BIT manuscript No.
(will be inserted by the editor)

On the Positivity of Poisson Integrators for the Lotka-Volterra Equations

Mélanie Beck · **Martin J. Gander**

Received: date / Accepted: date

Abstract Over the last decade, the field of geometric integration has rapidly established itself as one of the core research areas in numerical ordinary differential equations. Geometric integrators are numerical methods which preserve some of the mathematical or physical properties of the system they are approximating. In the case of the Lotka-Volterra equations, which are a Poisson system, a good geometric integrator should also be a Poisson integrator. There is however another important property of solutions of the Lotka-Volterra equations: they are non-negative, since they represent population densities. We study in this paper the conditions under which two Poisson integrators for the Lotka-Volterra equations lead to positive approximate solutions.

Keywords Poisson Integrators · Positivity · Lotka-Volterra equations

Mathematics Subject Classification (2000) 65P10 · 37M15 · 65L20 · 65L70

1 Introduction

1.1 The Lotka-Volterra system

We study numerical methods for solving the Lotka-Volterra equations

$$\begin{cases} \dot{u} = u(b-v), \\ \dot{v} = v(u-a). \end{cases} \tag{1.1}$$

These equations model the evolution of a prey with density $u(t)$ and its predator with density $v(t)$, and a and b are positive constants.

Mélanie Beck
Dawson College, Montreal
E-mail: beck.melanie@gmail.com

Martin J. Gander
University of Geneva
E-mail: martin.gander@unige.ch

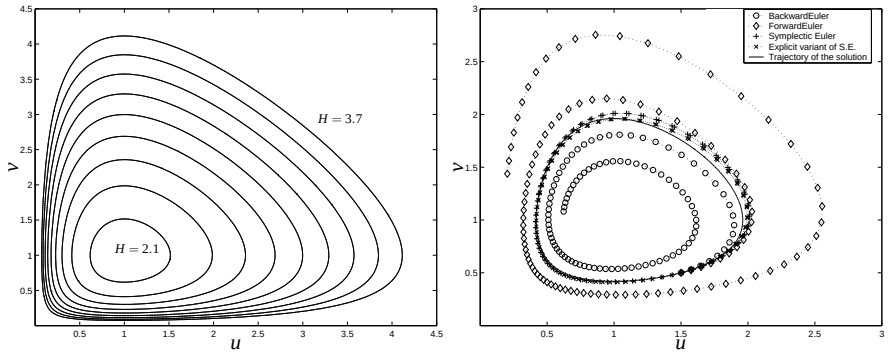

Fig. 1.1 On the left, some level curves of the Hamiltonian of the Lotka-Volterra system, from $H = 2.1$ to $H = 3.7$. On the right, illustration of several methods applied to the Lotka-Volterra system, with $u_0 = 1.5$, $v_0 = 0.5$, $a = b = 1$ and $h = 0.1$.

Defining the function $H : \mathbb{R}^2 \longrightarrow \mathbb{R}$ by

$$H(u,v) := u - a\ln u + v - b\ln v, \tag{1.2}$$

the Lotka-Volterra system (1.1) can be rewritten as

$$\begin{cases} \dot{u} = -uvH_v(u,v), \\ \dot{v} = uvH_u(u,v), \end{cases}$$

where H_u and H_v denote the partial derivatives of H with respect to u and v. Dividing the two equations of this system and separating the variables, we get

$$H_u(u,v)\dot{u} + H_v(u,v)\dot{v} = 0.$$

By integrating this equation, we obtain that the Hamiltonian $H(u,v)$ is an invariant of the system (1.1). This shows that the solution of (1.1) always lies on a level curve of H, and since the level sets of H are closed in the first quadrant, as shown in Figure 1.1, solutions of (1.1) are cyclic. Unlike classical Hamiltonian systems, there is a factor uv in front of the derivatives of the Hamiltonian H, and such systems were called in [4] *non-canonical* Hamiltonian systems. More generally, this system is a Poisson system, see [3]. Note that by applying the transformation $\psi(u,v) = (\ln u, \ln v) = (p,q)$, the Lotka-Volterra system becomes Hamiltonian with $K(p,q) = -H(u,v) = -H(e^p, e^q)$.

It is important that numerical simulations of the system (1.1) show the same qualitative behavior as the exact solution, in particular approximate solutions should also be cyclic and positive. As one can see on Figure 1.1, the results obtained by the forward Euler and the backward Euler methods spiral outwards or inwards, so specific methods have to be used to avoid this.

1.2 Poisson integrators

It is well-known that Hamiltonian systems are symplectic, which in our two dimensional case means area-preserving, and that suitable methods to get good qualitative

behavior are symplectic methods, that is numerical methods satisfying

$$\left(\frac{\partial(u_{n+1},v_{n+1})}{\partial(u_n,v_n)}\right)^T \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \left(\frac{\partial(u_{n+1},v_{n+1})}{\partial(u_n,v_n)}\right) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{1.3}$$

whenever they are applied to a smooth Hamiltonian system.

However, as we said in Section 1.1, the Lotka-Volterra system is not Hamiltonian but its structure is similar to a Hamiltonian system. In fact, the right hand sides are only multiplied by a factor uv. In other words, we can write the Lotka-Volterra system as

$$\dot{\mathbf{y}} = B(\mathbf{y})\nabla H(\mathbf{y}), \tag{1.4}$$

where $\mathbf{y} = (u,v)$, $H(\mathbf{y}) = u - a\ln u + v - b\ln v$ and

$$B(\mathbf{y}) = \begin{pmatrix} 0 & -uv \\ uv & 0 \end{pmatrix}. \tag{1.5}$$

The generalization (1.4) of a Hamiltonian system is called a *Poisson system*.

Definition 1.1 ([3]) If a matrix $B(\mathbf{y})$ is skew-symmetric and satisfies

$$\sum_{l=1}^{n} \left(\frac{\partial b_{ij}(\mathbf{y})}{\partial y_l} b_{lk}(\mathbf{y}) + \frac{\partial b_{jk}(\mathbf{y})}{\partial y_l} b_{li}(\mathbf{y}) + \frac{\partial b_{ki}(\mathbf{y})}{\partial y_l} b_{lj}(\mathbf{y}) \right) = 0, \qquad \text{for all } i,j,k, \tag{1.6}$$

then the formula

$$\{F,G\}(\mathbf{y}) = \sum_{i,j=1}^{n} \frac{\partial F(\mathbf{y})}{\partial y_i} b_{ij}(y) \frac{\partial G(\mathbf{y})}{\partial y_j} \tag{1.7}$$

is said to represent a general *Poisson bracket*. The corresponding differential system (1.4) is a *Poisson system*. We continue to call H the Hamiltonian.

Since the Lotka-Volterra system can be written in the form (1.4), where $B(\mathbf{y})$, defined in (1.5), is skew-symmetric and satisfies (1.6), it is a Poisson system. To study such systems, the notion of Poisson maps is essential.

Definition 1.2 ([3]) A transformation $\varphi : U \to \mathbb{R}^n$ (where U is an open set in $\mathbb{R}^n$) is called a *Poisson map* with respect to the Poisson bracket (1.7), if its Jacobian matrix satisfies

$$\varphi'(\mathbf{y}) B(\mathbf{y}) \varphi'(\mathbf{y})^T = B(\varphi(\mathbf{y})).$$

We observe, of course, a similarity with symplectic maps. The following theorem, whose proof can be found in [3], explains the relation between Poisson systems and Poisson maps.

Theorem 1.1 *([3]) If $B(\mathbf{y})$ is the structure matrix of a Poisson bracket, the flow $\varphi_t(\mathbf{y})$ of the differential system*

$$\dot{\mathbf{y}} = B(\mathbf{y})\nabla H(\mathbf{y})$$

is a Poisson map.

It would of course be interesting to choose numerical methods which exhibit the same characteristics as the flow $\varphi_t(\mathbf{y})$ when solving this kind of problems. This motivates the introduction of the notion of *Poisson integrators*.

