 CMES, vol.12, no.2, pp.83-107, 2006

Preserving Constraints of Differential Equations by Numerical Methods Based on Integrating Factors

Chein-Shan Liu[1]

Abstract: The system we consider consists of two parts: a purely algebraic system describing the manifold of constraints and a differential part describing the dynamics on this manifold. For the constrained dynamical problem in its engineering application, it is utmost important to developing numerical methods that can preserve the constraints. We embed the nonlinear dynamical system with dimensions n and with k constraints into a mathematically equivalent $n+k$-dimensional nonlinear system, which including k integrating factors. Each subsystem of the k independent sets constitutes a Lie type system of $\dot{\mathbf{X}}_i = \mathbf{A}_i\mathbf{X}_i$ with $\mathbf{A}_i \in so(n_i,1)$ and $n_1+\cdots+n_k = n$. Then, we can apply the exponential mapping technique to integrate the augmented systems and use the k freedoms to adjust the k integrating factors such that the k constraints are satisfied. A similar procedure is also applied to the case when one integrates the k augmented systems by the fourth-order Runge-Kutta method. Since all constraints are included in the newly developed integrating schemes, it is guaranteed that all algebraic equations that describe the manifold are satisfied up to an accuracy that is used to integrate these dynamical equations and hence a drift from the solution manifold can be avoided. Several numerical examples, including differential algebraic equations (DAEs), are investigated to confirm that the new numerical methods are effective to integrate the constrained dynamical systems by preserving the constraints.

keyword: Nonlinear dynamical system, Preserving constraints, Integrating factors, Cones, Minkowski space, Group preserving scheme

1 Introduction

Many practical engineering problems of interest can be modeled by systems of differential equations whose solutions satisfy some invariants, usually defined explicitly by algebraic equations. In the past several decades particular attention has been paid to developing numerical methods which approximate the solution of such a system while preserving invariants to a machinery precision; see, e.g., Baumgarte (1972), Führer and Leimkuhler (1991), Ascher and Petzold (1991), März (1991, 2002), Ascher, Chin and Reich (1994), Campbell and Moore (1995), Ascher (1997), Chan, Chartier and Murua (2002), Arevalo, Campbell and Selva (2004), and references therein.

Systems of coupled differential equations and algebraic equations often occur as a set of differential equations that are subject to constraints. The constraints may be linear or nonlinear and may be imposed on partial varables or on all variables. The constrained differential equations systems arise frequently as initial value problems in the computer-aided design and modeling of mechanical systems subject to constraints, for example, multi-body system, circuit simulation, chemical process modeling, material plasticity, friction system, and in many other applications. The numerical integration of constrained dynamical systems may be more complicated than that of ordinary differential equations without constraint.

In this paper we will provide an effective numerical method to integrate the following nonlinear differential equations system:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x},t),\ \ \mathbf{x}(0) = \mathbf{x}_0,\ \ \mathbf{x} \in \mathbb{R}^n,\ \ t \in \mathbb{R}^+, \tag{1}$$

which may subject to constraints. A superimposed dot signifies a time differentiation with respect to t. For a prescribing $\mathbf{x}_0$ at time $t = 0$, it constitutes an initial-value-problem. We first assume that the solution $\mathbf{x}(t)$ of Eq. (1) is subjected to a constraint:

$$\rho(\mathbf{x}(t),\mathbf{x}_0) := H(\mathbf{x}(t)) - H(\mathbf{x}_0) = 0, \tag{2}$$

where $H(\mathbf{x})$ is an invariant of Eq. (1). After developing the constraint preserved numerical integrating methods

[1] Department of Mechanical and Mechatronic Engineering, Taiwan Ocean University, Keelung, Taiwan

for Eqs. (1) and (2), we extend our methods to Eq. (1), which is subjected to several constraints.

Without considering the constraint (2), Liu (2001) has explored a Minkowski frame of the dynamical system (1) by introducing

$$\mathbf{n} := \frac{\mathbf{x}}{\|\mathbf{x}\|} \tag{3}$$

for $\mathbf{x} \neq \mathbf{0}$, where $\|\mathbf{x}\| := \sqrt{\mathbf{x}\cdot\mathbf{x}}$ is the Euclidean norm of $\mathbf{x}$, and a dot between two vectors, say $\mathbf{x}\cdot\mathbf{y}$, denotes their inner product.

From Eqs. (1) and (3) it is verified mathematically equivalent to the following dynamical system:

$$\dot{\mathbf{X}} = \mathbf{A}\mathbf{X}. \tag{4}$$

In the augmented homogeneous coordinates

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^s \\ X^0 \end{bmatrix} := \begin{bmatrix} X^0\mathbf{n} \\ X^0 \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \|\mathbf{x}\| \end{bmatrix}, \tag{5}$$

$X^0 = \|\mathbf{x}\|$ is an integrating factor, and

$$\mathbf{A} := \begin{bmatrix} \mathbf{0}_n & \mathbf{A}_0^s \\ (\mathbf{A}_0^s)^{\mathrm{T}} & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{0}_n & \frac{\mathbf{f}(\mathbf{x},t)}{\|\mathbf{x}\|} \\ \frac{\mathbf{f}^{\mathrm{T}}(\mathbf{x},t)}{\|\mathbf{x}\|} & 0 \end{bmatrix} \tag{6}$$

is a Lie algebra of the proper orthochronous Lorentz group $SO_o(n,1)$ satisfying

$$\mathbf{A}^{\mathrm{T}}\mathbf{g} + \mathbf{g}\mathbf{A} = \mathbf{0}, \tag{7}$$

where

$$\mathbf{g} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{1\times n} & -1 \end{bmatrix} \tag{8}$$

is a Minkowski metric. In the above, $\mathbf{I}_n$ is an identity matrix of order n, and the superscript T denotes the transpose.

It is obvious that the first equation is the same as the original equation (1), but the introduction of the second equation in the new formulation leads to a Minkowskian structure for the *augmented nonlinear system* (4). In the later it would be clear that this new formulation with an extra integrating factor X^0 has the advantage to fit the constraint equation by adjusting the length $X^0 = \|\mathbf{x}\|$ of state variables.

2 Group preserving scheme by exponential mapping

The numerical scheme would provide a medium to calculate the value of $\mathbf{X}$ at time $t = t_{\ell+1}$ when $\mathbf{X}$ is already known at time $t = t_\ell$. The evolution of $\mathbf{X}$ is governed by the dynamical law (4) with matrix $\mathbf{A}$ given by Eq. (6). However, due to the presence of $X^0 = \|\mathbf{x}\|$ and $X^0\mathbf{n} = \mathbf{x}$, $\mathbf{A}$ is not a constant matrix, and we may approximate the solution of the dynamical law (4) by considering $X^0 = \|\mathbf{x}\|$ and $X^0\mathbf{n} = \mathbf{x}$ constant in each single time step with stepsize $\Delta t = t_{\ell+1} - t_\ell$. Under such additional hypotheses, the matrix $\mathbf{A}$ is constant, and so the evolution of Eq. (4) is known to be

$$\mathbf{X}(\ell+1) = \mathbf{G}(\ell)\mathbf{X}(\ell), \tag{9}$$

where

$$\begin{aligned}\mathbf{G}(\ell) &:= \exp[\Delta t\mathbf{A}(\ell)] \\ &= \begin{bmatrix} \mathbf{I}_n + \frac{a(\ell)-1}{\|\mathbf{A}_0^s(\ell)\|^2}\mathbf{A}_0^s(\ell)(\mathbf{A}_0^s)^{\mathrm{T}}(\ell) & \frac{b(\ell)}{\|\mathbf{A}_0^s(\ell)\|}\mathbf{A}_0^s(\ell) \\ \frac{b(\ell)}{\|\mathbf{A}_0^s(\ell)\|}(\mathbf{A}_0^s)^{\mathrm{T}}(\ell) & a(\ell) \end{bmatrix},\end{aligned} \tag{10}$$

in which

$$a(\ell) := \cosh(\Delta t\|\mathbf{A}_0^s(\ell)\|), \tag{11}$$

$$b(\ell) := \sinh(\Delta t\|\mathbf{A}_0^s(\ell)\|). \tag{12}$$

A numerical algorithm is called a *group preserving scheme* (GPS) if for every time increment the mapping $\mathbf{G}(\ell)$ from $\mathbf{X}(\ell)$ to $\mathbf{X}(\ell+1)$ preserves the following group properties (Liu, 2001):

$$\mathbf{G}^{\mathrm{T}}\mathbf{g}\mathbf{G} = \mathbf{g}, \tag{13}$$

$$\det\mathbf{G} = 1, \tag{14}$$

$$G_0^0 \geq 1, \tag{15}$$

where det is the shorthand of determinant, and G_0^0 is the 00-th mixed component of $\mathbf{G}$.

3 Modified group preserving scheme by considering constraint

From Eqs. (9), (10) and (5) it follows a numerical scheme for $\mathbf{n}$:

$$\mathbf{n}(\ell+1)=\frac{\begin{array}{c}\|\mathbf{A}_0^s(\ell)\|^2\mathbf{n}(\ell)+[(a(\ell)-1)(\mathbf{A}_0^s)^{\mathrm{T}}(\ell)\mathbf{n}(\ell)\\ +b(\ell)\|\mathbf{A}_0^s(\ell)\|]\mathbf{A}_0^s(\ell)\end{array}}{b(\ell)\|\mathbf{A}_0^s(\ell)\|(\mathbf{A}_0^s)^{\mathrm{T}}(\ell)\mathbf{n}(\ell)+a(\ell)\|\mathbf{A}_0^s(\ell)\|^2}. \tag{16}$$

It is easy to check that

$$\|\mathbf{n}(\ell)\|=1\Longrightarrow\|\mathbf{n}(\ell+1)\|=1, \tag{17}$$

which means that scheme (16) preserves the unit length of $\mathbf{n}$. Corresponding to the symmetry $\mathbf{G}\in SO_o(n,1)$, the symmetry preserved by scheme (16) is denoted by $PSO_o(n,1)$, a projection of $SO_o(n,1)$.

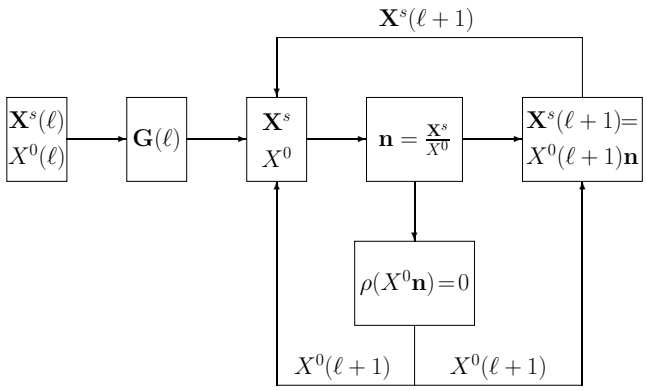

Figure 1 : Instead of the open-loop scheme to calculate $(\mathbf{X}^s(\ell+1),X^0(\ell+1))$ from $(\mathbf{X}^s(\ell),X^0(\ell))$ by left-multiplying $\mathbf{G}(\ell)$, in the modified group preserving scheme we use a closed-loop method to enhance the preservation of constraint.

In order to match the constraint exactly, we impose the condition (2) on the numerical solutions of $\mathbf{X}(\ell+1)$, which by Eq. (5) leads to

$$\rho(X^0(\ell+1)\mathbf{n}(\ell+1),\mathbf{x}_0)=0, \tag{18}$$

where $\mathbf{n}(\ell+1)$ is already calculated by scheme (16). Substituting $\mathbf{n}(\ell+1)$ into the above equation and solving it by the Newton-Raphson method we may obtain a new $\|\mathbf{x}(\ell+1)\|=X^0(\ell+1)$. With this new $\|\mathbf{x}(\ell+1)\|=X^0(\ell+1)$ we update $\mathbf{x}(\ell+1)=\mathbf{X}^s(\ell+1)$ to a new $\mathbf{x}(\ell+1)=\mathbf{X}^s(\ell+1)=X^0(\ell+1)\mathbf{n}(\ell+1)=\|\mathbf{x}(\ell+1)\|\mathbf{n}(\ell+1)$, which satisfies the constraint (2) within a specified error tolerance. Here we will call such a scheme the *modified group preserving scheme* (MGPS), which preserves the symmetry $PSO_o(n,1)$ as well as retains the constraint (2). Figure 1 shows the numerical processes by a closed-loop diagram.

In order to increase the accuracy we can apply the fourth-order Runge-Kutta method (RK4) to the augmented nonlinear system (4), instead of to the nonlinear system (1), such that we can calculate $(\mathbf{x},\|\mathbf{x}\|)$ for each assigned time step. Then we can calculate the orientation vector $\mathbf{n}$ in Eq. (3), and substitute the result of $\|\mathbf{x}\|\mathbf{n}$ into the constraint (2) to solve $\|\mathbf{x}\|$. Upon returning to $\|\mathbf{x}\|\mathbf{n}=\mathbf{x}$ we obtain the numerical solution of $\mathbf{x}$, which satisfies the constraint (2) within a specified error tolerance. In the later we will call this scheme a *modified RK4* (MRK4).

The above technique to satisfy the constraint is originated from the idea of introducing an extra variable of integrating factor $X^0=\|\mathbf{x}\|$ in the new augmented system. Such that we obtain an extra degree of freedom to adapt the factor of $X^0=\|\mathbf{x}\|$ by enhancing the constraint.

4 Numerical examples of one constraint

4.1 Example 1

Now, let us apply the above MGPS method to a certain example, which is a two-dimensional predator-prey equation of Lotka-Volterra type:

$$\dot{x}=-x+xy, \tag{19}$$

$$\dot{y}=y-xy, \tag{20}$$

where x is the population of predator and y is the population of prey. A prime feature of the above system is that its fixed point $(\overline{x},\overline{y})=(1,1)$ is neutral stable, and the conserved constraint of the above system is

$$\begin{aligned}\rho(x,y,x_0,y_0)&=H(x,y)-H(x_0,y_0)\\&=\ln x-x+\ln y-y-(\ln x_0-x_0+\ln y_0-y_0)\\&=0,\end{aligned} \tag{21}$$

where $x_0>0$ and $y_0>0$ are the initial values prescribed at time $t=0$, and after that $x>0$ and $y>0$ for all $t>0$ are direct results of Eqs. (19) and (20).

It is very difficult to construct finite difference schemes that give the proper periodic solutions behavior about the fixed point; see, for example, Sanz-Serna (1994). The conventional schemes almost give numerical solution points that either spiral in towards the fixed point or spiral out of the fixed point.

