

Feedback Integrators

Dong Eui Chang^{*} Fernando Jiménez[†] Matthew Perlmutter[‡]

To appear in *Journal of Nonlinear Science*. <http://dx.doi.org/10.1007/s00332-016-9316-7>

Abstract

A new method is proposed to numerically integrate a dynamical system on a manifold such that the trajectory stably remains on the manifold and preserves first integrals of the system. The idea is that given an initial point in the manifold we extend the dynamics from the manifold to its ambient Euclidean space and then modify the dynamics outside the intersection of the manifold and the level sets of the first integrals containing the initial point such that the intersection becomes a unique local attractor of the resultant dynamics. While the modified dynamics theoretically produces the same trajectory as the original dynamics, it yields a numerical trajectory that stably remains on the manifold and preserves the first integrals. The big merit of our method is that the modified dynamics can be integrated with any ordinary numerical integrator such as Euler or Runge-Kutta. We illustrate this method by applying it to three famous problems: the free rigid body, the Kepler problem and a perturbed Kepler problem with rotational symmetry. We also carry out simulation studies to demonstrate the excellence of our method and make comparisons with the standard projection method, a splitting method and Störmer-Verlet schemes.

Contents

1	Introduction	2
2	Theory	2
3	Applications	6
3.1	The Free Rigid Body	6
3.2	The Kepler Problem	9
3.3	A Perturbed Kepler Problem with Rotational Symmetry	12
4	Simulations	15
4.1	The Free Rigid Body	15
4.2	The Kepler Problem	16
4.3	A Perturbed Kepler Problem with Rotational Symmetry	16
5	Conclusions and Future Work	17

^{*}Corresponding author. Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. dechang@uwaterloo.ca

[†]Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada. fjimenez@uwaterloo.ca

[‡]Departamento de Matemática, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. matthew@mat.ufmg.br

1 Introduction

Given a dynamical system on a manifold with first integrals, it is important for a numerical integrator to preserve the manifold structure and the first integrals of the equations of motion. This has been the focus of much effort in the development of numerical integration schemes [2]. In this paper we do not propose any specific numerical integration scheme, but rather propose a new paradigm of integration that can faithfully preserve conserved quantities with existing numerical integration schemes.

The main idea in our paradigm is as follows. Consider a dynamical system on a manifold M with first integrals $f_i : M \rightarrow \mathbb{R}$, $i = 1, \dots, \ell$. Assume that we can embed the manifold M into Euclidean space \mathbb{R}^n and extend the first integrals to a neighborhood U of M in \mathbb{R}^n . For an arbitrary point $x_0 \in M$, consider the set

$$\Lambda = \{x \in U \mid x \in M, f_i(x) = f_i(x_0), i = 1, \dots, \ell\}$$

which is the intersection of M with all the level sets of the first integrals containing the point x_0 , and is an invariant set of the dynamical system. We then extend the dynamical system from M to U and then modify the dynamics outside of Λ such that the set Λ becomes a unique local attractor of the extended, modified system. Since the dynamics have not changed on Λ by the extension and modification to U , both the original system on M and the extended, modified system on U produce the same trajectory for the initial point $x_0 \in \Lambda$. Numerically, however, integrating the extended system has the following advantage: if the trajectory deviates from Λ at some numerical integration step, then it will get pushed back toward the attractor Λ in the extended, modified dynamics, thus remaining on the manifold M and preserving all the first integrals. It can be rigorously shown that the discrete-time dynamical system derived from any one-step numerical integrator with uniform step size h for the extended, modified continuous-time system indeed has an attractor Λ_h that contains the set Λ in its interior and converges to Λ as $h \rightarrow 0+$. In this paper we shall use the word, *preserve*, in this sense. It is noteworthy that the numerical integration of the extended dynamics can be carried out with any ordinary integrator and is done in one global Cartesian coordinate system on \mathbb{R}^n . We find conditions for applicability of this method and implement the result on the following three examples: the free rigid body dynamics, the Kepler problem, and a perturbed Kepler problem with rotational symmetry. We also carry out simulation studies to show the excellence of our new paradigm of integration for numerical preservation of conserved quantities in comparison with other well-known integration schemes, such as projection and splitting methods and symplectic Störmer-Verlet integrators.

2 Theory

Consider a dynamical system on an open subset U of \mathbb{R}^n :

$$\dot{x} = X(x), \tag{1}$$

where X is a C^1 vector field on U . Let us make the following assumptions:

A1. There is a C^2 function $V : U \rightarrow \mathbb{R}$ such that $V(x) \geq 0$ for all $x \in U$, $V^{-1}(0) \neq \emptyset$, and

$$\nabla V(x) \cdot X(x) = 0 \tag{2}$$

for all $x \in U$.

A2. There is a positive number c such that $V^{-1}([0, c])$ is a compact subset of U .

A3. The set of all critical points of V in $V^{-1}([0, c])$ is equal to $V^{-1}(0)$.

Adding the negative gradient of V to (1), let us consider the following dynamical system on U :

$$\dot{x} = X(x) - \nabla V(x). \quad (3)$$

Since 0 is the minimum value of V , $\nabla V(x) = 0$ for all $x \in V^{-1}(0)$. Hence, the two vector fields X and $X - \nabla V$ coincide on $V^{-1}(0)$.

Theorem 2.1. *Under assumptions A1 – A3, every trajectory of (3) starting from a point in $V^{-1}([0, c])$ stays in $V^{-1}([0, c])$ for all $t \geq 0$ and asymptotically converges to the set $V^{-1}(0)$ as $t \rightarrow \infty$. Furthermore, $V^{-1}(0)$ is an invariant set of both (1) and (3).*

Proof. Let $x(t)$ be a trajectory of (3) starting from a point in $V^{-1}([0, c])$. By A1

$$\frac{d}{dt}V(x(t)) = \nabla V(x(t)) \cdot (X(x(t)) - \nabla V(x(t))) = -|\nabla V(x(t))|^2 \leq 0 \quad (4)$$

for all t . Hence, $V^{-1}([0, c])$ is a positively invariant set of (3). From (4) and A3, it follows that $\{x \in V^{-1}([0, c]) \mid \dot{V}(x) = 0\} = \{x \in V^{-1}([0, c]) \mid \nabla V(x) = 0\} = V^{-1}(0)$. Hence, by LaSalle's invariance principle [5], $x(t)$ converges asymptotically to $V^{-1}(0)$ as $t \rightarrow \infty$, where A2 is used for compactness of $V^{-1}([0, c])$. The invariance of $V^{-1}(0)$ follows from (2) and the coincidence of (1) and (3) on $V^{-1}(0)$. \square

Let us find a higher-order condition than that in assumption A3 so that A3 can be relaxed. For the function V and the vector field X in the statement of assumption A1, which are now both assumed to be of C^∞ , let

$$S = \left\{ x \in U \mid X^k \frac{\partial V}{\partial x^i} = 0 \ \forall k \geq 0, 1 \leq i \leq n \right\}, \quad (5)$$

where $x = (x^1, x^2, \dots, x^n)$, and $X^k \frac{\partial V}{\partial x^i}$ denotes the k -th order directional derivative of $\partial V / \partial x^i$ along X , i.e.,

$$X^0 \frac{\partial V}{\partial x^i} = \frac{\partial V}{\partial x^i}; \quad X \frac{\partial V}{\partial x^i} = X \cdot \nabla \frac{\partial V}{\partial x^i}; \quad X^k \frac{\partial V}{\partial x^i} = X \left(X^{k-1} \frac{\partial V}{\partial x^i} \right), \quad k \geq 2.$$

Consider the following assumption in place of A3:

$$A3'. \quad S \cap V^{-1}([0, c]) \subset V^{-1}(0).$$

The following theorem generalizes Theorem 2.1:

Theorem 2.2. *Under assumptions A1, A2 and A3', every trajectory of (3) starting in $V^{-1}([0, c])$ stays in $V^{-1}([0, c])$ for all $t \geq 0$ and asymptotically converges to the set $V^{-1}(0)$ as $t \rightarrow \infty$. Furthermore, $V^{-1}(0)$ is an invariant set of both (1) and (3).*

Proof. Consider the dynamics (3). It is easy to show that $V^{-1}([0, c])$ is a positively invariant set of the dynamics. Let \mathcal{M} be the largest invariant set in $\mathcal{E} = \{x \in U \mid \dot{V}(x) = 0\} \cap V^{-1}([0, c])$. Let $x(t)$ be an arbitrary trajectory in \mathcal{M} . Since $\mathcal{E} = \{x \in U \mid \nabla V(x) = 0\} \cap V^{-1}([0, c])$ as shown in the proof of Theorem 2.1, the trajectory $x(t)$ satisfies $\nabla V = 0$, i.e.,

$$\frac{\partial V}{\partial x^i}(x(t)) = 0 \quad (6)$$

for all $t \in \mathbb{R}$ and $1 \leq i \leq n$. Since $\nabla V = 0$ along $x(t)$, the trajectory $x(t)$ satisfies

$$\dot{x}(t) = X(x(t)) \quad (7)$$

for all $t \in \mathbb{R}$. By differentiating (6) repeatedly in t and using (7) on each differentiation, we can show that the trajectory $x(t)$ satisfies

$$X^k \frac{\partial V}{\partial x^i} = 0$$

for all $t \in \mathbb{R}$, $k \geq 0$ and $1 \leq i \leq n$. Thus, the entire trajectory $x(t)$ is contained in the set S defined in (5), implying $\mathcal{M} \subset S$, from which and A3' it follows $\mathcal{M} \subset V^{-1}(0)$. Hence, by LaSalle's invariance principle, every trajectory starting in $V^{-1}([0, c])$ asymptotically converges to \mathcal{M} and thus to $V^{-1}(0)$ as $t \rightarrow \infty$.

The invariance of $V^{-1}(0)$ follows from (2) and the coincidence of (1) and (3) on $V^{-1}(0)$. \square

Remark 2.3. 1. If condition (2) is replaced by $\nabla V(x) \cdot X(x) \leq 0$ in assumption A1, then Theorems 2.1 and 2.2 still hold provided that the invariance of $V^{-1}(0)$ is replaced by positive invariance in the statement of the theorems.

2. Theorems 2.1 and 2.2 still hold with the use of the following modified dynamics

$$\dot{x} = X(x) - A(x)\nabla V(x)$$

instead of (3), where $A(x)$ is an $n \times n$ matrix-valued function with its symmetric part $(A(x) + A^T(x))$ positive definite at each $x \in \mathbb{R}^n$.

