A FINITE ELEMENT METHOD FOR VARIOUS KINDS OF INITIAL VALUE PROBLEMS

Fumio Kikuchi*

Yoshio Ando**

University of Tokyo

Tokyo, Japan

This paper presents some basic considerations to a most fundamental finite element scheme for initial value problems. The stability and the convergence of the approximate solution are obtained theoretically under some fundamental assumptions on the spatial operator, and it is shown that this scheme is applicable to wide and important classes of evolution equations. Numerical experiments are also performed to various kinds of evolution equations in order to demonstrate the validity of the present method.

SYMBOLS

H	: a real or complex Hilbert space (usually L_2)
X	: a Hilbert space in which the problem is considered
A	: an operator in X; $A{:}X \quad X$
D(A)	: domain of A; $D(A) \quad X$
R(A)	: range of A; $R(A) \quad X$
u, v etc.	: elements of X; if u is a vector with m components, it is designated as follows; $u = (u^1, \ldots\ldots , u^m)^T$ (T: transpose)
(,)	: inner product of X; sometimes $(\; , \;)_X$ is also employed.
$\| \;\; \|$	: norm of X; sometimes $\| \;\; \|_X$ is also employed.
B	: a self-adjoint operator in H; it is often positive bounded below.
H_B	: derived energy space by a self adjoint positive bounded below operator B.
[,]	: inner product of H_B (energy product)
$\vert \;\; \vert$	: norm of H_B (energy norm)
S_n	: subspace of X in which approximate solutions are sought (n=1,2,. . .)
u_n, v_n, etc.	: elements of S_n
A_n	: continuous approximation operator of A; $A_n{:}\ S_n \quad S_n$
t	: time
Δt, Δx	: time and spatial mesh sizes
$\underset{\sim}{\theta}$	: parameter; $0.0 \leqq \theta \leqq 1.0$
$\underset{\sim}{t}$	: $t + \theta \; \Delta \; t$
u	: $\theta u(t+\Delta t) + (1-\theta)\, u\, (t)$
u_n^*	: reference element in S_n for the check of consistency

*Graduate Student, Graduate School of Engineering

**Professor, Department of Nuclear Engineering, Faculty of Engineering

I. INTRODUCTION

Recently, finite element methods have been employed not only in structural problems but also in non-structural ones, and a range of their applicability is enlarged from linear boundary value problems and eigenvalue ones to nonlinear problems and initial ones. In fact, equilibrium or steady states in nature can exist only approximately and linearity is also idealization of nonlinearity. Thereforce, these effects should be taken together and analyzed at the same time in rigorous treatment, and as a matter of course much consideration should be now given to nonlinear and dynamic problems in the finite element methods (References 1 and 2).

However, the foundation of such analysis does not seem to be sufficiently developed, though some basic investigations have already been made mainly in wave and heat equations (References 3–11). For example, the instability of numerical solutions which one encounters in the analysis of dynamic problems by the finite element methods is an important thing that must be solved, and the situation will become more complicated if nonlinearity is introduced. It is also to be noted that the finite element methods should be valid irrespective of types of differential equations such as hyperbolic and parabolic, because different kinds of phenomena must be treated at the same time in the analysis of general complex systems.

This paper is intended to give some basic considerations and to present a unified treatment to a certain extent to a fundamental class of linear and/or nonlinear initial value problems from the above-mentioned standpoint.

The finite element method for initial value problems heretofore developed can be probably classified as follows, where the space is discretized by the usual finite element technique in all cases;

(1) finite difference methods: time is discretized by suitable finite difference schemes. (References 3–7, 9–11).

(2) direct generalization of the finite element methods in space: time and space is simultaneously discretized by similar methods to the usual finite element methods (Reference 12).

(3) some special methods based on the linearity or other character of the considered systems; mode-superposition methods and Laplace transform are examples of such methods.

(4) other methods: for example, combination of (1) and (2) may be sometimes possible. (Reference 13).

Though the methods belonging to (3) are very convenient for special classes of problems and have been often employed, they have common fatal defect of difficult application to nonlinear and time-varying systems. On the other hand the methods belonging to (1) and (2) can be used to nonlinear problems at least formally.

This paper first presents some considerations to the methods belonging to (2), and then discusses mainly a special fundamental type of method belonging to (1) both theoretically and experimentally. That is, the popular method in which the first order finite difference quotient used in the discretization of the time is exclusively discussed. This method has often been employed in special classes of problems, but some extension is made so that it can deal with wider classes of problems by imposing some fundamental properties on the spatial operator independent of linearity and nonlinearity of the problem. As a result, it is shown that the present method is applicable to fairly wide classes of evolution equations such as heat, wave, and Schrodinger equations, and unified treatment becomes possible to a certain extent. Furthermore, it is also shown that the class of problems treated in this formulation nearly corresponds to the class to which the theory of nonlinear semigroups in Hilbert space can be applied.

Numerical experiments are also conducted to some evolution equations in order to show the validity of the present theory, and it is demonstrated that the present method is actually applicable to various kinds of initial value problems.

In the course of the present theory, the elemental part of the functional analysis is used, and the concepts and techniques of the finite difference methods are also introduced if they are considered to be necessary.

II. PRELIMINARIES

In this section, some preliminaries are given briefly in preparation for the subsequent sections.

II — 1. General

Let X (or sometimes H) be a real or complex separable Hilbert space and the inner product and the norm of X be denoted by (,) and ‖ ‖ respectively (Reference 14).

Let A be an operator with both its domain and range in X;

$$A : X \qquad X \tag{1}$$

A is generally unbounded and nonlinear. The domain and the range of A are designated by D(A) and R(A) respectively. Boundary conditions associated with A are assumed to be homogeneous in the theoretical treatment of this paper. In the present theory, A is required to satisfy the following condition;

$$\text{(CA)} \quad \mathrm{Re}(Au - Av,\ u - v) > \quad \|u - v\|^2 \quad \text{for} \quad u,\ v \epsilon D(A) \tag{2}$$

where is a real constant independent of u and v. If ≥ 0 (>0), then A is called accretive (strictly accretive). In such cases, –A, is called dissipative (strictly dissipative) (References 14–17).

The condition (CA) can be regarded as a kind of stability condition. That is, the system is stable if A is strictly accretive, and, even if ≤ 0, it can be stabilized by adding λI to A, where I is the identity operator in X and $\lambda > -$. Therefore, this condition may not be sufficiently general but can be regarded as fundamental.

If X is a general Banach space in which the inner product is no longer available, (CA) is replaced by the next condition;

$$\text{(CA')}\ \| u - v + \lambda\,(Au - Av) \| \geqq (1 + \quad \lambda)\, \| u - v \| \quad \text{for} \quad u,\ v \epsilon D(A) \qquad (3)$$

where λ is an arbitrary positive number that satisfies $1 + \quad \lambda > 0$.

Let S_n be a finite-dimensional subspace of X ($n = 1,2, \ldots\ldots$). It is not necessarily required that $S_n \quad S_m$ if $n < m$, but the next condition is always required;

for $u \epsilon X$, at least one $u_n \epsilon S_n$ can be chosen for each $n (= 1,2, \ldots)$ in such a way that

$$\lim_{n \to \infty} \| u_n - u \| = 0 \qquad (4)$$

Let us consider a continuous approximation operator of A;

$$A_n : S_n \qquad S_n, \quad D(A_n) = S_n \qquad (5)$$

It is required that

for $u \epsilon D(A)$, at least one $u_n \epsilon S_n$ can be chosen for each n such that

$$\lim_{n \to \infty} \| A_n u_n - Au \| = 0, \qquad \lim_{n \to \infty} \| u_n - u \| = 0 \qquad (6)$$

Moreover, A_n conserves (CA) in S_n;

$$\mathrm{Re}(A_n u_n - A_n v_n,\ u_n - v_n) \geqq \quad \| u_n - v_n \|^2 \quad \text{for} \quad u_n,\ v_n \epsilon S_n \qquad (7)$$

where is not necessarily equal to , but it is independent of n, and $\quad \geqq 0$ ($\quad > 0$) if $\quad \geqq 0$ ($\quad > 0$).

The existence of such subspaces and approximation operators is assumed in this paper, but the condition that X is a separable Hilbert space is usually necessary.

II – 2. Approximate method for boundary value problems

Let us consider the following boundary value problem;

$$\text{(BVP)} \qquad Au = f \qquad \text{for } f \epsilon X \qquad (8)$$

where A is the operator given in II – 1 and it is assumed to be strictly accretive in this case. The uniqueness and existence of the solution are assumed here, but Galerkin's method is often employed as a mean of constructive proofs (Reference 15).

The approximate equation to (BVP) is given as follows by use of A_n and S_n introduced in <u>II – 1;</u>

$$A_n u_n = f_n \qquad \text{for} \qquad f_n \in S_n \tag{9}$$

where f_n satisfies the following condition;

$$\lim_{n \to \infty} \| f_n - f \| = 0 \tag{10}$$

The uniqueness of u_n in (9) is apparent because $\quad > 0$, and its existence can be proved by use of Brouwer's fixed point theorem because A_n is continuous in S_n (Reference 15). In linear problems, the existence is derived from the uniqueness.

A simple error estimation of u_n in (9) is given as follows;

$$\| u_n - u \| = \min_{u^*_n \in S_n} (\| u^*_n - u \| + {}^{1} \quad \| A_n u^*_n - f_n \|) \tag{11}$$

This estimation is exact because u^*_n can be taken as u_n and the second term in the right side of (11) vanishes for u_n. The convergence is assured if the conditions (6), (7) and (10) hold, where the next relation is used;

$$\| A_n u^*_n - f_n \| \leqq \| A_n u^*_n - f \| + \| f - f_n \| \tag{12}$$

The formula (11) is useful if a simple u^*_n can be easily found. In the finite difference methods, the restriction of u of (8) to nodal points is usually used as u^*_n, but its analog does not seem to be always effective in the finite element method.

In the finite element method, S_n is usually constructed as follows;

(a) $S_n \quad D(A)$

(b) for $\quad u \in D(A)$, at least one $u_n \in S_n$ can be chosen for each n such that

$$\lim_{n \to \infty} \| Au_n - Au \| = 0, \qquad \lim_{n \to \infty} \| u_n - u \| = 0 \tag{13}$$

Then A_n and f_n are decided as follows;

$$A_n = P_n A, \qquad f_n = P_n f \tag{14}$$

where P_n is the projection operator from X into S_n. In this case, the conditions (6) and (10) are assured because $\| P_n \| \leqq 1$ and (13) holds, and in (7) can be taken as because

$$\mathrm{Re}(A_n u_n - A_n v_n, u_n - v_n) = \mathrm{Re}(Au_n - Av_n, u_n - v_n) \geqq \quad \| u_n - v_n \|^2 \tag{15}$$

for $\quad u_n, v_n \in S_n$. The continuity of A_n is assured if A is hemi-continuous (Reference 15).

Sometimes, the present conditions for S_n and A_n are too strong for practical uses. If A is a self-adjoint positive bounded below operator in X (or H), then the theory of energy space can be employed and the conditions (13) can be replaced by the following ones;

(a') $S_n \quad H_A$

(b') for $u \epsilon H_A$, at least one $u_n \epsilon S_n$ can be chosen for each n such that

$$\lim_{n \to \infty} | u_n - u | = 0 \tag{16}$$

where H_A is the energy space derived by A, and $| \quad |$ is the energy norm. A_n is defined as follows;

$$(A_n u_n, v_n) = [u_n, v_n] \qquad \text{for} \quad u_n, v_n \epsilon S_n \tag{17}$$

where [,] is the energy product (References 18 and 19). In this case can be taken as again. A simple error estimation is, as is well known, given as follows;

$$| u_n - u | = \min_{u_n^* \epsilon S_n} | u_n^* - u | \qquad (u, u_n = \text{solutions of (8) and (9) respectively})$$

$$\| u_n - u \| \leqq | u_n - u | / \tag{18}$$

Therefore, u_n itself can be used as u_n^* in (6), though it seems meaningless for the present problem (BVP).

The theory presented in this section is an abstract theory of approximate methods for (BVP) in Hilbert spaces, and it is effective not only to the finite element methods but also to some other methods such as the finite difference methods. It is also to be noted that the non-conforming solutions in the (pseudo-) finite element methods are still valid if the corresponding A_n satisfies the above-mentioned conditions.

III. INITIAL VALUE PROBLEMS AND SOME APPROXIMATE METHODS

In this section, the evolution equation to be treated in this paper is introduced and some approximate methods for it are given.

III – 1. Initial value problem

In this paper, initial value problems are treated in the form of the following evolution equation;

$$\text{(IVP)} \quad \begin{aligned} &\frac{du}{dt} + A(t)\, u = f(t) \quad , \qquad 0 \leqq t \leqq T \\ &u(0) = u_0 \end{aligned} \tag{19}$$

where t indicates time and T is a positive constant which can be usually taken arbitrarily. It is assumed that $u_0 \epsilon D(A)$ and $f(t) \epsilon X$ at each t. A(t) is an operator in X at each t and is admitted to depend on t smoothly. Furthermore, A(t) satisfies (CA) at each t, and is dependent only on T. If (IVP) is considered in $0 \leqq t < \infty$, then is assumed to be constant. The term f(t) is separated from A(t)u for convenience sake.

If some subsidiary conditions are imposed on A, f and u_0, then the well-posedness of (IVP) is assured by the theory of nonlinear semi-groups (References 16 and 17). The unified treatment of nonlinear evolution equations is available to the extent of the above-mentioned classes, as far as the authors know. It is also to be noted that representative initial value problems in mathematical physics can be included to the present classes as will be seen in Section V.

In order to solve (IVP) approximately, the following discretizations are generally needed;

(a) discretization with respect to t

(b) discretization with respect to X

In the subsequent sub-sections, some approximate methods will be shown briefly.

III – 2. Boundary value techniques for initial value problems

Generally speaking, initial value problems can be interpreted as a kind of generalized boundary value problems if the initial conditions are regarded as boundary values. Therefore, the method given in II – 2 can be employed to them if they satisfy the conditions given there. In such treatment, the Hilbert space should be constructed on X x [0,T].

Of course, the popular step by step methods are usually more convenient because less numbers of unknowns are needed at each step. However, the boundary value techniques may be effective to periodic and short transient problems, in which long-time calculation is not necessary. Moreover, this approach can be regarded as a direct extension of the usual finite element methods based on Rayleigh–Ritz–Galerkin's method. This type of method is proposed by Zienkiewicz and Parekh (Reference 12), and a finite difference method using a similar (a little different, though) approach is discussed by Caraso and Parter (Reference 20). Some results will also be given by the present authors (Reference 21).

In order to treat (IVP) as (BVP), the next inner product and norm are introduced;

$$((u, v)) = \int_0^T (u, v)\, dt \quad , \qquad ||| u ||| = \sqrt{((u, u))} \tag{20}$$

Thus a Hilbert space can be constructed on X x [0,T].

In the case that A is strictly accretive , it can be proved that the next operator is strictly accretive in this new Hilbert space;

$$\overset{\circ}{A} = \frac{d}{dt} + A \quad , \tag{21}$$

The proof is;

$$\mathrm{Re}((\mathring{A}u - \mathring{A}v,\ u - v)) = ½ \,\| u - v \|^2 \Big|_0^T + \mathrm{Re} \int_0^T (Au - Av,\ u - v)\ dt$$

$$\geqq \int_0^T \| u - v \|^2 \ dt = \ ||| u - v |||^2 \tag{22}$$

where the following assumption is made;

$$u = v \qquad \text{at} \quad t = 0 \tag{23}$$

In the case $\leqq 0$, the transformation $u = \exp(\lambda t)\, v$ gives the next problem;

$$\frac{dv}{dt} + \exp(-\lambda t)\ A\ \exp(\lambda t)\ v + \lambda v = \exp(-\lambda t)\ f(t), \quad v(0) = u(0) \tag{24}$$

and $\mathring{A} = d/dt + \exp(-\lambda t)\ A\ \exp(\lambda t) + \lambda I$ is strictly accretive if $\lambda > -$.

Of course some consideration is necessary with respect to the domain of $A\ \exp(\lambda t)$ as was pointed by Kato (Reference 17), but this method is applicable at least formally and is certainly valid if A is linear.

III — 3. Semi-discrete methods

In this sub-section, two types of semi-discrete methods (Reference 22) are given.