Definition 1.3 ([3]) A numerical method $\mathbf{y}_1 = \Phi_h(\mathbf{y}_0)$ is a *Poisson integrator* for the structure matrix $B(\mathbf{y})$, if the transformation $\mathbf{y}_0 \mapsto \mathbf{y}_1$ respects the Casimirs and if it is a Poisson map whenever the method is applied to the corresponding differential system (1.4).

Since for $B(\mathbf{y})$ given by (1.5), there is no Casimir, we do not give the definition here (one is given in [3]), and a numerical method is a Poisson integrator for $B(\mathbf{y})$ if and only if it is a Poisson map whenever applied to the Poisson system (1.4), in other words a numerical method is a Poisson integrator for $B(\mathbf{y})$ given by (1.5) if and only if it satisfies

$$\left(\frac{\partial(u_{n+1},v_{n+1})}{\partial(u_n,v_n)}\right)^T \begin{pmatrix} 0 & -u_nv_n \\ u_nv_n & 0 \end{pmatrix} \left(\frac{\partial(u_{n+1},v_{n+1})}{\partial(u_n,v_n)}\right) = \begin{pmatrix} 0 & -u_{n+1}v_{n+1} \\ u_{n+1}v_{n+1} & 0 \end{pmatrix}. \tag{1.8}$$

The most interesting property of Poisson integrators is related to backward error analysis which is the topic of Section 3.

2 Properties of the Symplectic Euler method and its explicit variant

2.1 Two specific methods

In this paper, we focus on the simplest symplectic method, the symplectic Euler method, and an explicit variant of it. The symplectic Euler method for the system

$$\begin{cases} \dot{u} = f(u,v), \\ \dot{v} = g(u,v), \end{cases} \tag{2.1}$$

is defined in [2] by

$$\begin{cases} u_{n+1} = u_n + h\,f(u_{n+1},v_n), \\ v_{n+1} = v_n + h\,g(u_{n+1},v_n), \end{cases} \tag{2.2}$$

and gives, when applied to the Lotka-Volterra system

$$\begin{cases} u_{n+1} = \frac{u_n}{1-h(b-v_n)}, \\ v_{n+1} = v_n + h\,v_n(u_{n+1}-a). \end{cases} \tag{2.3}$$

An explicit variant of this method, defined by

$$\begin{cases} u_{n+1} = u_n + h\,f(u_n,v_n), \\ v_{n+1} = v_n + h\,g(u_{n+1},v_n), \end{cases} \tag{2.4}$$

was introduced by Gander in [1]. Applied to the Lotka-Volterra system, it becomes

$$\begin{cases} u_{n+1} = u_n + h\,u_n(b-v_n), \\ v_{n+1} = v_n + h\,v_n(u_{n+1}-a). \end{cases} \tag{2.5}$$

Both the symplectic Euler method and its explicit variant are Poisson integrators for the Lotka-Volterra system, more precisely we have the two following theorems.

Proposition 2.1 *The symplectic Euler method* (2.2) *is a Poisson integrator for Poisson systems with $B(\mathbf{y})$ given by* (1.5) *and any separable Hamiltonian H such that $1+hv_n(H_v-u_{n+1}H_{uv})$ is not zero. This condition is always satisfied if h is chosen small enough.*

Proof To prove that the condition (1.8) is satisfied whenever we apply the symplectic Euler method to the Poisson system (1.4) we differentiate

$$\begin{cases} u_{n+1} = u_n - h\,u_{n+1}v_nH_v(u_{n+1},v_n), \\ v_{n+1} = v_n + h\,u_{n+1}v_nH_u(u_{n+1},v_n) \end{cases}$$

with respect to (u_n,v_n) and write the result as a matrix equation

$$\begin{pmatrix} 1+hv_n(H_v-u_{n+1}H_{uv}) & 0 \\ -hv_n(H_u+u_{n+1}H_{uu}) & 1 \end{pmatrix} \begin{pmatrix} \frac{\partial u_{n+1}}{\partial u_n} & \frac{\partial u_{n+1}}{\partial v_n} \\ \frac{\partial v_{n+1}}{\partial u_n} & \frac{\partial v_{n+1}}{\partial v_n} \end{pmatrix} = \begin{pmatrix} 1 & -hu_{n+1}(H_v+v_nH_{vv}) \\ 0 & 1+hu_{n+1}(H_u+v_nH_{vu}) \end{pmatrix},$$

where the matrices H_{uv},H_{uu},H_{vv} of partial derivatives are evaluated at (u_{n+1},v_n). Assuming $1+hv_n(H_v-u_{n+1}H_{uv})$ is not zero, the first matrix is invertible and we can compute

$$\partial\Phi_h\,B(u_n,v_n)\,\partial\Phi_h^T = \begin{pmatrix} 0 & \frac{-u_nv_n(1+hu_{n+1}(H_u+v_nH_{vu}))}{1+hv_n(H_v-u_{n+1}H_{uv})} \\ \frac{u_nv_n(1+hu_{n+1}(H_u+v_nH_{vu}))}{1+hv_n(H_v-u_{n+1}H_{uv})} & 0 \end{pmatrix}.$$

Therefore the symplectic Euler method is a Poisson integrator for $B(\mathbf{y})$ if

$$\frac{u_nv_n(1+hu_{n+1}(H_u+v_nH_{vu}))}{1+hv_n(H_v-u_{n+1}H_{uv})} = u_{n+1}v_{n+1}.$$

Replacing u_{n+1} and v_{n+1} by their definitions we obtain the condition

$$H_{uv}(1+hv_nH_v) = -H_{uv}(1+hu_{n+1}H_u),$$

which is satisfied for any separable Hamiltonian $H(u,v)=T(u)+S(v)$.

Proposition 2.2 *The explicit variant of the symplectic Euler method* (2.4) *is a Poisson integrator for Poisson systems with $B(\mathbf{y})$ defined in* (1.5) *and any separable Hamiltonian H.*

Proof To prove that the method is a Poisson integrator, we proceed as above and get the matrix equation

$$\begin{pmatrix} 1 & 0 \\ -hv_n(H_u+u_{n+1}H_{uu}) & 1 \end{pmatrix} \begin{pmatrix} \frac{\partial u_{n+1}}{\partial u_n} & \frac{\partial u_{n+1}}{\partial v_n} \\ \frac{\partial v_{n+1}}{\partial u_n} & \frac{\partial v_{n+1}}{\partial v_n} \end{pmatrix} = \begin{pmatrix} 1-hv_n(H_v+u_nH_{uv}) & -hu_n(H_v+v_nH_{vv}) \\ 0 & 1+hu_{n+1}(H_u+v_nH_{vu}) \end{pmatrix},$$

where the matrices H_{vu} and H_{uu} are evaluated at (u_{n+1}, v_n), whereas the matrices H_{uv} and H_{vv} are evaluated at (u_n, v_n). Since the first matrix is invertible, we can compute the matrix of derivatives $\partial\Phi_h$ and we get

$$\partial\Phi_h B(u_n, v_n)\,\partial\Phi_h^T = \begin{pmatrix} 0 & A \\ -A & 0 \end{pmatrix},$$

with $A := u_n v_n(1 - hv_n(H_v + u_n H_{uv}))(1 + hu_{n+1}(H_u + v_n H_{uv}))$. Since

$$u_{n+1}v_{n+1} = u_n v_n(1 - hv_n H_v)(1 + hu_{n+1}H_u),$$

and

$$A = u_n v_n(1 - hv_n H_v - hv_n u_n H_{uv}))(1 + hu_{n+1}H_u + hu_{n+1}v_n H_{uv})),$$

the explicit variant of the symplectic Euler method is a Poisson integrator for $B(\mathbf{y})$ defined in (1.5) and any separable Hamiltonian.

Since both methods are Poisson integrators for the Lotka-Volterra system, we expect them to give good numerical results. This explains the excellent performance we observed on Figure 1.1.

Apart from the fact that they are Poisson integrators, an interesting property would be that for suitable values of the step-size, the numerical result stays in the first quadrant. We give such a result for the symplectic Euler method in Section 2.2. For the explicit variant of the symplectic Euler method, it is much more difficult to obtain such a result, which will be the main part of this paper.

2.2 Condition for positivity of the symplectic Euler method

For the symplectic Euler method applied to the Lotka-Volterra system, a very simple condition yields the desired result.

