Bull. Korean Math. Soc. **35** (1998), No. 1, pp. 157–172

GENERALIZED RUNGE-KUTTA METHODS FOR DYNAMICAL SYSTEMS

DONG WON YU

ABSTRACT. A numerical method is proposed for dynamical systems. We utilize the fact that special matrix exponentials can be exactly evaluated by the intrinsic library functions. Numerical examples are given, which show that the relative errors of the proposed method converge to a small constant and that the method faithfully approximates the dynamics of the nonlinear differential equations.

1. Introduction

We consider numerical approximate solutions for the system of nonlinear ordinary defferential equations

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \qquad \mathbf{y}(0) = \mathbf{y}_0, \tag{1.1}$$

where $\mathbf{f}(t, \mathbf{y})$ is a continuous vector-valued function on $[0, \infty) \times \mathbb{R}^m$ satisfying a Lipschitz condition.

Using $\mathbf{z}(t) = \exp(-tA)\mathbf{y}(t)$, (1.1) is transformed into

$$\mathbf{z}'(t) = \exp(-tA)\{\mathbf{f}(t, \exp(tA)\mathbf{z}(t)) - A\exp(tA)\mathbf{z}(t)\}, \ \mathbf{z}(0) = \mathbf{y}_0. \tag{1.2}$$

Lawson [3] applied the classical fourth order Runge-Kutta method to (1.2) and reverted to the original state by $\mathbf{z}_n = \exp(-t_n A)\mathbf{y}_n$. In this way, Lawson's generalized Runge-Kutta method was obtained.

The choice of A is important. Lawson [3] did not demand that $\partial\mathbf{f}/\partial\mathbf{y} - A = O$, but merely wished to make the eigenvalues of $\partial\mathbf{f}/\partial\mathbf{y} -$

Received June 7, 1997. Revised February 19, 1998.

1991 Mathematics Subject Classification: 65L06, 65L05.

Key words and phrases: Generalized Runge-Kutta methods, Order of convergence, Dynamical systems, Discrete dynamical systems.

This research was partially supported by Chung-Ang University Research Fund, 1996.

A are small enough so that accuracy rather than stability should dictate the step-size in solving (1.2).

The calculation of $\exp(\pm tA)$ is crucial for implementation of Lawson's generalized Runge-Kutta method. Lawson [3] and Ehle & Lawson [1] substituted the Padé approximation for the matrix exponential $\exp(\pm tA)$. Lawson [3] observed that the diagonal Padé approximation provides a set of unconditionally stable approximations to the matrix exponential function and indicated that Lawson's generalized Runge-Kutta method is not a suitable form for practical computation due to the presence of matrix exponentials.

In this paper, we propose a numerical method in which the matrix exponentials are exactly expressed by the library functions and show that the proposed method gives reasonablly good numerical results.

The paper is organized as follows. In Section 2, we prove that $\exp(W)$ can be exactly expressed by the library functions if W is a special scalar skew-symmetric matrix. In Section 3, we give a method how to choose such a scalar skew-symmetric matrix W. In Section 4, we analyze the order of convergence of the proposed method. In Section 5, numerical results of two examples are given. Example 1 shows that the relative error of the proposed method converges to a small constant as n increases. Example 2 shows that the dynamics of the approximate solutions obtained by the proposed method can faithfully describe the true dynamics of nonlinear systems.

2. Preliminaries

It is clear that the set of all $m \times m$ skew-symmetric matrices becomes a $\frac{m(m-1)}{2}$ dimensional vector space with a basis $\mathfrak{B} = \{S_{i,j}\}$, where $i < j, \quad 1 \le i \le m-1, \quad 2 \le j \le m$, and $S_{i,j}$ is a matrix whose (i,j)-element is -1, (j,i)-element is 1, and others are 0.

Choose a set of mutually commutable matrices $S_{i_1,j_1}, S_{i_2,j_2}, \ldots, S_{i_q,j_q}$ in $\mathfrak{B}$ and consider the scalar skew-symmetric matrix of the form

$$W = \omega I + \sum_{k=1}^{q} \mu_k S_{i_k,j_k}, \tag{2.1}$$

where I is the identity matrix and $\omega, \mu_1, \mu_2, \ldots, \mu_q$ are real constants. Such a matrix W will be called an *s-scalar* matrix.

THEOREM 2.1. *For any s-scalar matrix W given in* (2.1), *let* $\exp(W) = (e_{ln})$. *Then each element e_{ln} is expressed as*

$$e_{ln} = \begin{cases} \exp(\omega) & \text{if } l = n \quad (\neq i_k \text{ and } \neq j_k), \\ \exp(\omega)\cos(\mu_k) & \text{if } l = n = i_k, \\ -\exp(\omega)\sin(\mu_k) & \text{if } l = i_k \quad \text{and} \quad n = j_k, \\ \exp(\omega)\sin(\mu_k) & \text{if } l = j_k \quad \text{and} \quad n = i_k, \\ \exp(\omega)\cos(\mu_k) & \text{if } l = n = j_k, \\ 0 & \text{otherwise,} \end{cases}$$

for $k = 1, 2, \ldots, q$.

