QUANTITATIVE COMPARISON AMONG SEVERAL DIFFERENCE SCHEMES*[1)]

NI LIN-AN (倪林安) WU XIONG-HUA (吴雄华)
WANG YONG (王 勇) ZHU YOU-LAN (朱幼兰)
(Computing Center, Academia Sinica, Beijing, China)

Abstract

When one compares a difference scheme with another, only a qualitative comparison is usually given. Such a comparison is not enough. For example, scheme A is more accurate than scheme B, but it would take more time to use scheme A than to use scheme B. In order to determine which scheme is the best, it is necessary to make a quantitative comparison among difference schemes.

When the state equation is non-convex, the numerical solution is sensitive to methods[1,2]. We make a numerical test with 10 different schemes for such a problem. From the computed results the following conclusions are obtained:

The physically relevant solutions can be obtained if the Godunov scheme and the first-order E-O scheme are used, but the solutions are not so accurate. In our problem, it is necessary to take at least 80000 mesh points in the space direction in order to obtain a solution with an error of 10^{-3}. The physically relevant solution can also be obtained by using the Lax scheme, but its accuracy is lower than those of the Godunov scheme and the E-O scheme.

The physically relevant solutions cannot be obtained by using the L-W scheme, the MacCormack scheme, the Murman scheme, the Richtmyer scheme, the Courant scheme and the second-order one-sided scheme.

The physically relevant solutions can also be obtained by using the second-order singularity-separating method (S-S scheme for short). For our problem, 15—25 mesh points in the space direction are enough for a solution whose error is 10^{-3}. That is, in this case the number of mesh points for the S-S method is $\frac{1}{3200}-\frac{1}{5300}$ of that for the Godunov scheme or the first-order E-O scheme. We know from the computation that the convergence rates of the Godunov scheme and the first-order E-O scheme are about $O(\Delta t^{1/2})$ in L_2 space, but that of the S-S method is $O(\Delta t^2)$[2)]. We can see that the higher the required accuracy, the larger the difference of the computation amount. Moreover, because of the rounding errors, we cannot make the computational error infinitely small. If we solve our problem using a computer with a word length of 32 bits (the length of mantissa is 24 bits), the smallest possible error is 7×10^{-3} for the E-O scheme, but it is 10^{-5} for the S-S method. This is because the amount of computation for our method is less, and the problem of rounding errors is not so serious.

§ 1. The Problem

We consider the following initial-boundary-value problem with a non-convex equation of state:

$$\begin{cases}\dfrac{\partial U}{\partial t}+\dfrac{\partial f(U)}{\partial x}=0,\\ U(x,0)=\begin{cases}0.656-200(x+0.001), & -0.001\leqslant x<-0.0005,\\ 0.656+200x, & -0.0005\leqslant x<0,\\ 0.014+170x, & 0\leqslant x<0.0005,\\ 0.014-170(x-0.001), & 0.0005\leqslant x\leqslant 0.001.\end{cases}\end{cases}$$

* Received December 3, 1983.
1) Projects supported by the Science Fund of the Chinese Academy of Sciences.
2) In § 4 the definitions of norms are given.

$$\begin{cases} U(-0.001,\ t)=0.656-4x_1(t), \\ U(0.001,\ t)=0.014-3.4x_2(t), \end{cases}$$

where the equation of state is

$$f(U)=U^4/2-19U^3/30+U^2/4-33U/1000,$$

and $x_1(t)$, $x_2(t)$ are implicitly given by the following formulas

$$x_1(t)=tf'(0.656-4x_1(t)),$$
$$x_2(t)=tf'(0.014-3.4x_2(t)).$$

§ 2. The Schemes

The following schemes are used for the above problem.

1. The singularity-separating method (the S-S method)[13,16]

This method has been proven both in theory and in practice to be a very good method for the initial-boundary-value problem of the first-order quasi-linear hyperbolic equations. The discontinuity conditions are used on the discontinuity lines and there is no difference across the discontinuity lines in the system of difference equations. More accurate solutions can be obtained by using a few mesh points because of the small truncation errors. The figure of the solution for our problem is shown in Fig. 1. In order to solve this problem by the S-S method, we introduce a new coordinate system through the following coordinate transformation:

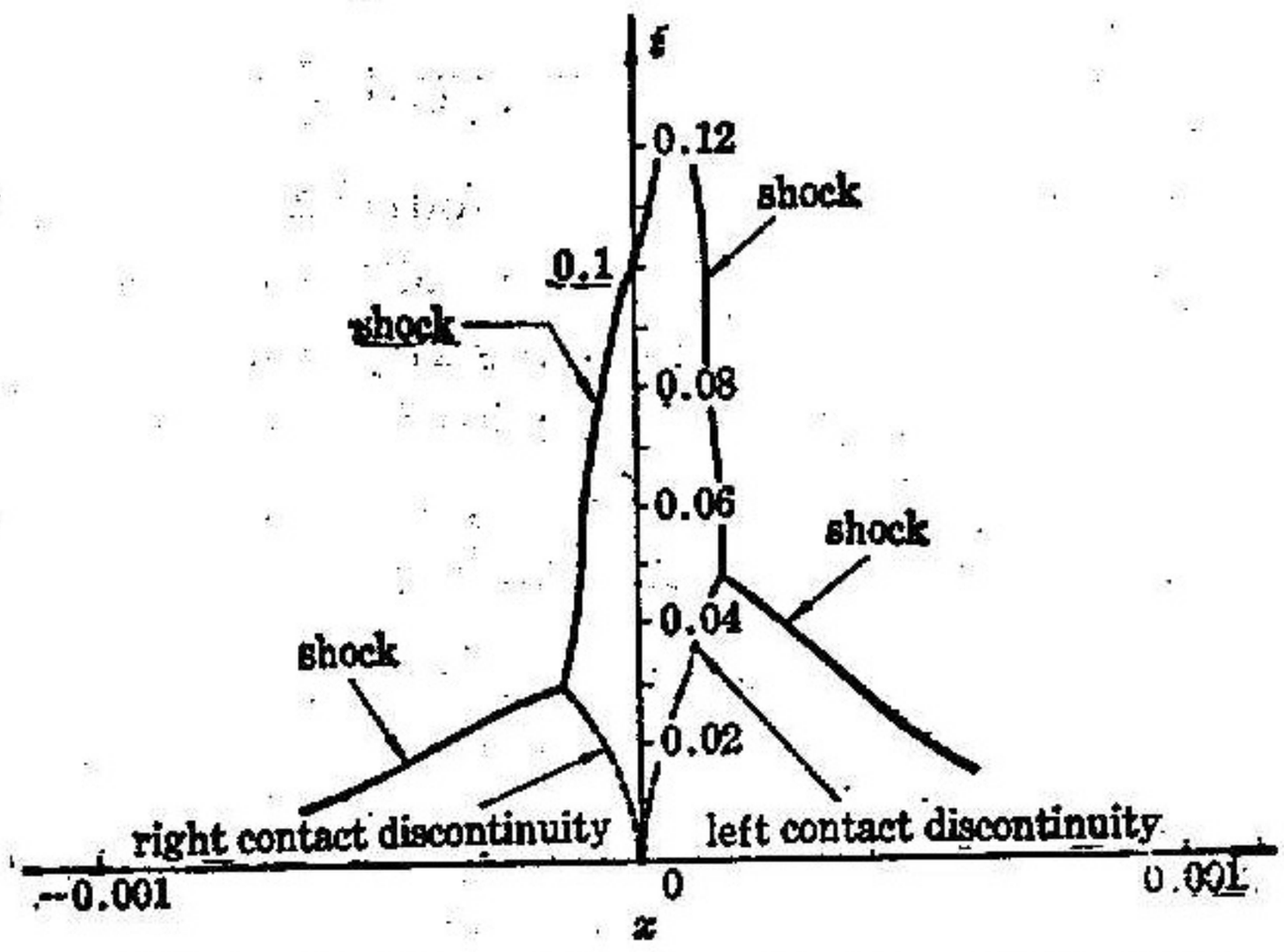

Fig. 1 Locations of the discontinuities of $U(x, t)$ in plane (x, t) (the S-S method)

$$\begin{cases} \xi=\dfrac{x-x_{l-1}(t)}{x_l(t)-x_{l-1}(t)}+l-1, & \text{if } x_{l-1}(t)\leqslant x\leqslant x_l(t), \\ t=t. \end{cases}$$

Here $x_l(t)$ represents a boundary line or a discontinuity line (From Fig. 1 it is clear that there are several discontinuity lines).

Through the above transformation, a problem with movable boundaries is changed into a problem with fixed boundaries. It is convenient for treating boundaries accurately. Suppose

$$\lambda=\frac{\partial\xi}{\partial t}+\frac{\partial f}{\partial U}\frac{\partial\xi}{\partial x},$$

then the equation in the new coordinate system can be written as:

$$\frac{\partial U}{\partial t}+\lambda\frac{\partial U}{\partial\xi}=0.$$

For this equation the following scheme is used in each subregion. Suppose

$$\sigma=\lambda\frac{\Delta t}{\Delta\xi},\quad \mu F_m=\frac{1}{2}(F_{m+1/2}+F_{m-1/2}),$$

$$\Delta_- F_m=F_m-F_{m-1},\quad \Delta_+ F_m=F_{m+1}-F_m,$$

$$(\sigma\Delta_0 U)_m=\begin{cases}\sigma_m\Delta_+ U_m, & \text{if } \sigma_m\leqslant 0,\\ \sigma_m\Delta_- U_m, & \text{if } \sigma_m\geqslant 0.\end{cases}$$

If $|\sigma_m|\leqslant 1$, the scheme

$$U_m^{k+1/2}=U_m^k-\frac{1}{2}(\sigma\Delta_0 U)_m^k,$$

$$U_m^{k+1}=U_{m\pm1}^k-(\mu\sigma_{m\pm1/2}^{k+1/2}\pm 1)\Delta_\pm U_m^{k+1/2}$$

is adopted.

If $|\sigma_m|\geqslant 1$, the following scheme is used:

$$\frac{1}{2}(U_m^{k+1/2}+U_{m\pm1}^{k+1/2})+\frac{1}{2}\mu\sigma_{m\pm1/2}^k\Delta_\pm U_{m\pm1}^{k+1/2}=\frac{1}{2}(U_m^k+U_{m\pm1}^k),$$

$$\frac{1}{2}(U_m^{k+1}+U_{m\pm1}^{k+1})+\frac{1}{2}\mu\sigma_{m\pm1/2}^{k+1/2}\Delta_\pm U_m^{k+1}=\frac{1}{2}(U_m^k+U_{m\pm1}^k)-\frac{1}{2}\mu\sigma_{m\pm1/2}^{k+1/2}\Delta_\pm U_m^k.$$

In the above scheme, there are fourteen signs $\pm$. The upper signs are used if $\lambda<0$, and the lower signs are used if $\lambda>0$; either can be used if $\lambda\approx 0$. g_m^k is the abbreviation of $g(m\Delta\xi, k\Delta t)$, where g represents any function of variables ξ and t. Computed results have been obtained for the problem in § 1 by using this scheme. In our computation each subregion is divided into 5, 10, 20, 40 or 80 parts.

2. The Godunov scheme[4, 14]

$$U_m^{k+1}=U_m^k-\frac{\Delta t_k}{\Delta x}(h^G(U_{m+1}^k, U_m^k)-h^G(U_m^k, U_{m-1}^k)),$$

where

$$h^G(U_{m+1}^k, U_m^k)=\begin{cases}\min\limits_{U_m^k<U<U_{m+1}^k} f(U), & \text{if } U_m^k<U_{m+1}^k,\\ \max\limits_{U_{m+1}^k<U<U_m^k} f(U), & \text{if } U_{m+1}^k<U_m^k.\end{cases}$$

In our computation, Δt_k satisfies the relation $\max\limits_m f|(U_m^k)|\frac{\Delta t_k}{\Delta x}=0.95$, and the computational region in the x direction is divided into n equal parts ($n=100$, 200, 400, 800).

3. The Engquist–Osher scheme (E–O scheme)[15]

$$U_m^{k+1}=U_m^k-\frac{\Delta t_k}{\Delta x}(\Delta_+ f_-(U_m^k)+\Delta_- f_+(U_m^k)),$$

where

$$f_-(U)=\int_0^U(1-\chi(s))f'(s)ds,$$

$$f_+(U)=\int_0^U\chi(s)f'(s)ds,$$

$$\chi(s)=\begin{cases}1, & \text{if } f'(s)\geqslant 0,\\ 0, & \text{if } f'(s)<0,\end{cases}$$

and

$$\Delta_+ F_m^k=F_{m+1}^k-F_m^k,$$

$$\Delta_- F_m^k=F_m^k-F_{m-1}^k.$$

In our computation, Δt_k satisfies the relation $\max|f'(U_m^k)|\frac{\Delta t_k}{\Delta x}=0.95$, and the

computational region in the x direction is divided into n equal parts ($n=100$, 200, 400, 800, 1600).

4. The Lax scheme[6]

$$U_m^{k+1}=\frac{1}{2}(U_{m+1}^k+U_{m-1}^k)-\frac{1}{2}\frac{\Delta t_k}{\Delta x}(f(U_{m+1}^k)-f(U_{m-1}^k)).$$

In our computation, Δt_k satisfies the relation $\max_m|f'(U_m^k)|\frac{\Delta t_k}{\Delta x}=0.95$, and the computational region in the x direction is divided into n equal parts ($n=100$, 200, 400, 800).