(a) discretization only to t; in this case, d/dt is replaced by its finite difference analog;

$$\frac{u(t + \Delta t) - u(t)}{\Delta t} + A(\tilde{t})\ \tilde{u}(t) = f(\tilde{t}), \quad u(0) = u_0 \tag{25}$$

where Δt : time mesh size ($t = m\ \Delta\ t$, $m = 0, l, 2, \ldots.$)

$\tilde{t} : t + \theta\ \Delta\ t$, (θ : parameter, $0 \leqq \theta \leqq 1.0$)

$\tilde{u}(t) : \theta u(t + \Delta t) + (1 - \theta)\, u(t)$ (Fig. 1)

(b) discretization only to X : in this case, X and A are replaced by S_n and A_n;

$$\frac{du_n}{dt} + A_n u_n = f_n(t), \qquad u_n(0) = u_{n0} \tag{26}$$

where f_n and u_{n0} are suitable approximations of f and u_0. This method is called Faedo—Galerkin's method if A_n and f_n defined in (14) or (17) are used (Reference 15).

The discussion of these methods are omitted here, but it is to be noted that they sometimes give useful information as the foundation of the perfectly discrete method given in the next section. Especially, the scheme (25) is valid if $0.5 \leq \theta \leq 1.0$. Furthermore, it is valid even in general Banach space if $\theta = 1.0$ (Reference 14).

IV. A FINITE ELEMENT METHOD FOR INITIAL VALUE PROBLEMS

In this section, a finite element scheme belonging to step by step methods is given and its validity is discussed.

IV – 1. Formulation

The equation (19) is now discretized with respect to both time and space;

$$\frac{u_n(t+\Delta t) - u_n(t)}{\Delta t} + A_n(\tilde{t})\, \tilde{u}_n(t) = f_n(\tilde{t}), \qquad u_n(0) = u_{n0} \tag{27}$$

where $\tilde{t}$, $\tilde{u}_n(t)$, A_n, f_n and u_{n0} are perfectly the same as those of (25) and (26). In order to solve (27), one begins from u_{n0} and proceeds step by step. The scheme for $\theta = 0.5$ is known as Crank–Nicolson's one in the finite difference methods (Reference 23). Practically, the schemes for $\theta = 0.0$, 0.5 and 1.0 are mainly employed. Of course, it is assumed that A_n satisfies (7) at each t.

Some similar schemes are also available;

$$\frac{u_n(t+\Delta t) - u_n(t)}{\Delta t} + \theta A_n(\tilde{t})\, u_n(t+\Delta t) + (1-\theta)A_n(\tilde{t})u_n(t) = f_n(\tilde{t}) \tag{28}$$

$$\frac{u_n(t+\Delta t) - u_n(t)}{\Delta t} + \theta A_n(t+\Delta t)u_n(t+\Delta t) + (1-\theta)A_n(t)u_n(t) = f_n(\tilde{t}) \tag{29}$$

where $f_n(\tilde{t})$ can be replaced by $\tilde{f}_n(t)$ if it is preferred. These schemes coincide with (27) if A_n is linear and independent of t, but the results obtained for (27) do not seem to hold generally except for $\theta = 0.0$ and 1.0.

IV – 2. On the validity of the present method

In order to give rigorous treatment to the approximate system, the next three concepts that have been mainly used in the finite difference methods are employed in a little modified forms (References 23, 24 and 25).

(a) consistency : this condition implies that the approximate system actually approximates the original system (19), and is defined as follows;

for the solution u of (19), at least one $u_n^* \in S_n$ can be chosen for each n in such a way that

$$\lim_{\substack{n\to\infty \\ \Delta t\ \ 0}} \|\tau_n^1\| = 0, \qquad \lim_{n\to\infty} \|\tau_n^2\| = 0 \tag{30}$$

where $$\tau_n^1 = \frac{u_n^*(t+\Delta t) - u_n^*(t)}{\Delta t} + A_n(\widetilde{t})\widetilde{u}_n^*(t) - f_n(\widetilde{t})$$

$$\tau_n^2 = u_n^* - u \tag{31}$$

The convergence in (30) is assumed to be uniform in $[0,T]$. If some relations are necessary between n and Δt to achieve consistency, then the scheme is called conditionally consistent. Otherwise, it is called unconditionally consistent.

(b) convergence : this means that the approximate solution u_n converges to the exact solution of (19) in the metric of X as $n \to \infty$ and Δt 0. The convergence is pointwise in $[0,T]$ and not necessarily uniform.

(c) stability : this means that the growth of the approximate solution is bounded to a certain extent. In the present case, only stability with respect to initial value and external source term is considered and it is given as follows;

$$\| u_n(t) \| \leqq C_1(T) \| u_n(0) \| + C_2(T) ||| f_n |||_\infty \quad 0 \leqq t \leqq T \tag{32}$$

where C_1 and C_2 are nonnegative constants dependent only on A and T, and

$$||| f_n |||_\infty = \sup_{0 \leqq t \leqq T} \| f_n(t) \| \tag{33}$$

Usually $u_n(0)$ and f_n in (32) can be replaced by u(0) and f. Sometimes a little stronger condition is preferred;

$$\| u_n(t) - v_n(t) \| \leqq C_1(T) \| u_n(0) - v_n(0) \| + C_2(T) ||| f_n - g_n |||_\infty \tag{34}$$

where $u_n(t)$ and $v_n(t)$ are solutions of (27) corresponding to the initial values $u_n(0)$ and $v_n(0)$ and the force terms f_n and g_n respectively. Equations (32) and (34) are identical in linear problems but seriously different in general nonlinear problems.

The existence and uniqueness of the approximate solution are also essential, though they can be derived from stability in linear problems. In general nonlinear problems, uniqueness can be obtained from the stability of (34) but existence is not assured from the above-mentioned three conditions. Therefore, a proof of the existence is given in the next section in the case of general approximate methods. It is also to be noted that convergence is assured if the stability of the type (34) and consistency hold. This is known as Lax-Kreiss' equivalence theorem in the case of linear problems. On the other hand, uniqueness and convergence are not generally obtained from the stability of (32).

The definition of consistency presented here is a little modified so that it is available to general purposes. This is mainly due to the fact that the restriction of the exact solution u to nodal points, which is usually employed in the finite difference methods, is no longer effective in general approximate methods. Furthermore, this consistency condition is necessary for convergence by the same reason in II – 2.

In the case of the finite element method, the consistency can be assured if S_n is chosen as follows (cf. II – 2);

(i) $S_n \quad D(A)$

(ii) for u of (19), there exists at least one u_n^* in each S_n such that

$$\text{(ii – 1)} \quad \lim_{n\to\infty} \| u_n^*(t) - u(t) \| = 0$$

$$\text{(ii – 2)} \quad \lim_{\substack{n\to\infty \\ \Delta t \quad 0}} \left\| \frac{u_n^*(t+\Delta t) - u_n^*(t)}{\Delta t} \quad \left.\frac{du}{dt}\right|_t \right\| = 0$$

$$\text{(ii – 3)} \quad \lim_{\substack{n\to\infty \\ \Delta t \quad 0}} \| A(\tilde{t})\tilde{u}_n^*(t) - A(\tilde{t})\tilde{u}(t) \| = 0 \tag{35}$$

$$\text{(ii – 4)} \quad \lim_{\Delta t \quad 0} \| A(\tilde{t})\tilde{u}(t) - A(\tilde{t})u(\tilde{t}) \| = 0$$

where the convergence is assumed to be uniform in $[0, T]$. (ii – 4) is a smoothness condition for A and u. Generally it can be expected that the value of (ii – 2) is smallest for $\theta = 0.5$ if u is sufficiently smooth.

As for the stability, it can be easily shown that the next conditions are sufficient ones;

$$\| u_n(t+\Delta t) \| \leqq (1 + C_3\Delta t) \ (\| u_n(t) \| + C_4\Delta t ||| f_n |||_\infty) \text{ for } (32) \tag{36}$$

$$\| w_n(t+\Delta t) \| \leqq (1 + C_3\Delta t) \ (\| w_n(t) \| + C_4\Delta t ||| h_n |||_\infty) \text{ for } (34) \tag{37}$$

where C_3 and C_4 are nonnegative constants independent of n, t and Δt, and dependent only on $'$, and $w_n = u_n - v_n$, $h_n = f_n - g_n$.

By using (7), the next main result of this paper with respect to existence and stability of approximate solutions can be obtained;

> In the case $0.5 \leqq \theta \leqq 1.0$, the approximate solution of (27) exists uniquely and the stability of (34) holds if $\Delta t < C_5 \; (\; ')$, where C_5 is a positive constant dependent only on $'$. (if $' \geqq 0$, then C_5 can be taken as ∞.)

Of course, the convergence is also assured if the consistency is guaranteed.

IV – 3. Sketches of proofs and some remarks

In this sub-section, the outline of the proofs of existence and stability is given.

(a) existence and uniqueness of the approximate solution ($0 \leqq \theta \leqq 1.0$)

For $\theta = 0.0$, the uniqueness and existence are apparent. In other cases, the procedure in each step is essentially to solve the next boundary value problem;

$$\tilde{u}_n + \theta \Delta t A_n \tilde{u}_n = e_n \tag{38}$$

where e_n is a known element in S_n. Because $\theta\Delta t$ is positive, the existence and uniqueness can be derived by use of the results in II – 2 if $' \geqq 0$. Same conclustion is derived if

$$\Delta t \leqq \frac{-1}{\theta\ '}, \qquad \text{for } \ ' < 0 \tag{39}$$

(b) stability ($0.5 \leqq \theta \leqq 1.0$)

The inequality (37) is to be derived. From the definitions of u_n, v_n, w_n and h_n, the next equation is obtained;

$$\frac{w_n(t+\Delta t) - w_n(t)}{\Delta t} + A_n(\tilde{t})\,\tilde{u}_n(t) - A_n(\tilde{t})\,\tilde{v}_n(t) = h_n(\tilde{t}) \tag{40}$$

The following inequalities can be derived by use of such fundamental inequalities as the triangle inequality and Cauchy–Schwartz' one;

$$\mathrm{Re}(w_n(t+\Delta t) - w_n(t),\ \tilde{w}_n(t)) \geqq ½ \left(\| w_n(t+\Delta t) \|^2 - \| w_n(t) \|^2 \right) \quad \text{for } 0.5 \leqq \theta \leqq 1.0 \tag{41}$$

$$\| \tilde{w}_n(t) \| \leqq \theta \| w_n(t+\Delta t) \| + (1-\theta) \| w_n(t) \| \quad \text{for } 0.0 \leqq \theta \leqq 1.0 \tag{42}$$

The next inequality is derived from (7);

$$\mathrm{Re}(A_n\tilde{u}_n(t) - A_n\tilde{v}_n(t),\ \tilde{w}_n(t)) \geqq \ ' \| \tilde{w}_n(t) \|^2 \tag{43}$$

By multiplying (40) by $\tilde{w}_n(t)$ and taking the real part of their inner product, the following inequality can be obtained with the aid of (41) and (43);

$$\begin{aligned} &½ \left(\| w_n(t+\Delta t) \|^2 - \| w_n(t) \|^2 \right) + \ '\Delta t \| \tilde{w}_n(t) \|^2 \\ &\leqq \Delta t \ ||| h_n |||_\infty \| \tilde{w}_n(t) \| \end{aligned} \tag{44}$$

In the case $\leqq 0$ and $\| w_n(t+\Delta t) \| \geqq \| w_n(t) \|$, the next inequality is derived by use of (42);

$$\begin{aligned} \| w_n(t+\Delta t) \|^2 \leqq{}& \| w_n(t) \|^2 - 2 \ '\Delta t \| w_n(t+\Delta t) \|^2 \\ &+ 2\Delta t \ ||| h_n |||_\infty \| w_n(t+\Delta t) \| \end{aligned} \tag{45}$$

Therefore,

$$\| w_n(t+\Delta t) \| \leqq \frac{1}{1+2\ \ '\Delta t}(\| w_n(t) \| + 2\,\Delta\, t\, |\| h_n |\|_{\infty}) \qquad (46)$$

if $\Delta t < \frac{-1}{2\ \ '}$. In the case $\| w_n(t) \| \geqq \| w_n(t+\Delta t) \|$, (46) is apparent. In a similar way, the next inequality can be derived in the case $\ \ ' \geqq 0$;

$$\| w_n(t+\Delta t) \| \leqq \| w_n(t) \| + 2\,\Delta\, t\, |\| h_n |\|_{\infty} \qquad (47)$$

In any case, (37) is valid because $1/(1+2\ \ '\Delta t) \leqq 1-2\ \ '\Delta t \quad (\ \ '<0)$ for sufficiently small Δt.

(c) Remarks

(c–1) As is seen from the above proof, C_3 in (37) can be taken to be 0 if $\ \ ' \geqq 0$. In such cases, the calculation by the present scheme is very stable and the accumulation of round-off error, which is not considered in this paper, is not so serious.

(c–2) In the case $h_n = 0$, the next estimation is obtained for $0.5 \leqq \theta \leqq 1.0$ by use of some inequalities;

$$\| w_n(t+\Delta t) \|^2 \leqq \frac{1 - 2\ \ '(2\theta - 1)\ (1-\theta)\Delta t}{1 + 2\ \ '\theta\ (2\theta - 1)\ \Delta t} \| w_n(t) \|^2 \quad \text{for} \ \ ' \geqq 0$$

$$\| w_n(t+\Delta t) \|^2 \leqq \frac{1 - 2\ \ '(1-\theta)\ \Delta t}{1+2\ \ '\ \theta\,\Delta\, t} \| w_n(t) \|^2 \quad \text{for} \ \ ' \leqq 0 \qquad (48)$$

Therefore, w_n grows at most exponentially with respect to t.

(c–3) The stability of type (32) can be derived in a similar way under the next assumption of A_n;

$$\mathrm{Re}(A_n u_n, u_n) \geqq \quad \| u_n \|^2 \qquad \text{for} \quad u_n \in S_n \qquad (49)$$

The result is obtained by Kreiss in his study of linear finite difference equations (Reference 26).

(c–4) In general Banach space, in which scalar product is generally no longer available, stability can be obtained in a similar way if $\theta = 1.0$ and A_n satisfies (3) in S_n. Therefore, the present theory is applicable even in general Banach space if A_n is linear because the existence can be derived from stability. As for nonlinear problems, the existence is not assured by the method given in this sub-section. Such consideration is not so useful for the finite element methods, which are usually constructed in Hilbert spaces, but may be effective in general approximate methods.

(c–5) In the case of schemes (28) and (29), the uniqueness and existence of u_n can be shown similarly. The stability is assured if

$$\Delta t \leqq \frac{2(\quad ' + C)}{(1-\theta)^2 M_n^2} \tag{50}$$

where C is an arbitrary nonnegative constant and $C > - \quad '$ if $\quad ' \leqq 0$. M_n is the Lipshitz constant in S_n (if it exists);

$$\| A_n u_n - A_n v_n \| \leqq M_n \| u_n - v_n \| \qquad \text{for} \quad u_n, v_n \in S_n \tag{51}$$

M_n generally increases as $n \to \infty$, and its existence is assured if A_n is linear because A_n is continuous in S_n. However, its existence is not guaranteed in nonlinear problems except special cases.

V. SOME FURTHER CONSIDERATIONS ON THE PRESENT METHOD

V – 1. Examples of evolution equations to which the present method is applicable

In this sub-section, some examples of evolution equations to which the present method can be employed are shown with brief explanations. The condition (CA) can be proved by use of symmetric properties of the system in the case of (b), (c), (d), and (e).

(a) heat equation : $\dfrac{du}{dt} + B\,u = f(t), \quad f \in X$ (external source term) (52)

B is a self-adjoint positive bounded below operator in X, the examples of which are certain classes of elliptic operators (Reference 18). In this case, (CA) is apparent because B is positive. This result can be extended to wider classes of elliptic operators such as strongly elliptic ones (Reference 14).

(b) wave equation : $\dfrac{d^2u}{dt^2} + B\,u = f(t), \quad f \in H$ (external force term) (53)

B is same as that of (a) and is independent of t. In order to transform (53) into the form of (19), the following two methods are available.

