Symplectic Methods Applied to the Lotka-Volterra System

Mélanie Beck

Department of Mathematics and Statistics,
McGill University, Montréal
Québec, Canada

June, 2003

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements of the degree of
Master of Science

Abstract

We analyse the preservation of physical properties of numerical approximations to solutions of the Lotka-Volterra system: its positivity and the conservation of the Hamiltonian. We focus on two numerical methods : the symplectic Euler method and an explicit variant of it. We first state under which conditions they are symplectic and we prove they are both Poisson integrators for the Lotka-Volterra system. Then, we study under which conditions they stay positive. For the symplectic Euler method, we derive a simple condition under which the numerical approximation always stays positive. For the explicit variant, there is no such simple condition. Using properties of Poisson integrators and backward error analysis, we prove that for initial conditions in a given set in the positive quadrant, there exists a bound on the step size, such that numerical approximations with step sizes smaller than the bound stay positive over exponentially long time intervals. We also show how this bound can be estimated. We illustrate all our results by numerical experiments.

Résumé

Nous analysons la préservation des propriétés physiques d'approximations numériques des solutions du système Lotka-Volterra : sa positivité et la conservation du Hamiltonien. Nous nous concentrons sur deux méthodes : la méthode d'Euler symplectique et une variante explicite de celle-ci. Nous énonçons d'abord sous quelles conditions ces méthodes sont symplectiques et nous prouvons qu'elles sont toutes deux des intégrateurs de Poisson pour le système Lotka-Volterra. Puis nous étudions sous quelles conditions elles sont positives. Pour la méthode d'Euler symplectique, nous obtenons une condition simple sous laquelle l'approximation numérique reste toujous positive. Pour la variante explicite, il n'y a pas de condition aussi simple. En utilisant les propriétś des intégrateurs de Poisson et l'analyse implicite de l'erreur ("backward error analysis"), nous prouvons que pour des conditions initiales appartenant à un ensemble donné dans le quadrant positif, il existe une borne sur le pas de temps, telle que les approximations numériques avec un pas de temps plus petit que cette borne restent positives sur des intervalles exponentiellement longs. Nous montrons également comment cette borne peut être estimée. Nous illustrons tous nos résultats pas des expériences numériques.

Acknowledgments

Je voudrais tout d'abord remercier mon superviseur, Martin J. Gander, pour son aide, son soutien, ses encouragements et ses suggestions ; le temps qu'il consacre à ses étudiants est impressionnant. Sans lui, je n'aurais jamais commencé de maîtrise, et j'aurais encore moins terminé cette maîtrise. He is such an excellent supervisor!

J'aimerais également remercier Ernst Hairer et Gerhard Wanner pour m'avoir invitée à participer au séminaire d'analyse numérique de l'Université de Genève, ainsi que pour les discussions concernant mon projet de recherche et les idées qu'ils m'ont apportées. Merci aussi pour leur ouvrage [5] si complet.

Merci à Olivier Dubois et Charles Fortin pour avoir relu ma thèse et suggéré des améliorations.

I would like to thank the professors of applied mathematics in the department, in particular Nilima Nigam, Paul Tupper, Georg Schmidt, Tony Humphries and Peter Bartello, for their efforts in making the department more and more friendly, pleasant and enjoyable for studying.

I also want to thank the Department of Mathematics and Statistics, l'Institut des Sciences Mathématiques and the Faculty of Graduate Studies, for their financial support.

Finalement je remercie ma famille et mes amis des deux côtés de l'Atlantique, grâce à qui ma vie d'étudiante est si agréable.

Table of Contents

Introduction

How can we preserve important physical properties of the solution of the Lotka-Volterra system when we solve it numerically? Geometric integration has been focusing on this kind of problems over the last decades. New categories of numerical integrators whose main advantage is to preserve the qualitative attributes of the solution as much as possible, have been developed. Well-known examples are symplectic integrators, energy preserving integrators, volume preserving integrators and Lie group integrators. Symplectic integrators, i.e. area preserving integrators in two dimensions, are well suited to approximate Hamiltonian systems of the form $\dot{p} = -H_q(p,q)$, $\dot{q} = H_p(p,q)$, where the Hamiltonian $H(p,q)$ represents the total energy and H_p and H_q are the vectors of partial derivatives. One can easily check that the Hamiltonian is an invariant of the solutions of the system and one can prove that the flows of Hamiltonian systems are symplectic maps. It has been observed that even if symplectic methods concentrate on the preservation of geometric properties, they give more accurate long-time integration than general-purpose methods.

The particularity of the Lotka-Volterra system, also called prey-predator system, is its similarity with Hamiltonian systems. To study this type of problems, an extansion of Hamiltonian systems has been created under the name Poisson systems. The solutions of the Lotka-Volterra system are periodic and positive, and the differential system itself is only valid for positive variables. Nevertheless, for a majority of numerical methods, it is impossible to be sure that the numerical results stay positive and consequently we can not expect a good long-time approximation. It is therefore important to study the possibilities offered by specific "Poisson integrators".

This thesis' focus is a specific method, the symplectic Euler method, and an explicit variant of it. Both methods are Poisson integrators for the Lotka-Volterra system and our interest lies in the preservation of the positivity. For the symplectic Euler method, very simple arguments yield the desired result : if the step-size is chosen smaller than a bound determined by the problem (namely, the minimum of the inverse of the equilibrium point's coordinates), the numerical solution stays positive for all time. In contrast, for an explicit variant of this method, much more work is needed and important properties of Poisson integrators have to be used to show a similar result. In particular, backward error analysis is the key tool. The final result shows how to compute, for given initial conditions, a bound h^* such that every numerical solution obtained with a step-size smaller than h^*, is really close to the exact solution and remains positive for exponentially long time intervals.

In the first chapter, after a short historical bibliography on the symplectic Euler method, we present the Lotka-Volterra system and its properties.

The second chapter contains illustrations of some classical methods applied to the Lotka-Volterra system together with the advantages and disadvantages of these methods.

We introduce, in the third chapter, the notions of symplecticity and of Poisson integrators, and we also study in more details two methods : the symplectic Euler method and an explicit variant of it. We study their symplecticity, whether or not they are Poisson integrators for the Lotka-Volterra system and finally we study their positivity (when applied to the Lotka-Volterra system).

The fourth chapter is devoted to the backward error analysis; after defining this concept, we state some properties of symplectic methods and Poisson integrators and compute the first terms of the numerical Hamiltonians of the symplectic Euler method and its explicit variant. We also study the structure of these numerical Hamiltonians.

The fifth and last chapter exploits backward error analysis. We focus on the explicit variant of the symplectic Euler method and prove the important theorem concerning the choice of the step-size ensuring the positivity of the numerical result for exponentially long time intervals.

Chapter 1

Preliminaries

1.1 Historical Bibliography

In a paper never published [2], Devogelaere introduced in 1956 for partitionned sytems

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

the numerical method defined by

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

He pointed out that it is area-preserving when applied to a Hamiltonian system, which is, as we will see, the characteristic of symplectic methods.

We have to wait until 1993 to find again this method in the lecture notes [7] of Kahan. In these notes, the method is presented under the name *unconventional numerical method.*

The following year, the method can be found in different papers. Sanz-Serna wrote an article [9] about the *unconventional symplectic integrator of W. Kahan*, and the method appeared in the book written by Sanz-Serna and Calvo, [6], devoted to the numerical approximation of Hamiltonian problems. However in this text the method is not given any specific name; it is presented as the first order symplectic Runge-Kutta method. In the mean time, Hairer introduced the method in [4] motivated by the backward error analysis and called it the *symplectic Euler method.* In [3], Gander studied particularly the Lotka-Volterra equation, and in order to have an explicit method, he defined the symplectic Euler method as

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

Later, in 2000, Meyer-Spasche and Gander studied several numerical integrators preserving physical properties in [8]. They continued studying the explicit variant of the symplectic Euler method and applied it to Hamiltonian problems with separable Hamiltonian and to the Lotka-Volterra system. The same year, two physicists, Sturgeon and Laird used the symplectic Euler method in [10] to define the Stormer-Verlet scheme, a composition of the symplectic Euler method and its adjoint.

The study of the symplectic Euler method continued in a few papers written in 2002. In a chapter of his thesis [11], Tupper explored the results obtained by different numerical methods applied to a Hamiltonian system on long time intervals and using large step sizes. He illustrated the excellent performance of the symplectic Euler method and the mediocrity of one-step-and-project methods in the context of long-time statistics. In Norway, Berland studied in his diploma thesis [1] numerical methods, including the symplectic Euler method, by the means of Lie group theory. Finally in that same year, Hairer, Lubich and Wanner published the most complete book written up to date on geometric numerical integration, [5]. Most of the results observed about the symplectic Euler method, and even more, can be found in this reference.

1.2 The Lotka-Volterra System

This thesis mainly focuses on the numerical approximation of the Lotka-Volterra system

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

which models the evolution of two animal species. Here $u(t)$ is the number of prey and $v(t)$ the number of predators. Actually, u and v are continuous variables since we consider densities and not numbers of individual. The constants a and b depend on the two animal species considered, b is the growing rate of preys, when there is no predator, and a represents the tendance of extinction of the predators when there is no prey. The term uv is related to the decreasing rate of preys due to predators in the first equation and to the rate of variation of predators corresponding to the quantity of available food in the second equation.

The Lotka-Volterra system is interesting due to its geometric property : every solution of (1.1) lies on a closed curve (actually one can even show that it is periodic).

If we divide the two equations of the Lotka-Volterra system (1.1), we obtain

$$\frac{\dot{u}}{\dot{v}} = \frac{u(b-v)}{v(u-a)},$$

which becomes, after separation of variables,

$$\frac{u-a}{u}\dot{u} - \frac{b-v}{v}\dot{v} = 0.$$

Integrating this equality we obtain an invariant of the system,

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

Hence every solution of (1.1) lies on the level curves of the function (1.2), and since these curves are closed, every solution is cyclic. Some level curves are plotted in Figure 1.1.

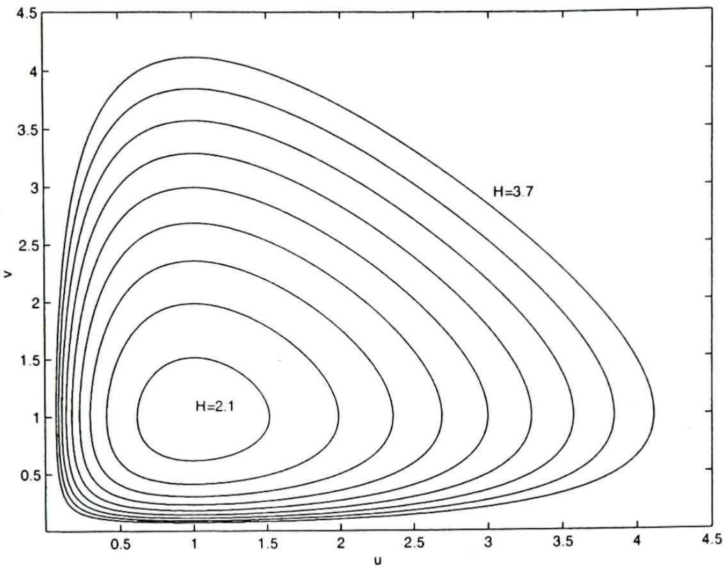

Figure 1.1: Some level curves of the Hamiltonian of the Lotka-Volterra system, from $H = 2.1$ to $H = 3.7$.

If H is defined by (1.2), the Lotka-Volterra system can be written as

$$\begin{cases} \dot{u} &= -uv\, H_v(u,v), \\ \dot{v} &= uv\, H_u(u,v), \end{cases}$$

where H_u and H_v denote the partial derivatives of H with respect to u and v. In other words, the system is not Hamiltonian but it is a *non-canonical* Hamiltonian system, or more generally a Poisson system (a more precise definition is given in Chapter 3). This explains why thereafter we call H the Hamiltonian of the system.

It is important that numerical simulations of the system (1.1) show the same qualitative behaviour as the exact solution, in particular its cyclicity. As one can see on Figure 1.2, the result obtained by the forward Euler and the backward Euler methods spiral outwards or inwards, so specific methods have to be used to avoid this.

Another problem of numerical methods applied to the Lotka-Volterra system is the positivity. Since the variables of the Lotka-Volterra system represent the density

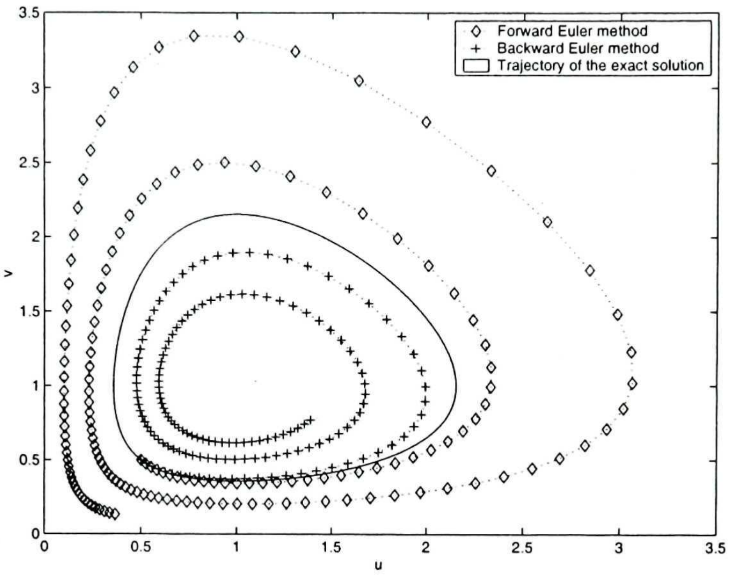

Figure 1.2: Illustration of the forward Euler and the backward Euler methods, with $u_0 = 0.5$, $v_0 = 0.5$, $a = b = 1$ and $h = 0.1$.

of certain species, they are supposed to be positive. The model is invalid whenever a variable is non-positive. Yet, it may occur that a numerical solution leaves the first quadrant. In this case, the numerical approximation and the method become useless. Examples are shown in Figure 1.3 and Figure 1.4.

Before applying different numerical methods to the system, we should study its linear stability. From the definition of the system (1.1) we compute the Jacobian

$$\nabla f = \begin{pmatrix} b - v & -u \\ v & u - a \end{pmatrix}. \tag{1.3}$$

We now study the behaviour of the equation close to its two distinct fixed points : the origin and the equilibrium point (a, b). At the origin, the Jacobian (1.3) becomes

$$\nabla f = \begin{pmatrix} b & 0 \\ 0 & -a \end{pmatrix}$$

whose eigenvalues are $b > 0$ and $-a < 0$, so that the origin is a saddle point, attracting

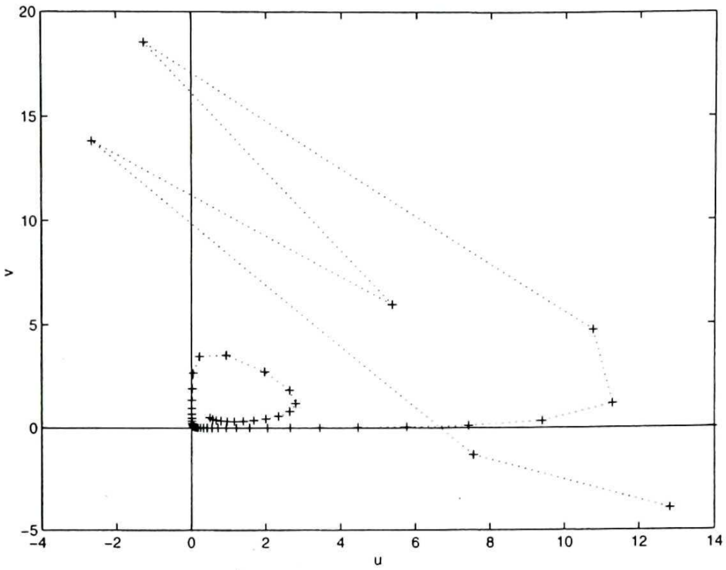

Figure 1.3: Illustration of the forward Euler method when the solution leaves the first quadrant. $u_0 = 0.5$, $v_0 = 0.5$ and $h = 0.3$.

along v and repulsive along u. At the equilibrium point the Jabobian becomes

$$\nabla f = \begin{pmatrix} 0 & -a \\ b & 0 \end{pmatrix}$$

and its eigenvalues are $\pm i\sqrt{ab}$. Hence the equilibrium point is hyperbolic and the solution is rotating around it. This analysis is confirmed by the shape of the level curves given in Figure 1.1.

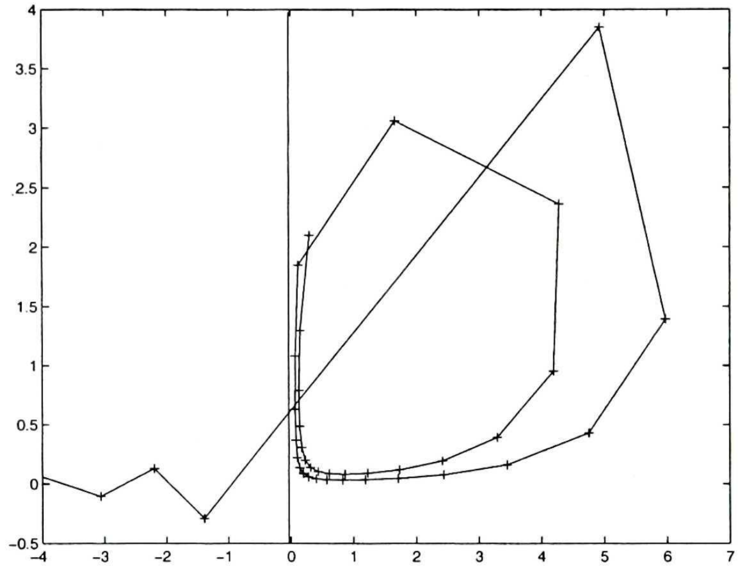

Figure 1.4: Illustration of an explicit variant of the symplectic Euler method when the solution leaves the first quadrant. $u_0 = 0.3$, $v_0 = 2.1$ and $h = 0.45$.

Chapter 2

Different Methods Applied to the Lotka-Volterra System

As an illustration, we apply classical numerical methods to the Lotka-Volterra system and observe their properties. To simplify the computation of implicit methods, we consider the system when a and b are both equal to one:

$$\begin{cases} \dot{u} &= u\,(1-v) \;=\; f(u,v), \\ \dot{v} &= v\,(u-1) \;=\; g(u,v). \end{cases}$$

2.1 Forward Euler

The forward Euler method, given by

$$.\begin{cases} u_{n+1} &= u_n + hf(u_n, v_n), \\ v_{n+1} &= v_n + hg(u_n, v_n), \end{cases}$$

is easy to implement because it is an explicit method. When we apply it to the Lotka-Volterra system, u_{n+1} and v_{n+1} are given by

$$\begin{cases} u_{n+1} &= u_n + hu_n\,(1-v_n), \\ v_{n+1} &= v_n + hv_n\,(u_n-1). \end{cases}$$

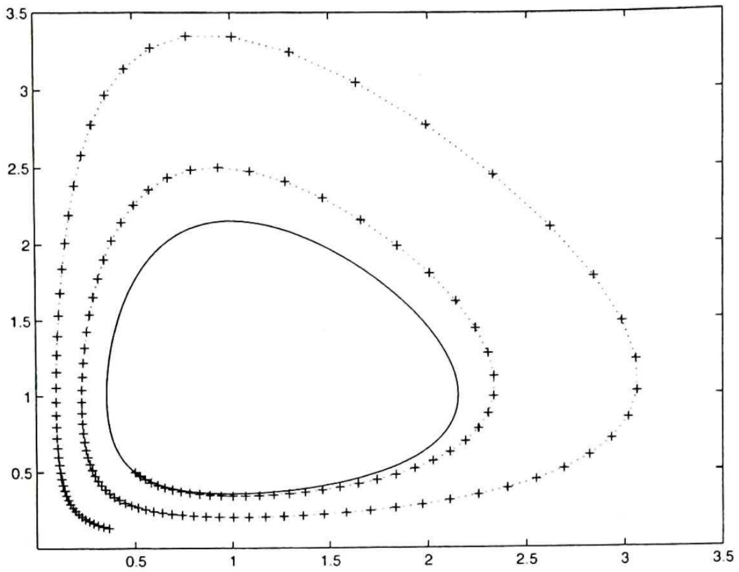

Figure 2.1: Illustration of the performance of the forward Euler method applied to the Lotka-Volterra system, with $u_0 = 0.5$, $v_0 = 0.5$ and $h = 0.1$.

Figure 2.1 shows the numerical solution obtained with the forward Euler method for $h = 0.1$. We observe that it spirals outwards whereas the exact solution should lie on a closed curve (the solid line on the figure).

2.2 Backward Euler

The backward Euler method is an implicit method given by

$$\begin{cases} u_{n+1} &= u_n + hf(u_{n+1}, v_{n+1}), \\ v_{n+1} &= v_n + hg(u_{n+1}, v_{n+1}); \end{cases}$$

yet, for the Lotka-Volterra system, one can explicitely advance it because of the simple form of f and g: to express u_{n+1} and v_{n+1} as functions of u_n and v_n, we first derive from the first equation of the method

$$u_{n+1} = \frac{u_n}{1 - h(1 - v_{n+1})}.$$

Substituting this definition of u_{n+1} into the second equation of the method,

$$v_{n+1} = v_n + hv_{n+1}(u_{n+1} - 1),$$

we obtain an equation of second order in v_{n+1}, whose solutions are

$$\frac{1}{2(1+h)}\left(- \left(\frac{1}{h} - h - u_n - v_n\right) \pm \sqrt{(\frac{1}{h} - h - u_n - v_n)^2 + 4(1+h)v_n(\frac{1}{h} - 1)}\,\right).$$

Now we have to choose one of these two solutions. As h goes to zero, the numerical result should converge to the exact solution, in particular v_{n+1} should stay bounded. Since the terms of $1/h$ blow up as h goes to zero, the correct root is the one with the positive sign, so that the terms $(1/h - h - u_n - v_n)$ balance themselves. One can indeed check, for example using Maple, that the first term of the expansion of the solution with the negative sign is $-1/h$ whereas the first term of the expansion of the solution with the positive sign is v_n.