For example, substituting Eq. (10) for $\mathbf{G}(\ell)$ into Eq. (9) and taking its first row by considering Eq. (5), we obtain

$$\mathbf{x}(\ell+1) = \mathbf{x}(\ell) + \eta(\ell)\mathbf{f}(\ell), \tag{22}$$

where the adaptive factor

$$\eta(\ell) := \frac{[a(\ell)-1]\mathbf{f}(\ell)\cdot\mathbf{x}(\ell) + b(\ell)\|\mathbf{x}(\ell)\|\|\mathbf{f}(\ell)\|}{\|\mathbf{f}(\ell)\|^2} \tag{23}$$

is varying step-by-step. Applying scheme (22) to Eqs. (19) and (20), we obtain a Jacobian matrix of the numerical mapping as follows:

$$\mathcal{J} := \frac{\partial \mathbf{x}(\ell+1)}{\partial \mathbf{x}(\ell)} = \mathbf{I}_2 + \mathbf{f}(\ell)\left(\frac{\partial \eta(\ell)}{\partial \mathbf{x}(\ell)}\right)^{\mathrm{T}} + \eta(\ell)\frac{\partial \mathbf{f}(\ell)}{\partial \mathbf{x}(\ell)}. \tag{24}$$

At the fixed point $\mathbf{f}(\ell) = \mathbf{0}$ and $\eta(\ell) = \Delta t$, and thus we have

$$\mathcal{J} = \begin{bmatrix} 1 & \Delta t \\ -\Delta t & 1 \end{bmatrix}. \tag{25}$$

The two eigenvalues of $\mathcal{J}$ are $\lambda = 1 \pm i\Delta t$, both of which have the magnitude of $|\lambda| = \sqrt{1+(\Delta t)^2} \neq 1$. The property of the fixed point is altered by the mapping (22). More precisely, the mapping (22) does not have the same neutral-type stability as the original system of Eqs. (19) and (20) has. Owing to this defect, the long term behavior of the original system is destroyed by this numerical scheme. In Figs. 2 and 3 we display the numerical results by applying scheme (22) to Eqs. (19) and (20) with time stepsizes of $\Delta t = 0.001$ sec and $\Delta t = 0.01$ sec, respectively, but with the same initial values of $x_0 = y_0 = 0.5$. As expected, the orbits of (x, y) as shown in Figs. 2(c) and 3(c) spiral out of the fixed point gradually. Figures 4(a) and 4(b) show the errors of invariant defined by $|\rho(\mathbf{x}, \mathbf{x}_0)|$ when utilizing $\Delta t = 0.001$ sec and $\Delta t = 0.01$ sec, respectively. When the above quantity is zero for numerical solutions we obtain an invariant-preserving scheme which preserves the constraint exactly. Obviously, scheme (22) gives the errors of invariant in the order of 10^{-2} for $\Delta t = 0.001$ sec and in the order of 10^{-1} for $\Delta t = 0.01$ sec. For the latter case the orbit of numerical solutions spirals out from the closed curve very quickly as shown in Fig. 3(c).

In order to reduce the error of invariant we apply the MGPS in Section 3 to Eqs. (19) and (20), where the error tolerance for applying the Newton-Raphson method to solve Eq. (18) is 10^{-4} for $\Delta t = 0.001$ sec and 10^{-3} for $\Delta t = 0.01$ sec. Under the same conditions as in the above, the numerical results obtained by this scheme were shown in Figs. 2-4 with dashed lines. It can be seen that the errors of invariant are greatly reduced to the orders of $10^{-11} - 10^{-5}$ for $\Delta t = 0.001$ sec and of $10^{-7} - 10^{-3}$ for $\Delta t = 0.01$ sec. The MGPS can produce solution points that stay on the closed curves as shown in Figs. 2(d) and 3(d).

Kahan has considered the following unconventional scheme for Eqs. (19) and (20):

$$\begin{aligned} &\frac{x(\ell+1)-x(\ell)}{\Delta t} \\ &= \frac{-1}{2}[x(\ell+1)+x(\ell)] \\ &+ \frac{1}{2}[x(\ell+1)y(\ell)+x(\ell)y(\ell+1)], \end{aligned} \tag{26}$$

$$\begin{aligned} &\frac{y(\ell+1)-y(\ell)}{\Delta t} \\ &= \frac{1}{2}[y(\ell+1)+y(\ell)] \\ &- \frac{1}{2}[x(\ell+1)y(\ell)+x(\ell)y(\ell+1)]. \end{aligned} \tag{27}$$

Some properties about this scheme have been demonstrated by Sanz-Serna (1994), who showed that Kahan's method is symplectic with respect to a noncanonical symplectic structure.

Due to the linearity, we can solve Eqs. (26) and (27) to obtain

$$x(\ell+1) = \frac{\begin{array}{c}(1-\tau)^2 x(\ell) + (1-\tau)\tau x^2(\ell) \\ +(1+\tau)\tau x(\ell)y(\ell)\end{array}}{1-\tau^2+(1+\tau)\tau x(\ell) - (1-\tau)\tau y(\ell)}, \tag{28}$$

$$y(\ell+1) = \frac{\begin{array}{c}(1+\tau)^2 y(\ell) - (1+\tau)\tau y^2(\ell) \\ -(1-\tau)\tau x(\ell)y(\ell)\end{array}}{1-\tau^2+(1+\tau)\tau x(\ell) - (1-\tau)\tau y(\ell)}, \tag{29}$$

where $\tau := \Delta t/2$. After a lengthy calculation, the Jacobian matrix of the above mapping at the fixed point is

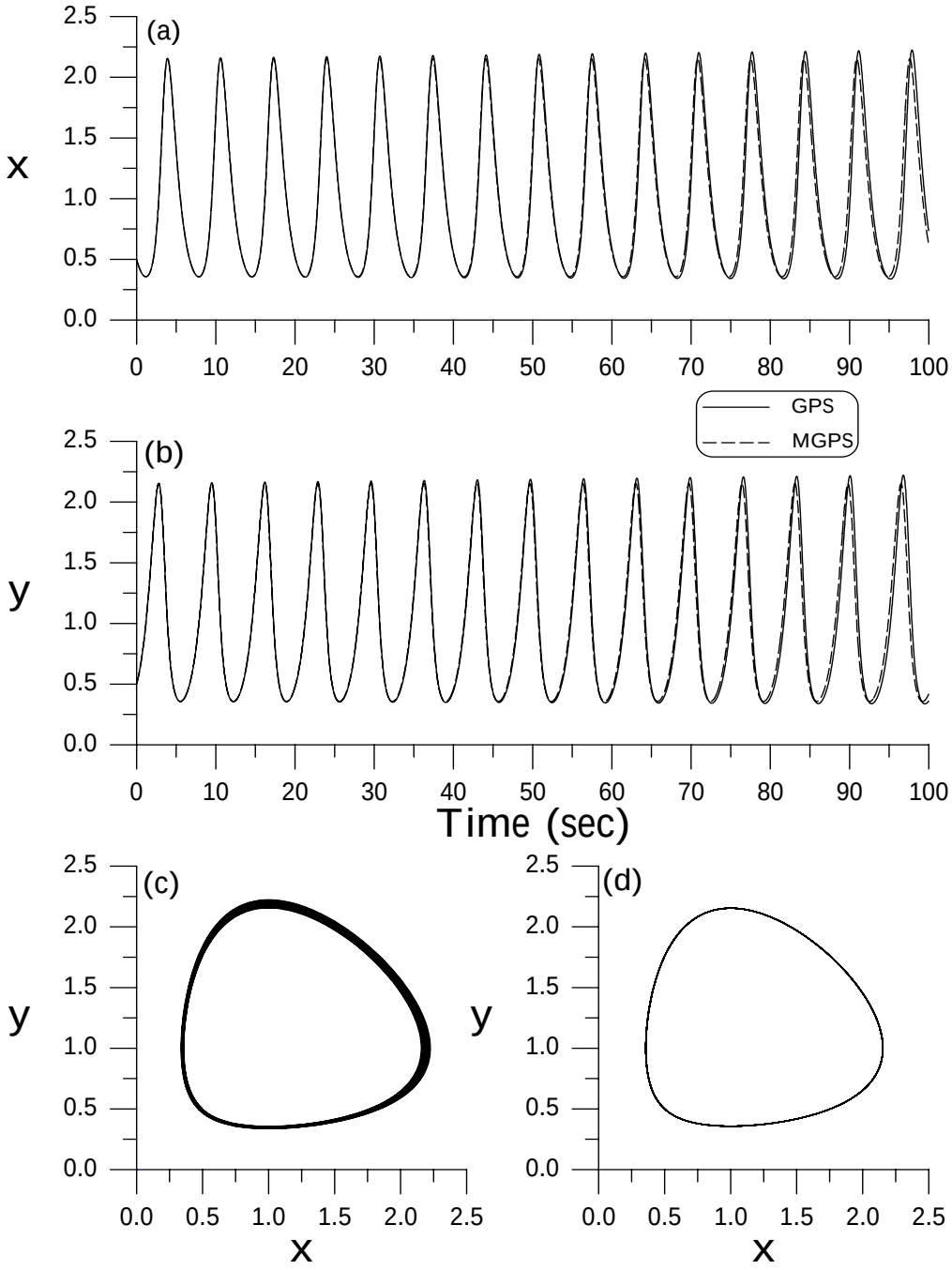

Figure 2 : Compare the numerical results of Example 1 calculated by GPS and MGPS with time stepsize $\Delta t = 0.001$: (a) the time history of x, (b) the time history of y, (c) the orbit of (x,y) calculated by GPS, and (d) the orbit of (x,y) calculated by MGPS.

found to be

$$\mathcal{J} = \begin{bmatrix} \frac{1-\tau^2}{1+\tau^2} & \frac{2\tau}{1+\tau^2} \\ \frac{-2\tau}{1+\tau^2} & \frac{1-\tau^2}{1+\tau^2} \end{bmatrix}. \tag{30}$$

The two eigenvalues of $\mathcal{J}$ are $\lambda = (1-\tau^2 \pm 2\tau i)/(1+\tau^2)$, both of which have the magnitude $|\lambda| = 1$. Thus, the property of the fixed point is not altered by the mappings (28) and (29). More precisely, the mappings (28) and (29) have the same neutral-type stability as the system

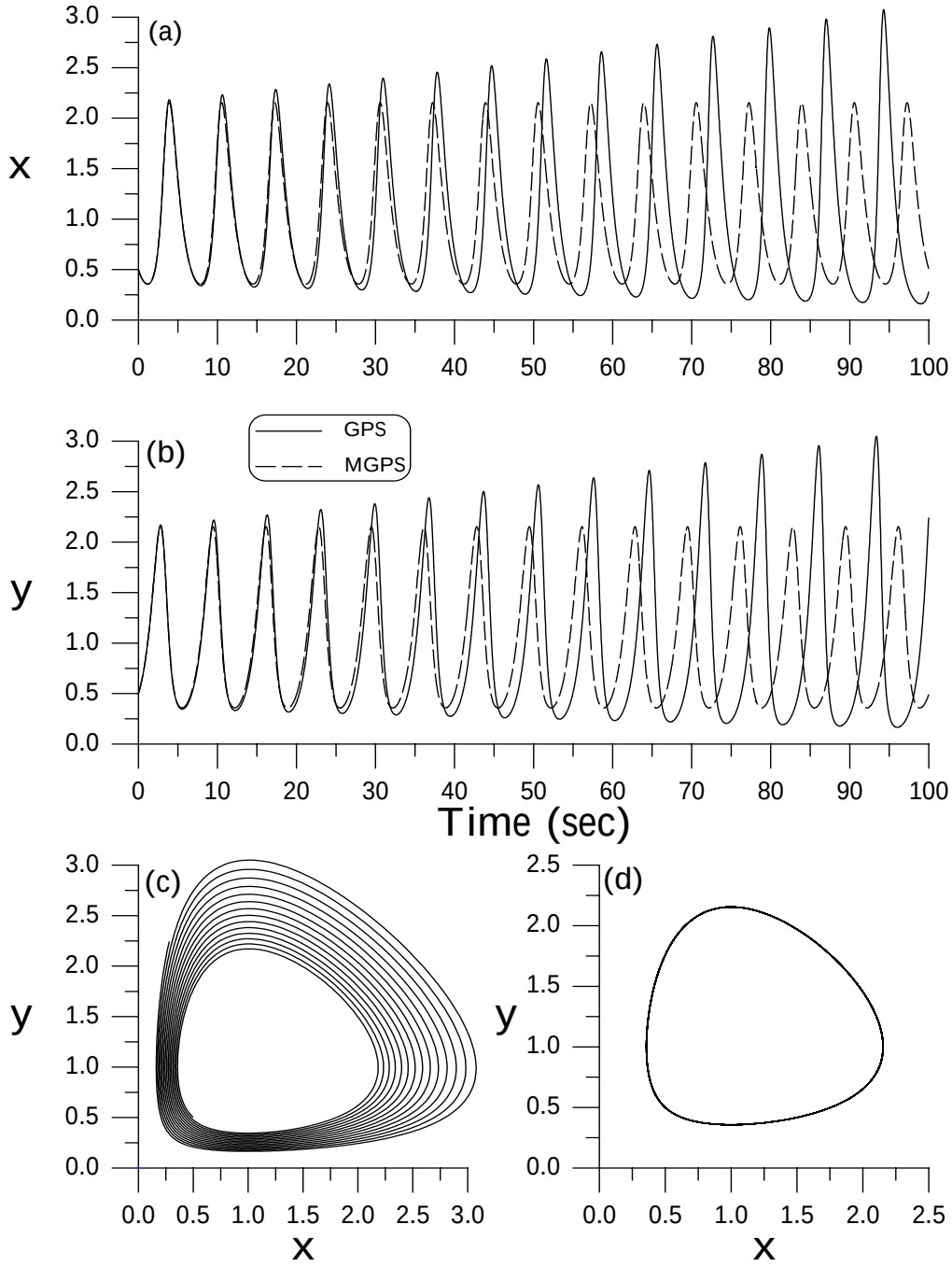

Figure 3 : Compare the numerical results of Example 1 calculated by GPS and MGPS with time stepsize $\Delta t = 0.01$: (a) the time history of x, (b) the time history of y, (c) the orbit of (x, y) calculated by GPS, and (d) the orbit of (x, y) calculated by MGPS.

of Eqs. (19) and (20) has. Numerical results computed by the above scheme coincide with that calculated by the MGPS as shown with dashed lines in Figs. 2(a) and 2(b). The error of invariant as shown in Fig. 4(a) is smaller than that of the MGPS in the orders of $10^{-13} - 10^{-7}$. In Section 5 we will provide a more accurate scheme than Kahan's method.