(b–1) decomposition of B :

Because B is self-adjoint and positive, it can be decomposed as follows;

$$B = T^*T \; ; \quad T : H \qquad H' \; , \qquad T^* : H' \qquad H \tag{54}$$

where H' is another Hilbert space and T* is the adjoint operator of T (Reference 27). Then the equation (53) can be transformed as follows;

$$\frac{du^1}{dt} + T^*u^2 = f(t), \qquad \frac{du^2}{dt} - T\,u^1 = 0 \quad \left(u^1 = \frac{du}{dt},\; u^2 = T\,u\right) \tag{55}$$

where X is taken as HxH' and its scalar product and norm is defined by

$$(u, v)_X = (u^1, v^1)_H + (u^2, v^2)_H, \quad , \quad \| u \|_X = \sqrt{(u, u)_X} \qquad (56)$$

The norm can be regarded as square root of the total energy (usually ½ is multiplied). The condition (CA) can be shown by use of the relation

$$(T^* u^2, u^1)_H = (u^2, Tu^1)_H, \qquad (57)$$

(b–2) use of energy space :

In this case, the equation (53) is transformed into the next form;

$$\frac{du^1}{dt} - u^2 = 0, \quad \frac{du^2}{dt} + B\ u^1 = f(t) \quad (u^1 = u,\ u^2 = \frac{du}{dt}) \qquad (58)$$

X should be taken as H_BxH, where H_B is the energy space derived by B, and its inner product and norm are defined as follows;

$$(u, v)_X = [u^1, v^1]_{H_B} + (u^2, v^2), \quad \| u \|_X = \sqrt{(u, u)_X} \qquad (59)$$

Again, this norm can be regarded as the square root of the total energy, and (CA) can be shown by use of the following fact;

$$[u^2, u^1]_{H_B} = (u^2, B\ u^1)_H \qquad (60)$$

(c) first order wave equation : $\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = f(t), \quad f \epsilon X = L_2$ (61)

It is assumed that u and f have compact support. The condition (CA) can be shown because

$$\text{Re}\ (\frac{\partial u}{\partial x}, u) = -\ \text{Re}\ (u, \frac{\partial u}{\partial x}) = 0 \qquad (62)$$

This result can be extended to general symmetric hyperbolic systems and the above-mentioned support condition can be moderated to a certain extent (Reference 28). A majority of differential equations describing reversible phenomena can be expressed in this form.

(d) Schrodinger's equation : $\frac{du}{dt} + i\ B\ u = 0$ (i = imaginary unit) (63)

In this case, the source term is usually absent and B is self-adjoint. (CA) can be derived by use of the fact that B is self-adjoint.

(e) equation of coupled sound and heat flow:

$$\frac{\partial u^1}{\partial t} - c \frac{\partial}{\partial x} (u^2 - (\gamma - 1) u^3) = 0, \qquad \frac{\partial u^2}{\partial t} - c \frac{\partial u^1}{\partial x} = 0$$
$$\frac{\partial u^3}{\partial t} - \sigma \frac{\partial^2 u^3}{\partial x^2} + c \frac{\partial u^1}{\partial x} = 0 \tag{64}$$

where c, γ (>1) and σ are positive constants. The physical meanings of (64) are given in Reference 23, and a similar equation is treated by Oden and Kross (Reference 29). This is a coupled type of equation of parabolic and hyperbolic, and (CA) can be shown by imposing suitable boundary conditions (e.g. $u^1 = 0$, $u^3 = 0$ on the boundary) and replacing u^3 by $u^3/\sqrt{\gamma - 1}$.

(f) semilinear heat equation; $\frac{du}{dt} + B\,u + h(u) = 0$ (65)

where B is same as that of (a) and h (nonlinear source term) is assumed to be smooth and non-decreasing to assure (CA). Sometimes certain classes of perfectly nonlinear heat equations can be treated similarly (Reference 15).

The present method is valid irrespective of linearity and non-linearity of A and many other examples can be found. However, it is seen from these examples that representative linear evolution equations in mathematical physics can be analyzed by the present method.

V – 2. Consistency in the case that the theory of energy space is applicable.

In the preceding sections, a general theory for initial value problems is presented. However, the conditions (35) are sometimes too severe in the finite element methods. In this sub-section, it will be shown that such a situation can be moderated to a certain extent in the case that the theory of energy space is applicable.

The equations treated here are : heat equation (a) (in V – 1), wave equation (b – 2) and Schrodinger equation (d). In all cases, B is assumed to be self-adjoint, positive bounded below and independent of t. S_n is so constructed as (a') and (b') in II – 2 and A_n is constructed as (17).

Then, the best approximation of u(t) (exact solution) in H_B can be used as u_n^* in (31) if u is sufficiently smooth. In this way, the error estimation of $B_n u_n^*$ becomes very easy.

It is also to be noted that $d^m u_n^*/dt^m$ ($m = 1, 2, \ldots$) is the best approximation of $d^m u/dt^m$ in H_B if it belongs to H_B. Furthermore, stability and convergence in H_B are assured in the case of heat equation, and stability in H_B is assured in the case of Schrodinger's equation.

V – 3. Conservation of norms of approximate solutions

Heretofore, the convergence and stability of the approximate solution have been mainly discussed. However, there is an important character for certain classes of initial value problems that a certain norm is conserved. Therefore, it is desirable that not only the approximate solution converges to the exact solution but also the corresponding norms of the approximate solution are conserved. Especially such a situation is favorable for long time calculation because an important physical value is conserved even if coarse mesh division is employed. In practical calculations, it is often observed that the error in the norm (not error of the norm) becomes large as time passes because of the deviation of the phase. Even in such cases, approximate solutions still can be regarded as good ones if such valuables as amplitude, period and norm are in good agreement with the exact values and the wave of decay is small (Fig. 2).

Such a situation can be realized in the following equations if θ is taken as 0.5. In the following examples, the source terms are absent in all cases.

(a) wave equation : (b–1) and (b–2) in V – 1. The corresponding norm is square root of total energy.

(b) first order wave equation : (c) in V – 1. The corresponding norm is L_2 – norm. This result can be extended to general symmetric hyperbolic systems with constant coefficients.

(c) Schrodinger's equation : (d) in V – 1. The corresponding norm is the norm in X (or H) (Reference 30). If B is positive bounded below, the norm of H_B is also conserved.

The proofs are omitted here, but they can be shown in a similar way to that of stability (section IV).

V – 4. Some considerations on the lumping of the mass matrix

As was discussed in the preceding sections, it is shown that the present method is valid at least in the case that $0.5 \leqq \theta \leqq 1.0$. However, the scheme is perfectly implicit in such cases, and, moreover, a new set of simultaneous equation must be solved at each step if A changes from step to step. This situation especially causes great computational difficulty if A is nonlinear.

To avoid this difficulty, the scheme for $\theta = 0.0$ can be used, though stability is not assured unless Δt is sufficiently small (see section V). In this case, a common set of linear simultaneous equations is required to be solved at all steps. However, the scheme thus obtained is still implicit, and use of the lumping the matrix derived from I (identity operator in X) is often made to make it explicit. This modified matrix is usually called lumped mass matrix in structural analysis, because it appears as the term of inertia.

In this sub-section, it will be shown in simple examples that more stable schemes can be sometimes obtained if lumping is suitably made. In the following examples, the piecewise linear shape function is employed and spatial mesh size Δx is assumed to be uniform. Furthermore, u is used instead of u_n.

(a) heat equation : $\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0$ (66)

The finite element scheme for $\theta = 0.0$ at interior mesh points is given as follows:

$$\frac{u(t+\Delta t,x+\Delta x)+4u(t+\Delta t,x)+u(t+\Delta t,x-\Delta x)-u(t,x+\Delta x)-4u(t,x)-u(t,x-\Delta x)}{6\,\Delta\,t}$$

$$-\frac{u(t,x+\Delta x)-2u(t,x)+u(t,x-\Delta x)}{\Delta\,x^2} = 0 \qquad (67)$$

This scheme is stable in L_2 if $\Delta t/\Delta x^2 \leqq 1/6$ (see V – 5). If lumping is ma made with respect to both t and t+Δt, the next scheme is obtained;

$$\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t} - \frac{u(t,x+\Delta x)-2u(t,x)+u(t,x-\Delta x)}{\Delta x^2} = 0 \qquad (68)$$

As is well known, this scheme is stable not only in L_2 but also in the uniform norm if $\Delta t/\Delta x^2 \leqq 0.5$ (Reference 23).

(b) first order wave equation: $\dfrac{\partial u}{\partial t} + \dfrac{\partial u}{\partial x} = 0$ (69)

The finite element scheme for $\theta = 0.0$ at interior mesh points is given as follows;

$$\frac{u(t+\Delta,x+\Delta x)+4u(t+\Delta t,x)+u(t+\Delta t,x-\Delta x)-u(t,x+\Delta x)-4u(t,x)-u(t,x-\Delta x)}{6\,\Delta\,t}$$

$$+\frac{u(t,x+\Delta x)-u(t,x-\Delta x)}{2\,\Delta\,x} = 0 \qquad (70)$$

This scheme is stable in L_2 if $\Delta t/\Delta x^2 \leqq C$ (C = arbitrary positive constant). If the same lumping as (a) is made, the stability condition is not improved essentially. However, if lumping is made only to t+Δt, then the next scheme is obtained;

$$\frac{6u(t+\Delta t,x)-u(t,x+\Delta x)-4u(t,x)-u(t,x-\Delta x)}{6\,\Delta\,t} + \frac{u(t,x+\Delta x)-u(t,x-\Delta x)}{2\,\Delta\,x} = 0 \qquad (71)$$

This scheme is consistent with (69) if $\Delta t/\Delta x$ is kept constant. As is easily proved, this scheme is stable in uniform norm if $\Delta t/\Delta x \leqq 1/3$.

In these examples, stability is improved by lumping. Similar result is obtained in the wave equation if two-level scheme is employed. However, such a conclusion does not hold generally because general method of lumping is not known. Anyway, the next assertion must be correct : by lumping, A_n is modified; if the modified A_n satisfies the conditions in II and IV, then the general theory of this paper can be adopted.

V – 5. Some results for $0.0 \leqq \theta < 0.5$ (linear problems)

In order to deal with the case $0.0 \leqq \theta < 0.5$, which is not referred to yet, the method of expansion by eigenfunctions is effective if A_n is linear and independent of t and any element of S_n can be expressed by linear combination of eigenfunctions of A_n. Such a situation is realized, for example, if A_n is self-adjoint in S_n.

Then, a simple stability condition is given by

$$\Delta t \leqq \min_i \frac{2(C + \mathrm{Re}\,\lambda_{ni})}{(1-2\theta)|\lambda_{ni}|^2} \quad \text{for } 0.0 \leqq \theta \leqq 0.5 \quad (1 \leqq i \leqq \dim(S_n)) \quad (72)$$

where λ_{ni} is the i'th eigenvalue of A_n and C is an arbitrary non—negative constant that satisfies $C + \mathrm{Re}\lambda_{ni} > 0$. It is also to be noted that $\mathrm{Re}\lambda_{ni} \geqq$ '. In the case that $\theta = 0.0$, $C = 0.0$ and $\mathrm{Re}\lambda_{ni} > 0$, this result is identical with that of Reference 22.

Similar methods such as method of Fourier series and Fourier transform are also available if spatial mesh size is uniform and the coefficients of A are constant (Reference 23).

Some results obtained by these methods are shown for typical linear evolution equations;

(a) heat equation : (a) in V – 1 $$\Delta t \leqq \frac{2}{(1-2\theta)\,\lambda_{n\ \max}} \quad (73)$$

(b) wave equation : (b – 2) in V – 1 $$\Delta t \leqq \frac{C}{(1-2\theta)\,\lambda_{n\ \max}} \quad (74)$$

(c) Schrodinger's equation : (d) in V – 1 $$\Delta t \leqq \frac{C}{(1-2\theta)\,\lambda^2_{n\ \max}} \quad (75)$$

(d) first order wave equation : (c) in V – 1 $$\Delta t \leqq \frac{C\Delta x^2}{1-2\theta} \quad (76)$$
(in this case, piecewise linear shape function with uniform spatial mesh size Δx is employed.)

where $\lambda_{n\ \max}$ is the largest eigenvalue of B_n, which is assumed to be self-adjoint and positive bounded below, and C is an arbitrary positive constant. If $B = -d^2/dx^2$ and piecewise linear shape function with uniform mesh size Δx is used, then $\lambda_{n\ \max} = 12/\Delta x^2$. In (74), (75) and (76), there is no definite threshold value for stability as in (73) because C is arbitrary, and Δt must be made sufficiently small until reasonable results are obtained in practical computation.

The present condition only assures that the stability is guaranteed if $n \to \infty$ and Δt 0 with the above-mentioned conditions preserved. In such cases, the order of C_3 in (36) or (37) is equal to that of C.

As is seen from these results, the stability condition changes greatly if the equation changes. This is an important point to be recognized especially in the calculation of mixed type phenomena, in which different types of equations must be treated together. Furthermore, the conditions required are often too severe to carry out practical calculations. In such cases, the scheme for $0.5 \leqq \theta \leqq 1.0$ should be used.

V – 6. Two–level scheme for wave equation

In this sub-section, some consideration is given to two-level scheme for wave equation because such scheme has been often discussed (References 3, 4, 5, and 7). The scheme treated here is given as follows;

$$\begin{aligned}\frac{u_n(t+\Delta t)-2u_n(t)+u_n(t-\Delta t)}{\Delta t^2} &+B_n[\theta_2\ \theta_1 u_n(t+\Delta t)+(1-\theta_1)u_n(t) && (77)\\ &+(1-\theta_2)\ \theta_1 u_n(t)+(1-\theta_1)u_n(t-\Delta t) && =0\end{aligned}$$

where θ_1 and θ_2 are constants $(0 \leqq \theta_1,\ \theta_2 \leqq 1)$, B_n is the approximation operator of B given in V – 1, and the source term f_n is omitted here because it does not affect the stability.

By use of a similar method to that of section IV, the unconditional stability in the following norm is obtained if $0.5 \leqq \theta_1,\ \theta_2 \leqq 1.0$;

$$\sqrt{\left\|\frac{u_n(t+\Delta t)-u_n(t)}{\Delta t}\right\|_H^2 + \left|\theta_1 u_n(t+\Delta t)+(1-\theta_1)u_n(t)\right|_{H_B}^2} \qquad (78)$$

Therefore, the stability in the total energy is obtained, where velocity is replaced by its finite difference analog.

If the method of expansion by eigenfunction is employed, a bit more information can be obtained;

$$\begin{array}{ll}\text{unconditionally stable} & \text{if } 3\theta_3+\theta_4 \geqq 1 \text{ and } \theta_3 \geqq \theta_4\\[2ex] \Delta t \leqq \dfrac{2}{\sqrt{1-3\theta_3-\theta_4)\ \lambda_{n\ max}}} & \text{if } 3\theta_3+\theta_4 < 1 \text{ and } \theta_3 \geqq \theta_4\end{array} \qquad (79)$$

where $\theta_3 = \theta_1\theta_2$ and $\theta_4 = (1-\theta_1)(1-\theta_2)$, and $\lambda_{n\ max}$ is the largest eigenvalue of B_n. This result includes some of the results given in References 3, 4, 5, and 7. The latter condition (79) is, however, weaker than the former (78) because this stability is essentially stability of u_n and not that of velocity.

V – 7. Some remarks on the present method for more general nonlinear problems

As was already referred to, the present method is valid to nonlinear problems if (CA) holds. However, nonlinear problems are very complex and it cannot be expected that A satisfies (CA) in general nonlinear problems. Therefore, there arises a question whether the present method is applicable to such problems, in which the global solution may not exist for arbitrary initial values and source terms. To give definite answer to it is probably very difficult, but the following consideration seems to still be possible.

The condition (CA) is fairly general for linear evolution equations, for it is a kind of stability condition for the system. On the other hand, such nonlinear equations are often obtained by linearization of the originally nonlinear ones.

Therefore, it can be expected that fairly wide classes of nonlinear evolution equations satisfy (CA) at least locally. If it is true, the present procedure can be continued until (CA) is violated.

Next, the computational procedure becomes very complicated because the approximate equations are usually nonlinear if A is nonlinear. To avoid this difficulty, such techniques as linearization at each step (piecewise linear procedure), extrapolation and iteration can be employed. Newton-Raphson method and predictor-corrector method are such examples (Reference 25). As for their validity, much investigation will be necessary.

In nonlinear problems, sometimes discontinuous solutions arise even for sufficiently smooth initial values. Therefore, approximate methods should be able to treat such problems correctly. The experience of the finite difference methods shows that the concept of artificial viscosity is effective especially to such problems. That is, the instability due to nonlinear effect can be annealed by such damping terms (References 31 and 32).