Proof. Let I_{i_k,j_k} be a matrix whose (i_k, i_k)- and (j_k, j_k)-elements are 1, and others are zero. Then we have

$$(S_{i_k,j_k})^{2n} = (-1)^n I_{i_k,j_k}, \qquad (S_{i_k,j_k})^{2n+1} = (-1)^n S_{i_k,j_k},$$

$$\begin{aligned} \exp(\mu_k S_{i_k,j_k}) &= I + \left\{ \sum_{n=1}^{\infty} (-1)^n \frac{\mu_k^{2n}}{(2n)!} \right\} I_{i_k,j_k} + \left\{ \sum_{n=0}^{\infty} (-1)^n \frac{\mu_k^{2n+1}}{(2n+1)!} \right\} S_{i_k,j_k} \\ &= I - I_{i_k,j_k} + \cos(\mu_k) I_{i_k,j_k} + \sin(\mu_k) S_{i_k,j_k}. \end{aligned}$$

Since $I_{i_k,j_k} S_{i_l,j_l} = O$, $I_{i_k,j_k} I_{i_l,j_l} = O$, and $S_{i_k,j_k} S_{i_l,j_l} = O$ for mutually different i_k, j_k, i_l, and j_l, we obtain

$$\begin{aligned} \exp(W) &= \exp(\omega) \prod_{k=1}^{q} \exp(\mu_k S_{i_k,j_k}) \\ &= \exp(\omega) \left\{ I - \sum_{k=1}^{q} I_{i_k,j_k} + \sum_{k=1}^{q} \cos(\mu_k) I_{i_k,j_k} + \sum_{k=1}^{q} \sin(\mu_k) S_{i_k,j_k} \right\}. \end{aligned}$$

This completes the proof. □

3. GS processes

In the previous section, we have shown that if W is an s-scalar matrix as in (2.1), then $\exp(tW)$ can be computed exactly by library functions. We will give a way how to choose such an s-scalar matrix W from (1.1).

Suppose that $\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(t, \mathbf{0})$ is a constant matrix A. Then (1.1) is rewritten by

$$\mathbf{y}' = A\mathbf{y}(t) + \mathbf{u}(t, \mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{y}_0, \tag{3.1}$$

where $\mathbf{u}(t, \mathbf{y}(t)) = \mathbf{f}(t, \mathbf{y}(t)) - A\mathbf{y}(t)$.

If A has real eigenvalues λ_j, $1 \le j \le p$, and complex eigenvalues $\lambda_{p+j} \pm i\mu_j$, $1 \le j \le q$, where $p+2q = m$, then there exists an invertible matrix

$$P = [\mathbf{v}_1, \ldots, \mathbf{v}_p, \mathbf{w}_1, \mathbf{v}_{p+1}, \ldots, \mathbf{w}_q, \mathbf{v}_{p+q}]. \tag{3.2}$$

Here $\mathbf{v}_j$ and $\mathbf{v}_{p+j} + i\mathbf{w}_j$ are eigenvectors or generalized eigenvectors of A corresponding to λ_j and $\lambda_{p+j} + i\mu_j$, respectively. Furthermore, there is only one way of expressing A as $S+N$, where S is semisimple, N is nilpotent, and $SN = NS$ (see [2] p. 116 and [4] p. 39). Hence the Jordan canonical form of A is splitted as

$$\bar{A} = P^{-1}AP = P^{-1}SP + P^{-1}NP = \bar{S} + \bar{N},$$

where $\bar{S}$ is represented by

$$\bar{S} = \sum_{j=1}^{p} \lambda_j I_{j,j} + \sum_{j=1}^{q} \{\lambda_{p+j} I_{p+2j-1,p+2j} + \mu_j S_{p+2j-1,p+2j}\},$$

and said to be an *s-diagonal* matrix.

Transforming (3.1) by $\mathbf{y}(t) = P\mathbf{x}(t)$, we have

$$\mathbf{x}'(t) = \{\bar{S} + \bar{N}\}\mathbf{x}(t) + P^{-1}\mathbf{u}(t, P\mathbf{x}(t)), \quad \mathbf{x}(0) = P^{-1}\mathbf{y}_0. \tag{3.3}$$

Let $x_j(t)$ and v_j denote the jth component of $\mathbf{x}(t)$ and $P^{-1}\mathbf{u}(t, P\mathbf{x}(t))$ of (3.3) as

$$\mathbf{x}(t) \equiv (x_1(t), x_2(t), \cdots, x_m(t))^T$$
$$P^{-1}\mathbf{u}(t, P\mathbf{x}(t)) \equiv (v_1, \cdots, v_p, v_{p+1}, \cdots, v_{p+2q})^T.$$

We first consider a method extracting an s-diagonal matrix like $\bar{S}$ in (3.3) from $P^{-1}\mathbf{u}(t, P\mathbf{x}(t))$.

(1) For each j $(1 \leq j \leq p)$, express v_j as

$$v_j = \varphi_j(t, \mathbf{x}(t))\, x_j(t) + \zeta_j(t, \mathbf{x}(t)),$$

where φ_j, $\zeta_j : [0, \infty) \times \mathbb{R}^m \to \mathbb{R}$ and $\zeta_j(t, \mathbf{x}(t))$ has no terms divisible by $x_j(t)$.

