Periodic orbits of the planar N–body problem with equal masses and all bodies on the same path

Carles Simó

Dept. Matemàtica Aplicada i Anàlisi, Univ. de Barcelona

1 Introduction

The differential equations for the planar Newtonian N–body problem are the well known

$$\ddot{q}_j = \sum_{0 \leq i < j \leq N-1} m_i (q_i - q_j) r_{i,j}^{-3}, \tag{1}$$

where $q_i \in \mathbb{R}^2$ and $r_{i,j} = |q_i - q_j|$, $|\quad|$ being the Euclidean norm. From now on we assume the center of mass located at the origin, by using a suitable reference system. The gravitational constant, G, is taken equal to 1.

While the Newtonian two–body problem is an elementary problem and the three–body problem is a major problem in Celestial Mechanics, only partially understood, the N–body problem, for $N > 3$, remains completely unsolved. Only a few solutions are known analytically, related to simple central configurations. In the planar case they give rise to the so–called *relative equilibrium solutions*, in which all the bodies rotate around the center of mass, with constant angular velocity, keeping the mutual distances constant.

In the simplest ones of these solutions the N bodies are located at the vertices of a regular N-gon. In the case $N = 3$ these solutions are the celebrated Lagrange equilateral solutions (Lagrange 1772), and they exist for any values of the masses m_1, m_2, m_3. In the general case the existence of a regular N-gon requires all masses equal.

From now on the masses will be taken all equal. We shall normalize them, so that all are taken equal to 1.

1.1 Motivation

The N-gon solutions offer, in particular, another interesting aspect: all bodies move periodically tracing the same curve on the plane. There is just a time shift in the position to pass from the position of a body to the position of the next one. We can assume, without restriction, that the period is 2π (see later) and, hence, there is a function of time

$$q : \mathbb{R} \to \mathbb{R}^2, \quad t \mapsto q(t) \tag{2}$$

such that the position of body k, $k = 0, \ldots, N-1$, is given by

$$q_k(t) = q(t - 2\pi k/N). \tag{3}$$

Note that, for convenience, we have labelled the bodies from 0 to $N-1$.

A natural question is to ask if, beyond the N-gon solutions, there are other solutions of the N–body problem with all masses equal, either planar or spatial (the spatial case being simpler) such that all the bodies move along the same path. We shall confine our study to the planar case.

1.2 Historical note

The first example of these solutions was found numerically by Moore (1993). A rigorous existence proof has been given by Chenciner and Montgomery (2000), making use of variational methods and of the symmetries of the problem. The solution turns out to be a *figure eight curve*, see Figure 1. A description of the dynamical properties of the figure eight solution, the "satellites" of the figure eight, the nearby solutions and bifurcations can be found in Chenciner et al. 2000, Simó 2000 and Simó 2001b. Soon after the announcement of the Chenciner–Montgomery result, in December 1999, J. Gerver found another such solution for $N = 4$ (Figure 1). At the beginning of 2000 I found a lot of solutions with curves of different shape and values of N ranging from 4 to several hundreds. See Chenciner et al. 2000 and Simó 2000 for additional references. The name *choreographic solutions* has been introduced for these solutions, inspired by the motion of the N bodies along the path, as seen in animations (see, e.g., `http://www.maia.ub.es/dsg/`). The N–gon solutions will be considered as the trivial choreographies.

Fixing the values of G and m_i still allows us to normalize time. We shall use a normalization such that the period of the choreographies is 2π. This fixes the levels of energy and angular momentum of the choreography.

Note that choreographies which are obtained one from the other by rotation or axial symmetry will be seen as the same one.