In the numerical experiments given in the next section, a simple nonlinear shock wave problem is treated by use of the above-mentioned methods and their effectiveness is investigated.

VI. NUMERICAL EXPERIMENTS

In order to demonstrate the validity of the present method, some numerical experiments were performed for several evolution equations. Some of the typical results are shown in this section. Especially some weak solutions are treated numerically, though the theoretical justification is not given in this paper.

In the present experiments, only one dimensional (in space) problems are treated by use of the piecewise linear shape function with uniform mesh size Δx. The initial values are taken such that they coincide with the exact initial values at nodal points.

The equations treated here are heat equation, wave equation, first order wave equation, equation of coupled sound and heat flow, Schrodinger's equation and nonlinear hyperbolic equation. The last one is treated purely experimentally because (CA) is not generally satisfied.

The main points of experiments to be observed are dependence of the scheme on θ and A. It can be generally expected that the scheme for $\theta = 0.5$ gives the best results to smooth solutions and $\theta = 1.0$ gives most stable results. It is also to be recognized that the method of solving the equation remains the same even if A differs.

In the present calculations, almost all the problems are solved by single precision (32 bits on HITAC 5020E) arithmetic and linear simultaneous equations are solved by Gaussian elimination method for band matrices.

(a) heat equation : $$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \ , \ u(0,x) = \sin(\pi x) \quad (0 \leqq x \leqq 1) \qquad (80)$$

The exact solution is $\sin(\pi x)\exp(-\pi^2 t)$. In Fig. 3, the results of $u(t,0.5)$ for $\Delta x = 0.1$ and $\Delta t = 0.01$ are shown in the case $\theta = 0.0$, 0.5 and 1.0. As is seen from them, the approximate solution for $\theta = 0.0$ is unstable. On the other hand, the approximate solutions for θ = 0.5 and 1.0 are stable and the results for θ = 0.5 is the best.

In Fig. 4, the results for θ = 0.0 are shown for Δx = 0.1 and several values of Δt. The theory given in V – 5. predicts that the scheme is stable if $\Delta t/\Delta x^2 \leqq 1/6$, and this is ascertained by this experiment. By the way, the present solution belongs to a special class; that is, the approximate solution is an eigenfunction of the approximation operator corresponding to the smallest eigenvalue and it coincides with the exact eigenfunction at nodes. Therefore, its convergence occurs even if the above-mentioned stability condition is not satisfied. However, this situation is violated fairly rapidly by roundoff errors, and the phenomena is well represented in this figure. If the calculation is performed by the single precision in the case of Δt = 0.01, the approximate solution becomes unstable in about 7 steps, while it becomes unstable in about 15 steps in the double precision. That is, the same principle as that of power method governs the present phenomena and higher modes become dominant after some steps as is shown in Fig. 5 in the case of Δt = 0.005. Such a situation is much severer if the initial value originally contains higher modes.

Table 1 shows the results for $\Delta t/\Delta x^2 = 1/6$ in. which the scheme is stable for all values of θ. It is seen from them that highly accurate result is obtained for θ = 1.0. Similar phenomena is observed in the case of the finite difference scheme (68) for $\Delta t/\Delta x^2 = 1/6$ (Reference 23), and the present result is approved because the finite element scheme coincides with the finite difference scheme (68) in this case. The high accuracy is due to the small local truncation error of this scheme, but such a situation is no longer expected if a source term exists. In general, the scheme for θ = 0.5 gives the best results.

(b) wave equation:
$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0 \ , \ u(0,x)=\sin(\pi x), \ \frac{\partial u}{\partial t} = 0(0 \leqq x \leqq 1) \qquad (81)$$

The exact solution is $\sin(\pi x)\cos(\pi t)$. In this case, the scheme for (b-2) in V-1 is employed and the scheme for θ = 0.5 coincides with one-step β-scheme for $\beta = 1/4$ (Reference 3).

Figure 6 shows the results of u(t, 0.5) in the case of $\Delta t = \Delta x = 0.1$. Again, the scheme is unstable for θ = 0.0 and best for θ = 0.5. For θ = 1.0, the result is highly contractive and much smaller value of Δt is needed to improve the result. On the other hand, the scheme for θ = 0.5 gives excellent result because of its conservation of total energy and of small local truncation error.

Figure 7 shows the effect of Δt for θ = 0.0, where Δx = 0.1 in all cases. The result can be improved and the scheme becomes stable gradually if the value of Δt decreases, but it does not seem that there exists a definite threshold value of Δt as in the case of heat equation.

(c) first order wave equation:
$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0 \ , \ u(0,x)=u_0(x) \ (-\infty < x < \infty) \qquad (82)$$

The exact solution is $u_0(x-t)$ and discontinuous solution arises if u_0 is discontinuous. In the following examples, the interval for numerical calculation is taken sufficiently broad so that it approximates the infinite interval.

Figure 8 shows some results of convergence of approximate solution for $\Delta t/\Delta x$ = 1.0. The initial value is so chosen that it approximates the step function with unit height. The procedure is unstable for θ = 0.0 as is predicted in V-5, and is stable for θ = 0.5 and 1.0. For θ = 0.5, overshoot and oscillation is observed in the neighborhood of the wave front, and the height of the overshoot does not seem to vanish even if mesh sizes are decreased. For θ = 1.0, such oscillation is not observed but the rise of the wave is not so sharp as that of θ = 0.5. Figure 9 shows the effect of the value of θ to the same problem. As is seen from it, the oscillation vanishes if the value of θ is taken a little larger than 0.5.

On the other hand, the scheme for θ = 0.5 gives best result for continuous solution as is shown in Fig. 10, where u_0 is given as follows:

$$u_0(x) = x(1-x) \quad \text{for } 0 \leqq x \leqq 1 \ , \ u_0(x) = 0 \quad \text{for } x<0 \text{ and } x>1$$

In general, the scheme for θ = 0.5 should be used for smooth solutions, but sometimes greater values of θ should be used for not sufficiently smooth solutions.

(d) Schrodinger's equation: $$\frac{\partial u}{\partial t} + i\left(-\frac{\partial^2 u}{\partial x^2} + V(x)\right) u = 0 \qquad (83)$$

Table 2 shows the results for a steady state solution, where V(x) = 0.0, and the initial value and the exact solution are $u(0,x) = \sin(\pi x)$ and $u(t,x) = \sin(\pi x) \exp(i\pi^2 t)$ respectively. As is seen from these results for $|u(t, 0.5)|^2$, the result for θ = 0.5 is best, while the result for θ = 0.0 is unstable and the result for θ = 1.0 is highly contractive.

Figures 11 and 12 show the calculated profile ($|u(t,x)|^2$) of Gaussian wave packet by the schemes for θ = 0.5 and 1.0. In this case, the scheme for θ = 0.0 gives rapidly diverging solution. The initial value and the potential are given as follows:

$$u(0,x) = (2\pi y^2)^{-1/4} \exp \quad -\frac{(x-0.16)^2}{4y^2} + i\ 50\pi x \quad (0 < x < 0.64)$$

$$V(x) = 5000\pi^2 \qquad (0.32 \leqq x \leqq 0.384)$$

$$= 0 \qquad (\text{otherwise})$$

where y = 0.035 and the values of u at both ends are 0. The mesh sizes are: $\Delta x = 0.004$, $\Delta t = \Delta x^2$. Figure 13 shows the change of the shape of the wave packet after bouncing back by potential and rigid wall. The tendency of the finite element solution for θ = 0.5 is in good agreement with that of the finite difference solution by Goldberg et al. (Reference 30). On the other hand, the result for θ = 1.0 is much more attractive again.

(e) equation fo coupled sound and heat flow: (e) in V-1

The initial value is given as follows:

$$u^1(0,x) = 1.0, \quad u^2(0,x) = 1/\sqrt{3}, \quad u^3(0,x) = 1/\sqrt{3} \quad \text{for } x < 0$$

$$u^1(0,x) = 0.0, \quad u^2(0,x) = 0.0 \ , \quad u^3(0,x) = 0.0 \quad \text{for } x > 0$$

In this case, the speed of propagation of shock wave is 1.0 if σ = 0.0. This infinitesimal shock decays gradually due to the heat conduction as is shown in Fig. 14, where the calculation was done by the scheme for θ = 0.5. Figure 15 shows the profile of the shock at t = 165, and the result is compared with the finite difference solution by Richtmyer and Morton (Reference 23). Their agreement is fairly good and it seems that the effect of coupling is well represented.

(f) nonlinear hyperbolic equation: $\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \epsilon \Delta x \frac{\partial^2 u}{\partial x^2} \quad (\epsilon \geqq 0)$ (84)

The last term in (84) is added as artificial viscosity and the equation is treated as a parabolic one if $\epsilon \neq 0.0$ (References 31 and 32). In this case, the stability of the type (32) is assured if u has compact support, because (49) can be proved. However, (34) is not generally expected.

In order to treat the nonlinearity, this equation is linearized in each step [t, t + Δt] by using U, the value of the approximate solution at t;

$$\frac{\partial u}{\partial t} + U \frac{\partial u}{\partial x} + u \frac{\partial U}{\partial x} - U \frac{\partial U}{\partial x} = \epsilon \Delta x \frac{\partial^2 u}{\partial x^2} \tag{85}$$

The problems treated here are propagation of shock wave and rarefaction wave where the initial values are taken as step functions with unit height.

Figure 16 shows the result of shock wave at t = 100 for several values of ϵ, where $\Delta t = 1.0$, $\Delta x = 1.0$, $\theta = 1.0$, and the shock speed is 0.5. It is seen that a fairly good result can be obtained for $\epsilon = 0.0$ in this case but better result can be obtained if the value of ϵ is suitably chosen. In this calculation, the shock wave calculated reaches its steady state after about 20 steps.

Figure 17 shows the rarefaction wave calculated by setting $\epsilon = 0.0$ and $\theta = 1.0$. In this case, the problem is not so serious as that of shock wave and a fairly good result is obtained without using the artificial viscosity.

VII. CONCLUDING REMARKS

Some consideration is made with respect to a simple finite element scheme for initial value problems, and a criterion for its convergence and stability is given. Its applicability is also demonstrated experimentally, and it is shown that fairly wide classes of evolution equations can be treated by a unified method. There are many things left to be done, but the functional analysis and the numerical experiments will be efficiently employed for such investigations.

REFERENCES

1. Zienkiewicz, O. C., Cheung, Y. K. The Finite Element Method in Structural and Continuum Mechanics, McGraw–Hill, 1967.

2. Oden, J. T. "A General Theory of Finite Elements: I. Topological Considerations; II. Applications", Int. J. Num. Meth. Engrg, I, 205–221 & 247–259, 1969.

3. Newmark, N. M. "A Method of Computation for Structural Dynamics", J. Eng. Mech. Div., Proc. ASCE, 85, 67–94, 1959.

4. Chan, S. P., Cox, H. L., Benfield, W. A. "Transient Analysis of Forced Vibrations and Complex Structural Mechanical Systems", J. Aeronaut. Soc., 66, 457–460, 1962.

5. Leech, J. W., Hsu, P. T., Mack, E. W. "Stability of a Finite-Difference Method for Solving Matrix Equations", AIAA J., 3, 2172–2173, 1965.

6. Johnson, D. E. "A Proof of the Stability of the Houbolt Method", AIAA J., 4, 1450–1451, 1966.

7. Nickell, R. E. "On the Stability of Approximation Operators in Problems of Structural Dynamics", Int. J. Solids Structures, 7, 301–319, 1971.

8. Visser, W. "A Finite Element-Method for the Determination of Non-Stationary Temperature Distribution and Thermal Deformations", Proc. Conf. Matrix. Meth. Struct. Mech., Wright-Patterson AFB, Ohio, 925–944, 1965.

9. Wilson, E. L., Nickell, R. E. "Application of the Finite Element Method to Heat Conduction Problems", Nucl. Engrg Des., 4, 276–286, 1966.

10. Douglas, J., Dupont, T. "Galerkin Methods for Parabolic Equations", SIAM J., Numer. Anal., 7, 575–626, 1970.

11. Descloux, J. "On the Numerical Integration of the Heat Equation", Num. Meth., 15, 371–381, 1970.

12. Zienkiewicz, O. C., Parekh, C. J. "Transient Field Problems; Two-dimensional Analysis by Isoparametric Finite Elements", Int. J. Num. Meth. Engrg, 2, 61–71, 1970.

13. Argyris, J. H., Scharpf, D. W. "Finite Elements in Time and Space", Nucl. Engrg. Des., 10, 456–464, 1969.

14. Yoshida, K. Functional Analysis, Springer, 1968.

15. Lions, J. L. Quelques Methodes de Resolution des problemes aux Limtes non Lineaires, Gauthier, 1969.

16. Komura, Y. "Nonlinear Semi-Groups in Hilbert Space", J. Math. Soc. Japan, 19, 493–507, 1967.

17. Kato, T. "Nonlinear Semigroups and Evolution Equations", J. Math. Soc. Japan, 19, 508–520, 1967.

18. Mikhlin, S. G. The Problem of the Minimum of a Quadratic Functional, Holden-Day, 1965.

19. Arantes Oliveira, E. R. "Theoretical Foundation of the Finite Element Method", Int. J. Solids Structures, 4, 929–952, 1968.

20. Caraso, A., Parter, S. V. "An Analysis of 'Boundary-Value Techniques' for Parabolic Problems", Math. Comp. 24, 315–340, 1970.

21. Kikuchi, F., Ando, Y. "A Finite Element Method for Friedrichs' Symmetric Positive Systems", a paper to be presented at 21st Japan Nat. Congr. Appl. Mech., Tokyo, 1971.

22. Varga, R. S. Matrix Iterative Analysis, Prentice-Hall, 1962.

23. Richtmyer, R. D., Morton, K. W. Difference Methods for Initial-Value Problems, Interscience, 1967.

24. Lax, P. D., Richtmyer, R. D. "Survey of the Stability of Linear Finite Difference Equation", CPAM, 9, 267–293, 1956.

25. Isaacson, E., Keller, H. B. Analysis of Numerical Methods, John Wiley & Sons, 1966.

26. Kreiss, H. O. "Uber Implizite Differenzmethoden fur partielle Differentialgleichungen", Num. Math., 5, 24–47, 1963.

27. Fujita, H. "Contribution to the Theory of Upper and Lower Bounds in Boundary Value Problems", J. Phys. Soc. Japan, 10, 1–8, 1955.

28. Mizohata, S. Theory of Partial Differential Equations (Japanese title: Hembibunhoteishiki-ron), Iwanami, 1965.

29. Oden, J. T., Kross, D. A. "Analysis of General Coupled Thermo-Elastic Problems by the Finite Element Method", Proc. 2nd Conf. Matrix Meth. Struct. Mech., Wright-Patterson AFB, Ohio, 1968.

30. Goldberg, A., Schey, H. M., Schwartz, J. L. "Computer Generated Motionpicture of One Dimensional Quantum Mechanical Transmission and Reflection", Am. J. Phys., 35, 177–186, 1967.

31. von Neumann, J., Richtmyer, R. D. "A Method for the Numerical Calculation of Hydrodynamic Shocks", J. Appl. Phys., 21, 232–237, 1950.

32. Lax, P. D. "Weak Solution of Nonlinear Hyperbolic Equations and Their Numerical Computation", CPAM, 7, 159–193, 1954.