(2) For each pair $p + 2k - 1$ and $p + 2k$ $(1 \leq k \leq q)$, express v_{p+2k-1} and v_{p+2k} as

$$\begin{aligned} v_{p+2k-1} &= \bar{\varphi}_{p+2k-1}(t, \mathbf{x}(t))\, x_{p+2k-1}(t) - \bar{\psi}_{p+2k-1}(t, \mathbf{x}(t))\, x_{p+2k}(t) \\ &\quad + \zeta_{p+2k-1}(t, \mathbf{x}(t)), \\ v_{p+2k} &= \bar{\psi}_{p+2k}(t, \mathbf{x}(t))\, x_{p+2k-1}(t) + \bar{\varphi}_{p+2k}(t, \mathbf{x}(t))\, x_{p+2k}(t) \\ &\quad + \zeta_{p+2k}(t, \mathbf{x}(t)), \end{aligned}$$

where $\bar{\varphi}_i, \bar{\psi}_i, \zeta_i : [0, \infty) \times \mathbb{R}^m \to \mathbb{R}$, and $\zeta_{p+2k-1}(t, \mathbf{x}(t))$ and $\zeta_{p+2k}(t, \mathbf{x}(t))$ have no terms divisible by $x_{p+2k-1}(t)$ and $x_{p+2k}(t)$.

If $\bar{\varphi}_{p+2k-1}(t, \mathbf{x}(t)) = \bar{\varphi}_{p+2k}(t, \mathbf{x}(t))$ and $\bar{\psi}_{p+2k-1}(t, \mathbf{x}(t)) = \bar{\psi}_{p+2k}(t, \mathbf{x}(t))$, let

$$\varphi_{p+k}(t, \mathbf{x}(t)) \equiv \bar{\varphi}_{p+2k-1}(t, \mathbf{x}(t)) = \bar{\varphi}_{p+2k}(t, \mathbf{x}(t)),$$
$$\psi_k(t, \mathbf{x}(t)) \equiv \bar{\psi}_{p+2k-1}(t, \mathbf{x}(t)) = \bar{\psi}_{p+2k}(t, \mathbf{x}(t)).$$

Otherwise, let $\varphi_{p+k}(t, \mathbf{x}(t)) \equiv 0$ and $\psi_k(t, \mathbf{x}(t)) \equiv 0$ for $k = 1, 2, \cdots, q$.

Define the s-diagonal matrix $S(t, \mathbf{x}(t))$ as follows:

$$
\begin{aligned}
S(t, \mathbf{x}(t)) &= \sum_{j=1}^{p} \varphi_j(t, \mathbf{x}(t))\, I_{j,j} \\
&+ \sum_{k=1}^{q} \Big\{ \varphi_{p+k}(t, \mathbf{x}(t))\, I_{p+2k-1,p+2k} + \psi_k(t, \mathbf{x}(t))\, S_{p+2k-1,p+2k} \Big\},
\end{aligned} \tag{3.4}
$$

where φ_i, ψ_i are obtained from (1) - (2).

If there is no $\varphi_i(t, \mathbf{x}(t)) = 0$ $(1 \le i \le p+q)$, and there is some k such that

$$\varphi_k(t, \mathbf{x}(t)) \ge \varphi_i(t, \mathbf{x}(t)) \quad i = 1, 2, \cdots, p+q, \quad t \ge 0,$$

let $\varphi(t, \mathbf{x}(t)) \equiv \varphi_k(t, \mathbf{x}(t))$. Otherwise, put $\varphi(t, \mathbf{x}(t)) \equiv 0$.

Finally, we define a constant s-scalar matrix W from the s-diagonal matrices $\bar{S}$ in (3.3) and $S(t, \mathbf{x}(t))$ in (3.4) as

$$W = \omega_n I + \frac{1}{2}\{\bar{S} - \bar{S}^T\} + \frac{1}{2}\{S(t_n, \mathbf{x}_n) - S^T(t_n, \mathbf{x}_n)\}, \tag{3.5}$$

where $\omega_n = \max_{1 \le j \le p+q}\{\lambda_j\} + \varphi(t_n, \mathbf{x}_n)$ (see Example 2 in §5).

Using such a matrix W, the system (3.3) can be rewritten by

$$\mathbf{x}'(t) = W\,\mathbf{x}(t) + \mathbf{U}(t, \mathbf{x}(t)), \qquad \mathbf{x}(t_n) = \mathbf{x}_n, \tag{3.6}$$

where $\mathbf{U}(t, \mathbf{x}(t)) = (\bar{S} + \bar{N} - W)\mathbf{x}(t) + P^{-1}\mathbf{u}(t, P\mathbf{x}(t))$.

We now propose a *GS process (Runge-Kutta-GS process)* which corresponds to a Runge-Kutta method by following steps:

(1) Apply the transformation $\mathbf{z}(t) = \exp(-tW')\mathbf{x}(t)$ to (3.6), then we have

$$\mathbf{z}'(t) = \exp(-tW)\mathbf{U}(t, \exp(tW)\mathbf{z}(t)), \qquad \mathbf{z}(t_n) = \mathbf{z}_n. \tag{3.7}$$

(2) Apply a Runge-Kutta method to (3.7), then we have

$$
\begin{aligned}
\mathbf{z}_{n+1} &:= \mathbf{z}_n + h\sum_{i=1}^{s} b_i \mathbf{k}_i, \quad \text{for} \quad n = 0, 1, 2, \ldots, \\
\mathbf{k}_i &:= \exp(-t_{n,i}W)\mathbf{U}(t_{n,i}, \exp(t_{n,i}W)\mathbf{z}_{n,i}), \\
\mathbf{z}_{n,i} &:= \mathbf{z}_n + h\sum_{j=1}^{s} d_{ij}\mathbf{k}_j,
\end{aligned} \tag{3.8}
$$

where $t_{n,i} = t_n + c_i h$, s is the number of stages, d_{ij} and b_i are constants, $\{c_i\}$ is monotone in $[0,1]$, and $c_i = \sum_{j=1}^{s} d_{ij}$ $(1 \le i, j \le s)$.