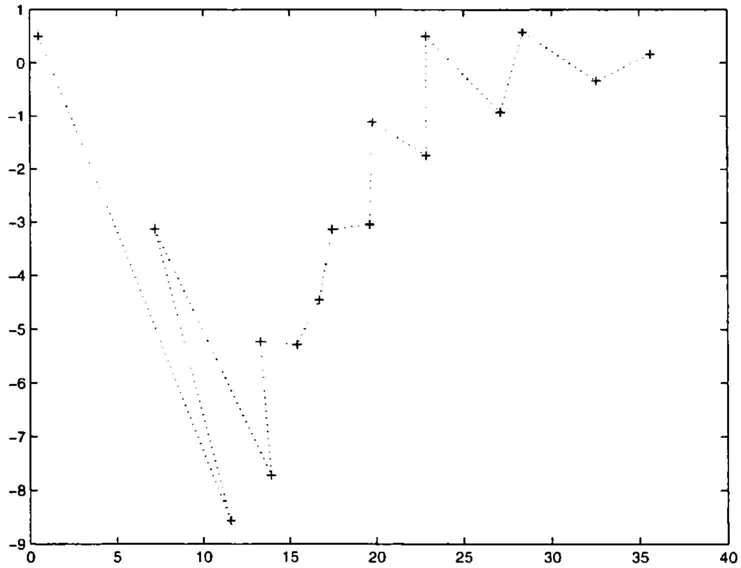

Figure 2.2: Illustration of the performance of the backward Euler method applied to the Lotka-Volterra system when we use the root with the negative sign for v_{n+1}, with $u_0 = 0.5$, $v_0 = 0.5$, and $h = 0.1$.

Figure 2.2 gives an illustration of the numerical approximation obtained if we use the root with the negative sign. Figure 2.3 shows the numerical solution obtained

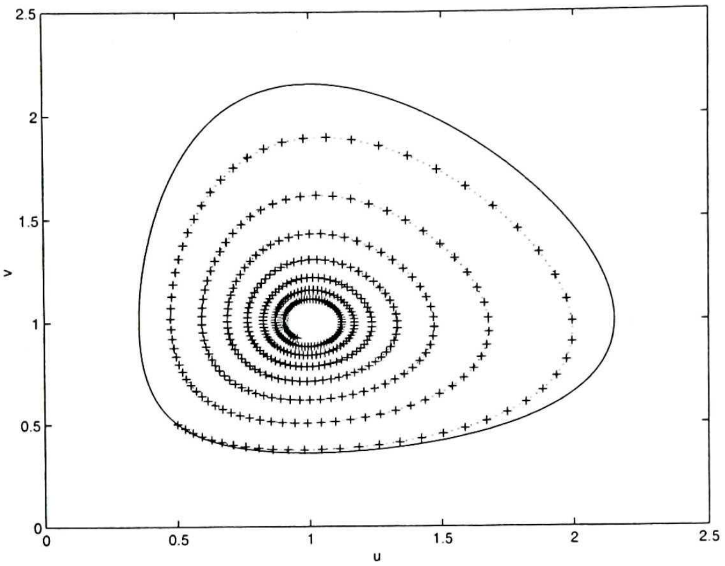

Figure 2.3: Illustration of the performance of the backward Euler method applied to the Lotka-Volterra system when we use the root with the positive sign for v_{n+1}, with $u_0 = 0.5$, $v_0 = 0.5$, and $h = 0.1$.

with this method when we use the root with the positive sign, together with the level curve of the Hamiltonian corresponding to the initial conditions. The solution obtained with the backward Euler method spirals inwards toward the steady state $(1, 1)$ whereas we saw in the previous section that the solution obtained using the forward Euler method spirals outwards. A direct consequence of this is that the numerical solution stays always positive.

In the two following sections, we present two first-order methods that are obtained by a slight modification of the forward Euler method. Both methods are easy to derive and the numerical approximations exhibit one correct qualitative behaviour, namely the cyclicity.

2.3 Symplectic Euler

The symplectic Euler method is defined in [4] by

$$\begin{cases} u_{n+1} &= u_n + h\, f(u_{n+1}, v_n), \\ v_{n+1} &= v_n + h\, g(u_{n+1}, v_n), \end{cases} \tag{2.1}$$

and gives, when applied to the Lotka-Volterra system

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

that is

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

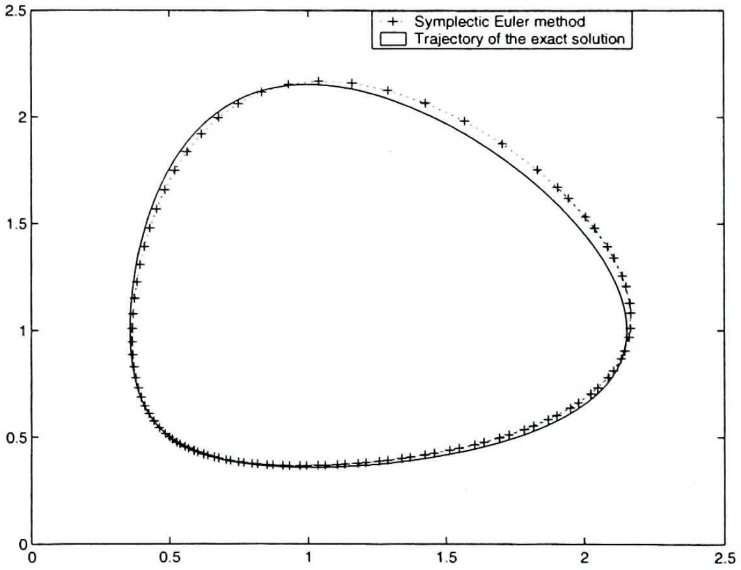

Figure 2.4: Illustration of the symplectic Euler method applied to the Lotka-Volterra system, with $u_0 = 0.5$, $v_0 = 0.5$ and $h = 0.1$.

An illustration of the excellent performance of the method is given in Figure 2.4. We observe that the numerical result stays on a closed curve, nearly the level curve

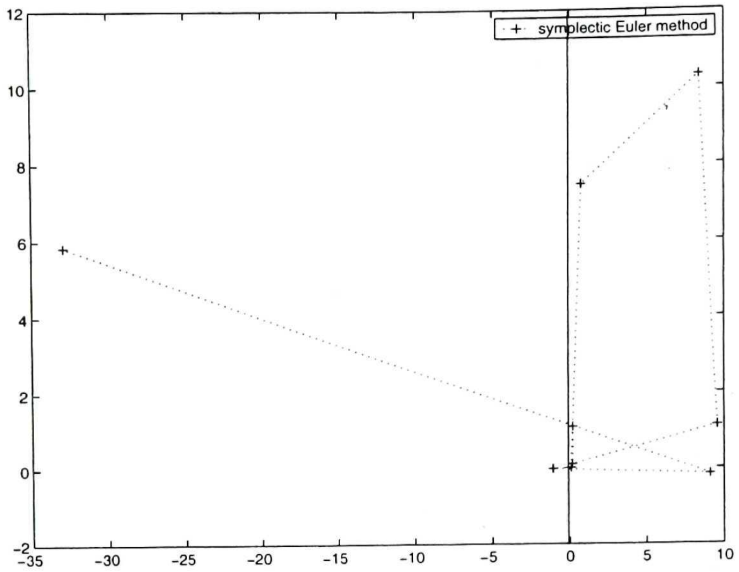

Figure 2.5: Illustration of the symplectic Euler method applied to the Lotka-Volterra system, with $u_0 = 0.2$, $v_0 = 1.1$ and $h = 1.1$.

of the Hamiltonian of the system. However it may happen, if we use a too large step-size, that the numerical simulation leaves the first quadrant. An example is given in Figure 2.5. To study the positivity of the numerical results, we plotted in Figure 2.6 the number of iterations needed for each point (u_0, v_0) to leave the first quadrant when applying the symplectic Euler method. It clearly illustrates that for some initial values and some step-size (here $h = 1$), solutions leave the first quadrant.

In the next chapter, we study in more details this method in order to explain its performance and we also give a condition on the step-size which ensures the positivity of the numerical approximations.

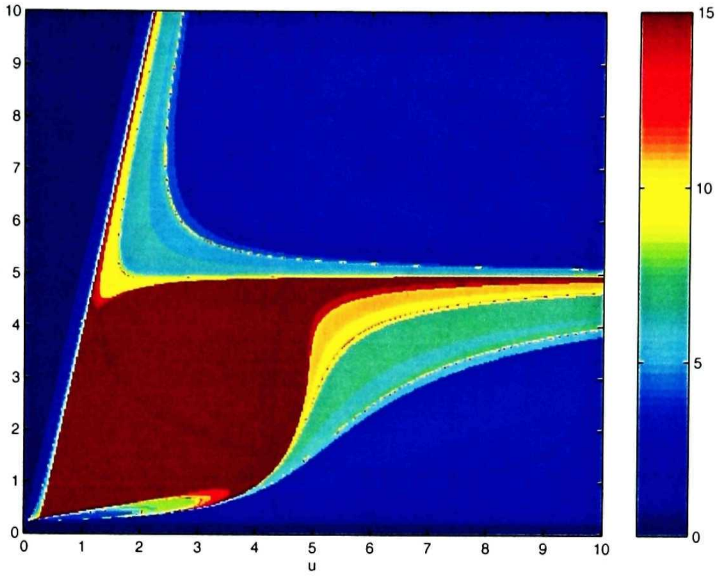

Figure 2.6: Number of iterations needed for each point (u_0, v_0) to leave the first quadrant when applying the symplectic Euler method to the Lotka-Volterra system with $h = 1$ and $a = b = 1.2$.

2.4 Explicit Variant of symplectic Euler

As we said in Chapter 1, an explicit variant of this method, defined by

$$\begin{cases} u_{n+1} &= u_n + h\, f(u_n, v_n), \\ v_{n+1} &= v_n + h\, g(u_{n+1}, v_n), \end{cases} \tag{2.3}$$

was introduced by Gander in [3]. One can note that we could also take $g(u_n, v_{n+1})$ instead of $g(u_{n+1}, v_n)$ in the second equation. Applied to the Lotka-Volterra system it becomes

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

In general, this method is well performant, as illustrated in Figure 2.7. However, as for the symplectic Euler method, it may happen that the numerical solution leaves the first quadrant, see an example on Figure 2.8. We will see in the next chapter that it is not as easy as for the symplectic Euler method to find values of the step-

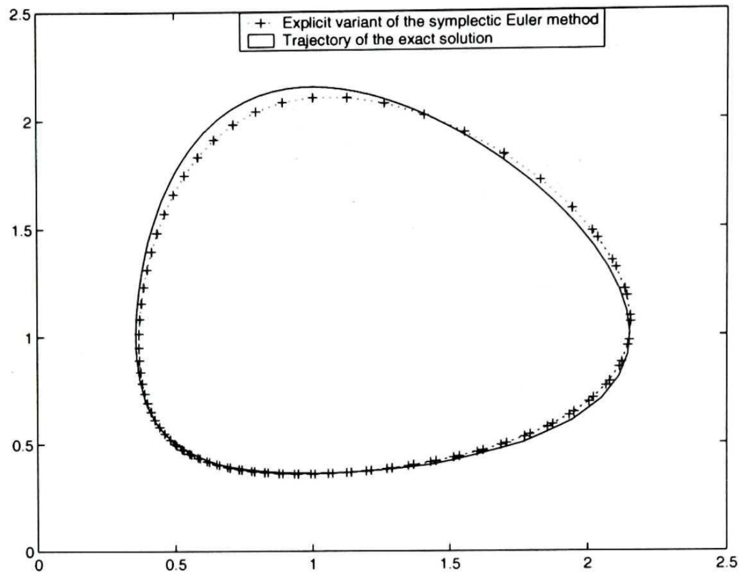

Figure 2.7: Illustration of the explicit variant of the symplectic Euler method applied to the Lotka-Volterra system, with $u_0 = 0.5$, $v_0 = 0.5$ and $h = 0.1$.

size for which the numerical solution stays positive. To study the positivity of the numerical results, we plotted in Figure 2.9 the picture corresponding to Figure 2.6 for the explicit variant of the symplectic Euler method. It appears that, in spite of the excellent performance of the results in general, we get, for h large, trajectories that leave the first quadrant and the region consisting of initial values for which the numerical result is positive for 100 iterations is really small for $h = 1$.

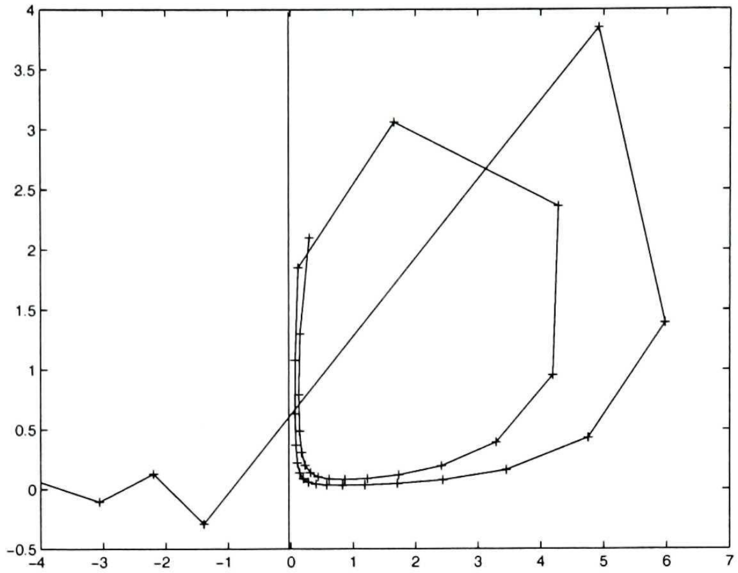

Figure 2.8: Illustration of the variant of the symplectic Euler method when the solution leaves the first quadrant. $u_0 = 0.3$, $v_0 = 2.1$ and $h = 0.45$.

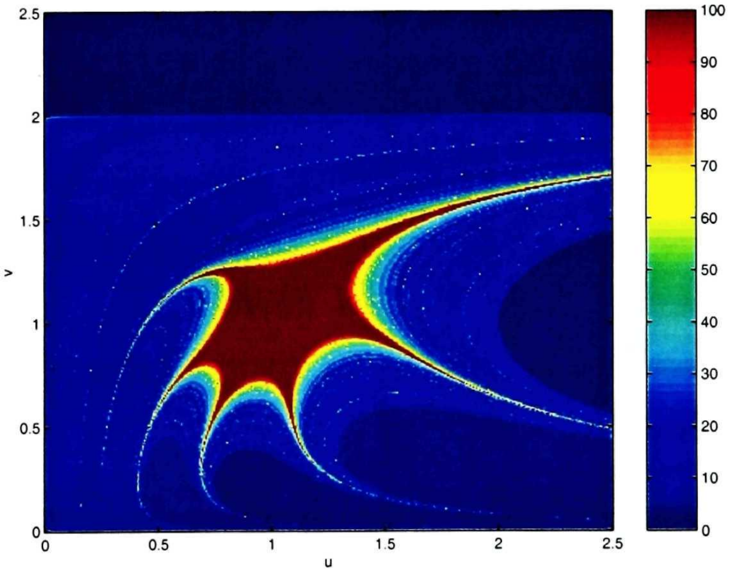

Figure 2.9: Number of iterations needed for each point (u_0, v_0) to leave the first quadrant when applying the explicit variant of the symplectic Euler method to the Lotka-Volterra system with $h = 1$ and $a = b = 1$.

Chapter 3

Symplectic Methods and Poisson Integrators

In this chapter, after introducing the notions of symplecticity and of Poisson integrators, we study in more details the symplectic Euler method and its explicit variant defined in Section 2.3 and Section 2.4. We study their symplecticity, under which conditions they are Poisson integrators and if it possible to find for which step-sizes the numerical solution stays positive.

3.1 Symplecticity

Before defining symplecticity, we need to introduce an important concept in the study of differential equations: the *flow* over time t. This map, denoted by ϕ_t, associates to any point y_0 in the phase space, the value $y(t)$ of the solution with initial value $y(0) = y_0$. In other words, it is defined by

$$\phi_t(y_0) = y(t) \qquad \text{if} \qquad y(0) = y_0.$$

As proved in [8], an interesting property of Hamiltonian systems of the form

$$\begin{cases} \dot{p} &= -\frac{\partial H}{\partial q}(p,q), \\ \dot{q} &= \frac{\partial H}{\partial p}(p,q), \end{cases} \tag{3.1}$$

where $H(p,q)$ is the Hamiltonian, is, when p and q are scalars, area preservation. Transformations that have this property are called *symplectic.* A generalization to higher dimensions of the definition of symplecticity is given in [5]:

Definition 3.1. A differentiable map $g : U \to \mathbb{R}^{2d}$ (where $U \subset \mathbb{R}^{2d}$ is an open set) is called *symplectic*, if the Jacobian matrix $g'(p,q)$ is everywhere symplectic:

$$g'(p,q)^T J\, g'(p,q) = J,$$

where

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

This definition allows us to consider systems of any dimensions: the oriented area of two-dimensional parallelograms P lying in $\mathbb{R}^2$, is replaced by the sum of the oriented areas of the projections of $2d$-dimensional parallelograms P onto the coordinate planes (p_i, q_i). Symplecticity is a characteristic of Hamiltonian systems, more precisely the flows of Hamiltonian systems are symplectic maps. This motivates the following definition of symplecticity of numerical methods.

Definition 3.2. A numerical method is called *symplectic*, if the one-step map $y_1 = \Phi_h(y_0)$ is symplectic whenever the method is applied to a smooth Hamiltonian system.

This means that a numerical method is symplectic if and only if

$$\left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}\right)^T J \left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}\right) = J \tag{3.2}$$

whenever it is applied to the smooth Hamiltonian system (3.1).

To check symplecticity in the case where p and q are scalars, there exists a simpler way (see for example [6]): if we consider a $\mathcal{C}^1$-transformation

$$\begin{pmatrix} u^* \\ v^* \end{pmatrix} = \psi \begin{pmatrix} u \\ v \end{pmatrix}$$

defined on a set $\mathbb{D}$, according to the standard rule for changing variables in an integral, ψ preserves area and orientation if and only if the Jacobian determinant is identically one, that is

$$\forall\,(u,v) \in \mathbb{D}, \quad \frac{\partial u^*}{\partial u}\frac{\partial v^*}{\partial v} - \frac{\partial u^*}{\partial v}\frac{\partial v^*}{\partial u} = 1. \tag{3.3}$$

This condition is equivalent to the matrix equation (3.2), in the case where p and q are scalars. Now we consider the differentials

$$du^* = \frac{\partial u^*}{\partial u}du + \frac{\partial u^*}{\partial v}dv \qquad \text{and} \qquad dv^* = \frac{\partial v^*}{\partial u}du + \frac{\partial v^*}{\partial v}dv$$

and we compute their wedge product (also called exterior product), $du \wedge dv$. This product is bilinear and skew-symmetric (i.e. $du \wedge du = dv \wedge dv = 0$ and $du \wedge dv = -dv \wedge du$), so we get

$$\begin{aligned} du^* \wedge dv^* &= \frac{\partial u^*}{\partial u}\frac{\partial v^*}{\partial u}du \wedge du + \frac{\partial u^*}{\partial u}\frac{\partial v^*}{\partial v}du \wedge dv + \frac{\partial u^*}{\partial v}\frac{\partial v^*}{\partial u}dv \wedge du + \frac{\partial u^*}{\partial v}\frac{\partial v^*}{\partial v}dv \wedge dv \\ &= \left(\frac{\partial u^*}{\partial u}\frac{\partial v^*}{\partial v} - \frac{\partial u^*}{\partial v}\frac{\partial v^*}{\partial u}\right) du \wedge dv. \end{aligned}$$

Consequently, according to the characterization of symplecticity (3.3), the method is symplectic if and only if

$$du_{n+1} \wedge dv_{n+1} = du_n \wedge dv_n \qquad \text{for all} \quad (u_n, v_n).$$

3.2 Poisson Integrators

As we said in Chapter 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 uv in addition. In other words, we can write the Lotka-Volterra system as

$$\dot{y} = B(y)\nabla H(y), \tag{3.4}$$

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

$$B(y) = \begin{pmatrix} 0 & -uv \\ uv & 0 \end{pmatrix}. \tag{3.5}$$

The generalization (3.4) of a Hamiltonian system is called a *Poisson system.*

Definition 3.3. If a matrix $B(y)$ is skew-symmetric and satisfies

$$\sum_{l=1}^{n} \left(\frac{\partial b_{ij}(y)}{\partial y_l} b_{lk}(y) + \frac{\partial b_{jk}(y)}{\partial y_l} b_{li}(y) + \frac{\partial b_{ki}(y)}{\partial y_l} b_{lj}(y) \right) = 0, \qquad \text{for all } i, j, k, \tag{3.6}$$

then the formula

$$\{F, G\}(y) = \sum_{i,j=1}^{n} \frac{\partial F(y)}{\partial y_i} b_{ij}(y) \frac{\partial G(y)}{\partial y_j} \tag{3.7}$$

is said to represent a general *Poisson bracket.* The corresponding differential system (3.4) is a *Poisson system.* We continue to call H the Hamiltonian.