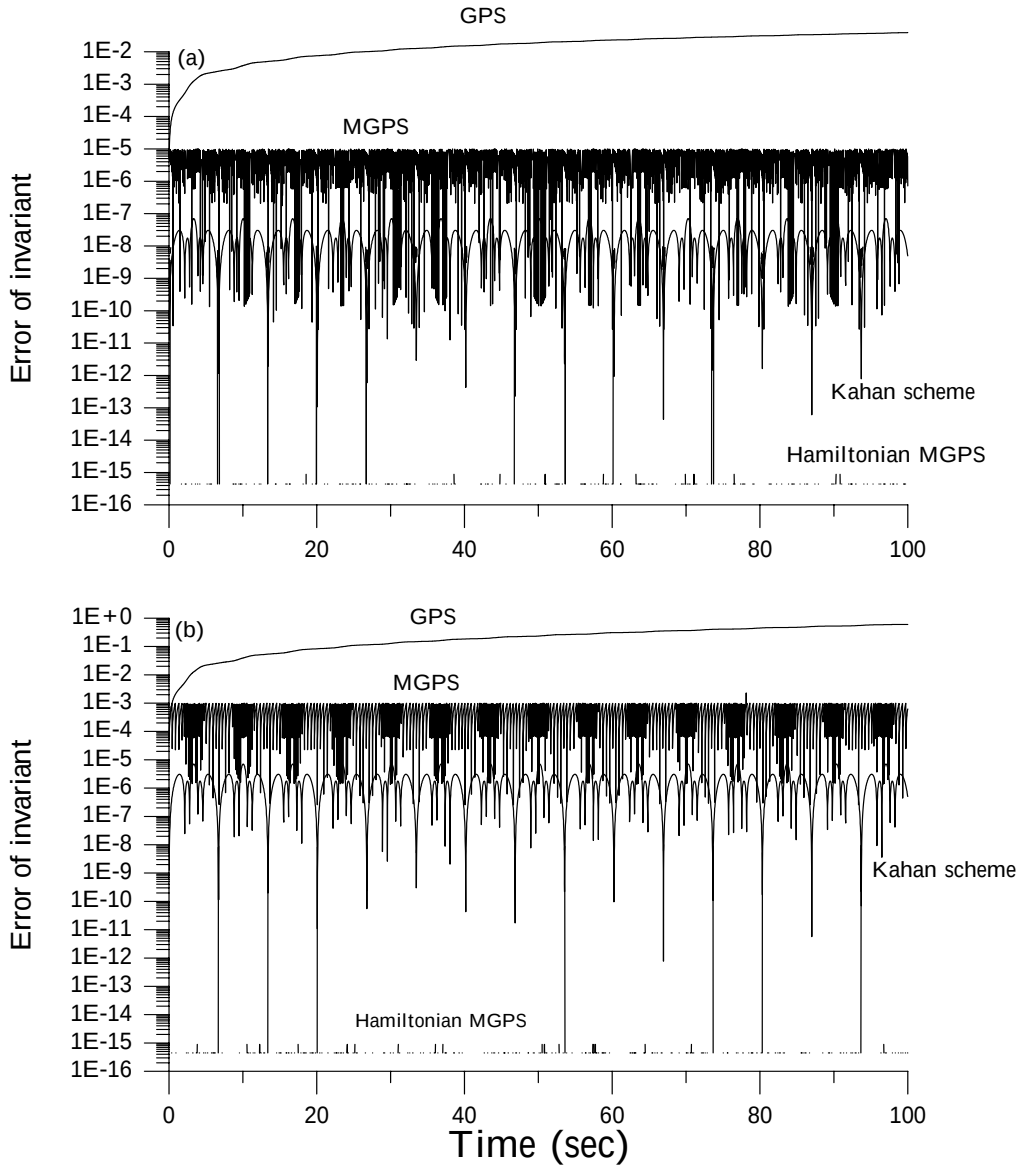

Figure 4 : Compare the errors of invariant for Example 1 by different numerical methods with time stepsizes of (a) $\Delta t = 0.001$ and (b) $\Delta t = 0.01$.

4.2 Example 2

Let us consider the following example:

$$\dot{x} = -2y - x\sin(xy), \tag{31}$$

$$\dot{y} = 2x + y\sin(xy). \tag{32}$$

A prime feature of the above system is that its fixed point $(\overline{x},\overline{y}) = (0,0)$ is neutral stable, and the conserved quantity of the above system is

$$\begin{aligned}\rho(x,y,x_0,y_0) &= H(x,y) - H(x_0,y_0)\\ &= x^2 + y^2 - \cos(xy) - x_0^2 - y_0^2 + \cos(x_0 y_0)\\ &= 0.\end{aligned} \tag{33}$$

Starting from the initial conditions of $(x_0,y_0) = (2,0)$ at $t = 0$ we apply the MGPS to the above system within the time of 10 seconds, where the time stepsize is taken to be $\Delta t = 0.005$ sec and the error tolerance used in the Newton-Raphson method to solve Eq. (18) with the

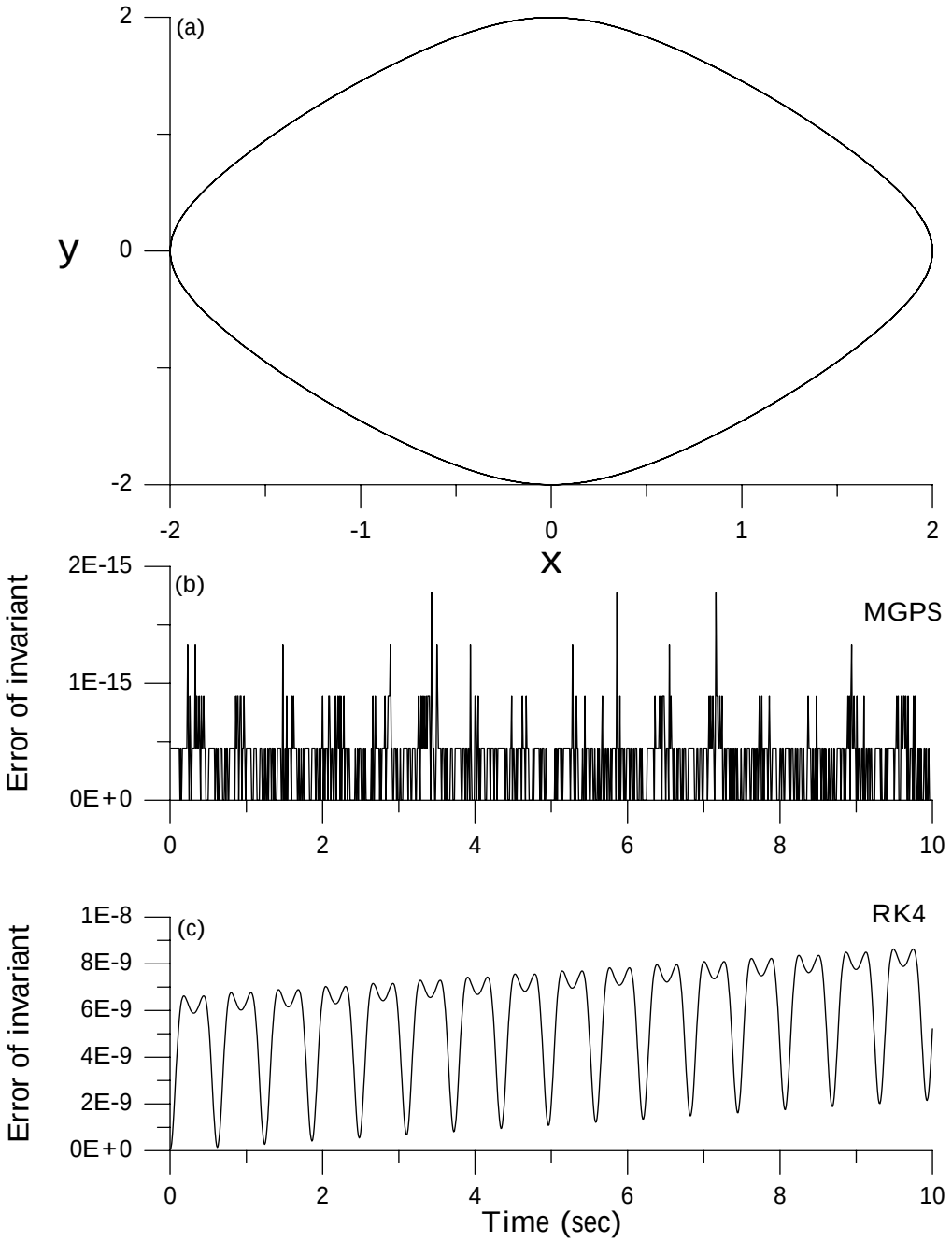

Figure 5 : Displaying the results for Example 2 calculated by MGPS and RK4: (a) the orbit of (x,y), (b) the error of invariant by MGPS, and (c) the error of invariant by RK4.

above ρ is 10^{-5}. Figure 5(a) shows the orbit of (x,y) and Fig. 5(b) shows the error of invariant. It can be seen that the error is smaller than 2×10^{-15}. At the same time we apply the RK4 method to integrate Eqs. (31) and (32), whose error of invariant is shown in Fig. 5(c). It can be seen that the error induced by RK4 is much larger than that by our method and is increasing with time gradually.

4.3 Example 3

The constraints of Examples 1 and 2 are of the constraint that imposed on all variables. In this example we consider a constraint on partial variables. The following sys-

tem:

$$\dot{\mathbf{x}} = \mathbf{y}, \tag{34}$$

$$\dot{\mathbf{y}} = -\frac{1}{m}\left(k\mathbf{x} + c\mathbf{y} + \mathbf{r} - \mathbf{p}\right), \tag{35}$$

$$\dot{\mathbf{r}} = k_d\mathbf{y} - \frac{k_d}{r_y^2}\mathbf{r}\cdot\mathbf{y}\mathbf{r} \tag{36}$$

describes the sliding phase motion of a two-dimensional Coulomb friction oscillator with mass m subjecting to external excitation $\mathbf{p} \in \mathbb{R}^2$ [Liu, Hong and Liou (2003)].

In above, $\mathbf{x}, \mathbf{y}, \mathbf{r} \in \mathbb{R}^2$ are respectively the displacement, velocity and friction force vectors. Especially, $\mathbf{r}$ is subjecting to the following constraint:

$$\|\mathbf{r}\|^2 = r_1^2 + r_2^2 = r_y^2. \tag{37}$$

Eqs. (34)-(36) constitute a six-dimensional cubic nonlinear system with a partial constraint (37) on $\mathbf{r}$. Here we let $p_1 = p_0\cos\omega_d t$ and $p_2 = p_0\sin\omega_d t$ and fix $m = 22500/\pi^2$ kN s^2/m, $c = 600/\pi$ kN s/m, $k_d = 50000$ kN/m, $k = 10000$ kN/m, $r_y = 50$ kN, $p_0 = 500$ kN, and $\omega_d = 4\pi$ rad/s. Figure 6(a) shows the orbit of (r_1, r_2) which tracing a circle with radius r_y and Fig. 6(b) shows the error of invariant. It can be seen that the error is smaller than 10^{-12}. At the same time we apply the RK4 method to integrate Eqs. (34)-(36), whose error of invariant is shown in Fig. 6(c). It can be seen that the error induced by RK4 is much larger than that by our method and has some peaks. In Fig. 7 we show the path of the two components of displacement, and phase portraits of (x_1, y_1) and (x_2, y_2).

4.4 Example 4

We consider an index 2 differential algebraic equation given by Maerz and Tischendorf (1994) and Rheinboldt (1997):

$$\dot{u}_1 + \sqrt{1-u_1^2} - \frac{1}{u_1^2} + w^2 + 1 = 0, \tag{38}$$

$$\dot{u}_2 + w = 0, \tag{39}$$

$$u_2 - \ln u_1 = 0. \tag{40}$$

For $(u_1(0), u_2(0)) = (1, 0)$, $w(0) = 0$, the exact solution is $u_1(t) = \cos t$, $u_2(t) = \ln\cos t$ and $w(t) = \tan t$.

Before applying the new schemes to the above equations we transform them to the following differential equations:

$$\dot{u}_1 = \frac{1}{u_1^2} - \sqrt{1-u_1^2} - w^2 - 1, \tag{41}$$

$$\dot{w} = \frac{u_1}{u_1 - 2w}\left[\left(\frac{\dot{u}_1}{u_1}\right)^2 + \frac{2\dot{u}_1}{u_1^4} - \frac{\dot{u}_1}{\sqrt{1-u_1^2}}\right], \tag{42}$$

subjecting to a constraint:

$$w^2 - u_1 w - \frac{1}{u_1^2} + 1 + \sqrt{1-u_1^2} = 0. \tag{43}$$

When u_1 and w are calculated by the above system, u_2 is calculated by $u_2 = \ln u_1$.

Maerz and Tischendorf (1994) and Rheinboldt (1997) have integrated the above system from $t = 0.5$ sec to $t = 1.5$ sec using accurate starting values with a fixed time stepsize of 10^{-5} sec by BEF-solver of order 2 and DAEN2, respectively. The resulting errors at $t = 1.5$ sec are compared in Table 1. Under the same accurate starting values and same stepsize, it can be seen that the MRK4 method proposed in Section 3 is accurate two orders than that calculated by Maerz and Tischendorf (1994) and Rheinboldt (1997). In Fig. 8 we show the numerical errors of u_1, u_2 and w by MGPS and MRK4.

4.5 Example 5

We consider an index 3 differential algebraic equations example given by Arévalo and Lotstedt (1995) and Sand (2002), which describes the position of a particle on a circular track:

$$\ddot{u}_1 = 2u_2 + \lambda u_1, \tag{44}$$

$$\ddot{u}_2 = -2u_1 + \lambda u_2, \tag{45}$$

$$u_1^2 + u_2^2 = 1. \tag{46}$$

For $(u_1(0), u_2(0)) = (0, 1)$, $\lambda(0) = 0$, the exact solution is $u_1(t) = \sin t^2$, $u_2(t) = \cos t^2$ and $\lambda(t) = -4t^2$.