Theorem 2.1 *If we apply the symplectic Euler method* (2.3) *to the Lotka-Volterra system with h smaller than* $1/a$ *and* $1/b$*, the numerical result stays in the first quadrant, that is* u_n *and* v_n *are positive for any n.*

Proof Assuming u_n and v_n are positive, u_{n+1} is positive if and only if $1 - h(b - v_n)$ is positive, that is

$$v_n > b - \frac{1}{h}.$$

For positive v_n, it is sufficient that h is smaller than $1/b$ to ensure the positivity of u_{n+1}. This also guarantees that the denominator of the first equation in (2.3) never vanishes. On the other hand, v_{n+1} is positive if and only if $1 + h(u_{n+1} - a)$ is positive, i.e.

$$u_{n+1} > a - \frac{1}{h}. \tag{2.6}$$

So if h is smaller than $1/b$ (so that u_{n+1} is positive) and smaller than $1/a$, the inequality (2.6) is satisfied and v_{n+1} is positive.

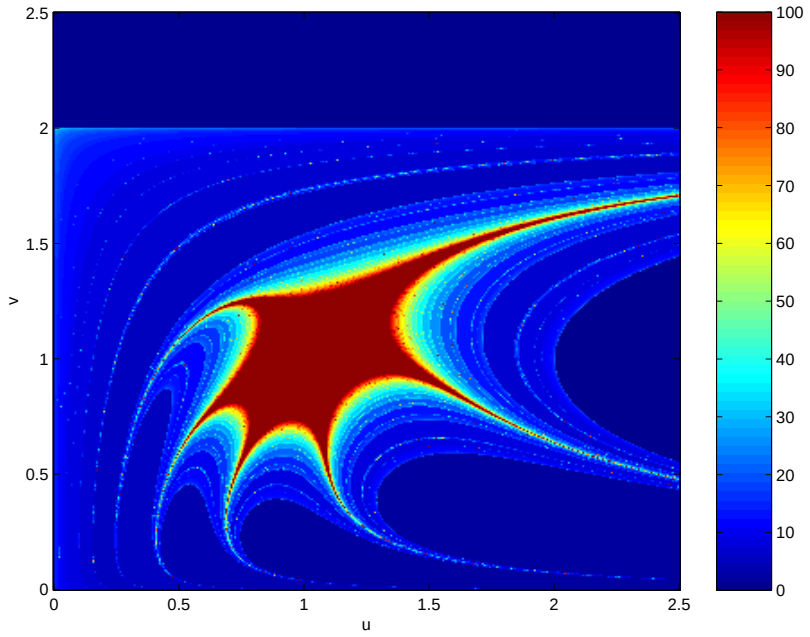

Fig. 2.1 Number of iterations needed for each point (u_0, v_0) to leave the first quadrant when applying the explicit variant of the symplectic Euler method to the Lotka-Volterra system with h=1 and $a = b = 1$.

On the other hand, if we want to apply the same argument to the explicit variant of the method, we obtain the same argument for v_{n+1}, but for u_{n+1} we get

$$u_{n+1} > 0 \qquad \Leftrightarrow \qquad v_n < b + \frac{1}{h}$$

and because this condition is not always satisfied, we can not predict in a simple way when a numerical result will stay in the first quadrant and when it will not.

Actually if we plot for $h = 1$ the number of iterations needed to leave the first quadrant for each initial value, the figure obtained, Figure 2.1, is aesthetically pleasing and very complicated. One can note that the condition

$$v_1 < b + \frac{1}{h} = 2$$

appears clearly in the figure. This figure indicates that for given initial conditions, it is always possible to find a step-size h for which the numerical results stay positive. We will use backward error analysis in order to prove such a result for exponentially long-time intervals.

3 Backward Error Analysis

The most interesting property of Poisson integrators is related to backward error analysis. For a given numerical equation $y_{n+1} = \Phi_h(y_n)$ applied to the differential equa-

tion $\dot{y} = f(y)$, we call *modified equation* the differential equation $\dot{\tilde{y}} = f_h(\tilde{y})$ of the form

$$\dot{\tilde{y}} = f(\tilde{y}) + h f_2(\tilde{t}) + h^2 f_3(\tilde{y}) + \ldots,$$

such that $y_n = \tilde{y}(nh)$. If we apply a Poisson integrator to a Poisson system, the modified equation is itself a Poisson system. The proof of these two results can be found in [3].

Theorem 3.1 *([3]) If a Poisson integrator $\Phi_h(\mathbf{y})$ is applied to the Poisson system (1.4), then the modified equation is locally a Poisson system. More precisely, for every $\mathbf{y}_0 \in \mathbb{R}^n$ there exists a neighborhood U and smooth functions $H_j : U \to \mathbb{R}$ such that on U, the modified equation is of the form*

$$\dot{\tilde{\mathbf{y}}} = B(\tilde{\mathbf{y}})(\nabla H(\tilde{\mathbf{y}}) + h\nabla H_2(\tilde{\mathbf{y}}) + \ldots). \tag{3.1}$$

This result, which is only considering the local structure of the modified equation, can be made more global under additional conditions on the differential equation.

Theorem 3.2 *([3]) If $H(\mathbf{y})$ and $B(\mathbf{y})$ are defined and smooth on a simply connected domain D, and if $B(\mathbf{y})$ is invertible on D, then a Poisson integrator $\Phi_h(\mathbf{y})$ has a modified equation* (3.1) *with smooth functions $H_j(\mathbf{y})$ defined on all of D.*

Since for the Lotka-Volterra system, the matrix $B(\mathbf{y})$ is invertible on $\mathbb{D} := \{\mathbf{y} = (u,v) : u > 0, v > 0\}$, whatever Poisson integrator you use to solve it, the modified equation is globally a Poisson system. We usually call the Hamiltonian of the modified system the *numerical Hamiltonian* of the original system and denote it by $\tilde{H}(\mathbf{y})$.

Following Sections IX.7 and IX.8 of [3], we bound the local error of the numerical result and using the above property of Poisson integrators, we bound the Hamiltonian error. This bound allows us to state our main theorem : given initial conditions, we can compute for which step-size values the numerical solution given by the explicit variant of the symplectic Euler method is ensured to remain positive over exponentially long time intervals.

3.1 Estimation of the Numerical Solution

Recall that the Lotka-Volterra system is a Poisson system whose Hamiltonian $H(u,v) = u - a\ln u + v - b\ln v$ is analytic on $E \times E$, where $E := \mathbb{C} \setminus \{z \,|\, \mathrm{Re}(z) \leq 0 \text{ and } \mathrm{Im}(z) = 0\}$. We shall denote for the rest of the chapter $\mathbf{y} := (u,v)$ and $\mathbf{f}_1(\mathbf{y}) := (u(b-v), v(u-a))^T$. We apply to the system the explicit variant of the symplectic Euler method $\Phi_h(\mathbf{y})$, defined in (2.5), with step size h.