Since the Lotka-Volterra system can be written in the form (3.4), where $B(y)$, defined in (3.5), is skew-symmetric and satisfies (3.6), it is a Poisson system. To study such systems, the notion of Poisson maps is essential.

Definition 3.4. 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 (3.7), if its Jacobian matrix satisfies

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

We observe, of course, a similarity with symplectic maps. The following theorem, whose proof can be found in [5], explains the relation between Poisson systems and Poisson maps.

Theorem 3.5. *If $B(y)$ is the structure matrix of a Poisson bracket, the flow $\varphi_t(y)$ of the differential system*

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

is a Poisson map.

It would of course be interesting to choose numerical methods which exhibit the same characteristics as the flow $\varphi_t(y)$ when solving this kind of problems. This motivates the introduction of the notion of *Poisson integrators*, but before stating its definition we need to introduce the *Casimir functions.*

Theorem 3.6. *Suppose that the matrix $B(y)$ defines a Poisson bracket and is of constant rank $n - q = 2m$ in a neighbourhood of $y_0 \in \mathbb{R}^n$. Then, there exist functions $P_1(y), \dots, P_m(y)$, $Q_1(y), \dots, Q_m(y)$, and $C_1(y), \dots, C_q(y)$ satisfying*

$$\begin{array}{lll} \{P_i, P_j\} \;= 0 & \{P_i, Q_j\} = -\delta_{ij} & \{P_i, C_l\} = 0 \\ \{Q_i, P_j\} \;= \delta_{ij} & \{Q_i, Q_j\} = 0 & \{Q_i, C_l\} = 0 \\ \{C_k, P_j\} \;= 0 & \{C_k, Q_j\} = 0 & \{C_k, C_l\} = 0 \end{array}$$

on a neighbourhood of y_0. The gradients of P_i, Q_j, C_k are linearly independent, so that $y \mapsto (P_i(y), Q_i(y), C_k(y))$ constitutes a local change of coordinates to canonical form.

The proof of this theorem can be found in [5]. The functions C_k are called *Casimirs* and the flow $\varphi_t(y)$ of a Poisson system respects them in the sense that $C_i(\varphi_t(y)) =$ *Const.* This motivates the following definition.

Definition 3.7. A numerical method $y_1 = \Phi_h(y_0)$ is a *Poisson integrator* for the structure matrix $B(y)$, if the transformation $y_0 \mapsto y_1$ respects the Casimirs and if it is a Poisson map whenever the method is applied to the corresponding differential system (3.4).

In the case of the Lotka-Volterra system, the matrix $B(y)$ is of rank 2 for all $y = (u, v) \in \mathbb{D} = \{(u, v) \ : \ u > 0, v > 0\}$, so there is no Casimir function (since

$q = 0$) and a numerical method is a Poisson integrator for $B(y)$ if and only if it is a Poisson map whenever applied to the Poisson system (3.4), in other words we need it to satisfy

$$\left(\frac{\partial(u_{n+1}, v_{n+1})}{\partial(u_n, v_n)}\right)^T \begin{pmatrix} 0 & -u_n v_n \\ u_n v_n & 0 \end{pmatrix} \left(\frac{\partial(u_{n+1}, v_{n+1})}{\partial(u_n, v_n)}\right) = \begin{pmatrix} 0 & -u_{n+1} v_{n+1} \\ u_{n+1} v_{n+1} & 0 \end{pmatrix}. \tag{3.8}$$

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

3.3 Symplectic Euler

The main characteristic of the symplectic Euler method (2.1) is its symplecticity.

Theorem 3.8. *If the matrix $I + hH_{pq}$, where I is the identity and H_{pq} is the matrix of partial derivatives evaluated at (p_{n+1}, q_n), is invertible, then the symplectic Euler method* (2.1) *is symplectic. The condition is always satisfied for h small enough.*

Proof. We have to prove that this method is symplectic in the sense of the definition given in [5], that is we have to prove the symplecticity characterization (3.2).

Applying the symplectic Euler method to a smooth Hamiltonian system gives

$$\begin{cases} p_{n+1} = p_n - h\,\frac{\partial H}{\partial q}(p_{n+1}, q_n), \\ q_{n+1} = q_n + h\,\frac{\partial H}{\partial p}(p_{n+1}, q_n), \end{cases}$$

and differentiating these expressions with respect to p_n and q_n, we obtain

$$\begin{cases} \frac{\partial p_{n+1}}{\partial p_n} = I - h\frac{\partial^2 H}{\partial p \partial q}(p_{n+1}, q_n)\frac{\partial p_{n+1}}{\partial p_n}, \\ \frac{\partial p_{n+1}}{\partial q_n} = -h\frac{\partial^2 H}{\partial q \partial q}(p_{n+1}, q_n) - h\frac{\partial^2 H}{\partial p \partial q}(p_{n+1}, q_n)\frac{\partial p_{n+1}}{\partial q_n}, \\ \frac{\partial q_{n+1}}{\partial p_n} = h\frac{\partial^2 H}{\partial p \partial p}(p_{n+1}, q_n)\frac{\partial p_{n+1}}{\partial p_n}, \\ \frac{\partial q_{n+1}}{\partial q_n} = I + h\frac{\partial^2 H}{\partial q \partial p}(p_{n+1}, q_n) + h\frac{\partial^2 H}{\partial p \partial p}(p_{n+1}, q_n)\frac{\partial p_{n+1}}{\partial q_n}. \end{cases}$$

This system can be written as a matrix equation

$$\begin{pmatrix} I + hH_{qp}^T & 0 \\ -hH_{pp} & I \end{pmatrix} \begin{pmatrix} \frac{\partial p_{n+1}}{\partial p_n} & \frac{\partial p_{n+1}}{\partial q_n} \\ \frac{\partial q_{n+1}}{\partial p_n} & \frac{\partial q_{n+1}}{\partial q_n} \end{pmatrix} = \begin{pmatrix} I & -hH_{qq} \\ 0 & I + hH_{qp} \end{pmatrix} \tag{3.9}$$

where the matrices H_{qp}, H_{pp}, H_{qq} of partial derivatives are evaluated at (p_{n+1}, q_n). To simplify notations, we define $A := I + hH_{qp}$. Assuming that the first matrix in equation (3.9) is invertible, that is $\det A \neq 0$, we can compute the matrix of derivatives

$$\partial\Phi_h := \begin{pmatrix} \frac{\partial p_{n+1}}{\partial p_n} & \frac{\partial p_{n+1}}{\partial q_n} \\ \frac{\partial q_{n+1}}{\partial p_n} & \frac{\partial q_{n+1}}{\partial q_n} \end{pmatrix} = \begin{pmatrix} A^T & 0 \\ -hH_{pp} & I \end{pmatrix}^{-1} \begin{pmatrix} I & -hH_{qq} \\ 0 & A \end{pmatrix},$$

where

$$\begin{pmatrix} A^T & 0 \\ -hH_{pp} & I \end{pmatrix}^{-1} = \begin{pmatrix} A^{-T} & 0 \\ hH_{pp}A^{-T} & I \end{pmatrix}.$$

We can now compute the matrix product

$$\begin{aligned} &\left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}\right)^T J \left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}\right) \\ &= \begin{pmatrix} I & 0 \\ -hH_{qq}^T & A^T \end{pmatrix} \begin{pmatrix} A^{-1} & A^{-1}hH_{pp}^T \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} A^{-T} & 0 \\ hH_{pp}A^{-T} & I \end{pmatrix} \begin{pmatrix} I & -hH_{qq} \\ 0 & A \end{pmatrix} \\ &= \begin{pmatrix} I & 0 \\ -hH_{qq}^T & A^T \end{pmatrix} \begin{pmatrix} 0 & A^{-1} \\ -A^{-T} & 0 \end{pmatrix} \begin{pmatrix} I & -hH_{qq} \\ 0 & A \end{pmatrix} \\ &= \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \end{aligned}$$

and the symplecticity of the method is proved.

□

Since the Lotka-Volterra system is a Poisson system, we can check whether or not the symplectic Euler method is a Poisson integrator for this system.

Theorem 3.9. *The symplectic Euler method* (2.1) *is a Poisson integrator for Poisson systems with $B(y)$ defined in* (3.5) *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. We need to prove that the condition (3.8) is satisfied whenever we apply the symplectic Euler method to a system of the form

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

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 results 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, we have

$$\begin{pmatrix} 1 + hv_n(H_v - u_{n+1}H_{uv}) & 0 \\ -hv_n(H_u + u_{n+1}H_{uu}) & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{1+hv_n(H_v - u_{n+1}H_{uv})} & 0 \\ \frac{hv_n(H_u + u_{n+1}H_{uu})}{1+hv_n(H_v - u_{n+1}H_{uv})} & 1 \end{pmatrix}$$

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 $B(y)$ 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} by $u_n/(1 + hv_nH_v)$ and v_{n+1} by $v_n(1 + hu_{n+1}H_u)$ 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 $H(u, v) = T(u) + S(v)$. Since the Hamiltonian of the Lotka-Volterra system is separable, the theorem is proved. □

Since the symplectic Euler method is a Poisson integrator for the Lotka-Volterra system, we can expect it to give good numerical results. This explains the excellent performance we observed on Figure 2.4.

The symplectic Euler method (2.1), applied to the Lotka-Volterra system, gives

$$\begin{cases} u_{n+1} = \frac{u_n}{1-h(b-v_n)}, \\ v_{n+1} = v_n + h\,v_n(u_{n+1} - a). \end{cases} \tag{3.10}$$

Apart from the fact that it is a Poisson integrator, an important property of this method is that if we carefully choose h, the numerical result stays in the first quadrant. This property is essential as we pointed out in Chapter 1.

Theorem 3.10. *If we apply the symplectic Euler method* (2.1) *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. To prove the theorem, we suppose that u_n and v_n are positive and check under which conditions u_{n+1} and v_{n+1} are also positive.

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

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

Since we know that v_n is positive, if $b - 1/h$ is negative, the above inequality is satisfied. Therefore, if h is smaller than $1/b$, u_{n+1} is positive. This also guarantees that the denominator of the first equation in (3.10) never vanishes. On the other hand, v_{n+1} is positive if and only if $1 + h(u_{n+1} - a)$ is positive which implies

$$u_{n+1} > a - \frac{1}{h}. \tag{3.11}$$

We just established that u_{n+1} is positive if h is smaller than $1/b$, so under this condition and if $a - 1/h$ is negative, that is h smaller than $1/a$, the inequality (3.11) is satisfied and v_{n+1} is positive. Hence if we choose

$$h < \min\left\{\frac{1}{a}, \frac{1}{b}\right\},$$

u_n and v_n are positive for all $n \in \mathbb{N}$. □

In Figure 2.5, the step-size used, $h = 1.1$, is larger than the minimum of $1/a$ and $1/b$ since a and b are both equal to one. That is why we obtain a numerical approximation that leaves the first quadrant.

We now study the linear stability of the map. From the equations of the method (2.2), we compute the Jacobian

$$\nabla(f,g) = \begin{pmatrix} \frac{1}{1-hv(b-v)} & \frac{-uh}{[1-h(b-v)]^2} \\ \frac{hv}{1-h(b-v)} & 1 + h\left[\frac{u}{1-h(b-v)} - a\right] - \frac{uvh^2}{[1-h(b-v)]^2} \end{pmatrix}.$$

At the origin, this Jacobian becomes

$$\nabla(f,g) = \begin{pmatrix} \frac{1}{1-hb} & 0 \\ 0 & 1 - ha \end{pmatrix},$$

so the eigenvalues are $1/(1 - hb)$ and $1 - ha$. Since we have $1/|1 - hb| < 1$ for h between zero and $2/b$ and $|1 - ha| < 1$ for h between zero and $2/a$, the origin is a saddle point attractive along v and repulsive along u if

$$h < \min\left\{\frac{2}{a}, \frac{2}{b}\right\}. \tag{3.12}$$

However, if $2/b < h < 2/a$, we have a sink, if $2/a < h < 2/b$, we have a source, and if h is larger than $2/a$ and $2/b$, we obtain a saddle point attractive along u and repulsive along v. But, since we have to choose h smaller than the minimum of $1/a$ and $1/b$ in the symplectic Euler method to guarantee a positive trajectory, the condition (3.12) is satisfied and the origin is a saddle point in the numerical method.

The study of the behaviour close to the equilibrium point (a, b) is slightly more complicated. The Jacobian at that point is

$$\nabla(f, g) = \begin{pmatrix} 1 & -ah \\ hb & 1 - abh^2 \end{pmatrix},$$

whose characteristic polynomial is

$$P(\lambda) = \lambda^2 + (abh^2 - 2)\lambda + 1.$$

If $abh^2 - 4$ is positive, the two eigenvalues are

$$\lambda_{1,2} = 1 - \frac{h^2ab}{2} \pm \frac{1}{2}\sqrt{abh^2(abh^2 - 4)} \in \mathbb{R}.$$

Some manipulations yield

$$|\lambda_1| < 1 \text{ and } \lambda_2 < -1,$$

so that we obtain a saddle point. However, we are mostly interested in what we obtain for small values of h, and for h smaller than $2/\sqrt{ab}$, that is $abh^2 - 4 < 0$, we have

$$\lambda_{1,2} = 1 - \frac{h^2ab}{2} \pm i\frac{1}{2}\sqrt{abh^2(4 - abh^2)} \in \mathbb{C}.$$

One can show that $|\lambda|^2 = 1$, which means that the equilibrium point is stable but not asymptotically stable and the solutions are rotating around it. Here again, the condition for the positivity of the numerical trajectory of the symplectic Euler method, i.e. $h < 1/a$ and $h < 1/b$, is stronger than the condition given by the linear stability, i.e. $h < 2/\sqrt{ab}$.

3.4 Explicit Variant of Symplectic Euler

Now, we study the explicit variant of the symplectic Euler method, defined by (2.3). This method is not symplectic in general.

Theorem 3.11. *The explicit variant of symplectic Euler method* (2.3) *is symplectic for separable Hamiltonians* $H(p,q) = S(p) + T(q)$.

Proof. We study the symplecticity of this method by checking whether or not the condition (3.2) is satisfied. Applying the explicit variant of symplectic Euler method to a smooth Hamiltonian system. gives

$$\begin{cases} p_{n+1} = p_n - h\,\frac{\partial H}{\partial q}(p_n, q_n), \\ q_{n+1} = q_n + h\,\frac{\partial H}{\partial p}(p_{n+1}, q_n), \end{cases}$$

and differentiating these expressions with respect to p_n and q_n, we obtain

$$\begin{cases} \frac{\partial p_{n+1}}{\partial p_n} = I - h\frac{\partial^2 H}{\partial p \partial q}(p_n, q_n), \\ \frac{\partial p_{n+1}}{\partial q_n} = -h\frac{\partial^2 H}{\partial q \partial q}(p_n, q_n), \\ \frac{\partial q_{n+1}}{\partial p_n} = h\frac{\partial^2 H}{\partial p \partial p}(p_{n+1}, q_n)\frac{\partial p_{n+1}}{\partial p_n}, \\ \frac{\partial q_{n+1}}{\partial q_n} = I + h\frac{\partial^2 H}{\partial q \partial p}(p_{n+1}, q_n) + h\frac{\partial^2 H}{\partial p \partial p}(p_{n+1}, q_n)\frac{\partial p_{n+1}}{\partial q_n}. \end{cases}$$

This system can be written as the matrix equation

$$\begin{pmatrix} I & 0 \\ -hH_{pp} & I \end{pmatrix} \begin{pmatrix} \frac{\partial p_{n+1}}{\partial p_n} & \frac{\partial p_{n+1}}{\partial q_n} \\ \frac{\partial q_{n+1}}{\partial p_n} & \frac{\partial q_{n+1}}{\partial q_n} \end{pmatrix} = \begin{pmatrix} I - hH_{pq} & -hH_{qq} \\ 0 & I + hH_{qp} \end{pmatrix}$$

where the matrices H_{qp}, H_{pp} of partial derivatives are evaluated at (p_{n+1}, q_n), whereas the matrices H_{pq}, H_{qq} are evaluated at (p_n, q_n). Since the first matrix of the equation is invertible with

$$\begin{pmatrix} I & 0 \\ -hH_{pp} & I \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ hH_{pp} & I \end{pmatrix}$$

we can compute the matrix of derivatives $\partial\Phi_h$ and then, using the fact that H_{pp} and H_{qq} are symmetric, and denoting by $A := I - hH_{pq}$, we obtain

$$\begin{aligned}
&\left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}\right)^T J \left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}\right) \\
&= \begin{pmatrix} A & -hH_{qq} \\ 0 & I + hH_{qp} \end{pmatrix}^T \begin{pmatrix} I & 0 \\ hH_{pp} & I \end{pmatrix}^T \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} I & 0 \\ hH_{pp} & I \end{pmatrix} \begin{pmatrix} A & -hH_{qq} \\ 0 & I + hH_{qp} \end{pmatrix} \\
&= \begin{pmatrix} A & -hH_{qq} \\ 0 & I + hH_{qp} \end{pmatrix}^T \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} A & -hH_{qq} \\ 0 & I + hH_{qp} \end{pmatrix} \\
&= \begin{pmatrix} 0 & (I - hH_{pq}^T)(I + hH_{qp}) \\ -(I + hHqp^T)(I - hH_{pq}) & h^2(H_{qp}^T H_{qq} - H_{qq}H_{qp}) \end{pmatrix}
\end{aligned}$$

and the method is symplectic if and only if

$$\begin{cases} (I - hH_{pq}^T)(I + hH_{qp}) = (I + hH_{qp}^T)(I - hH_{pq}) = I, \\ H_{qp}^T H_{qq} - H_{qq}H_{qp} = 0. \end{cases}$$

Considering that the matrix H_{pq} is composed of partial derivatives evaluated at (p_n, q_n) whereas the matrix H_{qp} is composed of partial derivatives evaluated at (p_{n+1}, q_n), these equalities are not satisfied in general and thus the method is not symplectic. However, if we consider a separable Hamiltonian $H(p, q) = S(p) + T(q)$, we have $H_{pq} = H_{qp} = 0$ and the above equations are satisfied (this is not surprising since in this situation, this method is equivalent to the symplectic Euler method).

□

Now we want to see if the explicit variant of the symplectic Euler method is a Poisson integrator for the Lotka-Volterra system.

Theorem 3.12. *The explicit variant of symplectic Euler method* (2.3) *is a Poisson integrator for Poisson systems with $B(y)$ defined in* (3.5) *and any separable Hamiltonian H.*

Proof. We have to prove that method is a Poisson integrator for the matrix $B(y)$ defined in (3.5), that is the equation (3.8) is satified whenever the method is applied to a problem of the form

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

We first differentiate the expressions giving u_{n+1} and v_{n+1}

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

with respect to u_n and v_n and write the resulting equations as one matrix equation

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

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 $A := u_n v_n(1 - hv_n(H_v + u_n H_{uv}))(1 + hu_{n+1}(H_u + v_n H_{uv}))$. We still have to check whether A equals $u_{n+1}v_{n+1}$. We have on one side

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

and on the other side

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

thus for any separable Hamiltonian, the explicit variant of the symplectic Euler method is a Poisson integrator for $B(y)$ defined in (3.5). □

An illustration of the method in Figure 2.7 shows that in general the numerical result is close to the exact solution and exhibits the right qualitative behaviour.

The explicit variant applied to the Lotka-Volterra system gives

$$\begin{cases} u_{n+1} &= u_n + h\, u_n(b - v_n), \\ v_{n+1} &= v_n + h\, v_n(u_{n+1} - a). \end{cases} \tag{3.13}$$

An essential property of the original symplectic Euler method is that under a simple condition on the step-size h, the numerical results are positive for any n. We now check if such a condition can also be given for the explicit variant of the symplectic Euler method.

Following the same arguments as for the symplectic Euler method, we end up with the same condition for v_{n+1}, however 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. One can see on Figure 2.8 an example where the solution leaves the first quadrant.

Actually if we plot for $h = 1$ the number of iterations needed to leave the first quadrant for every initial values, the figure obtained, Figure 2.9, is estetically pleasing and very complicated. One can note that the condition

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

appears clearly on the figure. If we choose h smaller, so that the condition $h < 1/a$ is satisfied, we still obtain a similar structure as one can see on Figure 3.1. From these figures, we suspect that, for given initial conditions, it is always possible to find a step-size h for which the numerical results stay positive. We later prove such a result for exponentially long-time intervals.

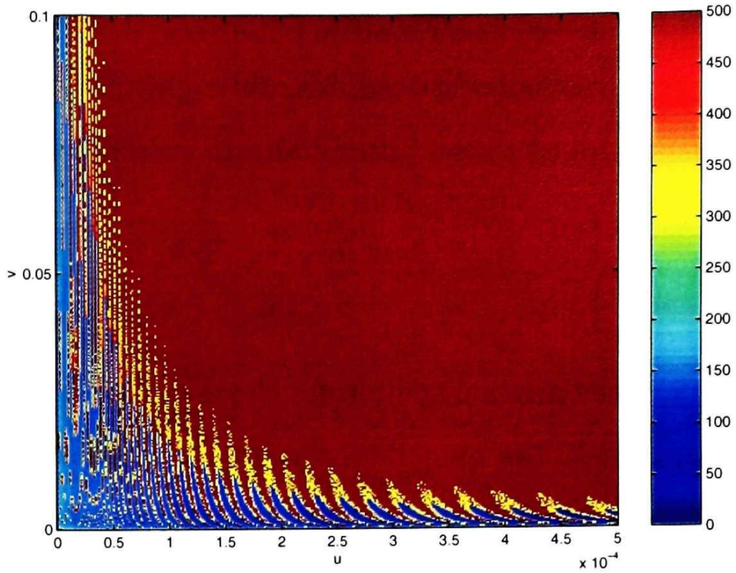

Figure 3.1: Number of iterations needed for each point (u_0, v_0) to leave the first quadrant when applying the explicit variant of the symlectic Euler method to the Lotka-Volterra system with h=0.1 and $a = b = 1$.