3. From the control viewpoint, the added term $-\nabla V(x)$ in (3) can be regarded as a negative feedback control $u(x) = -\nabla V(x)$ to asymptotically stabilize the set $V^{-1}(0)$ for the control system $\dot{x} = X(x) + u$ with control u .

Suppose that assumptions A1, A2 and A3 (or A3' instead of A3) hold and that we want to integrate the dynamics (1) for an initial point $x(0) \in V^{-1}(0)$. Since $V^{-1}(0)$ is positively invariant, the trajectory must remain in $V^{-1}(0)$ for all $t \geq 0$. Recall that the two dynamics (1) and (3) coincide on $V^{-1}(0)$, so we can integrate (3) instead of (1) for the initial condition. Though there is no theoretical difference between the two integrations, integrating (3) has a numerical advantage over integrating (1). Suppose that the trajectory numerically deviates from the positively invariant set $V^{-1}(0)$ during integration. Then the dynamics (3) will push the trajectory back toward $V^{-1}(0)$ since $V^{-1}(0)$ is the attractor of (3) in $V^{-1}([0, c])$ whereas the dynamics (1) will leave the trajectory outside of $V^{-1}(0)$ which would not happen in the exact solution. It is noteworthy that this integration strategy is independent of the choice of integration schemes. In the Appendix we show that any one-step numerical integrator, as a discrete-time dynamical system, with uniform step size h for (3) has an attractor Λ_h that contains $V^{-1}(0)$ in its interior and converges to $V^{-1}(0)$ as $h \rightarrow 0+$.

Let us now apply this integration strategy to numerically integrate dynamics on a manifold while preserving its first integrals and the domain manifold. Consider a manifold M and dynamics

$$\dot{x} = X(x) \tag{8}$$

on M that have ℓ first integrals $f_i : M \rightarrow \mathbb{R}$, $i = 1, \dots, \ell$. Suppose that M is an embedded manifold in \mathbb{R}^n as a level set of a function $f_0 : \mathbb{R}^n \rightarrow \mathbb{R}^r$ for some r , and that both the dynamics (8) and the functions f_i , $i = 0, \dots, \ell$ extend to an open neighborhood U of M in \mathbb{R}^n . Our goal is to numerically integrate (8) with an initial condition $x(0) = x_0 \in M$ while preserving the manifold M and the first integrals. Let

$$f = (f_0, f_1, \dots, f_\ell) : \mathbb{R}^n \rightarrow \mathbb{R}^{r+\ell} \tag{9}$$

and define a function $V : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$ by

$$V(x) = \frac{1}{2}(f(x) - f(x_0))^T K(f(x) - f(x_0)), \tag{10}$$

where K is an $(r + \ell) \times (r + \ell)$ constant symmetric positive definite matrix. Notice that

$$V^{-1}(0) = \{x \in U \mid x \in M, f_i(x) = f_i(x_0), i = 1, \dots, \ell\},$$

and that $V^{-1}(0)$ is invariant under the flow of (8). Or, more generally we can define a function $V(x)$ as $V(x) = W(f_0(x), f_1(x), \dots, f_\ell(x))$ where $W : \mathbb{R}^{r+\ell} \rightarrow \mathbb{R}$ is a non-negative function that takes the value of 0 only at $(f_0(x_0), f_1(x_0), \dots, f_\ell(x_0))$. If the function V satisfies assumptions A1, A2 and A3 (or A3' instead of A3), then by Theorem 2.1 (or Theorem 2.2), $V^{-1}(0)$ is the local attractor of the modified dynamics

$$\dot{x} = X(x) - \nabla V(x) \tag{11}$$

that coincide with the original dynamics (8) on $V^{-1}(0)$.

The following lemma provides a sufficient condition under which the function V defined in (10) satisfies assumptions A2 and A3:

Lemma 2.4. *Consider the functions f and V defined in (9) and (10). If $V^{-1}(0)$ is compact and the Jacobian matrix $Df(x)$ of f has rank $(r + \ell)$ for all $x \in V^{-1}(0)$, then there is a number $c > 0$ such that assumptions A2 and A3 hold.*

Proof. By compactness of $V^{-1}(0)$ and the regularity of Df , there is a bounded open set X such that $V^{-1}(0) \subset X \subset \text{cl}(X) \subset U$, and $Df(x)$ has rank $r + \ell$ for all $x \in X$, where $\text{cl}(X)$ denotes the closure of X . Consider now the gradient of V . An easy calculation shows that,

$$\nabla V(x) = Df(x)^T K(f(x) - f(x_0)).$$

Now, since for all $x \in X$, $Df(x)$ is onto as a linear map, $Df(x)^T$ is therefore one to one. It follows that, for $x \in X$,

$$\nabla V(x) = 0 \iff f(x) - f(x_0) = 0 \iff x \in V^{-1}(0). \tag{12}$$

In other words, the set of all critical points of V in X is equal to $V^{-1}(0)$. Since the boundary ∂X of X , being closed and bounded, is compact and $\partial X \cap V^{-1}(0) = \emptyset$, the minimum value, denoted by d , of V on ∂X is positive. If necessary, restrict the function V to X , replacing its original domain U with X . Then, there is a positive number c less than d such that $V^{-1}([0, c]) \subset X$. Therefore, assumption A3 holds for this number c . Since the closed set $V^{-1}([0, c])$ is contained in the bounded set X , it is compact, which implies that assumption A2 holds. \square

Theorem 2.5. *For the functions f and V defined in (9) and (10), if V satisfies (2) for all $x \in U$, the set $V^{-1}(0)$ is compact and the Jacobian matrix $Df(x)$ is onto for all $x \in V^{-1}(0)$, then there is a number $c > 0$ such that every trajectory starting in $V^{-1}([0, c])$ remains in $V^{-1}([0, c])$ for all $t \geq 0$ and asymptotically converges to $V^{-1}(0)$ as $t \rightarrow \infty$.*

Theorem 2.6. *For the functions f and V defined in (9) and (10), if V satisfies (2) for all $x \in U$, the set $V^{-1}(0)$ is compact and there is an open subset X of U containing $V^{-1}(0)$ such that the Jacobian matrix $Df(x)$ is onto for all $x \in X \setminus V^{-1}(0)$, then there is a number $c > 0$ such that every trajectory starting in $V^{-1}([0, c])$ remains in $V^{-1}([0, c])$ for all $t \geq 0$ and asymptotically converges to $V^{-1}(0)$ as $t \rightarrow \infty$.*

Proof. Modify the proof of Lemma 2.4 appropriately. \square

As discussed above, we can integrate (11) instead of (8) for the initial condition $x(0) = x_0 \in V^{-1}(0)$, which will yield a trajectory that is expected to numerically well remain on the manifold M and preserve the values of the first integrals $f_i, i = 1, \dots, \ell$. It is noteworthy that the integration is carried out in one Cartesian coordinate system on \mathbb{R}^n rather than over local charts on the manifold M which would take additional computational costs for coordinate changes between local charts. In the following section, we will apply this strategy to the free rigid body dynamics, the Kepler problem and a perturbed Kepler problem with rotational symmetry to integrate the dynamics preserving their first integrals and domain manifolds.

3 Applications

3.1 The Free Rigid Body

Consider the free rigid body dynamics:

$$\dot{R} = R \hat{\Omega}, \tag{13a}$$

$$\dot{\Omega} = \mathbb{I}^{-1} ((\mathbb{I}\Omega) \times \Omega), \tag{13b}$$

where $(R, \Omega) \in \text{SO}(3) \times \mathbb{R}^3$; \mathbb{I} is the moment of inertia matrix; and

$$\hat{\Omega} = \begin{bmatrix} 0 & -\Omega_3 & \Omega_2 \\ \Omega_3 & 0 & -\Omega_1 \\ -\Omega_2 & \Omega_1 & 0 \end{bmatrix} \tag{14}$$

for

$$\Omega = \begin{bmatrix} \Omega_1 \\ \Omega_2 \\ \Omega_3 \end{bmatrix}.$$

Since $\text{SO}(3) \subset \mathbb{R}^{3 \times 3}$, from here on we assume that the rigid body dynamics are defined on the Euclidean space $\mathbb{R}^{3 \times 3} \times \mathbb{R}^3$ and that the matrix R denotes a 3×3 matrix, not necessarily in $\text{SO}(3)$. This is the *extension* of the dynamics step.

Define two functions $E : \mathbb{R}^3 \rightarrow \mathbb{R}$ and $\pi : \mathbb{R}^{3 \times 3} \times \mathbb{R}^3 \rightarrow \mathbb{R}^3$ by

$$E(\Omega) = \frac{1}{2} \Omega^T \mathbb{I} \Omega, \tag{15}$$

$$\pi(R, \Omega) = R \mathbb{I} \Omega, \tag{16}$$

where E represents the kinetic energy of the free rigid body and π the spatial angular momentum vector when $R \in \text{SO}(3)$. These quantities are first integrals of (13). Choose any

$$R_0 \in \text{SO}(3), \quad \Omega_0 \in \mathbb{R}^3 \setminus \{(0, 0, 0)\},$$

and let

$$E_0 = E(\Omega_0) > 0, \quad \pi_0 = \pi(R_0, \Omega_0) \in \mathbb{R}^3 \setminus \{(0, 0, 0)\}. \tag{17}$$

Define an open set U by

$$U = \{(R, \Omega) \in \mathbb{R}^{3 \times 3} \times \mathbb{R}^3 \mid \det(R) > 0\}$$

and a function $V : U \subset \mathbb{R}^{3 \times 3} \times \mathbb{R}^3 \rightarrow \mathbb{R}$ by

$$V(R, \Omega) = \frac{k_0}{4} \|R^T R - I\|^2 + \frac{k_1}{2} |E(\Omega) - E_0|^2 + \frac{k_2}{2} |\pi(R, \Omega) - \pi_0|^2 \tag{18}$$

for $(R, \Omega) \in U \subset \mathbb{R}^{3 \times 3} \times \mathbb{R}^3$, where $k_i > 0$, $i = 0, 1, 2$ are constants, and $\|\cdot\|$ is the 2-norm defined by $\|A\| = \sqrt{\text{trace}(A^T A)}$ for a matrix A . Observe that we are endowing the space $\mathbb{R}^{3 \times 3} \times \mathbb{R}^3$ with the standard inner product, and that the trace norm is precisely the norm induced on $\mathbb{R}^{3 \times 3}$ by this inner product. We compute all gradients that follow with respect to this inner product. Notice that

$$V^{-1}(0) = \{(R, \Omega) \in \mathbb{R}^{3 \times 3} \times \mathbb{R}^3 \mid R \in \text{SO}(3), E(\Omega) = E_0, \pi(R, \Omega) = \pi_0\}.$$

Lemma 3.1. *The gradient $(\nabla_R V, \nabla_\Omega V) \in \mathbb{R}^{3 \times 3} \times \mathbb{R}^3$ of the function V (18) is given by*

$$\nabla_R V = k_0 R(R^T R - I) + k_2(\pi(R, \Omega) - \pi_0)\Omega^T \mathbb{I}, \quad (19a)$$

$$\nabla_\Omega V = k_1(E(\Omega) - E_0)\mathbb{I}\Omega + k_2\mathbb{I}R^T(\pi(R, \Omega) - \pi_0). \quad (19b)$$

Proof. Straightforward. \square

The following lemma shows that the function V satisfies assumption A1 stated in §2.

