SOME REMARKS ON NUMERICAL METHODS FOR NONLINEAR HEAT EQUATIONS WITH NEAR SINGULAR SPECIFIC HEATS

Anthony Miller

Consider the one-dimensional nonlinear heat conduction equation

$$(1)\qquad \begin{cases} c(u)\,\dfrac{\partial u}{\partial t} = \dfrac{\partial}{\partial x}\left(k(u)\,\dfrac{\partial u}{\partial x}\right) + q(x,t)\ , & x \in (0,1),\ t > 0 \\ u(0,t) = u(1,t) = 0\ , & t > 0 \\ u(x,0) = u_0(x)\ , & x \in (0,1)\ , \end{cases}$$

where $c(\cdot)$ and $k(\cdot)$ are continuous functions on $\mathbb{R}$ satisfying

$$(2a)\qquad \exists c^* > 0 \text{ such that } c(u) \geq c^* > 0 \quad \forall u \in \mathbb{R}$$

$$(2b)\qquad k(u) > 0 \quad \forall u \in \mathbb{R}\ .$$

Of particular interest are cases where $c(u)$ varies greatly over a small temperature range. Such behaviour can arise in simple models of phase changes in alloys ("near Stefan problems"). It would be desirable to have a numerical method for approximating the solution of (1) whose accuracy was in some sense independent of the behaviour of the coefficients $c(\cdot)$ and $k(\cdot)$. It is however unreasonable to expect this much since the accuracy of any approximation will clearly be influenced by the regularity of the exact solution. This regularity can vary considerably, depending on the coefficient $c(\cdot)$, $k(\cdot)$ as well as the initial temperature data u_0 and the source data q. A more reasonable request would be that the stability properties of a numerical method be uniform for all $c(\cdot)$ and $k(\cdot)$ satisfying (2). That is, we would like to be able to assert something of the form:

$$\sup \frac{\|e\|}{|||u|||} < M < \infty ,$$

where $\|e\|$ denotes some norm of the error e of the approximate solution, $|||u|||$ is some measure of the approximability (i.e. regularity) of the exact solution u, and the supremum is taken over all coefficients $c(\cdot)$ and $k(\cdot)$ satisfying (2). This is the matter we wish to discuss briefly here.

For simplicity of exposition we shall only talk in terms of finite difference schemes for (1) with uniform spatial and time mesh spacings $\Delta x = \frac{1}{N}$, Δt respectively. Most of what we shall say extends to more general settings with minor modifications.

Typical of the standard discretization methods for nonlinear parabolic equations that one may think of using is the classical (fully) implicit method. Applied to (1), and assuming that $k(\cdot)$ is a constant, it gives

$$(3) \quad \begin{cases} c\left(u_i^n\right) \left(\dfrac{u_i^{n+1} - u_i^n}{\Delta t}\right) = k\, \delta^2 u_i^{n+1} + q_i^n , & (i = 1,\ldots,N-1;\ n = 0,1,\ldots) \\ u_0^n = u_N^n = 0 , & (n = 0,1,\ldots) \\ u_i^0 = u_0\left(\dfrac{i}{N}\right) , & (i = 1,\ldots,N-1) , \end{cases}$$

where

$$\left.\begin{aligned} \delta^2 u_i^n &= \frac{1}{(\Delta x)^2}\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right) \\ q_i^n &= q\left(\frac{i}{N} , n\Delta t\right) . \end{aligned}\right. \qquad (i = 1,\ldots,N-1;\ n = 0,1,\ldots)$$

The standard stability analysis of the difference scheme (3) requires some assumption that controls the slope of $c(\cdot)$ (e.g. Lipschitz continuity). Counterexamples show that in the absence of any such assumption the stability of the scheme can degenerate as more extreme choices for $c(\cdot)$ are made. Intuitively, at least one reason for this is clear: any perturbation in u^n say, may cause perturbations in the $c\left(u_i^n\right)$, and consequently in u^{n+1}, which are unable to be uniformly controlled by the original perturbation in u^n .