We now study the linear stability of this method. From the equations (3.13), we can compute the Jacobian

$$\nabla(f,g) = \begin{pmatrix} 1 + h(b-v) & -uh \\ hv(1+h(b-v)) & 1 + h[u + hu(b-v) - a] - uvh^2 \end{pmatrix},$$

which becomes at the origin

$$\nabla(f,g) = \begin{pmatrix} 1+hb & 0 \\ 0 & 1-ha \end{pmatrix}.$$

Since $|1+hb|$ is larger than one whenever h is positive and $|1-ha|$ is smaller than one for h between zero and $2/a$, we obtain a saddle point attracting along v and repulsive along u for $h < 2/a$. Otherwise, we have a source.

Since the Jacobian at the equilibrium point is the same as for the symplectic Euler method, that is

$$\nabla(f,g) = \begin{pmatrix} 1 & -ah \\ hb & 1 - abh^2 \end{pmatrix},$$

the conclusions are the same : for h smaller than $2/\sqrt{ab}$, the numerical solution is rotating around the equilibrium point. We will see later that we need to choose h much smaller than $2/a$ or $2/\sqrt{ab}$ to ensure a positive numerical solution, thus we will have a saddle point at the origin and a center at the equilibrium point.

Chapter 4

Backward Error Analysis

In this chapter, we introduce the notion of backward error analysis, a very useful tool to study the qualitative behaviour of numerical methods over long time intervals. The idea of backward error analysis is to search for a *modified* differential equation of the form

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + h^2 f_3(\tilde{y}) + \dots, \tag{4.1}$$

such that the solution $\tilde{y}$ of this modified equation corresponds to the numerical solution of $\dot{y} = f(y)$, that is $y_n = \tilde{y}(nh)$. Of course the modified equation depends on the method applied and usually, the series in (4.1) diverges, so one has to truncate it suitably.

4.1 Properties of Symplectic Methods and Poisson Integrators

The most important property of symplectic methods is that if such a method is applied to a Hamiltonian system with a smooth Hamiltonian, then the modified equation (4.1) is also Hamiltonian. Before stating this result we need to prove an important Lemma, often called *Integrability Lemma.* The proof we give is essentially the same as in [5].

Lemma 4.1. *Let $D \subset \mathbb{R}^n$ be open and $f : D \to \mathbb{R}^n$ be continuously differentiable, and assume that the Jacobian $f'(y)$ is symmetric for all $y \in D$. Then, for every $y_0 \in D$ there exists a neighbourhood of y_0 and a function $H(y)$ such that*

$$f(y) = \nabla H(y)$$

on this neighbourhood.

Proof. Consider $y_0 \in D$ and a ball around y_0 which is contained in D. Then we define on this ball

$$H(y) = \int_0^1 (y - y_0)^T f(y_0 + t(y - y_0))dt + Const. \tag{4.2}$$

Differentiating H with respect to y_k, the kth component of the vector y, and using the symmetry assumption $\frac{\partial f_i}{\partial y_k} = \frac{\partial f_k}{\partial y_i}$ (which implies $\nabla f_k = \frac{\partial f}{\partial y_k}$) yields

$$\begin{aligned}\frac{\partial H}{\partial y_k}(y) &= \int_0^1 f_k(y_0 + t(y - y_0)) + (y - y_0)^T \frac{\partial f}{\partial y_k}(y_0 + t(y - y_0))\, t\, dt \\ &= \int_0^1 \frac{d}{dt}(t\, f_k(y_0 + t(y - y_0)))dt \\ &= f_k(y),\end{aligned}$$

which proves the lemma. □

The important point of this proof is that it shows that for star-shaped regions D, or convex sets D, the function H is globally defined : we fix y_0 such that for all y in D and all t between zero and one, we have

$$y_0 + t(y - y_0) \in D$$

and H defined by (4.2) is thus defined on all D.

Theorem 4.1. *If a symplectic method $\Phi_h(y)$ is applied to a Hamiltonian system with a smooth Hamiltonian $H : D \subset \mathbb{R}^{2d} \to \mathbb{R}$, where D is simply connected, then* *the modified equation (4.1) is also Hamiltonian. More precisely, there exist smooth functions $H_j : D \to \mathbb{R}$ for $j = 2, 3, \ldots$, such that $f_j(y) = J^{-1}\nabla H_j(y)$.*

This result can be generalized to any arbitrary open set D, however since the Lotka-Voltera system is defined on a convex set, we don't need to study further symplectic methods. The proof of this theorem and the proofs of the following ones can be found in [5]. As stated in the following theorems, the previous result can be generalized to Poisson integrators.

Theorem 4.2. *If a Poisson integrator $\Phi_h(y)$ is applied to the Poisson system (3.4), then the modified equation is locally a Poisson system. More precisely, for every $y_0 \in \mathbb{R}^n$ there exist a neighbourhood U and smooth functions $H_j : U \to \mathbb{R}$ such that on U, the modified equation is of the form*

$$\dot{\tilde{y}} = B(\tilde{y})(\nabla H(\tilde{y}) + h\nabla H_2(\tilde{y}) + \ldots). \tag{4.3}$$

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 4.3. *If $H(y)$ and $B(y)$ are defined and smooth on a simply connected domain D, and if $B(y)$ is invertible on D, then a Poisson integrator $\Phi_h(y)$ has a modified equation* (4.3) *with smooth functions $H_j(y)$ defined on all of D.*

Since for the Lotka-Volterra system, the matrix $B(y)$ is invertible on $\mathbb{D} = \{y = (u, v) : u > 0, v > 0\}$, 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.

4.2 The Symplectic Euler Method

4.2.1 First Order Term

In this section, we derive the first order term of the numerical Hamiltonian corresponding to the symplectic Euler method applied to the Lotka-Volterra system,

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

The first step is to find what the method gives when we expand $u(t_{n+1}) = u(t_n+h)$, for h small. We consider

$$u(t_{n+1}) = \frac{u(t_n)}{1-h(b-v(t_n))},$$

and using Taylor series

$$f(\varepsilon) = \frac{1}{1-\varepsilon} = \sum_{n\geq 0} \varepsilon^n,$$

we obtain the expansion

$$u(t_{n+1}) = u(t_n) \sum_{p\geq 0} \left[\, h(b-v(t_n))\, \right]^p, \tag{4.4}$$

which means the term of order h^p is $u(b-v)^p\, h^p$.

To find the first order term of the numerical Hamiltonian, we use the *ansatz*

$$\begin{cases} \dot{u} &= f(u,v) + h\, f_2(u,v), \\ \dot{v} &= g(u,v) + h\, g_2(u,v), \end{cases}$$

where $f(u,v) = u(b-v)$ and $g(u,v) = v(u-a)$, and substitute it into the Taylor expansion of $u(t_{n+1})$,

$$\begin{aligned}
u(t_{n+1}) &= u(t_n) + h\dot{u}(t_n) + \frac{h^2}{2}\ddot{u}(t_n) + O(h^3)\\
&= u(t_n) + hf(u(t_n),v(t_n)) + h^2 f_2(u(t_n),v(t_n))\\
&\quad + \frac{h^2}{2}\left(\frac{\partial f}{\partial u}(u(t_n),v(t_n))\dot{u} + \frac{\partial f}{\partial v}(u(t_n),v(t_n))\dot{v}\right) + O(h^3),\\
&= u(t_n) + hf(u(t_n),v(t_n)) + h^2 f_2(u(t_n),v(t_n))\\
&\quad + \frac{h^2}{2}\left(\frac{\partial f}{\partial u}(u(t_n),v(t_n))f(u(t_n),v(t_n))\right.\\
&\qquad\qquad \left. + \frac{\partial f}{\partial v}(u(t_n),v(t_n))g(u(t_n),v(t_n))\right) + O(h^3).
\end{aligned}$$

Comparing this expansion with the one of the method (4.4) that we can write as

$$u(t_{n+1}) = u + hf(u,v) + h^2 f(u,v)\frac{\partial f}{\partial u}(u,v) + O(h^3),$$

we find that for an $O(h^3)$ residual, we need f_2 to satisfy

$$f_2(u,v) + \frac{1}{2}\left(\frac{\partial f}{\partial u}(u,v)f(u,v) + \frac{\partial f}{\partial v}(u,v)g(u,v)\right) = f(u,v)\frac{\partial f}{\partial u}(u,v),$$

in other words

$$f_2(u,v) = \frac{1}{2}\left(\frac{\partial f}{\partial u}f - \frac{\partial f}{\partial v}g\right) = \frac{1}{2}\ [u(b-v)^2 + uv(u-a)].$$

Similarly we obtain the Taylor expansion for v_{n+1} ,

$$\begin{aligned}
v(t_{n+1}) &= v(t_n) + hg(u(t_n),v(t_n)) + h^2 g_2(u(t_n),v(t_n))\\
&\quad + \frac{h^2}{2}\left(\frac{\partial g}{\partial u}(u(t_n),v(t_n))f(u(t_n),v(t_n))\right.\\
&\qquad\qquad \left. + \frac{\partial g}{\partial v}(u(t_n),v(t_n))g(u(t_n),v(t_n))\right) + O(h^3),
\end{aligned}$$

and for v the method gives

$$\begin{aligned}
v(t_{n+1}) &= v(t_n) + hv(t_n)[u(t_{n+1}) - a],\\
&= v(t_n) + hv(t_n)[u + u(b-v)h + u(b-v)^2h^2 + \cdots - a]\\
&= v(t_n) + hv(u-a) + h^2uv(b-v) + h^3uv(b-v)^2 + \ldots,
\end{aligned}$$

that is, the term of order h^p is $[uv(b-v)^{p-1}]h^p$. Therefore, to obtain an $O(h^3)$ residual, we must have

$$g_2(u,v) + \frac{1}{2}\left(\frac{\partial g}{\partial u}(u,v)f(u,v) + \frac{\partial g}{\partial v}(u,v)g(u,v)\right) = \frac{\partial g}{\partial u}(u,v)f(u,v),$$

or

$$g_2(u,v) = \frac{1}{2}\left(\frac{\partial g}{\partial u}f - \frac{\partial g}{\partial u}g\right) = \frac{1}{2}[uv(b-v) - v(u-a)^2].$$

Putting these results together, we obtain the modified equation

$$\begin{cases} \dot{u} = u(b-v) + \frac{h}{2}\,[\,u(b-v)^2 + uv(u-a)\,], \\ \dot{v} = v(u-a) + \frac{h}{2}\,[\,uv(b-v) - v(u-a)^2\,]. \end{cases} \tag{4.5}$$

To obtain the numerical Hamiltonian, some algebra is needed. We first divide the two equations of the modified system,

$$\frac{\dot{u}}{\dot{v}} = \frac{u(b-v) + \frac{h}{2}\,[\,u(b-v)^2 + uv(u-a)\,]}{v(u-a) + \frac{h}{2}\,[\,uv(b-v) - v(u-a)^2\,]},$$

to obtain

$$du\left(v(u-a) + \frac{h}{2}\,[\,uv(b-v) - v(u-a)^2\,]\right) = dv\left(u(b-v) + \frac{h}{2}\,[\,u(b-v)^2 + uv(u-a)\,]\right).$$

Dividing this equality by uv, we get

$$du\left(1 - \frac{a}{u} + \frac{h}{2}\,[\,2a + b\boxed{-v} - u - \frac{a^2}{u}\,]\right) + dv\left(1 - \frac{b}{v} + \frac{h}{2}\,[\,a\boxed{-u} - \frac{b^2}{v} + 2b - v\,]\right) = 0$$

and integrating it, we obtain the Hamiltonian of the modified system (4.5)

$$\begin{aligned} H_h(u,v) &= u - a\ln u + \frac{h}{2}(2ua + bu\boxed{-uv} - \frac{u^2}{2} - a^2\ln u) + v - b\ln v \\ &\quad + \frac{h}{2}(av - b^2\ln v + 2bv - \frac{v^2}{2}). \\ &= H(u,v) - \frac{h}{2}[\frac{u^2}{2} + uv + \frac{v^2}{2} - (2a+b)u - (a+2b)v \\ &\quad + a^2\ln u + b^2\ln v]. \end{aligned}$$

We summarize these results in the following lemma.

Lemma 4.2. *If the symplectic Euler method is applied to the Lotka-Volterra system, the modified system is of the form*

$$\begin{cases} \dot{\tilde{u}} & = f(\tilde{u},\tilde{v}) + hf_2(\tilde{u},\tilde{v}) + O(h^2), \\ \dot{\tilde{v}} & = g(\tilde{u},\tilde{v}) + hg_2(\tilde{u},\tilde{v}) + O(h^2), \end{cases}$$

with

$$f_2(\tilde{u},\tilde{v}) = \frac{1}{2}[\tilde{u}(b-\tilde{v})^2 + \tilde{u}\tilde{v}(\tilde{u}-a)],$$

and

$$g_2(\tilde{u},\tilde{v}) = \frac{1}{2}\,[\tilde{u}\tilde{v}(b-\tilde{v}) - (\tilde{u}-a)^2\tilde{v}].$$

Furthermore an invariant of this modified system is

$$H_h(\tilde{u},\tilde{v}) = H(\tilde{u},\tilde{v}) - \frac{h}{2}[\frac{\tilde{u}^2}{2} + \tilde{u}\tilde{v} + \frac{\tilde{v}^2}{2} - (2a+b)\tilde{u} - (a+2b)\tilde{v} \\ + a^2\ln\tilde{u} + b^2\ln\tilde{v}] + O(h^2).$$

Figure 4.1 shows an illustration of the numerical Hamiltonian of order one.

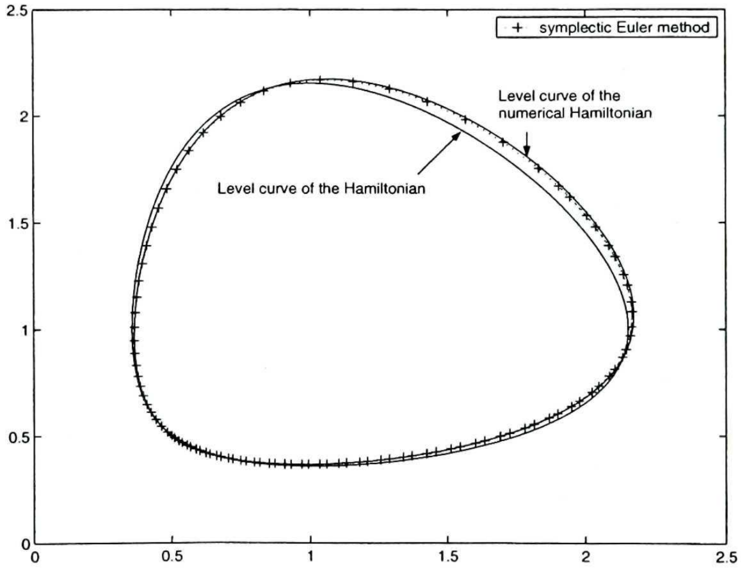

Figure 4.1: Illustration of the numerical Hamiltonian of first order of the symplectic Euler method applied to the Lotka-Volterra system with $h = 0.1$.

4.2.2 Second order term

Following the same steps as in the previous section, we derive the second-order term of the numerical Hamiltonian. To do so, we use the *ansatz*

$$\begin{cases} \dot{u} &= f(u,v) + h\, f_2(u,v) + h^2 f_3(u,v), \\ \dot{v} &= g(u,v) + h\, g_2(u,v) + h^2 g_3(u,v), \end{cases} \tag{4.6}$$

where f_2 and g_2 are those computed in the previous section. To compute f_3 and g_3 we need the Taylor expansions of $u(t_{n+1})$ and $v(t_{n+1})$ up to order h^4. From the expression giving $\dot{u}$ in the *ansatz* (4.6), we obtain $\ddot{u} = f'(u,v) + hf_2'(u,v) + O(h^2)$ and $\dddot{u} = f''(u,v) + O(h)$ and thus (omitting t_n, u and v when there is no ambiguity)

$$\begin{aligned} u(t_{n+1}) &= u(t_n) + h\dot{u}(t_n) + \frac{h^2}{2}\ddot{u}(t_n) + \frac{h^3}{6}\dddot{u}(t_n) + O(h^4) \\ &= u + hf + h^2 f_2 + \frac{h^2}{2}\left[\frac{\partial f}{\partial u} f + \frac{\partial f}{\partial v} g\right] \\ &\quad + h^3\Bigg(f_3 + \frac{1}{2}\left[\frac{\partial f}{\partial u} f_2 + \frac{\partial f}{\partial v} g_2 + \frac{\partial f_2}{\partial u} f + \frac{\partial f_2}{\partial v} g\right] \\ &\quad + \frac{1}{6}\Bigg[\frac{\partial^2 f}{\partial u^2} f^2 + 2\frac{\partial^2 f}{\partial u \partial v} g + \frac{\partial^2 f}{\partial v^2} g^2 + \left(\frac{\partial f}{\partial u}\right)^2 f \\ &\qquad + \frac{\partial f}{\partial u}\frac{\partial f}{\partial v} g + \frac{\partial f}{\partial v}\frac{\partial g}{\partial u} f + \frac{\partial f}{\partial v}\frac{\partial g}{\partial v} g\Bigg]\Bigg) + O(h^4). \end{aligned}$$

This result (and the following ones) can be easily obtained using Maple. To obtain a residual of order h^4, we simply have to set

$$f_3(u,v) = \frac{1}{3}u(b-v)^3 + uv\left(-\frac{1}{3}(u-a)^2 + \frac{1}{2}(u-a)(b-v) + \frac{1}{6}u(b-v)\right).$$

Similarly, we obtain the expression for g_3,

$$g_3(u,v) = \frac{1}{3}v(u-a)^3 + uv\left(\frac{1}{3}(b-v)^2 - \frac{1}{2}(b-v)(u-a) + \frac{1}{6}v(u-a)\right).$$

The next step is to consider $\dot{u}/\dot{v}$, where $\dot{u}$ and $\dot{v}$ are given by the *ansatz* (4.6) and f_3 and g_3 are the ones we just derived. After simplifications, we obtain the numerical

Hamiltonian

$$\begin{aligned}H_h = H(u,v) &- \frac{h}{2}\left[\frac{u^2}{2} + uv + \frac{v^2}{2} - (2a+b)u - (a+2b)v + a^2\ln u + b^2\ln v\right] \\ &+ \frac{h^2}{3}\left[\frac{u^3}{3} + u^2v + uv^2 + \frac{v^3}{3} - (3a + \frac{3}{2}b)\frac{u^2}{2} - (2a+2b)uv - (\frac{3}{2}a + 3b)\frac{v^2}{2}\right. \\ &\left. + (3a^2 + \frac{3}{2}ab + b^2)u + (a^2 + \frac{3}{2}ab + 3b^2)v - a^3\ln u - b^3\ln(v)\right] + O(h^3).\end{aligned}$$

Higher order terms can be computed following the same procedure.

4.3 The Structure of the Numerical Hamiltonian

As one can see from the expansion of the numerical Hamiltonian we just derived, it seems that each term of the expansion consists of a sum of a term in $\ln u$, one in $\ln v$ and a polynomial in u and v. It is interesting to study the structure of this expansion further, which we do in this section. More precisely, we prove that the term of order n of the expansion is of the form

$$-\frac{h^n}{n+1}\,(\,a^{n+1}\ln u + b^{n+1}\ln v + \text{polynomial in } u \text{ and } v\,). \tag{4.7}$$

Theorem 4.4. *When we use the ansatz*

$$\begin{cases} \dot u &= f + hf_2 + h^2 f_3 + \ldots + h^n f_{n+1}, \\ \dot v &= g + fg_2 + h^2 g_3 + \ldots + h^n g_{n+1}. \end{cases} \tag{4.8}$$

the coefficients f_{n+1} and g_{n+1} are of the form

$$f_{n+1} = \frac{1}{n+1}u(b-v)^{n+1} + uv \times P_{n+1}(u,v) \tag{4.9}$$

and

$$g_{n+1} = \frac{(-1)^n}{n+1}v(u-a)^{n+1} + uv \times Q_{n+1}(u,v) \tag{4.10}$$

where P_{n+1} and Q_{n+1} are polynomials.

Once this theorem is proved, a simple manipulation yields the numerical Hamiltonian. However before proving this theorem, we first need to establish several lemmas.

4.3.1 Notations

To simplify the expansion of f_{n+1} and g_{n+1}, we introduce some notation. We show the details only for the functions f_i, similar notations can be easily deduced for the functions g_i. We may often, to simplify formulas, denote f by f_1, so that f_i is well-defined for $i = 1, ..., n+1$. Usually we use the notation $\partial_u f$ to denote the partial derivative of f with respect to u, whereas the dot and the prime correspond to the derivative with respect to t.

From the *ansatz* (4.8), one can find the higher order derivatives of u with respect to t :

$$\ddot{u} = f_1' + hf_2' + \ldots + h^n f_{n+1}', \qquad \dddot{u} = f_1'' + hf_2'' + \ldots + h^n f_{n+1}'', \qquad \text{etc.}$$

Since f_i is a function of u and v, we have

$$f_i' = \partial_u f_i \, \dot{u} + \partial_v f_i \, \dot{v}, \tag{4.11}$$

which becomes, when we substitute the *ansatz* (4.8) into it,

$$f_i' = f \, \partial_u f_i + g \, \partial_v f_i + h(f_2 \, \partial_u f_i + g_2 \, \partial_v f_i) + \ldots + h^n (f_{n+1} \, \partial_u f_i + g_{n+1} \, \partial_v f_i). \tag{4.12}$$

Since each derivative of the function f_i with respect to t is a partial sum of a series in h, we introduce a new notation, $f_{i,j}'$, to denote each term of the sum,

$$f_i' = f_{i,1}' + hf_{i,2}' + \ldots + h^n f_{i,n+1}' = \sum_{j=0}^{n} h^j f_{i,j+1}'.$$

Similarly, we introduce the same notation for the higher derivatives,

$$f_i^{(k)} = f_{i,1}^{(k)} + hf_{i,2}^{(k)} + \ldots + h^n f_{i,n+1}^{(k)} = \sum_{j=0}^{n} h^j f_{i,j+1}^{(k)}, \qquad \text{for } k = 1...n. \tag{4.13}$$

In other words, $f_{i,j}^{(k)}$ corresponds to the term of order h^{j-1} of the kth derivative of f_i with respect to t.

