A Modified Finite Difference Scheme
for the Stefan Problem

By R. E. White

Abstract. In this paper we describe an algorithm which can be used to approximate the solution to the enthalpy formulation of the Stefan problem. We allow the thermal properties to have a space and temperature dependence. The algorithm is not explicit in the time variable and, hence, the stability condition on $\Delta t$ is not too severe. A proof of convergence is given and two numerical examples are presented.

1. Introduction. In this paper we describe an algorithm that may be used to approximate the solution of

$$E + \Delta t A(E) \beta(E) = \eta > 0,$$

where $E, \beta(E) = (\beta_j(E_j)), \eta \in \mathbb{R}^L, L =$ number of nodes and $A(E) = (a_{ij}(E))$ is an $L \times L$ matrix. The matrix is associated with an elliptic boundary value problem. This algebraic problem evolves from an implicit time discretization of the enthalpy formulation of the Stefan problem. In (1) $E =$ enthalpy and $\beta(E) =$ temperature. As the thermal conductivity will be temperature-dependent, the components of $A(E)$ must depend on the temperature and, hence, the enthalpy.

Originally J. Stefan [11] formulated the problem for the solidification of water. Since then there have been many other applications. D. R. Atthey [1] studied the welding problem in which an explicit time discretization is used. A stability condition on $\Delta t$ was developed. N. Shamsundar and E. M. Sparrow [7] study thermal energy storage units which utilize phase change materials. As this process is over a much longer duration, they use an implicit time discretization. J. A. Wheeler [12] simulates the behavior of permafrost adjacent to the Alaskan pipeline. Here the time interval is very large as compared to the welding problem. Recently, A. D. Solomon [10] discussed simulations of cryosurgery. In this report it is noted that the thermal properties vary with temperature, position and types of tissues. An explicit time discretization of the enthalpy formulation is used. Also, the reader may wish to consult the following texts for additional applications: L. I. Rubinstein [6], J. Ockendon and W. Hodgkins [3], and D. G. Wilson, A. Solomon and P. T. Boggs [15].

Before describing the algorithm let us review the enthalpy formulation. In the classical heat equation the principal unknown is the temperature. In M. Rose [5] the enthalpy is the principal unknown. Consider the enthalpy function, $H$, and its "inverse", $\beta$, as illustrated in Figure 1. In the graphs of $H$ and $\beta$ the specific heat, $c,$
and density, $\rho$, are assumed constant in the solid, $s$, and liquid, $l$, phases. $\rho L$ = latent heat/volume. The thermal conductivity may be viewed as a function of temperature, $\beta(E)$, or as a function of enthalpy, $E$. Figure 2 illustrates the thermal conductivity $K$ when it is constant in the solid and liquid phases.

The enthalpy formulation of the Stefan problem is (2.1)–(2.4) where $d/d\nu = \text{conormal derivative}$.

\begin{align*}
(2.1) \quad & E_t - \nabla \cdot K \nabla \beta(x, E) = f(x, t, E) \quad \text{on } \Omega \times (0, T), \\
(2.2) \quad & \beta(x, E) = g_1(x, t) \quad \text{on } \Gamma_1 \times (0, T), \Gamma_1 \cup \Gamma_2 = \partial \Omega, \\
(2.3) \quad & \left( \frac{d}{d\nu} + \sigma(x, t) \right) \beta(x, E) = g_2(x, t) \quad \text{on } \Gamma_2 \times (0, T), \Gamma_1 \cap \Gamma_2 = \emptyset, \\
(2.4) \quad & E(x, 0) = \Psi(x) \quad \text{on } \Omega.
\end{align*}
An implicit time discretization of (2.1) yields

\[ E - \Delta t \nabla \cdot K \nabla \beta(x, E) = \Delta tf(x, E) + \bar{E}, \]

where the superscripts for the time have been deleted and \( \bar{E} \) is the enthalpy at the previous time step. Equation (1) evolves from Eq. (3), when finite differences are used. \( A(E) \) is the matrix associated with the elliptic operator \( -\nabla \cdot K \nabla \beta(E) \) and the boundary conditions (2.2) and (2.3).