θ STEP	0.0	0.5	1.0	EXACT
1	9.834E–01	9.836E–01	9.837E–01	9.837E–01
5	9.198E–01	9.204E–01	9.210E–01	9.210E–01
10	8.460E–01	8.472E–01	8.483E–01	8.483E–01
50	4.334E–01	4.364E–01	4.394E–01	4.393E–01
100	1.878E–01	1.904E–01	1.930E–01	1.930E–01
150	8.138E–02	8.309E–02	8.481E–02	8.480E–02
200	3.527E–02	3.626E–02	3.726E–02	3.726E–02
250	1.528E–02	1.582E–02	1.637E–02	1.637E–02
300	6.623E–03	6.904E–03	7.192E–03	7.192E–03

$\Delta x = 1/10$

$\Delta t = 1/600$

Table 1: Comparison of u(0.5) for θ = 0.0, 0.5 and 1.0 —— heat equation ——

Δt	0.1			0.05		
θ	0.0	0.5	1.0	0.0	0.5	1.0
t = 0.1	1.99E+00	1.00E+00	5.02E–01	1.56E+00	1.00E+00	6.43E–01
0.5	9.71E+04	1.00E+00	3.20E–02	3.31E+19	1.00E+00	1.09E–01
1.0	3.42E+25	1.00E+00	1.03E–03		1.00E+00	1.20E–02
2.0	overflow	1.00E+00	1.05E–06	overflow	1.00E+00	1.44E–04
3.0		1.00E+00	1.08E–09		1.00E+00	1.72E–06
4.0		1.00E+00	1.11E–12		1.00E+00	2.07E–08
5.0		1.00E+00	1.13E–15		1.00E+00	2.48E–10
EXACT	1.0000					

$\Delta x = 0.1$

Table 2: Comparison of $|u(0.5)|^2$ for θ = 0.0, 0.5 and 1.0 — Schrodinger's equation —

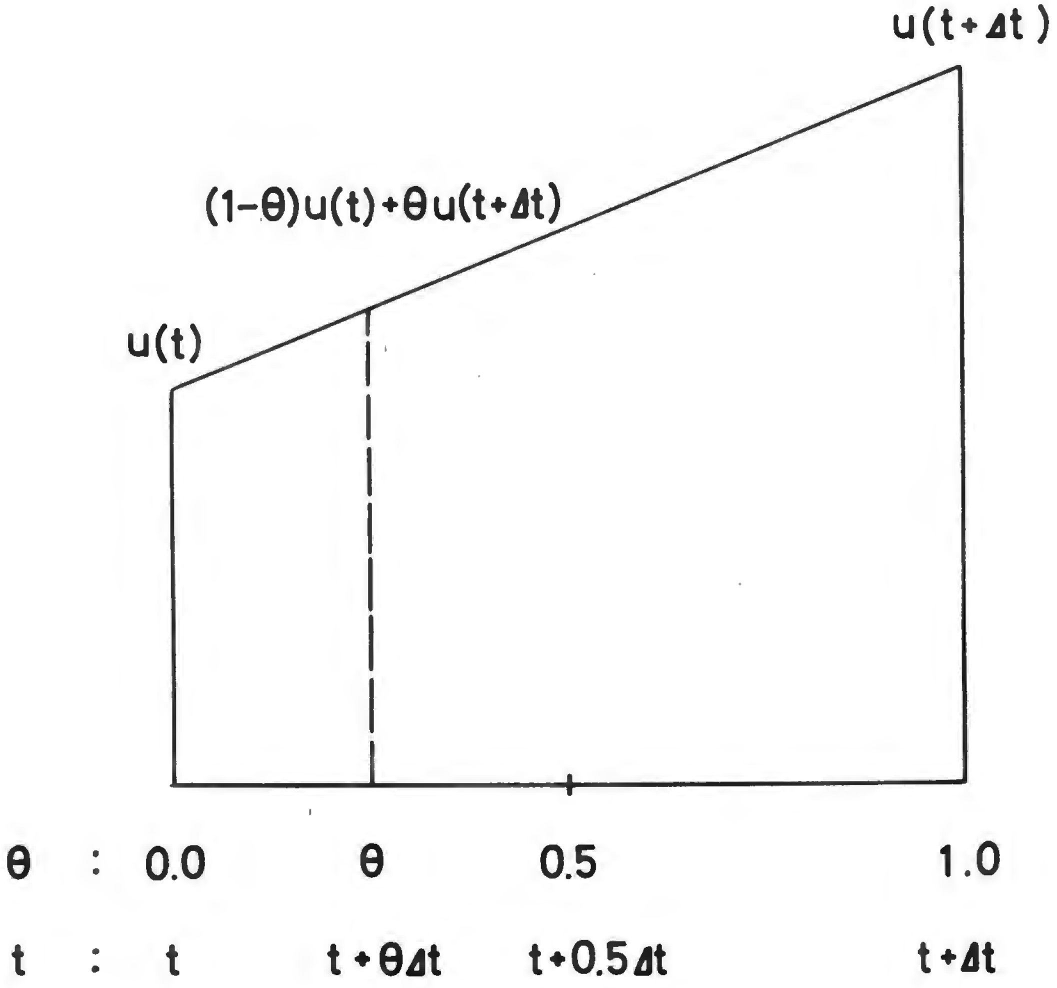

Figure 1. Parameter "θ"

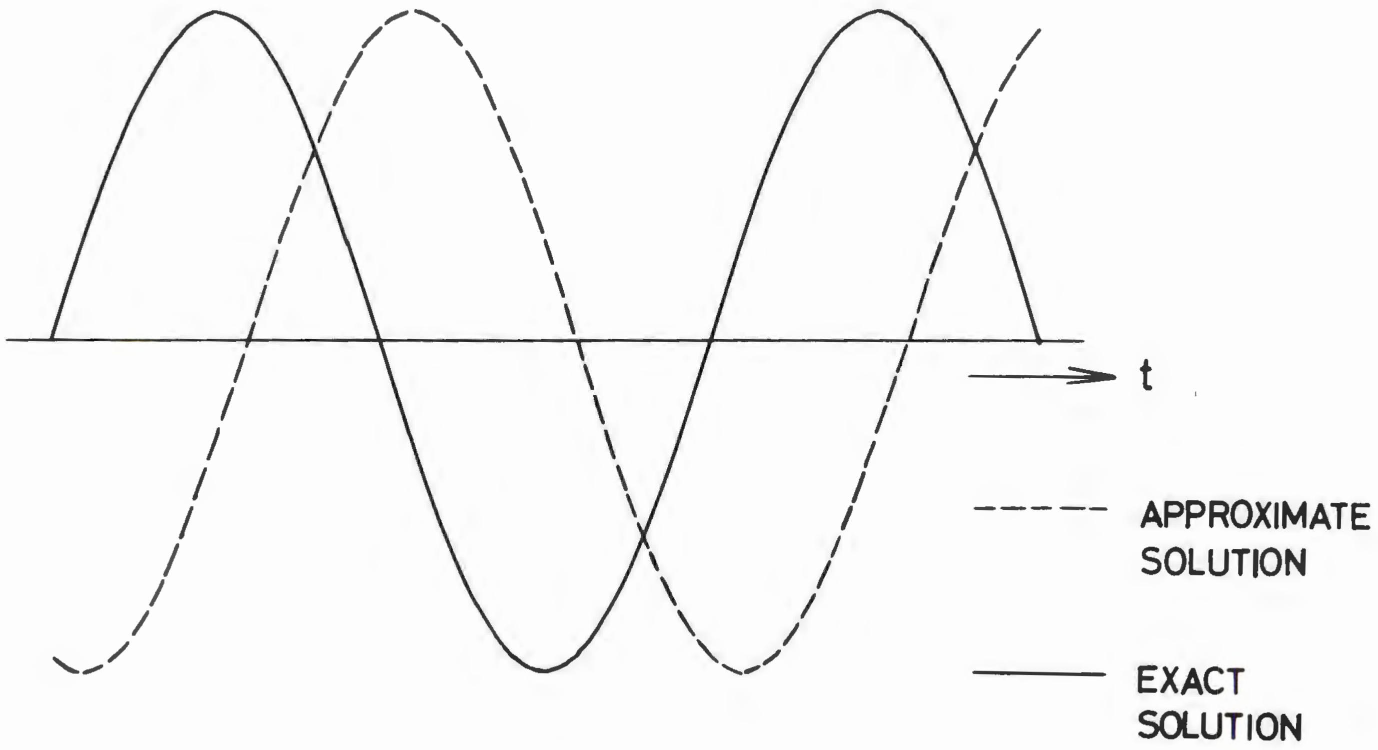

Figure 2. Behavior of an Approximate Solution When the Norm Is Conserved

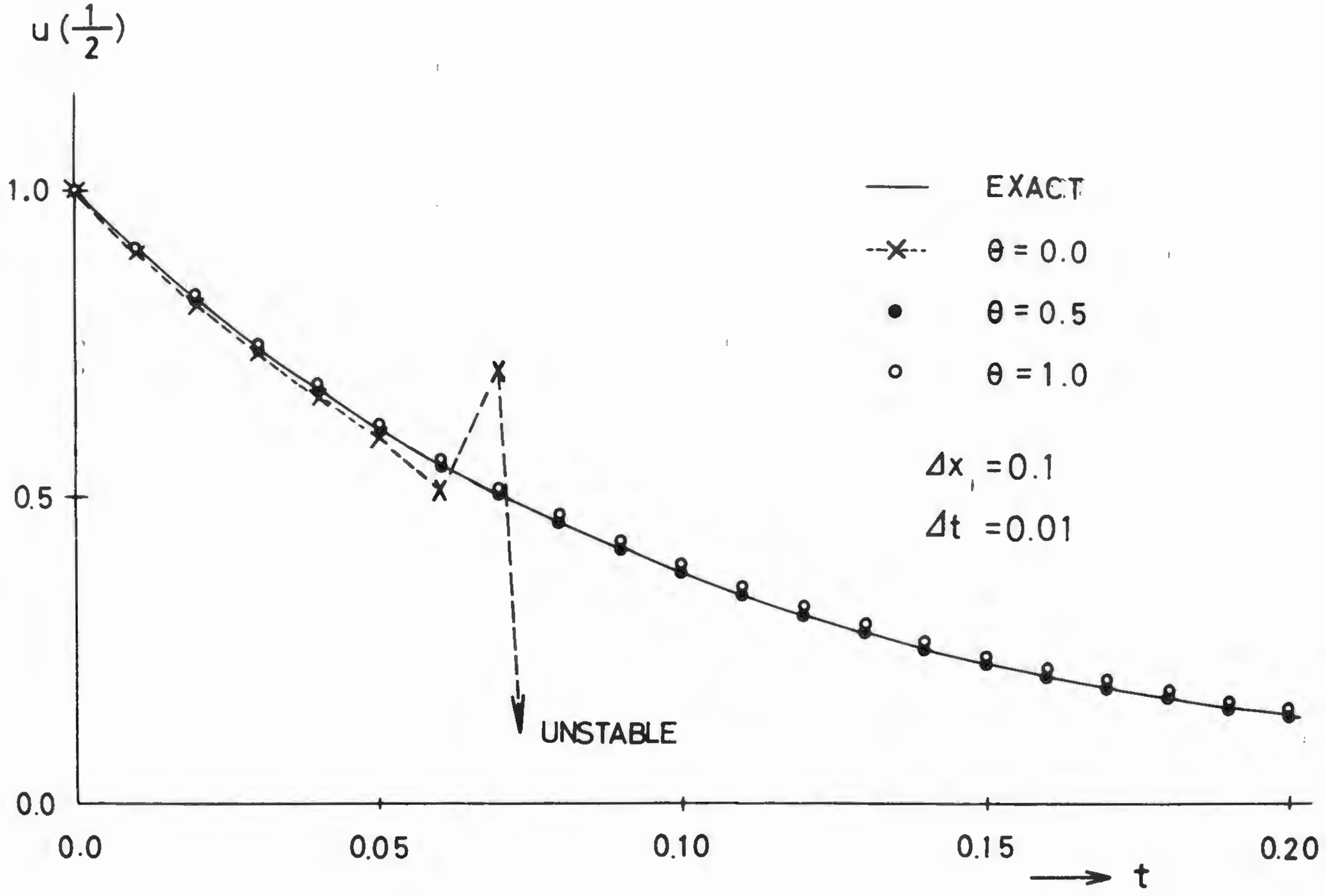

Figure 3. Approximate Solutions of a One-Dimensional Heat Flow Problem

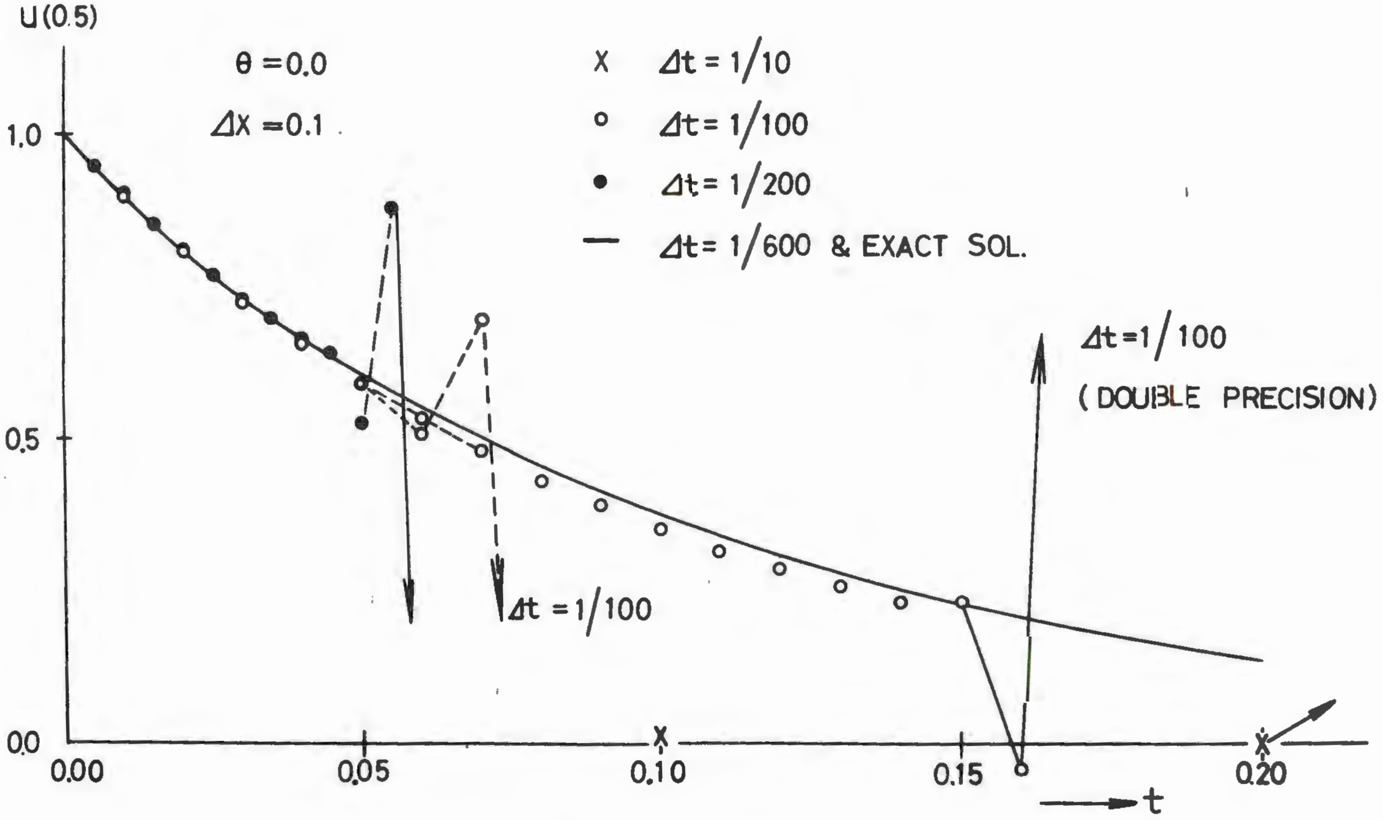

Figure 4. Approximate Solutions of a One-Dimensional Heat Flow Problem — θ = 0.0 —