Rather than base the discretization directly on the formulation (1), introduce a new dependent variable $h = h(u)$ defined by

$$h(u) = \int_0^u c(s)\,ds \ .$$

The problem (1) may now be reformulated as

$$(4) \qquad \begin{cases} \dfrac{\partial h}{\partial t} = \dfrac{\partial}{\partial x}\left(k(u)\dfrac{\partial u}{\partial x}\right) + q(x,t) \ , & x \in (0,1),\ t > 0 \\ u(0,t) = u(1,t) = 0 \ , & t > 0 \\ h(x,0) = h(u_0(x)) = h_0(x) \quad \text{say} \ , & x \in (0,1) \ , \end{cases}$$

where

$$h(x,t) = h(u(x,t)) \ .$$

Physically h represents the specific (volumetric) enthalpy. Notice that by (2a) $h(u)$ is strictly increasing, and so

$$(5) \qquad (h(u_1) - h(u_2))(u_1 - u_2) \geq 0 \quad \forall u_1,u_2 \in \mathbb{R} \ .$$

There is no loss of generality in supposing $k(u) = 1$ in (4). For if not we may define the Kirchhoff temperature,

$$\theta(u) = \int_0^u k(s)\,ds$$

and rewrite (4) as

$$\begin{cases} \dfrac{\partial h}{\partial t} = \dfrac{\partial^2\theta}{\partial x^2} = q(x,t) \ , & x \in (0,1),\ t > 0 \\ \theta(0,t) = \theta(1,t) = 0 \ , & t > 0 \\ h(x,0) = h_0(x) \ , & x \in (0,1) \ , \end{cases}$$

where $\theta(x,t) = \theta(u(x,t))$. Moreover, since $\theta(u)$ is strictly increasing by (2b), h regarded as a function of θ is also strictly increasing. Thus the problem (4) with arbitrary k , if thought of in terms of h and

θ , is of precisely the same form as (4), (5) with $k = 1$. We shall from now on only consider this case.

We may think of applying obvious generalizations of the standard discretization methods to (4). Two cases will be considered:

Fully Implicit (FI) method:

$$(6) \qquad \frac{h_i^{n+1} - h_i^n}{\Delta t} = \delta^2 u_i^{n+1} + q_i^n \, , \qquad (i = 1,\dots,N-1; \; n = 0,1,\dots) \, .$$

Crank-Nicolson (C-N) method:

$$(7) \qquad \frac{h_i^{n+1} - h_i^n}{\Delta t} = \tfrac{1}{2}\left(\delta^2 u_i^{n+1} + \delta^2 u_i^n\right) + \tfrac{1}{2}\left(q_i^{n+1} + q_i^n\right) \, , \qquad (i = 1,\dots,N-1; \; n = 0,1,\dots)$$

with in both cases

$$h_i^0 = h_0\left(\frac{i}{N}\right) \, , \qquad (i = 1,\dots,N-1)$$

$$u_0^n = u_N^n = 0 \, , \qquad (n = 0,1,\dots)$$

$$h_i^n = h\left(u_i^n\right) \, , \qquad (i = 1,\dots,N-1; \; n = 0,1,\dots) \, .$$

We wish to examine the stability properties of FI and C-N. However, let us first mention that for both methods, various norms of the discrete solutions u_i^n , h_i^n can be bounded independently of the discretization parameters Δx , Δt . Using standard compactness arguments, it can then be shown that the discrete solutions converge in some sense (see e.g. [2], [3], [4]). However such arguments only establish convergence in rather weak norms. Moreover they do not provide any form of estimate for the error in the discrete solution. This makes it difficult to develop any theoretical understanding of the equality of the methods.

We show that for the FI method h and u are stable in a discrete L_1

sense. More specifically we have

THEOREM: *Let* h_i^n , u_i^n *and* $\tilde{h}_i^n$, $\tilde{u}_i^n$ *be two solutions of* (6) *corresponding to initial and source data* h_i^0 , q_i^n *and* $\tilde{h}_i^0$, $\tilde{q}_i^n$ *respectively* $(i = 0,\dots,N;\ n = 0,1,\dots)$, *then*

$$\sum_{i=0}^{N} \left(|u_i^n - u_i^n| + |\tilde{h}_i^n - \tilde{h}_i^n|\right) \le c\left(\sum_{i=0}^{N} |h_i^0 - \tilde{h}_i^0| + \Delta t \sum_{m=1}^{n-1} \sum_{i=0}^{N} |q_i^m - \tilde{q}_i^m|\right) \qquad n = 0,1,\dots$$

where the constant c *is independent of* Δx , Δt *and can be selected uniformly for all* $c(\cdot)$ *satisfying* (2a).