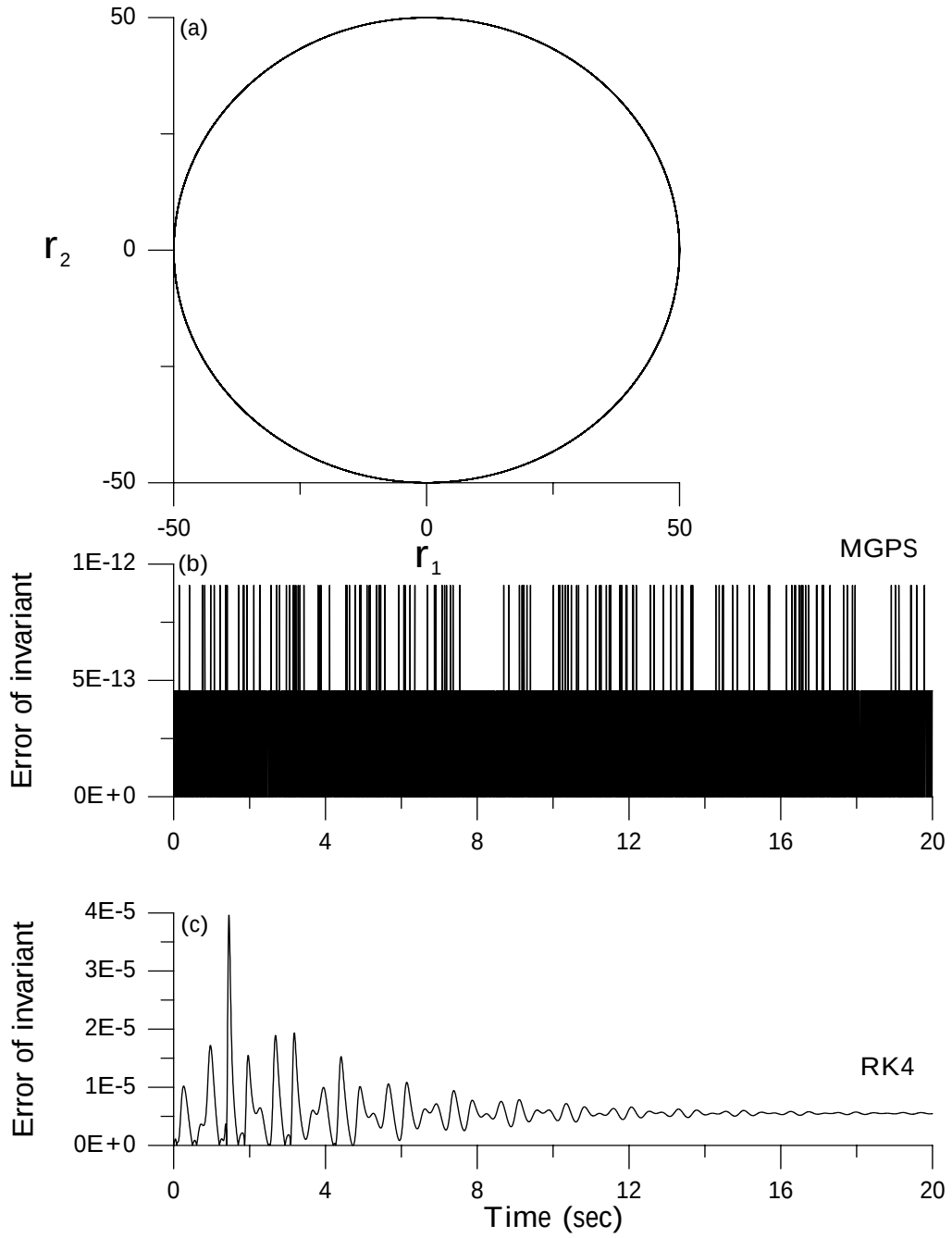

Figure 6 : Displaying the results for Example 3 calculated by MGPS and RK4: (a) the orbit of (r_1, r_2), (b) the error of invariant by MGPS, and (c) the error of invariant by RK4.

Table 1 : Errors of Example 4 for an index 2 problem by using different numerical schemes.

method	Error(u_1)	Error(u_2)	Error(w)
Maerz and Tischendorf (1994)	0.560×10^{-9}	0.791×10^{-8}	0.112×10^{-6}
Rheinboldt (1997)	2.734×10^{-10}	3.115×10^{-9}	5.476×10^{-8}
MRK4	3.738×10^{-12}	5.212×10^{-11}	7.286×10^{-10}

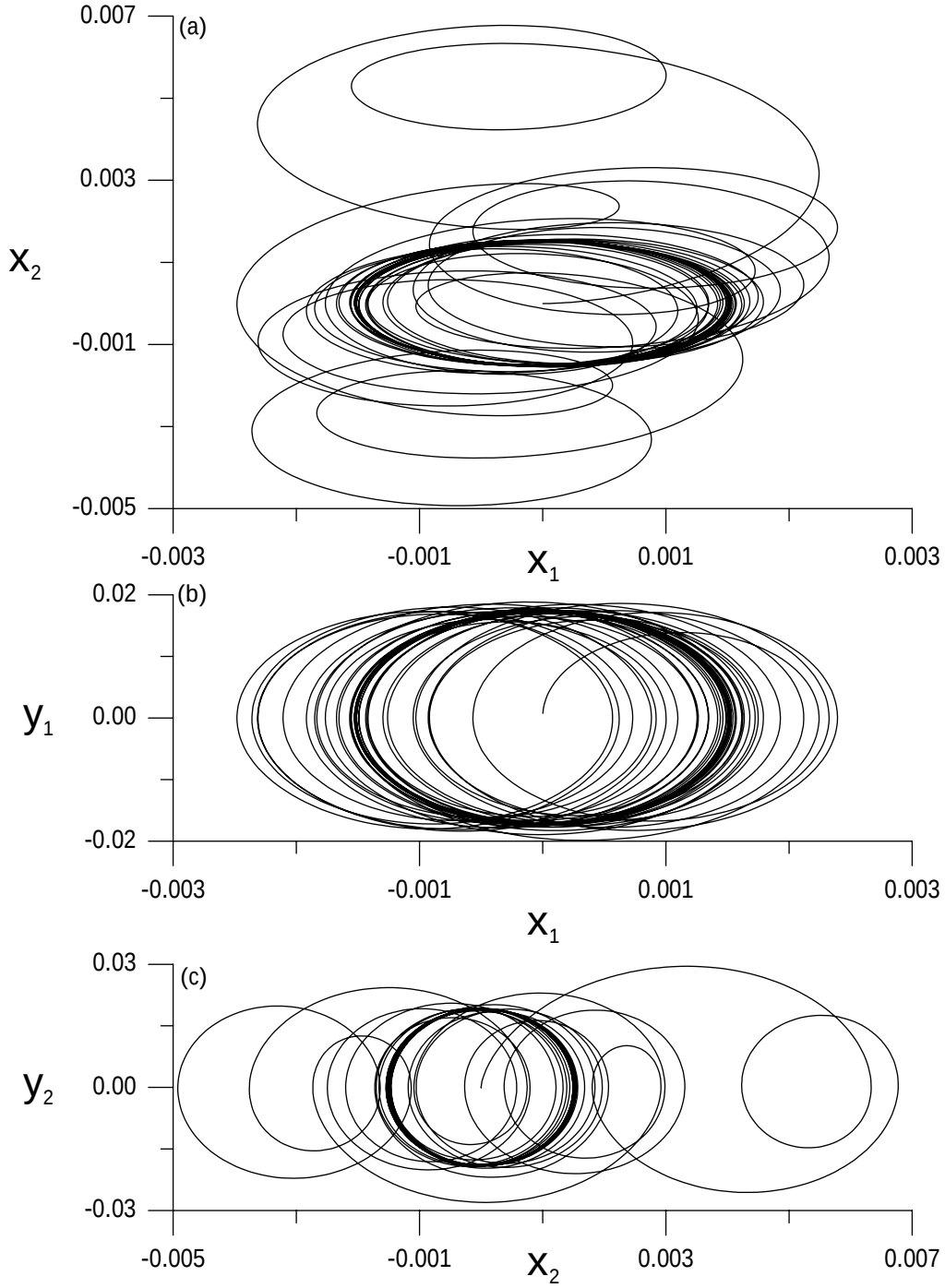

Figure 7 : Displaying the results for Example 3 calculated by MGPS: (a) the orbit of (x_1, x_2), (b) the orbit of (x_1, y_1) and, (c) the orbit of (x_2, y_2).

If we let $x_1 = u_1$, $x_2 = \dot{u}_1$, $x_3 = u_2$ and $x_4 = \dot{u}_2$, then $\dot{x}_2 = 2x_3 - x_1(x_2^2 + x_4^2)$, (48)
through some derivations we find that the above system
can be transformed to

$\dot{x}_1 = x_2$, (47) $\dot{x}_3 = x_4$, (49)

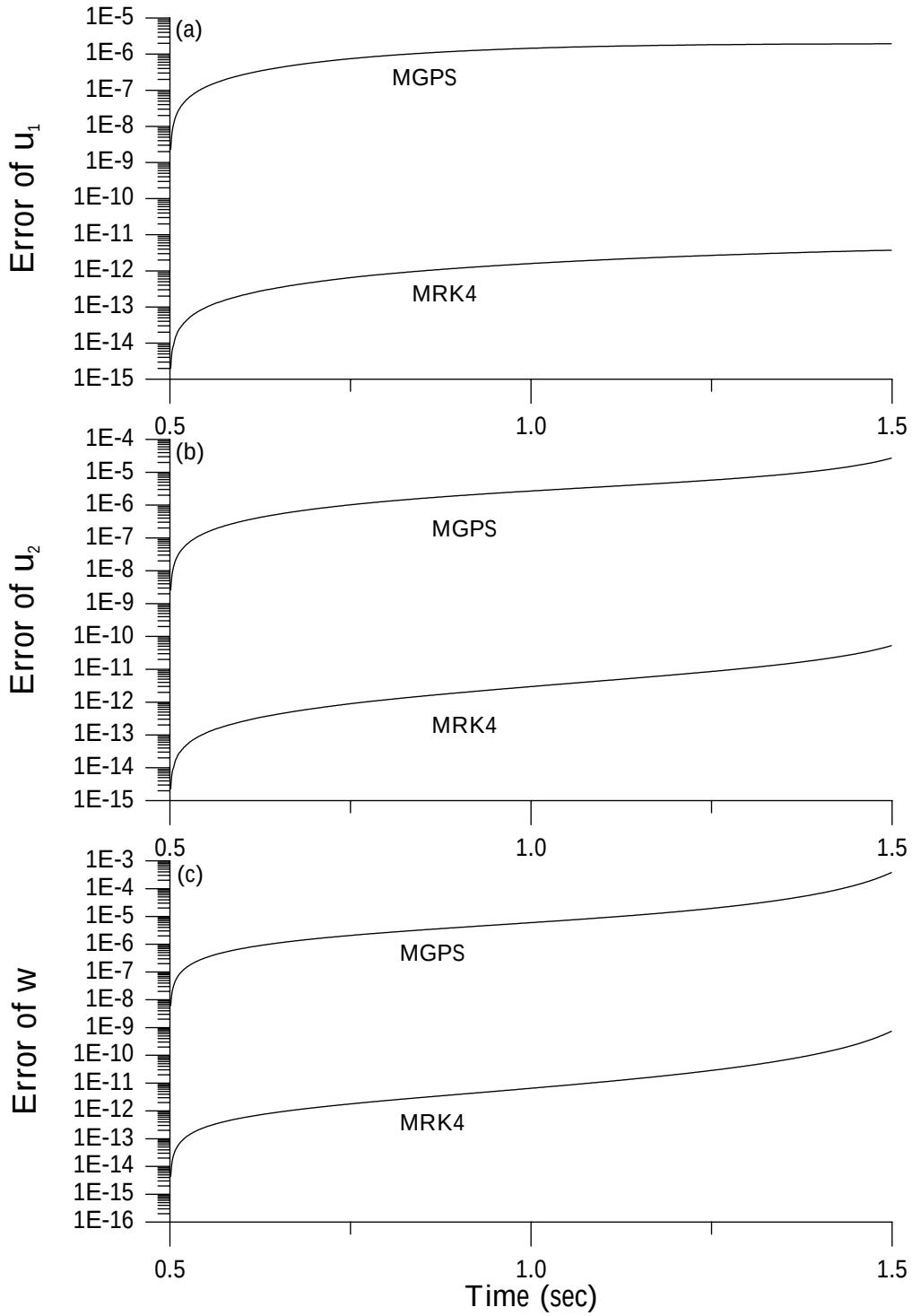

Figure 8 : Comparing the numerical errors for Example 4 calculated by MGPS and MRK4: (a) the error of u_1, (b) the error of u_2, and (c) the error of w.

$$\dot{x}_4 = -2x_1 - x_3(x_2^2 + x_4^2), \tag{50}$$

subjecting to a constraint on (x_1, x_3):

$$x_1^2 + x_3^2 = 1, \tag{51}$$

and the Lagrange multiplier λ is calculated by

$$\lambda = -x_2^2 - x_4^2. \tag{52}$$

In Fig. 9 we show the numerical errors of u_1, u_2 and λ by MGPS and MRK4 with a fixed time stepsize of $\Delta t = 0.005$ sec.

5 Numerical scheme for generalized Hamiltonian system

It is well known that any autonomous system (1), subjected to the constraint (2), can be written in a skew-gradient form:

$$\dot{\mathbf{x}} = \mathbf{J}\nabla_{\mathbf{x}} H, \tag{53}$$

where H is a generalized Hamiltonian function of $\mathbf{x}$, $\nabla_{\mathbf{x}}$ denotes the gradient with respect to $\mathbf{x}$, and $\mathbf{J}$ is an $n \times n$ skew-symmetric matrix function; see, e.g., Iserles and Zanna (2000) and Liu (2002). For example, Eqs. (19) and (20) can be written as

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 & -xy \\ xy & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial H}{\partial x} \\ \frac{\partial H}{\partial y} \end{bmatrix}, \tag{54}$$

with $H(x,y) = \ln x - x + \ln y - y$ and a noncanonical symplectic metric:

$$\mathbf{J} := \begin{bmatrix} 0 & -xy \\ xy & 0 \end{bmatrix}. \tag{55}$$

After developing numerical scheme for Eq. (53) we will return to this example again.

Theorem 1. *For the generalized Hamiltonian system (53) with its H not dependent on t and being a regular and strictly convex function of* $\mathbf{x}$*, we can obtain a nonlinear Lorentzian system as follows*:

$$\dot{\mathbf{X}} = \mathbf{A}\mathbf{X}, \tag{56}$$

where $\mathbf{A} \in so(n,1)$ is a matrix function of $\mathbf{X}$, which satisfying the cone condition $\mathbf{X}^T\mathbf{g}\mathbf{X} = 0$.

Proof. Let us introduce a unit n-dimensional orientation vector,

$$\mathbf{n} := \frac{\nabla_{\mathbf{x}} H}{\|\nabla_{\mathbf{x}} H\|}, \tag{57}$$

which is well-defined according to the assumption of the regularity of H. Because $\nabla_{\mathbf{x}} H$ is a strictly monotonic operator by the assumption of H strictly convex, i.e.,

$$\mathbf{H}(\mathbf{x}) := \nabla_{\mathbf{x}}^2 H(\mathbf{x}) > \mathbf{0}, \tag{58}$$

there exists a homeomorphism $\mathbf{F}$ between $\mathbf{x}$ and $\nabla_{\mathbf{x}} H = \|\nabla_{\mathbf{x}} H\|\mathbf{n}$, such that

$$\mathbf{x} = \mathbf{F}(\nabla_{\mathbf{x}} H) = \mathbf{F}(\|\nabla_{\mathbf{x}} H\|\mathbf{n}). \tag{59}$$

In the following a methodology will be developed to embed the pair $(\mathbf{n}, \|\nabla_{\mathbf{x}} H\|)$ into the Minkowski space, and a system of equations to calculate $(\mathbf{n}, \|\nabla_{\mathbf{x}} H\|)$ will be derived.