Of course an explicit time discretization avoids the system (1). It was first used by A. Solomon [8] and later in [1], [10] and others. In order to avoid the stability criterion, G. H. Meyer [2] smoothed \( H(u) \) to \( H_s(u) \) and used the implicit time
discretization to solve for the temperature. The smoothing of $H(u)$ was necessary because he utilized the standard Gauss-Seidel iterative scheme, for which continuity of $H(u)$ is necessary, to solve $Au = -H(u) + \Delta t f + H(\bar{u})$. N. Shamsundar and E. M. Sparrow [7] used a different iterative scheme to solve (1) where $A(E)$ is independent of $E$. The convergence proof of this scheme is presented in R. E. White [13]. A. D. Solomon [9] modifies (3) with $\beta(x, E) = \beta(E)$ by writing $\nabla \beta(E) = \beta'(E) \nabla E$. This generates an equation similar to (1) where $A(E)$ is slightly different and $\beta(E)$ is replaced by $E$. The algorithm in this paper is similar to Solomon’s algorithm.

The advantages of the implicit time discretization are (i) it avoids the stability condition on $\Delta t$ and (ii) it allows the Crank-Nicolson scheme, which is considered to be more accurate, to be used. Hence, the algorithm given in line (4) can be used on the algebraic problem (1) which results from the Crank-Nicolson scheme, and from problems with temperature and space dependent thermal properties.

In Section 2 we define the algorithm and give a convergence proof. Section 3 contains two numerical examples.

2. The Algorithm. Consider Eq. (1).

Definition. Let $\tilde{A}(E) \equiv (a_{ij}(E)\beta_j(E_j)/E_j)$ and $\delta(E) \equiv I + \Delta t \tilde{A}(E)$.

$$E^{k+1} \equiv \delta(E^k)^{-1} \eta$$ and $E^0 = \bar{E}$.

We shall show that $E^{k+1} \to E \in +RL$, $+R \equiv (0, \infty)$, and $E$ is a solution to (1). This is done by a simple application of the contraction mapping theorem to $G E \equiv \delta(E)^{-1} \eta$. We will need the following lemma which is proved in J. M. Ortega and W. C. Rheinboldt [4, p. 54].

Comparison Lemma. Let $A^1 = D - Bx$ be an $M$-matrix. If $D_2 > 0$, $B_2 > 0$ and $B_2 < B_1$, then $A \equiv D_1 + D_2 - (B_1 - B_2)$ is an $M$-matrix and $0 < A^{-1} \leq A_1^{-1}$.

Assumptions.

1. $\beta_j: +R \to +R$ satisfy
   (a) $0 < \beta_j(x) < Cx$, $x > 0$.
   (b) $\beta_j$ are Lipschitz continuous on $+R$.
2. $A(E) = (a_{ij}(E))$ and $a_{ij}: +RL \to R$ satisfy
   (a) $\sum_{j \neq i} |a_{ij}(E)| \leq |a_{ii}(E)|$ for all $E \in +RL$.
   (b) $|a_{ii}(E)| \leq M < \infty$ for all $E \in +RL$.
   (c) $a_{ij}(E) \leq 0$ for all $i \neq j$, and $a_{ii}(E) > 0$.
   (d) $|a_{ij}(E) - a_{ij}(\bar{E})| \leq m_{ij}||E - \bar{E}||_{\infty}$, where $\Sigma_j m_{ij} = \Sigma_{j \neq i} m_{ij} \leq m < \infty$.

Theorem. Let $E^{k+1}$ be given by (4) with $\eta$, $E^0 \in +RL$. If assumptions 1 and 2 hold, then there exists a $\delta > 0$ such that if $\Delta t \leq \delta$, then $E^{k+1} \to E \in +RL$ and $E$ is the unique solution of (1).