(3) Revert (3.8) by $\mathbf{z}_{n,i} = \exp(-t_{n,i}W)\mathbf{x}_{n,i}$, then we have

$$
\begin{aligned}
\mathbf{x}_{n+1} &:= \exp(hW)\left\{\mathbf{x}_n + h\sum_{i=1}^{s} b_i \exp(-c_i hW)\bar{\mathbf{k}}_i\right\}, \\
\bar{\mathbf{k}}_i &:= \mathbf{U}(t_{n,i}, \mathbf{x}_{n,i}), \\
\mathbf{x}_{n,i} &:= \exp(c_i hW)\left\{\mathbf{x}_n + h\sum_{j=1}^{s} d_{ij}\exp(-c_j hW)\bar{\mathbf{k}}_j\right\}.
\end{aligned} \tag{3.9}
$$

Here $\exp(\pm c_j hW)$ can be exactly expressed by Theorem 2.1.

(4) Find the solution of (3.1) by $\mathbf{y}_n = P\mathbf{x}_n$.

REMARK 3.1. If we apply the Euler method to (3.7) and revert by $\mathbf{z}_n = \exp(-t_n W)\mathbf{x}_n$, then its corresponding GS process *(Euler-GS process)* is derived as follows:

$$
\mathbf{x}_{n+1} = \exp(hW)\Big\{\mathbf{x}_n + h\mathbf{U}(t_n, \mathbf{x}_n)\Big\}. \tag{3.10}
$$

REMARK 3.2. Let a linear multistep method be specified by a positive integer s and constants α_i, β_i, $i = 0, 1, \ldots, s$ with $\alpha_s = 1$. If we apply the linear multistep method to (3.7) and revert by $\mathbf{z}_n = \exp(-t_n W)\mathbf{x}_n$, then its corresponding GS process *(linear multistep-GS process)* is derived as follows:

$$
\sum_{i=0}^{s} \alpha_i \exp(-ihW)\mathbf{x}_{n+i} = h\sum_{i=0}^{s} \beta_i \exp(-ihW)\mathbf{U}(t_{n+i}, \mathbf{x}_{n+i}).
$$

4. Order of convergence

Let us consider the local truncation error $\mathbf{E}_n = \mathbf{y}(t_n) - \hat{\mathbf{y}}_n$. Here $\hat{\mathbf{y}}_n$ represents the numerical solution evaluated with the exact initial value $\mathbf{y}(t_{n-1})$.

THEOREM 4.1. *If a numerical method is an rth order method, then its corresponding GS process is also an rth order method.*

Proof. If an rth order method is applied to (3.7), then its local truncation error is given by

$$\mathbf{z}(t_{n+1}) - \hat{\mathbf{z}}_{n+1} = \frac{h^{r+1}}{(r+1)!}\mathbf{z}^{(r+1)}(t_n + \xi) \quad \text{for some} \quad \xi \in (0, h).$$

From the fact $\mathbf{z}(t) = \exp(-tW)\mathbf{x}(t)$, we obtain

$$\mathbf{z}^{(r+1)}(t_n+\xi) = \exp(-(t_n+\xi)W)\sum_{j=0}^{r+1}(-1)^j\binom{r+1}{j}W^j\mathbf{x}^{(r+1-j)}(t_n+\xi).$$

Hence

$$\mathbf{z}(t_{n+1}) - \hat{\mathbf{z}}_{n+1} = \frac{h^{r+1}}{(r+1)!}\exp(-(t_n+\xi)W)\sum_{j=0}^{r+1}(-1)^j\binom{r+1}{j}W^j\mathbf{x}^{(r+1-j)}(t_n+\xi).$$

Then

$$\mathbf{E}_{n+1} = P\{\mathbf{x}(t_{n+1}) - \hat{\mathbf{x}}_{n+1}\} = P\exp(t_{n+1}W)\{\mathbf{z}(t_{n+1}) - \hat{\mathbf{z}}_{n+1}\}.$$

Since $\mathbf{x}^{(r+1-j)}(t_n+\xi) = P^{-1}\mathbf{y}^{(r+1-j)}(t_n+\xi)$, we have

(4.1)

$$\mathbf{E}_{n+1} = \frac{h^{r+1}}{(r+1)!}P\exp((h-\xi)W)\sum_{j=0}^{r+1}(-1)^j\binom{r+1}{j}W^jP^{-1}\mathbf{y}^{(r+1-j)}(t_n+\xi)$$

for some $\xi \in (0, h)$. The proof is completed. □

In particular, if the problem (3.1) is homogeneous and linear, then (3.1) can be transformed into

$$\mathbf{x}'(t) = (\bar{S} + \bar{N})\mathbf{x}(t), \tag{4.2}$$

and its local truncation error of an rth order GS process is obtained by (4.1) as

$$\mathbf{E}_{n+1} = \frac{h^{r+1}}{(r+1)!} P \exp[(h-\xi)W] \times \left\{ \sum_{j=0}^{r+1} (-1)^j \binom{r+1}{j} W^j (\bar{S}+\bar{N})^{r+1-j} \right\} P^{-1}\mathbf{y}(t_n+\xi) \tag{4.3}$$

for some $\xi \in (0, h)$.