5. The Murman scheme[7]

$$U_m^{k+1}=U_m^k-\frac{\Delta t_k}{\Delta x}(h^M(U_{m+1}^k, U_m^k)-h^M(U_m^k, U_{m-1}^k)),$$

where
$$h^M(U_{m+1}^k, U_m^k)=\frac{1}{2}[f(U_{m+1}^k)+f(U_m^k)-a_{m+1/2}^k(U_{m+1}^k-U_m^k)],$$

$$a_{m+1/2}^k=\left|\frac{\Delta_- f(U_{m+1}^k)}{\Delta_- U_{m+1}^k}\right|=\left|\frac{f(U_{m+1}^k)-f(U_m^k)}{U_{m+1}^k-U_m^k}\right|.$$

In our computation, Δt_k satisfies the relation $\max_m|f'(U_m^k)|\frac{\Delta t_k}{\Delta x}=0.95$, and the computational region in the x direction is divided into n equal parts ($n=50$, 100, 200, 400, 800).

6. The Courant scheme[8]

$$U_m^{k+1}=U_m^k-\frac{\Delta t_k}{\Delta x}h^0(U_m^k),$$

where
$$h^0(U_m^k)=\begin{cases}f(U_{m+1}^k)-f(U_m^k), & \text{if } f'(U_m^k)<0,\\ f(U_m^k)-f(U_{m-1}^k), & \text{if } f'(U_m^k)>0.\end{cases}$$

In our computation Δt_k satisfies the relation $\max_m|f'(U_m^k)|\frac{\Delta t_k}{\Delta x}=0.95$ (or 0.75), and the computational region in the x direction is divided into n equal parts ($n=100$, 200).

7. The second-order one-sided scheme[9]

$$U_m^{k+1/2}=U_m^k-\frac{1}{2}\frac{\Delta t_k}{\Delta x}\Delta_\pm f(U_m^k),$$

$$U_m^{k+1}=U_{m\pm1}^k\pm\Delta_\pm U_m^{k+1/2}-\frac{\Delta t_k}{\Delta x}\Delta_\pm f(U_m^{k+1/2}),$$

where
$$\Delta_+F_m=F_{m+1}-F_m,\quad \Delta_-F_m=F_m-F_{m-1}.$$

In the above scheme, there are several signs $\pm$. The upper signs are used if $f'(U_m^k)<0$, and the lower signs are used if $f'(U_m^k)>0$. In our computation Δt_k satisfies the relation $\max_m|f'(U_m^k)|\frac{\Delta t_k}{\Delta x}=1.9$ (or 0.95, or 0.5), and the computational region in the x direction is divided into 100 equal parts.

8. The Lax-Wendroff scheme[10]

$$U_m^{k+1}=U_m^k-\frac{1}{2}\frac{\Delta t_k}{\Delta x}(f(U_{m+1}^k)-f(U_{m-1}^k))+\frac{1}{2}\left(\frac{\Delta t_k}{\Delta x}\right)^2\Delta_-(a_{m+1/2}^k\Delta_+f(U_m^k)),$$

where
$$a_{m+1/2}^k=\frac{1}{2}(f'(U_{m+1}^k)+f'(U_m^k)).$$

In our computation, Δt_k satisfies the following relation

$$\max_m |f'(U_m^k)| \frac{\Delta t_k}{\Delta x} = 0.95,$$

and the computational region in the x direction is divided into 100 equal parts.

9. The MacCormack scheme[11]

$$\begin{cases} U_m^{k+1/2} = U_m^k - \dfrac{\Delta t_k}{\Delta x}(f(U_{m+1}^k) - f(U_m^k)) + s\dfrac{\Delta t_k}{\Delta x}(f(U_{m+1}^k) - 2f(U_m^k) + f(U_{m+1}^k)), \\ U_m^{k+1} = \dfrac{1}{2}(U_m^{k+1/2} + U_m^k) - \dfrac{1}{2}\dfrac{\Delta t_k}{\Delta x}(f(U_m^{k+1/2}) - f(U_{m-1}^{k+1/2})) - \dfrac{s}{2}\dfrac{\Delta t_k}{\Delta x}(f(U_{m+1}^{k+1/2}) \\ \qquad -2f(U_m^{k+1/2}) + f(U_{m-1}^{k+1/2})). \end{cases}$$

where $s=0$ or $s=1$. For the problem in § 1, we take

$$s = \begin{cases} 1, & \text{if } K = \text{an odd number}, \\ 0, & \text{if } K = \text{an even number}. \end{cases}$$

In our computation, Δt_k satisfies the following relation

$$\max_m |f'(U_m^k)| \frac{\Delta t_k}{\Delta x} = 0.95 \text{ (or } 0.75, \text{ or } 0.5).$$

The computational region in the x direction is divided into n equal parts ($n=100$, 200, 400).

10. The Richtmyer scheme[12]

$$\begin{cases} U_{m+1/2}^{k+1/2} = \dfrac{1}{2}(U_{m+1}^k + U_m^k) - \dfrac{1}{2}\dfrac{\Delta t_k}{\Delta x}(f(U_{m+1}^k) - f(U_m^k)), \\ U_m^{k+1} = U_m^k - \dfrac{\Delta t_k}{\Delta x}(f(U_{m+1/2}^{k+1/2}) - f(U_{m-1/2}^{k+1/2})). \end{cases}$$

In our computation, Δt_k satisfies the relation $\max_m |f'(U_m^k)| \dfrac{\Delta t_k}{\Delta x} = 0.95$, and the computational region in the x direction is divided into 100 equal parts.

§ 3. Computation of the Boundary Values

Because the boundary condition is not an explicit equation, a nonlinear equation in x_1^{k+1} (or x_2^{k+1}) must be solved to obtain the boundary values U at time $(k+1)\Delta t$. The secant iterative method is used for the equation. The initial value of variable $x_1^{k+1}(x_2^{k+1})$ is $x_1^k(x_2^k)$, which is the value at the last time step, and another value is near $x_1^k(x_2^k)$. When x_1^{k+1}(or x_2^{k+1}) is obtained, $U(-0.001, (k+1)\Delta t)$ (or $U(0.001, (k+1)\Delta t)$) can be obtained immediately.

§ 4. Numerical Results and Analysis

The schemes in § 2 have been used for our problem. The errors at $t=0.066$ in the sense of L_2–norm:

$$\|\sigma\| = \sqrt{\frac{1}{b-a}\int_a^b (U(x,t) - U^*(x,t))^2 dx}$$

have been given in Table 1. Here, $a=-0.001$, $b=0.001$ and U^* represents the exact solution. The distributions of U at $t=0.066$ obtained by using these schemes are given in Figs. 2—12. In what follows, we shall discuss and analyse these results.

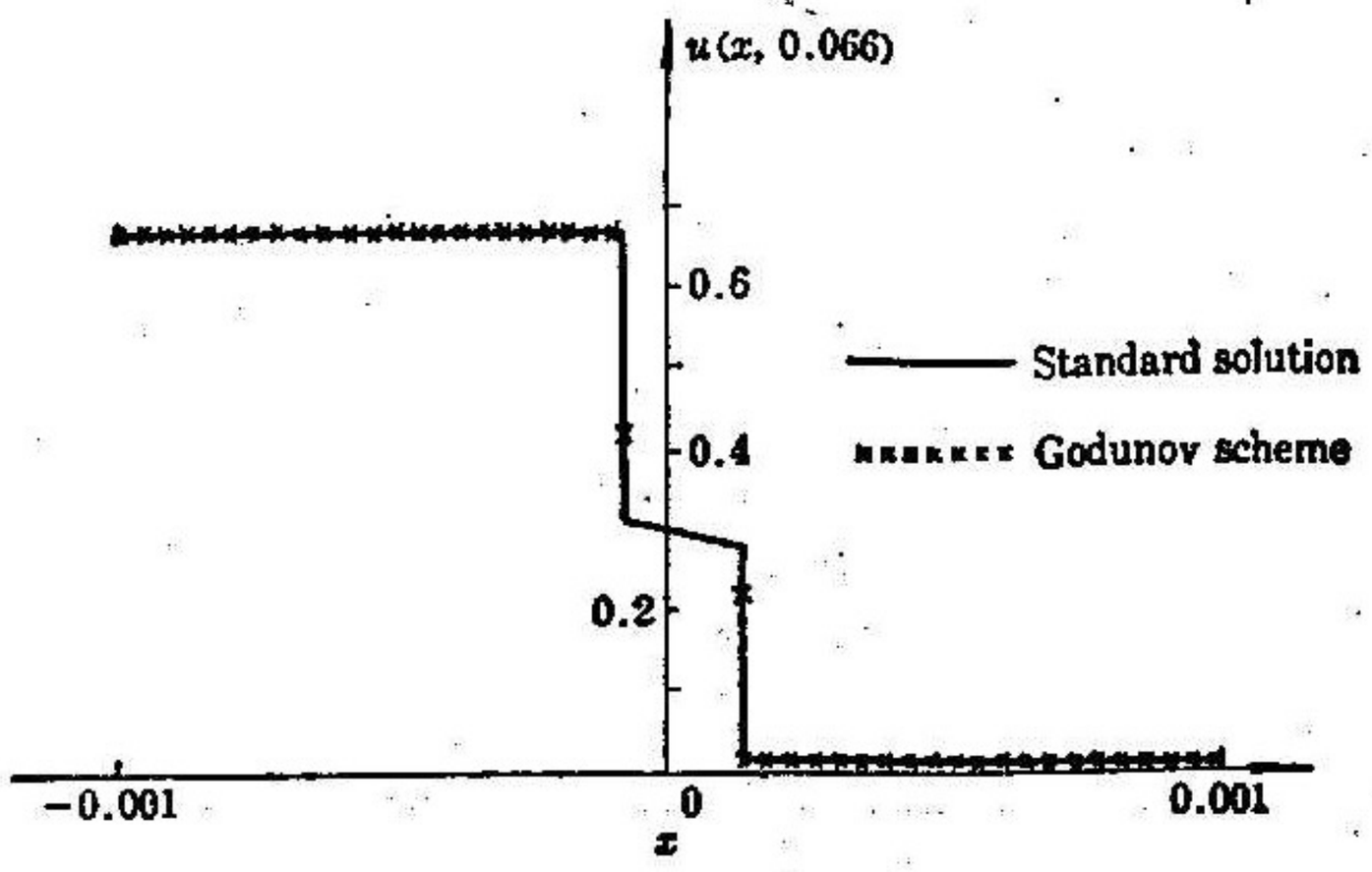

Fig. 2 Comparison between the solution of the Godunov scheme (800 equal parts) and the standard solution

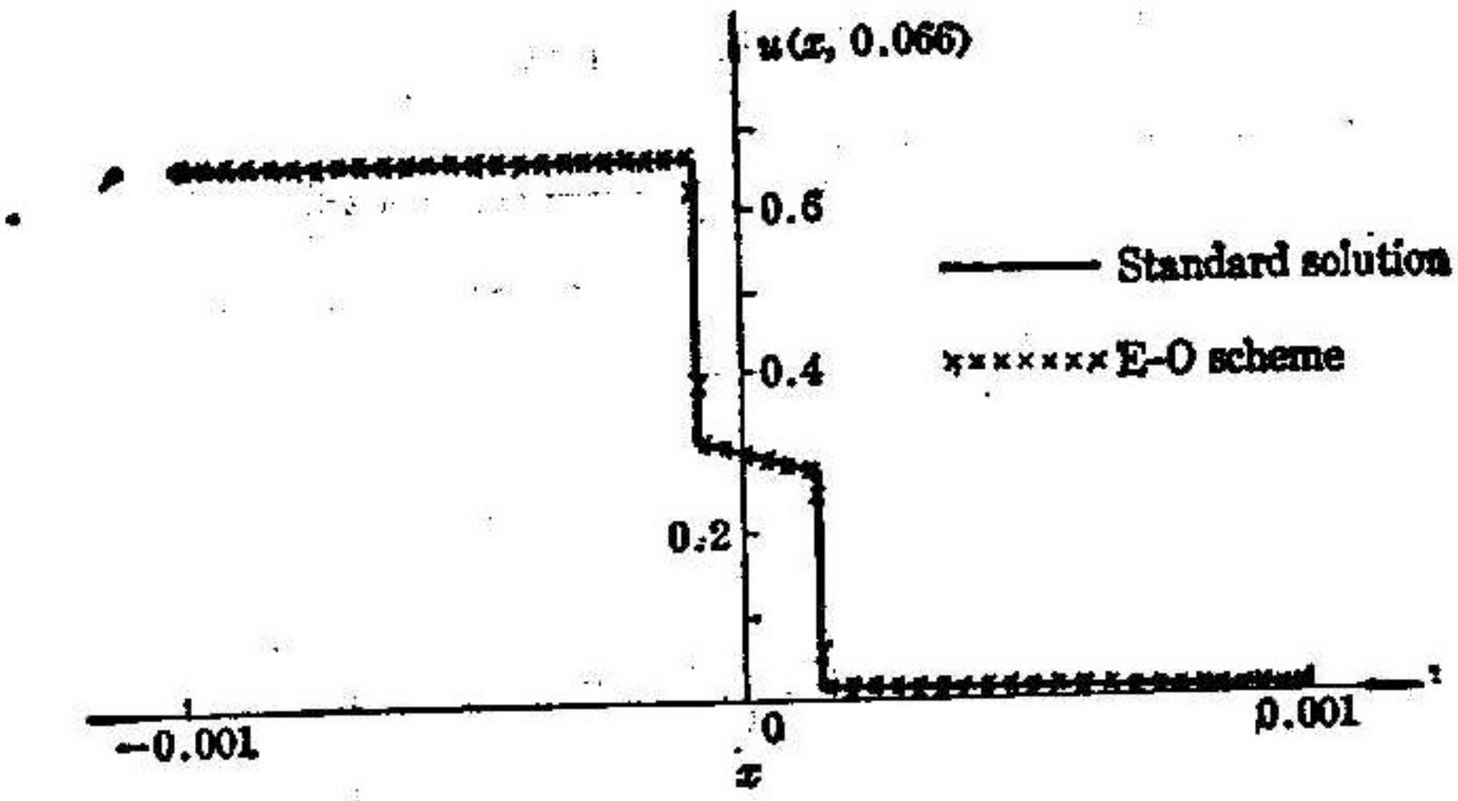

Fig. 3 Comparison between the solution of the E–O scheme (800 equal parts) and the standard solution