Proof. The proof is an application of the contraction mapping theorem. We shall show that $G(E) \equiv \delta(E)^{-1} \eta$, where $\delta(E) \equiv I + \Delta t \tilde{A}(E)$ is contractive. Note that (4) is just $E^{k+1} = G(E^k)$ and $E = G(E)$ satisfies (1). First, we demonstrate that for small enough $\Delta t$, $\delta(E)^{-1}$ exists for all $E \in +RL$. By assumptions (1a) and (2c) the components of $\delta(E)$ will also satisfy (2c). By assumptions (1a), (2a), and (2b) the
components $\tilde{A}(E)$ are also uniformly bounded. Consider the following

$$\sum_j \Delta t |a_{ij}(E)| \frac{\beta_j(E_j)}{E_j} \leq \sum_j \Delta t |a_{ij}(E)| C, \quad \text{assumption (1a)},$$

$$\leq \Delta t MC, \quad \text{assumption (2a), (2b)},$$

$$< 1 + \Delta t a_{ii}(E) \frac{\beta_i(E_i)}{E_i},$$

$$\Delta t \leq 1/(2MC) \quad \text{for } \Delta t < \delta_1 \equiv 1/(2MC).$$

Hence, $\varphi(E)$ is uniformly strictly diagonally dominant and is by assumption (2c) an $M$-matrix. So, $\varphi(E)^{-1} \geq 0$ exists for all $E \in \mathbb{R}^L$.

Second, we show there exists an $\epsilon > 0$ such that for $\Delta t \leq \delta_1$, $\varphi(E)^{-1} \eta \geq \epsilon > 0$ for all $E \in \mathbb{R}^L$. This is done by using the comparison lemma with

$$\varphi(E) = A_1 = (I + \Delta t D(E)) - (\Delta t (L(E) + U(E))), \quad D_1 \equiv I + \Delta t D(E),$$

$$B_1 \equiv \Delta t (L(E) + U(E)), \quad D_2 \equiv (I + \Delta t MC I) - (I + \Delta t D(E)),$$

$B_2 \equiv B_1$, and $D, L, U$ denote the diagonal, lower and upper parts of $\tilde{A}$. Hence, for $\eta \geq 0$ and all $E \in \mathbb{R}^L$,

$$(5) \quad \left[\varphi(E)^{-1}\right]_{i,j} \geq \left[(I + \Delta t MC I)^{-1}\right]_{i,j} \geq \frac{\min \eta_i}{1 + \Delta t MC} \geq \frac{\min \eta_i}{1 + \delta_1 MC} \equiv \epsilon > 0.$$

Third, we establish the following formula for $\varphi(E)^{-1}$

$$(6) \quad \varphi(E)^{-1} = \sum_{i=0}^{\infty} (-\Delta t \tilde{A}(E))^i.$$

It suffices to show $||-\Delta t \tilde{A}(E)||_{\infty} < \frac{1}{2}$ for suitably small $\Delta t$. Recall the norm

$$(7) \quad ||-\Delta t \tilde{A}(E)||_{\infty} = \max_i \sum_j \Delta t |a_{ij}(E)| \frac{\beta_j(E_j)}{E_j} \leq \Delta t 2 MC,$$

where (7) follows from assumptions (1a), (2a) and (2b). So, let $\delta_2 \equiv 1/(4MC) < \delta_1$. Then (7) and $\Delta t \leq \delta_2$ yield $||-\Delta t \tilde{A}(E)||_{\infty} < \frac{1}{2}$ and

$$(8) \quad ||\varphi(E)^{-1}||_{\infty} \leq \sum_{i=0}^{\infty} ||-\Delta t \tilde{A}(E)||_{\infty} \leq 2.$$

Also,

$$(9) \quad ||G(E)||_{\infty} = ||\varphi(E)^{-1}\eta||_{\infty} \leq 2||\eta||_{\infty}.$$

Equations (5) and (9) state that

$$(10) \quad G: \mathbb{R}^L \to [\epsilon, 2||\eta||_{\infty}]^L.$$

The fourth and final step is to show that $G$ is contractive. By line (10) we may view $G$ as $G: [\epsilon, 2||\eta||_{\infty}]^L \to [\epsilon, 2||\eta||_{\infty}]^L$; if $\min_j E_j^0 < \epsilon$, then we may replace $\epsilon$ by $\min(\min_j E_j^0, \epsilon)$. 