COROLLARY 4.1. (1) *If $\bar{S}$ is an s-scalar matrix and $\bar{N} = O$ in* (4.2), *then its solution can be accurately evaluated by GS process.*

(2) *Its local truncation error is zero for each time step.*

Proof. (1) Since $W = \bar{S}$ and $\mathbf{U}(t, \mathbf{x}) = \mathbf{0}$ in (3.6), we have

$$\mathbf{x}_{n+1} = \exp(hW)\mathbf{x}_n = \{\exp(hW)\}^{n+1}\mathbf{x}(0) = \exp(t_{n+1}W)\mathbf{x}(0) = \mathbf{x}(t_{n+1}),$$

where $\{\exp(hW)\}^{n+1}$ is computed by Theorem 2.1. Hence $\mathbf{y}_{n+1} = \mathbf{y}(t_{n+1})$.

(2) Since $\bar{A} = P^{-1}AP = \bar{S} = W$, it follows from (4.3) that

$$\mathbf{E}_{n+1} = \frac{h^{r+1}}{(r+1)!} P \exp[(h-\xi)hW]\{\bar{S}-W\}^{r+1} P^{-1}\mathbf{y}(t_n+\xi) = O. \quad \square$$

COROLLARY 4.2. *If $\bar{S}$ is an s-scalar matrix and the degree of nilpotency of $\bar{N}$ is less than or equal to $r+1$ in* (4.2), *then the local truncation error of an rth order GS process is zero.*

Proof. Since $\bar{A} = P^{-1}AP = \bar{S} + \bar{N}$ and $\bar{S}\bar{N} = \bar{N}\bar{S}$, we have $\bar{A}\bar{S} = \bar{S}\bar{A}$. Since $W = \bar{S}$ and $\bar{N}^{r+1} = O$, we arrive from (4.3) at

$$\mathbf{E}_{n+1} = \frac{h^{r+1}}{(r+1)!} P \exp((h-\xi)W)\{\bar{A}-W\}^{r+1} P^{-1}\mathbf{y}(t_n+\xi) = O. \quad \square$$

5. Numerical examples

For the convenience of numerical computation we consider the Euler method which is a special case of Runge-Kutta methods and linear multistep methods.

We compare numerical results of Euler method, Lawson's generalized Euler method, Euler-GS process (3.10), and Heun's method. We also compare the dynamics of numerical solutions obtained by the Euler method, Sanz-Serna's recursion [5], and Euler-GS process (3.10). Numerical computations are done by double precision using a personal computer.

EXAMPLE 1. Consider a nonlinear initial-value problem

$$\frac{d}{dt}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a & b \\ b & a \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} u^2 - v^2 \\ u^2 - v^2 \end{pmatrix}, \quad \begin{pmatrix} u(0) \\ v(0) \end{pmatrix} = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix}, \tag{5.1}$$

where a and b are real. The general solution of (5.1) is given by

$$\begin{pmatrix} u(t) \\ v(t) \end{pmatrix} = \begin{pmatrix} c_1 \exp[(a+b)t + \frac{4c_2}{a-b}\exp((a-b)t)] + c_2 \exp[(a-b)t] \\ c_1 \exp[(a+b)t + \frac{4c_2}{a-b}\exp((a-b)t)] - c_2 \exp[(a-b)t] \end{pmatrix}.$$

Since the matrix P of $A = \frac{\partial}{\partial \mathbf{x}}\mathbf{f}(t, \mathbf{0}) = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$ is given by $P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, the system (5.1) is transformed by $\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$. Then we have

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a+b & 0 \\ 0 & a-b \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 4xy \\ 0 \end{pmatrix}, \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} 0.5(u_0 + v_0) \\ 0.5(u_0 - v_0) \end{pmatrix}$$

It follows (3.5) that the system (3.6) is given by

$$W = \begin{pmatrix} \alpha[A] & 0 \\ 0 & \alpha[A] \end{pmatrix} \quad \text{and} \quad \mathbf{U}(t, \mathbf{x}(t)) = \begin{pmatrix} (a+b-\alpha[A])x + 4xy \\ (a-b-\alpha[A])y \end{pmatrix},$$

where $\alpha[A] = \max\{a+b, a-b\}$.

The following table shows the numerical results for (5.1) with $h = 0.01, a = -2$, $b = 3$, $u(0) = 1.4493$, and $v(0) = -0.55067$.

n	e_n	$\hat{e}_n$	$\bar{e}_n$	$e_n[Heun]$
1	.95939E-03	-.47246E-04	.12051E-02	.12653E-04
100	.20182E-01	-.11771E-01	.16056E-01	.55134E-04
500	.39882E-01	-.11933E-01	.16625E-01	.11501E-03
1000	.63433E-01	-.11385E-01	.16625E-01	.19771E-03
2000	.10882E+00	.93226E+00	.16625E-01	.36308E-03
3000	.15200E+00	.10000E+01	.16625E-01	.52845E-03
4000	.19293E+00	.10000E+01	.16625E-01	.36175E-03
5000	.99495E+00	.10000E+01	.16625E-01	.99496E+00
6000	.10000E+01	.10000E+01	.16625E-01	.10000E+01
10000	.10000E+01	.10000E+01	.16625E-01	.10000E+01

In the table, $e_n \equiv \frac{u(nh)-u_n}{u(nh)}$, $\hat{e}_n$, $\bar{e}_n$, and $e_n[Heun]$ denote the relative errors of the Euler method, Lawson's generalized Euler method, Euler-GS process (3.10), and Heun's method at $t = nh$, respectively. The Padé approximation $E(A) = (I - 3A/4 + A^2/4 - A^3/24)^{-1}(I + A/4)$ of $\exp(A)$ is used in Lawson's method.