4.3.2 Derivation of f_{n+1} and g_{n+1}

The first step to prove Theorem 4.4 is to express f_{n+1} and g_{n+1} as functions of $f_{i,j}^{(k)}$ and $g_{i,j}^{(k)}$.

Lemma 4.3. *In the notation of* (4.13), *we have*

$$f_{n+1} = u(b-v)^{n+1} - \sum_{k=1}^{n} \frac{1}{(k+1)!}\left(\sum_{j=1}^{n-k+1} f_{j,n-k-j+2}^{(k)}\right), \qquad \textit{for } n \geq 1$$

and

$$g_{n+1} = uv(b-v)^{n} - \sum_{k=1}^{n} \frac{1}{(k+1)!}\left(\sum_{j=1}^{n-k+1} g_{j,n-k-j+2}^{(k)}\right), \qquad \textit{for } n \geq 1.$$

Proof. Using the notation introduced in (4.13), one can write down explicitly the expansion of $u(t_{n+1})$:

$$\begin{aligned}
u(t_{n+1}) &= u + h\dot{u} + \frac{h^2}{2!}\ddot{u} + \ldots + \frac{h^n}{n!}u^{(n)} + \frac{h^{n+1}}{(n+1)!}u^{(n+1)} + O(h^{n+2}) \\
&= u + h(f + hf_2 + h^2 f_3 + \ldots + h^n f_{n+1}) \\
&\quad + \frac{h^2}{2!}(f' + hf_2' + \ldots + h^{n-1}f_n') \\
&\quad \vdots \\
&\quad + \frac{h^n}{n!}(f^{(n-1)} + hf_2^{(n-1)}) \\
&\quad + \frac{h^{n+1}}{(n+1)!}(f^{(n)}) \quad + O(h^{n+2}) \\
&= u + hf_1 + h^2\left(f_2 + \frac{1}{2!}f_{1,1}'\right) + h^3\left(f_3 + \frac{1}{2!}(f_{1,2}' + f_{2,1}') + \frac{1}{3!}f_{1,1}''\right) \\
&\quad + h^4\left(f_4 + \frac{1}{2!}(f_{1,3}' + f_{2,2}' + f_{3,1}') + \frac{1}{3!}(f_{1,2}'' + f_{2,1}'') + \frac{1}{4!}f_{1,1}^{(3)}\right) \\
&\quad \vdots \\
&\quad + h^{n+1}\Bigg(f_{n+1} + \frac{1}{2!}(f_{1,n}' + f_{2,n-1}' + \ldots + f_{n,1}') + \frac{1}{3!}(f_{1,n-1}'' \\
&\quad + \ldots + f_{n-1,1}'') + \ldots + \frac{1}{n!}(f_{1,2}^{(n-1)} + f_{2,1}^{(n-1)}) + \frac{1}{(n+1)!}f_{1,1}^{(n)}\Bigg) \\
&\quad + O(h^{n+2}).
\end{aligned}$$

From this expansion, we obtain an explicit expression for f_{n+1} as a function of the derivatives of f_i, $i = 1, ..., n$: since the method gives

$$u_{n+1} = u_n \sum_{p \geq 0} [h\,(b - v_n)]^p,$$

we should have

$$f_{n+1} = u(b-v)^{n+1} - \frac{1}{2!}(f'_{1,n} + \ldots + f'_{n,1}) - \ldots - \frac{1}{n!}(f^{(n-1)}_{1,2} + f^{(n-1)}_{2,1}) - \frac{1}{(n+1)!} f^{(n)}_{1,1}, \quad (4.14)$$

for $n \geq 1$. A similar argument is used to derive g_{n+1}. The only difference comes from the expansion of the method; we have

$$v_{n+1} = v_n + hv(u-a) + \sum_{p \geq 2} [uv(b-v)^{p-1}]h^p,$$

therefore, we obtain, for $n \geq 1$,

$$g_{n+1} = uv(b-v)^n - \frac{1}{2!}(g'_{1,n} + \ldots + g'_{n,1}) - \ldots - \frac{1}{n!}(g^{(n-1)}_{1,2} + g^{(n-1)}_{2,1}) - \frac{1}{(n+1)!} g^{(n)}_{1,1},$$

and the lemma is proved. □

This result, although interesting, is not directly usable, since the derivatives of f_i and g_i become more and more complicated. A method enabling us to find the functions f_i and g_i explicitly is to express the functions in terms of trees. More details and results about this method can be found in [5] (Theorem 9.4 on page 319). In the next chapter we give an expansion of f_{n+1} as a function of Lie derivatives. Yet, even if we are not able to find f_{n+1} and g_{n+1} explicitly in a simple way, we can use those results to study the structure of the functions f_i and g_i and prove Theorem 4.4. The advantage of this notation is that it is easier to follow exactly what each term corresponds to.

4.3.3 The Structure of $f_{i,j}^{(k)}$ and $g_{i,j}^{(k)}$

From (4.14), we see that f_{n+1} consists of a linear combination of derivatives with respect to t of the f_i. When we replace these derivatives by their counterparts composed of derivatives with respect to u and v, we obtain a polynomial whose terms are a product of f_i and g_i and their derivatives, but the essential observation is that, as soon as there is at least one derivative with respect to u in a product, then at least one f_i appears in this product, and similarly a factor g_i is necessary if the product contains a derivative with respect to v. We have exactly the same result for g_{n+1}.

By our induction hypothesis, we have for $k = 1, ..., n$,

$$\partial_u f_k = \frac{1}{k}(b-v)^k + v \times \text{polynomial of } (u,v),$$

$$\partial_v f_k = -u(b-v)^{k-1} + u \times \text{polynomial of } (u,v),$$

and

$$\partial_u g_k = (-1)^{k-1} v(u-a)^{k-1} + v \times \text{polynomial of } (u,v),$$

$$\partial_v g_k = \frac{(-1)^{k-1}}{k}(u-a)^k + u \times \text{polynomial of } (u,v).$$

Then, the higher derivatives with respect to u (and only u) of f_k and g_k are all of the form $v \times$ polynomial of (u,v), and the higher derivatives with respect to v (and only v) of f_k and g_k are all of the form $u \times$ polynomial of (u,v).

Because we want to prove (4.9) and (4.10), we are only interested in terms that do not contain the product $u \times v$: any term containing this product is included in the second part of (4.9) or (4.10). We just saw that any product containing a derivative with respect to u contains also one f_i and any product containing a derivative with respect to v contains also one g_i, thus any product containing a derivative of u *and* v contains also one f_i and one g_i. Moreover, since, by the induction hypothesis, f_i is of the form $u \times$ a polynomial of (u,v) and g_i is of the form $v \times$ a polynomial of (u,v), we know that any product containing a derivative with respect to u and v is of the form $u \times v \times$ polynomial of (u,v).

Moreover, apart from $\partial_u f_i$, the derivatives of f_i and g_i with respect to u are of the form $v \times$ polynomial of (u, v), so multiplied by a function f_i we obtain again $u \times v \times$ a polynomial of (u, v). Similarly, the derivatives of f_i and g_i with respect to v are, apart from $\partial_v g_i$, of the form $u \times$ polynomial of (u, v), so multiplied by g_i, we obtain $u \times v \times$ a polynomial of (u, v).

Consequently, the only terms not included in the second part of (4.9) or (4.10) are the ones composed only of f_i's and/or first derivatives of f_i with respect to u, or composed only of g_i's and/or first derivatives of g_i with respect to v.

Our task is now to find where these terms appear in $f_{i,j}^{(k)}$ and $g_{i,j}^{(k)}$. From (4.12), we already know that

$$f'_{i,j} = \partial_u f_i \, f_j + \partial_v f_i \, g_j,$$

and

$$g'_{i,j} = \partial_u g_i \, f_j + \partial_v g_i \, g_j.$$

To find the higher derivatives, we use (4.11) to obtain

$$f''_i = \partial_{uu} f_i \, \dot{u}^2 + 2\partial_{uv} f_i \, \dot{u}\dot{v} + \partial_{vv} f_i \, \dot{v}^2 + \partial_u f_i \, \ddot{u} + \partial_v f_i \, \ddot{v},$$

and

$$g''_i = \partial_{uu} g_i \, \dot{u}^2 + 2\partial_{uv} g_i \, \dot{u}\dot{v} + \partial_{vv} g_i \, \dot{v}^2 + \partial_u g_i \, \ddot{u} + \partial_v g_i \, \ddot{v},$$

but as we said before the only terms we are interested in are the ones containing only the first derivative with respect to u of f_i, or the ones containing only the first derivative with respect to v of g_i. So we should write

$$f''_i = \partial_u f_i \, \ddot{u} + \text{ other derivatives of } f_i,$$

and

$$g''_i = \partial_v g_i \, \ddot{v} + \text{ other derivatives of } g_i.$$

Similarly, we obtain for the next derivatives

$$f_i^{(k)} = \partial_u f_i \, u^{(k)} + \text{ other derivatives of } f_i,$$

and

$$g_i^{(k)} = \partial_v g_i \, v^{(k)} + \text{ other derivatives of } g_i.$$

Since $u^{(k)}$ and $v^{(k)}$ are given by

$$u^{(k)} = f_1^{(k-1)} + hf_2^{(k-1)} + \ldots + h^n f_{n+1}^{(k-1)},$$

and

$$v^{(k)} = g_1^{(k-1)} + hg_2^{(k-1)} + \ldots + h^n g_{n+1}^{(k-1)},$$

we obtain, using the notation introduced in (4.13),

$$f_{i,j}^{(k+1)} = \partial_u f_i \left[f_{1,j}^{(k)} + f_{2,j-1}^{(k)} + \ldots + f_{j,1}^{(k)} \right] + \text{ functions of other derivatives,} \tag{4.15}$$

and

$$g_{i,j}^{(k+1)} = \partial_v g_i \left[g_{1,j}^{(k)} + g_{2,j-1}^{(k)} + \ldots + g_{j,1}^{(k)} \right] + \text{ functions of other derivatives.}$$

From this last result, we can obtain $f_{i,j}^{(k)}$ and $g_{i,j}^{(k)}$ by induction on k.

Lemma 4.4. *In the notation of* (4.13), *we have*

$$f_{i,j}^{(k+1)} = \frac{1}{i} \, C_j^{(k+1)} \, u(b-v)^{i+j+k} + uv \times \textit{polynomial in } u \textit{ and } v,$$

and

$$g_{i,j}^{(k+1)} = (-1)^{i+j} \, \frac{1}{i} C_j^{(k+1)} \, v(u-a)^{i+j+k} + uv \times \textit{polynomial in } u \textit{ and } v,$$

where $C_j^{(k)}$ *is defined recursively by* $C_j^{(1)} = 1/j$ *and*

$$C_j^{(k+1)} = \sum_{p=1}^{j} \frac{1}{p} C_{j-p+1}^{(k)} \, .$$

Proof. As suggested by the expansion (4.15) of $f_{i,j}^{(k+1)}$, we use an induction argument. We first consider the case $k = 0$ where we have

$$f'_{i,j} = \partial_u f_i \, f_j + \ldots = \frac{1}{i}(b-v)^i \frac{1}{j} u(b-v)^j + uv \ldots = \frac{1}{ij} u(b-v)^{i+j} + uv \ldots,$$

and we define C'_j to be $C'_j := \frac{1}{j}$ so that

$$f'_{i,j} = \frac{1}{i}\, C'_j\, u(b-v)^{i+j} + uv \ldots .$$

Now we consider

$$\begin{aligned} f''_{i,j} &= \partial_u f_i [f'_{1,j} + f'_{2,j-1} + \ldots + f'_{j,1}] + \ldots \\ &= \partial_u f_i \sum_{p=1}^{j} f'_{p,j+1-p} + \ldots \\ &= \frac{1}{i}(b-v)^i \sum_{p=1}^{j} \frac{1}{p}\, C'_{j+1-p}\, u(b-v)^{p+j+1-p} + \ldots \\ &= \frac{1}{i} \sum_{p=1}^{j} \frac{1}{p}\, C'_{j+1-p}\;\; u(b-v)^{i+j+1} + \ldots , \end{aligned}$$

so defining $C''_j := \sum_{p=1}^{j} \frac{1}{p}\, C'_{j+1-p}$, we obtain

$$f''_{i,j} = \frac{1}{i}\, C''_j\, u(b-v)^{i+j+1} + uv \ldots .$$

This example suggests to define $C_j^{(k+1)}$ by

$$C_j^{(k+1)} := \sum_{p=1}^{j} \frac{1}{p} C_{j+1-p}^{(k)} .$$

Then, using the induction hypothesis we get

$$\begin{aligned} f_{i,j}^{(l+1)} &= \partial_u f_i \sum_{p=1}^{j} f_{p,j+1-p}^{(l)} + \ldots \\ &= \frac{1}{i}(b-v)^i \sum_{p=1}^{j} \frac{1}{p} C_{j+1-p}^{(l)}\, u(b-v)^{p+j+1-p+l-1} + uv \ldots \\ &= \frac{1}{i} \sum_{p=1}^{j} \frac{1}{p} C_{j+1-p}^{(l)}\, u(b-v)^{i+j+l} + uv \ldots \\ &= \frac{1}{i}\, C_j^{(l+1)}\;\; u(b-v)^{i+j+l} + uv \ldots , \end{aligned}$$

and the first part of the lemma is proved.

A similar procedure leads us to the formula corresponding to $g_{i,j}^{(k+1)}$. For $k = 0$, we have

$$\begin{aligned} g'_{i,j} &= g_j \partial_v g_i + \dots \\ &= \frac{(-1)^{(i-1)}}{i}(u-a)^i \frac{(-1)^{j+1}}{j} v(u-a)^j + uv\dots \\ &= \frac{(-1)^{i+j}}{ij} v(u-a)^{i+j} + uv\dots \\ &= (-1)^{i+j}\,\frac{1}{i}\,C'_j\,v(u-a)^{i+j}, \end{aligned}$$

and using the induction hypothesis, we obtain

$$\begin{aligned} g_{i,j}^{(l+1)} &= \partial_v g_i \sum_{p=1}^{j} g_{p,j-p}^{(l)} + \dots \\ &= \frac{(-1)^{(i-1)}}{i}(u-a)^i \sum_{p=1}^{j} (-1)^{p+j+1-p}\,\frac{1}{p}\,C_{j+1-p}^{(l)}\,v(u-a)^{p+j+1-p+l-1} + uv\dots \\ &= \frac{(-1)^{i+j}}{i} \sum_{p=1}^{j} \frac{1}{p}\,C_{j+1-p}^{(l)}\,v(u-a)^{i+j+l} + uv\dots \\ &= (-1)^{i+j}\,\frac{1}{i}\,C_j^{(l+1)}\,v(u-a)^{i+j+l} + uv\dots, \end{aligned}$$

which concludes the proof. □

4.3.4 Proof of Theorem 4.4

The first step for proving Theorem 4.4 consists of checking whether or not the induction hypotheses are satisfied for $n = 0$ and $n = 1$. We already know from the previous sections that $f = u(b-v)$, $g = v(u-a)$,

$$f_2 = \frac{1}{2}u(b-v)^2 + uv\frac{1}{2}(u-a) \qquad \text{and} \qquad g_2 = -\frac{1}{2}v(u-a)^2 + uv\frac{1}{2}(b-v),$$

so they satisfy the induction hypothesis (4.9) and (4.10). Then, using the results of Lemma 4.4, we have

$$\begin{aligned} f^{(k)}_{j,n-k-j+2} &= \frac{1}{j} C^{(k)}_{n-k-j+2}\, u(b-v)^{j+n-k-j+2+k-1} + uv\ldots \\ &= \frac{1}{j} C^{(k)}_{n-k-j+2}\, u(b-v)^{n+1} + uv\ldots, \end{aligned}$$

so that

$$\begin{aligned} \sum_{j=1}^{n-k+1} f^{(k)}_{j,n-k-j+2} &= \sum_{j=1}^{n-k+1} \frac{1}{j} C^{(k)}_{n-k-j+2}\, u(b-v)^{n+1} + uv\ldots \\ &= \left(\sum_{j=1}^{n-k+1} \frac{1}{j} C^{(k)}_{n-k-j+2} \right) u(b-v)^{n+1} + uv\ldots \\ &= C^{(k+1)}_{n-k+1}\, u(b-v)^{n+1} + uv\ldots. \end{aligned}$$

Substituting this into the expansion of f_{n+1} given in Lemma 4.3 we obtain

$$\begin{aligned} f_{n+1} &= u(b-v)^{n+1} - \sum_{k=1}^{n} \frac{1}{(k+1)!} \left(C^{(k+1)}_{n-k+1} u(b-v)^{n+1} + uv\ldots \right), \\ &= \left(1 - \sum_{k=1}^{n} \frac{1}{(k+1)!}\, C^{(k+1)}_{n-k+1} \right)\, u(b-v)^{n+1} + uv\ldots. \end{aligned}$$

Similarly, we have for g_{n+1}

$$\begin{aligned} g_{n+1} &= uv(b-v)^n - \sum_{k=1}^{n} \frac{1}{(k+1)!} \left(\sum_{j=1}^{n-k+1} g^{(k)}_{j,n-k-j+2} \right) \\ &= uv(b-v)^n - \sum_{k=1}^{n} \frac{1}{(k+1)!} \left(\sum_{j=1}^{n-k+1} (-1)^{n-k} \frac{1}{j} C^{(k)}_{n-k-j+2} v(u-a)^{n+1} + uv\ldots \right) \\ &= \sum_{k=1}^{n} -\frac{(-1)^{n-k}}{(k+1)!}\, C^{(k+1)}_{n-k+1}\, v(u-a)^{n+1} + uv\ldots. \end{aligned}$$

The parts we denoted by "$+uv\ldots$" in the above results are polynomials, so it only remains to show that

$$1 - \sum_{k=1}^{n} \frac{1}{(k+1)!}\, C^{(k+1)}_{n-k+1} = \frac{1}{n+1}$$

and

$$\sum_{k=1}^{n} -\frac{(-1)^{n-k}}{(k+1)!} C_{n-k+1}^{(k+1)} = \frac{(-1)^n}{n+1}.$$

Lemma 4.5. *If we define*

$$C_j^{(1)} = \frac{1}{j} \qquad \text{and} \qquad C_j^{(k+1)} = \sum_{p=1}^{j} \frac{1}{p} C_{j-p+1}^{(k)}, \qquad j \geq 1,$$

the two following identities hold

$$\sum_{k=1}^{n} \frac{1}{(k+1)!} C_{n-k+1}^{(k+1)} = \frac{n}{n+1}, \qquad \text{and} \qquad \sum_{k=1}^{n} -\frac{(-1)^k}{(k+1)!} C_{n-k+1}^{(k+1)} = \frac{1}{n+1}.$$

Proof. The key is to introduce the generating function (suggested by Ernst Hairer)

$$a^{(k)}(\zeta) = \sum_{j \geq 0} C_{j+1}^{(k)} \zeta^j.$$

For $k = 1$, we obtain

$$a^{(1)}(\zeta) = \sum_{j \geq 0} \frac{1}{j+1} \zeta^j = \frac{1}{\zeta} \sum_{j \geq 0} \frac{1}{j+1} \zeta^{j+1} = \frac{-1}{\zeta} \ln(1-\zeta),$$

and on the other hand we have

$$a^{(1)}(\zeta) a^{(k)}(\zeta) = \sum_{j \geq 0} \sum_{p=0}^{j} C_{p+1}^{(1)} C_{j-p+1}^{(k)} \zeta^j = \sum_{j \geq 0} C_{j+1}^{(k+1)} \zeta^j = a^{(k+1)}(\zeta),$$

therefore

$$a^{(k)}(\zeta) = (-1)^k \left[\frac{\ln(1-\zeta)}{\zeta} \right]^k = \sum_{j \geq 0} C_{j+1}^{(k)} \zeta^j.$$

From this identity we obtain

$$\sum_{k \geq 2} \sum_{j \geq 0} \frac{(-1)^k}{k!} C_{j+1}^{(k)} \zeta^{j+k} = \sum_{k \geq 2} \frac{1}{k!} [\ln(1-\zeta)]^k \tag{4.16}$$

as well as

$$\sum_{k \geq 2} \sum_{j \geq 0} \frac{1}{k!} C_{j+1}^{(k)} \zeta^{j+k} = \sum_{k \geq 2} \frac{1}{k!} [-\ln(1-\zeta)]^k. \tag{4.17}$$

If we define n to be $n = j + k$, the left-hand side of (4.16) can be rewritten as

$$\begin{aligned}\sum_{k\geq 2}\sum_{j\geq 0}\frac{(-1)^k}{k!}C^{(k)}_{j+1}\,\zeta^{j+k} &= \sum_{k\geq 2}\sum_{n\geq k}\frac{(-1)^k}{k!}C^{(k)}_{n-k+1}\,\zeta^{n}\\ &= \sum_{k\geq 1}\sum_{n\geq k}\frac{(-1)^{k+1}}{(k+1)!}C^{(k+1)}_{n-k+1}\,\zeta^{n+1}\\ &= \sum_{n\geq 1}\left[\sum_{k=1}^{n}\frac{(-1)^{k+1}}{(k+1)!}C^{(k+1)}_{n-k+1}\right]\zeta^{n+1}.\end{aligned}$$