Lemma 3.2. *The function V satisfies*

$$\langle (\nabla_R V, \nabla_\Omega V), (R\hat{\Omega}, \mathbb{I}^{-1}((\mathbb{I}\Omega) \times \Omega)) \rangle = 0. \quad (20)$$

Proof. One can compute

$$\begin{aligned} \langle \nabla_R V, (R\hat{\Omega}) \rangle &= \text{trace}(\hat{\Omega}^T R^T (k_0 R(R^T R - I) + k_2(\pi(R, \Omega) - \pi_0)\Omega^T \mathbb{I})) \\ &= -k_0 \text{trace}(\hat{\Omega} R^T R(R^T R - I)) - k_2 \text{trace}(\hat{\Omega} R^T (\pi(R, \Omega) - \pi_0)\Omega^T \mathbb{I}) \\ &= -k_2 \Omega^T \mathbb{I} \hat{\Omega} R^T (\pi(R, \Omega) - \pi_0), \end{aligned}$$

where, in the third equality we use the fact that for A symmetric and B antisymmetric, $\text{trace}(AB) = 0$.

Next, we compute,

$$\begin{aligned} \langle \nabla_\Omega V, \mathbb{I}^{-1}((\mathbb{I}\Omega) \times \Omega) \rangle &= \langle k_1(E(\Omega) - E_0)\mathbb{I}\Omega + k_2\mathbb{I}R^T(\pi(R, \Omega) - \pi_0), \mathbb{I}^{-1}((\mathbb{I}\Omega) \times \Omega) \rangle \\ &= k_1(E(\Omega) - E_0)\langle (\mathbb{I}\Omega) \times \Omega, \Omega \rangle + k_2\langle R^T(\pi(R, \Omega) - \pi_0), (\mathbb{I}\Omega) \times \Omega \rangle \\ &= k_2\langle \mathbb{I}\Omega, \Omega \times R^T(\pi(R, \Omega) - \pi_0) \rangle \\ &= k_2 \Omega^T \mathbb{I} \hat{\Omega} R^T (\pi(R, \Omega) - \pi_0). \end{aligned}$$

Hence,

$$\langle (\nabla_R V, \nabla_\Omega V), (R\hat{\Omega}, \mathbb{I}^{-1}((\mathbb{I}\Omega) \times \Omega)) \rangle = \langle \nabla_R V, R\hat{\Omega} \rangle + \langle \nabla_\Omega V, \mathbb{I}^{-1}((\mathbb{I}\Omega) \times \Omega) \rangle = 0. \quad \square$$

The following lemma shows that the function V satisfies assumptions A2 and A3 stated in §2.

Lemma 3.3. *There is a number c satisfying*

$$0 < c < \min\{k_0/4, k_1|E_0|/2, k_2|\pi_0|^2/2\} \quad (21)$$

such that $V^{-1}([0, c])$ is a compact subset of U and the set of all critical points of V in $V^{-1}([0, c])$ is equal to $V^{-1}(0)$.

Proof. It is obvious that there is a number c satisfying (21) such that $V^{-1}([0, c])$ becomes a compact set in U . For such a number c , the matrix R is invertible for every $(R, \Omega) \in V^{-1}([0, c])$. Since 0 is the minimum value of V , every point in $V^{-1}(0)$ is a critical point of V .

Let (R, Ω) be a critical point of V in $V^{-1}([0, c]) \setminus V^{-1}(0)$. By Lemma 3.1 it satisfies

$$k_0 R(R^T R - I) + k_2(\pi - \pi_0)\Omega^T \mathbb{I} = 0, \quad (22a)$$

$$k_1(E - E_0)\mathbb{I}\Omega + k_2\mathbb{I}R^T(\pi - \pi_0) = 0, \quad (22b)$$

where

$$\pi = \pi(R, \Omega), \quad E = E(R, \Omega).$$

Post-multiplying (22a) by R^T and pre-multiplying (22b) by R yield

$$k_0 R(R^T R - I)R^T + k_2(\pi - \pi_0)\pi^T = 0, \quad (23a)$$

$$k_1(E - E_0)\pi + k_2 R\mathbb{I}R^T(\pi - \pi_0) = 0, \quad (23b)$$

since $\pi = R\mathbb{I}\Omega$. Notice that $\Omega = 0$ would imply $V(R, \Omega) \geq \frac{k_2}{2}|\pi_0|^2 > c$, contradicting $(R, \Omega) \in V^{-1}([0, c])$. Hence, $\Omega \neq 0$. It follows from (22) that if any of the three equations

$$R^T R - I = 0, \quad \pi - \pi_0 = 0, \quad E - E_0 = 0$$

holds, then the three of them all hold. Thus

$$R^T R \neq I, \quad \pi \neq \pi_0, \quad E \neq E_0 \quad (24)$$

since $(R, \Omega) \notin V^{-1}(0)$. Since the matrix $(\pi - \pi_0)\Omega^T \mathbb{I}$ in (22a) has rank 1 and the matrix $(R^T R - I)$ is symmetric, there exist a unit vector $u \in \mathbb{R}^3$ and a number $\kappa \neq 0$ such that

$$R^T R - I = \kappa u u^T. \quad (25)$$

Substitution of (25) into (22a) and (23a) yields

$$\begin{aligned} k_0 \kappa R u u^T + k_2(\pi - \pi_0)\Omega^T \mathbb{I} &= 0, \\ k_0 \kappa R u u^T R^T + k_2(\pi - \pi_0)\pi^T &= 0, \end{aligned}$$

which implies

$$R u \parallel \pi \parallel \pi_0, \quad u \parallel \mathbb{I}\Omega, \quad (26)$$

where the symbol \parallel means ‘is parallel to.’ Hence, we can express R and π as

$$R = w_1 u_1^T + w_2 u_2^T + a e_{\pi_0} u^T, \quad (27)$$

$$\pi = b \pi_0, \quad (28)$$

for some numbers $a \neq 0$, $b \neq 1$ and vectors $u_1, u_2, w_1, w_2 \in \mathbb{R}^3$, where $e_{\pi_0} = \pi_0/|\pi_0|$ and the vectors u_1 and u_2 can be any vectors such that $\{u_1, u_2, u\}$ becomes an orthonormal basis for \mathbb{R}^3 . Substitution of (27) into (25) implies that $\{w_1, w_2, e_{\pi_0}\}$ is an orthonormal basis for \mathbb{R}^3 . Substitution of (27) and (28) into (22b) implies $\mathbb{I}\Omega \parallel \mathbb{I}u$, which together with $u \parallel \mathbb{I}\Omega$ in (26), implies $u \parallel \mathbb{I}u$, i.e., u is an eigenvector of \mathbb{I} . We can now choose or re-define the unit vectors u_1 and u_2 such that they become eigenvectors of the symmetric matrix \mathbb{I} , too. In the orthonormal basis $\{u_1, u_2, u\}$, we can now write the moment of inertia matrix \mathbb{I} as

$$\mathbb{I} = I_1 u_1 u_1^T + I_2 u_2 u_2^T + I_3 u u^T,$$

where I_1, I_2, I_3 are the eigenvalues of \mathbb{I} , which are all positive, corresponding to the eigenvectors u_1, u_2, u , respectively. It is then easy to see that equations (23) imply

$$k_0 a^2 (a^2 - 1) + k_2 |\pi_0|^2 b (b - 1) = 0, \quad (29a)$$

$$k_1 \left(\frac{|\pi_0|^2 b^2}{2I_3 a^2} - E_0 \right) b + k_2 I_3 a^2 (b - 1) = 0, \quad (29b)$$

where we have used $E = (1/2)\Omega^T \mathbb{I} \Omega = (1/2)\pi^T (R \mathbb{I} R^T)^{-1} \pi = |\pi_0|^2 b^2 / 2I_3 a^2$.

We consider the following two separate cases: $E_0 = |\pi_0|^2 / 2I_3$ and $E_0 \neq |\pi_0|^2 / 2I_3$. Suppose $E_0 = |\pi_0|^2 / 2I_3$. If $b \leq 0$, then

$$V(R, \Omega) \geq \frac{k_2}{2} |\pi - \pi_0|^2 = \frac{k_2}{2} (|b| + 1)^2 |\pi_0|^2 > c$$

by (21), which contradicts $(R, \Omega) \in V^{-1}([0, c])$. Hence, $b > 0$. If $b > 1$, then equation (29a) implies $a^2 < 1$, but equation (29b) implies $b^2 < a^2$, implying $b^2 < 1$. This cannot be compatible with $b > 1$. Hence, $b > 1$ is ruled out. Similarly, $0 < b < 1$ can be ruled out. Hence, $b = 1$, which implies $\pi = \pi_0$ contradicting (24). Thus, when $E_0 = |\pi_0|^2 / 2I_3$, there are no critical points of V in $V^{-1}([0, c]) \setminus V^{-1}(0)$.

Suppose $E_0 \neq |\pi_0|^2 / 2I_3$. We analyze equations (29) using a continuity argument. At $a^2 = 1$, (29a) implies $b = 0$ or 1 , neither of which satisfies (29b) at $a^2 = 1$. Thus, by continuity there exists a number δ with $0 < \delta < 1$ such that for any a with $|a^2 - 1| < \delta$ there is no number b satisfying both (29a) and (29b). Hence, $|a^2 - 1| \geq \delta$. We now shrink the number c such that it not only satisfies (21) but also $c < k_0 \delta^2 / 4$. For such a number c , we have

$$V(R, \Omega) \geq \frac{k_0}{4} \|R^T R - I\|^2 = \frac{k_0}{4} \|(a^2 - 1)uu^T\|^2 \geq \frac{k_0}{4} \delta^2 > c,$$

which contradicts $(R, \Omega) \in V^{-1}([0, c])$. Hence, when $E_0 \neq |\pi_0|^2 / 2I_3$, there are no critical points of V in $V^{-1}([0, c]) \setminus V^{-1}(0)$ for some $c > 0$.