EXAMPLE 2. Consider a complex equation

$$\frac{dz}{dt} = (i + s - |z|^2)z, \quad z(0) = z_0, \tag{5.2}$$

where s is a real parameter. Letting $z = x + iy$, the real system corresponding to (5.2) is given by

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} s & -1 \\ 1 & s \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} x(x^2+y^2) \\ y(x^2+y^2) \end{pmatrix}, \quad \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}. \tag{5.3}$$

Due to the rotational symmetry of (5.2), it is possible to derive a scalar real equation for the evolution of the variable $q = |z|^2 = x^2 + y^2$, namely,

$$\frac{dq}{dt} = 2(s - q)q, \quad q(0) = q_0 = x_0^2 + y_0^2. \tag{5.4}$$

Figure 1 dipicts the true dynamics of (5.4) described in Sanz-Serna [5].

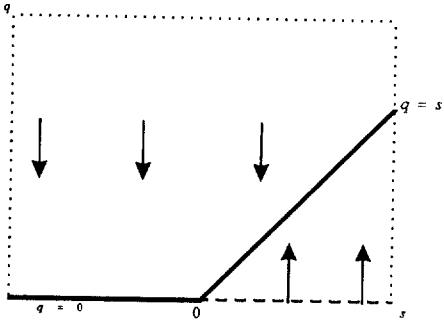

FIGURE 1

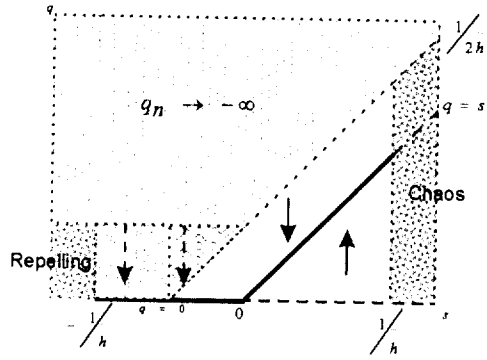

FIGURE 2

In order to approximate the true dynamics of (5.4) by a discrete dynamical system, we must have that

$$(5.5)\quad \begin{cases} \text{if} \quad s < 0, & \text{then} \quad q_n > q_{n+1} \geq 0, \\ \text{if} \quad s > 0, & \text{then} \quad q_n > q_{n+1} \geq s \quad \text{or} \quad 0 < q_n < q_{n+1} \leq s. \end{cases}$$

Application of Euler method to (5.4) yields a discrete dynamical system

$$(5.6)\qquad q_{n+1} = \alpha(q_n) = \rho(s, h, q_n)q_n := \{1 + 2h(s - q_n)\}q_n.$$

Its dynamics is described as follows:

For $s < 0$, the function α has a fixed point $q^* = 0$.
- If $-\frac{1}{h} < s < 0$, then $|\alpha'(q^*)| < 1$ and q^* is an attracting fixed point.
- If $s < -\frac{1}{h}$, then $|\alpha'(q^*)| > 1$ and q^* is a repelling fixed point.

For $s > 0$, the function α has two fixed points $q^* = 0$ and $q^+ = s$.
Since $|\alpha'(q^*)| > 1$, q^* is a repelling fixed point.
For the fixed point q^+, there are two subcases to be considered.
- For $0 < s < \frac{1}{h}$, $|\alpha'(q^+)| < 1$. Hence q^+ is an attracting fixed point.
- For $s > \frac{1}{h}$, $|\alpha'(q^+)| > 1$. Hence q^+ is unstable. The dynamics in the region $\{(s, q) | s > \frac{1}{h}, 0 < q < \infty\}$ can be very complicated including chaos.

Figure 2 shows that the dynamics of Euler approximation with $\bar{h}$ fixed and s varying. The behavior of the orbit $\{q_n\}$ is determined by the

value of $\rho(\bar{s}, \bar{h}, q_n)$. Here we note that

(1) If $(\bar{s}, q_0) \in \{(s,q) \mid s > 0, q > s + \frac{1}{2\bar{h}}\} \cup \{(s,q) \mid s < 0, q > \frac{1}{2\bar{h}}\}$, then $\rho(\bar{s}, \bar{h}, q_0) < 0$, $q_1 = \rho(\bar{s}, \bar{h}, q_0)q_0 < 0$, $\rho(\bar{s}, \bar{h}, q_1) = 1 + 2\bar{h}\bar{s} - 2\bar{h}q_1 > 1$, and $q_2 = \rho(\bar{s}, \bar{h}, q_1)q_1 < q_1 < 0$. Hence $q_n \to -\infty$ as $n \to \infty$.