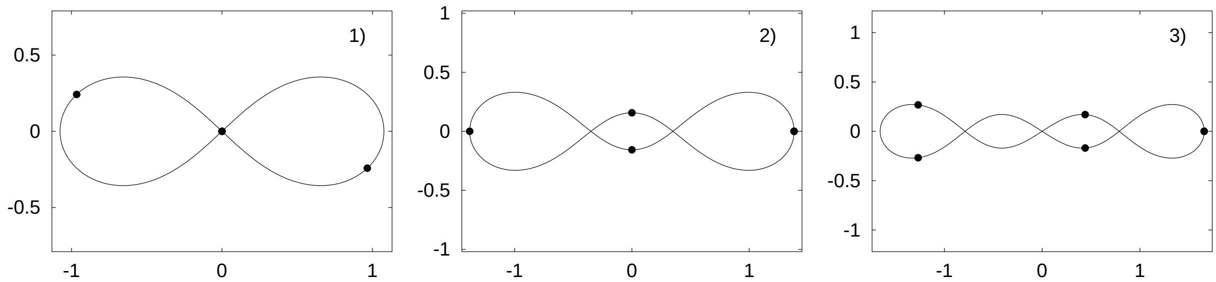

Figure 1. *The first choreographies found, with 3, 4 and 5 bodies, respectively. The dots show the position of the bodies at some initial time.*

1.3 Other potentials

As it will be clear in what follows, it is also important to consider a more general problem. Instead of the Newtonian potential in $1/r$ we can consider other potentials. The homogeneous potentials in $1/r^a$ (or even $\log r$) are particularly relevant and we shall make use of them in the subsequent sections. If the potential is of the form

$$U(q_0, \dots, q_{N-1}) = \sum_{0 \leq i < j \leq N-1} f(r_{ij}) \tag{4}$$

and $f(r)$ in (4) behaves like $1/r^a$, $a \geq 2$ as $r \to 0$, the potential is known as a *strong–force potential*. Strong–force potentials are easy to deal with in order to show the existence of choreographies. For the homogeneous cases we can scale always the period of the choreographies to 2π.

1.4 Theoretical approach

The first question to study is the existence of choreographies. As the problem is of global nature, that is, we cannot rely on local or perturbation methods, it is natural to consider a global approach. This is provided by the variational formalism. The problem is rephrased in terms of the action functional. We should look for extremals of the action.

The space of paths used to make the action extremal must exclude collisions. Otherwise we have solutions of (1) which fail to satisfy the differential equations at collisions.

We shall introduce some choreography classes as suitable sets where we look for extremals of the action. As mentioned, strong–force potentials will lead to simpler solutions. On the other side, for non strong–force potentials we fail to obtain existence proofs by variational methods. Only the figure eight choreography has been proved to exist for the Newtonian potential, by using these methods. The difficulty relies in the fact that we need *a priori* estimates of the action, on some "candidate" to choreography, to show that minimizing the action we do not tend to some path for which collisions occur.

Another relevant point is the requirement that all masses be equal. It is still an open question the existence of choreographies with unequal masses. But Chenciner (2000) has shown that this is impossible for $N \leq 5$.

1.5 Numerical methods

Having existence results it appears the problem of how to find choreographies explicitly. Even if we do not have existence proofs, we can try to approximate these solutions. We have to recur to numerical methods. As it is customary in Celestial Mechanics and other areas in Dynamical Systems, the more efficient method to find periodic solutions (in this case with period 2π) is to look for initial conditions such that we return to them after integration of (1) for one period. That is, we have to look for the zeros of some function. Newton's method is suitable to this end.

In our case we have to use, also, that the time shift from one particle to next one is $2\pi/N$. Furthermore any solution gives rise to other solutions by shifting the origin of time. Hence, we have to "normalize" the initial conditions (e.g., by asking that a given particle has $y = 0$ and $\dot{y} > 0$ at $t = 0$, if (x, y) denote the coordinates in $\mathbb{R}^2$.)

But an essential difficulty, for the convergence of Newton's method, is to have a good starting point. We shall deal with this by a numerical implementation of the minimization of the action. This can give the desired initial conditions for Newton's method.