Therefore, there exists a number $c > 0$ such that $V^{-1}(0)$ is the set of all critical points of V in $V^{-1}([0, c])$. \square

Consider the dynamics

$$\dot{R} = R\hat{\Omega} - k_0 R(R^T R - I) - k_2 (\pi(R, \Omega) - \pi_0) \Omega^T \mathbb{I}, \quad (30a)$$

$$\dot{\Omega} = \mathbb{I}^{-1}((\mathbb{I}\Omega) \times \Omega) - k_1 (E(\Omega) - E_0) \mathbb{I} \Omega - k_2 \mathbb{I} R^T (\pi(R, \Omega) - \pi_0), \quad (30b)$$

which correspond to (3). From Theorem 2.1 and Lemmas 3.2 and 3.3 comes the following theorem:

Theorem 3.4. *There is a number $c > 0$ such that every trajectory of (30) starting from a point in $V^{-1}([0, c])$ stays in $V^{-1}([0, c])$ for all $t \geq 0$ and asymptotically converges to the set*

$$V^{-1}(0) = \{(R, \Omega) \in \mathbb{R}^{3 \times 3} \times \mathbb{R}^3 \mid R \in \text{SO}(3), E(\Omega) = E_0, \pi(R, \Omega) = \pi_0\}$$

as $t \rightarrow \infty$, where the function V is defined in (18). Furthermore, $V^{-1}(0)$ is an invariant set of both (13) and (30).

3.2 The Kepler Problem

The two-body dynamics in the Kepler problem are given in the usual barycentric coordinates by

$$\dot{x} = v, \quad (31a)$$

$$\dot{v} = -\mu \frac{x}{|x|^3}, \quad (31b)$$

where $x \in \mathbb{R}_0^3 := \mathbb{R}^3 \setminus \{(0, 0, 0)\}$ is the position vector, $v \in \mathbb{R}^3$ is the velocity vector and μ is the gravitational parameter. Define two functions $L : \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}^3$ and $A : \mathbb{R}_0^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}^3$ by

$$L(x, v) = x \times v, \quad (32)$$

$$A(x, v) = v \times (x \times v) - \mu \frac{x}{|x|}, \quad (33)$$

where L is called the angular momentum vector and A is called the Laplace-Runge-Lenz vector. It is known that both L and A are first integrals of the two-body dynamics (31) and they are orthogonal to each other, i.e.,

$$L(x, v) \perp A(x, v)$$

for all $(x, v) \in \mathbb{R}_0^3 \times \mathbb{R}^3$. The energy function

$$E(x, v) = \frac{1}{2}|v|^2 - \frac{\mu}{|x|}$$

satisfies

$$|A(x, v)|^2 = \mu^2 + 2E(x, v)|L(x, v)|^2 \quad (34)$$

for all $(x, v) \in \mathbb{R}_0^3 \times \mathbb{R}^3$, implying that the energy E is also a first integral of the two-body dynamics (31). It is also known that a non-degenerate elliptic Keplerian orbit is uniquely determined by a pair (L, A) that satisfies $L \perp A$, $|L| \neq 0$ and $|A| < \mu$ [1].

Fix a non-degenerate elliptic Keplerian orbit, i.e., a pair of vectors (L_0, A_0) that satisfies

$$L_0 \perp A_0, \quad |L_0| \neq 0, \quad |A_0| < \mu.$$

Define a function $V : \mathbb{R}_0^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}$ by

$$V(x, v) = \frac{k_1}{2}|L(x, v) - L_0|^2 + \frac{k_2}{2}|A(x, v) - A_0|^2 \quad (35)$$

for $(x, v) \in \mathbb{R}_0^3 \times \mathbb{R}^3$, where $k_1 > 0$ and $k_2 > 0$. Notice that

$$V^{-1}(0) = \{(x, v) \in \mathbb{R}_0^3 \times \mathbb{R}^3 \mid L(x, v) = L_0, A(x, v) = A_0\},$$

which is the non-degenerate Keplerian elliptic orbit whose angular momentum vector and Laplace-Runge-Lenz vector are L_0 and A_0 , respectively.

Lemma 3.5. *The gradient $(\nabla_x V, \nabla_v V) \in \mathbb{R}^3 \times \mathbb{R}^3$ of the function V defined in (35) is given by*

$$\begin{aligned} \nabla_x V &= k_1 v \times \Delta L + k_2 \left(v \times (\Delta A \times v) - \frac{\mu}{|x|} \Delta A + \frac{\mu}{|x|^3} x x^T \Delta A \right), \\ \nabla_v V &= k_1 \Delta L \times x + k_2 ((x \times v) \times \Delta A + x \times (v \times \Delta A)), \end{aligned}$$

where $\Delta L = L(x, v) - L_0$ and $\Delta A = A(x, v) - A_0$.

The following lemma shows that the function V defined in (35) satisfies assumptions A1 and A2 stated in §2.

Lemma 3.6. *1. The function V satisfies*

$$\langle (\nabla_x V, \nabla_v V), (v, -\mu x/|x|^3) \rangle = 0.$$

2. For any number c satisfying

$$0 < c < \min\{k_1|L_0|^2/2, k_2(\mu - |A_0|)^2/2\}, \quad (36)$$

the set $V^{-1}([0, c])$ is a compact set in $\mathbb{R}_0^3 \times \mathbb{R}^3$.

Proof. The first fact is a straightforward calculation using the previous Lemma. For the second, the essential idea is that the fibers of V are homeomorphic to circles, corresponding to the elliptic orbits, and are therefore compact. For a detailed proof of the second statement, refer to Corollary 2.2 in [1]. \square

The following lemma shows that the function V defined in (35) satisfies assumption A3 stated in §2.

Lemma 3.7. *For any number c satisfying (36) the set of all critical points of V in $V^{-1}([0, c])$ is equal to $V^{-1}(0)$.*

Proof. Choose an arbitrary number c satisfying (36). Let (x, v) be an arbitrary critical point of V in $V^{-1}([0, c])$. For notational convenience, let us write

$$L = L(x, v), \quad A = A(x, v)$$

suppressing the dependence on (x, v) . By Lemma 3.5, the critical point (x, v) satisfies

$$0 = k_1 v \times \Delta L + k_2 \left(v \times (\Delta A \times v) - \frac{\mu}{|x|} \Delta A + \frac{\mu}{|x|^3} x x^T \Delta A \right), \quad (37a)$$

$$0 = k_1 \Delta L \times x + k_2 ((x \times v) \times \Delta A + x \times (v \times \Delta A)). \quad (37b)$$

If $|L| = 0$, then $V(x, v) \geq k_1 |L_0|^2 / 2 > c$, contradicting $(x, v) \in V^{-1}([0, c])$. Hence, $|L| \neq 0$, which together with (32) implies that the three vectors x, v, L form a basis for \mathbb{R}^3 . The dot product of (37b) with x yields

$$0 = x \cdot ((x \times v) \times \Delta A) = \Delta A \cdot (x \times L),$$

so there are numbers a and b such that

$$\Delta A = ax + bL. \quad (38)$$

Substitution of (38) into (37) gives

$$0 = v \times \left(k_1 \Delta L + k_2 \left(aL - bv \times L + \frac{b\mu}{|x|} x \right) \right),$$

$$0 = (k_1 \Delta L + k_2 (2aL - bv \times L)) \times x.$$

It follows that there are numbers d and f such that

$$k_1 \Delta L + k_2 \left(aL - bv \times L + \frac{b\mu}{|x|} x \right) = dv, \quad (39a)$$

$$k_1 \Delta L + k_2 (2aL - bv \times L) = fx. \quad (39b)$$

From (39), we obtain

$$\left(\frac{bk_2\mu}{|x|} + f \right) x - dv - ak_2L = 0.$$

By linear independence of $\{x, v, L\}$,

$$a = 0, \quad d = 0, \quad f = -bk_2\mu/|x|.$$

Substitution of these into (38) and (39b) gives

$$\Delta A = bL, \quad \Delta L = \frac{bk_2}{k_1} A,$$

where we have used the definition of A given in (33). Hence,

$$A_0 = A - bL, \quad L_0 = L - \frac{bk_2}{k_1}A. \quad (40)$$

From (40) and the orthogonality $A_0 \perp L_0$ and $A \perp L$, it follows that

$$0 = A_0 \cdot L_0 = -b \left(|L|^2 + \frac{k_2}{k_1} |A|^2 \right).$$

Since $|L| \neq 0$, and recalling that $k_1 > 0$ and $k_2 > 0$, we have $b = 0$. Substitution of $b = 0$ into (40) yields

$$L = L_0, \quad A = A_0,$$

which implies $(x, v) \in V^{-1}(0)$. Thus, every critical point of V in $V^{-1}([0, c])$ is contained in $V^{-1}(0)$.

Since 0 is the minimum value of V , every point in $V^{-1}(0)$ is a critical point of V . Therefore, the set of all critical points of V in $V^{-1}([0, c])$ is $V^{-1}(0)$. \square

Choose a non-degenerate Keplerian elliptic orbit and let (x_0, v_0) be a point on the orbit. Set

$$L_0 = L(x_0, v_0), \quad A_0 = A(x_0, v_0)$$

to be the angular momentum vector and the Laplace-Runge-Lenz vector of the orbit, respectively. Consider the dynamics:

$$\dot{x} = v - k_1 v \times \Delta L - k_2 \left(v \times (\Delta A \times v) - \frac{\mu}{|x|} \Delta A + \frac{\mu}{|x|^3} x x^T \Delta A \right), \quad (41a)$$

$$\dot{v} = -\mu \frac{x}{|x|^3} - k_1 \Delta L \times x - k_2 ((x \times v) \times \Delta A + x \times (v \times \Delta A)), \quad (41b)$$

where $\Delta L = L(x, v) - L_0$ and $\Delta A = A(x, v) - A_0$, which correspond to (3). From Theorem 2.1 and Lemmas 3.6 and 3.7 comes the following theorem:

Theorem 3.8. *For any $c > 0$ satisfying (36), every trajectory of (41) starting from a point in $V^{-1}([0, c])$ stays in $V^{-1}([0, c])$ for all $t \geq 0$ and asymptotically converges to the set*

$$V^{-1}(0) = \{(x, v) \in \mathbb{R}_0^3 \times \mathbb{R}^3 \mid L(x, v) = L_0, A(x, v) = A_0\}$$

as $t \rightarrow \infty$, where the function V is defined in (35). Furthermore, $V^{-1}(0)$ is an invariant set of both (31) and (41).