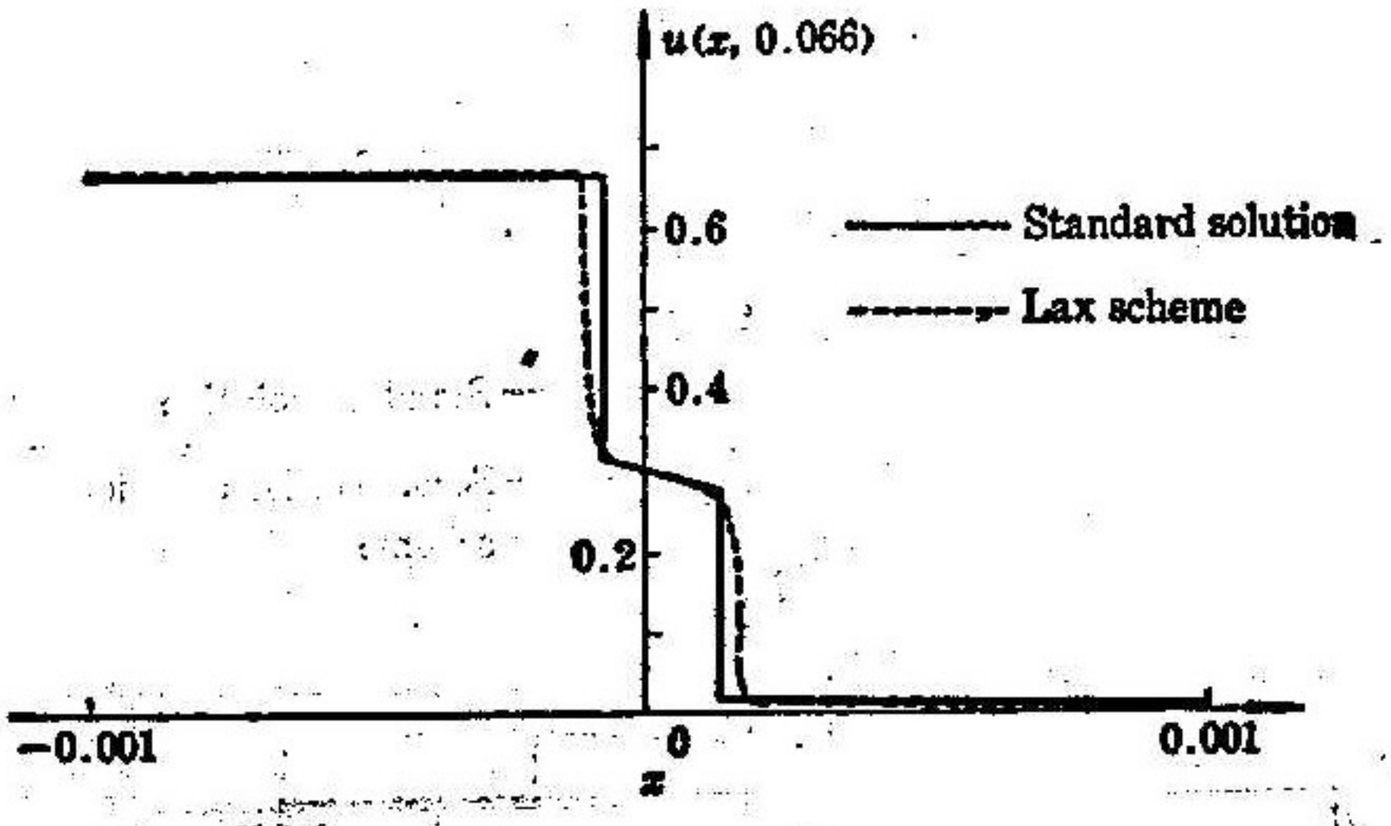

Fig. 4 Comparison between the solution of the Lax scheme (800 equal parts) and the standard solution

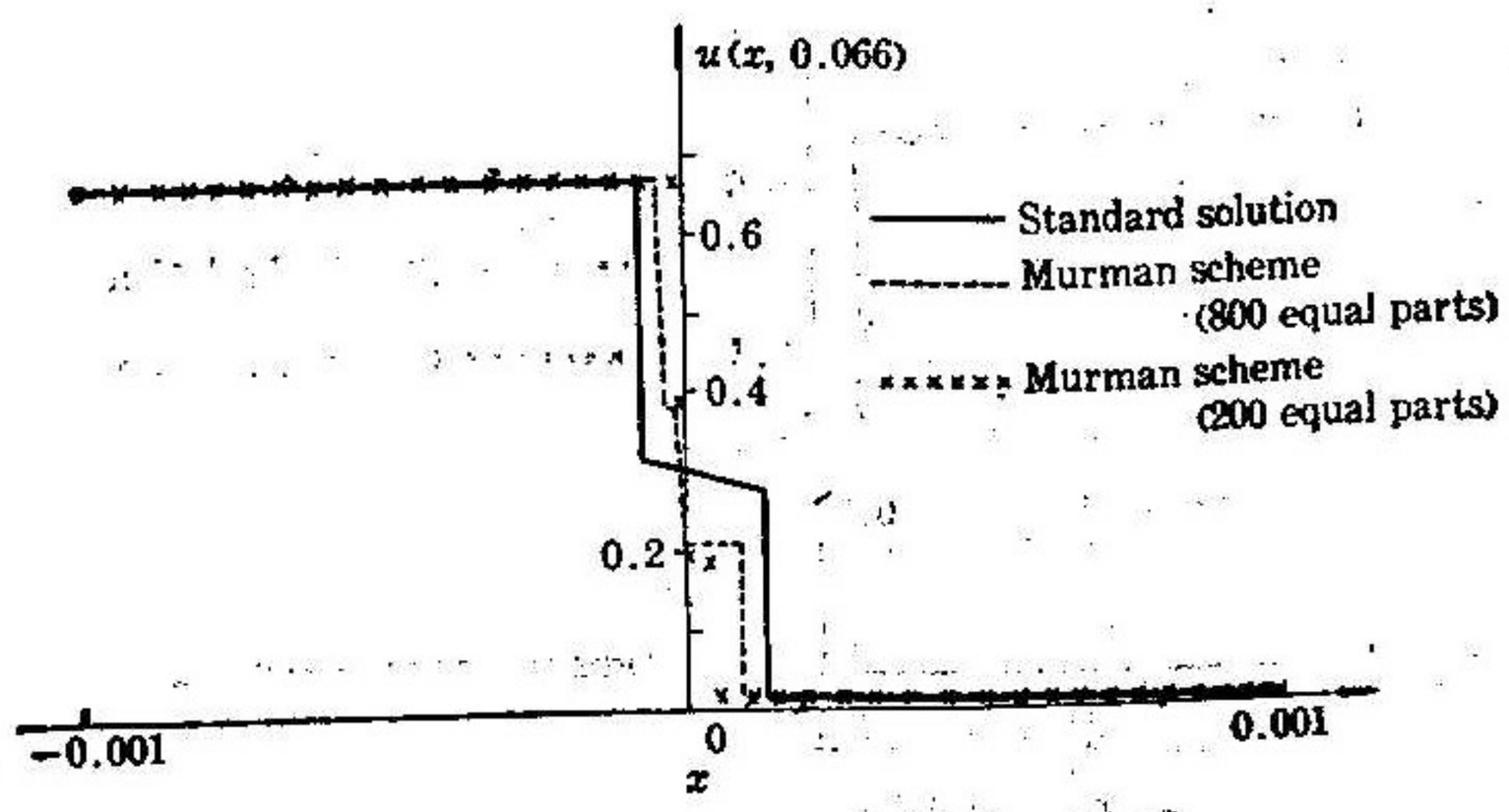

Fig. 5 Comparison between the solution of the Murman scheme (800 equal parts and 200 equal parts) and the standard solution

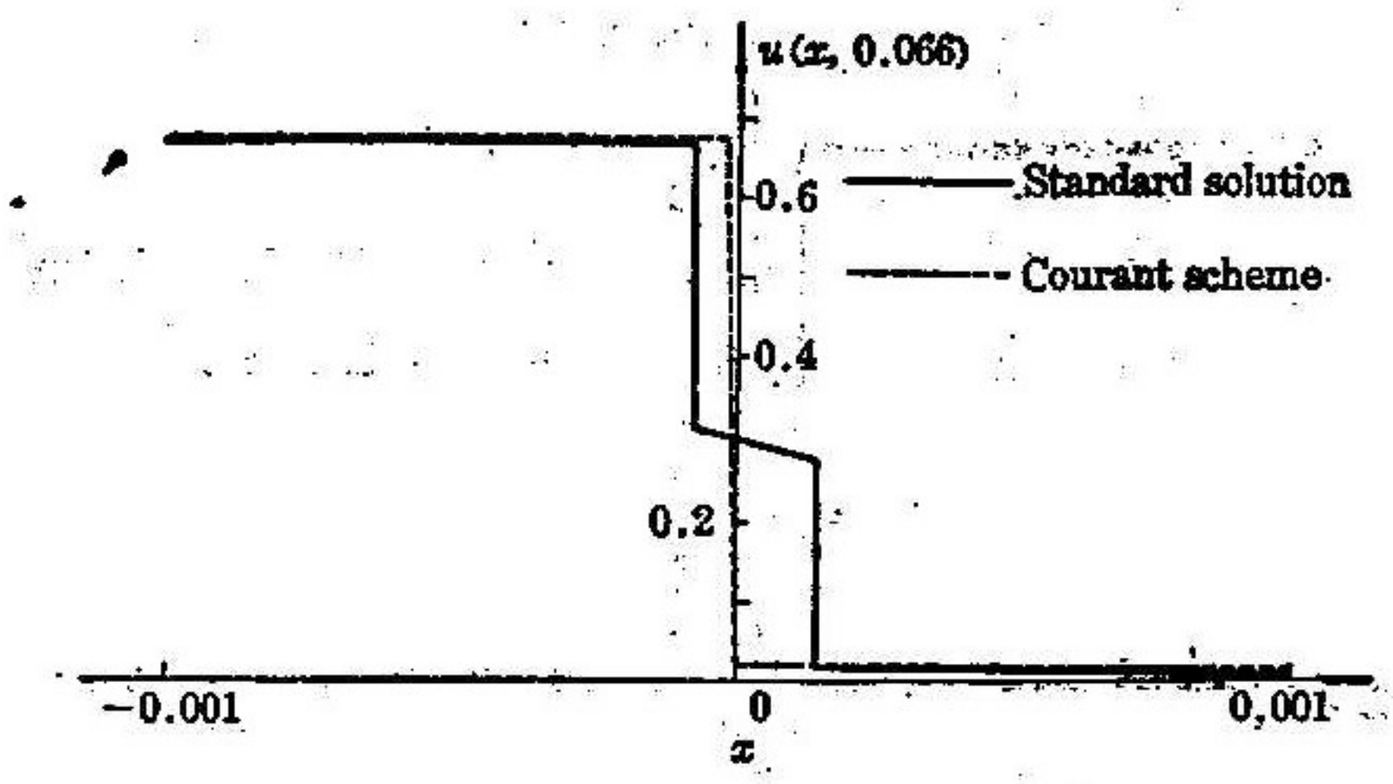

Fig. 6 Comparison between the solution of the Courant scheme (100 equal parts and 200 equal parts) and the standard solution