We fix a compact set $K \subset \mathbb{D} := \{(u,v) \in \mathbb{R}^2 \,|\, u > 0, v > 0\}$ and define

$$R := \alpha \ \mathrm{distance}(K, \mathbb{D}^c),$$

where $0 < \alpha < 1$ can be arbitrarily chosen (we usually obtain better results if α is large), so that for all $\mathbf{y}_0 \in K$ and all $\mathbf{y}$ such that $\|\mathbf{y} - \mathbf{y}_0\| \leq R$, $\mathbf{y}$ belongs to $\mathbb{D}$, see Figure 3.1. Denoting by $\tilde{K}$ the compact set $\tilde{K} := \{\mathbf{y} \,|\, \exists \mathbf{y}_0 \in K \text{ such that } \|\mathbf{y} - \mathbf{y}_0\| \leq R\}$

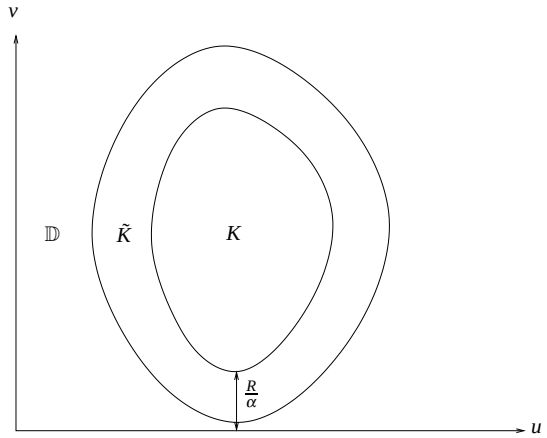

Fig. 3.1 Illustration of the sets K and $\tilde{K}$.

and by M the bound $M := \max\{\|\mathbf{f}_1(\mathbf{y})\| : \mathbf{y} \in \tilde{K}\}$, we now have, for all $\mathbf{y}_0 \in K$,

$$\|\mathbf{f}_1(\mathbf{y})\| \le M \qquad \text{for} \qquad \|\mathbf{y}-\mathbf{y}_0\| \le R. \tag{3.2}$$

Note that we can write the explicit variant of the symplectic Euler method applied to the Lotka-Volterra system as

$$\Phi_h(\mathbf{y}) = \mathbf{y} + h\mathbf{f}_1(\mathbf{y}) + h^2\mathbf{d}_2(\mathbf{y}), \tag{3.3}$$

where $\mathbf{d}_2(\mathbf{y}) = (0, uv(b-v))^T$ is analytic. Since we have to bound $\mathbf{d}_2(\mathbf{y})$ for $\mathbf{y} \in \tilde{K}$, we simply define $M_2 := \max\{|uv(b-v)| : (u,v) \in \tilde{K}\}$ so that for all $\mathbf{y}_0 \in K$,

$$\|\mathbf{d}_2(\mathbf{y})\| = |uv(b-v)| \le M_2 \qquad \text{for} \qquad \|\mathbf{y}-\mathbf{y}_0\| \le R. \tag{3.4}$$

3.2 Estimation of the Coefficients of the Modified Equation

The next step is to bound the functions $\mathbf{f}_j$ of the modified equation

$$\dot{\tilde{\mathbf{y}}} = \sum_{n\ge 1} h^{n-1}\mathbf{f}_n(\tilde{\mathbf{y}}) = \mathbf{f}_1(\tilde{\mathbf{y}}) + h\mathbf{f}_2(\tilde{\mathbf{y}}) + h^2\mathbf{f}_3(\tilde{\mathbf{y}}) + \dots. \tag{3.5}$$

The key idea to obtain an explicit formula for these functions is to introduce the Lie derivative

$$D_j = \sum_k \mathbf{f}_j^{[k]}(\mathbf{y})\frac{\partial}{\partial \mathbf{y}^{[k]}},$$

where $\mathbf{y}^{[k]}$ denotes the kth component of the vector $\mathbf{y}$; in particular, for any differentiable function $\mathbf{g}$,

$$D_j\mathbf{g}(\mathbf{y}) = \mathbf{g}'(\mathbf{y})\mathbf{f}_j(\mathbf{y}).$$

Using Lie derivatives and denoting $\mathbf{y} := \tilde{\mathbf{y}}(t)$, we can write the solution of the modified equation (3.5) expanded into a Taylor series as

$$\tilde{\mathbf{y}}(t+h) = \mathbf{y} + h\mathbf{f}_1(\mathbf{y}) + h^2[\mathbf{f}_2(\mathbf{y}) + \frac{1}{2!}(D_1\mathbf{f}_1)(\mathbf{y})] + h^3[\mathbf{f}_3(\mathbf{y}) + \frac{1}{2!}(D_1\mathbf{f}_2 + D_2\mathbf{f}_1)(\mathbf{y}) + \frac{1}{3!}(D_1^2\mathbf{f}_1)(\mathbf{y})] + \ldots.$$

In other words, the solution of the modified equation (3.5), with initial value $\mathbf{y}(t) = \mathbf{y}$ can be formally written as

$$\tilde{\mathbf{y}}(t+h) = \mathbf{y} + \sum_{i\geq 1} \frac{h^i}{i!} D^{i-1}\mathbf{F}(\mathbf{y}),$$

where $\mathbf{F}(\mathbf{y}) = \sum_{n\geq 1} h^{n-1}\mathbf{f}_n(\mathbf{y})$ stands for the modified equation, and $hD = \sum_{n\geq 1} h^n D_n$ for the corresponding Lie operator.

Expanding the formal sum, we obtain

$$\tilde{\mathbf{y}}(t+h) = \mathbf{y} + \sum_{i\geq 1} \frac{1}{i!} \left[\sum_{k_1,\ldots,k_i \geq 1} h^{k_1+\cdots+k_i} (D_{k_1} \ldots D_{k_{i-1}} \mathbf{f}_{k_i})(\mathbf{y}) \right], \tag{3.6}$$

and then we can compare like powers of h in the numerical method (3.3) and the expansion of the exact solution (3.6) to obtain

$$\sum_{i\geq 1} \frac{1}{i!} \left[\sum_{k_1+\cdots+k_i=2} (D_{k_1} \ldots D_{k_{i-1}} \mathbf{f}_{k_i})(\mathbf{y}) \right] = \mathbf{d}_2(\mathbf{y})$$

and for all $j \geq 3$,

$$\sum_{i\geq 1} \frac{1}{i!} \left[\sum_{k_1+\cdots+k_i=j} (D_{k_1} \ldots D_{k_{i-1}} \mathbf{f}_{k_i})(\mathbf{y}) \right] = 0.$$

In other words,

$$\mathbf{f}_2(\mathbf{y}) = \mathbf{d}_2(\mathbf{y}) - \frac{1}{2}(D_1\mathbf{f}_1)(\mathbf{y}), \tag{3.7}$$

and

$$\mathbf{f}_j(\mathbf{y}) = -\sum_{i=2}^{j} \frac{1}{i!} \left[\sum_{k_1+\cdots+k_i=j} (D_{k_1} \ldots D_{k_{i-1}} \mathbf{f}_{k_i})(\mathbf{y}) \right], \tag{3.8}$$

for $j \geq 3$, so if we want to get bounds for $\|\mathbf{f}_j(\mathbf{y})\|$, we have to estimate first $\|(D_j\mathbf{g})(\mathbf{y})\|$ and for this we use the following variant of Cauchy's estimate proved in [3].

Lemma 3.1 *For analytic functions $\mathbf{f}_j(\mathbf{y})$ and $\mathbf{g}(\mathbf{y})$ we have for $0 \leq \sigma < \rho$ the estimate*

$$\|D_j\mathbf{g}\|_\sigma \leq \frac{1}{\rho - \sigma} \cdot \|\mathbf{f}_j\|_\sigma \cdot \|\mathbf{g}\|_\rho.$$

Here, $\|\mathbf{g}\|_\rho := \max\{\|\mathbf{g}(\mathbf{y})\| : \mathbf{y} \in B_\rho(\mathbf{y}_0)\}$ and $\|\mathbf{f}_j\|_\sigma, \|D_j\mathbf{g}\|_\sigma$ are defined similarly.

The following theorem gives an explicit bound for the functions $\mathbf{f}_j(\mathbf{y})$, for $\mathbf{y} \in \tilde{K}_2 := \{\, \mathbf{y} \,|\, \exists \mathbf{y}_0 \in K \text{ s.t. } \|\mathbf{y}-\mathbf{y}_0\| \le R/2 \,\}$. Note that this bound is only valid when we apply the explicit variant of the symplectic Euler method to the Lotka-Volterra system.

Lemma 3.2 *For all $\mathbf{y}_0 \in K$, $\mathbf{f}(\mathbf{y})$ and the coefficients $\mathbf{f}_j(\mathbf{y})$ of the modified differential equation (3.5) are analytic in $B_R(\mathbf{y}_0)$, so if the bounds (3.2) and (3.4) are satisfied, we have for the coefficients $\mathbf{f}_j$, $j \ge 2$,*

$$\|\mathbf{f}_j(\mathbf{y})\| \le \ln 2 \, \frac{\eta M}{2} \left(\frac{\eta M (j-1)}{R} \right)^{j-1} \qquad \textit{for} \qquad \mathbf{y} \in \tilde{K}_2, \tag{3.9}$$

where $\eta := 2/(2\ln 2 - 1) + M_2 R/M^2$.