Another possibility is to use minimization of the action for a homogeneous strong–force potential. This converges to a solution of (1). As soon as a sufficiently good approximation is found, we shift to Newton's method. Then, we use continuation with respect to the exponent of the potential until we reach, if it is possible, the desired exponent (e.g., $a = 1$ for the Newtonian case).

1.6 Extensions

It is also interesting to look for *relative choreographies*. Relative means here that they can be seen as choreographies in rotating axes, but not with respect to a fixed reference frame. Hence, the particles move on a 2D torus instead of moving on a closed curve. For suitable angular velocity of the rotating axes (i.e., rotation with rational frequency) we can have periodic solutions of the N–body problem. These solutions can be choreographies or not, depending on whether the different particles follow the same path or not.

A given choreography can be seen as a fixed point of a canonical transformation (Poincaré map) on a transversal section with fixed energy. It can happen that the differential of the Poincaré map at the fixed point has some elliptic component. Around the given solution we can find *satellite solutions* moving around the basic solution and closing after k revolutions. These solutions can be *satellite choreographies* in case all bodies move on the same path.

See Chenciner et al. 2000 for several examples of relative and satellite choreographies and further discussion.

In what follows we shall concentrate on choreographies for $N \geq 4$. The stronger emphasis will be on methods to find these solutions. We believe that these methods can be useful for many other applications to Celestial Mechanics.

2 The variational formalism

2.1 The action functional

Let us denote by K and $-U$ the kinetic and potential energy of the N–body problem. We have

$$K = \frac{1}{2} \sum_{i=0}^{N-1} m_i |\dot{q}_i|^2, \tag{5}$$

while U is given by (4) with $f(r)$ depending on the chosen potential. Let us introduce the Lagrangian $L = K + U$. It is well known that classical solutions of (1) (i.e., solutions without singularities) must come from *extremals of the action* A. In particular, to find 2π periodic solutions we can look for extremals of

$$A = \int_0^{2\pi} L(t)dt, \quad L(t) = K(\dot{q}_0(t), \ldots, \dot{q}_{N-1}(t)) + U(q_0(t), \ldots, q_{N-1}(t)). \tag{6}$$

If we want to find choreographies we must consider functions $q_j(t)$ as defined by (3) being $q(t)$ as in (2). As q is a 2π periodic function the image $\{q(t),\, t \in [0, 2\pi)\}$ is a closed curve on the plane. That is, the image of a circle $\mathbb{S}^1$. In what follows we shall denote such functions q simply as a *path*. It must be emphasized that the relevant object is not only the curve $q(\mathbb{S}^1)$, obtained as image of $\mathbb{S}^1$ by q, but the *parameterized curve*, that is, the curve and how it is travelled with respect to time.

2.2 The functional space

However not all such paths are admissible. The first condition is that the integral in (6) must be well defined. To this end we require that q belongs to the space $\mathcal{H}$ of functions with *square integrable first derivative.*

On the other side we must prevent from collisions. A collision will occur if there are two values of t, say t_1 and t_2, such that $q(t_1) = q(t_2)$ with $t_1 - t_2 = 2\pi j/N$. In that case bodies numbers 0 and j coincide at the same point when $t = t_1$, and the same happens after time shift $2\pi/N$ with bodies 1 and $j+1$, etc. The subset of functions in $\mathcal{H}$ giving rise to collisions will be denoted as Δ. Hence we must search for $q \in \mathcal{H} \setminus \Delta$.

