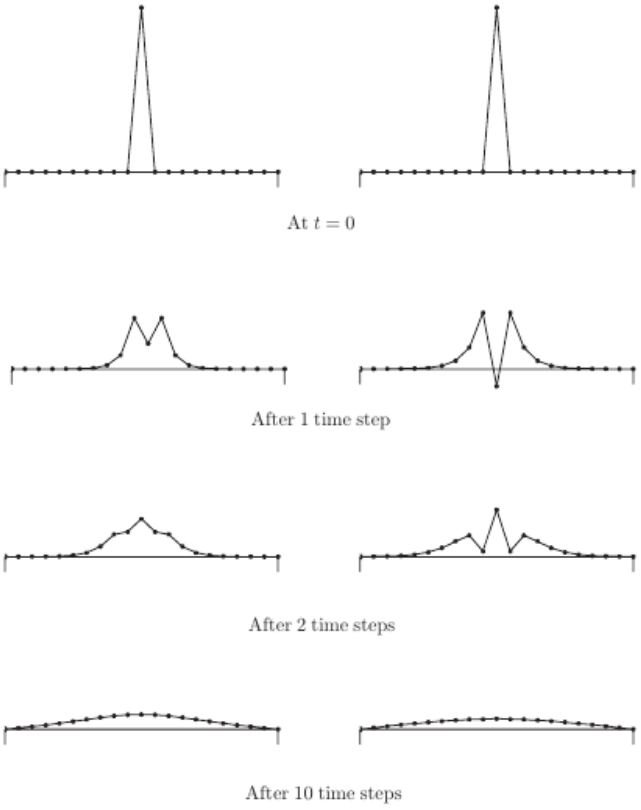
and this tends to zero following from Lemma 1.3. Suppose now that there are errors in the initial and boundary values of U_i^n and denote them by ϵ_i^0, ϵ_0^n and ϵ_{m+1}^n with $i = 0, 1, \cdots m+1$ and $n = 0, 1, \cdots$. Then the errors E_i^n satisfy the recurrence relation (19) with initial and boundary values

$$\begin{aligned} E_i^0 &= \epsilon_i^0, \quad i = 0, 1, \cdots m+1, \\ E_0^n &= \epsilon_0^n, \ E_{m+1}^n = \epsilon_{m+1}^n, \quad n = 0, 1, \cdots . \end{aligned} \tag{24}$$

By Duhamel's principle, E_i^n can be written as the sum of two terms. The first term satisfies (19) with zero initial and boundary values; this term is bounded by (23). The second term satisfies the homogeneous form of (19), with the term of τ omitted, and with the given non-zero initial and boundary values. By the discrete maximum principle this term must lie between the maximum and minimum values of these initial and boundary values. Thus the error of the numerical solution will tend to zero, as required, provided that the initial and boundary values are consistent; that is, the errors in the initial and boundary values also tend to zero as k, h tend to zero. □

Example 1.12. The condition for discrete maximum principle, $r(1-\theta) \leq \frac{1}{4}$, is very much more restrictive than that needed in stability, $r(1-2\theta) \leq \frac{1}{4}$. For example, the Crank–Nicolson method always satisfies the stability condition, but only if $r \ \leq \frac{1}{2}$ does it satisfy the condition given for a discrete maximum principle.

Consider the model problem solved by the Crank–Nicolson method. The boundary conditions specify that the solution is zero at each end of the range, and the initial condition gives the values of U_i^0 to be zero except at the mid-point; the value at the mid-point is unity. This corresponds to a function with a sharp spike at $x = \frac{1}{2}$.



The Crank–Nicolson method applied to the heat equation where the initial distribution has a sharp spike at the mid-point; $m = 19, h = 0.05$; left: $r = \frac{1}{2}$, right: $r = 1$.

In the case $r = 1$ the discrete maximum principle does not hold, and we see that at the first time level the numerical solution becomes negative at the mid-point. This would normally be regarded as unacceptable. When $r = \frac{1}{2}$ the discrete maximum principle holds, and the numerical values all lie between 0 and 1, as required. However, at the first time level the numerical solution shows two peaks, one each side of the mid-point; the exact solution of the problem will have only a single maximum for all t.

These results correspond to a rather extreme case, and the unacceptable behaviour only persists for a few time steps; thereafter the solution becomes very smooth in each case. However, they show that in a situation where we require to model some sort of rapid variation in the solution we shall need to use a value of r somewhat smaller than the stability limit.