$$\left(\max_m |f'(U_m^k)| \frac{\Delta t_k}{\Delta x} = 0.95,\ 0.75\right)$$

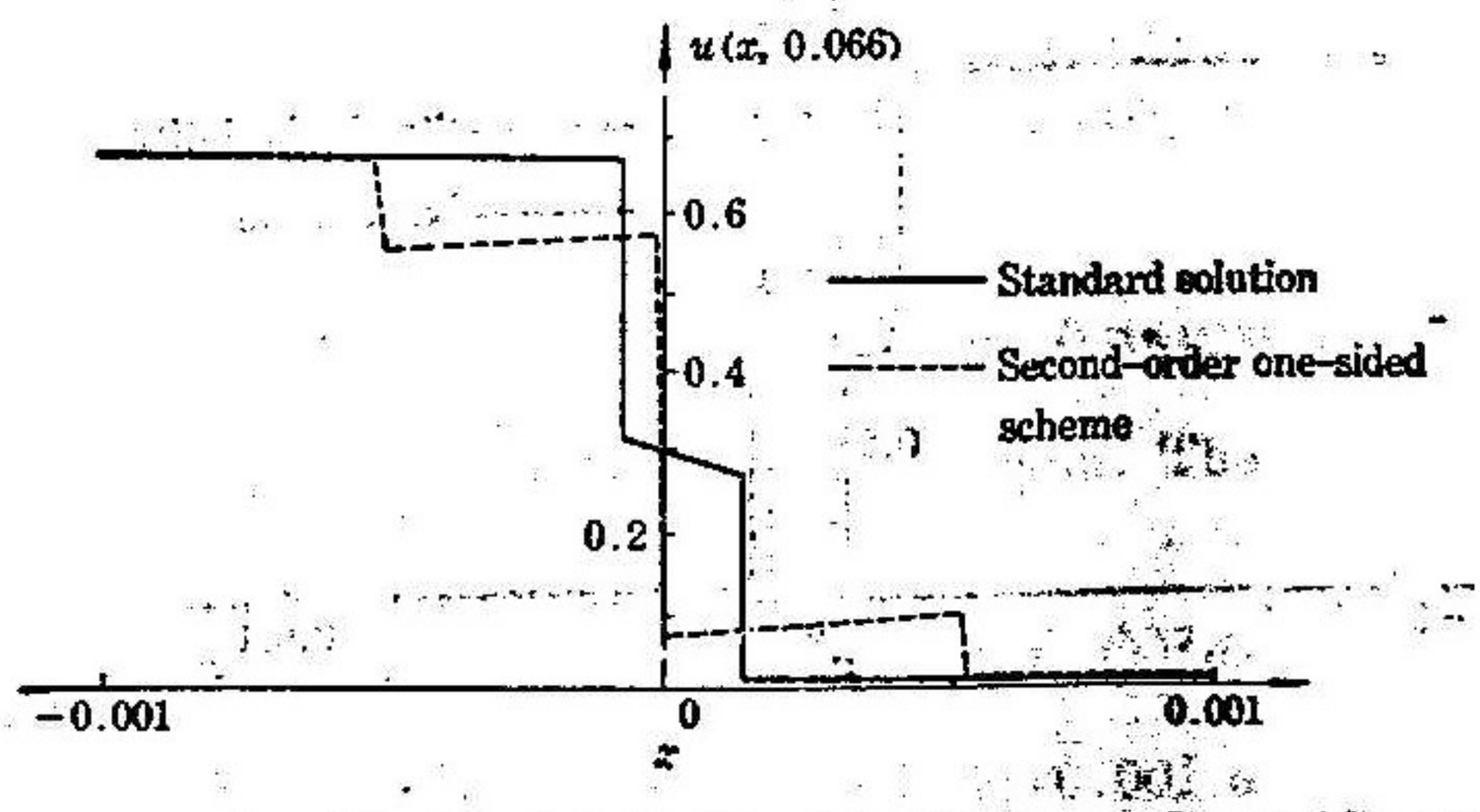

Fig. 7 Comparison between the solution of the second-order one-sided scheme (100 equal parts) and the standard solution

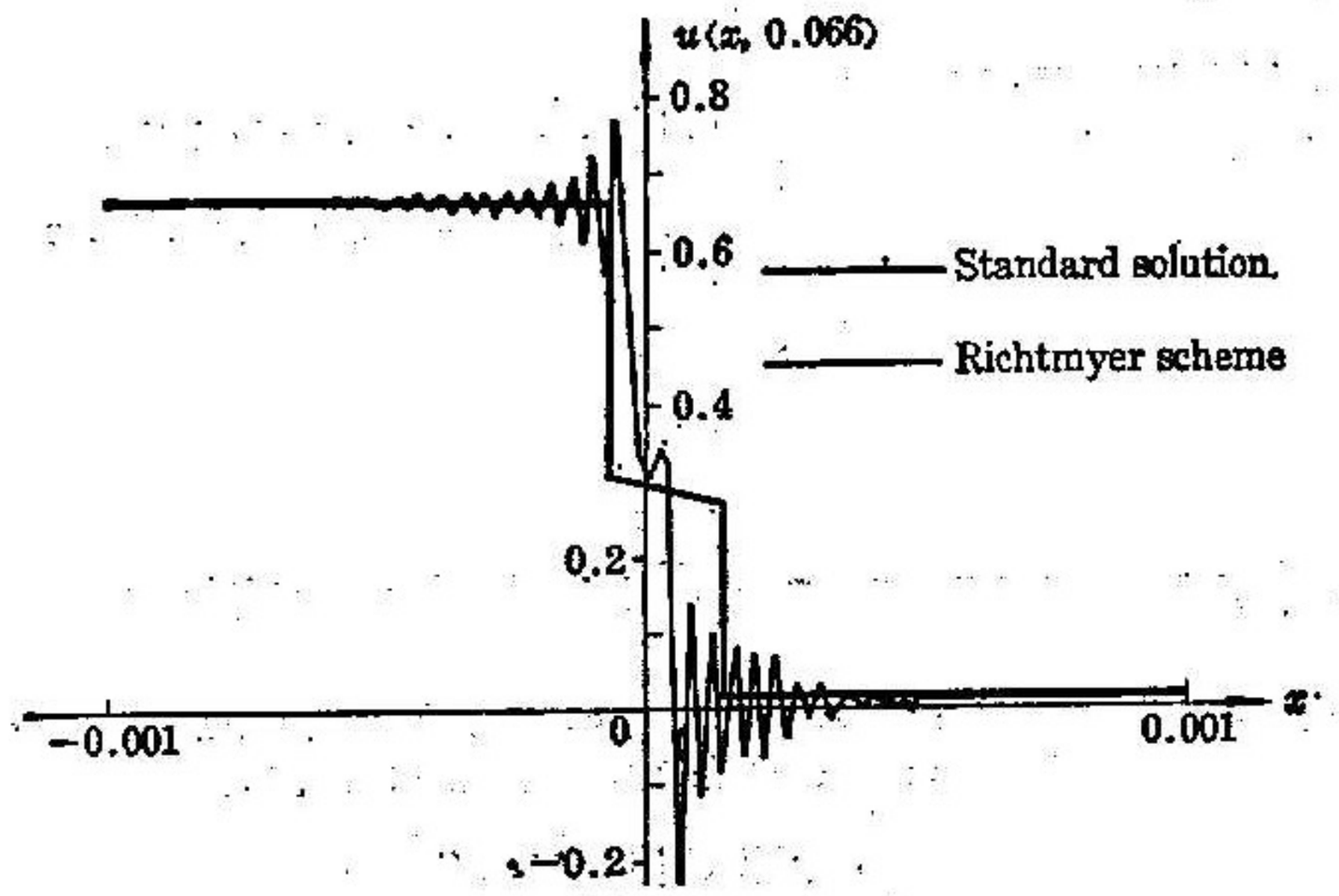

Fig. 8 Comparison between the solution of the Richtmyer scheme (100 equal parts) and standard solution

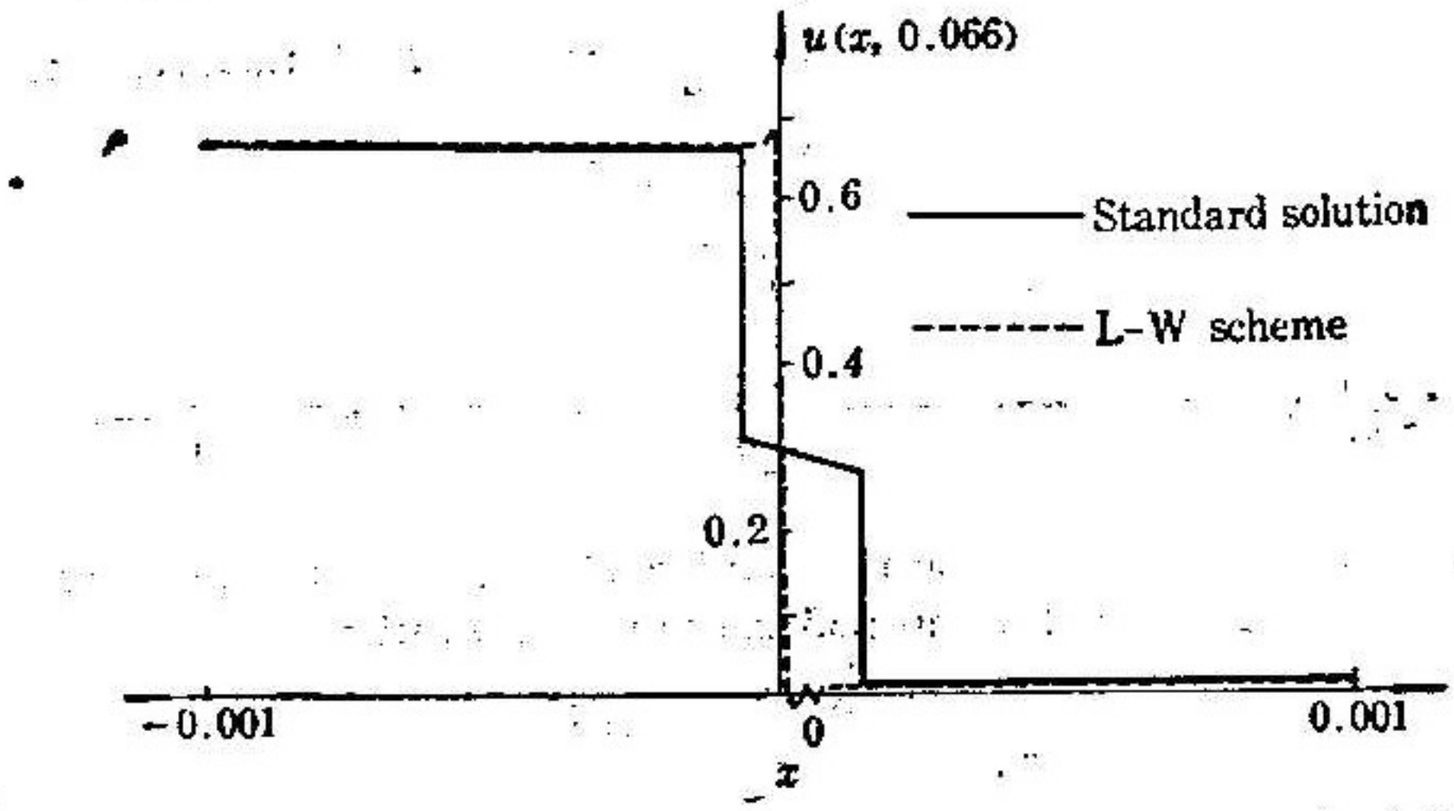

Fig. 9 Comparison between the solution of the L–W scheme (100 equal parts) and the standard solution

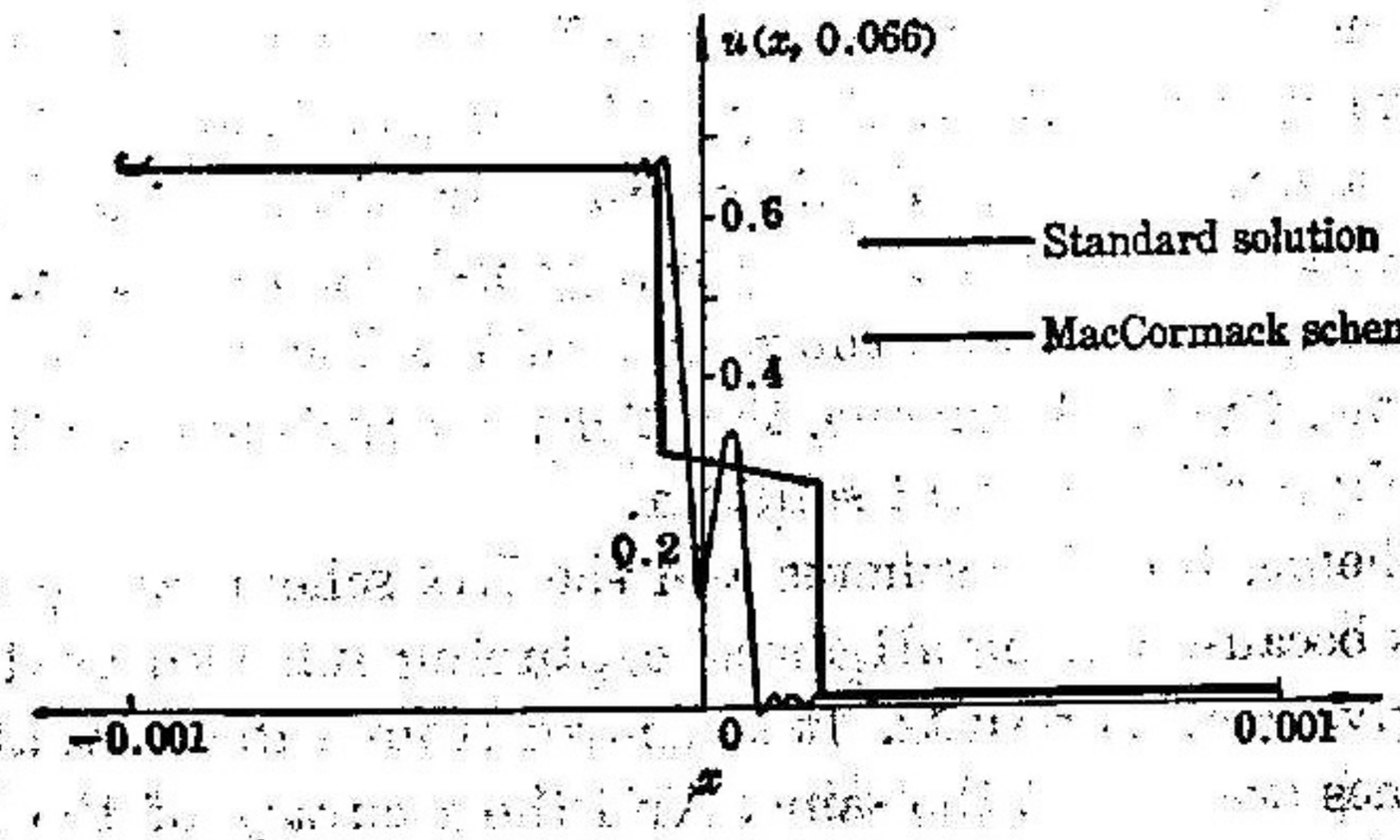

Fig. 10 Comparison between the solution of the MacCormack scheme (100 equal parts) and the standard solution