For further use let us study the double points in $q(\mathbb{S}^1)$. Let t_1 and t_2 time values such that $q(t_1) = q(t_2)$. Then, going from $q(t_1)$ to $q(t_2)$ by increasing (or decreasing) t we have a loop associated to the double point. Let us measure the *time length of the loop*, that is, $|t_2 - t_1|$ taking $2\pi/N$ as the unit of measure. Let $\ell = |t_2 - t_1|/(2\pi/N)$. Then we say that the length of the loop is ℓ. Of course, we can also take $N - \ell$ instead of ℓ (this depends of which sense we use to go from $q(t_1)$ to $q(t_2)$. For concreteness, we shall assume from now on (except otherwise stated) that we take as length of the loop associated to the double point (or, simply, length of the double point) the minimum between ℓ and $N - \ell$. We shall denote again the minimum length as ℓ. We use the symbol $\bar{\ell}$ to denote the integer part of ℓ and refer to $\bar{\ell}$ as the *integer length associated to the double point.* In the exceptional case that a point has multiplicity higher than 2, then several lengths can be associated to that point in a natural way.

It is clear that if $\ell = \bar{\ell}$ then $q \in \Delta$. Hence, to avoid collisions we should restrict the *admissible paths* to those such that the lengths of the double points are not integers.

Note that double points with $\bar{\ell} = 0$ are non-essential. They can be removed (or introduced) by a deformation of the parameterization $\{q(t)\}$ without passing through a double point of integer length during the deformation.

Given q it is possible that by deformation we obtain new paths having a tangency (or loosing it). This is allowed provided the tangency point has not an integer value of ℓ.

2.3 Choreographic classes

This defines a partition of $\mathcal{H} \setminus \Delta$ in connected components, that will be denoted as *choreographic classes.* In other words: Two admissible parameterized closed curves $\{q^{(0)}(t)\}$ and $\{q^{(1)}(t)\}$ belong to the same class if there exist a one–parameter family of parameterizations, $\{q^{(\varepsilon)}(t)\}$, such that no double point of integer length exist for $q^{(\varepsilon)}$ for any $\varepsilon \in [0, 1]$ and it reduces to $\{q^{(0)}(t)\}$ for $\varepsilon = 0$ and to $\{q^{(1)}(t)\}$ for $\varepsilon = 1$. Hence, both $q^{(0)}$ and $q^{(1)}$ have the same topology (letting aside small loops created by non-essential double points and allowed tangencies) and the integer lengths of the corresponding double points are the same. The boundary of a class consists of paths having double points of integer length.

2.4 Results

It is clear that if we multiply a curve $\{q(t)\}$ by a factor ρ then the kinetic part is multiplied by ρ^2 and the potential one by ρ^{-1} (by ρ^{-a} in case $f(r) = 1/r^a$). This prevents ρ from going to ∞ or to 0 and sets bounds on the size of the choreographies (with prescribed masses and period). The same reason shows that the extremals of A cannot be maxima. Therefore, only minima or saddles are possible.

Then the idea is to take an initial function q belonging to the desired class, and start a minimization process. Certainly in this way the probability to reach a saddle is zero, but we shall return later to this point. To minimize A we can follow (theoretically) the gradient flow of A in q in the sense which decreases A.

The problem is that we cannot exclude that this gradient flow tends to a collision. Indeed, if $f(r) = 1/r$, assume that on $q(t)$ there is a collision. At collision both K and U become unbounded. Let r_{min} be the minimum distance between two of the N bodies. By shifting the origin of time we can assume that the collision occurs at $t = 0$. When $t \to 0$ we have $r_{min} \to 0$. But we know that at a collision of the Newtonian N–body problem r_{min} behaves as $|t|^{2/3}$. Furthermore the functions K and U differ in the total energy of the system ($K - U = h$). Hence, the contribution of a collision to the integral in (6) is finite. There is nothing against having a local minimum of A for a q having collisions (see, however, Simó 2000 for some constraints that must be satisfied by the shape of q near collision in order that a q minimizing $A(q)$ contains a collision).

This is impossible for the strong–force cases. Indeed, if $a = 2$ then r_{min} behaves as $|t|^{1/2}$. Then U behaves, locally, as t^{-1} and we have divergence of A. This fact was realized by Poincaré (1896) who was able to prove existence of periodic solutions of the N–body problem for strong–force cases using variational methods.

Using the ideas sketched here the following is proved in Chenciner et al. 2000.