Proof We fix an index $J > 1$ and we consider $\|\mathbf{f}_J\|_{R/2} = \max\{\, \|\mathbf{f}_J(\mathbf{y})\| : \mathbf{y} \in B_{R/2}(\mathbf{y}_0)\}$. The trick of the proof is to introduce $\delta := R/(2(J-1))$ and to estimate $\|\mathbf{f}_j\|_{R-(j-1)\delta}$.

In order to simplify the notation, we abbreviate $\|\cdot\|_{R-(j-1)\delta}$ by $\|\cdot\|_j$. Applying repeatedly Cauchy's estimate given in Lemma 3.1, we obtain for $k_1 + \cdots + k_i = j$,

$$\|D_{k_1} D_{k_2} \ldots D_{k_{i-1}} \mathbf{f}_{k_i}\|_j \le \frac{1}{\delta^{i-1}} \|\mathbf{f}_{k_1}\|_j \cdot \|\mathbf{f}_{k_2}\|_{j-1} \cdot \ldots \cdot \|\mathbf{f}_{k_{i-1}}\|_{j-i+2} \cdot \|\mathbf{f}_{k_i}\|_{j-i+1}.$$

By definition, for $k < j$ we have $B_{R-(j-1)\delta} \subset B_{R-(k-1)\delta}$, so that $\|\mathbf{g}\|_j \le \|\mathbf{g}\|_k$, so from $k_1, k_2, \ldots, k_i \le j-i+1$, we obtain

$$\|D_{k_1} D_{k_2} \ldots D_{k_{i-1}} \mathbf{f}_{k_i}\|_j \le \frac{1}{\delta^{i-1}} \|\mathbf{f}_{k_1}\|_{k_1} \cdot \|\mathbf{f}_{k_2}\|_{k_2} \cdot \ldots \cdot \|\mathbf{f}_{k_{i-1}}\|_{k_{i-1}} \cdot \|\mathbf{f}_{k_i}\|_{k_i}.$$

We now apply this inequality to the expansions of the functions $\mathbf{f}_j$ given by (3.7) and (3.8) and obtain

$$\|\mathbf{f}_2\|_2 \;\le\; \|\mathbf{d}_2\|_2 + \frac{1}{2}\|D_1 \mathbf{f}_1\|_2 \;\le\; \|\mathbf{d}_2\|_2 + \frac{1}{2\delta}\|\mathbf{f}_1\|_1^2,$$

and

$$\begin{aligned} \|\mathbf{f}_\mathbf{j}\|_j &\le \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\cdots+k_i=j} \|D_{k_1} \ldots D_{k_{i-1}} \mathbf{f}_{k_i}\|_j \\ &\le \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\cdots+k_i=j} \frac{1}{\delta^{i-1}} \|\mathbf{f}_{k_1}\|_{k_1} \cdot \|\mathbf{f}_{k_2}\|_{k_2} \cdot \ldots \cdot \|\mathbf{f}_{k_i}|_{k_i}. \end{aligned}$$

We define, by induction,

$$\beta_j = \frac{M}{\delta} \left(\frac{M_2}{M} \right)^{j-1} + \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\cdots+k_i=j} \beta_{k_1} \beta_{k_2} \ldots \beta_{k_i},$$

so that $\|\mathbf{f}_j\|_j \leq \beta_j \delta$, for $1 \leq j \leq J$, and we consider the generating function

$$\begin{aligned} b(\zeta) &= \sum_{j\geq 1} \beta_j \zeta^j \\ &= \sum_{j\geq 1} \frac{M}{\delta}\left(\frac{M_2}{M}\right)^{j-1} \zeta^j + \sum_{j\geq 2}\sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\cdots+k_i=j} \beta_{k_1}\beta_{k_2}\ldots\beta_{k_i}\zeta^j \\ &= \frac{M\zeta}{\delta}\sum_{j\geq 1}\left(\frac{\zeta M_2}{M}\right)^{j-1} + \sum_{j\geq 2}\frac{1}{j!}(b(\zeta))^j. \end{aligned}$$

Denoting by $\gamma := M/\delta$ and $q := M_2/M$ and assuming $|q\zeta| < 1$ we obtain

$$b(\zeta) = \frac{\gamma\zeta}{1-q\zeta} + e^{b(\zeta)} - b(\zeta) - 1.$$

So denoting by $x := b(\zeta)$, we have to solve

$$\Phi(\zeta, x) := e^x - 2x - 1 + \frac{\gamma\zeta}{1-q\zeta} = 0. \tag{3.10}$$

We can apply the implicit function theorem whenever $e^x = e^{b(\zeta)} \neq 2$ i.e. $x \neq B := \ln 2 + 2k\pi i$, so we need

$$e^{b(\zeta)} - 2b(\zeta) = 1 - \frac{\gamma\zeta}{1-q\zeta} \neq 2 - 2B.$$

Solving this last equation, we obtain that

$$\zeta \neq \frac{2B-1}{\gamma + q(2B-1)}.$$

So finally $b(\zeta)$ is analytic in a disc with radius $\frac{1}{\nu} := \frac{2\ln 2-1}{\gamma+q(2\ln 2-1)}$ centered at the origin. One can note that since $\frac{1}{\nu} < \frac{1}{q}$, the sum $\sum_{j\geq 0}(q\zeta)^j$ in the derivation of $b(\zeta)$ is well defined.

Now we want to prove that on the disc $|\zeta| < \frac{1}{\nu}$, the solution $b(\zeta)$ of (3.10) with $b(0) = 0$ is bounded by $\ln 2$. Applying the conformal maps $\zeta \mapsto \frac{1}{\zeta}$, $\zeta \mapsto \zeta - q$ and again $\zeta \mapsto \frac{1}{\zeta}$ to the disc $|\zeta| < \frac{1}{\nu}$ we obtain the disc of radius $\frac{\nu}{|q^2-\nu^2|}$ centered at $\frac{-q}{q^2-\nu^2}$ if $q \neq \nu$ (which is always the case since γ cannot be zero). Finally we multiply by $(-\gamma)$ and obtain the disc centered at $\frac{\gamma q}{q^2-\nu^2}$ and of radius $\frac{\gamma\nu}{|q^2-\nu^2|}$. Since $\nu > q$, the center of the disc is negative and the largest point of the disc is

$$w^- = \frac{\gamma q}{q^2-\nu^2} + \frac{\gamma\nu}{q^2-\nu^2} = \frac{\gamma}{q-\nu} = -(2\ln 2 - 1).$$

So now we have to consider the image of the disc $|w| \leq 2\ln 2 - 1$ centered at the origin under the mapping $b(w)$ defined by $e^b - 1 - 2b = w$ and $b(0) = 0$. One can prove (see, for example, [3] page 309) that it is completely contained in the disc $|b| \leq \ln 2$.

Applying Cauchy's estimate to $b(\zeta) = \sum_{j\geq 1} \beta_j \zeta^j$, we now obtain

$$|\beta_j| = \left| \frac{b^{(j)}(0)}{j!} \right| \leq \ln 2 \, \nu^j,$$

and thus $\|\mathbf{f}_J\|_{R/2} = \|\mathbf{f}_J\|_J \leq \delta\beta_J \leq \ln 2 \, \delta \, \nu^J$. By definition of ν, we have

$$\nu = \frac{M(J-1)}{R}\left(\frac{2}{2\ln 2 - 1} + \frac{RM_2}{M^2(J-1)}\right) \leq \frac{M(J-1)}{R}\left(\frac{2}{2\ln 2 - 1} + \frac{RM_2}{M^2}\right),$$

so defining $\eta := 2/(2\ln 2 - 1) + RM_2/M^2$, we obtain

$$\delta\nu = \frac{R}{2(J-1)}\nu \leq \frac{R}{2(J-1)}\frac{M(J-1)\eta}{R} = \frac{M\eta}{2},$$

so that we get the result for $J > 1$.