Taking the time differential of Eq. (57) we obtain

$$\dot{\mathbf{n}} = \frac{\mathbf{H}\dot{\mathbf{x}}}{\|\nabla_{\mathbf{x}} H\|} - \frac{(\mathbf{n}\cdot\mathbf{H}\dot{\mathbf{x}})\mathbf{n}}{\|\nabla_{\mathbf{x}} H\|}. \tag{60}$$

Applying the operator $\mathbf{H}/\|\nabla_{\mathbf{x}} H\|$ to Eq. (53), leads to

$$\frac{\mathbf{H}\dot{\mathbf{x}}}{\|\nabla_{\mathbf{x}} H\|} = \mathbf{A}_0^s, \tag{61}$$

where

$$\mathbf{A}_0^s := \mathbf{H}\mathbf{J}\mathbf{n} =: \mathbf{K}\mathbf{n}. \tag{62}$$

The inner product of Eq. (61) with $\mathbf{n}$ generates

$$\frac{\mathbf{n}\cdot(\mathbf{H}\dot{\mathbf{x}})}{\|\nabla_{\mathbf{x}} H\|} = \mathbf{A}_0^s \cdot \mathbf{n}, \tag{63}$$

which renders obviously

$$\frac{\mathbf{n}\cdot(\mathbf{H}\dot{\mathbf{x}})}{\|\nabla_{\mathbf{x}} H\|}\mathbf{n} = (\mathbf{A}_0^s \cdot \mathbf{n})\mathbf{n}. \tag{64}$$

Substituting Eqs. (64) and (61) into Eq. (60) we obtain

$$\dot{\mathbf{n}} = \mathbf{A}_0^s - (\mathbf{A}_0^s \cdot \mathbf{n})\mathbf{n}. \tag{65}$$

Upon defining the integrating factor

$$X^0(t) := \|\nabla_{\mathbf{x}} H(\mathbf{x}(0))\| \exp\left[\int_0^t [\mathbf{A}_0^s(\xi)\cdot\mathbf{n}(\xi)]d\xi\right], \tag{66}$$

Eqs. (65) and (66) become, respectively,

$$\frac{d}{dt}(X^0\mathbf{n}) = X^0\mathbf{A}_0^s, \tag{67}$$

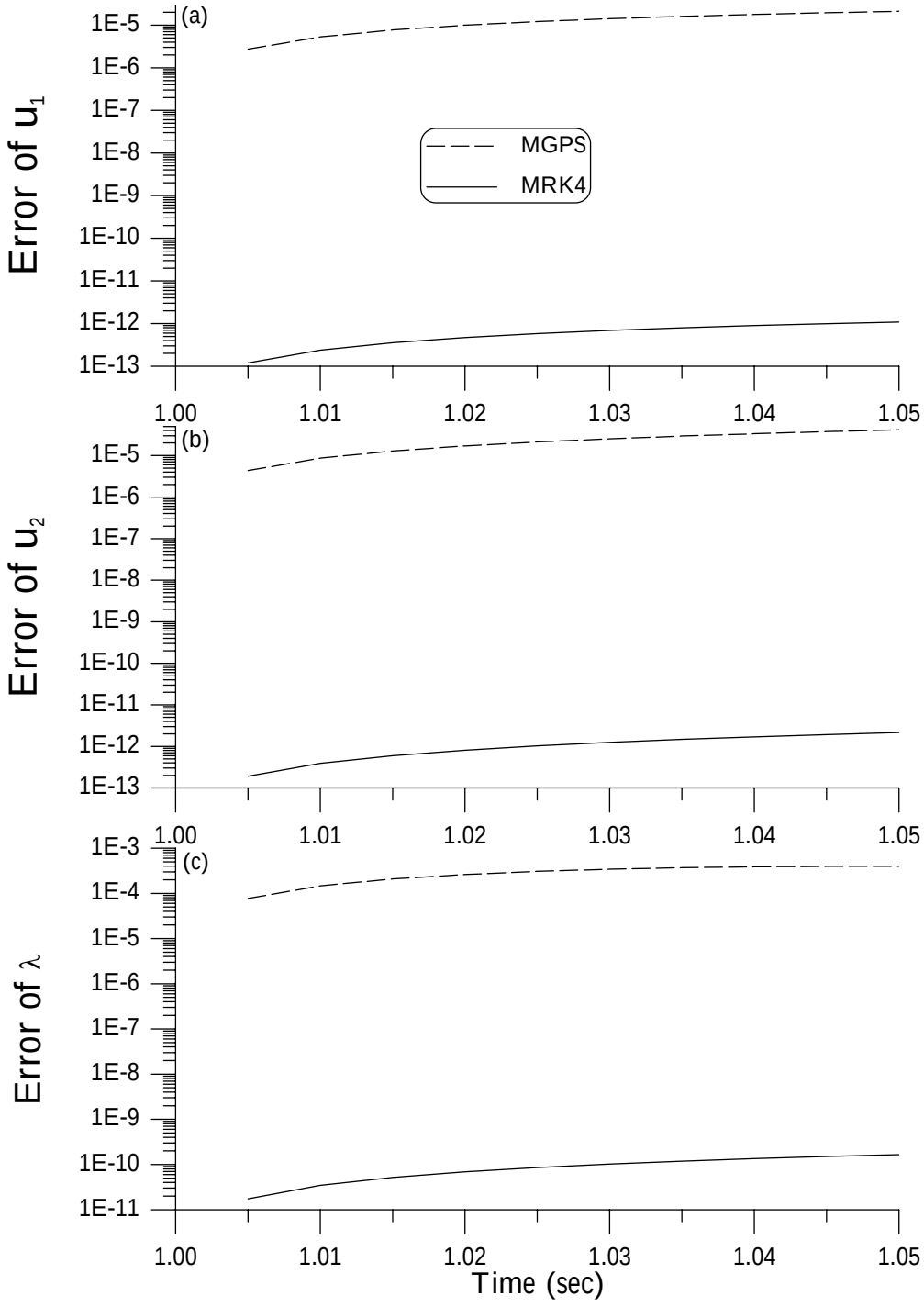

Figure 9 : Comparing the numerical errors for Example 5 calculated by MGPS and MRK4: (a) the error of u_1, (b) the error of u_2, and (c) the error of λ.

$$\frac{d}{dt}X^0 = X^0 \mathbf{A}_0^s \cdot \mathbf{n}. \tag{68}$$

In terms of the homogeneous coordinates

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^s \\ X^0 \end{bmatrix} := X^0 \begin{bmatrix} \mathbf{n} \\ 1 \end{bmatrix}, \tag{69}$$

Eqs. (67) and (68) together leads to Eq. (56) with

$$\mathbf{A} := \begin{bmatrix} \mathbf{0}_n & \mathbf{A}_0^s \\ (\mathbf{A}_0^s)^{\mathrm{T}} & 0 \end{bmatrix} \tag{70}$$

satisfying the Lie algebra condition (7).

Substituting Eq. (62) into Eq. (68) it follows that

$$\frac{\dot{X}^0}{X^0} = \mathbf{n} \cdot \mathbf{K}\mathbf{n}. \tag{71}$$

On the other hand, with the aid of Eqs. (57) and (58) we obtain

$$\frac{d}{dt}\|\nabla_{\mathbf{x}} H\| = \mathbf{n} \cdot \mathbf{H}\dot{\mathbf{x}}, \tag{72}$$

which, via Eqs. (53) and (62), becomes

$$\frac{d}{dt}\|\nabla_{\mathbf{x}} H\| = \mathbf{n} \cdot \mathbf{K}\nabla_{\mathbf{x}} H. \tag{73}$$

Dividing the above equation by $\|\nabla_{\mathbf{x}} H\|$, using Eq. (57) and comparing the resultant with Eq. (71) we obtain

$$\frac{\dot{X}^0}{X^0} = \frac{d\|\nabla_{\mathbf{x}} H\|/dt}{\|\nabla_{\mathbf{x}} H\|}. \tag{74}$$

Integrating and using the initial condition $X^0(0) = \|\nabla_{\mathbf{x}} H(\mathbf{x}(0))\|$ derived from Eq. (66), gives us a meaningful relation,

$$X^0 = \|\nabla_{\mathbf{x}} H\|. \tag{75}$$

With this and Eqs. (57) and (69) the following identity is verified:

$$\mathbf{X}^s = \nabla_{\mathbf{x}} H. \tag{76}$$

If $\mathbf{X}$ is available from Eq. (56) then using Eq. (57) and the strict convexity of the Hamiltonian function one can obtain $\mathbf{x}$ through the homeomorphism between $\mathbf{X}^s$ and $\mathbf{x}$ as shown in Eq. (59). In summary, we have a nonlinear equations system (56) with

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^s \\ X^0 \end{bmatrix} = \begin{bmatrix} \nabla_{\mathbf{x}} H \\ \|\nabla_{\mathbf{x}} H\| \end{bmatrix}, \tag{77}$$

and the following $\mathbf{A}_0^s$:

$$\mathbf{A}_0^s = \frac{\mathbf{K}\nabla_{\mathbf{x}} H}{\|\nabla_{\mathbf{x}} H\|}, \tag{78}$$

which is obtained from Eq. (62) to replace $\mathbf{n}$ by $\nabla_{\mathbf{x}} H/\|\nabla_{\mathbf{x}} H\|$, and $\mathbf{K}$ is a function of $\nabla_{\mathbf{x}} H$ through the homeomorphism (59).

From the definition (69) it is very natural to endow a *cone* in the Minkowski space,

$$\mathbf{X}^{\mathrm{T}}\mathbf{g}\mathbf{X} = 0. \tag{79}$$

Furthermore, in terms of $(\nabla_{\mathbf{x}} H, \|\nabla_{\mathbf{x}} H\|)$ through the identification (77), the following condition is obvious

$$\nabla_{\mathbf{x}} H \cdot \nabla_{\mathbf{x}} H - \|\nabla_{\mathbf{x}} H\|^2 = \|\nabla_{\mathbf{x}} H\|^2 - \|\nabla_{\mathbf{x}} H\|^2 = 0, \tag{80}$$

which is a natural condition that we can impose on system (53). The above ends the proof. □

Similarly, we can apply scheme (16) to the $\mathbf{n}$ defined by Eq. (57) but with $\mathbf{A}_0^s$ calculated from Eq. (62). If $\mathbf{n}(\ell+1)$ is available, which together with Eqs. (59) and (75) being substituted into the constraint (2) leads to a nonlinear equation for X^0:

$$\rho(\mathbf{F}(X^0(\ell+1)\mathbf{n}(\ell+1)), \mathbf{x}_0) = 0. \tag{81}$$

Substituting $\mathbf{n}(\ell+1)$ into the above equation and solving it by the Newton-Raphson method we may obtain $X^0(\ell+1)$. With the new $X^0(\ell+1)$ we update $\mathbf{X}^s(\ell+1)$ to a new $\mathbf{X}^s(\ell+1) = X^0(\ell+1)\mathbf{n}(\ell+1)$, such that by Eqs. (59) and (76) we can calculate $\mathbf{x}(\ell+1)$. $\mathbf{A}_0^s$ defined in Eq. (62) can be calculated, and then use scheme (16) to calculate the next $\mathbf{n}$ and Eq. (81) the next X^0. Here we call such a scheme the *Hamiltonian MGPS*, which preserves the Hamiltonian function invariant. This technique has been applied by Liu and Chang (2004) to the computation of a convex plasticity equation.

For demonstration let us return to Example 1 in Section 4.1 again, of which we have

$$\begin{aligned} \mathbf{x} &= \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{1+X^0 n_1} \\ \frac{1}{1+X^0 n_2} \end{bmatrix}, \\ \nabla_{\mathbf{x}} H &= \begin{bmatrix} \frac{1}{x} - 1 \\ \frac{1}{y} - 1 \end{bmatrix} = \begin{bmatrix} X^0 n_1 \\ X^0 n_2 \end{bmatrix}, \end{aligned} \tag{82}$$

$$\begin{aligned} \mathbf{K} &= \begin{bmatrix} 0 & \frac{1+X^0 n_1}{1+X^0 n_2} \\ -\frac{1+X^0 n_2}{1+X^0 n_1} & 0 \end{bmatrix}, \\ \mathbf{A}_0^s &= \begin{bmatrix} \frac{(1+X^0 n_1) n_2}{1+X^0 n_2} \\ \frac{-(1+X^0 n_2) n_1}{1+X^0 n_1} \end{bmatrix}. \end{aligned} \tag{83}$$

If n_1 and n_2 are calculated we need to solve the following equation for X^0:

$$\ln(1+X^0 n_1)+\frac{1}{1+X^0 n_1}+\ln(1+X^0 n_2) + \frac{1}{1+X^0 n_2}+\ln x_0 - x_0 + \ln y_0 - y_0 = 0. \tag{84}$$

In order to reduce the error of invariant we apply the Hamiltonian MGPS to Eqs. (19) and (20), where the error tolerance for applying the Newton-Raphson method to solve Eq. (84) is taken to be 10^{-8}. Under the same conditions as that given in Section 4.1, the numerical errors of invariant by this scheme were shown in Figs. 4(a) and 4(b). It can be seen that the errors of invariant are greatly reduced to the order 10^{-16} for $\Delta t = 0.001$ sec and the order 10^{-15} for $\Delta t = 0.01$ sec.

6 Multiple-constraint preserving schemes

There are many physical systems that are subjected to multiple constraints; for example, the Euler equations of rigid body dynamics (Example 6):

$$\frac{d}{dt}\begin{bmatrix}\Pi_1\\ \Pi_2\\ \Pi_3\end{bmatrix} = \begin{bmatrix} 0 & \frac{\Pi_3}{I_3} & \frac{-\Pi_2}{I_2}\\ \frac{-\Pi_3}{I_3} & 0 & \frac{\Pi_1}{I_1}\\ \frac{\Pi_2}{I_2} & \frac{-\Pi_1}{I_1} & 0\end{bmatrix}\begin{bmatrix}\Pi_1\\ \Pi_2\\ \Pi_3\end{bmatrix}, \tag{85}$$

where $I_1, I_2, I_3 > 0$ are the three principal moments of inertia of the rigid body, and Π_1, Π_2, Π_3 are the three components of the rigid body angular momentum.