Theorem 1 *Given any choreographic class, there is a periodic solution of any planar strong–force N–body problem which belongs to this class.*

Remark 1 *Note that nothing is said about the uniqueness of the solution. In fact the numerical results show that there are classes with more than one solution, at least in the Newtonian case and also for all values of a close to 1.*

2.5 Symmetries

The simplest choreographies (see Figures 1, 2 and 3) have one or several symmetries. It is easy to take this into account both in the variational formulation and in the effective computations. For instance, in Figure 1 the number of essential initial conditions to determine the solution in the three examples displayed is 2, 3 and 6, respectively. However we shall not use any such reduction in the effective computations. Choreographies without any symmetry have been found starting with $N = 6$.

2.6 Linear chains

Some choreographies present a very simple structure. They are symmetric with respect to the x axis and all double points are on the x axis. Furthermore, if $q(0)$ is the rightmost point on $q(\mathbb{S}^1)$ then the first component, $x(t)$, of q is decreasing from $t = 0$ to $t = \pi$ and, hence, has a single minimum and a single maximum. These choreographies will be denoted as *linear chains*. This is the case in Figure 1, cases 1 and 2 in Figure 2, cases 1 to 4 in Figure 3 and case 3 in Figure 4.

They are characterized by the increasing sequence of integer lengths of the double points (taking the suitable determination, by choosing ℓ or $N - \ell$ as appropriate). A simple count (see Simó 2000) shows that the number of possible linear chains increases exponentially with N. All of them exist for strong–force potentials. It seems reasonable to conjecture that they also exist for the Newtonian potential.

It can also be interesting to name linear chains as *direct* linear chains and to use the name *folded* linear chains if the number of extrema is greater than two. This is what happens in Figure 2 cases 3 to 6 (note that one has to rotate case 5 by $\pi/2$). All of them can be seen as a linear chain with some loops "folded": the left half loop in case 3, the left loop in case 4, the extreme loops folded inside in case 5, one and one half loops in case 6. This is also true in several cases in Figure 4. Note that in Figure 2 case 6 we have 4 double points on the x axis and only 4 bodies. This seems not possible for direct linear chains.

2.7 Open questions

At this point it is convenient to summarize some open questions, without being exhaustive.

- The most important one is to characterize which choreographic classes are realized for the Newtonian potential.
- Which are the folded linear chains which can be realized? How to characterize them?
- Find obstructions to the existence of solutions inside some class for the Newtonian potential.
- Which is the behaviour of a choreography which exists for strong–force potentials but fails to exist for the Newtonian one, when we change continuously the exponent in the potential?
- Up to which values of the exponent can the choreographies be continued?
- To study, for general choreographies, if to have one such periodic solution one should require that all masses are equal (see the point 1.4 above).

Other related questions refer to relative or satellite choreographies.

The strongest difficulty to answer these questions is that the use of variational methods relies on how to avoid collisions. This requires "good enough" candidates to a minimum of the action and a priori estimates, to rule out the existence of collisions. But the estimates depend strongly on the unknown details of the shape of the curve and the time parameterization.

3 Numeric variational methods

We pass to the effective computation of choreographies. The method can be quite different if we have approximate initial conditions (i.e, the values $q(2\pi j/N), \dot{q}(2\pi j/N)$, $j = 0, \ldots, N-1$) or not. We assume, first, that no information on the choreography is available. Only a desired topology and a rough idea of the values of ℓ associated to the double points which make the parameterized curve admissible.

3.1 A representation problem

To look for q is not a numerical problem: we want to find a function. This must be reduced to a numerical one by representing an approximation $\hat{q}$ of the function by a finite set of numbers. As the components (x, y) of q are 2π–periodic functions, a convenient representation for the components $(\hat{x}, \hat{y})$ of $\hat{q}$ is by means of a truncated Fourier series (trigonometrical polynomials)

$$\hat{x}(t) \approx \sum_{k=1}^{K} (a_k^{(x)} \cos kt + b_k^{(x)} \sin kt), \qquad \hat{y}(t) \approx \sum_{k=1}^{K} (a_k^{(y)} \cos kt + b_k^{(y)} \sin kt), \tag{7}$$

But other representations can be useful. It is easy to take into account symmetries in (7). For instance, (direct) linear chains can be parameterized in such a way that x (resp. y) contains only cosine (resp. sine) terms.