(2) If $(\bar{s}, q_0) \in \{(s,q) \mid -\frac{1}{\bar{h}} < s < -\frac{1}{2\bar{h}}, 0 < q < \frac{1}{2\bar{h}}\}$ then $-1 < \rho(\bar{s}, \bar{h}, q_0) < 0$, $-q_0 < q_1 = \rho(\bar{s}, \bar{h}, q_0)q_0 < 0$, $-1 < \rho(\bar{s}, \bar{h}, q_1) < 0$, and $-q_1 > q_2 = \rho(\bar{s}, \bar{h}, q_1)q_1 > 0$. Hence $\{q_n\}$ is swinging from one side of 0 to the other and the magnitude of the swing is shrinking. So $\{q_n\}$ oscilates and converges to q^*.

(3) If $\{(s,q) \mid -\frac{1}{2\bar{h}} < s < 0, s + \frac{1}{2\bar{h}} < q < \frac{1}{2\bar{h}}\}$, then $-1 < \rho(\bar{s}, \bar{h}, q_0) < 0$, $-q_0 < q_1 = \rho(\bar{s}, \bar{h}, q_0)q_0 < 0$, $1 - 2\bar{h}q_0 > 0$, $-\frac{1}{2\bar{h}} < \bar{s} - q_1 < 0$, $0 < \rho(\bar{s}, \bar{h}, q_1) < 1$, $q_1 < q_2 = \rho(\bar{s}, \bar{h}, q_1)q_1 < 0$, and $0 < \rho(\bar{s}, \bar{h}, q_2) < 1$. So $\{q_n\}$ converges to q^* from below.

Application of Euler method to (5.3) is transformed into the following Sanz-Serna's recursion

$$q_{n+1} = \mu(q_n) = \varphi(s, h, q_n)q_n := [\{1 + h(s - q_n)\}^2 + h^2]q_n.$$

Its dynamics with $\bar{h}$ fixed and s varying is described in Sanz-Serna [5]. In this case, the fixed points of the function μ are given by

$$q^* = 0, \quad q^s = s - s_l, \quad \text{and} \quad q^a = s - s_r,$$

where $s_l = -\frac{1+\sqrt{1-h^2}}{h}$ and $s_r = -\frac{1-\sqrt{1-h^2}}{h}$. Here s_l, s_r, and $s_u = \frac{2-h^2}{h\sqrt{1-h^2}} - \frac{1}{h}$ are evaluated by $h = \bar{h}$.

Figure 3 depicts the dynamics of Sanz-Serna's approximation with $\bar{h}$ fixed and s varying. The behavior of the orbit $\{q_n\}$ is determined by the value of $\varphi(\bar{s}, \bar{h}, q)$. Here we note that

(1) $\varphi(s, h, q) > 0$ for all s, h, and q.
(2) The range of the possible step size is $0 < h < h_0 \approx 0.5$.
(3) q^s has no counterpart in (5.4), i.e., q^s is a spurious equilibrium.
(4) q^a is an $\mathcal{O}(h)$ approximation to the equilibrium q^+ of (5.4).
(5) If $\bar{h} \to 0$, then $s_l \to -\infty$, $s_r \to 0$ and $s_u \to \infty$.

(6) If $(\bar{s}, q_0) \in \{(s,q) \mid s < s_l,\ q > 0\} \cup \{(s,q) \mid s > s_l,\ q > s - s_l\}$, then $\varphi(\bar{s}, \bar{h}, q_0) > 1$. Hence $q_n < q_{n+1}$ and $q_n \to +\infty$ as $n \to \infty$.

(7) In the region $\{(s,q) \mid s_r < s < s_u, s - s_r < q < s - s_l\}$, $\varphi(\bar{s}, \bar{h}, q_0)$ is increasing to 1 as $q_0 \to s - s_l$ or $q_0 \to s - s_r$, and $\varphi(\bar{s}, \bar{h}, q_0)$ is decreasing to $\bar{h}^2 < 0.25$ as $q_0 \to \bar{s} + \frac{1}{\bar{h}}$. Hence it is difficult to estimate the behavior of the orbit $\{q_n\}$.

For the Euler-GS method, the s-scalar matrices for the problems (5.3) and (5.4) are chosen by (3.5) as

$$W \equiv \begin{pmatrix} s - x_n^2 - y_n^2 & -1 \\ 1 & s - x_n^2 - y_n^2 \end{pmatrix} \quad \text{and} \quad W' \equiv 2(s - q_n),$$

respectively. Applications of Euler-GS process (3.10) to (5.3) and (5.4) yield the discrete dynamical systems

$$(5.7) \quad \begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \exp(h(s - x_n^2 - y_n^2)) \begin{pmatrix} \cos(h) & -\sin(h) \\ \sin(h) & \cos(h) \end{pmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix},$$

and

$$(5.8) \quad q_{n+1} = \gamma(q_n) = \psi(s, h, q_n) q_n := \exp(2h(s - q_n)) q_n.$$

The dynamics of (5.8) is described as follows

$$\begin{cases} \text{For } s < 0, & \text{the dynamics of } \gamma \text{ is the same as the true dynamics.} \\ \text{For } s > 0, & \text{the dynamics of } \gamma \text{ is the same as the dynamics of } \alpha. \end{cases}$$