$$\left(\max_{m}|f'(U_m^k)|\frac{\Delta t_k}{\Delta x}=0.5\right)$$

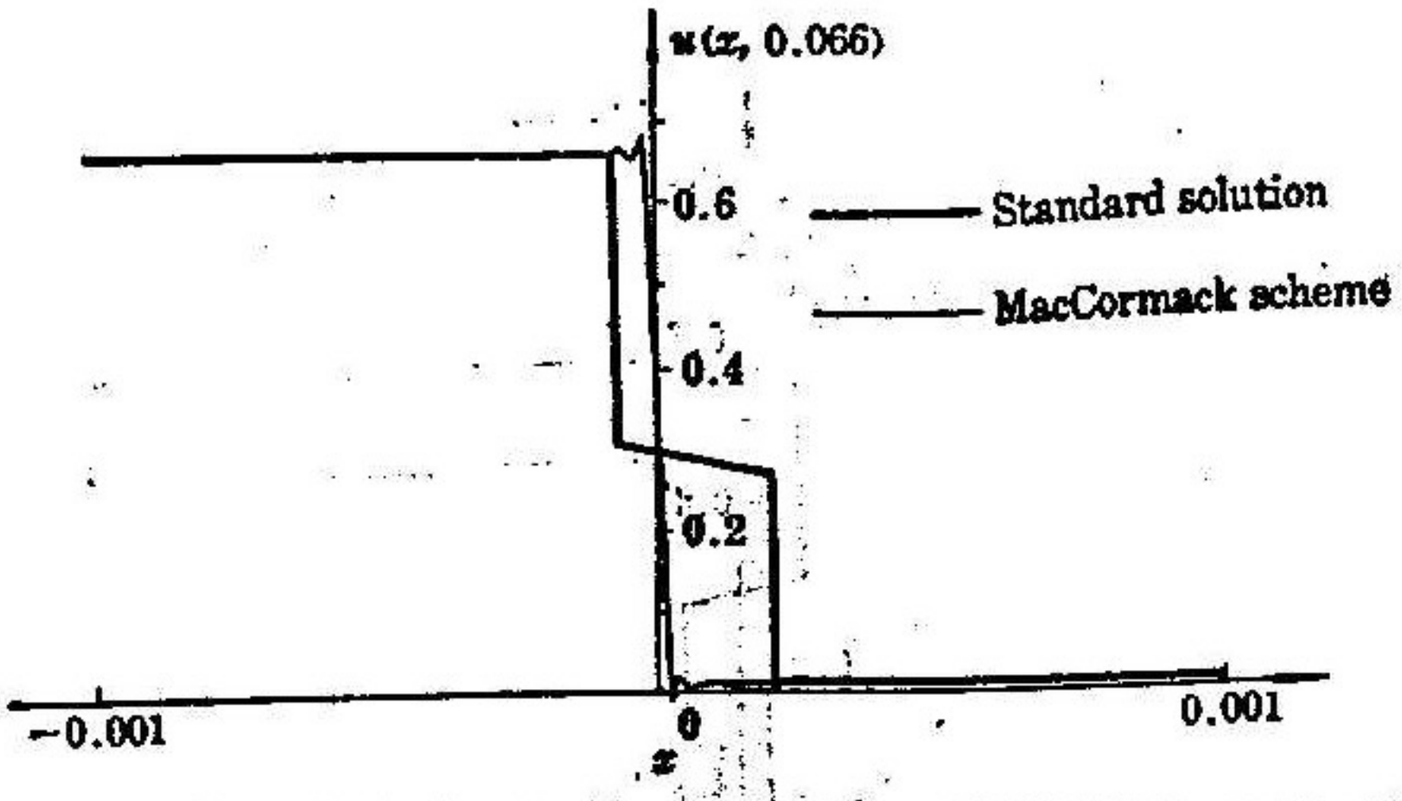

Fig. 11 Comparison between the solution of the MacCormack scheme (100 equal parts) and the standard solution

$$\left(\max_m |f'(U_m^k)| \frac{\Delta t_k}{\Delta x} = 0.75\right)$$

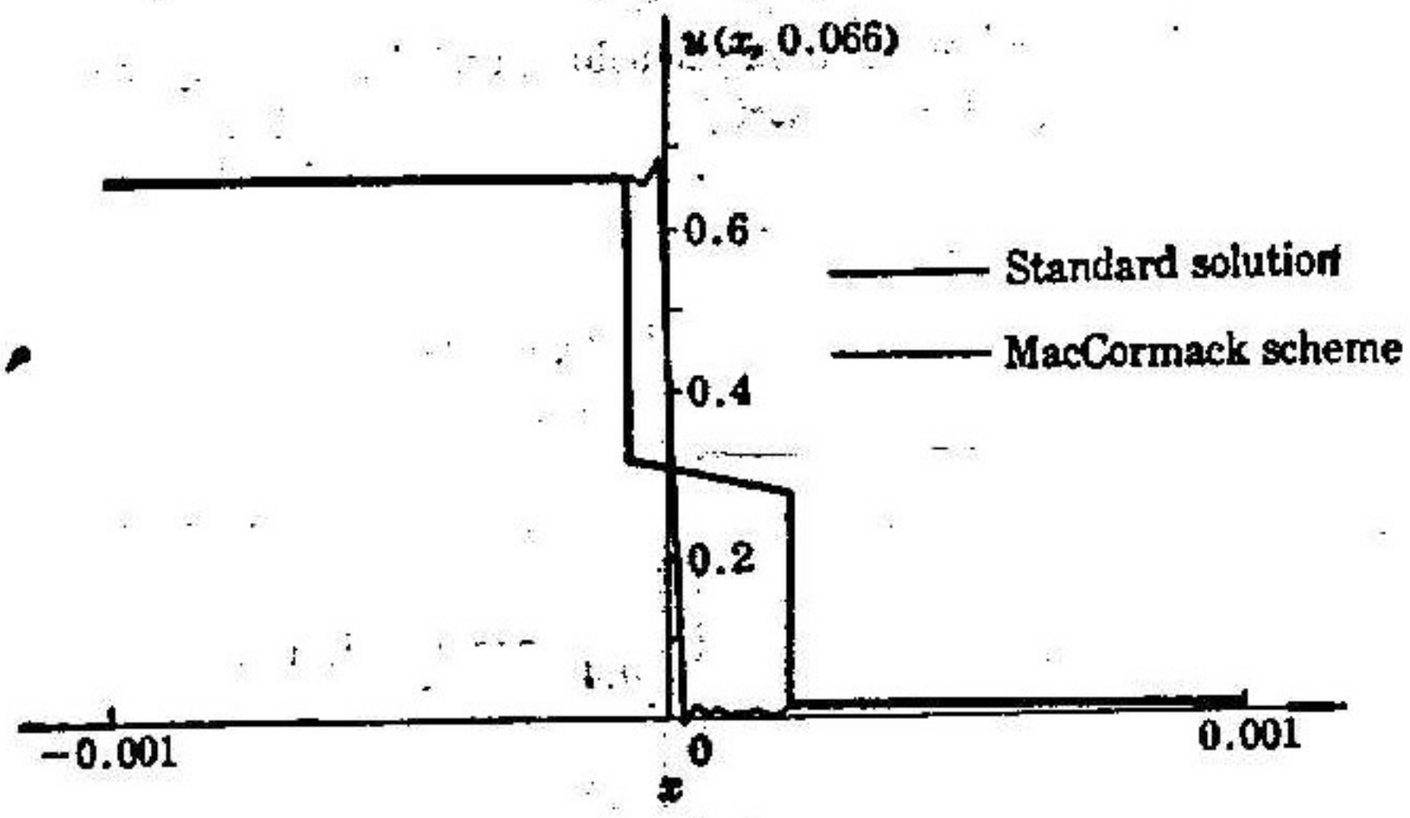

Fig. 12 Comparison between the solution of the MacCormack scheme (100 equal parts) and the standard solution

$$\left(\max_m f'(U_m^k) \frac{\Delta t_k}{\Delta x} = 0.95\right)$$

From Table 1 it is known that the more the mesh points, the less the errors of the results corresponding to the singularity-separating method, the Godunov scheme, the E–O scheme[1] and the Lax scheme. And we see from Figs. 2—4 that the results of the Godunov scheme, the E–O scheme and the Lax scheme are in good agreement with the physically relevant solution. Though the errors of the results corresponding to the Murman scheme decrease slightly if the number of mesh points increases, we know from Fig. 5 that they are not the physically relevant solution. From Table 1 and Figs. 6—12, it is also known that the results of the Courant scheme, the second-order one-sided scheme, the L–W scheme, the Richtmyer scheme and the MacCormack scheme are not the physically relevant solution.

The Godunov scheme, the E–O scheme and the Lax scheme are good schemes for nonconvex problems because among all shock-capturing schemes used in this paper only these schemes give correct results. It is known from Table 1 that the accuracy of the first two schemes are almost the same. And the accuracy of the Lax scheme is lower than those of the other two schemes. For the same accuracy, the number of

1) The result of the "so-called E–O scheme" in [18] is actually a result of the Courant scheme. We made a mistake in that paper.

Table 1 The errors at $t=0.066$ in the sense of L_2-norm

Schemes	Numbers of mesh points	Errors($\|\sigma\|$)
S-S method	15(5 equal parts in each subregion)	0.767×10^{-3}*
	30(10 equal parts in each subregion)	0.19×10^{-3}
	60(20 equal parts in each subregion)	0.48×10^{-4}
Godunov scheme	50	0.365×10^{-1}
	100	0.260×10^{-1}
	200	0.182×10^{-1}
	400	0.142×10^{-1}
	800	0.92×10^{-2}
E-O scheme	50	0.383×10^{-1}
	100	0.281×10^{-1}
	200	0.202×10^{-1}
	400	0.151×10^{-1}
	800	0.103×10^{-1}
	1600	0.71×10^{-2}
Lax scheme	100	0.885×10^{-1}
	200	0.699×10^{-1}
	400	0.547×10^{-1}
	800	0.429×10^{-1}
Murman scheme	50	0.934×10^{-1}
	100	0.960×10^{-1}
	200	0.944×10^{-1}
	400	0.834×10^{-1}
	800	0.727×10^{-1}
Courant scheme	100 $\left(\max_m\|f'(U_m^k)\|\frac{\Delta t_k}{\Delta x}=0.95\right)$	1.029×10^{-1}
	200	1.035×10^{-1}
	100 $\left(\max_m\|f'(U_m^k)\|\frac{\Delta t_k}{\Delta x}=0.75\right)$	1.029×10^{-1}
	200	1.035×10^{-1}
Second-order one-sided scheme	100 $\left(\max_m\|f'(U_m^k)\|\frac{\Delta t_k}{\Delta x}=1.9\right)$	1.406×10^{-1}
	100 $\left(\max_m\|f'(U_m^k)\|\frac{\Delta t_k}{\Delta x}=0.95\right)$	1.407×10^{-1}
	100 $\left(\max_m\|f'(U_m^k)\|\frac{\Delta t_k}{\Delta x}=0.5\right)$	1.407×10^{-1}
L-W scheme	100	1.009×10^{-1}

(Table 1 *Continued*)

Schemes	Numbers of mesh points	Errors($\|\sigma\|$)
Richtmyer scheme	100	0.938×10^{-1}
MacCormack scheme	100 $\left(\max_m \|f'(U_m^k)\|\frac{\Delta t_k}{\Delta x}=0.5\right)$	0.662×10^{-1}
	200	0.689×10^{-1}
	100	0.986×10^{-1}
	200 $\left(\max_m \|f'(U_m^k)\|\frac{\Delta t_k}{\Delta x}=0.75\right)$	1.005×10^{-1}
	400	1.021×10^{-1}
	100	1.008×10^{-1}
	200 $\left(\max_m \|f'(U_m^k)\|\frac{\Delta t_k}{\Delta x}=0.95\right)$	1.005×10^{-1}
	400	1.022×10^{-1}

* For the S–S method the definition of the error is

$$\|\sigma\|=\left[\frac{1}{b-a}\int_a^b (U-U^*)^2dx+(x_1-x_1^*)^2+(x_2-x_2^*)^2\right]^{1/2},$$

where x_1^*, x_2^* are the exact locations of the two shocks and x_1, x_2 are approximate ones.

mesh points in the x–direction needed by the Lax scheme is at least ten times more than that needed by the E–O scheme. Therefore only the E–O scheme is further analysed in the following.