Remark 2 *Due to the center of mass condition, a choreography with N bodies cannot have harmonics of any order multiple of N in the (x, y) components.*

But it is not necessary to impose this condition a priori. The variational method will lead these coefficients to zero.

3.2 The starting point

We need some initial guess of the form (7) to start with a minimization process. One can start, for instance with a handmade sketch of the curve done with the mouse on a graphical screen and some indications on timing. The values of (x, y) on the marked points can be converted to equispaced data with respect to time (by interpolation). Then a Fourier analysis (and some filter to skip irregularities of hand drawing) is used to get the initial approximation.

With these data the position of the center of mass will be not constant, in general. But, as said before, this is irrelevant for the initial data. Any method we use will take care of this and lead it to a constant position. This one can be shifted to the origin by a translation. Or we can skip the coefficients of the constant part of $\hat{x}$ and $\hat{y}$ from the plot giving rise to our initial guess.

3.3 Minimizing the action

To do a numerical implementation of the variational method the simplest approach is to mimic the ideas of the proof. We want to approximate the function q by a function

$\hat{q}$, either trigonometric polynomials like in (7), the values at a set of equispaced points or some other method. Let us denote as $\mathcal{Q} \in \mathbb{R}^M$ the finite set of parameters (Fourier coefficients, coordinates of points, etc) needed to represent $\hat{q}$. The action A as defined in (6) is approximated by a *discretized* map

$$\hat{A} : \mathbb{R}^M \to \mathbb{R}, \tag{8}$$

where the integral is computed by a numerical quadrature formula. The time derivatives are computed by differentiation of the trigonometric polynomial, or by finite differences (which introduce an additional error in $\hat{A}$), etc. Due to the time shift appearing in (3), which must be preserved by the discretization, it is enough to do the integrals from 0 to $2\pi/N$. Taking into account that collision–free solutions are analytical, the trapezoid rule is highly convenient if trigonometric polynomials are used. In that case any time step can be used to compute the integral, while in the case of a discrete set of points we can use only the data at the prescribed times for which we know approximations of q. Using trigonometric polynomials one has to compute the components of $\hat{q}(t_j)$, for the required t_j, before evaluation of $\hat{A}$, but this can be done quickly by using FFT methods.

Then we proceed to minimize $\hat{A}(\mathcal{Q})$ using a gradient method, some variant or a combination of variants.

3.4 Advantages of the method

The method has clear advantages:

- It is quite robust. It allows very rough starting values, $\mathcal{Q}^{(0)}$, obtained as described before, and leads to some local minimum of $\hat{A}$,
- It is easy to program and to use with any potential,
- As we have an explicit formula for the discretized action $\hat{A}$, the gradient can be obtained from $\mathcal{Q}$ without need of numerical differentiation,
- When an approximate minimum is reached we can do checks, looking for the invariance of the energy for the different time steps. We can also compute the residual acceleration: the difference between the acceleration of the masses computed from $\hat{q}$ (by derivating twice) and using the equations of motion (1). This gives a clear indication of the goodness of the solution found.

3.5 Difficulties of the method

But the variational implementation has several problems:

- Except in quite simple cases M has to be selected large, specially if there are passages close to collision. In this case there are sudden changes of $q(t)$ and this means that the number of harmonics must be large or that the number of equispaced (in time) points must be very large. Typical values of M are in the range $[10^3, 10^4]$. It is possible to use different spacings in time in different parts of the curve, but this adds some complexity to the algorithm,

- The function $\hat{A}$ is very flat, possibly with lots of extrema. The number of iterations to achieve a good approximation is also large (say, again in $[10^3, 10^4]$). This slows down the process.