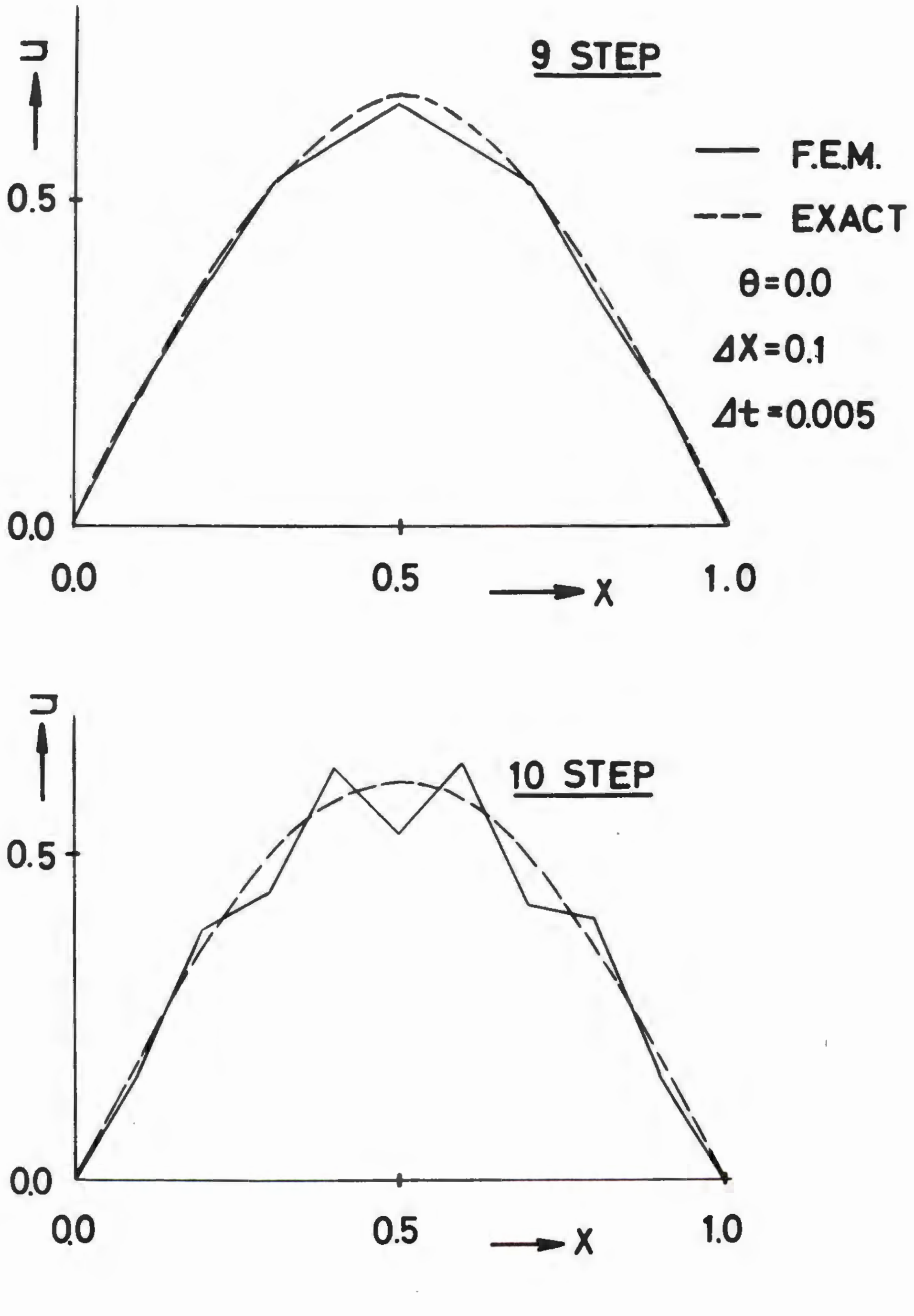

Figure 5. Instability Due to Mingling of Higher Modes — Heat Equation, θ = 0.0 —

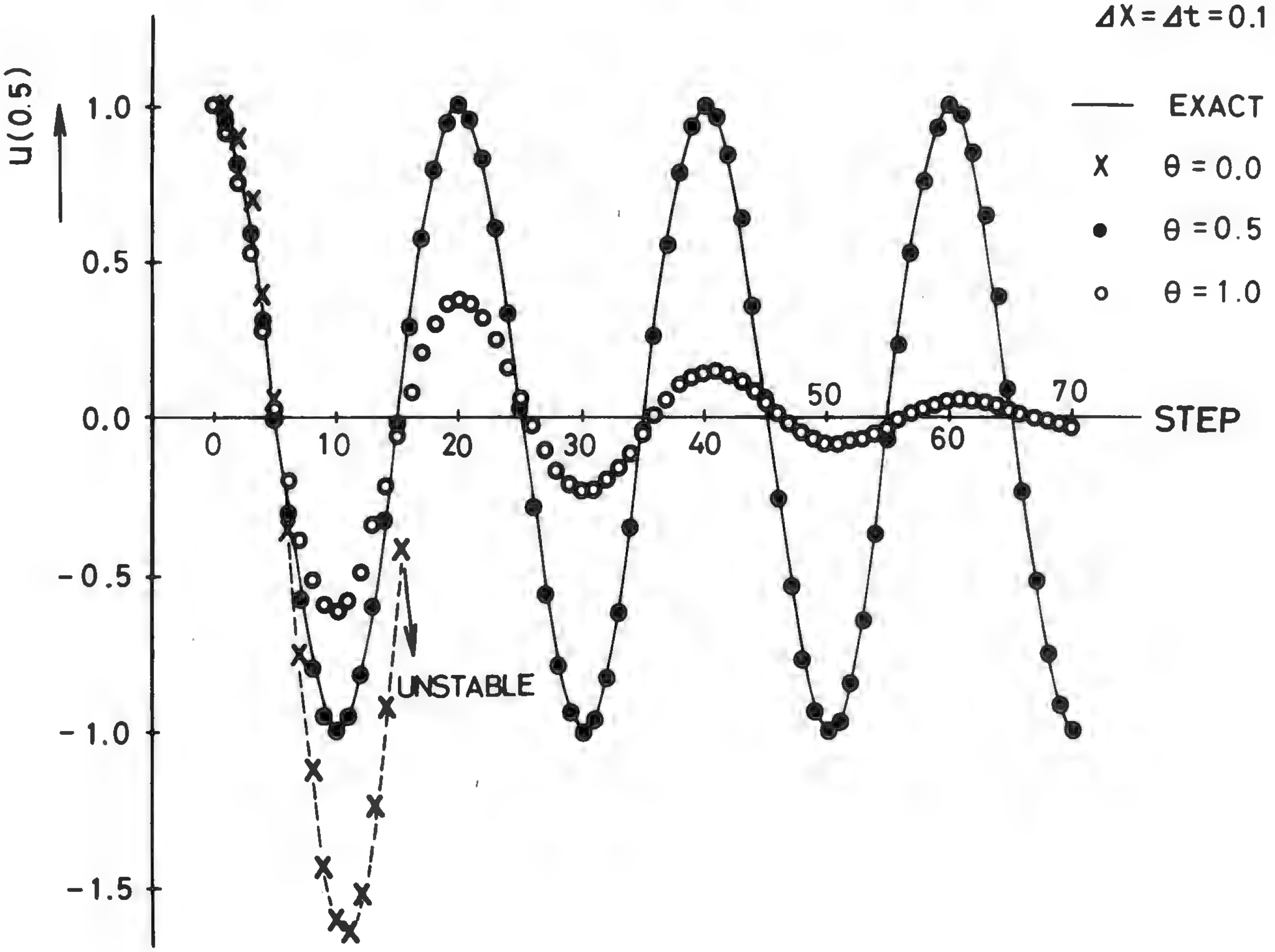

Figure 6. Approximate Solutions of a One-Dimensional Vibration Problem

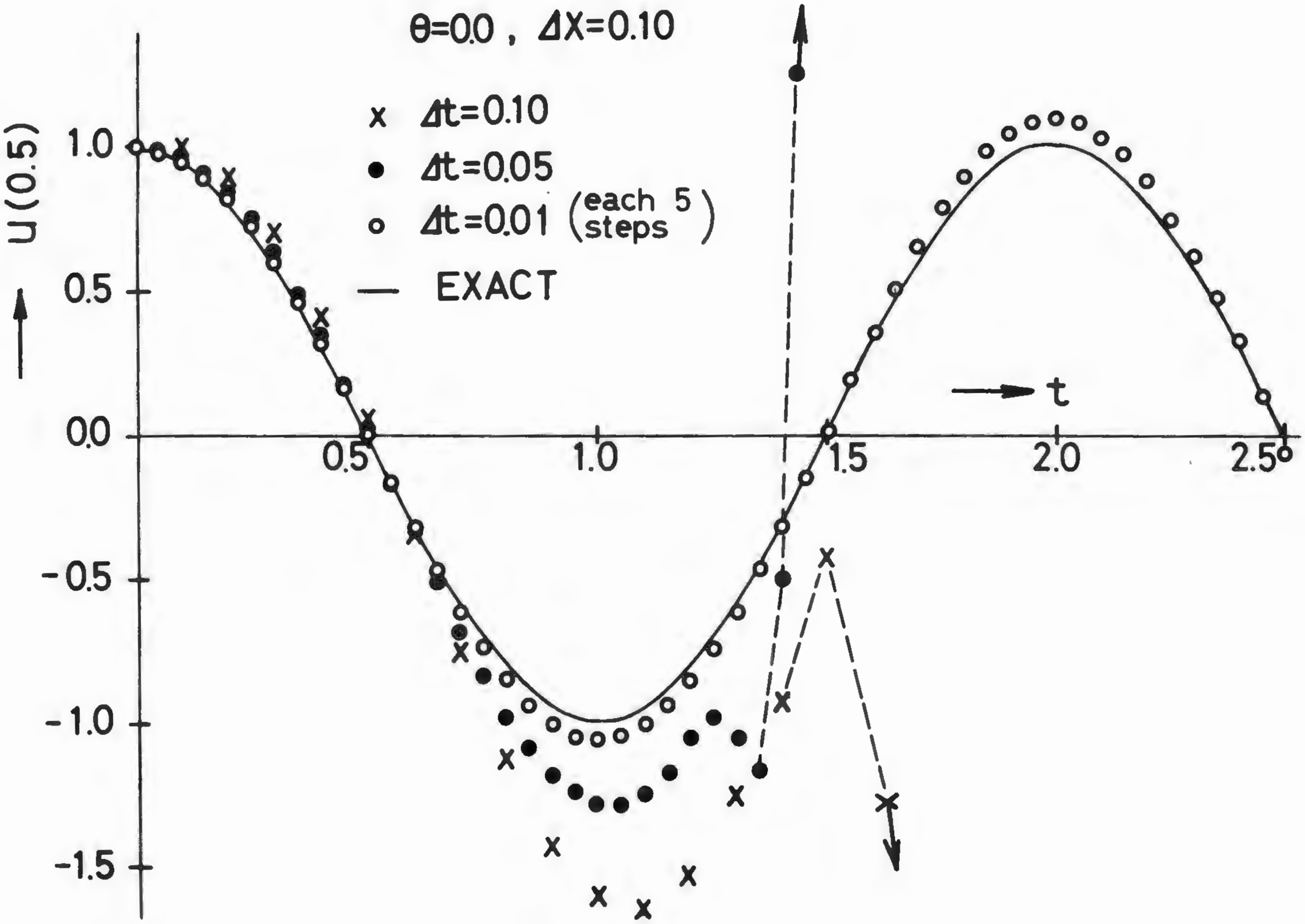

Figure 7. Approximate Solutions of a One-Dimensional Vibrational Problem — θ = 0.0 —