From Table 1, it is known that the error of the result of the S–S method will reduce to about 1/4 of the original one if Δt reduces to one half of the original one, that is, the convergence rate is $O(\Delta t^2)$[15]. And the errors of the results of the Godunov scheme and the E–O scheme will only reduce to about $1/\sqrt{2}$ of the original one in the same case, that is, the convergence rate is $O(\Delta t^{1/2})$. When 1600 equal parts are taken in the x direction, the error of the results of the E–O scheme is 0.71 $\times10^{-2}$. Because the convergence rate is only $O(\Delta t^{1/2})$, $1600\times(7.1)^2\approx80000$ equal parts in the x direction must be taken if we want the error to be equal to 10^{-3}. For the S–S method 5 equal parts in each subregion is enough. At $t=0.066$ there are 3 subregions; this means that 15 parts in the whole computational region are enough. The ratio of the mesh points of two methods for this error is 5300. (At the very beginning there are 5 subregions for the S–S method; so 25 parts are taken in the whole region. Therefore, the ratio for those time levels is 3200.) Because the convergence rate for the S–S method is $O(\Delta t^2)$ and that for the E–O scheme is $O(\Delta t^{1/2})$, for an error equal to 10^{-k} $(k\geqslant3)$, the ratio of the mesh points will be

$$\frac{80000\times(10^{k-3})^2}{15\times(10^{k-3})^{1/2}}\approx5300\times(10^{k-3})^{3/2}.$$

This means that the higher the expected accuracy, the larger the ratio. For example, if $k=5$, the ratio will be 5300000.

The computer–times which are needed for the E–O scheme and the S–S method in different cases are given in Table 2. It is known from the table that for the E–O scheme the CPU time is directly proportional to the square of the number of parts in

Table 2

E–O scheme		S–S method	
Numbers of points	Time	Numbers of points	Time
100 parts	13.42 sec.	15 (5 equal parts in each subregion)	12.44 sec.
		30 (10 equal parts in each subregion)	18.65 sec.
200 parts	47.52 sec.	60 (20 equal parts in each subregion)	37.75 sec.
		120 (40 equal parts in each subregion)	98.57 sec.
400 parts	179.45 sec.	240 (80 equal parts in each subregion)	316.18 sec.
Compilation	6.85 sec.	Compilation	18.59 sec.

the x direction. Because the CPU time for 400 parts is 179.45 sec., it will be 7.2×10^6 sec. for 80000 parts. For this accuracy (10^{-3}), only 12.44 sec. is needed for the S–S method. The ratio of times between the two methods is 5.8×10^5. If the number of the mesh points is great enough, the CPU time will also be directly proportional to the square of the number of parts in the x direction for the S–S method. It is seen that for an error equal to $10^{-k}(k\geqslant3)$, the ratio of the CPU times of the two methods is

$$5.8\times10^5\times\frac{(10^{k-3})^{2\times2}}{(10^{k-3})^{(1/2)\times2}}=5.8\times10^5\times(10^{k-3})^3.$$

Therefore, the ratio grows very fast while the required accuracy raises. For example, if the required error is 10^{-4} or 10^{-5}, the ratio of the CPU times will be 5.8×10^8 or 5.8×10^{11}. However, if the required error is much less than 10^{-3}, the ratio will be much less than 5.8×10^5. Moreover, the error we discuss here is the total error and it comes mainly from the region near the discontinuity lines. Therefore, the number of mesh points for the E–O scheme may greatly reduce if the solution near the discontinuity lines is not so important. In many practical problems the required error is not very small and sometimes the solution near the discontinuity lines is not important. In those cases, the E–O scheme, the Godunov scheme, etc., will give required results by use of a usual grid, i.e., they are quite efficient in those cases.

The errors of the E–O scheme in two computers whose word lengths are 32 and 48 are given in Table 3. The difference between the two errors shows the rounding error on the first computer whose word length is 32. It is seen from the results that if the number of mesh points in the x direction increases by 2^k times, the difference will increase by 3^k times. When 800 parts are taken, the difference is 0.00023. If k parts are taken ($k>800$), the difference will be $0.00023\times3^{\log_2(k/800)}$. The truncation error should be $0.0103/\sqrt{\frac{k}{800}}$. Therefore, the minimal error for this method is

$$\min_k\left\{0.00023\times3^{\log_2(k/800)}+0.0103/\sqrt{\frac{k}{800}}\right\}.$$

It is not difficult to show that this value is nearly equal to 0.007. The above analysis

Table 3

Numbers of mesh points	Errors (32)*	Errors (48)	Differences between two errors	Ratios of differences
100	0.028133	0.028125	0.000008	3.25
200	0.020292	0.020266	0.000026	2.885
400	0.015218	0.015143	0.000075	3.04
800	0.010541	0.010313	0.000228	

* Errors (32) means the errors on a computer whose word length is 32 bits.

tells us that for our problem the minimal error for the E–O scheme in a computer whose word length is 32 is about 7×10^{-5}.

Therefore, the physically relevant solution with some accuracy can be obtained by using the E–O scheme, but it will take much CPU time to raise the accuracy and it will soon almost be impossible to further raise the accuracy because of the rounding error. The main problem is that its convergence rate is only $O(\Delta t^{1/2})$. This problem exists in all shock-capturing methods.

The compilation times of two routines are also given in Table 2. The time for the S–S method is almost three times as much as that for the E–O scheme. This means that the routine for the S–S method is more complicated than that for the E–O scheme. It is the shortcoming of the S–S method. Another shortcoming is that if the number of mesh points are the same, the CPU time needed by the S–S method is 5—10 times as much as that needed by the E–O scheme. This can be remedied by its high accuracy if the S–S method is used, i.e., by the fact that only a few mesh points are needed. In fact for the same accuracy, the CPU time needed by the S–S method is much less than that needed by the E–O scheme. Because the computation amount is small and the rounding error is not so large for the S–S method, it is possible to obtain a result of our problem with an accuracy of 10^{-5} in a computer whose word length is 32 bits.

It is shown in Figs. 5—12 that the non-physically relevant solutions are given by the Murman scheme, the Courant scheme, the second-order one-sided scheme, the L–W scheme, the Richtmyer scheme and the MacCormack scheme. It is worth notice that the weak solutions given by the Murman scheme, the Courant scheme and the second-order one-sided scheme are very "sharp", and the structure of the solutions of both the Courant scheme and the second-order one-sided scheme is not related to the step lengths and the ratio of the step lengths to a certain extent, but the difference among these solutions and the physically relevant solution is large. Now we analyse the reason why the Courant scheme and the second-order one-sided scheme give such results. It is known from Fig. 1 that there are a left contact discontinuity and a right one at $t\approx0$ in the physical picture. After some time there appear a right-facing shock wave in the left region and a left-facing shock wave in the right region. Later the right-facing shock wave interacts with the right contact discontinuity, changing into a new right-facing shock wave; and the left-facing shock wave does so with the left contact discontinuity, changing into a new left-facing shock wave.

It is clear that the direction of the space difference depends on the value of $f'(U_m^k)$ in the Courant scheme and in the second-order one-sided scheme. Therefore, if $f'(U_m^k)>0$, $f'(U_{m+1}^k)<0$ at some point $t=t_k$, $x=x_m$ and if the difference between U_m^k and U_{m+1}^k is large, then the sharp jump of U would remain until there appears some k^* for which the condition $f'(U_m^{k^*})>0$ or $f'(U_{m+1}^{k^*})<0$ is not satisfied, no matter how big the step lengths and the ratio of the step lengths are. In this problem the initial value at $t=0$ has a jump at the point $x=0$

$$U|_{x=0-}=0.656,\quad U|_{x=0+}=0.014,$$

and it is known from Fig. 13 that $f'(0.656)>0$ and $f'(0.014)<0$. At $t>0$, the numerical solutions of the two schemes will keep the above properties, so the numerical solutions of these methods would keep the sharp jump at the point $x=0$. That is, the left contact discontinuity and the right one, which exist in the physically relevant solution, do not appear for these methods.

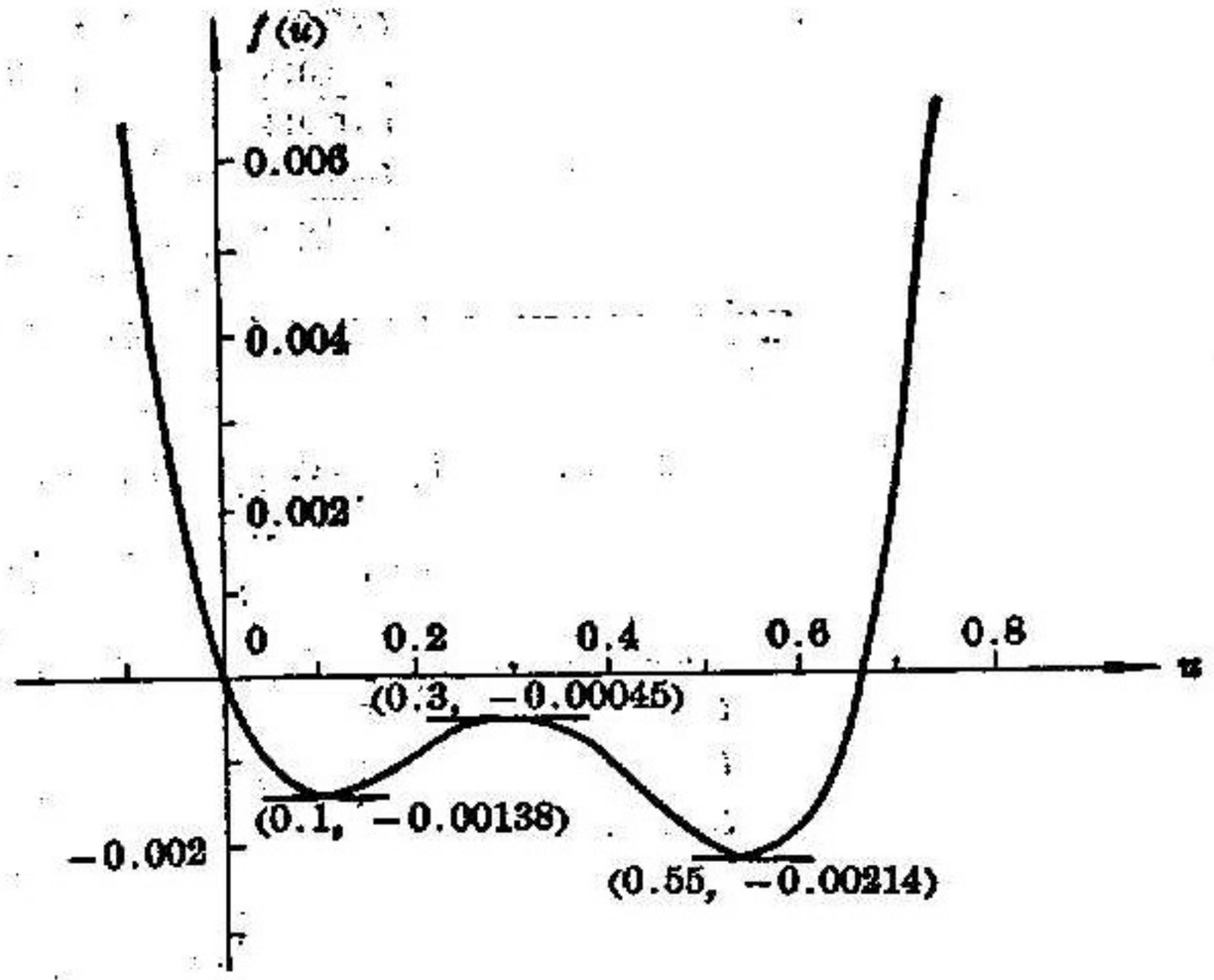

Fig. 13 The equation of state
$f(U)=\frac{U^4}{2}-\frac{19}{30}U^3+\frac{1}{4}U^2-\frac{33}{1000}U$

As mentioned above, after some time there appear gradually two shocks in the right and the left regions. Fig. 14 and Fig. 15 show that the two shocks can be obtained if we use these schemes. In the results of the Courant scheme the values of U both in front of and behind the left shock wave are larger than 0.55. And it is known from Fig. 13 that the corresponding values of f' are all larger than zero. Therefore, the shock wave continues to move to the right until it meets the discontinuity line $x=0$. The values of U both in front of and behind the right shock wave are less than 0.1; so the corresponding values of f' are all less than zero. Thus, the shock wave continues to move to the left until it meets the discontinuity line $x=0$. That is, the solution of the Courant scheme possesses the structure shown in Fig. 16. Because the discontinuity at the point $x=0$ does not satisfy the entropy condition, this solution is not a physically relevant solution. In the results of the second order one-sided scheme at $t=0.012$ for a certain $m(x\approx -0.0005)$, $U_m^k>0.55$ and $U_{m+1}^k<0.55$ (for the correct solution both U_m^k and U_{m+1}^k should be larger than 0.55) and the difference between U_m^k and U_{m+1}^k is very large. Moreover, it is known from Fig. 13 that $f'(U_m^k)>0$, and $f'(U_{m+1}^k)<0$. It is clear that one-sided schemes often make such discontinuities keep sharp and immovable. Therefore, there appears an immovable shock in the left region which should originally move to the right. At $t=0.018$ for a certain $m(x\approx 0.0005)$, $U_m^k>0.1$, $U_{m+1}^k<0.1$ (for the correct solution both U_m^k and U_{m+1}^k should be less than 0.1) and the difference between U_m^k and U_{m+1}^k is very large. It is known from Fig. 13 that there again appear $f'(U_m^k)>0$ and $f'(U_{m+1}^k)<0$. Therefore, in the right region there also appears a stationary shock which should originally move to the left. That is, the solution given by the second-order one-sided scheme has three