License or copyright restrictions may apply to redistribution; see http://www.ams.org/journal-terms-of-use
The following inequalities, (12) and (13), shall be useful. By assumptions (1b) and (2d) there exists \( b_j \) such that

\[
(11) \quad \left| a_{ij}(E) \frac{\beta_j(E)}{E_j} - a_{ij}(\bar{E}) \frac{\beta_j(\bar{E})}{E_j} \right| \leq m_{ij} b_j \|E - \bar{E}\|_\infty.
\]

Let \( b = \max b_j \), and use assumption (2d) and Eq. (11) to obtain

\[
(12) \quad \|\tilde{A}(E) - \tilde{A}(\bar{E})\|_\infty \leq \max \sum_j m_{ij} b_j \|E - \bar{E}\|_\infty \leq 2mb\|E - \bar{E}\|_\infty.
\]

Also, by mathematical induction on \( l \)

\[
(13) \quad \|A^l - B^l\|_\infty \leq \|A - B\|_\infty 3l^l - 1
\]

when \( \|A\|_\infty, \|B\|_\infty \leq \frac{1}{2} \), and \( A \) and \( B \) are any square matrices. In order to establish the contractive property of \( G \), use the geometric series representation given in (6) and

\[
(14) \quad \|G(E) - G(\bar{E})\|_\infty = \left\| \sum_{l=0}^{\infty} (-\Delta t \tilde{A}(E))^l - \sum_{l=0}^{\infty} (-\Delta t \tilde{A}(\bar{E}))^l \right\|_\infty
\]

\[
\leq \sum_{l=0}^{\infty} \|(-\Delta t \tilde{A}(E))^l - (-\Delta t \tilde{A}(\bar{E}))^l\|_\infty \|\eta\|_\infty
\]

\[
\leq \|\Delta t \tilde{A}(E) - \Delta t \tilde{A}(\bar{E})\|_\infty \left( \sum_{l=0}^{\infty} 3l^l - 1\right) \|\eta\|_\infty, \quad \text{by (13)}
\]

\[
\leq \Delta t 2mb\|\eta\|_\infty \|E - \bar{E}\|_\infty, \quad \text{by (12)}.
\]

The constant \( B = 3 \sum_{l=1}^{\infty} l(\frac{1}{2})^{l-1} \) is finite by comparison with the integral \( \int_1^{\infty} 3x(\frac{1}{2})^{x-1} \, dx < \infty \). Let \( \delta_2 \equiv 1/(4mb\|\eta\|_\infty) \) and \( \delta \equiv \min(\delta_2, \delta_3) \). By Eq. (14) if \( \Delta t \leq \delta \), then \( \|G(E) - G(\bar{E})\|_\infty \leq \frac{1}{2} \|E - \bar{E}\|_\infty \). Hence, for \( \Delta t \leq \delta \) the assumptions of the contractive mapping theorem hold and the sequence given by (4) for any \( E^0 \in ^*R^l \) converges to a unique solution in \( ^*R^l \) of (1).

**Remarks.**

1. In the theorem we require \( \eta \in ^*R^l \). This imposes conditions on \( \Delta t \), \( \Delta x \) and the data. However, since \( E \) is large, e.g. see the first example in Section 3, this is not severe.

2. In numerical experiments the constraints on \( \Delta t \) that are given in the proof of the theorem are not necessary for convergence. Apparently, less severe constraints may be imposed on \( \Delta t \).

3. If \( \Delta t \) is too large, uniqueness may fail, but the existence may continue to hold. Consider the simple example for \( L = 1 \), \( \beta(E) = E \), \( \eta = \Delta t \) and

\[
A(E) = \begin{cases} 
3, & E \leq 1/4, \\
\frac{1}{E} - 1, & 1/4 < E \leq 1/2, \\
1, & 1/2 < E.
\end{cases}
\]

If \( \Delta t = 1 \), then any element of \([1/4, 1/2]\) will be a solution to

\[
E + \Delta t A(E) = \eta.
\]

If \( \Delta t < 1 \), then the only positive solution is \( E = \Delta t/(1 + 3\Delta t) < 1/4 \).
4. In problems in which the algebraic problem (1) results from a one space variable Stefan problem \(A(E)\) will be tridiagonal and, hence, \(\varrho(E)^{-1}\eta\) will be tridiagonal. \(\varrho(E)^{-1}\eta\) is easily computed directly by the tridiagonal (or Thomas) algorithm. In case (1) results from a two or three space variable Stefan problem, \(\varrho(E)^{-1}\eta\) may be computed by an alternating direction method. Also, other iterative methods such as the Gauss-Seidel method may be used, as \(\varrho(E)\) is an \(M\)-matrix, to compute \(\varrho(E)^{-1}\eta\).