3.3 Estimation of the Local Error

Since the modified differential equation series usually diverges, we have to work with a truncated equation

$$\dot{\tilde{\mathbf{y}}} = \mathbf{F}_N(\tilde{\mathbf{y}}), \qquad \mathbf{F}_N(\tilde{\mathbf{y}}) = \mathbf{f}(\tilde{\mathbf{y}}) + h\mathbf{f}_2(\tilde{\mathbf{y}}) + \cdots + h^{N-1}\mathbf{f}_N(\tilde{\mathbf{y}}), \tag{3.11}$$

with initial value $\tilde{\mathbf{y}_0} = \mathbf{y}_0$. Supposing that $hN \leq h_0$ with $h_0 := \frac{R}{e\eta M}$ and using the bound (3.9), we estimate for $\mathbf{y} \in \tilde{K_2}$,

$$\|\mathbf{F_N}\| \leq \|\mathbf{f}\| + h\|\mathbf{f}_2\| + \cdots + h^{N-1}\|\mathbf{f}_N\| \leq M\left[1 + \eta\frac{\ln 2}{2}\sum_{j=1}^{N-1}\left(\frac{j}{eN}\right)^j\right],$$

and since the sum in the last line is maximal for $N = 2$ and bounded by 0.184, we obtain

$$\|\mathbf{F}_N\| \leq M\left[1 + 1.0022\,\eta\frac{\ln 2}{2}\right] \leq M\left[1 + 0.064\eta\right]. \tag{3.12}$$

This estimation allows us to bound the local error.

Lemma 3.3 *If $h \leq h_0/3$ with $h_0 = R/(e\eta M)$, then there exists $N = N(h)$ (namely N equal to the largest integer satisfying $hN \leq h_0$) such that, for any $y_0 \in K$, the difference between the numerical solution $\mathbf{y}_1 = \Phi_h(\mathbf{y}_0)$ and the exact solution $\tilde{\phi}_{N,t}(\mathbf{y}_0)$ of the truncated modified equation (3.11) satisfies*

$$\|\Phi_h(\mathbf{y_0}) - \tilde{\phi}_{N,h}(\mathbf{y}_0)\| \leq h\gamma M e^{-h_0/h},$$

where $\gamma = e(2 + \frac{eh_0M_2}{3M} + 0.064\eta)$.

Proof For any $\mathbf{y}_0 \in K$ fixed, we consider the analytic function

$$\mathbf{g}(h) := \Phi_h(\mathbf{y}_0) - \tilde{\phi}_{N,h}(\mathbf{y}_0).$$

By definition of the functions $\mathbf{f}_j(\mathbf{y})$ of the modified equation, the coefficients of the Taylor series for $\Phi_h(\mathbf{y}_0)$ and $\tilde{\phi}_{N,h}$ are the same up to the h^N-term, but not further due to the truncation of the modified equation. Hence the function $\mathbf{g}(h)$ contains the factor h^{N+1} and we can apply the maximum principle for analytic functions to $\frac{\mathbf{g}(h)}{h^{N+1}}$. If $\mathbf{g}(z)$ is analytic for $|z| \le \varepsilon$, we have for $0 \le h \le \varepsilon$,

$$\left\| \frac{\mathbf{g}(h)}{h^{N+1}} \right\| \le \frac{1}{\varepsilon^{N+1}} \max_{|z|\le\varepsilon} \|\mathbf{g}(z)\|. \tag{3.13}$$

Since $\mathbf{g}(h)$ is analytic for any h, we can choose $\varepsilon = eh_0/N$.

On the other hand we have

$$\|\Phi_z(\mathbf{y}_0) - \mathbf{y}_0\| = \|z\mathbf{f}(\mathbf{y}_0) + z^2\mathbf{d}_2(\mathbf{y}_0)\| \le |z|M + |z|^2 M_2.$$

Moreover $\|\mathbf{F}_N\| \le M(1+0.064\eta)$ is valid for any $\mathbf{y} \in \tilde{K}_2$ and any $|h| \le \varepsilon$, so we have

$$\|\tilde{\phi}_{N,z}(\mathbf{y}_0) - \mathbf{y}_0\| = \|\tilde{\mathbf{y}}(z) - \tilde{\mathbf{y}}(0)\| \le |z| \cdot \|\dot{\tilde{\mathbf{y}}}(z)\| = |z| \cdot \|\mathbf{F}_N(\tilde{\mathbf{y}})\| \le |z|M(1+0.064\eta),$$

as long as $\tilde{\phi}_{N,z}(\mathbf{y}_0) = \tilde{\mathbf{y}}(z)$ stays in $\tilde{K}_2$. In fact, because

$$\varepsilon M(1+0.064\eta) = \frac{eh_0}{N} M(1+0.064\eta) = \frac{R}{N}\left(\frac{1}{\eta} + 0.064\right) \le \frac{R}{2},$$

since $\eta \ge 5$ and $N \ge 3$, the solution $\tilde{\phi}_{N,z}$ stays in the ball $B_{R/2}(\mathbf{y}_0) \subset \tilde{K}_2$ for all $|z| \le \varepsilon$.

Finally we go back to (3.13). Since

$$\|\mathbf{g}(z)\| \le \|\Phi_z(\mathbf{y}_0) - \mathbf{y}_0\| + \|\tilde{\phi}_{N,z}(\mathbf{y}_0) - \mathbf{y}_0\| \le |z|M\left(1 + |z|\frac{M_2}{M} + 1 + 0.064\eta\right),$$

we have

$$\begin{aligned} \|\mathbf{g}(h)\| &\le \frac{h^{N+1}}{\varepsilon^{N+1}} \max_{|z|\le\varepsilon} \left[|z|M\left(2 + |z|\frac{M_2}{M} + 0.064\eta\right) \right] \\ &\le \left(\frac{hN}{eh_0}\right)^N hM\left[2 + \frac{eh_0M_2}{3M} + 0.064\eta\right]. \end{aligned}$$

Then, because $hN \le h_0$, we obtain

$$\|\mathbf{g}(h)\| \le e^{-N} hM\left[2 + \frac{eh_0M_2}{3M} + 0.064\,\eta\right].$$

Finally, since $N \le h_0/h < N+1$, we have $e^{-h_0/h} \ge e^{-(N+1)}$ and the theorem is proved.

3.4 Estimates of the Variation of the Hamiltonian

We are now in a position to prove that if the numerical result stays in a compact set, then it is really close to the exact trajectory for exponentially long time intervals.

Lemma 3.4 *If the numerical solution stays in the compact set $\tilde{K}_2 \subset \mathbb{D} = \{u > 0, v > 0\}$ and if $h \le h_0/3$, with $h_0 = R/(e\eta M)$, then there exists $N = N(h)$ (the largest integer satisfying $hN \le h_0$) such that, over exponentially long time intervals $nh \le e^{h_0/2h}$,*

$$\begin{aligned} |\tilde{H}(\mathbf{y}_n) - \tilde{H}(\mathbf{y}_0)| &\le L\gamma M e^{-h_0/2h}, \\ |H(\mathbf{y}_n) - H(\mathbf{y}_0)| &\le L\,\gamma M e^{-h_0/2h} + 2hC, \end{aligned} \tag{3.14}$$

with $L := \frac{M(1+0.064\,\eta)}{(u_{\min}-R/2)(v_{\min}-R/2)}$ and $C := \frac{0.277M^2\eta^2}{(u_{\min}-R/2)(v_{\min}-R/2)}$, where $u_{\min} := \min\{u : (u,v) \in K\}$ and $v_{\min}$ is defined in a similar way.