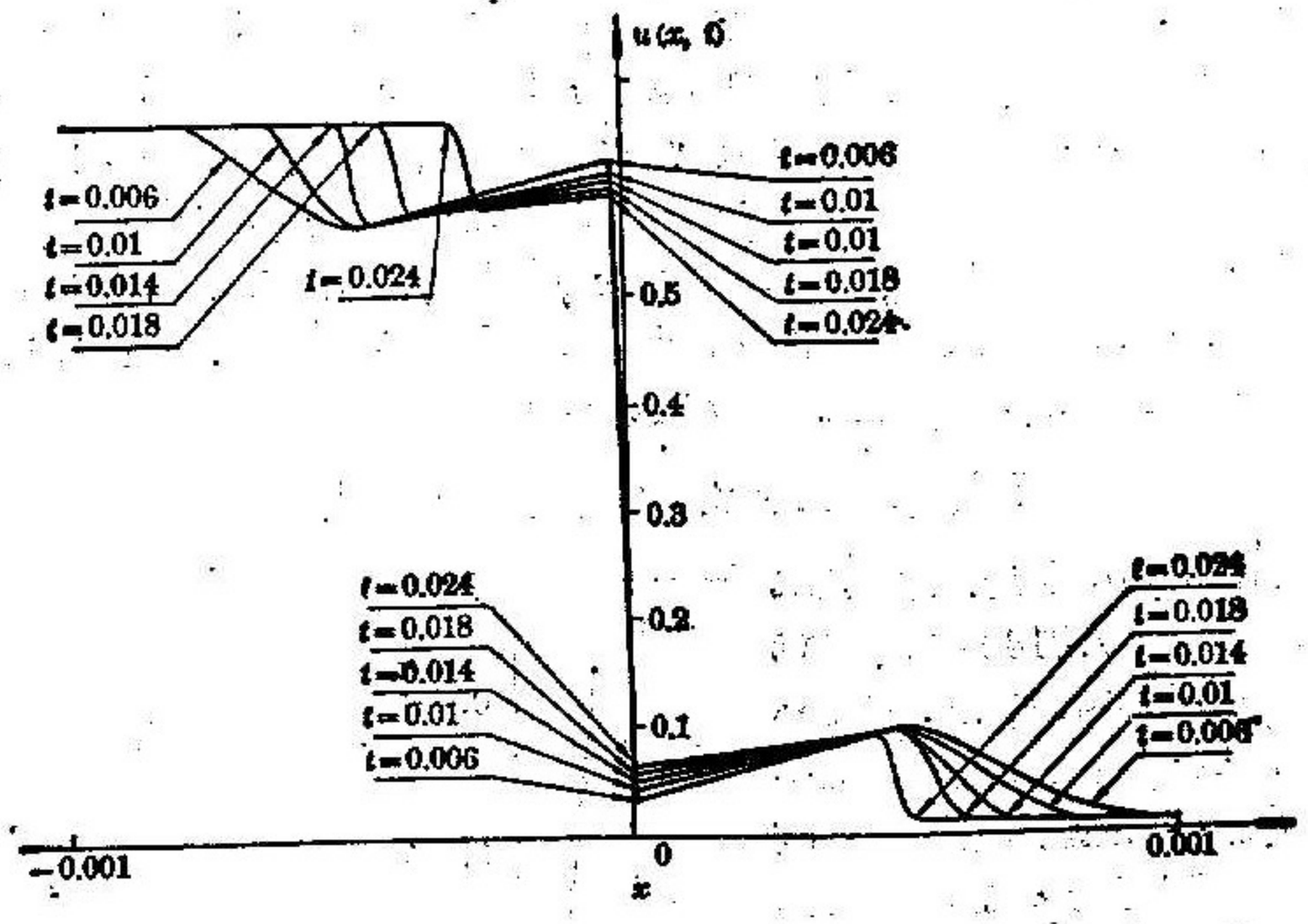

Fig. 14 The results of the Courant scheme at different times

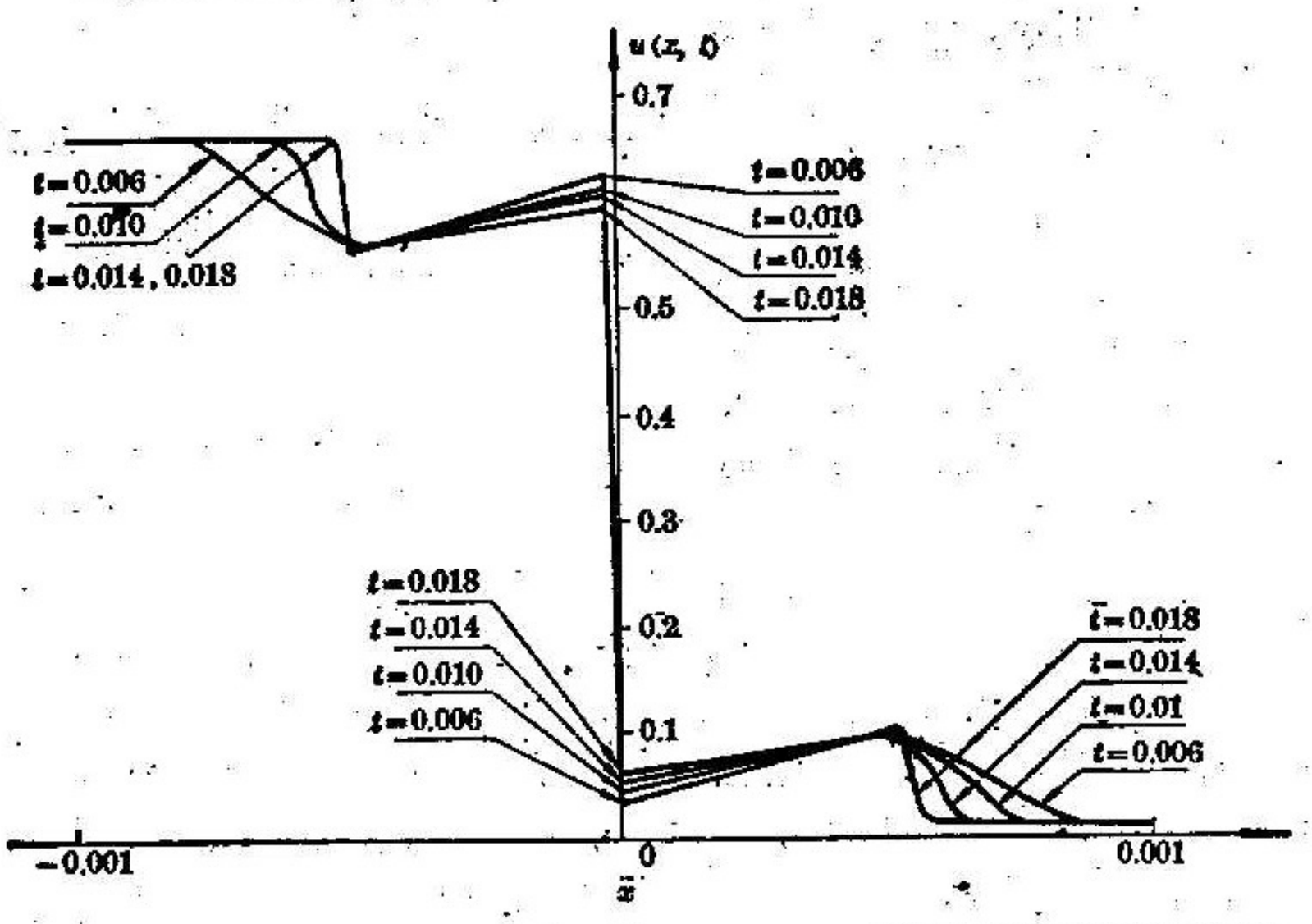

Fig. 15 The results of the second-order one-sided scheme at different times

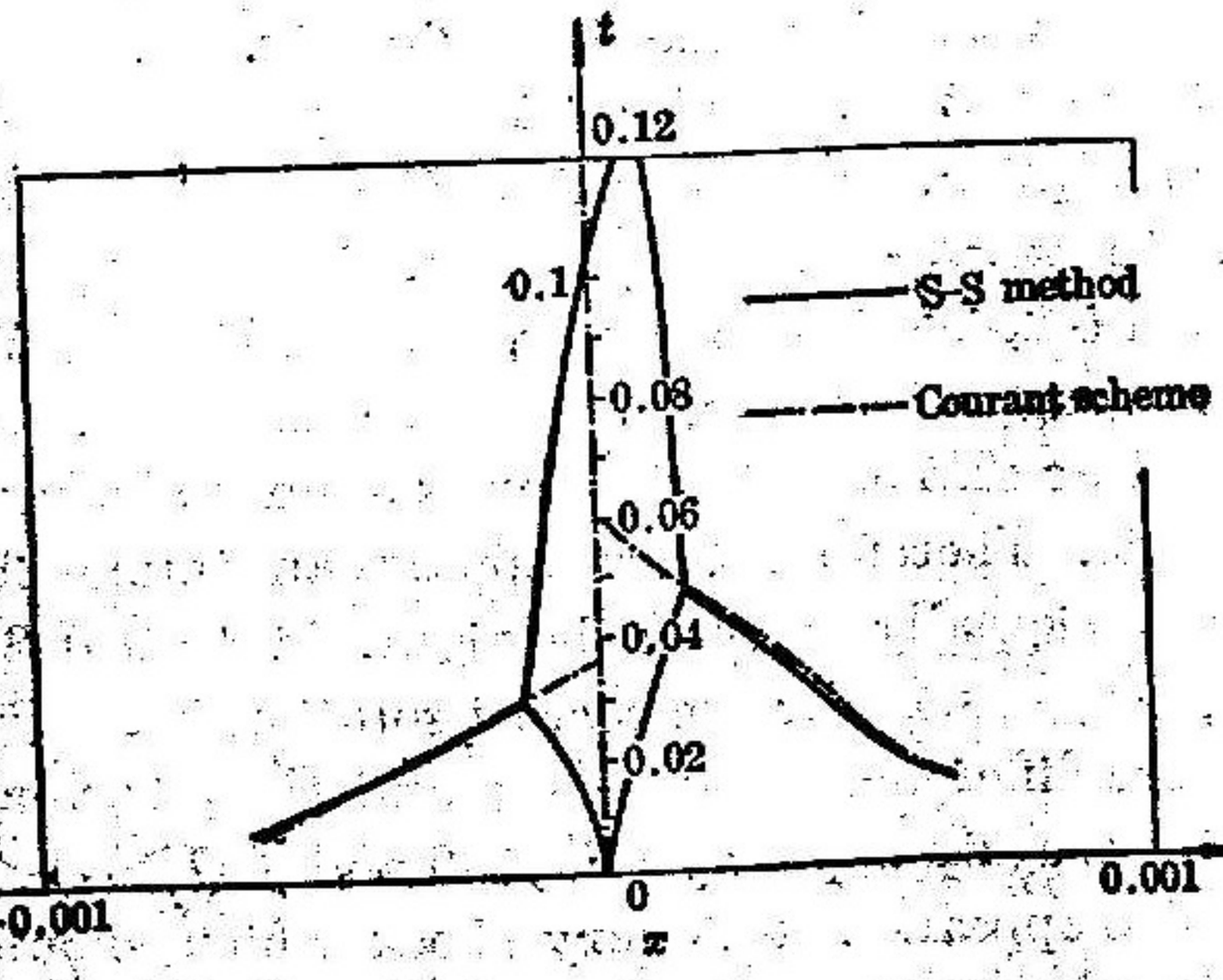

Fig. 16 Comparison between discontinuities of the Courant scheme and the S-S method