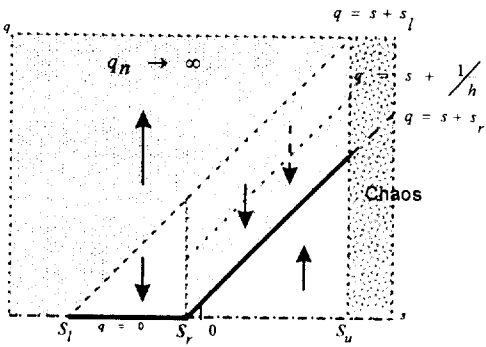

FIGURE 3

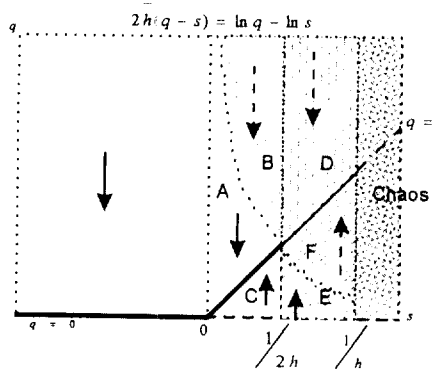

FIGURE 4

In Figure 4, the behavior of the orbit $\{q_n\}$ is determined by the value of $\psi(\bar{s}, \bar{h}, q)$.

For the regions

$$A = \{(s,q) \,|\, 0 < s < \tfrac{1}{2\bar{h}},\ q > s, 2\bar{h}(s-q) > \ln(s) - \ln(q)\},$$
$$B = \{(s,q) \,|\, 0 < s < \tfrac{1}{2\bar{h}},\ q > s, 2\bar{h}(s-q) < \ln(s) - \ln(q)\},$$
$$C = \{(s,q) \,|\, 0 < s < \tfrac{1}{2\bar{h}},\ 0 < q < s\},$$
$$D = \{(s,q) \,|\, \tfrac{1}{2\bar{h}} < s < \tfrac{1}{\bar{h}},\ q > s\},$$
$$E = \{(s,q) \,|\, \tfrac{1}{2\bar{h}} < s < \tfrac{1}{\bar{h}},\ q < s,\ 2\bar{h}(s-q) < \ln(s) - \ln(q)\}, \text{ and}$$
$$F = \{(s,q) \,|\, \tfrac{1}{2\bar{h}} < s < \tfrac{1}{\bar{h}},\ 0 < q < s,\ 2\bar{h}(s-q) > \ln(s) - \ln(q)\},$$

we note that

(1) $\psi(s,h,q) > 0$ for all $s, h,$ and q.
(2) if $(\bar{s}, q_0) \in A$, then $q_n > q_{n+1} \geq \bar{s}$ and $\{q_n\}$ converges to $\bar{s}$ from above.
(3) if $(\bar{s}, q_0) \in B$, then $(\bar{s}, q_1) \in C$,
(4) if $(\bar{s}, q_0) \in C$, then $2\bar{h}(\bar{s} - q_n) < \ln(\bar{s}) - \ln(q_n)$, $q_n < q_{n+1} \leq \bar{s}$, and $\{q_n\}$ converges to $\bar{s}$ from below.
(5) if $(\bar{s}, q_0) \in D$, then $2\bar{h}(\bar{s} - q_n) < \ln(\bar{s}) - \ln(q_n)$, $(\bar{s}, q_1) \in E$ or $(\bar{s}, q_1) \in F$,
(6) if $(\bar{s}, q_0) \in E$, then $q_0 < q_1 < \bar{s}$ but $q_n \in F$ for some n.
(7) if $(\bar{s}, q_0) \in F$, then $(\bar{s}, q_1) \in D$.

The dynamics in the white regions of each figures coincides with the true dynamics.

From Figure 4 we conclude for any given $\bar{s}$ and q_0 that

(1) if $\bar{s} < 0$, then for all h, or
(2) if $\bar{s} > 0$, then for small h such that $h < \min\left\{\frac{1}{2\bar{s}}, \frac{\ln \bar{s} - \ln q_0}{2(\bar{s} - q_0)}\right\}$,

the discrete dynamical system (5.8) faithfully approximates the true dynamics of (5.4).

The problem (5.3) is equivalent to (5.4), the discrete dynamical system (5.7) is equivalent to (5.8), and the problem (5.4) is faithfully approximated by the discrete dynamical system (5.8). Hence the periodicity of the problem (5.3) can be faithfully represented by the discrete dynamical system (5.7).

References

[1] B. L. Ehle and J. D. Lawson, *Generalized Runge-Kutta processes for stiff initial-value problems*, J. Inst. Maths. Applics. **16** (1975), 11-21.

[2] M. W. Hirsch and S. Smale, *Differential equations, dynamical systems, and linear algebra*, Academic press, New York, San Francisco, London, 1974.

[3] J. D. Lawson, *Generalized Runge-Kutta processes for stable systems with large Lipschitz constants*, SIAM J. Numer. Anal. **4** (1967), 372-380.

[4] L. Perko, *Differential equations and dynamical systems*, Springer-Verlag, New York, Inc., 1991.

[5] J. M. Sanz-Serna, *Numerical ordinary differential equations vs. dynamical systems. In "The dynamics of numerics and the numerics of dynamics" (D.S. Broomhead and A. Iserles, eds.)*, Clarendon Press, Oxford. (1992), 81-106.

Department of Mathematics, Chung-Ang University, Seoul 156-756, Korea
E-mail: dwyu@cau.ac.kr