Proof Let $\tilde{\phi}_{N,t}(\mathbf{y}_0)$ be the flow of the truncated modified equation (3.5). As stated in Theorem 3.2, this differential equation is a Poisson system whose Hamiltonian is $\tilde{H} = H + hH_2 + h^2H_3 + \cdots + h^{N-1}H_N$, so that

$$\tilde{H}(\tilde{\phi}_{N,t}(\mathbf{y}_0)) = \tilde{H}(\mathbf{y}_0), \qquad \forall\, t.$$

Our first goal is to bound $\nabla\tilde{H}$. By Theorem 3.2 and using the bound on $\|\mathbf{F}_N\|$ derived in (3.12), we have

$$\|\mathbf{F}_N(\mathbf{y})\| = \|B(\mathbf{y})\nabla\tilde{H}(\mathbf{y})\| \le M(1+0.064\eta).$$

On the other hand, since we consider the Lotka-Volterra system, we have

$$\|B(\mathbf{y})\nabla\tilde{H}(\mathbf{y})\| = \left\| \begin{matrix} uv\tilde{H}_v(\mathbf{y}) \\ -uv\tilde{H}_u(\mathbf{y}) \end{matrix} \right\| = |uv| \left\| \begin{matrix} \tilde{H}_v(\mathbf{y}) \\ -\tilde{H}_u(\mathbf{y}) \end{matrix} \right\| = |uv|\,\|\nabla\tilde{H}(\mathbf{y})\|,$$

so that

$$\|\nabla\tilde{H}(\mathbf{y})\| \le \frac{M(1+0.064\eta)}{|uv|} \le \frac{M(1+0.064\eta)}{(u_{\min}-R/2)(v_{\min}-R/2)} =: L,$$

since $\min\{u : (u,v) \in \tilde{K}_2\} = u_{\min} - R/2$ and similarly for v. The bound L is in fact a global h-independent Lipschitz constant for $\tilde{H}$ and

$$\|\tilde{H}(\mathbf{y}_{n+1}) - \tilde{H}(\tilde{\phi}_{N,h}(\mathbf{y}_n))\| \le L\,\|\mathbf{y}_{n+1} - \tilde{\phi}_{N,h}(\mathbf{y}_n)\| \le L\,h\gamma M e^{-h_0/h},$$

by Lemma 3.3.

We are now in a position to bound

$$\begin{aligned} |\tilde{H}(\mathbf{y}_n) - \tilde{H}(\mathbf{y}_0)| &= \left| \sum_{j=1}^{n} \left[\tilde{H}(\mathbf{y}_j) - \tilde{H}(\mathbf{y}_{j-1})\right] \right| = \left| \sum_{j=1}^{n} \left[\tilde{H}(\mathbf{y}_j) - \tilde{H}(\tilde{\phi}_{N,h}(\mathbf{y}_{j-1}))\right] \right| \\ &\le \sum_{j=1}^{n} |L\,h\gamma M e^{-h_0/h}| = nhL\,\gamma M e^{-h_0/h}, \end{aligned}$$

so that for $nh < e^{h_0/2h}$, we get the first inequality we wanted to prove.

It remains to show an equivalent result for the Hamiltonian. Since

$$\tilde{H}(\mathbf{y}) = H(\mathbf{y}) + h[H_2(\mathbf{y}) + hH_3(\mathbf{y}) + \cdots + h^{N-2}H_N(\mathbf{y})],$$

we have to prove that $H_2(\mathbf{y}) + hH_3(\mathbf{y}) + \cdots + h^{N-2}H_N(\mathbf{y})$ is uniformly bounded on K independently of h and N. By Theorem 3.2, we have for all j

$$\mathbf{f}_j(\mathbf{y}) = B(\mathbf{y})\mathbf{g}_j(\mathbf{y}) = B(\mathbf{y})\nabla H_j(\mathbf{y}),$$

so that, using the bound (3.9), we have for $j \geq 2$ and $\mathbf{y} \in \tilde{K}_2$

$$\|\mathbf{f}_j(\mathbf{y})\| = |uv| \cdot \|\nabla H_j(\mathbf{y})\| = |uv| \cdot \|\mathbf{g}_j(\mathbf{y})\| \leq \ln 2 \, \frac{\eta M}{2} \left(\frac{\eta M(j-1)}{R} \right)^{j-1}.$$

On the other hand, one can check by direct differentiation (and using the symmetry assumption $\frac{\partial \mathbf{f}^{[i]}}{\partial \mathbf{y}^{[k]}} = \frac{\partial \mathbf{f}^{[k]}}{\partial \mathbf{y}^{[i]}}$) that we get $\mathbf{f}(\mathbf{y}) = \nabla H(\mathbf{y})$ for H defined by

$$H_j(\mathbf{y}) = \int_0^1 (\mathbf{y} - \mathbf{z}_0)^T \mathbf{g}_j(\mathbf{z}_0 + t(\mathbf{y} - \mathbf{z}_0))dt$$

for any $\mathbf{z}_0 \in K$. So we can choose $\mathbf{z}_0$ such that $\|\mathbf{y} - \mathbf{z}_0\| \leq R/2$ which yields

$$\begin{aligned}
\|H_j(\mathbf{y})\| &= \| \int_0^1 (\mathbf{y} - \mathbf{z}_0)^T \mathbf{g}_j(\mathbf{z}_0 + t(\mathbf{y} - \mathbf{z}_0))dt \, \| \\
&\leq \frac{R}{2} \frac{1}{|uv|} \frac{\ln 2M\eta}{2} \left(\frac{\eta M(j-1)}{R} \right)^{j-1} \\
&\leq \frac{R \ln 2M\eta}{4(u_{\min} - R/2)(v_{\min} - R/2)} \left(\frac{\eta M(j-1)}{R} \right)^{j-1}
\end{aligned}$$

and then

$$\begin{aligned}
&\|H_2(\mathbf{y}) + hH_3(\mathbf{y}) + \cdots + h^{N-2}H_N(\mathbf{y})\| \\
&\qquad \leq \sum_{j=2}^{N} \frac{R \ln 2M\eta}{4(u_{\min} - R/2)(v_{\min} - R/2)} \left(\frac{\eta M(j-1)}{R} \right)^{j-1} h^{j-2} \\
&\qquad \leq \frac{\ln 2M^2\eta^2}{4(u_{\min} - R/2)(v_{\min} - R/2)} \sum_{j=1}^{N-1} eN \left(\frac{j}{eN} \right)^{j},
\end{aligned}$$

and since the sum $\sum_{j=1}^{N-1} N(\frac{j}{eN})^j$ is maximal for $N = 4$ and is bounded by 0.588 we define

$$C := \frac{0.588\, e \ln 2M^2\eta^2}{4(u_{\min} - R/2)(v_{\min} - R/2)} = \frac{0.277 M^2\eta^2}{(u_{\min} - R/2)(v_{\min} - R/2)}$$

and $H_2(\mathbf{y}) + hH_3(\mathbf{y}) + \cdots + h^{N-2}H_N(\mathbf{y})$ is uniformly bounded on K by C. Finally we have for $nh < e^{h_0/2h}$

$$|H(\mathbf{y}_n) - H(\mathbf{y}_0)| \leq L\, \gamma M e^{-h_0/2h} + 2hC.$$

3.5 Application

We can apply Lemma 3.4 to our problem, namely "how to be sure that the numerical result remains positive". The constructive proof of the following theorem gives a routine which enables us to determine the step-size h^* for which the numerical result stays in the first quadrant over exponentially long time intervals.

Theorem 3.3 *Let (u_0, v_0) be given initial conditions, and let h^* be the minimum of $h_0/3$ and the unique solution of*

$$L\,\gamma M e^{-h_0/2h} + 2hC = H_{\max} - H_0,$$

where the constants h_0,L,γ,M,C,$H_{\max}$ and H_0 are defined in the proof below. Then if we apply the explicit variant of the symplectic Euler method with a step-size h smaller than h^*, the numerical solution stays positive over exponentially long time intervals $t = nh^* \leq e^{h_0/2h^*}$.

Proof The constants a, b, u_0 and v_0 are fixed, so that we can compute $H_0 := H(u_0, v_0)$. The level curve of H_0 defines the compact set K. Then we compute the maximum values and the minimum values of u and v in K and we obtain the values of

$$\begin{aligned} u_{\max} &:= \max\{u : (u,v) \in K\}, \\ u_{\min} &:= \min\{u : (u,v) \in K\}, \\ v_{\max} &:= \max\{v : (u,v) \in K\}, \\ v_{\min} &:= \min\{v : (u,v) \in K\}. \end{aligned}$$

Then we set $R := \alpha \min\{u_{\min}, v_{\min}\}$ with, for example, $\alpha = 0.9$, so that we can define the compact set $\tilde{K} = \{\mathbf{y} : \exists \mathbf{y_0} \in K$ such that $\|\mathbf{y} - \mathbf{y}_0\| \leq R\}$ and in a similar way, $\tilde{K}_2$.