Similarly the left-hand side of (4.17) can be rewritten as

$$\sum_{k\geq 2}\sum_{j\geq 0}\frac{1}{k!}C^{(k)}_{j+1}\,\zeta^{j+k} = \sum_{n\geq 1}\left[\sum_{k=1}^{n}\frac{1}{(k+1)!}C^{(k+1)}_{n-k+1}\right]\zeta^{n+1}.$$

Now if we consider the right-hand side of (4.16), we have

$$\begin{aligned}\sum_{k\geq 2}\frac{1}{k!}[\,\ln(1-\zeta)\,]^k &= \exp(\ln(1-\zeta)) - 1 - \ln(1-\zeta)\\ &= 1-\zeta-1-\ln(1-\zeta) = \frac{\zeta^2}{2}+\frac{\zeta^3}{3}+\dots\\ &= \sum_{n\geq 2}\frac{\zeta^n}{n} = \sum_{n\geq 1}\frac{\zeta^{n+1}}{n+1}\end{aligned}$$

and the equation (4.16) becomes

$$\sum_{n\geq 1}\left[\sum_{k=1}^{n}\frac{(-1)^{k+1}}{(k+1)!}C^{(k+1)}_{n-k+1}\right]\zeta^{n+1} = \sum_{n\geq 1}\frac{\zeta^{n+1}}{n+1},$$

so that we have

$$\sum_{k=1}^{n}\frac{(-1)^{k+1}}{(k+1)!}C^{(k+1)}_{n-k+1} = \frac{1}{n+1},$$

which is the first identity we wanted to prove. Finally since the right-hand side of (4.17) can be rewritten as

$$\begin{aligned}\sum_{k\geq 2}\frac{1}{k!}[\,-\ln(1-\zeta)\,]^k &= \frac{1}{1-\zeta} - 1 + \ln(1-\zeta)\\ &= 1+\zeta+\zeta^2+\dots-1-\frac{\zeta^2}{2}-\frac{\zeta^3}{3}-\dots\\ &= \sum_{n\geq 1}\frac{n}{n+1}\zeta^{n+1},\end{aligned}$$

the equation (4.17) becomes

$$\sum_{n\geq 1}\left[\sum_{k=1}^{n}\frac{1}{(k+1)!}C_{n-k+1}^{(k+1)}\right]\zeta^{n+1}=\sum_{n\geq 1}\frac{n}{n+1}\zeta^{n+1}$$

and we obtain the second identity. □

4.3.5 Conclusion

To find the numerical Hamiltonian, we consider the quotient

$$\frac{\dot{u}}{\dot{v}}=\frac{f+hf_2+h^2f_3+\ldots+h^nf_{n+1}}{g+hg_2+h^2g_3+\ldots+h^ng_{n+1}},$$

which gives

$$du\,(g+hg_2+h^2g_3+\ldots+h^ng_{n+1})=dv\,(f+hf_2+h^2f_3+\ldots+h^nf_{n+1}).$$

We are only interested in the part of order h^n since we want to prove (4.7), that is we consider

$$\frac{1}{uv}\Big(f_{n+1}dv-g_{n+1}du\Big)=0,$$

and using the expressions of f_{n+1} and g_{n+1} given by Theorem 4.4, we can write

$$du\bigg(\frac{(-1)^n}{n+1}\frac{1}{u}(u-a)^{n+1}+P_{n+1}(u,v)\bigg)-dv\bigg(\frac{1}{n+1}\frac{1}{v}(b-v)^{n+1}+Q_{n+1}(u,v)\bigg)=0,$$

which becomes when we expand the products

$$du\bigg(\frac{(-1)^n}{n+1}\frac{(-a)^{n+1}}{u}+\tilde{P}_{n+1}(u,v)\bigg)-dv\bigg(\frac{1}{n+1}\frac{b^{n+1}}{v}+\tilde{Q}_{n+1}(u,v)\bigg)=0.$$

To obtain the term of the numerical Hamiltonian, we simply integrate this equality (which is possible by Theorem 4.3), and we get

$$\frac{(-1)^{2n+1}}{n+1}a^{n+1}\ln u-\frac{1}{n+1}b^{n+1}\ln v+\text{polynomial in }(u,v)=\mathit{Const.},$$

and then the new term of the expansion of the numerical Hamiltonian is of the form

$$-\frac{h^n}{n+1}[a^{n+1}\ln u+b^{n+1}\ln v+\text{polynomial of }u\text{ and }v],$$

which is what we wanted to prove.

4.4 The Explicit Variant of the Symplectic Euler Method

4.4.1 First Order Term

We now derive the term of order h of the numerical Hamiltonian corresponding to the explicit variant of the symplectic Euler method. Considering the Lotka-Volterra system (1.1), the variant of the symplectic Euler method is given by (2.3) with $f(u,v) = u(b-v)$ and $g(u,v) = v(u-a)$. To find the first order term of the numerical Hamiltonian, we use the *ansatz*

$$\begin{cases} \dot{u} &= f(u,v) + h\, f_2(u,v), \\ \dot{v} &= g(u,v) + h\, g_2(u,v). \end{cases}$$

Since we have

$$\begin{aligned} u(t_{n+1}) &= u(t_n) + h\dot{u}(t_n) + \frac{h^2}{2}\ddot{u}(t_n) + O(h^3) \\ &= u(t_n) + hf(u(t_n),v(t_n)) + h^2 f_2(u(t_n),v(t_n)) \\ &\quad + \frac{h^2}{2}\left(\frac{\partial f}{\partial u}(u(t_n),v(t_n))\dot{u} + \frac{\partial f}{\partial v}(u(t_n),v(t_n))\dot{v}\right) + O(h^3), \\ &= u(t_n) + hf(u(t_n),v(t_n)) + h^2 f_2(u(t_n),v(t_n)) \\ &\quad + \frac{h^2}{2}\left(\frac{\partial f}{\partial u}(u(t_n),v(t_n))f(u(t_n),v(t_n))\right. \\ &\qquad \left. + \frac{\partial f}{\partial v}(u(t_n),v(t_n))g(u(t_n),v(t_n))\right) + O(h^3), \end{aligned}$$

and the method gives

$$u(t_{n+1}) = u(t_n) + hf(u(t_n),v(t_n)),$$

we must have

$$f_2(u,v) + \frac{1}{2}\left(\frac{\partial f}{\partial u}(u,v)f(u,v) + \frac{\partial f}{\partial v}(u,v)g(u,v)\right) = 0,$$

to obtain an $O(h^3)$ residual, that is

$$\begin{aligned} f_2(u,v) &= -\frac{1}{2}\left(\frac{\partial f}{\partial u}(u,v)f(u,v) + \frac{\partial f}{\partial v}(u,v)g(u,v)\right), \\ &= -\frac{1}{2}[u(b-v)^2 - uv(u-a)]. \end{aligned}$$

Similarly we have for $v(t_{n+1})$

$$\begin{aligned} v(t_{n+1}) = v(t_n) &+ hg(u(t_n), v(t_n)) + h^2 g_2(u(u_n), v(t_n)) \\ &+ \frac{h^2}{2}\left(\frac{\partial g}{\partial u}(u(t_n), v(t_n))f(u(t_n), v(t_n))\right. \\ &\qquad \left. + \frac{\partial g}{\partial v}(u(t_n), v(t_n))g(u(t_n), v(t_n))\right) + O(h^3), \end{aligned}$$

but since for v the method gives

$$\begin{aligned} v(t_{n+1}) &= v(t_n) + hg(u(t_{n+1}), v(t_n)), \\ &= v(t_n) + hg(u(t_n) + hf(u(t_n), v(t_n)) + O(h^2), v(t_n)), \\ &= v(t_n) + hg(u(t_n), v(t_n)) + h^2\frac{\partial g}{\partial u}(u(t_n), v(t_n))f(u(t_n), v(t_n)) + O(h^3), \end{aligned}$$

the condition g_2 must satisfy is

$$g_2(u,v) + \frac{1}{2}\left(\frac{\partial g}{\partial u}(u,v)f + \frac{\partial g}{\partial v}(u,v)g\right) = \frac{\partial g}{\partial u}(u,v)f(u,v),$$

that is

$$\begin{aligned} g_2(u,v) &= \frac{\partial g}{\partial u}f - \frac{1}{2}\left(\frac{\partial g}{\partial u}f + \frac{\partial g}{\partial v}g\right) = \frac{1}{2}\frac{\partial g}{\partial u}f - \frac{1}{2}\frac{\partial g}{\partial u}g \\ &= \frac{1}{2}\left[uv(b-v) - (u-a)^2 v\right]. \end{aligned}$$

Finally, we obtain the numerical Hamiltonian of the Lotka-Volterra system

$$\begin{aligned} H_h = H(u,v) - \frac{h}{2}[\frac{u^2}{2} + uv - \frac{v^2}{2} - (2a+b)u - (a-2b)v \\ + a^2 \ln u - b^2 \ln v] + O(h^2). \end{aligned}$$

We summarize these results in the following lemma.

Lemma 4.6. *If the explicit variant of the symplectic Euler method is applied to the Lotka-Volterra system, the modified system is*

$$\begin{cases} \dot{\tilde{u}} &= f(\tilde{u},\tilde{v}) + hf_2(\tilde{u},\tilde{v}) + O(h^2), \\ \dot{\tilde{v}} &= g(\tilde{u},\tilde{v}) + hg_2(\tilde{u},\tilde{v}) + O(h^2), \end{cases}$$

with

$$f_2(\tilde{u},\tilde{v}) = -\frac{1}{2}[\tilde{u}(b-\tilde{v})^2 - \tilde{u}\tilde{v}(\tilde{u}-a)],$$

and

$$g_2(\tilde{u},\tilde{v}) = \frac{1}{2}\,[\tilde{u}\tilde{v}(b-\tilde{v}) - (\tilde{u}-a)^2\tilde{v}].$$

Furthermore an invariant of this modified system is

$$H_h(\tilde{u},\tilde{v}) = H(\tilde{u},\tilde{v}) - \frac{h}{2}[\frac{\tilde{u}^2}{2} + \tilde{u}\tilde{v} - \frac{\tilde{v}^2}{2} + (-2a-b)\tilde{u} + (-a+2b)\tilde{v} \\ + a^2\ln\tilde{u} - b^2\ln\tilde{v}] + O(h^2).$$

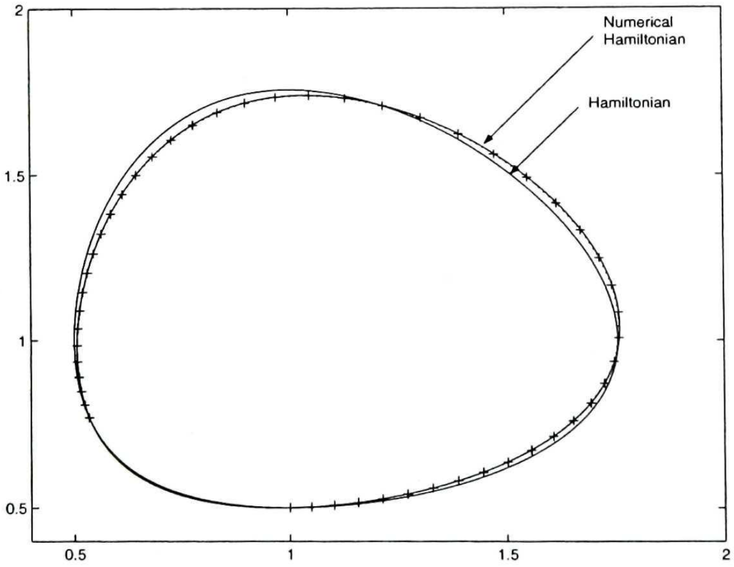

Figure 4.2: Illustration of the numerical Hamiltonian of order one of the explicit variant of the symplectic Euler method applied to the Lotka-Volterra system with $h = 0.1$.

4.5 Structure of the Numerical Hamiltonian

As one can see from the expansion of the numerical Hamiltonian we just derived, it is very similar to the one we derived for the symplectic Euler method. We can prove, following the same steps as in Section 4.3, that each term of the expansion is of the form

$$-\frac{h^n}{n+1}\,[\,a^{n+1}\ln u+(-1)^n\,b^{n+1}\ln v+\text{polynomial of } u \text{ and } v\,].$$

The theorem corresponding to Theorem 4.4 is the following.

Theorem 4.5. *The coefficients f_{n+1} and g_{n+1} are of the form*

$$f_{n+1}=\frac{(-1)^n}{n+1}u(b-v)^{n+1}+uv\times P_{n+1}(u,v)$$

and

$$g_{n+1}=\frac{(-1)^n}{n+1}v(u-a)^{n+1}+uv\times Q_{n+1}(u,v),$$

where P_{n+1} and Q_{n+1} are polynomials, when we use the ansatz

$$\begin{cases}\dot{u} &= f+hf_2+h^2f_3+\ldots+h^nf_{n+1},\\ \dot{v} &= g+fg_2+h^2g_3+\ldots+h^ng_{n+1}.\end{cases}$$

The expansion of $u(t_{n+1})$ given in the proof of Lemma 4.3 is still valid, however, since the method gives

$$u_{n+1}=u_n+hf(u_n,v_n),$$

with no term of order h^n, $n>1$, the term of order h^n in the expansion of $u(t_{n+1})$ vanishes, and we obtain

$$\begin{aligned}f_{n+1}&=-\frac{1}{2!}(f'_{1,n}+\ldots+f'_{n,1})-\ldots-\frac{1}{n!}(f^{(n-1)}_{1,2}+f^{(n-1)}_{2,1})-\frac{1}{(n+1)!}f^{(n)}_{1,1}\\ &=-\sum_{k=1}^{n}\frac{1}{(k+1)!}\left(\sum_{j=1}^{n-k+1}f^{(k)}_{j,n-k-j+2}\right),\qquad \text{for } n>1.\end{aligned}$$

For v_{n+1}, the method gives

$$v_{n+1}=v_n+hg(u_{n+1},v_n)=v_n+hg(u_n+hf(u_n,v_n),v_n),$$

which gives with a Taylor expansion

$$v_{n+1} = v_n + h\Big(g + h\,\partial_u g\,\dot{u} + \frac{h^2}{2}(\partial_{uu} g\,\dot{u}^2 + \partial_u g\,\ddot{u}) + \\ \frac{h^3}{3!}(\partial_{uuu} g\,\dot{u}^3 + 3\,\partial_{uu} g\,\dot{u}\,\ddot{u} + \partial_u g\,\dddot{u}) + \ldots\Big),$$

but since we know that $\partial_{uu} g$ and the higher derivatives of g are zero, this becomes

$$v_{n+1} = v_n + h\Big(g + h\partial_u g\dot{u} + \frac{h^2}{2}\partial_u g\ddot{u} + \frac{h^3}{3!}\partial_u g\,\dddot{u} + \ldots + \frac{h^n}{n!}\partial_u g\,u^{(n)}\Big) + O(h^{n+2}).$$

We can rewrite this expansion as

$$v_{n+1} = v_n + hg + h\partial_u g\Big(h\dot{u} + \frac{h^2}{2}\ddot{u} + \frac{h^3}{3!}\dddot{u} + \ldots + \frac{h^n}{n!}u^{(n)}\Big) + O(h^{n+2}),$$

which is very similar to the expansion of $u(t_{n+1})$ derived in the proof of Lemma 4.3, and it becomes

$$\begin{aligned} v_{n+1} = v_n &+ hg + h^2\partial_u g f_1 + h^3\partial_u g\Big(f_2 + \frac{1}{2!}f'_{1,1}\Big) \\ &+ h^4\partial_u g\Big(f_3 + \frac{1}{2!}(f'_{1,2} + f'_{2,1}) + \frac{1}{3!}f''_{1,1}\Big) \\ &\vdots \\ &+ h^{n+1}\partial_u g\Big(f_n + \frac{1}{2!}(f'_{1,n-1} + \ldots + f'_{n-1,1}) + \frac{1}{3!}(f''_{1,n-2} + \ldots \\ &\quad \ldots + f''_{n-2,1}) + \ldots + \frac{1}{(n-1)!}(f^{(n-2)}_{1,2} + f^{(n-2)}_{2,1}) + \frac{1}{n!}f^{(n-1)}_{1,1}\Big) \\ &+ O(h^{n+2}). \end{aligned}$$

By the induction hypothesis, we have $f_{m+1} = -\sum_{k=1}^{m}\frac{1}{(k+1)!}\sum_{j=1}^{m-k+1} f^{(k)}_{j,m-k-j+2}$ for $m = 1, \ldots, n-1$, so we obtain that all the terms of order h^k, for $k = 3$ to $n+1$ will vanish, so that the method becomes

$$v_{n+1} = v_n + hg + h^2\partial_u g f_1 + O(h^{n+2}).$$

Therefore, we obtain, for n greater that 2, the function g_{n+1}, essentially the same as f_{n+1},

$$\begin{aligned} g_{n+1} &= -\frac{1}{2!}(g'_{1,n} + \ldots + g'_{n,1}) - \ldots - \frac{1}{n!}(g^{(n-1)}_{1,2} + g^{(n-1)}_{2,1}) - \frac{1}{(n+1)!}g^{(n)}_{1,1}, \\ &= -\sum_{k=1}^{n} \frac{1}{(k+1)!}\Bigg(\sum_{j=1}^{n-k+1} g^{(k)}_{j,n-k-j+2} \Bigg). \end{aligned}$$

Then, considering the remarks made in Section 4.3.3, we easily obtain the following lemma.

Lemma 4.7. *In the notation of* (4.13), *we have*

$$f^{(k+1)}_{i,j} = (-1)^{i+j}\,\frac{1}{i}C^{(k+1)}_j\, u(b-v)^{i+j+1} + uv \times \textit{polynomial in } u \textit{ and } v,$$

and

$$g^{(k+1)}_{i,j} = (-1)^{i+j+1}\,\frac{1}{i}C^{(k+1)}_j\, v(u-a)^{i+j+1} + uv \times \textit{polynomial in } u \textit{ and } v,$$

where $C^{(k)}_j$ is defined in Lemma 4.4.

Now we can prove Theorem 4.5: one can easily check that the induction hypotheses are satisfied for $n = 0$ and $n = 1$, then Lemma 4.5 and Lemma 4.7 conclude the proof. To find the numerical Hamiltonian, we consider the quotient $\dot{u}/\dot{v}$, which yields, as in Section 4.3.5,

$$\frac{1}{uv}\Bigg(f_{n+1}dv - g_{n+1}du \Bigg) = 0,$$

and using the expressions of f_{n+1} and g_{n+1} we just derived, we can write

$$du\Bigg(\frac{(-1)^n}{n+1}\frac{1}{u}(u-a)^{n+1} + P_{n+1}(u,v) \Bigg) - dv\Bigg(\frac{(-1)^n}{n+1}\frac{1}{v}(b-v)^{n+1} + Q_{n+1}(u,v) \Bigg) = 0,$$

which gives, if we expand the products,

$$du\Bigg(\frac{(-1)^n}{n+1}\frac{(-a)^{n+1}}{u} + \tilde{P}_{n+1}(u,v) \Bigg) - dv\Bigg(\frac{(-1)^n}{n+1}\frac{b^{n+1}}{v} + \tilde{Q}_{n+1}(u,v) \Bigg) = 0.$$

To obtain the numerical Hamiltonian, we simply integrate this equality which gives

$$\frac{(-1)^{2n+1}}{n+1}a^{n+1}\ln u - \frac{(-1)^n}{n+1}b^{n+1}\ln v + \text{polynomial in } (u,v) = Const.,$$

and then the term of order h^n of the expansion of the numerical Hamiltonian is of the form

$$-\frac{h^n}{n+1}[a^{n+1}\ln u + (-1)^n\, b^{n+1}\ln v + \text{polynomial of } u \text{ and } v],$$

which is what we wanted to prove.

4.6 Concluding Remarks

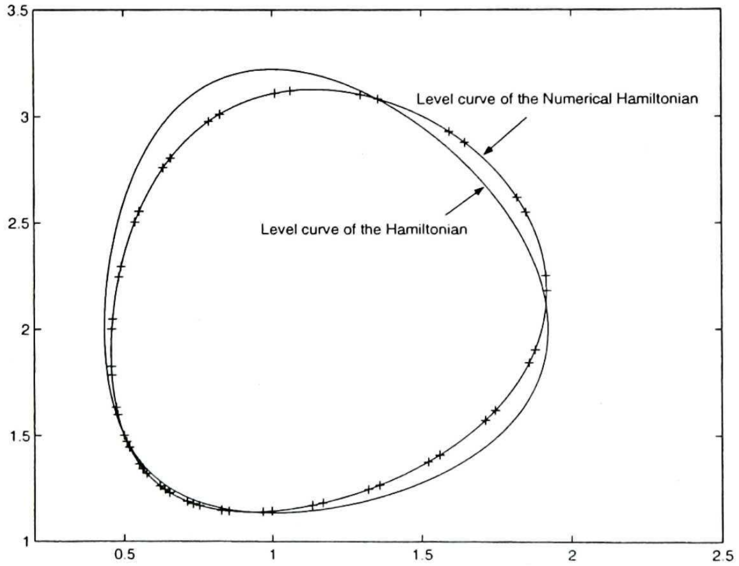

Figure 4.3: Illustration of the numerical Hamiltonian of order 5 of the explicit variant of the symplecitc Euler method (h=0.2).

By definition of the numerical Hamiltonian, we know that the numerical trajectory obtained applying the symplectic Euler method and its explicit variant to a Lotka-Volterra system should stay on a level curve of their respective numerical Hamiltonian.