Proof: Write

$$H_i^n = h_i^n - \tilde{h}_i^n , \qquad U_i^n = u_i^n - \tilde{u}_i^n \quad \text{and} \quad Q_i^n = q_i^n - \tilde{q}_i^n .$$

Subtracting the respective cases of (6) gives

$$H_i^{n+1} - H_i^n = \Delta t\ \delta^2\ U_i^{n+1} + \Delta t\ Q_i^n .$$

Multiply this equation by

$$\operatorname{sgn} U_i^{n+1} = \begin{cases} 1 & \text{if } U_i^{n+1} > 0 \\ 0 & \text{if } U_i^{n+1} \\ -1 & \text{if } U_i^{n+1} < 0 \end{cases}$$

and sum over $i = 0,\dots,N$ to obtain

$$\text{(8)}\quad \sum_i H_i^{n+1} \operatorname{sgn} U_i^{n+1} = \sum_i H_i^n \operatorname{sgn} U_i^{n+1} + \sum_i \Delta t\ \delta^2 U_i^{n+1} \operatorname{sgn} U_i^{n+1} + \sum_i \Delta t\, Q_i^n \operatorname{sgn} U_i^{n+1}$$
$$= \Sigma^{(1)} + \Sigma^{(2)} + \Sigma^{(3)} \quad \text{say} .$$

From (5) it follows that

$$H_i^{n+1}\ U_i^{n+1} = \left(h\left(u_i^{n+1}\right) - h\left(\tilde{u}_i^{n+1}\right)\right)\left(u_i^{n+1} - \tilde{u}_i^{n+1}\right) \ge 0$$

and so the left hand side of (8) becomes

$$\sum_i H_i^{n+1} \operatorname{sgn} U_i^{n+1} = \sum_i |H_i^{n+1}| ,$$

On the other hand

$$|\Sigma^{(1)}| \le \sum_i |H_i^n|$$

$$|\Sigma^{(3)}| \le \Delta t \sum_i |Q_i^n|$$

while after a summation by parts (noting that $U_0^{n+1} = U_N^{n+1} = 0)$,

$$\Sigma^{(2)} = -\frac{\Delta t}{(\Delta x)^2} \sum_{i=0}^{N-1} \left(U_{i+1}^{n+1} - U_i^{n+1}\right)\left(\operatorname{sgn} U_{i+1}^{n+1} - \operatorname{sgn} U_i^{n+1}\right) \le 0 .$$

Thus (8) gives

$$\sum_i |H_i^{n+1}| \le \sum_i |H_i^n| + \Delta t \sum_i |Q_i^n| .$$

Iterating this result back to $n = 0$ gives the theorem, having noted that by (2a)

$$|U_i^{n+1}| \le \frac{1}{c^*} |H_i^{n+1}| . \qquad //$$

The natural question to ask now is whether a similar stability result holds for C-N. This question is of some practical interest since, by analogy with the case of the linear heat equation say, if some reasonable form of stability holds for C-N then C-N can be expected to be markedly more accurate than FI. However the following simple counterexample shows an estimate analogous to that of the theorem cannot hold for C-N.

Suppose the enthalpy-temperature relation is given by

$$h(u) = \begin{cases} u & \text{if } u \ge 0 \\ \beta u & \text{if } u \le 0 , \end{cases}$$

where $\beta > 0$ is a constant. Take $\Delta x = .5$ (i.e. $N = 2$) and consider solutions of (7) corresponding to data

$$\tilde{h}_i^0 = 0 , \quad \tilde{q}_i^n = 0 \qquad (i = 0,1,2; \; n = 0,1,\ldots)$$

and $$h_i^0 = \varepsilon , \quad q_i^n = 0 .$$

Obviously $\tilde{h}_i^n = 0$ $(i = 0,1,2;\ n = 0,1,\ldots)$, whereas provided

$$\alpha = \frac{\Delta t}{(\Delta x)^2} > 1$$

is satisfied, then

$$h_1^1 = \varepsilon\left(\frac{1-\alpha}{1+\frac{\alpha}{\beta}}\right) .$$

Clearly an estimate of the form $|h_1^1| \le C\varepsilon$ cannot hold with C independent of α and β . In other words, the estimate of the theorem cannot apply in the C-N case with C independent of Δx , Δt and uniform for all $c(\cdot)$ satisfying (2a).

Let us remark however that if we impose the extra condition

$$\frac{1}{c^*}\frac{\Delta t}{(\Delta x)^2} \le 1 \tag{9}$$

then the theorem can be shown to hold for C-N. Note that this condition was specifically violated in the above counterexample.