3.3 A Perturbed Kepler Problem with Rotational Symmetry

Consider a perturbed Kepler problem with rotational symmetry whose equations of motion are given by

$$\dot{x} = v, \quad (42a)$$

$$\dot{v} = -U'(|x|) \frac{x}{|x|}, \quad (42b)$$

where $x \in \mathbb{R}_0^3 := \mathbb{R}^3 \setminus \{(0, 0, 0)\}$ is the position vector, $v \in \mathbb{R}^3$ is the velocity vector, and $U : (0, \infty) \rightarrow \mathbb{R}$ is the potential function that depends only on the radial distance from the

origin. The total energy $E : \mathbb{R}_0^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}$ and the angular momentum vector $L : \mathbb{R}_0^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}^3$ are defined by

$$E(x, v) = \frac{1}{2}|v|^2 + U(|x|), \quad (43)$$

$$L(x, v) = x \times v \quad (44)$$

and they are conserved quantities of the dynamics (42). Take any point $(x_0, v_0) \in \mathbb{R}_0^3 \times \mathbb{R}^3$ such that

$$x_0 \times v_0 \neq 0.$$

Let

$$E_0 = E(x_0, v_0), \quad L_0 = L(x_0, v_0) \neq 0.$$

Define a function $V : \mathbb{R}_0^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}$ by

$$V(x, v) = \frac{k_1}{2}|E(x, v) - E_0|^2 + \frac{k_2}{2}|L(x, v) - L_0|^2$$

with $k_1 > 0$ and $k_2 > 0$. Then,

$$V^{-1}(0) = \{(x, v) \in \mathbb{R}_0^3 \times \mathbb{R}^3 \mid E(x, v) = E_0, L(x, v) = L_0\}.$$

The gradient $(\nabla_x V, \nabla_v V)$ of V is given by

$$\nabla_x V = k_1 \Delta E U'(|x|) \frac{x}{|x|} + k_2 v \times \Delta L,$$

$$\nabla_v V = k_1 \Delta E v + k_2 \Delta L \times x,$$

where $\Delta E = E(x, v) - E_0$ and $\Delta L = L(x, v) - L_0$. Trivially, V satisfies (2), i.e.

$$\langle (\nabla_x V, \nabla_v V), (v, -U'(|x|)x/|x|) \rangle = 0 \quad (45)$$

for all $(x, v) \in \mathbb{R}_0^3 \times \mathbb{R}^3$. The modified dynamics, which correspond to (3), are computed as

$$\dot{x} = v - k_1 \Delta E U'(|x|) \frac{x}{|x|} - k_2 v \times \Delta L, \quad (46a)$$

$$\dot{v} = -U'(|x|) \frac{x}{|x|} - k_1 \Delta E v - k_2 \Delta L \times x. \quad (46b)$$

Theorem 3.9. *Suppose that $V^{-1}(0)$ is compact and there is no common solution $r > 0$ to the following two equations:*

$$E_0 = \frac{1}{2}rU'(r) + U(r), \quad (47)$$

$$|L_0|^2 = r^3U'(r). \quad (48)$$

Then, assumptions A2 and A3 hold and there is a number $c > 0$ such that every trajectory of (46) starting in $V^{-1}([0, c])$ remains in $V^{-1}([0, c])$ for all $t \geq 0$ and asymptotically converges to $V^{-1}(0)$ as $t \rightarrow \infty$.

Proof. Define a function $f : \mathbb{R}_0^3 \times \mathbb{R}^3 \rightarrow \mathbb{R} \times \mathbb{R}^3$ by

$$f(x, v) = \begin{bmatrix} E(x, v) \\ L(x, v) \end{bmatrix}.$$

Then,

$$Df(x, v)^T = \begin{bmatrix} U'(|x|) \frac{x}{|x|} & \hat{v} \\ v & -\hat{x} \end{bmatrix},$$

where the over-hat symbol $\hat{\cdot}$ denotes the hat map defined in (14). We want to show that the 6×4 matrix $Df(x, v)^T$ is one-to-one for all $(x, v) \in V^{-1}(0)$. Fix an arbitrary point $(x, v) \in V^{-1}(0)$. It follows

$$E_0 = \frac{1}{2}|v|^2 + U(|x|), \quad (49)$$

$$L_0 = x \times v \neq 0. \quad (50)$$

Take any point $(a, w) \in \mathbb{R} \times \mathbb{R}^3$ from the kernel of $Df(x, v)^T$. Then,

$$0 = aU'(|x|) \frac{x}{|x|} + v \times w, \quad (51a)$$

$$0 = av - x \times w. \quad (51b)$$

Suppose $a \neq 0$. Taking the inner product of (51a) with x and of (51b) with v , we obtain

$$0 = aU'(|x|)|x| + L_0 \cdot w,$$

$$0 = a|v|^2 + L_0 \cdot w,$$

from which it follows that

$$|x|U'(|x|) = |v|^2. \quad (52)$$

Taking the inner product of (51b) with x , we get $x \cdot v = 0$ which implies

$$|L_0| = |x| \cdot |v|. \quad (53)$$

From (49), (52) and (53), we obtain

$$E_0 = \frac{1}{2}|x|U'(|x|) + U(|x|), \quad (54)$$

$$|L_0|^2 = |x|^3U'(|x|). \quad (55)$$

By hypothesis, there cannot be any $x \in \mathbb{R}_0^3$ that satisfies both (54) and (55). Hence, we cannot have $a \neq 0$.

Substitute $a = 0$ into (51). It follows that w is parallel to $x \times v$. Hence, there is a number b such that $w = bL_0$. Substituting this in (51b) yields $bx \times L_0 = 0$. Taking the cross product of this with x yields $b|x|^2L_0 = 0$ since $x \cdot L_0 = 0$. Since $x \neq 0$ and $L_0 \neq 0$, we have $b = 0$, so $w = 0$. It follows that $(a, w) = (0, 0)$, which implies that $Df(x, v)^T$ is one-to-one for all $(x, v) \in V^{-1}(0)$. In other words, $Df(x, v)$ is onto for all $(x, v) \in V^{-1}(0)$. Hence, the conclusion of the theorem follows from Lemma 2.4, equation (45), and Theorem 2.5. \square

Remark 3.10. Consider a special case in which the potential function $U(r)$ is of the form

$$U(r) = -\frac{\mu}{r} - \frac{\delta}{r^3}, \quad (56)$$

where $\mu > 0$ and $\delta > 0$. Then equations (47) and (48) become

$$E_0 = -\frac{\mu}{2r} + \frac{\delta}{2r^3}, \quad (57)$$

$$|L_0|^2 = \mu r + \frac{3\delta}{r}. \quad (58)$$

Given E_0 and L_0 , it is then easy to check if there is no common solution $r > 0$ to (57) and (58).

4 Simulations

4.1 The Free Rigid Body

Consider the free rigid body dynamics in §3.1 with the moment of inertia matrix $\mathbb{I} = \text{diag}(3, 2, 1)$ and the initial condition

$$R(0) = I, \quad \Omega(0) = (1, 1, 1). \quad (59)$$

The values of the energy E and the spatial angular momentum vector $\pi = (\pi_1, \pi_2, \pi_3)$ corresponding to the initial condition are

$$E(0) = 3, \quad \pi(0) = (3, 2, 1).$$

The period T_Ω of the trajectory of the body angular velocity vector $\Omega(t)$ is computed approximately to be $T_\Omega = 6.4227$.

We integrate the dynamics over the time interval $[0, 10^3] = [0, 155.7T_\Omega]$ with step size $\Delta t = 10^{-4}$, using the following four integration methods: a feedback integrator with the Euler scheme, a projection method with the Euler scheme, a splitting method with three rotations splitting, and the ordinary Euler method. The feedback integrator with the Euler scheme denotes the Euler method applied to the modified free rigid dynamics (30) with the following values of the parameters k_0 , k_1 , and k_2

$$k_0 = 50, \quad k_1 = 100, \quad k_2 = 50.$$

The projection method is the standard one explained on pp.110–111 in [2]. In order to solve constraint equations for projection at each step of integration in the projection method, we use the Matlab command *fsolve* with the parameter *TolFun*, which is termination tolerance on the function value, set equal to 10^{-4} , which is the same as the integration step size Δt . The splitting method is the one explained on pp.284–285 in [2]. The three of the projection method, the splitting method and the ordinary Euler method are applied to the original free rigid body dynamics (13).

The trajectories of the body angular velocity vector $\Omega(t)$, the energy error $|\Delta E(t)| = |E(t) - E(0)|$, the error $|\Delta \pi(t)| = |\pi(t) - \pi(0)|$ in spatial angular momentum, and the deviation $\|R(t)^T R(t) - I\|$ of the rotation matrix $R(t)$ from $\text{SO}(3)$ are plotted in Figures 1, 2, 3 and 4, respectively. In Figure 1, it is observed that the trajectories of $\Omega(t)$ generated by the feedback integrator and the projection method maintain a periodic shape well whereas those by the splitting method and the Euler method drift away significantly from the periodic shape. In Figure 2, it is observed that the feedback integrator and the projection method keep the energy error sufficiently small whereas the energy errors by the other two methods increase in time. Although the two trajectories of energy error by the splitting method and the Euler method seem to coincide in Figure 2, an examination of the numerical data shows that the energy of the Euler method gets larger than that of the splitting method in time. For example, at $t = 1000$, the energy of the Euler method is bigger than that of the splitting method by 1.767×10^{-3} . In Figures 3 and 4, it is observed that the feedback method preserves the spatial angular momentum vector and the manifold $\text{SO}(3)$ sufficiently well. In terms of computation time, the projection method takes much more time than the others, which is due to the steps of solving the constraint equations for projection. The splitting method is symplectic and of order 2 whereas the other methods are of order 1. All of these observations lead us to the conclusion that the feedback integrator overall has produced the best outcome in the simulation of the free rigid body dynamics.