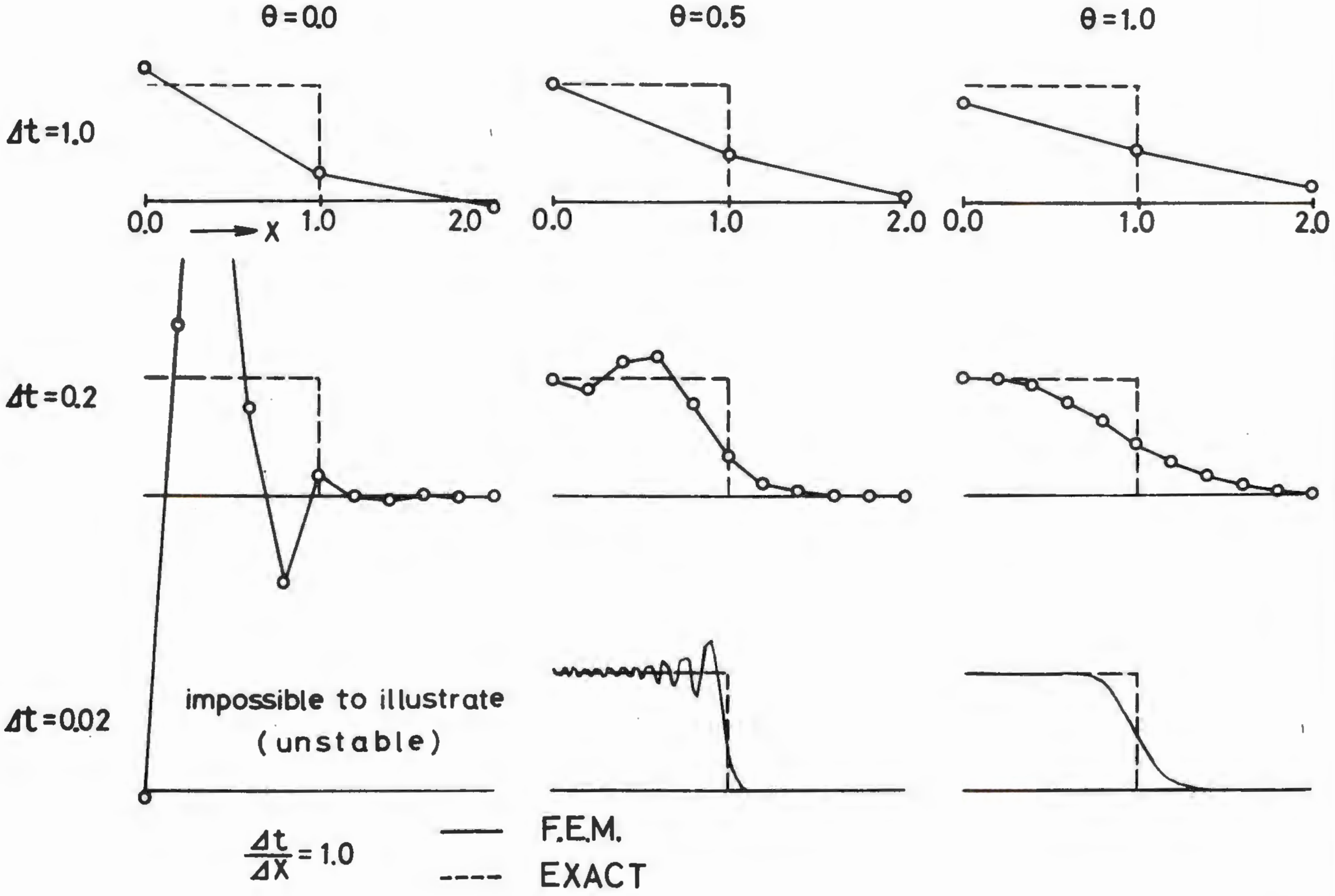

Figure 8. Propagation of a Discontinuous Wave — First Order Wave Equation —

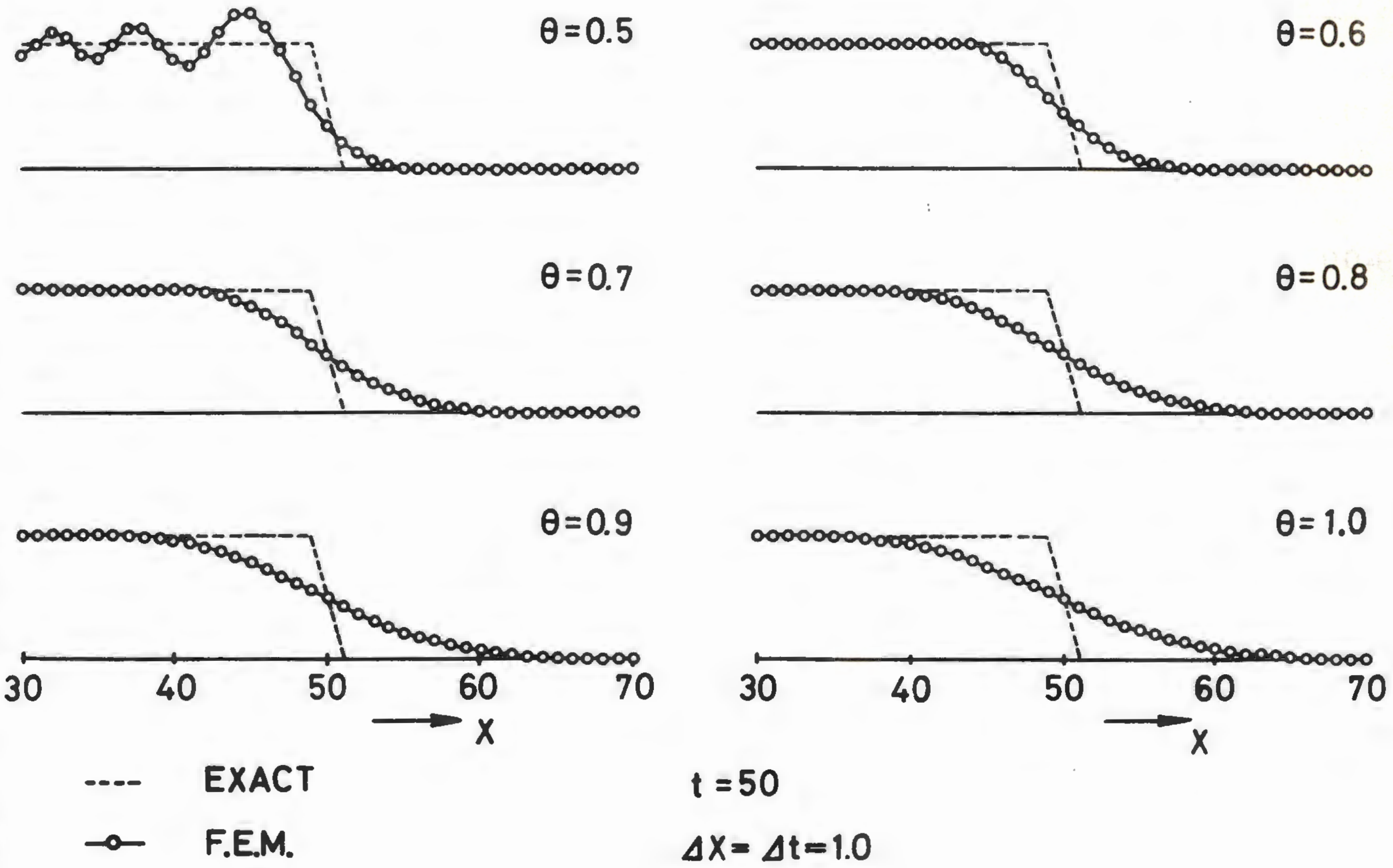

Figure 9. Effect of θ — First Order Wave Equation —

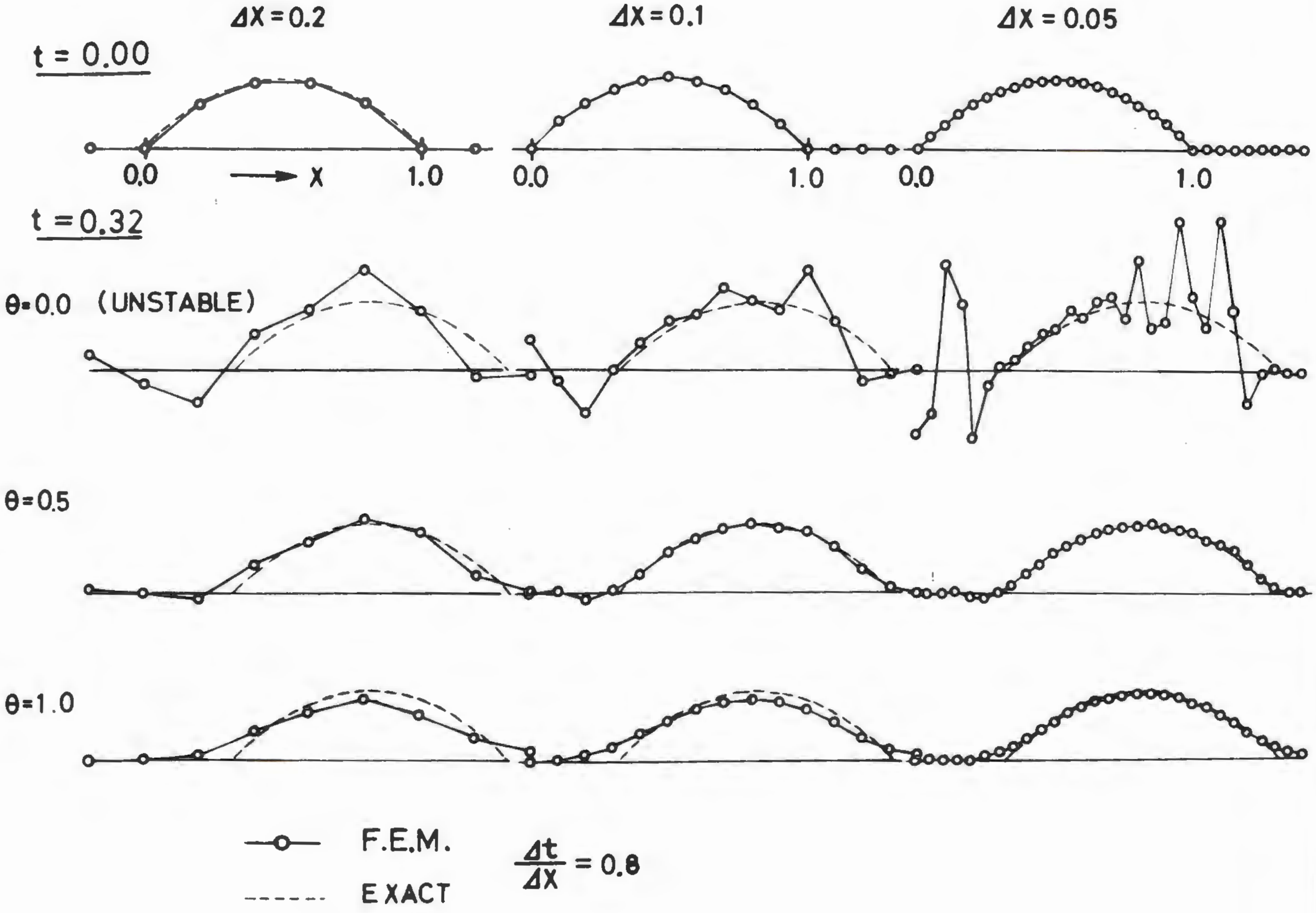

Figure 10. Propagation of a One-Dimensional Wave With Compact Support
— First Order Wave Equation —

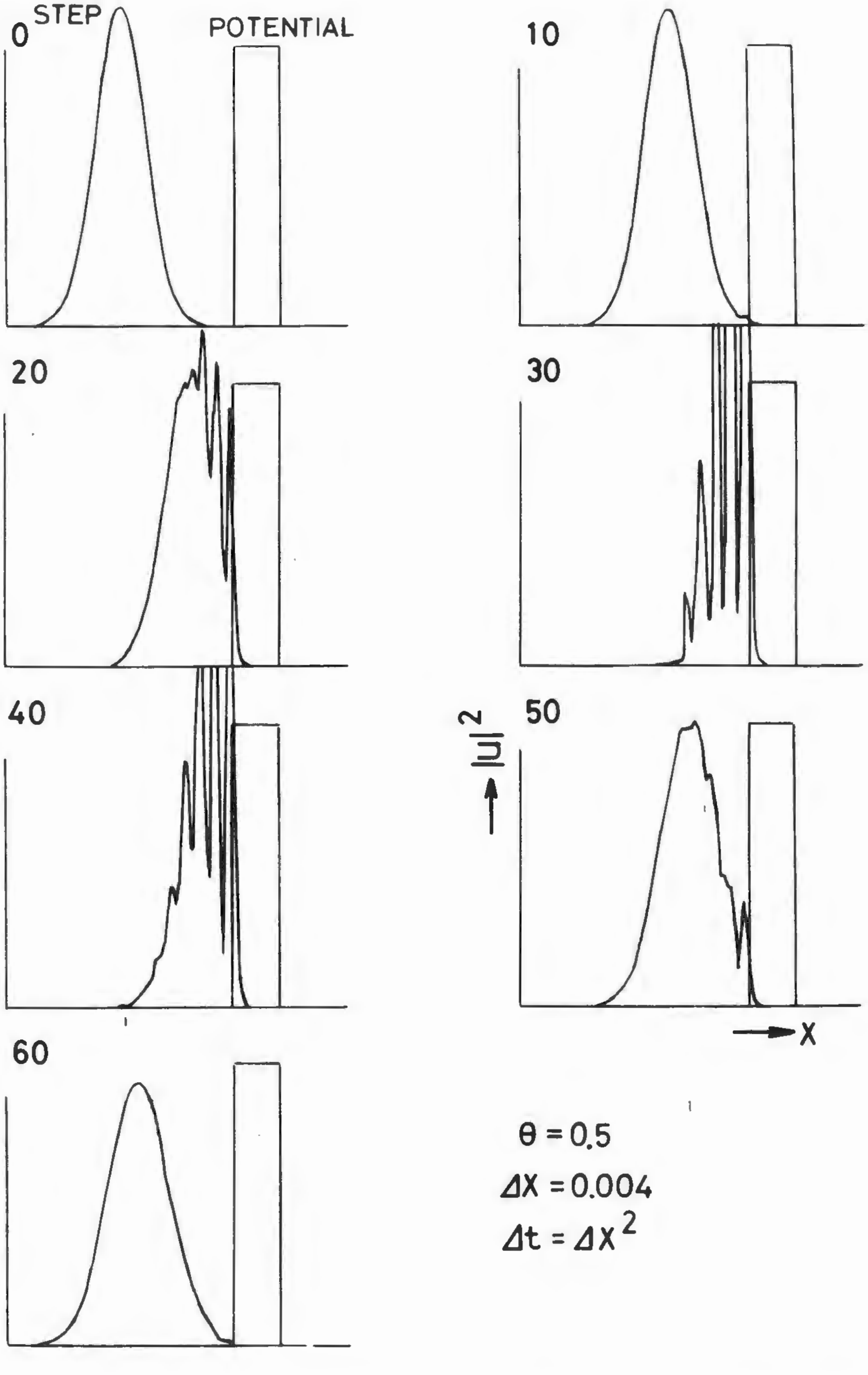

Figure 11. Gaussian Wave Packet Scattering From A Square Barrier — θ = 0.5 —

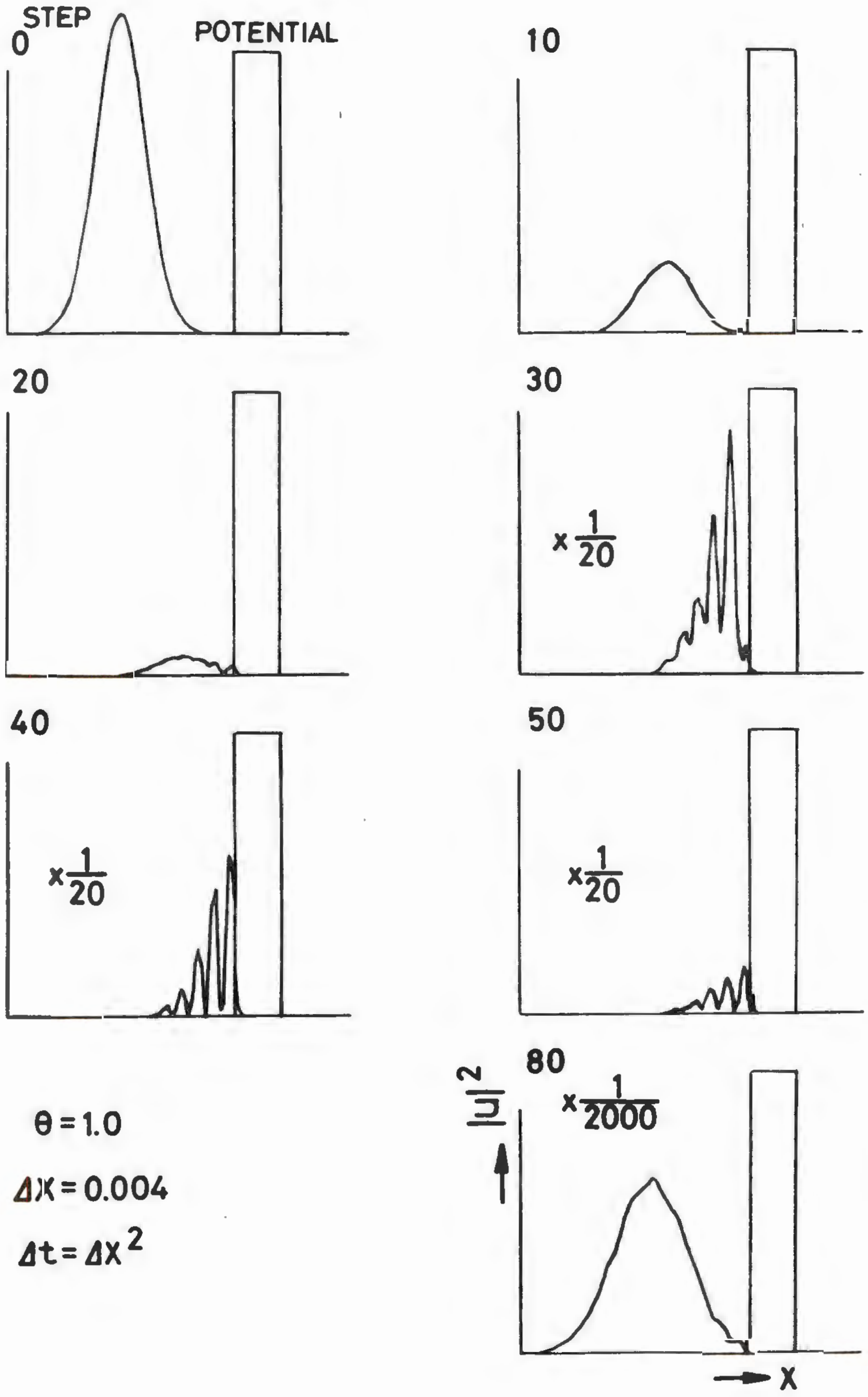

Figure 12. Gaussian Wave Packet Scattering From A Square Barrier — θ = 1.0 —

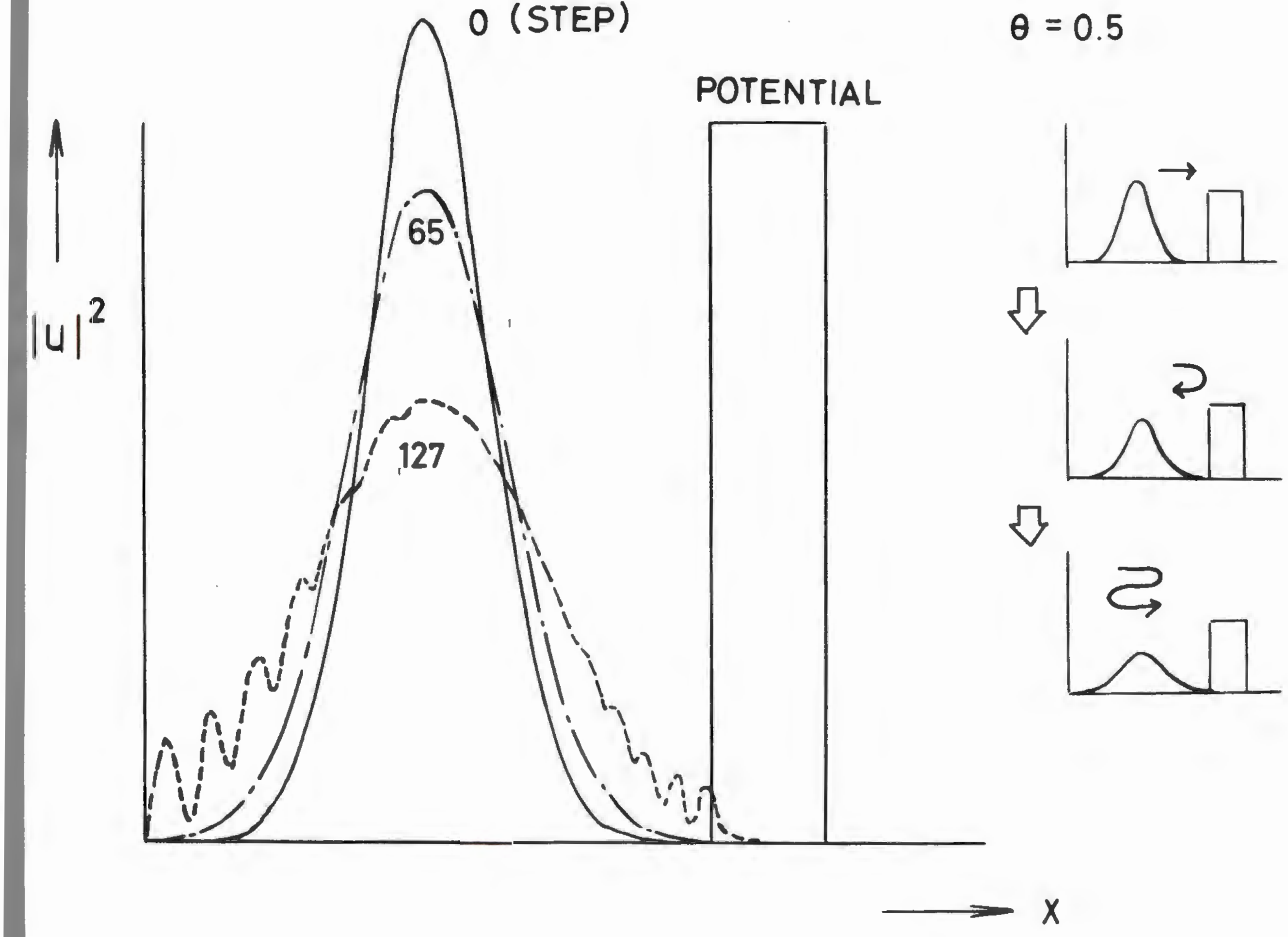

Figure 13. Gaussian Wave Packet Scattering from a Square Barrier and a Rigid Wall
— θ = 0.5 —