We know that the system of Euler equations possesses two invariants; the first is the momentum, a Casimir function:

$$C := \frac{1}{2}\|\mathbf{\Pi}\|^2, \tag{86}$$

and the second is the energy, a Hamiltonian:

$$H := \frac{1}{2}\mathbf{\Pi}\cdot\mathbf{J}^{-1}\mathbf{\Pi}, \tag{87}$$

where $\mathbf{J}$ is the inertia tensor of the body:

$$\mathbf{J} := \begin{bmatrix} I_1 & 0 & 0\\ 0 & I_2 & 0\\ 0 & 0 & I_3\end{bmatrix}. \tag{88}$$

Figures 10(a) and 10(b) show the errors for the above two invariants due to the difference between the results calculated by applying the MGPS to Eq. (85) and the exact values. Since in the modified scheme we are solved Eq. (2) for X^0 by substituting the H of Eq. (87), it can be seen that the energy error is in the order of 10^{-16} (due to a machinery round-off error), but that the momentum error is rather large in the order of 10^{-4}, and is gradually increasing. At the same time we also applied RK4 to this example, and the errors of invariants were also plotted in Figs. 10(a) and 10(b). Due to its high accuracy RK4 gave very small errors. However, it does not truly preserve the constaints. In order to genuinely retain the constaints, let us extend the single-constraint preserving numerical methods developed in Section 3 to the multiple-constraint dynamical systems.

6.1 Scheme one: the group-preserving method

In order to obtain a much better scheme to calculate the dynamical system (1) which subjecting to k constraints, $2 \le k < n$:

$$\rho_i(\mathbf{x}(t),\mathbf{x}_0) = 0, \quad i = 1,2,\ldots,k, \tag{89}$$

let us first divide the variables $\mathbf{x}$ into k independent sets, such that Eq. (1) can be written as:

$$\begin{aligned}&\dot{\mathbf{x}}_1 = \mathbf{f}_1(\mathbf{x}_1,\ldots,\mathbf{x}_k,t),\\ &\vdots\\ &\dot{\mathbf{x}}_k = \mathbf{f}_k(\mathbf{x}_1,\ldots,\mathbf{x}_k,t),\end{aligned} \tag{90}$$

with $\mathbf{x}_1 \in \mathbb{R}^{n_1}$, ..., $\mathbf{x}_k \in \mathbb{R}^{n_k}$, $1 \le n_i < n$, $i = 1,\ldots,k$, and $n_1 + \ldots + n_k = n$. Correspondingly, the constraints in Eq. (89) are written as

$$\begin{aligned}&\rho_1(\mathbf{x}_1,\ldots,\mathbf{x}_k,\mathbf{x}_0) = 0,\\ &\vdots\\ &\rho_k(\mathbf{x}_1,\ldots,\mathbf{x}_k,\mathbf{x}_0) = 0.\end{aligned} \tag{91}$$

For $\mathbf{x}_i$, $i = 1,\ldots,k$, we introduce the orientation vectors and integrating factors by:

$$\mathbf{n}_i := \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}, \quad i = 1,\ldots,k, \tag{92}$$

$$X_i^0 := \|\mathbf{x}_i\|, \quad i = 1,\ldots,k, \tag{93}$$

such that we have k mathematically equivalent dynamical systems:

$$\dot{\mathbf{X}}_i = \mathbf{A}_i\mathbf{X}_i, \quad i = 1,\ldots,k \ \ (i \ \text{not summed}) \tag{94}$$

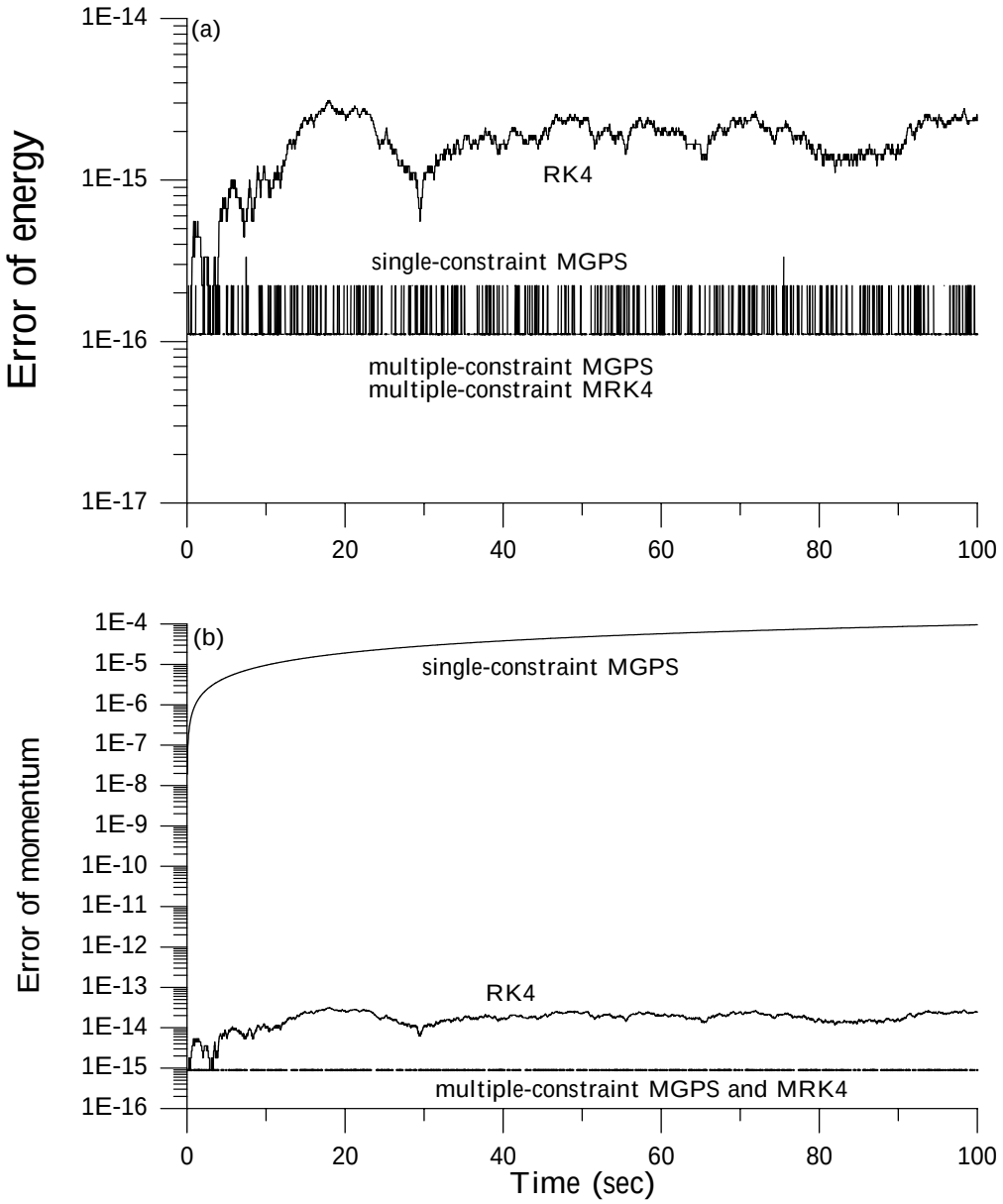

Figure 10 : Comparing the errors of invariants for Example 6 calculated by different numerical methods: (a) the error of energy, and (b) the error of momentum.

with

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{X}_i^s \\ X_i^0 \end{bmatrix} := \begin{bmatrix} X_i^0 \mathbf{n}_i \\ X_i^0 \end{bmatrix}, \ i = 1, \ldots, k \ (i \text{ not summed}), \tag{95}$$

and

$$\mathbf{A}_i := \begin{bmatrix} \mathbf{0}_{n_i} & \frac{\mathbf{f}_i(\mathbf{x}_1,\ldots,\mathbf{x}_k,t)}{\|\mathbf{x}_i\|} \\ \frac{\mathbf{f}_i^{\mathrm{T}}(\mathbf{x}_1,\ldots,\mathbf{x}_k,t)}{\|\mathbf{x}_i\|} & 0 \end{bmatrix}. \tag{96}$$

It can be seen that $\mathbf{A}_i$ is a Lie algebra of the proper orthochronous Lorentz group $SO_o(n_i,1)$ satisfying

$$\mathbf{A}_i^{\mathrm{T}} \mathbf{g}_i + \mathbf{g}_i \mathbf{A}_i = \mathbf{0}, \tag{97}$$

with

$$\mathbf{g}_i = \begin{bmatrix} \mathbf{I}_{n_i} & \mathbf{0}_{n_i \times 1} \\ \mathbf{0}_{1 \times n_i} & -1 \end{bmatrix}. \tag{98}$$

Therefore, for the dynamical system with k constraints we observe that the internal symmetry group is the di-

rect product of the k groups with $SO_o(n_1,1)\otimes\ldots\otimes SO_o(n_k,1)$, which left acts on the totally k cones:

$$\mathbf{X}_1^{\mathrm{T}}\mathbf{g}_1\mathbf{X}_1=0,\ldots,\mathbf{X}_k^{\mathrm{T}}\mathbf{g}_k\mathbf{X}_k=0 \tag{99}$$

in the product Minkowski space of $\mathbb{M}^{n_1+1}\otimes\ldots\otimes\mathbb{M}^{n_k+1}$.

Now, by applying the modified group preserving scheme for each dynamical equation in Eq. (94), and each $\mathbf{n}_i,\ i=1,\ldots,k$, being calculated from the scheme and then substituting the resultants into the k constraints in Eq. (91) we obtain

$$\begin{aligned}&\rho_1(X_1^0\mathbf{n}_1,\ldots,X_k^0\mathbf{n}_k,\mathbf{x}_0)=0,\\&\vdots\\&\rho_k(X_1^0\mathbf{n}_1,\ldots,X_k^0\mathbf{n}_k,\mathbf{x}_0)=0.\end{aligned} \tag{100}$$

Simultaneously solving the above k algebraic equations for $X_i^0,\ i=1,\ldots,k$, then returning to $\mathbf{x}_i=X_i^0\mathbf{n}_i,\ i=1,\ldots,k$, we obtain the solutions of $\mathbf{x}$ which satisfy the k constraints exactly.

6.2 Scheme two: the fourth-order Runge-Kutta method

When dynamical system (1) is subjecting to k constraints as shwon in Eq. (89) we first embed the n-dimensional system to an $n+k$-dimensional system by introducing the k lengths of k independent sets of variables: $(\mathbf{x}_1,\ldots,\mathbf{x}_k)$ as shown in Eq. (93), such that

$$\begin{aligned}&\dot{\mathbf{x}}_1=\mathbf{f}_1(\mathbf{x}_1,\ldots,\mathbf{x}_k,t),\\&\dot{X}_1^0=\frac{1}{X_1^0}\mathbf{x}_1\cdot\mathbf{f}_1(\mathbf{x}_1,\ldots,\mathbf{x}_k,t),\\&\vdots\\&\dot{\mathbf{x}}_k=\mathbf{f}_k(\mathbf{x}_1,\ldots,\mathbf{x}_k,t),\\&\dot{X}_k^0=\frac{1}{X_k^0}\mathbf{x}_k\cdot\mathbf{f}_k(\mathbf{x}_1,\ldots,\mathbf{x}_k,t).\end{aligned} \tag{101}$$

The above differential equations constitute an augmented differential equations system for the $n+k$ augmented variables of $(\mathbf{x}_1,X_1^0,\ldots,\mathbf{x}_k,X_k^0)$.

Applying the fourth-order Runge-Kutta scheme to the system in Eq. (101), instead of to the system (1), we can calculate $(\mathbf{x}_1,X_1^0,\ldots,\mathbf{x}_k,X_k^0)$ for each assigned time step. Then we can calculate the orientation vectors by

$$\mathbf{n}_i:=\frac{\mathbf{x}_i}{X_i^0},\quad i=1,\ldots,k. \tag{102}$$

For each $\mathbf{x}_i$ we substitute the results of $X_i^0\mathbf{n}_i$ into the k constraints to obtain the k simultaneous equations as shown in Eq. (100) for the k extra variables of $X_i^0,\ i=1,\ldots,k$. Similarly, solving the k algebraic equations for $X_i^0,\ i=1,\ldots,k$, then returning to $\mathbf{x}_i=X_i^0\mathbf{n}_i,\ i=1,\ldots,k$, we obtain the solutions of $\mathbf{x}$ which satisfy the k constraints exactly.

7 Numerical examples with multiple-constraint

7.1 Example 6

In order to assess the performance of the newly developed schemes let us return to Example 6. For this example we divide the three independent variables into two independent sets: $\{\Pi_1,\Pi_2\}$ and $\{\Pi_3\}$, and solve the constraints equations to obtain

$$X_1^0=\sqrt{\frac{I_1I_2(2HI_3-2C)}{n_1^2(I_2I_3-I_1I_2)+n_2^2(I_1I_3-I_1I_2)}}, \tag{103}$$

$$X_2^0=\sqrt{\frac{2C-(X_1^0n_1)^2-(X_1^0n_2)^2}{n_3^2}}, \tag{104}$$

where $n_1=\Pi_1/X_1^0$, $n_2=\Pi_2/X_1^0$ and $n_3=\Pi_3/X_2^0$.