5. If \(\eta = \eta(E) = (\eta_j(E_j))\), then an iterative scheme \(E^{k+1} = \varrho(E^k)^{-1}\eta(E^k)\) can be defined. If the \(\eta_j\) satisfy certain conditions, e.g. Lipschitz continuous and uniformly bounded, and \(\Delta t\) is suitably small (see Eqs. (6)-(9)), then \(E^{k+1} \to E\) and \(E\) is a solution of \(E + \Delta tA(E)\beta(E) = \eta(E)\).

6. Work on comparison with other methods and their rates of convergence is in progress. Because of the lack of smoothness of \(\beta(E)\) at the solid-liquid interface, convergence is slow near this region.

3. Examples. The first example illustrates \(H(u), \beta(u), A(E)\) for algorithm (4). It models the freezing of water starting from 310K and going down to 73K, the temperature of liquid nitrogen. Thus the thermal properties vary with temperature. A comparison with the explicit time discretization used in [10] is given. The second example has a thermal conductivity which is temperature and space dependent. The computations agree with those of two other algorithms given in R. E. White [14].

**Example 1.** The following data for water was taken from [10] and the units are cgs. Assume temperature \(u < 373\).

\[
\rho(u) = \begin{cases} 
920, & u \leq 273 = u_f, \\
1,000, & u > 273.
\end{cases}
\]

\[
c(u) = \begin{cases} 
.007,16u + .138, & u \leq 273, \\
4.18668, & u > 273.
\end{cases}
\]

\[
K(u) = \begin{cases} 
.002,24 + .000,005,95(273 - u)^{1.156}, & u \leq 273, \\
.000,101,7 + .000,001,695u, & u > 273.
\end{cases}
\]

\[
H(u) = \int_0^u \rho(\bar{u})c(\bar{u})d\bar{u} = \begin{cases} 
3.293,6u^2 + 126.96u, & u \leq 273, \\
4186.8(u - 273) + 613,858.79, & u > 273.
\end{cases}
\]

\[
\rho_fL = 333,730,
\]

\[
\beta(E) = \begin{cases} 
-19.274 + \sqrt{371.477 + .30362E}, & E \leq 280,128.79, \\
273, & 280,128.79 < E < 613,858.79, \\
.000,238,86E + 126.411, & 613,858.79 \leq E.
\end{cases}
\]

\[
K(E) = \begin{cases} 
.002,24 + .000,005,95(273 - \beta(E))^{1.156}, & E \leq 280,128.79, \\
-.000,000,005,022,77(E - 280,128.79) + .002,4, & 280,128.79 < E < 613,858.79, \\
.000,101,7 + .000,001,695\beta(E), & 613,858.79 \leq E.
\end{cases}
\]
We considered the problem (15.1)--(15.4)

\begin{align*}
(15.1) & \quad E_x - (K(E)\beta(E))_x = 0, \\
(15.2) & \quad \beta(E(0, t)) = 73, \\
(15.3) & \quad \beta(E(.05, t)) = 0, \\
(15.4) & \quad E(x, 0) = H(310).
\end{align*}

We used $\Delta x = .05/20$ and $\Delta t = .5$ for $0 \leq t \leq 6.0$, and $\Delta t = 6.0$ for $t > 6.0$. The matrix $\partial(E)$ is an $M$-matrix and tridiagonal. The computation $\partial(E)^{-1}\eta$ is easily done by the tridiagonal (Thomas) algorithm. Convergence is given when $|E^{k+1}_i - E^k_i| < 1000$, for all $i$, and was usually obtained in 5 to 8 iterations. The results are given in Table 1 where they are compared with those in [10]. In [10] an explicit scheme was used i.e.,

$$\rho(E) \frac{E^{k+1}_i - E^k_i}{\Delta t} - \nabla \cdot (K(E) \nabla \beta(E)) = 0$$

and $\rho$ does not appear in $\beta$. Note the savings in computation time.