We have by definition $M := \max\{\|\mathbf{f}(u,v)\| \,|\, (u,v) \in \tilde{K}\}$, however it is much easier to use

$$M = \max\{\tilde{u}_{\max}(\tilde{v}_{\max} - b), \tilde{v}_{\max}(\tilde{u}_{\max} - a)\},$$

where $\tilde{u}_{\max} = \max\{u \,|\, (u,v) \in \tilde{K}\} = u_{\max} + R$ and $\tilde{v}_{\max} = v_{\max} + R$. Similarly we use

$$M_2 = \tilde{u}_{\max}\tilde{v}_{\max}(\tilde{v}_{\max} - b)$$

instead of $\max_{\mathbf{y}\in\tilde{K}} \|\mathbf{d}_2(\mathbf{y})\|$. Once we have these values, we can compute

$$\eta = \frac{2}{2\ln 2 - 1} + \frac{RM_2}{M^2} \qquad \text{and} \qquad h_0 = \frac{R}{e\eta M}$$

as well as $\gamma = e[2 + \frac{eh_0M_2}{3M} + 0.064\eta]$,

$$\tilde{L} = \frac{M(1 + 0.064\,\eta)}{(u_{\min} - R/2)(v_{\min} - R/2)}$$

and

$$C = \frac{0.277M^2\eta^2}{(u_{\min} - R/2)(v_{\min} - R/2)}.$$

The next step is to choose h smaller than $h_0/3$. Once h is chosen, we check whether or not it is small enough to ensure that the numerical solution stays in $\tilde{K}_2$. We know that the bound (3.14) is valid if and only if $\mathbf{y}_n$ is in $\tilde{K}_2$, so we need to know that it does stay in this compact set. Defining

$$H_{\max} := \min\{H(u,v) \,|\, (u,v) \in \partial\tilde{K}_2\} = \min\{\, H(u_{\min} - R/2, b), H(a, v_{min} - R/2)\,\},$$

we know that $\mathbf{y}_n$ stays in $\tilde{K}_2$, if $H(\mathbf{y}_n) < H_{\max}$ that is, using the bound (3.14), if

$$L\,\gamma M e^{-h_0/2h} + 2hC \leq H_{\max} - H_0. \tag{3.15}$$

Now all the constants in the above expression are positive, and since any function of the form

$$f(h) = \alpha e^{-h_0/2h} + \beta h - \delta,$$

where α, β, and δ are positive, is strictly increasing, h satisfies inequality (3.15) if and only if h is smaller than the unique solution of

$$L\,\gamma M e^{-h_0/2h} + 2hC = H_{\max} - H_0. \tag{3.16}$$

In other words, the bound h^* we are looking for is given by the minimum of $h_0/3$ and the solution of (3.16). Moreover, the numerical solution stays in $\tilde{K}_2$ for at least $t = nh^* \leq e^{h_0/2h^*}$.

3.6 Example

As an illustration, we consider the problem

$$\begin{cases} \dot{u} = u(1-v), \\ \dot{v} = v(u-2), \end{cases} \tag{3.17}$$

with the initial condition $u(0) = 1.5$, $v(0) = 0.5$.

The exact solution of the system stays on the level curve of the Hamiltonian with $H_0 = 1.8822$, so we define $K = \{(u,v) \,|\, |H(u,v)| \leq 1.8822\}$. Since the solution of (3.16) is 0.000113, we conclude that for values of h smaller that 0.000113, we are sure that the numerical Hamiltonian is well-conserved and that the numerical solution stays positive and exhibits the right qualitative behavior, for a time $t = nh \sim 10^{15}$. As one can see in Figure 3.2, the estimate for h is really pessimistic. For values as large as $h = 0.1$, the Hamiltonian is extremely well-conserved, with $\max_{n\geq 1} |H(\mathbf{y}_n) - H_0| = 0.0127$ and the numerical simulation only starts to leave the first quadrant for values of h greater or equal to 0.7.

An important remark is that we not only proved that the numerical solution of the problem (3.17) with the initial condition $u(0) = 1.5$, $v(0) = 0.5$, will stay in the first quadrant if we use a step size smaller than 0.000113, but we also proved this result for any similar problem with initial condition $\mathbf{y}_0 = (u(0), v(0)) \in K$, since the initial condition was only used to define the compact set K, and all results are true for all $\mathbf{y}_0 \in K$.

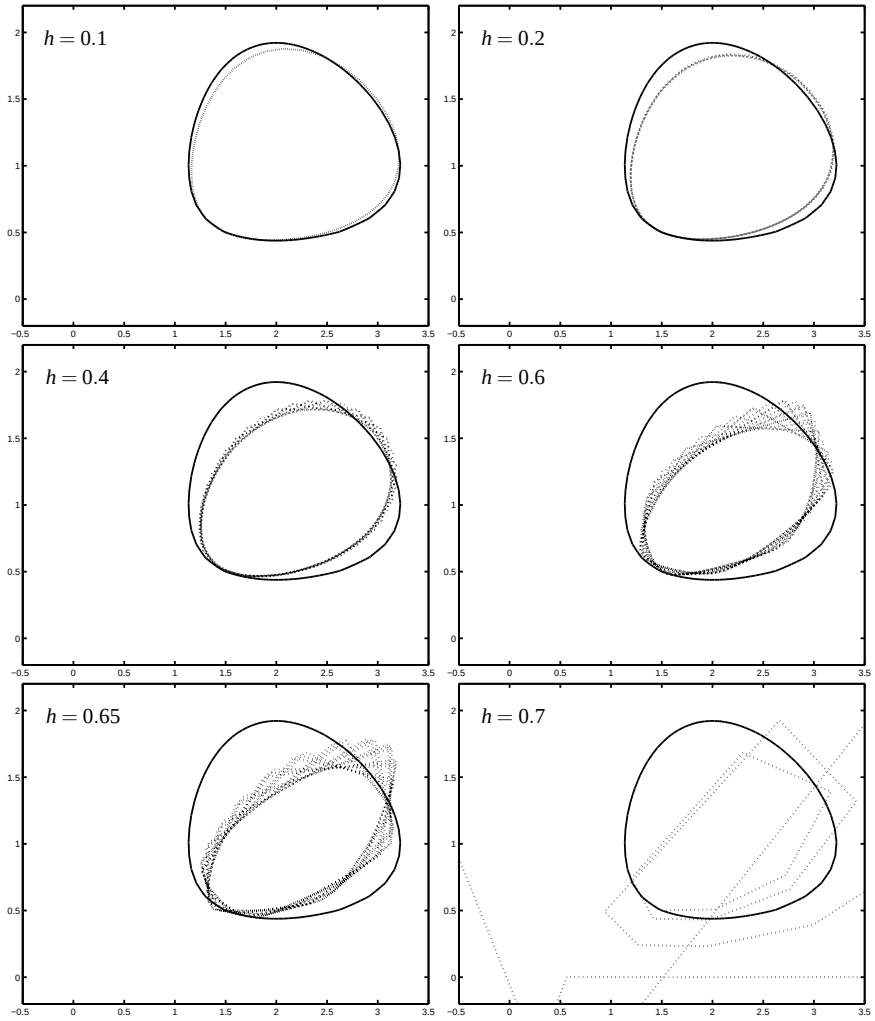

Fig. 3.2 Numerical solutions obtained using large step-size.

References

1. M. GANDER. A non-spiraling integration for the Lotka-Volterra equation. *Il Volterriano*, 4:21–28, 1994.
2. E. HAIRER. Backward analysis of numerical integrators and symplectic methods. *Annals of Numerical Mathematics*, 1(1-4):107–132, 1994.
3. E. HAIRER, C. LUBICH, and G.WANNER. *Geometric Numerical Integration*. Springer-Verlag. Berlin, 2002.
4. W. KAHAN. *Unconventional numerical methods for trajectory calculations*. Lecture Notes. CS Division, Department of EECS, University of California at Berkeley, 1993.