4.2 The Kepler Problem

Consider the Kepler problem in §3.2 with $\mu = 1$ and the initial condition

$$x(0) = (1, 0, 0), \quad v(0) = (0, \sqrt{1.8}, 0).$$

The corresponding initial values of the angular momentum vector and the Laplace-Runge-Lenz vector are

$$L(0) = (0, 0, \sqrt{1.8}), \quad A(0) = (0.8, 0, 0).$$

The period T and the eccentricity e of the Kepler orbit containing the initial point are

$$T = 70.2481, \quad e = 0.8.$$

We integrate the Kepler dynamics over the time interval $[0, 1000T]$ with step size $\Delta t = 0.005$, using the following four integration methods: a feedback integrator with the Euler scheme, the standard projection method with the Euler scheme, and two Störmer-Verlet schemes. The feedback integrator with the Euler scheme denotes the Euler method applied to (41) with $k_1 = 4$ and $k_2 = 2$. The standard projection method is explained on pp.110–111 in [2]. To solve the constraint equations for projection, we use the Matlab command *fsolve* with the parameter *TolFun* set equal to 0.005, which is the same as the integration step size Δt . The two Störmer-Verlet schemes are those in (3.4) and (3.5) on pp. 189–190 in [2], and we call them Störmer-Verlet-A and Störmer-Verlet-B, respectively, for convenience. The Störmer-Verlet schemes are symplectic methods of order 2.

The trajectories of the planar orbit $x(t) = (x_1(t), x_2(t), 0)$, the error of the Laplace-Runge-Lenz vector, $|\Delta A(t)| = |A(t) - A(0)|$, and the error of the angular momentum vector, $|\Delta L(t)| = |L(t) - L(0)|$, are plotted in Figures 5, 6 and 7. In Figure 5 it is observed that the planar trajectories $x(t) = (x_1(t), x_2(t), 0)$ generated by the feedback integrator and the projection method maintain the elliptic shape well whereas those by the Störmer-Verlet schemes precess. This can be also verified in Figure 6, where the feedback integrator and the projection method preserve the Laplace-Runge-Lenz vector well, but the Störmer-Verlet schemes cause the Laplace-Runge-Lenz vector to noticeably precess. In Figure 7, it is observed that the Störmer-Verlet schemes preserve the angular momentum vector exceptionally well in comparison with the other two methods. In Figures 6 and 7, we can see that the precision of the feedback integrator is comparable with that of the projection method. However, the feedback integrator takes much less computation time than the projection method. The feedback integrator and the projection method used here are of order 1, whereas the Störmer-Verlet schemes are of order 2. All of these observations lead us to conclude that the feedback integrator has produced the best result overall.

4.3 A Perturbed Kepler Problem with Rotational Symmetry

Consider the perturbed Kepler problem in §3.3 with the potential function U given in (56) with $\mu = 1$ and $\delta = 0.0025$, which is the one used in Example 4.3 on p. 111 in [2]. We use the initial conditions

$$x(0) = (1 - e, 0, 0), \quad v(0) = (0, \sqrt{(1+e)/(1-e)}, 0)$$

with eccentricity $e = 0.6$ as in [2]. The corresponding values of the energy and the angular momentum vector are

$$E(0) = -0.5390625, \quad L(0) = (0, 0, 0.8).$$

We integrate the perturbed Kepler dynamics over the time interval $[0, 200]$ with step size $\Delta t = 0.03$, just as on p. 111 in [2], using the following four integration methods: a feedback

integrator with the Euler scheme, the standard projection method with the Euler scheme, the Störmer-Verlet scheme in (3.4) on p. 189 in [2], and the Matlab command, *ode45*. The feedback integrator with the Euler scheme denotes the Euler method applied to (46) with $k_1 = 2$ and $k_2 = 3$, and it is straightforward to verify that the hypotheses in Theorem 3.9 hold true. The other three methods are applied to (42). The Matlab command *fsolve* is used in the projection method with the parameter *TolFun* set equal to 10^{-8} . The options of $RelTol = AbsTol = 10^{-10}$ are used for the Matlab integrator, *ode45*, so the result generated by *ode45* can be used as a reference.

The trajectories of the planar orbit $x(t) = (x_1(t), x_2(t), 0)$, the energy error $|\Delta E(t)| = |E(t) - E(0)|$ and the error $|\Delta L(t)| = |L(t) - L(0)|$ in angular momentum are plotted in Figures 8, 9 and 10. In Figure 8 it is observed that the orbits generated by the feedback integrator and the Störmer-Verlet scheme are similar to that by *ode45*, but the orbit by the projection method precesses too much which is a very poor result. The projection method excels only at preserving the energy and the angular momentum as expected in view of the nature of the projection method and the small tolerance parameter value, $TolFun = 10^{-8}$, used for the Matlab command, *fsolve*. In Figure 9, it is observed that the feedback integrator is comparable with the Störmer-Verlet scheme in energy conservation. The feedback integrator also preserves the angular momentum well in view of the step size $\Delta t = 0.03$, as can be seen in Figure 10. The feedback integrator and the projection method used here are of order 1 whereas the Störmer-Verlet scheme is of order 2. From all of these observations, we conclude that the feedback integrator has produced the best result overall.

5 Conclusions and Future Work

We have developed a theory to produce numerical trajectories of a dynamical system on a manifold that stably remain on the manifold and preserve first integrals of the system. Our theory is not a numerical integration scheme but rather a modification of the original dynamics by feedback. The actual numerical integration in our framework can be done with any usual integrator such as Euler and Runge-Kutta. Our method is successfully applied to the free rigid body, the Kepler problem and a perturbed Kepler problem with rotational symmetry, and its excellent performance is demonstrated by simulation studies in comparison with the standard projection method, two Störmer-Verlet schemes and a splitting method via three rotations splitting.

As future work, we plan to apply our theory to various mechanical systems with symmetry and non-holonomic systems. We also plan to carry out a quantitative study of the effect of the parameters in the Lyapunov function on the performance of our method.

Appendix

We show, using results in [3], that any discrete-time dynamical system derived from a one-step numerical integration scheme with uniform step size h for the modified dynamical system (3) has an attractor Λ_h that contains $V^{-1}(0)$ in its interior and converges to $V^{-1}(0)$ as $h \rightarrow 0+$. Let us first review some definitions from [3]. Let A and B be nonempty, compact subsets of \mathbb{R}^n and x a point in \mathbb{R}^n . The distance between x and A is defined by

$$\text{dist}(x, A) = \inf\{|x - a|, a \in A\}.$$

The Hausdorff separation of A from B is defined by

$$H^*(A, B) = \max\{\text{dist}(a, B), a \in A\}.$$

The Hausdorff distance between A and B is defined by

$$H(A, B) = \max\{H^*(A, B), H^*(B, A)\}.$$

The Hausdorff distance is a metric on the space of nonempty compact subsets of \mathbb{R}^n . For $r > 0$, let

$$S(A, r) = \{x \in \mathbb{R}^n \mid \text{dist}(x, A) < r\}$$

denote an r -neighborhood of A .

We say that a nonempty, compact subset Λ of \mathbb{R}^n is uniformly stable for an autonomous dynamical system if for each $\epsilon > 0$ there exists a $\delta = \delta(\epsilon) > 0$ such that

$$[x_0 \in S(\Lambda, \delta) \text{ and } t \geq 0] \Rightarrow x(t; x_0) \in S(\Lambda, \epsilon),$$

where $x(t; x_0)$ is the solution of the given dynamical system with initial condition $x(0) = x_0$. A set Λ is said to be positively invariant for an autonomous dynamical system if $x(t; x_0) \in \Lambda$ for all $x_0 \in \Lambda$ and $t \geq 0$. A nonempty, compact subset Λ of \mathbb{R}^n is called uniformly asymptotically stable for an autonomous dynamical system if it is positively invariant and uniformly stable for the dynamical system, and additionally satisfies the following property: there is a $\delta_0 > 0$ and for each $\epsilon > 0$ a time $T(\epsilon) > 0$ such that

$$[x_0 \in S(\Lambda, \delta_0) \text{ and } t \geq T(\epsilon)] \Rightarrow x(t; x_0) \in S(\Lambda, \epsilon).$$

Lemma 5.1. *Suppose that assumptions A1, A2 and A3 (or A3' instead of A3) stated in §2 hold true. Then, the set $V^{-1}(0)$ is uniformly asymptotically stable for the modified dynamical system (3).*

Proof. Since the three assumptions are satisfied, the conclusions of Theorem 2.1 (or, 2.2) hold true. For convenience, let $\Lambda = V^{-1}(0)$, which is invariant under (3) by Theorem 2.1 (or, 2.2). Let $c > 0$ be the number c in assumption A2. Using compactness of $V^{-1}([0, c])$ and continuity of V , it is easy to show that for any $\epsilon > 0$ there is a $b = b(\epsilon) > 0$ such that $V^{-1}([0, b]) \subset S(\Lambda, \epsilon)$. It is also easy to show that for any $b > 0$ there is an $\epsilon = \epsilon(b) > 0$ such that $S(\Lambda, \epsilon) \subset V^{-1}([0, b])$. Hence, we can use the family of sets $\{V^{-1}([0, b]), b > 0\}$ instead of the family of open sets $\{S(\Lambda, \epsilon), \epsilon > 0\}$ to show uniform stability and uniform asymptotic stability of Λ for (3).

Let us first show uniform stability of Λ for (3). Given any $\epsilon > 0$, take any δ such that $0 < \delta \leq \min\{\epsilon, c\}$. Then, for any $x_0 \in V^{-1}([0, \delta])$, $x(t; x_0) \in V^{-1}([0, \delta]) \subset V^{-1}([0, \epsilon])$ for all $t \geq 0$ since V is decreasing along the trajectory of $x(t; x_0)$ of (3). Hence, Λ is uniformly stable for (3).

Let us now show uniform asymptotic stability of Λ for (3). Take any δ_0 such that $0 < \delta_0 \leq c$. By continuous dependence of $x(t; x_0)$ on initial point x_0 , compactness of $V^{-1}([0, \delta_0])$, continuity of the function V , and the property that $V(x(t; x_0))$ decreases to 0 as $t \rightarrow \infty$ for any $x_0 \in V^{-1}([0, c])$, it is easy to show that for any $\epsilon > 0$ there is a time $T(\epsilon) > 0$ such that for any $x_0 \in V^{-1}([0, \delta_0])$ we have $x(t; x_0) \in V^{-1}([0, \epsilon])$ for all $t \geq T(\epsilon)$. Hence, Λ is uniformly asymptotically stable for (3). □

Suppose the vector field X is C^p and the function V is C^{p+1} in the modified dynamical system (3). Consider a discrete analogue of (3) described by any one-step numerical method of p th order

$$x_{k+1} = x_k + hY_h(x_k) \tag{60}$$

with uniform step size $h > 0$, where $Y_h : \mathbb{R}^n \rightarrow \mathbb{R}^n$ for each h .