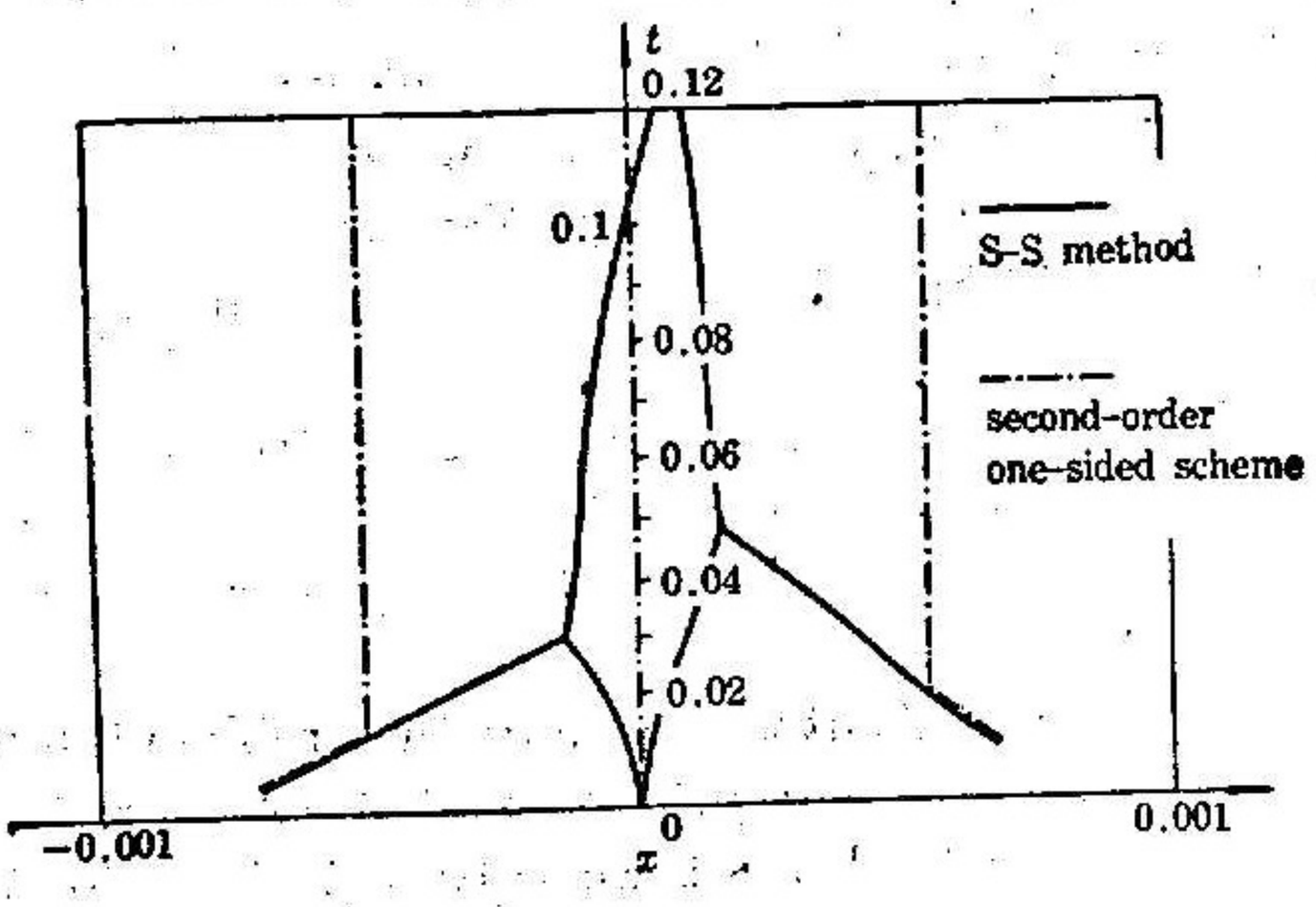

Fig. 17 Comparison between discontinuities of the second-order one-sided scheme and the S-S method

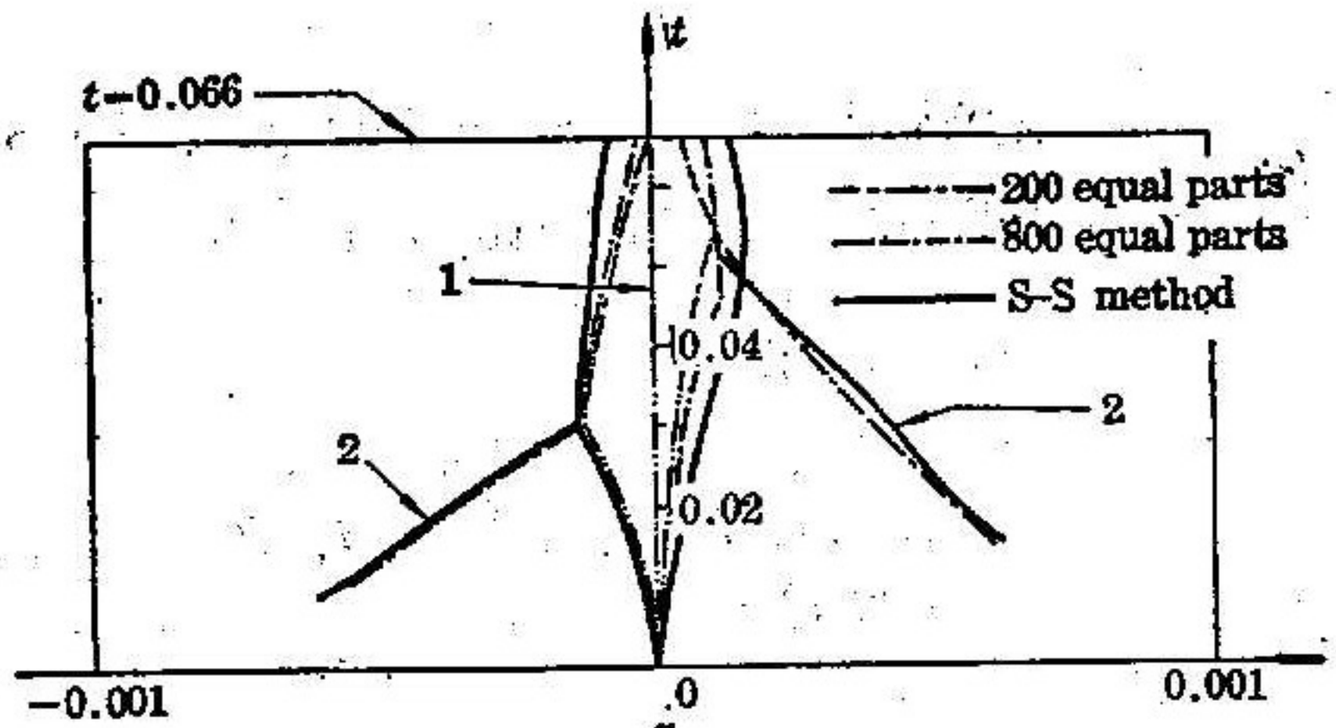

Fig. 18 Comparison between discontinuities of the Murman scheme (800 and 200 equal parts) and the S-S method
(1) The result of 800 equal parts and that of 200 equal parts overlap;
(2) The result of 800 equal parts and that of the S-S method overlap.

sharp discontinuity lines as shown in Fig. 15 and it is also far from the physically relevant solution. In order to test whether the structure of these results is related to the step lengths or the ratio of step lengths, for the case $\max_m |f'(U_m)| \frac{\Delta t}{\Delta x} = 0.95$, different numbers of mesh points, including 100, 200, 400, 800, 1600, are taken and when the number of mesh points is 100, different ratioes of step lengths, including $\max_m |f'(U_m)| \frac{\Delta t}{\Delta x} = 0.5$, 0.95, 1.9, are taken. All the results give the same physical picture (see Fig. 17).

Though the Murman scheme is basically a one-sided scheme, the different result from those of the Courant scheme and the second-order one-sided scheme is obtained. This is because its algorithm in the place where f' changes its sign is different from the Courant scheme. From Fig. 5 and Fig. 13, it is known that for its result there is only one transitional point on the shock wave and that there appears a rarefaction shock wave. Fig. 18 gives the physical picture of the solution of the Murman scheme when different step lengths are taken. The figure shows that the result is closer to the physically relevant solution than those of the Courant scheme and the second-order one-sided scheme.

As well known, the main problem of the Richtmyer scheme is that there usually appears quite strong oscillation. f' increases monotonously when $U>0.55$ and $U<0.1$; so $\max|f'|$ must be very large when the strong oscillation appears. By the Courant condition, Δt must be very small in this case, and the computation has to be stopped sometimes.

It is seen from Figs. 9—12 that the phenomenon that a shock splits up into two discontinuities does not appear in the results of the L–W scheme and the MacCormack scheme. Figs. 10—12 also show that different solutions are given by the MacCormack scheme with different step lengths.

Finally, it must be pointed out that because the problem is much more complicated, it is difficult to give an exact analytic expression for $U^*(x, t)$. In the process of computing the errors given in Table 1 we substitute the result given by the S–S method with a grid of 40 equal parts in each subregion for the exact solution. In what follows, we shall show that such a substitution does not influence the data given in Table 1.

Because we take $\frac{\Delta t}{\Delta \xi}=\text{const.}$ in the S–S method and the S–S method has the second order accuracy, the following method can be used to estimate the exact solution accurately.

Let the exact solution of the problem be $U^*(x, t)$ and the solution obtained by our difference scheme be $\bar{U}(x, t, \Delta t)$. Suppose that $\bar{U}$ tends to U^* at a convergence rate of $O((\Delta t)^2)$ in the maximum norm, i.e., $\bar{U}(x, t, \Delta t)-U^*(x, t)=c(x, t, \Delta t)\Delta t^2$. We also suppose $c(x, t, \Delta t)=a(x, t)+b(x, t)\Delta t+O((\Delta t)^2)$. Therefore there is the following formula

$$\bar{U}(\Delta t)=U^*+a(\Delta t)^2+b(\Delta t)^3+O((\Delta t)^4).$$

Clearly, for $\bar{U}(\Delta t^*)$, $\bar{U}\left(\frac{\Delta t^*}{2}\right)$, $\bar{U}\left(\frac{\Delta t^*}{4}\right)$, which stand for the difference solutions corresponding to the grids of 10, 20, 40 equal parts in each subregion respectively, we have the relations

$$\bar{U}(\Delta t^*)=U^*+a(\Delta t^*)^2+b(\Delta t^*)^3+O((\Delta t^*)^4),$$

$$\bar{U}\left(\frac{\Delta t^*}{2}\right)=U^*+a\left(\frac{\Delta t^*}{2}\right)^2+b\left(\frac{\Delta t^*}{2}\right)^3+O((\Delta t^*)^4),$$

$$\bar{U}\left(\frac{\Delta t^*}{4}\right)=U^*+a\left(\frac{\Delta t^*}{4}\right)^2+b\left(\frac{\Delta t^*}{4}\right)^3+O((\Delta t^*)^4);$$

so there is the following relation

$$U^*=\frac{32}{21}\bar{U}\left(\frac{\Delta t^*}{4}\right)-\frac{12}{21}\bar{U}\left(\frac{\Delta t^*}{2}\right)+\frac{1}{21}\bar{U}(\Delta t^*)+O((\Delta t^*)^4).$$

According to this expression and our numerical results, we have the estimate

$$\left\|U^*-\bar{U}\left(\frac{\Delta t^*}{4}\right)\right\|\approx 0.000005.$$

Therefore though our result is not the exact solution, the error caused by substituting $\bar{U}\left(\frac{\Delta t^*}{4}\right)$ for U^* does not influence the correctness of the values given in Table 1 because

$$\|U^*-\bar{U}\|\leqslant\left\|U^*-\bar{U}\left(\frac{\Delta t^*}{4}\right)\right\|+\left\|\bar{U}\left(\frac{\Delta t^*}{4}\right)-\bar{U}\right\|,$$

where $\bar{U}$ represents a difference solution obtained by one of the schemes in Table 1. However, there is an exception. If we substitute $U\left(\frac{\Delta t^*}{4}\right)$ for U^*, the second significant digit of the error of the S–S method with a grid of 20 equal parts in each subregion will not be correct since the error itself is very small. In order to give a correct value of the error, we substitute $\bar{U}\left(\frac{\Delta t^*}{8}\right)$ for U^* in this case.

We have had a useful discussion with Prof. Teng Zhen-huan on some problems of this paper. The authors wish to express their appreciation to him.

References

[1] A. Harten, J. M. Hyman, P. D. Lax, B. Keyfitz, On finite-difference approximation and entropy condition for shocks, *Comm. Pure and Appl. Math.*, **29** (1976), 297—322.

[2] Zhou, Y. l., Li, D. y., Gong, J. f., The calculation of physical solutions of quasilinear equations of the first order, *Journal on Numerical Methods and Computer Applications*, Vol. 1, No. 1, 1980, 16—25.

[3] Zhu, Y. l., Zhong, X. c., Chen, B. m., Zhang, Z. m., Difference Methods for Initial-Boundary-Value Problems and Flow around Bodies, Science Press, Beijing, 1980.

[4] S. K. Godunov, A finite-difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics, *Mat. Sb.*, **47** (1959), 271—290.

[5] B. Engquist, S. Osher, Stable and entropy-condition-satisfying approximations for transonic flow calculations, *Math. Comp.*, **34** (1980), 45—75.

[6] P. D. Lax, Weak solution of non-linear hyperbolic equations and their numerical computation, *Comm. Pure Appl. Math.*, **7**:1 (1954), 159—193.

[7] E. M. Murman, Analysis of embedded shock waves calculated by relaxation methods, *AIAA J.*, **12** (1974), 626—633.

[8] R. Courant, E. Isaacson, M. Rees, On the solution of nonlinear hyperbolic differential equations by finite differences, *Comm. Pure Appl. Math.*, **5** (1952), 243—255.

[9] Zhu, Y. l., An application of uncentered schemes to computation of unsteady flows, *Journal on Numerical Methods and Computer Applications*, **1**:4 (1980), 239—242.

[10] P. D. Lax, B. Wendroff, Systems of conservation laws, *Comm. Pure Appl. Math.*, **13**:2 (1960), 217—237.

[11] R. W. MacCormack, The effects of viscosity in hypervelocity impact cratering, *AIAA Paper*, 69-354, 1969.

[12] R. D. Richtmyer, A survey of difference methods for non-steady fluid dynamics, NCAR TN 63-2, 1962.

[13] Zhu, Y. l., The singularity-separating method, The Proceedings of the Fourth International Symposium on Finite Element Methods in Flow Problems, July 26—29-th, 1982.

[14] S. Osher, Riemann solvers, the entropy condition, and difference approximations (submitted to SINUM).

[15] Zhu, Y. l., Stability and convergence of difference schemes for linear initial-boundary-value problems, *Mathematicae Numericae Sinica*, **4**:1 (1982), 98—108.

[16] Wu, X. h., Zhu Y. l., A scheme of the Singularity-Separating method for the nonconvex problem. (to appear)