The above counterexample does not preclude the possibility of a slightly weaker form of stability than that of the theorem holding for C-N (e.g. discrete L_1 stability for just u). However the following numerical example suggests that this is unlikely.

Consider the classical Neumann solution of the Stefan problem in the one-dimensional half space $x > 0$ (see [1]). Concentrate on the interval (1,2), and use the exact solution to obtain boundary values at the end-points of this interval and the initial value when the moving phase boundary passes the left hand endpoint of the interval. Transforming (1,2) to a (0,1) and choosing a new time origin we can obtain a formulation for the problem of the form of (4). More specifically the exact solution is given by

$$u(x,t) = \begin{cases} \frac{1}{\operatorname{erf}\lambda}\operatorname{erf}(\xi) & 0 < \xi < \lambda \\ 2 - \frac{1}{\operatorname{erfc}\lambda}\operatorname{erfc}(\xi) & \lambda < \xi \end{cases}$$

where $\lambda = 0.05$ and $\xi = \dfrac{x+1}{2(t+100)^{\frac{1}{2}}}$ $(x \in (0,1),\ t > 0)$. Erf(•) and erfc(•) are the usual error and complementary error functions respectively. This solution satisfies a slightly modified form of (4) with $k = 1$, $q = 0$ and enthalpy-temperature relation given by

$$h(u) = \begin{cases} u & u < 1 \\ [1,\ 188.738] & u = 1 \\ 188.738 + (u-1) & u > 1 \ . \end{cases}$$

Note that here u has non-homogeneous Dirichlet boundary conditions at $x = 0,1$, and that $h(u)$ is set valued at $u = 1$. These modifications however introduce no major new features to the problem and our previous discussion is unaffected by them.

This problem was solved for two sets of mesh parameters, $\Delta x = .05$, $\Delta t = 10$ and $\Delta x = .025$, $\Delta t = 5$. The pointwise errors in the numerical solution at times $t = 100$ and $t = 250$ at selected nodal points are shown in Table I.

A comparison of the errors for the two methods reasonably suggests that FI is more reliable than C-N. While the errors in the FI solution increase steadily towards the phase transition (this occurs at $x = .414$ for $t = 100$ and $x = .87$ for $t = 250$), those in the C-N solution oscillate with relatively large amplitudes. Decreasing both Δx and Δt does not seem to effect the comparative behaviour of the two methods.

The inferior performance of C-N in this example seems to suggest that the stability properties of C-N are appreciably weaker than those of FI, since presumably there is little significant difference between the discretization errors made in the FI and C-N equations (6) and (7). Indeed, if anything, the expectation would be that the discretization error in the C-N equation (7) would be less than that in the FI equation, at least in the limit as $\Delta x, \Delta t \to 0$. Note incidentally that the quantity $\frac{1}{c^*} \frac{\Delta t}{(\Delta x)^2}$ of (9) takes the values 4×10^3 and 8×10^3 for the meshes considered here.

TABLE I: Errors in Numerical Solution of Test Problem.

$\Delta x = .05,\ \Delta t = 10$			$\Delta x = .025,\ \Delta t = 5$		
t = 100			t = 100		
x	FI	CN	x	FI	CN
.1	6 (E-4)*	-105	.1	-1	-57
.2	14	-280	.2	-1	-36
.3	18	-107	.3	-2	35
.4	27	100	.4	-3	56
t = 250			t = 250		
x	FI	CN	x	FI	CN
.1	3	86	.1	-2	-66
.2	7	247	.2	-3	-57
.3	11	63	.3	-4	10
.4	14	-155	.4	-7	34
.5	18	125	.5	-10	-8
.6	22	-215	.6	-13	90
.7	26	48	.7	-16	133
.8	29	202	.8	-19	2

[*: All quantities in the FI and CN columns should be multiplied by this factor.]

REFERENCES

[1] H.S. Carslaw and J.C. Jaeger, *Conduction of Heat in Solids*, Oxford University Press, London, 1959.

[2] J.F. Ciavaldini, 'Analyse numérique d'un problème de Stefan à deux phases par une méthode d'éléments finis', *SIAM J. Num. Anal.* <u>12</u> (1975), 464-487.

[3] C.M. Elliot and J.R. Ockendon, *Weak and Variational Methods for Moving Boundary Problems*, Pitman Research Notes in Mathematics No. 59.

[4] R.E. White, 'An enthalpy formulation of the Stefan problem', *SIAM J. Num. Anal.* 19 (1982), 1129-1157.