Of course, since we only derived a truncated numerical Hamiltonian and not the exact one (which may not exist), we cannot expect the numerical trajectories obtained to stay *exactly* on a level curve of the numerical Hamiltonians derived in Sections 4.4.1 and 4.2.2. Nevertheless we know that it should be close to it up to $O(h^n)$. As one can see on Figure 4.3, we indeed have a great improvement. On Figure 4.4, we plotted the numerical Hamiltonian error of the explicit variant of the symplectic Euler method; we considered different approximations of the Hamiltonian to clearly show the improvement at each step.

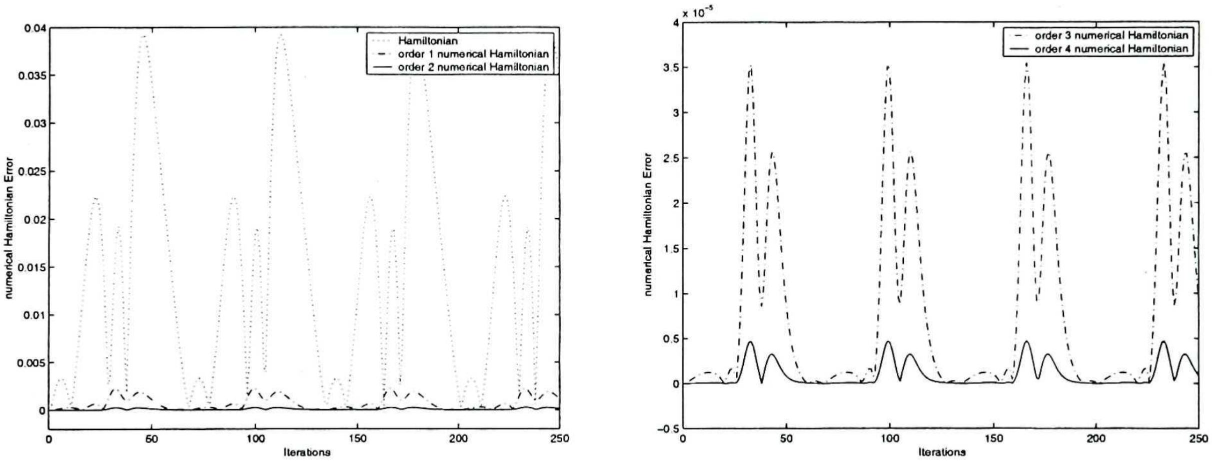

Figure 4.4: Error of the numerical Hamiltonian. On the left figure, the Hamiltonians used are the original Hamiltonian and the numerical Hamiltonian of order 1 and 2. On the right figure, we used the numerical Hamiltonian of order 3 and 4.

The structure of the numerical Hamiltonians derived in Sections 4.3 and 4.5 suggests that the series of the modified differential equation do not converge for some values of h. We plotted in Figures 4.5 and 4.6, the level curves of the numerical Hamiltonians of order one to four for the two methods. Even if one cannot draw conclusions based on the observation of the first few terms of a series, these figures suggest that the series oscillate.

Another important observation is that the modified systems obtained in Section

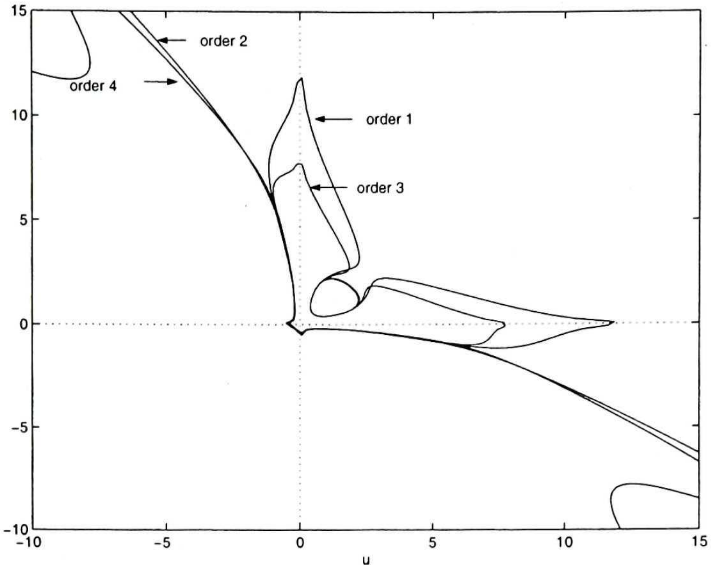

Figure 4.5: Level curves of the numerical Hamiltonian of order 1 to 4 of the symplecitc Euler method for $u_0 = 1.5$, $v_0 = 1.5$ and $h = 0.5$.

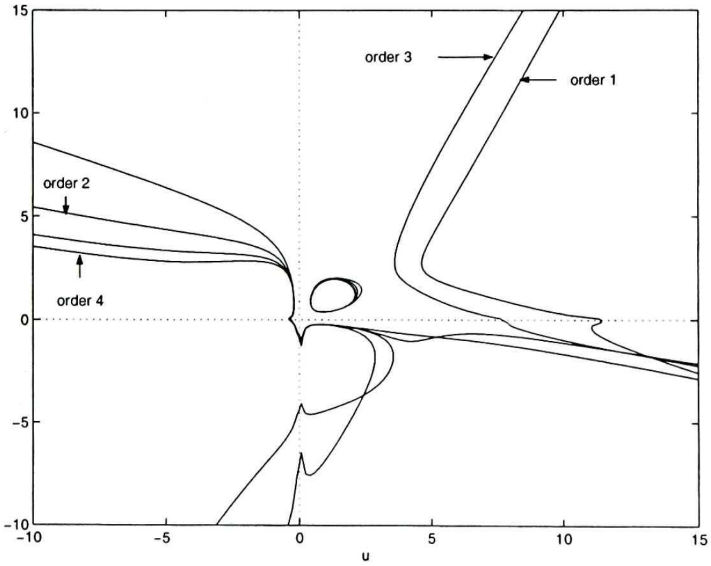

Figure 4.6: Level curves of the numerical Hamiltonian of order 1 to 4 of the explicit variant of the symplectic Euler method for $u_0 = 1.5$, $v_0 = 1.5$ and $h = 0.5$.

4.4.1 and 4.2.2 are Poisson systems as we expected from Theorem 4.3.

Chapter 5

Exploitation of the Numerical Hamiltonian

The aim of this chapter is to study the positivity of the numerical results given by the explicit variant of the symplectic Euler method (EVSE) using backward error analysis.

We mostly follow Sections IX.7 and IX.8 of [5] where the study focuses on Hamiltonian systems. Some modifications are necessary to take into account the fact that the Lotka-Volterra system is not a Hamiltonian but a Poisson system. We also transform some general results into ones more specific to EVSE in order to achieve better bounds.

The procedure is to bound the local error of the numerical result and to use this result and properties of Hamiltonian systems to bound the Hamiltonian error $|H(y_n) - H(y_0)|$. 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 EVSE is ensured to remain positive over exponentially long time intervals. We also give the procedure to find these step-sizes and illustrate it by a numerical example.

5.1 Estimation of the Numerical Solution

We consider the Lotka-Volterra system

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

which is a Poisson system whose Hamiltonian

$$H(u,v) = u - a\ln u + v - b\ln v$$

is analytic on $E \times E$, where $E := \mathbb{C}\backslash \{z \,|\, \mathrm{Re}(z) \leq 0 \text{ and } \mathrm{Im}(z) = 0\}$. We shall denote for the rest of the chapter $y := (u,v)$ and $f(y) = (u(b-v), v(u-a))^T$. We apply to the system the explicit variant of the symplectic Euler method $\Phi_h(y)$, defined by (2.3), 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

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

where $0 < \alpha < 1$ can be arbitrarily chosen (we usually obtain better results if α is large), so that for all $y_0 \in K$ and all y such that $\|y - y_0\| \leq R$, y belongs to $\mathbb{D}$. Denoting by $\tilde{K}$ the compact set $\tilde{K} := \{\, y \,|\, \exists\, y_0 \in K \text{ such that } \|y - y_0\| \leq R\}$ and by M the bound $M := \max\{f(u,v) \,:\, (u,v) \in \tilde{K}\}$, we now have, for all $y_0 \in K$,

$$\|f(y)\| \leq M \qquad \text{for} \qquad \|y - y_0\| \leq R. \tag{5.1}$$

As we saw in Section 4.4.1, we can write the explicit variant of the symplectic Euler method applied to the Lotka-Volterra system as

$$\Phi_h(y) = y + hf(y) + h^2 d_2(y), \tag{5.2}$$

where $d_2(y) = (0, uv(b-v))^T$ is analytic. Since we have to bound $d_2(y)$ for $y \in \tilde{K}$, we simply define $M_2 := \max\{uv(b-v) \,:\, (u,v) \in \tilde{K}\}$ so that for all $y_0 \in K$,

$$\|d_2(y)\| = |uv(b-v)| \leq M_2 \qquad \text{for} \qquad \|y - y_0\| \leq R. \tag{5.3}$$

5.2 Estimation of the Coefficients of the Modified Equation

Our task is now to bound the functions f_j of the modified equation

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + h^2 f_3(\tilde{y}) + \dots. \tag{5.4}$$

In order to simplify formulas, we may often denote f by f_1. The key idea to obtain an explicit formula for these functions, is to introduce the Lie derivative

$$D_j = \sum_k f_j^{[k]}(y)\frac{\partial}{\partial y^{[k]}}$$

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

$$D_j g(y) = g'(y) f_j(y).$$

Using Lie derivatives and denoting $y := \tilde{y}(t)$, we can write the solution of the modified equation (5.4) expanded into a Taylor series as

$$\begin{aligned}
\tilde{y}(t+h) &= y + h(f(y) + hf_2(y) + h^2 f_3(y) + \dots) \\
&\quad + \frac{h^2}{2!}(f'(y) + hf_2'(y) + \dots)(f(y) + hf_2(y) + \dots) + \dots \\
&= y + hf(y) + h^2[f_2(y) + \frac{1}{2!}f'f(y)] \\
&\quad + h^3[f_3(y) + \frac{1}{2!}(f_2'f(y) + f'f_2(y)) + \frac{1}{3!}(f'f'f(y) + f"ff(y))] + \dots \\
&= y + hf(y) + h^2[f_2(y) + \frac{1}{2!}(D_1 f_1)(y)] \\
&\quad + h^3[f_3(y) + \frac{1}{2!}(D_1 f_2 + D_2 f_1)(y) + \frac{1}{3!}(D_1^2 f_1)(y)] + \dots.
\end{aligned}$$

In other words, the solution of the modified equation (5.4), with initial value $y(t) = y$ can be formally written as

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

where $F(y) = f_1(y) + hf_2(y) + h^2 f_3(y) + \ldots$ stands for the modified equation, and $hD = hD_1 + h^2 D_2 + h^3 D_3 + \ldots$ for the corresponding Lie operator.

Expanding the formal sum, we obtain

$$\tilde{y}(t+h) = y + \sum_{i\geq 1} \frac{1}{i!} \left[\sum_{k_1,\ldots,k_i} h^{k_1+\cdots+k_i} (D_{k_1} \ldots D_{k_{i-1}} f_{k_i})(y) \right], \tag{5.5}$$

where all $k_m \geq 1$, and then we can compare like powers of h in the numerical method (5.2) and the expansion of the exact solution (5.5) to obtain for $j \geq 3$,

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

and for $j = 2$

$$\sum_{i\geq 1} \frac{1}{i!} \left[\sum_{k_1+\cdots+k_i=2} (D_{k_1} \ldots D_{k_{i-1}} f_{k_i})(y) \right] = d_2(y).$$

In other words,

$$f_2(y) = d_2(y) - \frac{1}{2}(D_1 f_1)(y) \tag{5.6}$$

and

$$f_j(y) = -\sum_{i=2}^{j} \frac{1}{i!} \left[\sum_{k_1+\cdots+k_i=j} (D_{k_1} \ldots D_{k_{i-1}} f_{k_i})(y) \right], \tag{5.7}$$

for $j \geq 3$, so if we want to get bounds for $\|f_j(y)\|$, we have to estimate first $\|(D_j g)(y)\|$ and for this we use the following variant of Cauchy's estimate given in [5].

Lemma 5.1. *For analytic functions $f_j(y)$ and $g(y)$ we have for $0 \leq \sigma < \rho$ the estimate*

$$\|D_i g\|_\sigma \leq \frac{1}{\rho - \sigma} \cdot \|f_i\|_\sigma \cdot \|g\|_\rho.$$

Here, $\|g\|_\rho := \max\{\, \|g(y)\| \;:\; y \in B_\rho(y_0)\}$ and $\|f_i\|_\sigma, \|D_i g\|_\sigma$ are defined similarly.

Proof. Fix $y \in B_\sigma(y_0)$ and consider $\alpha(z) := g(y + z f_i(y))$. This function is analytic for $|z| \leq \varepsilon := \frac{\rho-\sigma}{M}$ with $M := \|f_i\|_\sigma$. We then have

$$\alpha'(0) = g'(y) f_i(y) = (D_i g)(y),$$

so applying Cauchy's estimate to α we obtain

$$\|(D_i g)(y)\| = \|\alpha'(0)\| \leq \frac{1}{\varepsilon} \sup_{|z|\leq\varepsilon} \|\alpha(z)\| = \frac{M}{\rho-\sigma}\|g\|_\rho$$

and finally

$$\|D_i g\| \leq \frac{1}{\rho-\sigma}\|f_i\|_\sigma \|g\|_\rho.$$

□

The following theorem gives an explicit bound for the functions $f_j(y)$, for $y \in \tilde{K}_2 := \{\, y \mid \exists y_0 \in K \text{ s.t. } \|y - y_0\| \leq R/2 \,\}$. Note that this bound is only valid when we apply the explicit variant of the symplectic Euler method to the Lotka-Volterra system.

Theorem 5.1. *For all $y_0 \in K$, $f(y)$ and the coefficients $f_j(y)$ of the modified differential equation (5.4) are analytic in $B_R(y_0)$, so if the bounds (5.1) and (5.3) are satisfied, we have for the coefficients f_j, $j \geq 2$,*

$$\|f_j(y)\| \leq \ln 2\, \frac{\eta M}{2} \left(\frac{\eta M (j-1)}{R} \right)^{j-1} \qquad \textit{for} \qquad y \in \tilde{K}_2, \tag{5.8}$$

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

Proof. We fix an index $J > 1$ and we want to bound $\|f_J\|_{R/2} = \max\{\, \|f_J(y)\| : y \in B_{R/2}(y_0)\}$. The trick of the proof is to introduce $\delta := R/(2(J-1))$ and estimate $\|f_j\|_{R-(j-1)\delta}$, so that for $j = J$, we obtain $\|f_j\|_{R-(j-1)\delta} = \|f_J\|_{R/2}$.

In order to simplify notations, we abbreviate $\|\cdot\|_{R-(j-1)\delta}$ by $\|\cdot\|_j$. Applying repeatedly Cauchy's estimate given in Lemma 5.1, we obtain for $k_1 + \cdots + k_i = j$,

$$\begin{aligned}
\|D_{k_1} D_{k_2} \ldots D_{k_{i-1}} f_{k_i}\|_j &\leq \frac{1}{\delta}\|f_{k_1}\|_j \cdot \|D_{k_2} \ldots D_{k_{i-1}} f_{k_i}\|_{j-1} \\
&\leq \frac{1}{\delta^2}\|f_{k_1}\|_j \cdot \|f_{k_2}\|_{j-1} \cdot \|D_{k_3} \ldots D_{k_{i-1}} f_{k_i}\|_{j-2} \\
&\leq \ldots \\
&\leq \frac{1}{\delta^{i-1}}\|f_{k_1}\|_j \cdot \|f_{k_2}\|_{j-1} \cdot \ldots \cdot \|f_{k_{i-1}}\|_{j-i+2} \cdot \|f_{k_i}\|_{j-i+1}.
\end{aligned}$$

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

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

We now apply this inequality to the expansions of the functions f_j given by (5.6) and (5.7) and obtain

$$||f_2||_2 \;\leq\; ||d_2||_2 + \frac{1}{2}||D_1 f_1||_2 \;\leq\; ||d_2||_2 + \frac{1}{2\delta}||f_1||_1^2$$

and

$$\begin{aligned} ||f_j(y)||_j &\leq \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\cdots+k_i=j} ||D_{k_1} \ldots D_{k_{i-1}} f_{k_i}||_j \\ &\leq \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1+\cdots+k_i=j} \frac{1}{\delta^{i-1}} ||f_{k_1}||_{k_1} \cdot ||f_{k_2}||_{k_2} \cdot \ldots \cdot ||f_{k_i}||_{k_i}. \end{aligned}$$

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 $||f_j||_j \leq \beta_j \delta$, 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 $y := b(\zeta)$, we have to solve

$$\Phi(\zeta, y) := e^y - 2y - 1 + \frac{\gamma\zeta}{1 - q\zeta} = 0 \tag{5.9}$$

and since $\partial\Phi/\partial y = e^y - 2$, we can apply the implicit function theorem whenever $e^y = e^{b(\zeta)} \neq 2$. We have $e^y = 2$ for $y = B := \ln 2 + 2k\pi i$, 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 (5.9) with $b(0) = 0$ is bounded by $\ln 2$. We consider first the map defined by $w = \frac{-\gamma\zeta}{1-q\zeta} = \frac{-\gamma}{1/\zeta - q}$ and decompose it. The image of the disc $|\zeta| \leq \frac{1}{\nu}$ under the mapping $\zeta \mapsto \frac{1}{\zeta}$ is the disc $|\zeta| \leq \nu$, then we apply the translation $\zeta \mapsto \zeta - q$ and we obtain the disc of radius ν centered at $-q$. Applying again the transformation $\zeta \mapsto \frac{1}{\zeta}$ 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 centre 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, [5] 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 $\|f_J\|_{R/2} = \|f_J\|_J \leq \delta\beta_J \leq \ln 2 \, \delta \, \nu^J$. By definition of ν, we have

$$\begin{aligned} \nu = q + \frac{\gamma}{2\ln 2 - 1} &= \frac{M_2}{M} + \frac{M}{\delta(2\ln 2 - 1)} = \frac{M_2}{M} + \frac{2M(J-1)}{R(2\ln 2 - 1)} \\ &= \frac{M(J-1)}{R}\left(\frac{2}{2\ln 2 - 1} + \frac{RM_2}{M^2(J-1)}\right) \leq \frac{M(J-1)}{R}\left(\frac{2}{2\ln 2 - 1} + \frac{RM_2}{M^2}\right), \end{aligned}$$

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

$$\nu \leq \frac{M(J-1)\eta}{R}.$$

On the other hand, we have

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

so that finally, for $J > 1$, we obtain

$$\|f_J\|_{R/2} \leq \ln 2 \, \frac{M\eta}{2} \left(\frac{M(J-1)\eta}{R}\right)^{J-1}.$$

□

5.3 Estimation of the Local Error

As we said in Chapter 4, the modified differential equation series usually diverges, so that we have to work with a truncated equation

$$\dot{\tilde{y}} = F_N(\tilde{y}), \qquad F_N(\tilde{y}) = f(\tilde{y}) + hf_2(\tilde{y}) + \cdots + h^{N-1} f_N(\tilde{y}) \tag{5.10}$$

with initial value $\tilde{y}_0 = y_0$. Supposing that $hN \leq h_0$ with $h_0 := \frac{R}{e\eta M}$ and using the bound (5.8), we estimate for $y \in \tilde{K}_2$,

$$\begin{aligned}
\|F_N(y)\| &= \|f(y) + hf_2(y) + \cdots + h^{N-1}f_N(y)\| \\
&\leq \|f\| + h\|f_2\| + \cdots + h^{N-1}\|f_N\| \\
&\leq M + \sum_{j=2}^{N} \ln 2\frac{\eta M}{2}\left(\frac{\eta M(j-1)h}{R}\right)^{j-1} \\
&\leq M\left[1 + \eta\frac{\ln 2}{2}\sum_{j=1}^{N-1}\left(\frac{\eta M j h_0}{RN}\right)^j\right] \\
&\leq M\left[1 + \eta\frac{\ln 2}{2}\sum_{j=1}^{N-1}\left(\frac{j}{eN}\right)^j\right],
\end{aligned}$$

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

$$\|F_N(y)\| \leq M\left[1 + 1.0022\,\eta\frac{\ln 2}{2}\right] \leq M\,[1 + 0.064\eta]\,. \tag{5.11}$$

This estimation allows us to bound the local error.

Theorem 5.2. *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 $y_1 = \Phi_h(y_0)$ and the exact solution $\tilde{\phi}_{N,t}(y_0)$ of the truncated modified equation (5.10) satisfies*

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

where $\gamma = e(2 + \frac{eh_0 M_2}{3M} + 0.064\eta)$.

Proof. For any $y_0 \in K$ fixed, we consider the function

$$g(h) := \Phi_h(y_0) - \tilde{\phi}_{N,h}(y_0),$$

which is analytic, since $\Phi_h(y_0) = y_0 + hf(y_0) + h^2 d_2(y_0)$ and $\tilde{\phi}_{N,h}$ in Section 4.5 are both analytic functions of h.