Theorem 5.2. *Suppose that the vector field X is C^p and the function V is C^{p+1} , and that assumptions A1, A2 and A3 (or A3' instead of A3) are satisfied. Then there is a number $h_2 > 0$ such that for each $0 < h < h_2$ the discrete-time dynamical system (60) has a compact, uniformly asymptotically stable set Λ_h which contains $V^{-1}(0)$ in its interior and converges to $V^{-1}(0)$ with respect to the Hausdorff metric as $h \rightarrow 0+$. Moreover, there is a bounded, open set U_0 , which is independent of h and contains Λ_h , and a time*

$$T_0(h) = A + Bp \log \frac{1}{h},$$

where A and B are constants depending on the stability characteristic of $V^{-1}(0)$, such that the iterates of (60) satisfy

$$x_k \in \Lambda_h$$

for all $kh \geq T_0(h)$, $x_0 \in U_0$ and $0 < h < h_2$.

Proof. We have only to show that the hypotheses in Theorem 1.1 of [3] hold. Since X is C^p and V is C^{p+1} , the vector field $X - \nabla V$ of (3) and its derivatives of order up to p are all continuous and bounded on the compact set $V^{-1}([0, c])$. The set $V^{-1}(0)$ is uniformly asymptotically stable for (3) by Lemma 5.1 in the above. Therefore, the conclusions of this theorem follow from Theorem 1.1 and Lemma 3.3 of [3]. \square

Refer to [3] to see how to obtain the set U_0 and values of the parameters h_2 , A and B that appear in the statement of the above theorem. The above theorem extends to multi-step numerical integrators; refer to [4] for detail.

Acknowledgement

This research was supported in part by DGIST Research and Development Program (CPS Global Center) funded by the Ministry of Science, ICT & Future Planning, Global Research Laboratory Program (2013K1A1A2A02078326) through NRF, and Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government (MSIP) (No. B0101-15-0557, Resilient Cyber-Physical Systems Research).

References

- [1] CHANG DE, CHICHKA DF AND MARSDEN JE, “Lyapunov-Based Transfer between Elliptic Keplerian Orbits,” *Discrete and Continuous Dynamical Systems – Series B*, **2**(1), pp. 57–67, (2002).
- [2] HAIRER E, LUBICH C AND WANNER G, “Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations,” *Springer Series in Computational Mathematics*, **31**, 2nd Ed., Springer, (2006).
- [3] KLOEDEN PE AND LORENZ J, “Stable Attracting Sets in Dynamical Systems and in Their One-Step Discretizations,” *SIAM J. Numer. Anal.*, **23**(5), pp. 986 – 995, (1986).
- [4] KLOEDEN PE AND LORENZ J, “A Note on Multistep Methods and Attracting Sets of Dynamical Systems,” *Numer. Math.*, **56**, pp. 667 – 673, (1990).
- [5] LASALLE JP, “Some Extensions of Liapunov’s Second Method,” *IRE Trans. Circuit Theory*, **7**(4), pp.520 – 527, (1960).

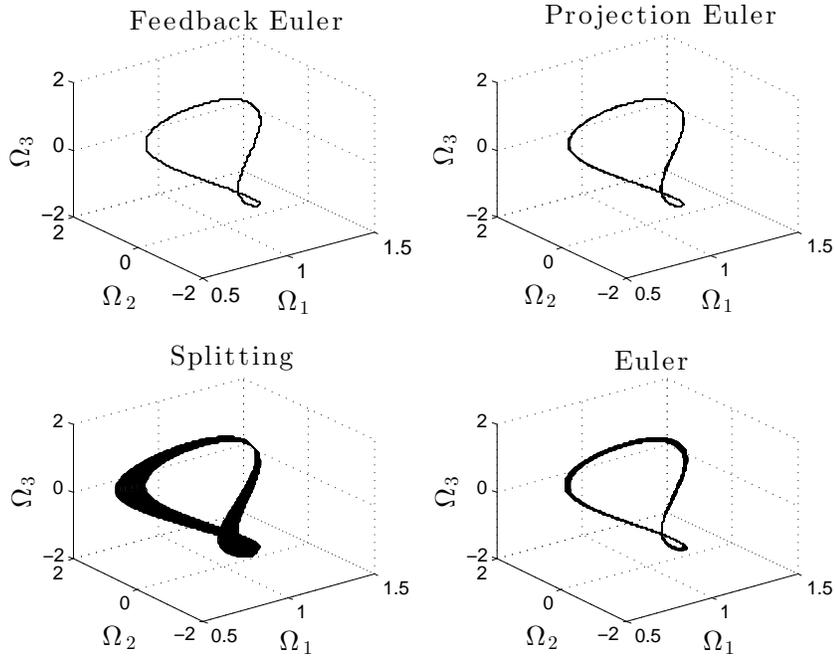


Figure 1: The trajectories of the body angular velocity $\Omega(t) = (\Omega_1(t), \Omega_2(t), \Omega_3(t))$, $0 \leq t \leq 1000$, of the free rigid body dynamics generated by four different methods with step size $\Delta t = 10^{-4}$: a feedback integrator with the Euler scheme, the standard projection method with the Euler scheme, a three rotations splitting method and the usual Euler method.

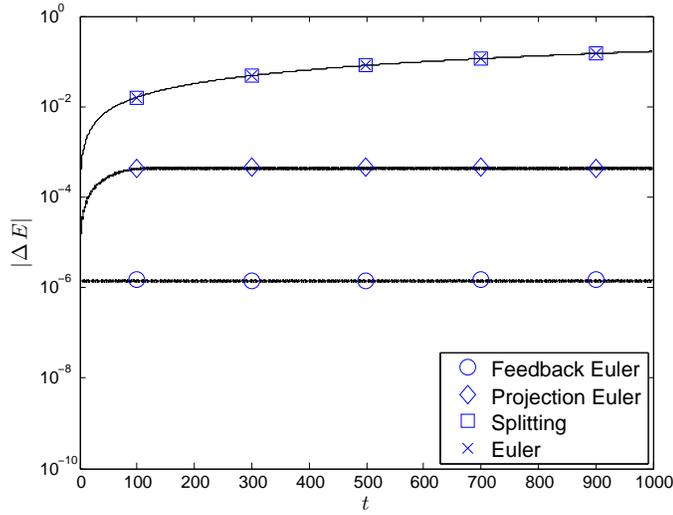


Figure 2: The trajectories of the energy error $|\Delta E(t)| = |E(t) - E(0)|$, $0 \leq t \leq 1000$, of the free rigid body dynamics generated by four different methods with step size $\Delta t = 10^{-4}$: a feedback integrator with the Euler scheme (\circ), the standard projection method with the Euler scheme (\diamond), a three rotations splitting method (\square) and the usual Euler method (\times).

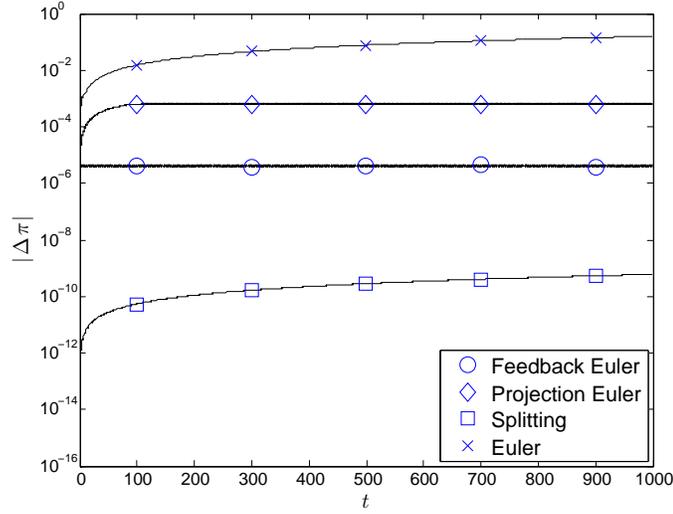


Figure 3: The trajectories of the spatial angular momentum error $|\Delta\pi(t)| = |\pi(t) - \pi(0)|$, $0 \leq t \leq 1000$, of the free rigid body dynamics generated by four different methods with step size $\Delta t = 10^{-4}$: a feedback integrator with the Euler scheme (\circ), the standard projection method with the Euler scheme (\diamond), a three rotations splitting method (\square) and the usual Euler method (\times).

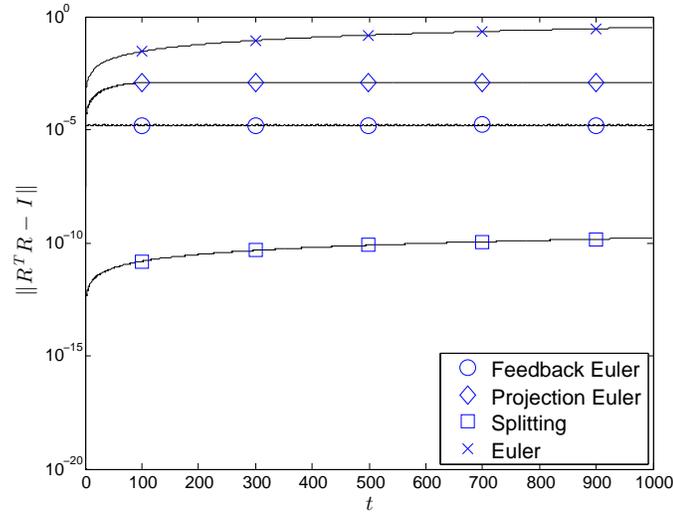


Figure 4: The trajectories of the deviation $\|R(t)^T R(t) - I\|$ of the rotation matrix $R(t)$ from $SO(3)$, $0 \leq t \leq 1000$, of the free rigid body dynamics generated by four different methods with step size $\Delta t = 10^{-4}$: a feedback integrator with the Euler scheme (\circ), the standard projection method with the Euler scheme (\diamond), a three rotations splitting method (\square) and the usual Euler method (\times).