For the special case of $I_1=I_2>I_3$, the closed-form solution of the Euler equations is available [see, *e.g.*, Marsden and Ratiu (1994)]:

$$\begin{aligned}\Pi_1(t)=&\Pi_1(0)\cos\frac{(I_3-I_1)\Pi_3(0)}{I_1I_3}t\\&-\Pi_2(0)\sin\frac{(I_3-I_1)\Pi_3(0)}{I_1I_3}t,\end{aligned} \tag{105}$$

$$\begin{aligned}\Pi_2(t)=&\Pi_2(0)\cos\frac{(I_3-I_1)\Pi_3(0)}{I_1I_3}t\\&+\Pi_1(0)\sin\frac{(I_3-I_1)\Pi_3(0)}{I_1I_3}t,\end{aligned} \tag{106}$$

$$\Pi_3(t)=\Pi_3(0). \tag{107}$$

In Fig. 10, the results calculated by using the new schemes were compared with the closed-form solutions while the momentum and energy errors were shown in Figs. 10(a) and 10(b). It can be seen that the multiple-constraint preserving schemes provide more accurate

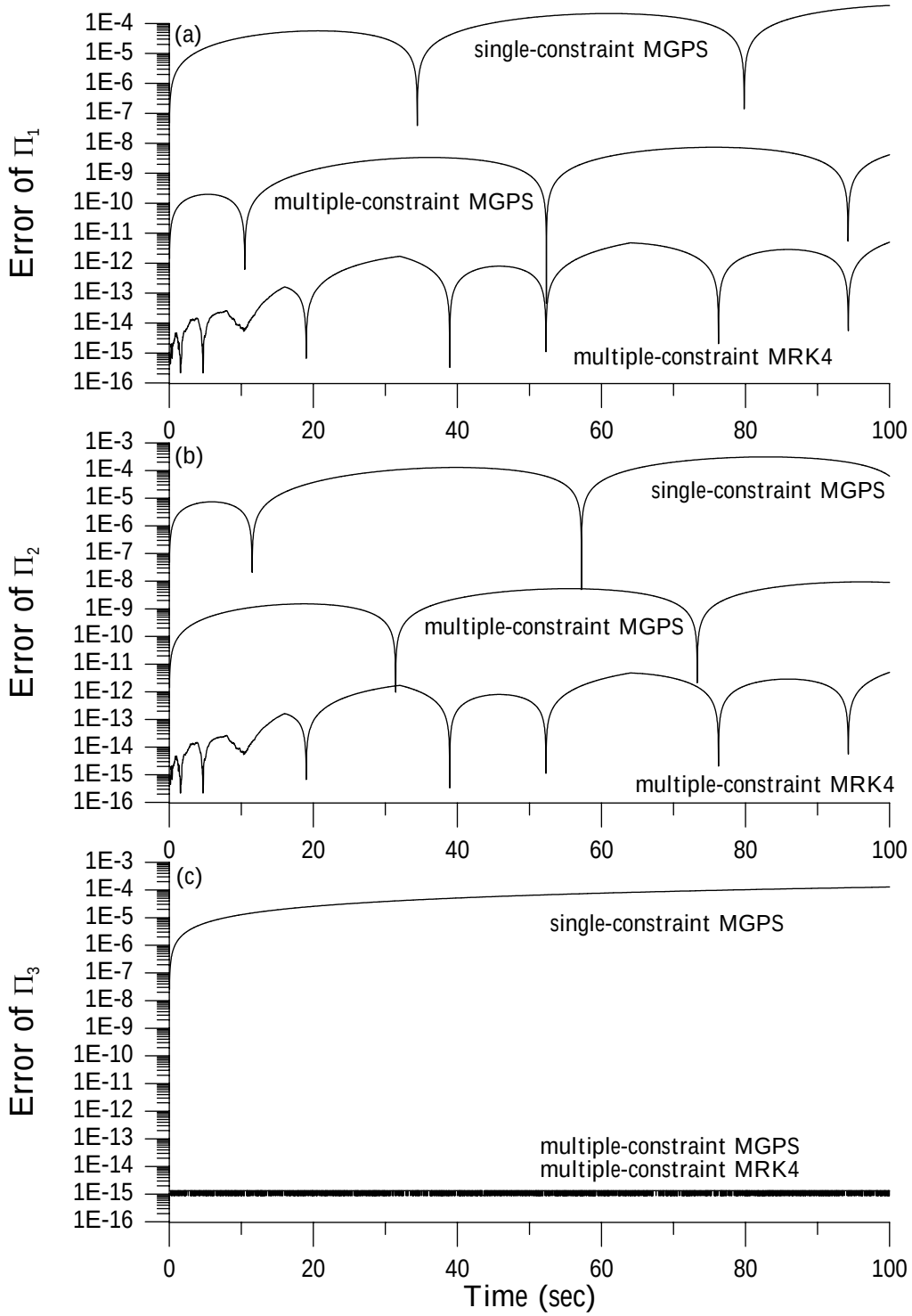

Figure 11 : Comparing the numerical errors for Example 6 calculated by different numerical methods: (a) the error of Π_1, (b) the error of Π_2, and (c) the error of Π_3.

results for the energy and momentum than does the single-constraint GPS. At the same time, as shown in Figs. 11(a)-11(c), the errors of $\Pi_1(t)$, $\Pi_2(t)$ and $\Pi_3(t)$ are also greatly reduced by the multiple-constraint MGPS

and MRK4.

7.2 Example 7

This example is a batch reactor model given by the Dow Chemical Company [Caracotsios and Stewart (1985), Biegler and Damiano (1986), Becerra, Roberts and Griffiths (2001)]:

$$\dot{u}_1 + k_2 u_2 w_2 = 0, \tag{108}$$

$$\dot{u}_2 + k_1 u_2 u_6 - k_{-1} w_4 + k_2 u_2 w_2 = 0, \tag{109}$$

$$\dot{u}_3 - k_2 u_2 w_2 - k_3 u_4 u_6 + k_{-3} w_3 = 0, \tag{110}$$

$$\dot{u}_4 + k_1 u_4 u_6 - k_{-3} w_3 = 0, \tag{111}$$

$$\dot{u}_5 - k_1 u_2 u_6 + k_{-1} w_4 = 0, \tag{112}$$

$$\dot{u}_6 + k_1 u_2 u_6 + k_3 u_4 u_6 - k_{-1} w_4 - k_{-3} w_3 = 0, \tag{113}$$

$$u_6 - w_1 + w_2 + w_3 + w_4 - a = 0, \tag{114}$$

$$w_2 - \frac{K_2 u_1}{K_2 + w_1} = 0, \tag{115}$$

$$w_3 - \frac{K_3 u_3}{K_3 + w_1} = 0, \tag{116}$$

$$w_4 - \frac{K_1 u_5}{K_1 + w_1} = 0, \tag{117}$$

where

$$k_1 = 21.893,\ k_{-1} = 2.14 \times 10^9,\ k_2 = 32.318,$$
$$k_3 = 21.893,\ k_{-3} = 1.07 \times 10^9,$$

$$K_1 = 7.65 \times 10^{-18},\ K_2 = 4.03 \times 10^{-11},$$
$$K_3 = 5.32 \times 10^{-18},\ a = 0.0131.$$

The initial conditions are

$$\mathbf{u}(0) = (1.5776, 8.32, 0.01, 0, 0.01, a),$$
$$\mathbf{w}(0) = (0.79735161 \times 10^{-5}, 0.79735161 \times 10^{-5}, 0, 0).$$

Taking the time derivatives of the last three constraints in Eqs. (115)-(117) we obtain

$$\dot{w}_2 = \frac{K_2 \dot{u}_1}{K_2 + w_1} - \frac{K_2 u_1 \dot{w}_1}{(K_2 + w_1)^2}, \tag{118}$$

$$\dot{w}_3 = \frac{K_3 \dot{u}_3}{K_3 + w_1} - \frac{K_3 u_3 \dot{w}_1}{(K_3 + w_1)^2}, \tag{119}$$

$$\dot{w}_4 = \frac{K_1 \dot{u}_5}{K_1 + w_1} - \frac{K_1 u_5 \dot{w}_1}{(K_1 + w_1)^2}, \tag{120}$$

which together with the time derivative of the first constraint in Eq. (114) leads to

$$\dot{w}_1 = \left[1 + \frac{K_2 u_1}{(K_2 + w_1)^2} + \frac{K_3 u_3}{(K_3 + w_1)^2} + \frac{K_1 u_5}{(K_1 + w_1)^2}\right]^{-1} \left[\dot{u}_6 + \frac{K_2 \dot{u}_1}{K_2 + w_1} + \frac{K_3 \dot{u}_3}{K_3 + w_1} + \frac{K_1 \dot{u}_5}{K_1 + w_1}\right]. \tag{121}$$

Eqs. (108)-(113), (121), (118)-(120) constitute a ten-dimensional differential equations system with four constraints in Eqs. (114)-(117).

For this example we divide the ten independent variables into four independent sets: $\{u_2, u_4, u_6, w_1, w_2, w_3, w_4\}$, $\{u_1\}$, $\{u_3\}$ and $\{u_5\}$, and solve the constraints equations to obtain

$$X_1^0 = \frac{a}{n_3 - n_4 + n_5 + n_6 + n_7}, \tag{122}$$

$$X_2^0 = \frac{w_2(K_2 + w_1)}{K_2 n_8}, \tag{123}$$

$$X_3^0 = \frac{w_3(K_3 + w_1)}{K_3 n_9}, \tag{124}$$

$$X_4^0 = \frac{w_4(K_1 + w_1)}{K_1 n_{10}}, \tag{125}$$

where $n_1 = u_2/X_1^0$, $n_2 = u_4/X_1^0$, $n_3 = u_6/X_1^0$, $n_4 = w_1/X_1^0$, $n_5 = w_2/X_1^0$, $n_6 = w_3/X_1^0$, $n_7 = w_4/X_1^0$, $n_8 = u_1/X_2^0$, $n_9 = u_3/X_3^0$ and $n_{10} = u_5/X_4^0$.

In Fig. 12, the results calculated by using the new scheme were shown by the errors of the above four constraints with a time stepsize of 0.001 sec. It can be seen that the multiple-constraint MRK4 provides very accurate results of the four constraints.

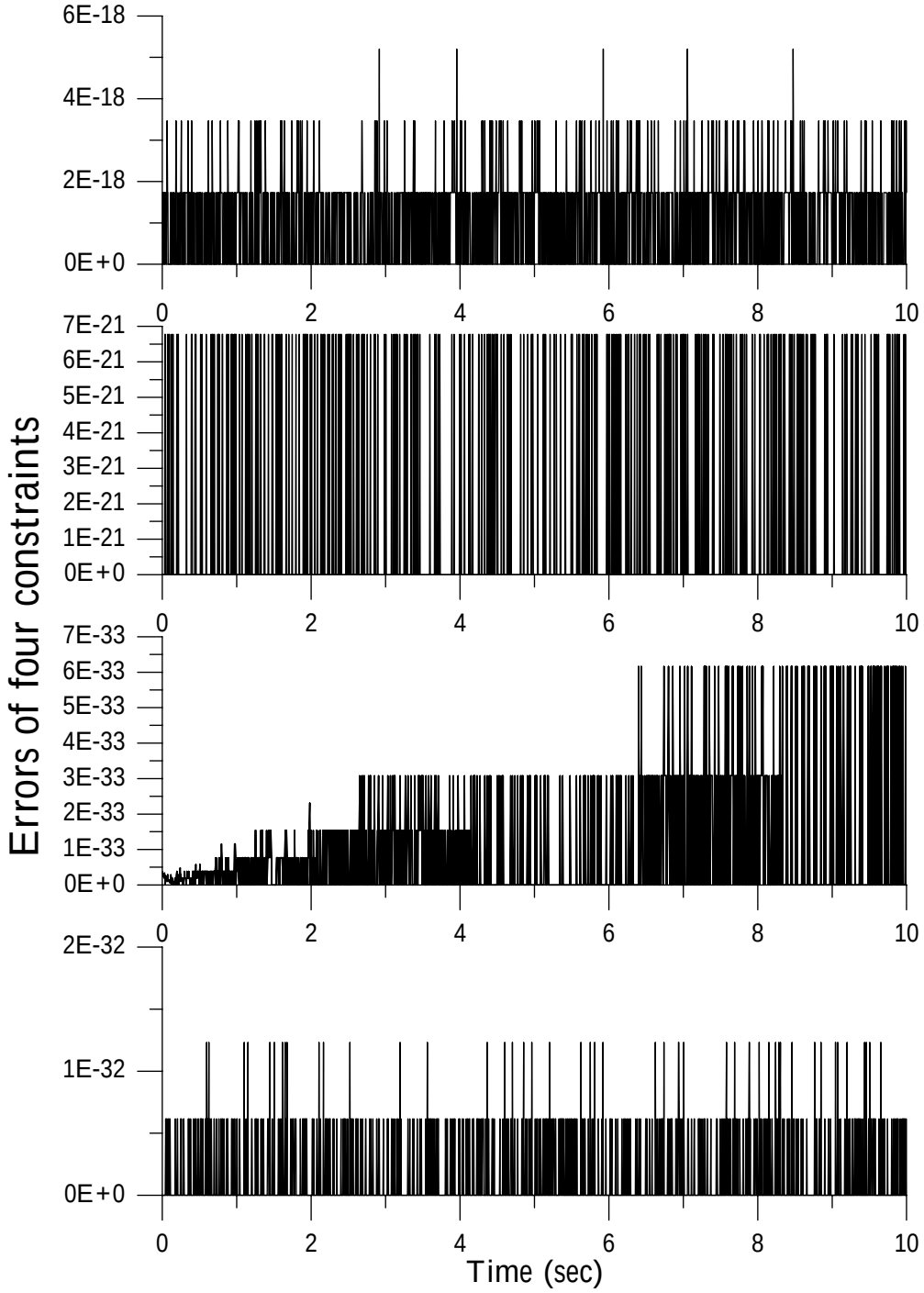

Figure 12 : Displaying the errors of invariants for Example 7 calculated by the multi-constraint MRK4 method.

7.3 Example 8

In many applications the Kepler potential has to be modified in various ways. Kirchgraber (1988) has proposed the following Hamiltonian for the modified Kepler problem:

$$H = \frac{p_1^2 + p_2^2}{2} - \frac{1}{r} - \frac{\varepsilon}{2r^3}, \tag{126}$$

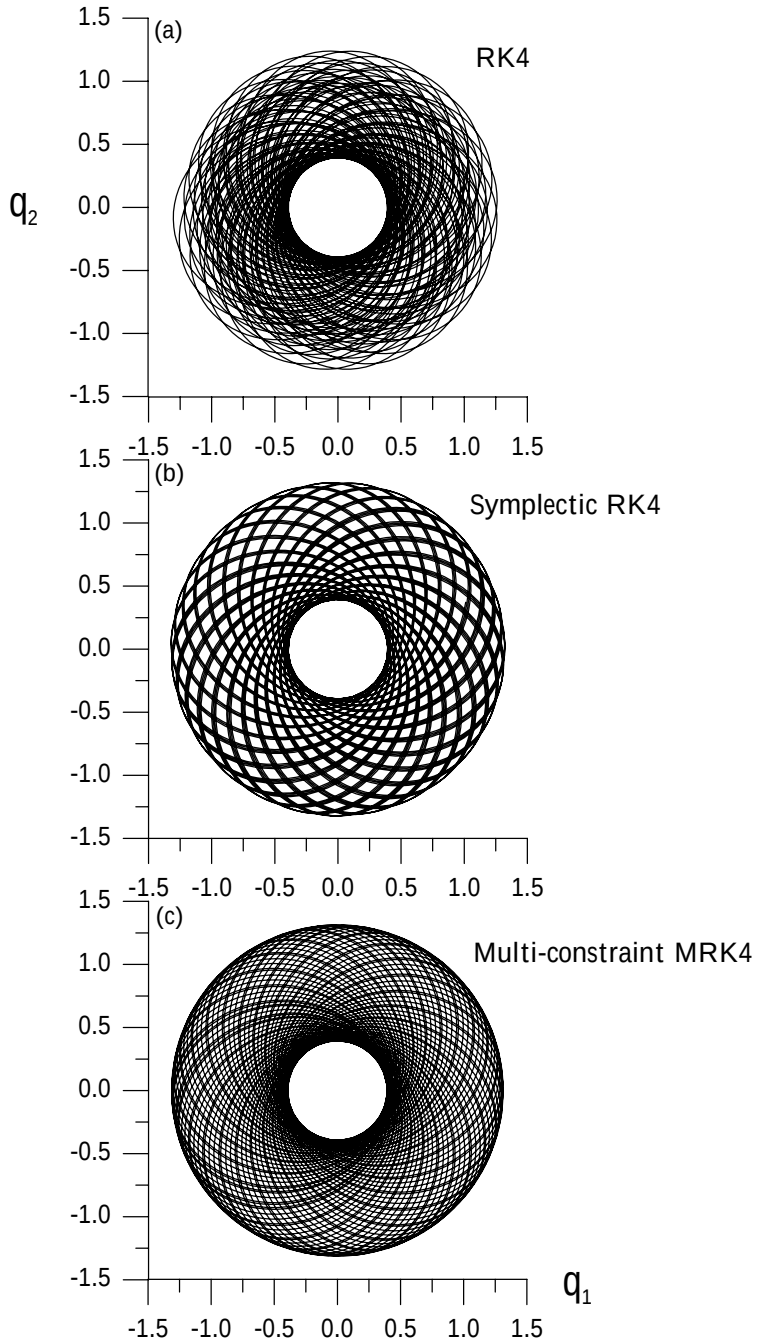

Figure 13 : Displaying the tori for Example 8 of the modified Kepler problem calculated by the RK4, symplectic RK4 and multi-constraint MRK4 methods.

where ε is a small perturbation parameter, and $r = \sqrt{q_1^2+q_2^2}$.