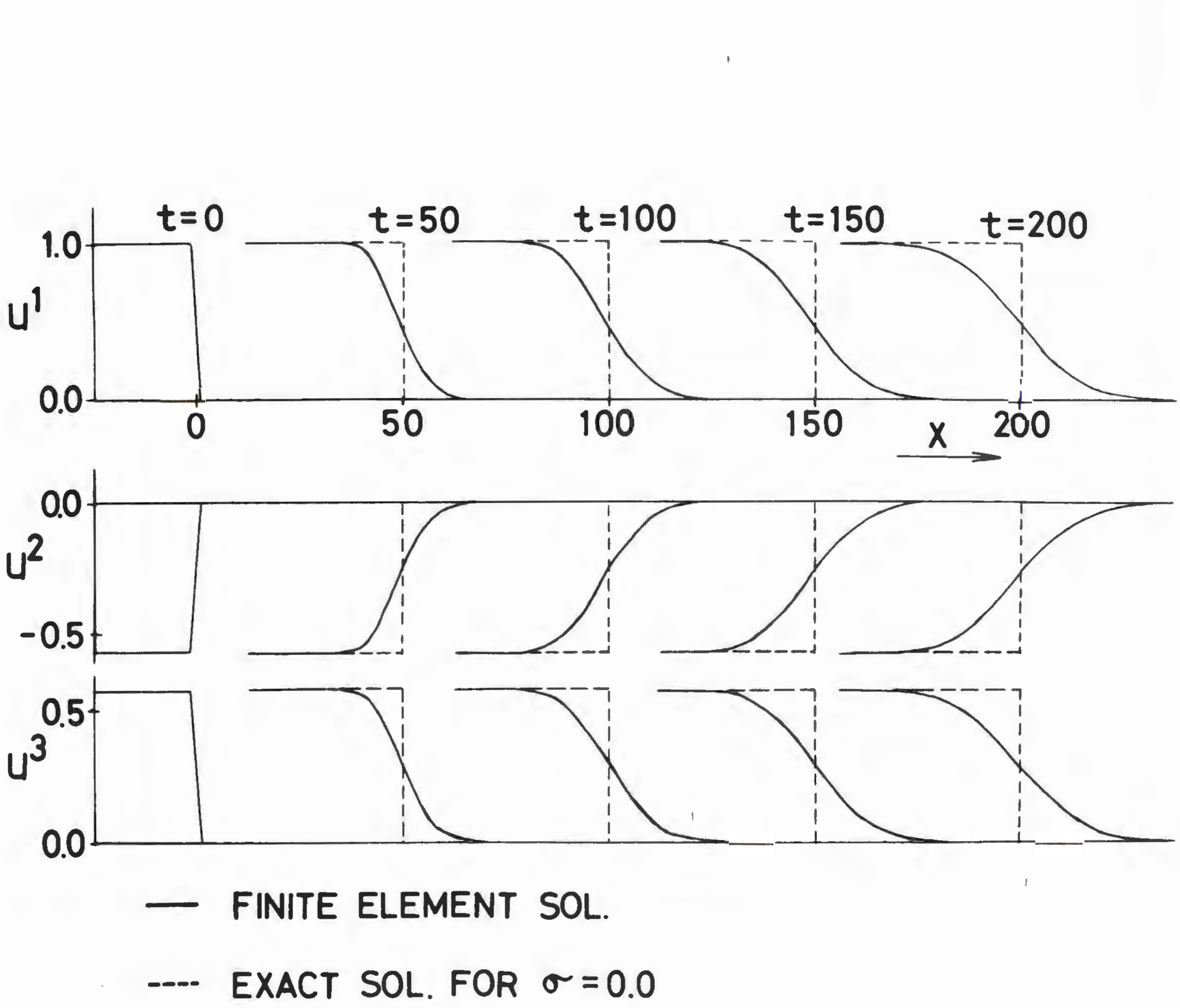

Figure 14. Calculated Profiles of Coupled Sound and Heat Flow — θ = 0.5 —

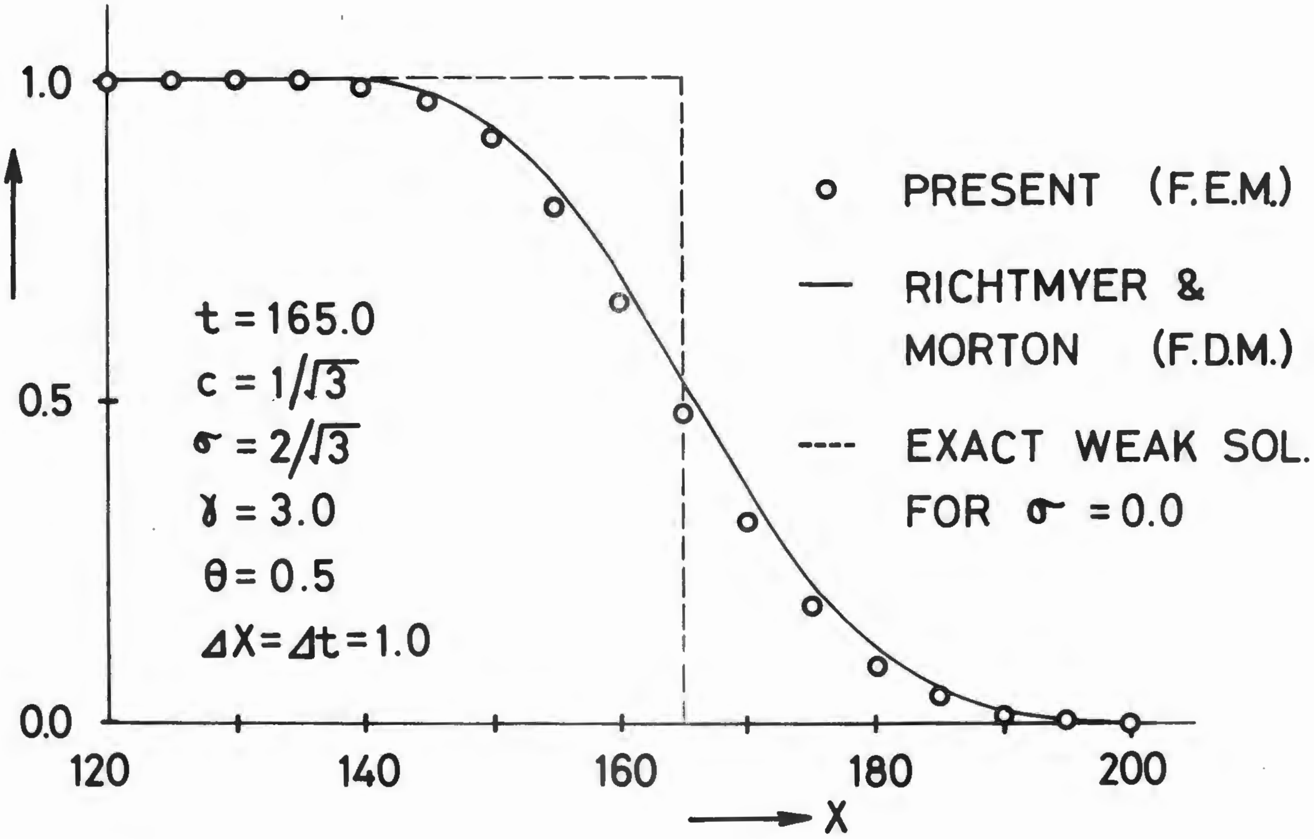

Figure 15. Calculated Profile of an Initially Sharp Sound Wave — $\theta = 0.5$ —

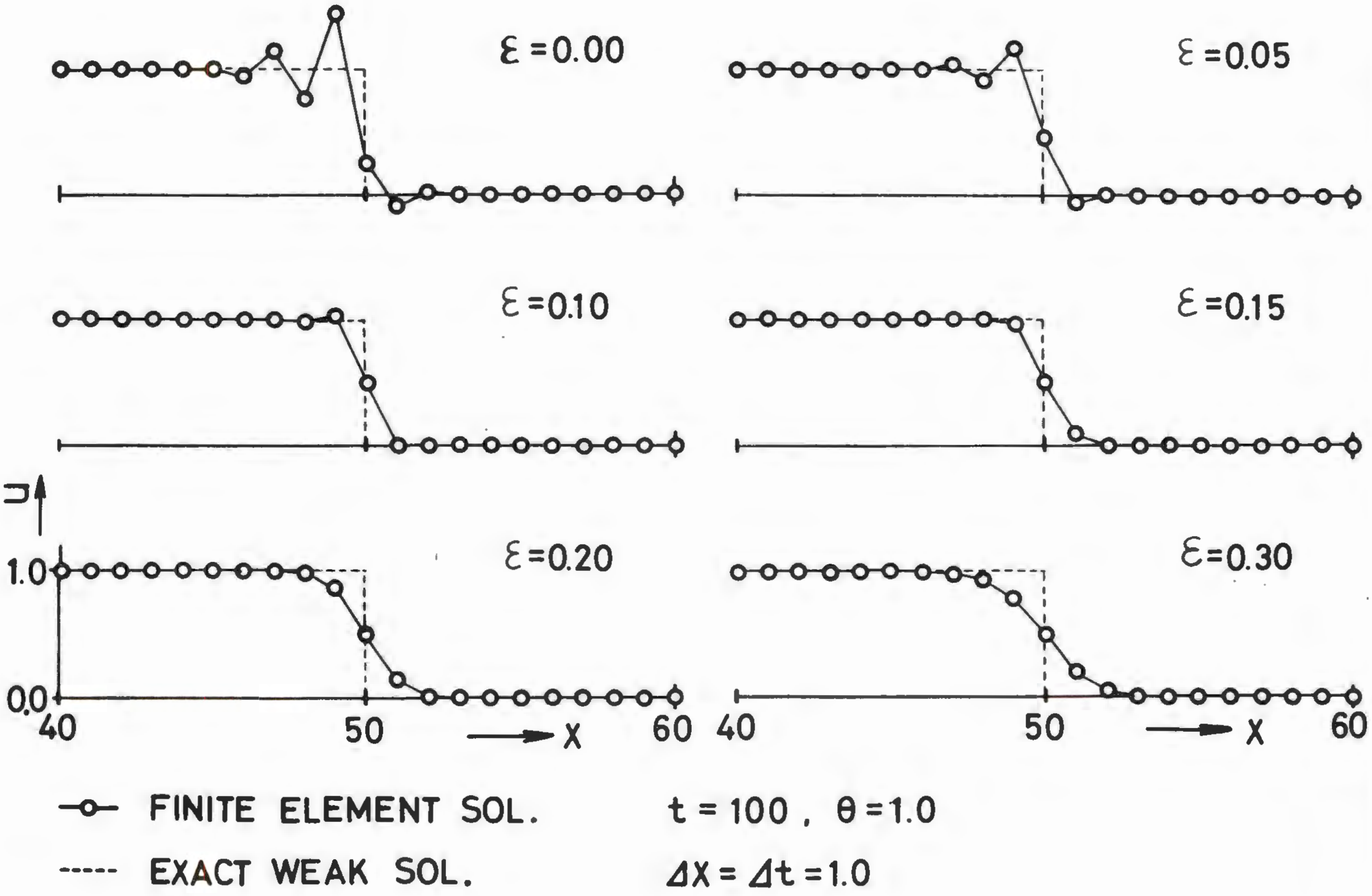

Figure 16. Calculated Profiles of a Nonlinear Shock Wave — θ = 1.0 —

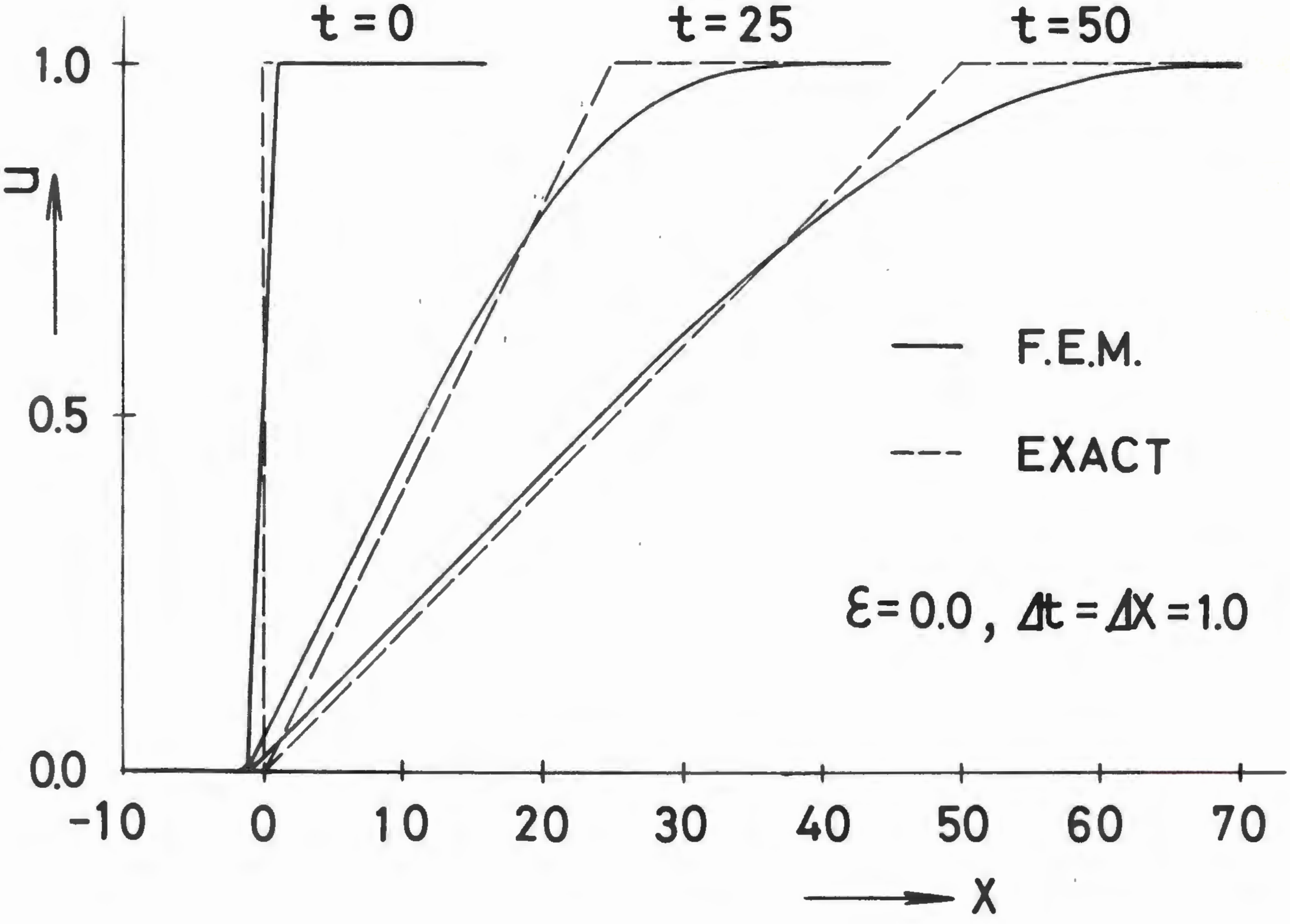

Figure 17. Calculated Profile of a Rarefaction Wave — θ = 1.0 —