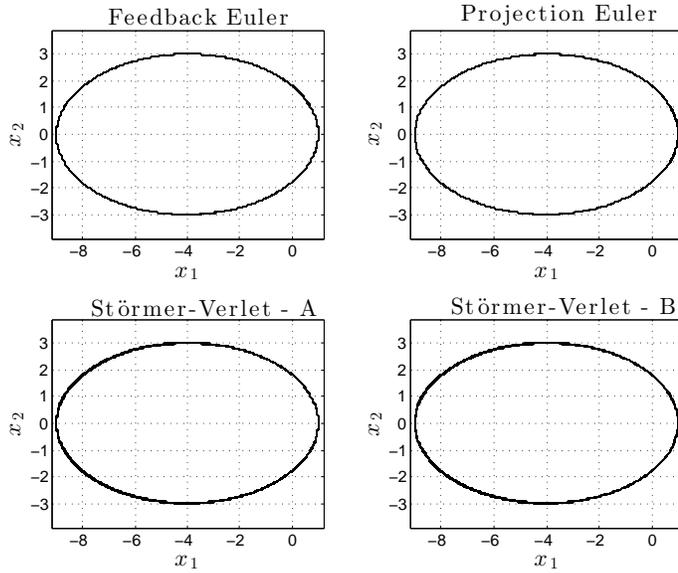


Figure 5: The trajectories of the planar orbit $x(t) = (x_1(t), x_2(t), 0)$, $0 \leq t \leq 70,248$, in the Kepler problem generated by four different methods with step size $\Delta t = 0.005$: a feedback integrator with the Euler scheme, the standard projection method with the Euler scheme, and two Störmer-Verlet schemes.

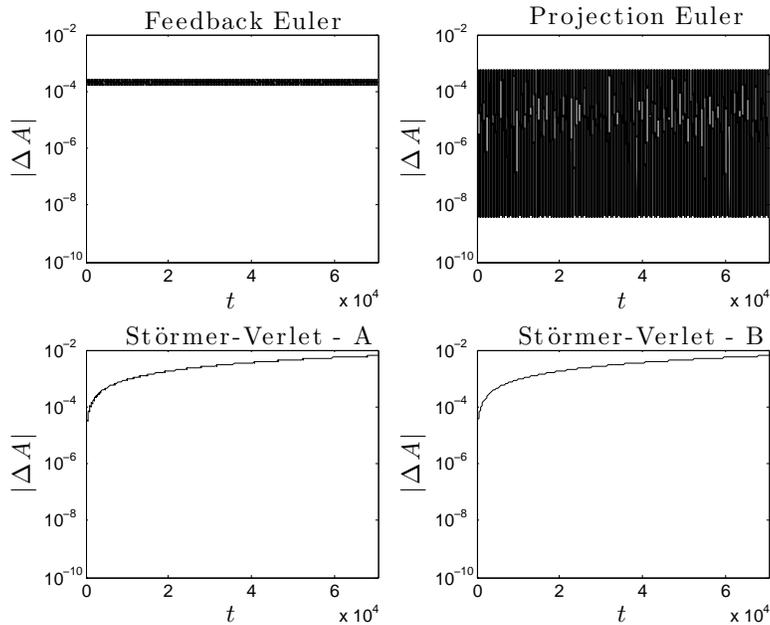


Figure 6: The trajectories of the error $|\Delta A(t)| = |A(t) - A(0)|$, $0 \leq t \leq 70,248$, of the Laplace-Runge-Lenz vector in the Kepler problem generated by four different methods with step size $\Delta t = 0.005$: a feedback integrator with the Euler scheme, the standard projection method with the Euler scheme, and two Störmer-Verlet schemes.

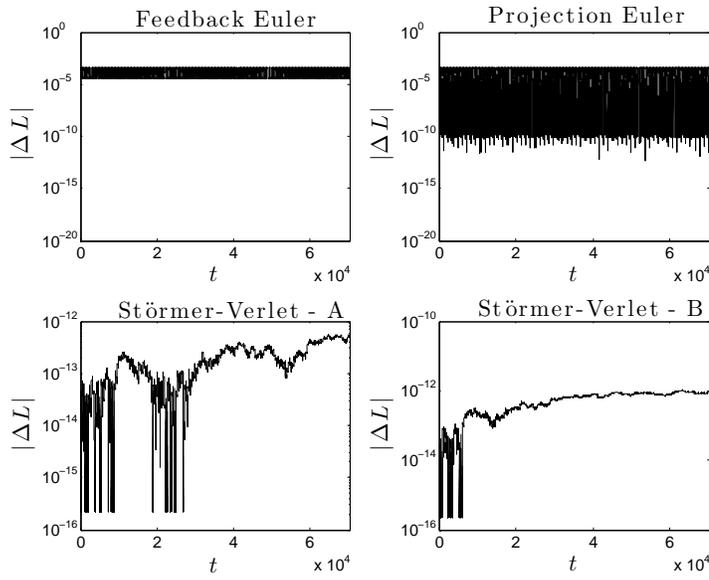


Figure 7: The trajectories of the angular momentum error $|\Delta L(t)| = |L(t) - L(0)|$, $0 \leq t \leq 70,248$, in the Kepler problem generated by four different methods with step size $\Delta t = 0.005$: a feedback integrator with the Euler scheme, the standard projection method with the Euler scheme, and two Störmer-Verlet schemes.

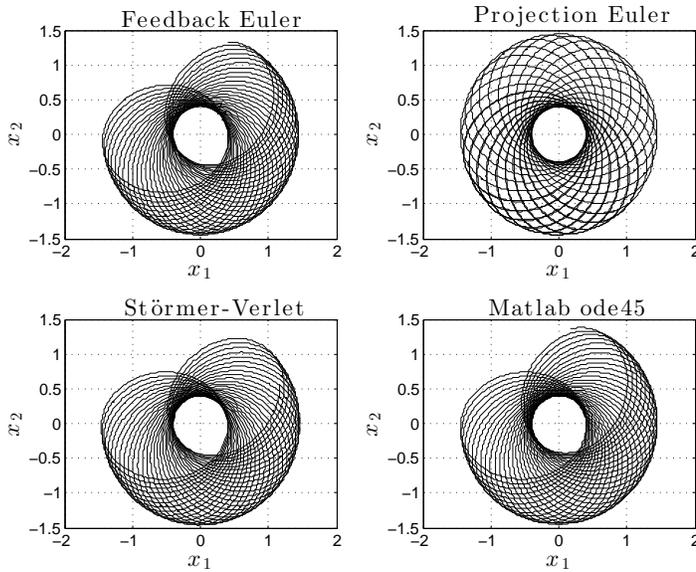


Figure 8: The trajectories of the planar orbit $x(t) = (x_1(t), x_2(t), 0)$, $0 \leq t \leq 200$, in the perturbed Kepler problem generated by four different methods: a feedback integrator with the Euler scheme, the standard projection method with the Euler scheme, a Störmer-Verlet scheme and the Matlab command `ode45`, where the step size $\Delta t = 0.03$ is used for the first three methods.

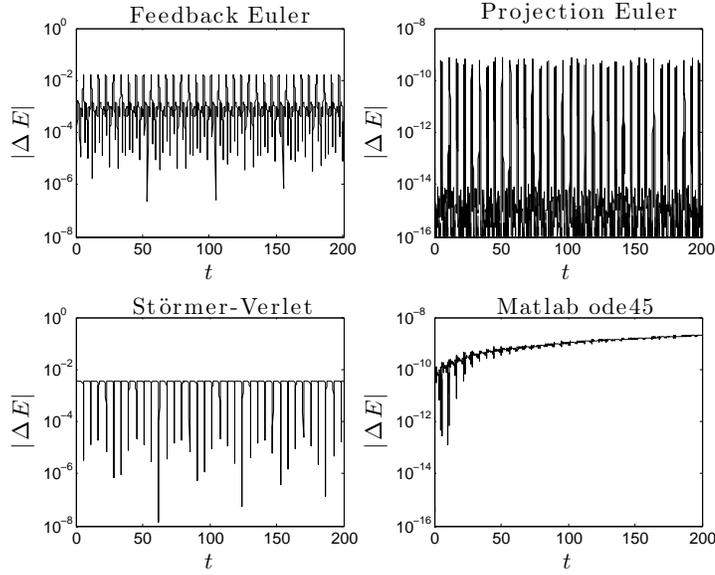


Figure 9: The trajectories of the energy error $|E(t)| = |E(t) - E(0)|$, $0 \leq t \leq 200$, in the perturbed Kepler problem generated by four different methods: a feedback integrator with the Euler scheme, the standard projection method with the Euler scheme, a Störmer-Verlet scheme and the Matlab command *ode45*, where the step size $\Delta t = 0.03$ is used for the first three methods.

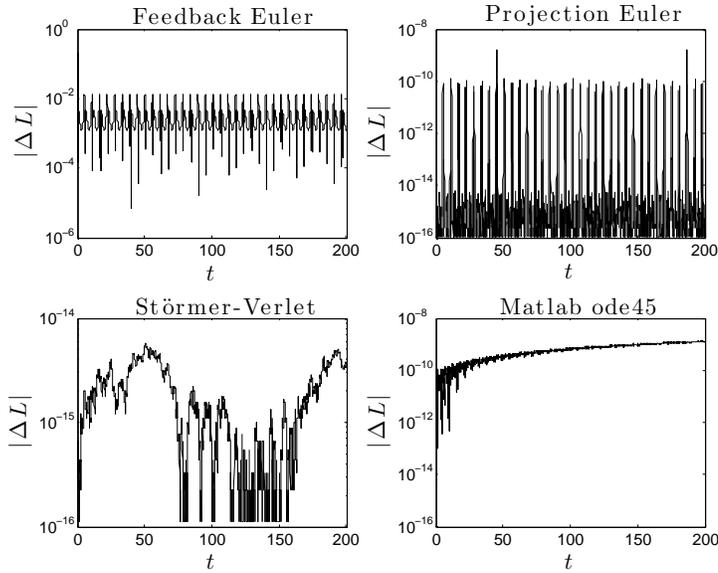


Figure 10: The trajectories of the angular momentum error $|\Delta L(t)| = |L(t) - L(0)|$, $0 \leq t \leq 200$, in the perturbed Kepler problem generated by four different methods: a feedback integrator with the Euler scheme, the standard projection method with the Euler scheme, and a Störmer-Verlet scheme and the Matlab command *ode45*, where the step size $\Delta t = 0.03$ is used for the first three methods.