The modified Kepler problem is a standard test problem for numerical methods that are constraints preservation [e.g., Ascher (1997) and Ascher, Chin and Reich (1994)]. According to the above Hamiltonian the differential system is:

$$\dot{q}_1 = p_1, \tag{127}$$

$$\dot{q}_2 = p_2, \tag{128}$$

$$\dot{p}_1 = -\left(\frac{q_1}{r^3} + \frac{3\varepsilon q_1}{2r^5}\right), \tag{129}$$

$$\dot{p}_2 = -\left(\frac{q_2}{r^3} + \frac{3\varepsilon q_2}{2r^5}\right). \tag{130}$$

For this problem there are two invariants:

$$\begin{aligned}\rho_1 = &\left[\frac{p_1^2(t)+p_2^2(t)}{2} - \frac{1}{r(t)} - \frac{\varepsilon}{2r^3(t)}\right] \\ &- \left[\frac{p_1^2(0)+p_2^2(0)}{2} - \frac{1}{r(0)} - \frac{\varepsilon}{2r^3(0)}\right] = 0,\end{aligned} \tag{131}$$

$$\begin{aligned}\rho_2 &= M(t) - M(0) \\ &= [q_1(t)p_2(t) - q_2(t)p_1(t)] \\ &\quad - [q_1(0)p_2(0) - q_2(0)p_1(0)] \\ &= 0.\end{aligned} \tag{132}$$

In addition H the angular momentum M is also an invariant. In the following numerical test the initial values are

$$q_1(0) = 1-c,\ \ q_2(0) = 0,\ \ p_1(0) = 0,\ \ p_2(0) = \sqrt{\frac{1+c}{1-c}}.$$

Since there are two constraints we divide the four independent variables into two independent sets: $\{q_1, q_2\}$ and $\{p_1, p_2\}$, and solve the constraints equations to obtain

$$\begin{aligned}&\frac{n_3^2+n_4^2}{2}(X_2^0)^2 - \frac{1}{X_1^0\sqrt{n_1^2+n_2^2}} \\ &\quad - \frac{\varepsilon}{2(X_1^0)^3\sqrt[3]{n_1^2+n_2^2}} - H(0) = 0,\end{aligned} \tag{133}$$

$$X_1^0 X_2^0 (n_1 n_4 - n_2 n_3) - M(0) = 0, \tag{134}$$

where $n_1 = q_1/X_1^0$, $n_2 = q_2/X_1^0$, $n_3 = p_1/X_2^0$ and $n_4 = p_2/X_2^0$. Substituting Eq. (134) for $1/X_1^0$ into Eq. (133) we obtain a third degree algebraic equation for X_2^0, which can be solved exactly. Then X_1^0 can be solved from Eq. (134).

To compare the numerical results in Table 2, we add the results obtained by Ascher (1997), Ascher, Chin and Reich (1994) and Ascher and Petzold (1991), which include midpoint scheme, second-order RK2, post-stabilization midpoint scheme, post-stabilization RK2, projected midpoint scheme, as well as a symplectic RK4 method [Sanz-Serna and Calvo (1994)]. This scheme is designed to preserve the symplectic structure of Hamiltonian systems. Our method is denoted by multi-constraint MRK4 (mul-cons MRK4). All runs were with uniform time stepsizes of $\Delta t = 0.01\pi$ and 0.001π, $\varepsilon = 0$ and $c = 0.6$. Because of $\varepsilon = 0$ the theoretical solution has a period 2π, and the errors in the numerical solutions can be measured at the integer multiples of 2π. From Table 2 it can be seen that both the projected midpoint scheme and the multi-constraint MRK4 retain the first constraint, and that both the symplectic RK4 scheme and the multi-constraint MRK4 retain the second constraint.

Table 2 demonstrates that a drastic improvement can be obtained when using the multi-constraint MRK4. It is accurate in the periodic solutions with three orders than the projected midpoint scheme when $\Delta t = 0.01\pi$, and four orders when $\Delta t = 0.001\pi$, and is also accurate than RK4 with two orders when $\Delta t = 0.01\pi$, and one order when $\Delta t = 0.001\pi$.

In Fig. 13 we compare the computed results of the modified Kepler problem with $\varepsilon = 0.01$ and $c = 0.6$ in the phase plane of (q_1, q_2) within the interval of $0 < t <$ 500 seconds. The discretizations to be compared all use the time step $\Delta t = 0.1$ sec. The RK4 gave a noisy data and the features of the torus are somewhat blurred as shown in Fig. 13(a). The errors in the Hamiltonian and momentum are about 0.15 and 0.29×10^{-1}. The symplectic RK4 led to a typically symplectic behavior as shown in Fig. 13(b) but is not accurate enough with the error in the Hamiltonian about 0.28×10^{-2} and the radius of the torus about 1.32 (the exact value is about 1.313). The multi-constraint MRK4 also led to a rather well-structured torus as shown in Fig. 13(c) and is very accurate, of which the errors in the Hamiltonian and momentum are both zero, and the radius with 1.311 of the torus is much better approximated to 1.313. Through the discussions above, the superiority of the multi-constraint MRK4 method is clear.

8 Conclusions

This paper has presented a new form of the n-dimensional nonlinear dynamical system endowed with k constraints, by converting it into an augmented $n+k$-dimensional differential equation system in the Minkowski space. In this space, k cone conditions are

Table 2 : Errors of Example 8 for Kepler problem by using different numerical schemes.

method	Δt	$\lvert q_2(2\pi)\rvert$	$\lvert q_2(4\pi)\rvert$	$\lvert q_2(20\pi)\rvert$	$\lvert q_2(50\pi)\rvert$	$\lvert\rho_1\rvert_\infty$	$\lvert\rho_2\rvert_\infty$
midpt	0.01π	0.16	0.30	0.72	0.10	0.42×10^{-2}	$\times$
RK2	0.01π	0.12	0.18	0.67	0.52	0.36×10^{-1}	$\times$
post-stab-midpt	0.01π	0.54×10^{-2}	0.11×10^{-1}	0.54×10^{-1}	0.13	0.81×10^{-7}	$\times$
post-stab-RK2	0.01π	0.40×10^{-2}	0.81×10^{-2}	0.40×10^{-1}	0.10	0.15×10^{-6}	$\times$
proj-midpt	0.01π	0.14×10^{-2}	0.28×10^{-2}	0.14×10^{-1}	0.34×10^{-1}	0	$\times$
RK4	0.01π	0.18×10^{-3}	0.49×10^{-3}	0.74×10^{-2}	0.42×10^{-1}	0.84×10^{-4}	0.15×10^{-4}
symp-RK4	0.01π	0.76×10^{-3}	0.15×10^{-2}	0.76×10^{-2}	0.19×10^{-1}	0.20×10^{-4}	0
mul-cons MRK4	0.01π	0.16×10^{-5}	0.33×10^{-5}	0.16×10^{-4}	0.41×10^{-4}	0	0
midpt	0.001π	0.16×10^{-2}	0.32×10^{-2}	0.16×10^{-1}	0.40×10^{-1}	0.42×10^{-4}	$\times$
RK2	0.001π	0.15×10^{-2}	0.29×10^{-2}	0.12×10^{-1}	0.20×10^{-1}	0.41×10^{-4}	$\times$
post-stab-midpt	0.001π	0.54×10^{-4}	0.11×10^{-3}	0.54×10^{-3}	0.14×10^{-2}	0.83×10^{-13}	$\times$
post-stab-RK2	0.001π	0.40×10^{-4}	0.81×10^{-4}	0.40×10^{-3}	0.10×10^{-2}	0.86×10^{-13}	$\times$
proj-midpt	0.001π	0.14×10^{-4}	0.29×10^{-4}	0.14×10^{-3}	0.36×10^{-3}	0	$\times$
RK4	0.001π	0.12×10^{-7}	0.26×10^{-7}	0.18×10^{-6}	0.69×10^{-6}	0.11×10^{-8}	0.15×10^{-9}
symp-RK4	0.001π	0.78×10^{-7}	0.16×10^{-6}	0.78×10^{-6}	0.16×10^{-5}	0.20×10^{-8}	0
mul-cons MRK4	0.001π	0.22×10^{-8}	0.45×10^{-8}	0.22×10^{-7}	0.56×10^{-7}	0	0

deduced for the augmented states. For the resulting systems $\dot{\mathbf{X}}_i = \mathbf{A}_i\mathbf{X}_i$, $i = 1,\ldots,k$, $\mathbf{A}_i \in so(n_i,1)$ is a local Lie algebra of the proper orthochronous Lorentz group $SO_o(n_i,1)$, and thus the internal symmetry of the nonlinear dynamical system with k constraints was revealed.

According to the inherent symmetry of the augmented system, we have applied the exponential mapping technique to integrate the augmented system and use the extra k freedoms of the system to adjust the k integrating factors such that the k constraints are satisfied. A similar procedure was also applied to the case when one integrates these augmented subsystems by the fourth-order Runge-Kutta method. Since all constraints were included in the newly developed integrating schemes it is guaranteed that all algebraic equations that describe the manifold of solutions are satisfied up to an accuracy that is used to integrate these dynamical equations and hence a drift from the solution manifold is avoided. Several numerical examples were investigated to confirm that the new numerical methods are effective to integrate the constrained dynamical systems and to preserve the constraints.

Since the new schemes are easy to implement numerically and have high computational efficiency and accuracy, they may be used for physical and engineering applications which particularly emphasized the constraints.

References

Arevalo, C.; Campbell, S. L.; Selva, M. (2004): Unitary partitioning in general constraint preserving DAE integrators. *Math. Comp. Model.*, vol. 40, pp. 1273-1284.

Ascher, U. M. (1997): Stabilization of invariants of discretized differential systems. *Num. Algor.*, vol. 14, pp. 1-24.

Ascher, U. M.; Chin, H.; Reich, S. (1994): Stabilization of DAEs and invariant manifolds. *Num. Math.*, vol. 67, pp. 131-149.

Ascher, U. M.; Petzold, L. (1991): Projected implicit Runge-Kutta methods for differential-algebraic equations. *SIAM J. Num. Anal.*, vol. 28, pp. 1097-1120.

Baumgarte, J. (1972): Stabilization of constraints and integrals of motion in dynamical systems. *Comp. Math. Appl. Mech. Eng.*, vol. 1, pp. 1-16.

Becerra, V. M.; Roberts, P. D.; Griffiths, G. W. (2001): Applying the extended Kalman filter to systems described by nonlinear differential-algebraic equations. *Cont. Eng. Pract.*, vol. 9, pp. 267-281.

Biegler, L. T.; Damiano, J. J. (1986): Nonlinear parameter estimation: A case study. *AIChE Journal*, vol. 32,

pp. 29-45.

Campbell, S. L.; Moore, E. (1995): Constraint preserving integrators for general nonlinear higher index DAEs. *Num. Math.*, vol. 69, pp. 383-399.

Caracotsios, M.; Stewart, E. (1985): Sensitivity analysis of initial value problems with mixed ODEs and algebraic equations. *Comp. Chem. Eng.*, vol. 9, pp. 359-365.

Chan, R. P. K.; Chartier, P.; Murua, A. (2002): Post-projected Runge-Kutta methods for index-2 differential-algebraic equations. *Appl. Num. Math.*, vol. 42, pp. 77-94.

Führer, C.; Leimkuhler, B. J. (1991): Numerical solution of differential-algebraic equations for constrained mechanical motion. *Num. Math.*, vol. 59, pp. 55-69.

Iserles, A.; Zanna, A. (2000): Preserving algebraic invariants with Runge-Kutta methods. *J. Comp. Appl. Math.*, vol. 125, pp. 69-81.

Kirchgraber, U. (1988): An ODE-solver based on the method of averaging. *Num. Math.*, vol. 53, pp. 621-652.

Liu, C.-S. (2001): Cone of non-linear dynamical system and group preserving schemes. *Int. J. Non-Linear Mech.*, vol. 36, pp. 1047-1068.

Liu, C.-S. (2002): Nonlinear Lorentzian and Hamiltonian formulations and their relations. *Int. J. Appl. Math.*, vol. 10, pp. 59-97.

Liu, C.-S.; Chang, C.-W. (2004): Lie group symmetry applied to the computation of convex plasticity constitutive equation. *CMES: Computer Modeling in Engineering & Sciences*, vol. 6, pp. 277-294.

Liu, C.-S.; Hong, H.-K.; Liou, D.-Y. (2003): Two-dimensional friction oscillator: group-preserving scheme and handy formulae. *J. Sound Vib.*, vol. 266, pp. 49-74.

Maerz, R.; Tischendorf, C. (1994): Solving more general index-2 differential equations. *Comp. Math. Appl.*, vol. 28, pp. 77-105.

Marsden, J. E.; Ratiu, T. S. (1994): *Introduction to Mechanics and Symmetry*. Springer-Verlag, New York.

März, R. (1991): Numerical methods for differential algebraic equations. *Acta Num.*, pp. 141-198.

März, R. (2002): Differential algebraic systems anew. *Appl. Num. Math.*, vol. 42, pp. 315-335.

Rheinboldt, W. C. (1997): Solving algebraically explicit DAEs with the MANPAK-manifold-algorithms. *Comp. Math. Appl.*, vol. 33, pp. 31-43.

Sand, J. (2002): On implicit Euler for high-order high-index DAEs. *Appl. Num. Math.*, vol. 42, pp. 411-424.

Sanz-Serna, J. M. (1994): An unconventional symplectic integrator of W. Kahan. *Appl. Num. Math.*, vol. 16, pp. 245-250.

Sanz-Serna, J. M.; Calvo, M. P. (1994): *Numerical Hamiltonian Problems*. Chapman & Hall, New York.