By definition of the functions $f_j(y)$ of the modified equation, the coefficients of the Taylor series for $\Phi_h(y_0)$ 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 $g(h)$ contains the factor h^{N+1} and we can apply the maximum principle for analytic functions to $\frac{g(h)}{h^{N+1}}$. If $g(z)$ is analytic for $|z| \le \varepsilon$, we have for $0 \le h \le \varepsilon$,

$$\left\| \frac{g(h)}{h^{N+1}} \right\| \le \frac{1}{\varepsilon^{N+1}} \max_{|z|\le\varepsilon} \|g(z)\|. \tag{5.12}$$

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

On the other hand we have

$$\|\Phi_z(y_0) - y_0\| = \|zf(y_0) + z^2 d_2(y_0)\| \le |z|M + |z|^2 M_2.$$

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

$$\|\tilde{\phi}_{N,z}(y_0) - y_0\| = \|\tilde{y}(z) - \tilde{y}(0)\| \le |z| \cdot \|\dot{\tilde{y}}(z)\| = |z| \cdot \|F_N(\tilde{y})\| \le |z|M(1 + 0.064\eta),$$

as long as $\tilde{\phi}_{N,z}(y_0) = \tilde{y}(z)$ stays in $\tilde{K}_2$. In fact, because

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

since $\eta \ge 5$ and $N \ge 3$, the solution $\tilde{\phi}_{N,z}$ stays in the ball $B_{R/2}(y_0) \subset \tilde{K}_2$ for all $|z| \le \varepsilon$.

Finally we go back to (5.12). Since

$$\begin{aligned} \|g(z)\| &\le \|\Phi_z(y_0) - y_0\| + \|\tilde{\phi}_{N,z}(y_0) - y_0\| \\ &\le |z|M + |z|^2 M_2 + M|z|(1 + 0.064\eta) \\ &\le |z|M\,(1 + |z|\frac{M_2}{M} + 1 + 0.064\eta), \end{aligned}$$

we have

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

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

$$\left(\frac{hN}{eh_0}\right)^{N} \leq e^{-N}$$

and

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

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

□

5.4 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.

Theorem 5.3. *If the numerical solution stays in the compact set $\tilde{K}_2 \subset \mathbb{D} = \{u > 0, v > 0\}$ and if $h \leq h_0/3$, with $h_0 = R/(e\eta M)$, then there exist $N = N(h)$ (the largest integer satisfying $hN \leq h_0$) such that, over exponentially long time intervals $nh \leq e^{h_0/2h}$,*

$$\begin{aligned}|\tilde{H}(y_n) - \tilde{H}(y_0)| &\leq L\gamma M e^{-h_0/2h} \\ |H(y_n) - H(y_0)| &\leq L\,\gamma M e^{-h_0/2h} + 2hC\end{aligned} \qquad (5.13)$$

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 $\tilde{\phi}_{N,t}(y_0)$ be the flow of the truncated modified equation (5.4). As stated in Theorem 4.3, 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

$$\tilde{H}(\tilde{\phi}_{N,t}(y_0)) = \tilde{H}(y_0), \qquad \forall\, t.$$

Our first goal is to bound $\nabla\tilde{H}$. By definition of Poisson systems, we have

$$F_N(y) = f(y) + hf_2(y) + \cdots + h^{N-1}f_N(y) = B(y)\nabla\tilde{H}(y),$$

and using the bound on $||F_N(y)||$ derived in (5.11), we have

$$||B(y)\nabla\tilde{H}(y)|| \leq \; M(1+0.064\eta).$$

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

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

so that

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

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

$$||\tilde{H}(y_{n+1}) - \tilde{H}(\tilde{\phi}_{N,h}(y_n))|| \leq L\; ||y_{n+1} - \tilde{\phi}_{N,h}(y_n)|| \leq L\; h\gamma M e^{-h_0/h},$$

by Theorem 5.2.

We are now in a position to bound

$$|\tilde{H}(y_n) - \tilde{H}(y_0)| = \left|\sum_{j=1}^{n}\left[\tilde{H}(y_j) - \tilde{H}(y_{j-1})\right]\right| = \left|\sum_{j=1}^{n}\left[\tilde{H}(y_j) - \tilde{H}(\tilde{\phi}_{N,h}(y_{j-1}))\right]\right|$$

$$\leq \sum_{j=1}^{n} |L\; h\gamma M e^{-h_0/h}| = nhL\;\gamma M e^{-h_0/h},$$

so that for $nh < e^{h_0/2h}$, we have

$$|\tilde{H}(y_n) - \tilde{H}(y_0)| \leq L\gamma M e^{-h_0/2h},$$

which is the first inequality we wanted to prove.

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

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

we have to prove that $H_2(y) + hH_3(y) + \cdots + h^{N-2}H_N(y)$ is uniformly bounded on K independently of h and N. By Theorem 4.3, we have for all j

$$f_j(y) = B(y)g_j(y) = B(y)\nabla H_j(y),$$

so that, using the bound (5.8), we have for $j \geq 2$ and $y \in \tilde{K}_2$

$$||f_j(y)|| = |uv| \cdot ||\nabla H_j(y)|| = |uv| \cdot ||g_j(y)|| \leq \ln 2 \frac{\eta M}{2} \left(\frac{\eta M(j-1)}{R}\right)^{j-1}.$$

On the other hand, the proof of the Integrability Lemma 4.1 gives

$$H_j(y) = \int_0^1 (y - z_0)^T g_j(z_0 + t(y - z_0))dt$$

for any $z_0 \in K$. So we can choose z_0 such that $||y - z_0|| \leq R/2$ and then

$$\begin{aligned} ||H_j(y)|| &= ||\int_0^1 (y - z_0)^T g_j(z_0 + t(y - z_0))dt\, || \\ &\leq \int_0^1 ||(y - z_0)|| \cdot ||g_j(z_0 + t(y - z_0))||dt \\ &\leq \frac{R}{2}\frac{1}{|uv|}\frac{\ln 2M\eta}{2}\left(\frac{\eta M(j-1)}{R}\right)^{j-1} \\ &\leq \frac{R \ln 2M\eta}{4(u_{\min} - R/2)(v_{\min} - R/2)}\left(\frac{\eta M(j-1)}{R}\right)^{j-1}. \end{aligned}$$

and then

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

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

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

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

$$
|H(y_n) - H(y_0)| \leq L\,\gamma M e^{-h_0/2h} + 2hC.
$$

□

5.5 Application

We can apply Theorem 5.3 to our problem, namely "how to be sure that the method 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 5.4. *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

$$H_0 := H(u_0, v_0) = u_0 - a\ln u_0 + v_0 - b\ln v_0.$$

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, $\alpha = 9/10$, so that we can define the compact set $\tilde{K} = \{\, y \,:\, \exists\, y_0 \in K$ such that $\|y - y_0\| \leq R\}$ and in a similar way, $\tilde{K}_2$.

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

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

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

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

instead of $\max_{y \in \tilde{K}} \|d_2(y)\|$.

Once we have these values, we can compute

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

as well as $\gamma = e[2 + \frac{eh_0 M_2}{3M} + 0.064\eta]$,

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

and

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

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 (5.13) is valid if and only if y_n is in $\tilde{K}_2$, so we need to know that it does stay in this compact set.

Defining $H_{\text{max}} := \min\{H(u,v) | (u,v) \in \partial\tilde{K}_2\} = \min\{\, H(u_{\text{min}} - R/2, b), H(a, v_{min} - R/2)\,\}$, we know that y_n stays in $\tilde{K}_2$, if

$$H(y_n) - H_0 < H_{\text{max}} - H_0$$

that is, using the bound (5.13), if

$$L\,\gamma M e^{-h_0/2h} + 2hC \leq H_{\text{max}} - H_0. \tag{5.14}$$

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 (5.14) if and only if h is smaller than the unique solution of

$$L\,\gamma M e^{-h_0/2h} + 2hC = H_{\text{max}} - H_0. \tag{5.15}$$

In other words, the bound h^* we are looking for is given by the minimum of $h_0/3$ and the solution of (5.15). Moreover, the numerical solution stays in $\tilde{K}_2$ for at least $t = nh^* \leq e^{h_0/2h^*}$.

□

5.6 Example

As an illustration, we consider the problem

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

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 $K = \{(u,v) \mid |H(u,v)| \leq 1.8822\}$ and then $u_{\text{min}} = 1.135, u_{\text{max}} = 3.222, v_{\text{min}} = 0.4343$ and $v_{\text{max}} = 1.921$. We obtain $R = 0.3909$ and we compute $h_0 = 0.0057$, so we have to choose h to be smaller than $h_0/3 = 0.0019$. Since the solution of

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

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 behaviour, for a time $t = nh \sim 10^{15}$.

It is interesting to note that, as one can see on Figure 5.1, the estimate for h is really pessimistic. For values as large as $h = 0.1$, the Hamiltonian is extremely well-conserved, with $\max_{n \geq 1} |H(y_n) - H_0| = 0.0127$ and the numerical simulation 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

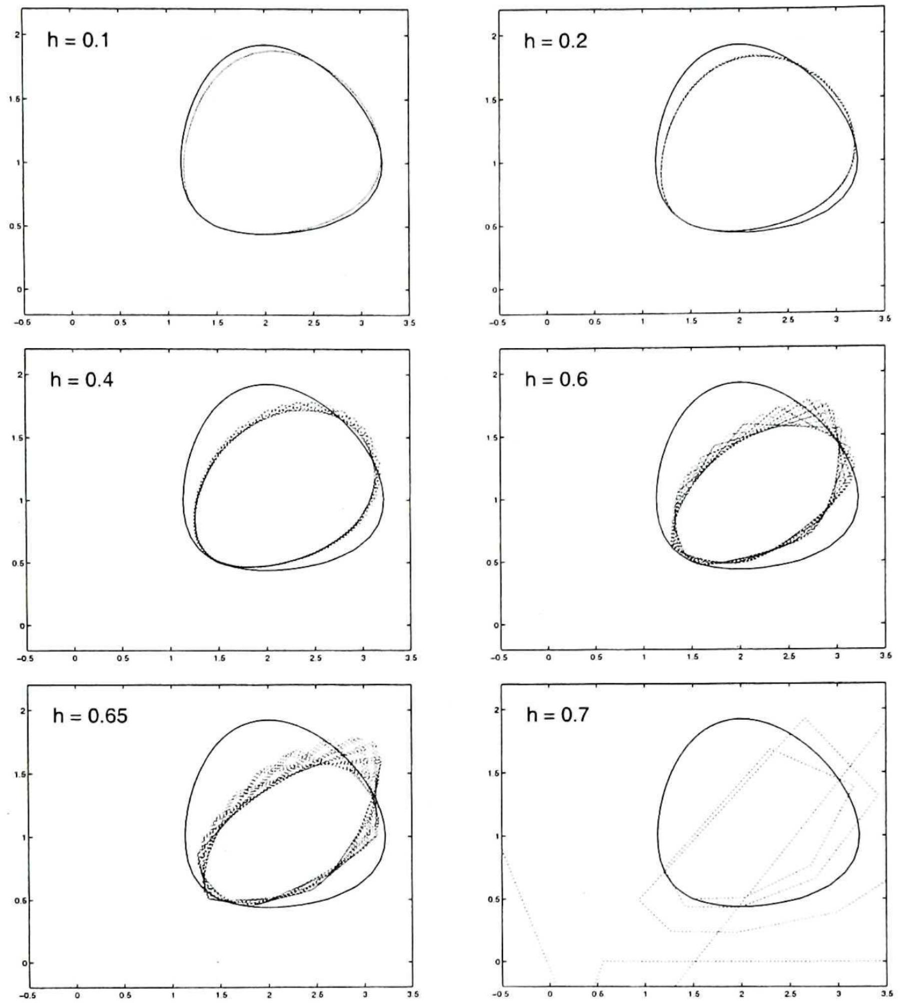

Figure 5.1: Numerical solutions obtained using large step-size.

the problem

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

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 0.00011, but we also proved this result for any similiar problem with initial condition $y_0 = (u(0), v(0)) \in K$, since the initial condition was only used to define the compact set K, and all results are true for all $y_0 \in K$.

Conclusion

In this thesis we mostly focused on two properties of the Lotka-Volterra system: its cyclicity, by the mean of its Poisson structure, and its positivity. As we pointed out in the introduction, geometric integration has been studying, among other topics, numerical methods preserving Poisson structure. Independently of geometric integration, there is current research focusing on the positivity of numerical methods. Our goal was to unify these two subjects. We first studied the symplecticity of the symplectic Euler method and its explicit variant and then we proved that they are Poisson integrators for the Lotka-Volterra system. The latter property explains their excellent performance in general. We finally focused on the positivity of the numerical results.

In the case of the Lotka-Volterra system, we derived simple conditions ensuring the positivity of the symplectic Euler method, whereas we should expect from Figure 2.9 which gives the number of iterations needed for each point to leave the first quadrant, that those conditions are not as simple for its explicit variant (2.3). To study these conditions, we used backward error analysis, but this analysis needed to be more refined than what is generally required. Since this concept is usually employed to study the qualitative behaviour of numerical methods over long time intervals, the idea is in general to prove that the order of the error is small. However for our problem, we needed precise bounds on the error, in Theorem 5.3 in particular, in order to ensure the positivity of the numerical results. The last chapter is an excellent illustration of a concrete application of backward error analysis for Poisson systems.

These results are, of course, very restricted as they are only valid for the symplectic Euler method and its variant when they are applied to the Lotka-Volterra system. However, nothing proves that it will not be possible to first generalize the results and the procedure given in Chapter 5 for the explicit variant of the symplectic Euler method applied to any Poisson system and then to find similar results for other methods.

We observed, in particular, that several second-order methods, such as the implicit midpoint rule, the trapezoidal rule and the Störmer-Verlet method, exhibit the same characteristics as the explicit variant of the symplectic Euler method. Even if, in general, the numerical approximations show the correct qualitative behaviour, they may become negative depending on the choice of the initial conditions and step size. By plotting the number of iterations needed for each point to leave the first quadrant when we apply one of these three methods to the Lotka-Volterra system with $h = 1$, we notice on Figures 5.2, 5.3 and 5.4 that we get trajectories leaving the first quadrant.

For the implicit midpoint rule and the trapezoidal rule, the region consisting of initial values for which the numerical result is positive for 100 iterations is large (almost $[0, 3] \times [0.3]$), whereas the one corresponding to the Störmer-Verlet method is much smaller and similar to the one corresponding to the explicit variant of the symplectic Euler method (see Figure 2.9). Using the tools we derived in this thesis, it should be possible to give conditions on the step size to ensure the positivity of the numerical approximations over long time intervals. One could start by the Störmer-Verlet method as it is a Poisson integrator for the Lotka-Volterra system. Then the results could be generalized to the implicit midpoint rule, which is symplectic but not a Poisson integrator. Finally one could study the case of the trapezoidal rule.

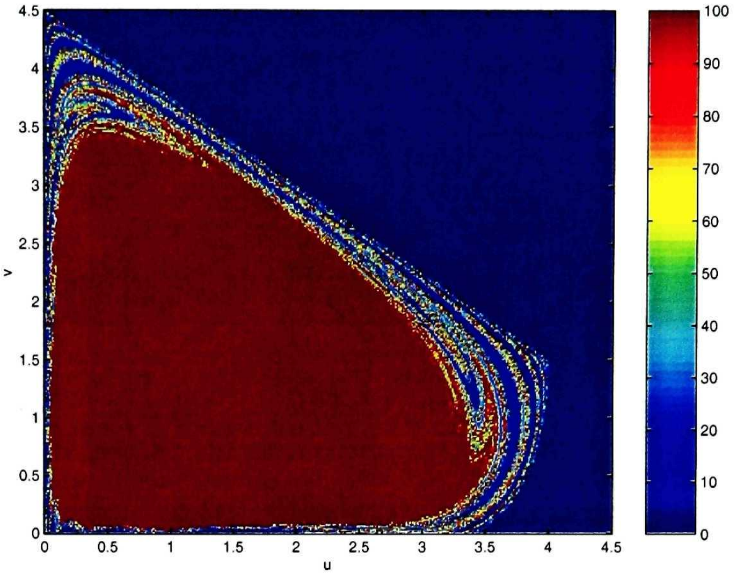

Figure 5.2: Number of iterations needed for each point (u_0, v_0) to leave the first quadrant, when applying the implicit midpoint rule to the Lotka-Volterra system with $h = 1$.

However, it is not obvious our tools still apply to this case since this method is non-symplectic.

There is another technique we briefly studied which however does not appear in this thesis. We transformed the system using a substitution of the type $u = \exp(x), v = \exp(y)$ or $u = x^2, v = y^2$, applied the symplectic Euler method or its variant to the new system, and finally did a backsubstitution to recover the output. The goal of these methods is to enforce the positivity of the numerical results. However, we may lose the symplecticity during the backsubstitution. For example, if we apply the exponential transformation to the symplectic Euler method, we lose symplecticity, but the resulting method is still a Poisson integrator for the Lotka-Volterra system. Even more surprising is the fact that it is a Poisson integrator for the matrix $B(y)$ defined in (3.5) even if the Hamiltonian is not separable. whereas the symplectic Euler method is a Poisson integrator for $B(y)$ only for separable Hamiltonians.

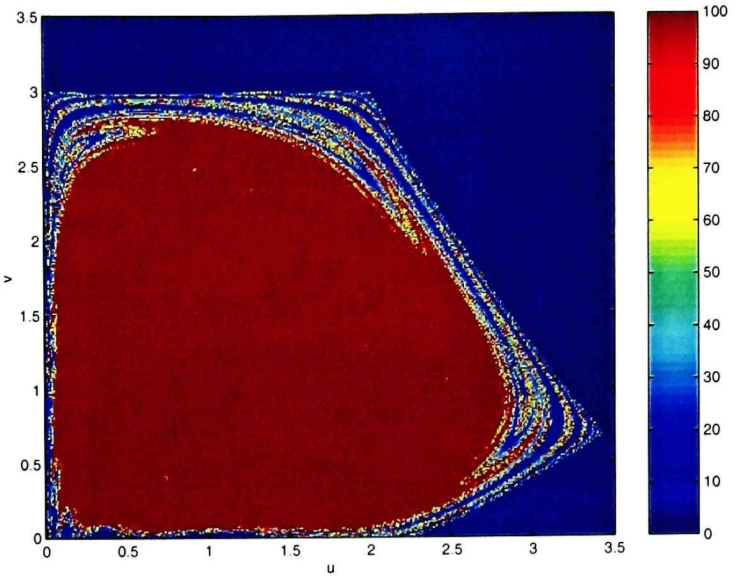

Figure 5.3: Number of iterations needed for each point (u_0, v_0) to leave the first quadrant, when applying the implicit trapezoidal rule, with $h = 1$.

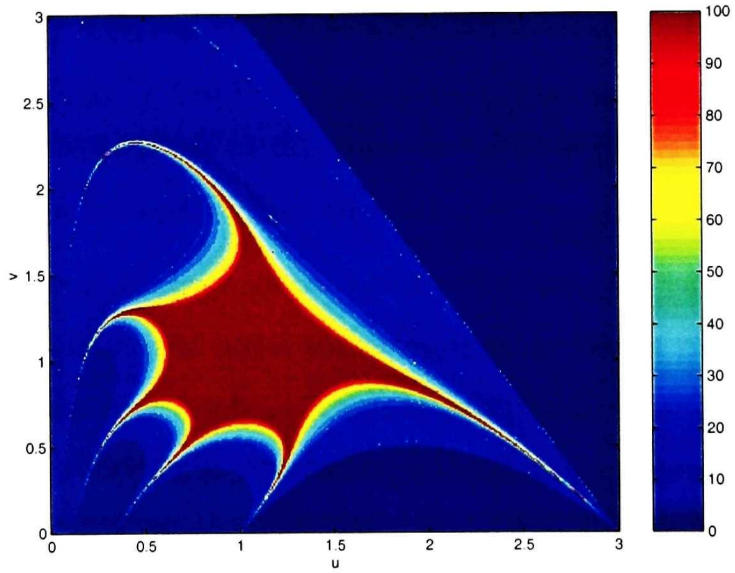

Figure 5.4: Number of iterations needed for each point (u_0, v_0) to leave the first quadrant, when applying the Störmer-Verlet scheme, with $h = 1$.

In future work, it would be interesting to study the opportunities offered by these transformations and to try to find integrators adapted to the transformed systems.

Bibliography

[1] H. BERLAND. *Isotropy in geometric integration.* Diploma thesis. Faculty for Physics, Informatics and Mathematics, Norwegian University of Science and Technology, 2002.

[2] R. DEVOGELAERE. Methods of integration which preserve the contact transformation property of the Hamilton equations. *Submitted for publication*, 1956.

[3] M. GANDER. A non-spiraling integration for the Lotka-Volterra equation. *Il Volterriano*, 4:21–28, 1994.

[4] E. HAIRER. Backward analysis of numerical integrators and symplectic methods. *Annals of Numerical Mathematics*, 1(1-4):107–132, 1994.

[5] E. HAIRER, C. LUBICH, and G.WANNER. *Geometric Numerical Integration.* Springer-Verlag. Berlin, 2002.

[6] M. CALVO. J.M. SANZ-SERNA. *Numerical Hamiltonian Problems.* Applied Mathematics and Mathematical Computation. Chapman and Hall, 1994.

[7] W. KAHAN. *Unconventionnal numerical methods for trajectory calculations.* Lecture Notes. CS Division, Department of EECS, University of California at Berkeley, 1993.

[8] R. MEYER-SPASCHE. M.J. GANDER. An introduction to numerical integrators preserving physical properties. In *Applications of Nonstandard Finite Difference Schemes*, pages 1–67. 2000.

[9] J. SANZ-SERNA. An unconventionnal symplectic integrator of W. Kahan. Technical report, Universidad de Valladolid, Spain, 1994. Report 1994/2.

[10] J. STURGEON and B. LAIRD. Symplectic algorithm for constant-pressure molecular dynamics using a Nose-Poincare thermostat. *Journal of Chemical Physics*, 112(8):3474–82, 2000.

[11] P. TUPPER. *Topics in the Numerical Analysis of Ordinary Differential Equations: Molecular Dynamics and Chemical Kinetics.* PhD thesis, Scientific Computing–Computational Mathematics Program, Stanford University, 2002.