<table>
<thead>
<tr>
<th>$x$</th>
<th>Algorithm (4)</th>
<th>Explicit ($\Delta x = .0025 \quad \Delta t = .25$)</th>
<th>Explicit ($\Delta x = .00125 \quad \Delta t = .0625$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.</td>
<td>73.00</td>
<td>73.00</td>
<td>73.00</td>
</tr>
<tr>
<td>0.05</td>
<td>94.17</td>
<td>93.79</td>
<td>93.74</td>
</tr>
<tr>
<td>.010</td>
<td>116.9</td>
<td>116.04</td>
<td>115.95</td>
</tr>
<tr>
<td>.015</td>
<td>141.5</td>
<td>139.91</td>
<td>139.84</td>
</tr>
<tr>
<td>.020</td>
<td>168.2</td>
<td>165.55</td>
<td>165.59</td>
</tr>
<tr>
<td>.025</td>
<td>197.8</td>
<td>193.12</td>
<td>193.36</td>
</tr>
<tr>
<td>.030</td>
<td>230.8</td>
<td>222.84</td>
<td>223.25</td>
</tr>
<tr>
<td>.035</td>
<td>267.0</td>
<td>255.32</td>
<td>255.56</td>
</tr>
<tr>
<td>.040</td>
<td>292.3</td>
<td>280.88</td>
<td>285.37</td>
</tr>
<tr>
<td>.045</td>
<td>304.5</td>
<td>301.92</td>
<td>303.01</td>
</tr>
<tr>
<td>.050</td>
<td>307.2</td>
<td>306.94</td>
<td>307.01</td>
</tr>
</tbody>
</table>

**Example 2.** Let $K$ be given by Figure 3 and $\beta(E)$ be given by Figure 1 with $\rho_s c_s = u_f = \rho_t c_t = 1$ and $\rho_s L = 2 = H$. Consider the one space variable version of (2.1)--(2.4).

\begin{align*}
(15.1) & \quad E_t - (K\beta(E))_x = A, \quad (x, t) \in (0, 1) \times (0, T), \\
(15.2) & \quad -K\beta(E)_x = B(TB - \beta(E)), \quad x = 0, \\
(15.3) & \quad \beta(E)_x = 0, \quad x = 1, \\
(15.4) & \quad E(x, 0) = .000,001.
\end{align*}
In the computations that are given by the graphs in Figure 4 we used the following constants: \( A = 10.0, \ DK = 1.0, \ AK = 1.0, H = 2.0, B = 2.0, TB = -1.0, DX = .05 \) and \( T = .5 \). The algorithm (4) was considered to have converged, at each time step when \(|E_{i}^{k+1} - E_{i}^{k}| < ER\) for all \( i = 1, \ldots, 21 \). \( ER = .001 \) was used. The tridiagonal algorithm was used to compute \( \mathcal{G}(E^{k})^{-1} \eta \) where \( \eta_{i} = DT^*A + \bar{E}_{i} \).

\[
K(x,E) = e^{-AK*x}K(E) \quad (a)
\]

\[
K(x,u) = e^{-AK*x}K(u) \quad (b)
\]

**Figure 3**

*Thermal conductivity for Example 2*

When \( DT = .0125 \), convergence was not obtained for \( k \leq 50 \). When \( DT = .00125 \) was used, convergence was usually obtained within 2 to 9 iterations. The computations agree within 1\% of the computations obtained by two other algorithms in R. E. White [14]. The reader should note that this \( DT \) is above the stability condition, that is required by an explicit time discretization,

\[
\frac{\max K}{\min \rho c} \frac{DT}{DX*DX} < \frac{1}{2},
\]

i.e., \( DT < .000,312,5 \).
Figure 4

Graph of $E(x, t)$ in Example 2