- It can happen that $\hat{q}$ jumps from a choreographic class to another one when the minimization process leads to the vicinity of a collision,

- The variational technique does not provide direct information on the stability properties of the periodic solutions.

- One should be aware, when looking at the residual acceleration, that a bad determination of the higher order harmonics can have a mild influence on the potential, a medium size influence on the kinetic energy and a large influence on the acceleration $\ddot{\hat{q}}$. This implies that after obtaining an approximation, it is suitable to look at the contribution of the highest order harmonics in (7) and to increase the number of harmonics to obtain a better approximation. Working with an increasing number of harmonics can also give significative savings in computing time.

The difficulties related to the number of harmonics and to changes of class are much less in case the potential is strong–force. Then one cannot come too close to a collision. This suggests (see later) that it can be faster to proceed to the computation of the approximate solution with $a = 2$ (even if one is interested in $a = 1$), then refine the solution by Newton's method and use continuation to go back to $a = 1$, if this is possible.

3.6 Looking for saddle points of the action

Up to now we have considered only minima as extremals of the action. The standing question is how to find saddle points. As before we are interested in obtaining approximations by using a variational method and then to shift to Newton's method for further refinement. Newton's method will not care whether the choreography is a minimum or a saddle of the action.

Assume we have two different local minima of A on the same choreographic class. The so–called "mountain pass lemma" ensures that there is an arc (in the functional space $\mathcal{H} \setminus \Delta$) joining both minima and such that the point of the arc on which the action has a maximum (along the arc) is a saddle point of the action. The intuitive idea is quite "geographical".

We can implement this idea numerically. Let us consider two approximate minima of the action $\hat{q}^{(0)}$ and $\hat{q}^{(1)}$. Let $\hat{q}^{(s)}$, for $s \in [0,1]$, be an arc joining $\hat{q}^{(0)}$ and $\hat{q}^{(1)}$. We can take a segment in the space of Fourier coefficients, for instance, but we should check that the elements of the arc do not leave the choreography class. (This can need more sophisticated arcs, like some polygonal in the space of approximations).

Then we can subdivide the arc in a prescribed number of points, equally spaced for instance $s_i = i/L$, $i = 1, \ldots, L-1$. Starting at each one of these points we can minimize the action. After every step of minimization of all the points in the arc, we can reexamine the distance between those points. The ones close to one of the minima will approach it. So, they can be discarded. The intermediate points can increase their distance, so that new points have to be added (by interpolation).

At every step we also check for the maximum of the action along the arc. The process finishes when in two (or more) consecutive steps the value of this maximum remains essentially constant. This gives the saddle point.

The process is time–consuming. It is very convenient the implementation on an array of processors, so that every processor deals with one (or a few) of the points on the arc.

4 Zero–finding methods

Let φ_t denote the flow time t of the differential equation (1), with the natural modifications in case we are dealing with a potential other than the Newtonian one. Let us denote as w_j the vector containing the components of q_j and $\dot{q}_j$ at $t = 0$ (4 components in the planar case). Let w the vector containing all the w_j components. The principle consists in searching for some w such that

$$\hat{G}(w) := \varphi_{2\pi}(w) - w = 0. \tag{9}$$

In this section we present methods to solve (9), the reductions to be done, difficulties which appear and how to overcome them. Finally we comment on how to turn the numerical finding of these periodic orbits into a computer assisted rigorous proof.

4.1 Fundamentals

To solve (9) we recall, as it is well known, that the solution, if it exists, is not unique. Given any vector w of initial conditions then $\varphi_t(w)$ gives another set of initial conditions. To single out one of them we can do the following: Consider the function q and look for a value of t^* such that $|q(t^*)|$ is maximal. By replacing t by $t-t^*$ and rotating the reference frame we can always assume that at $t = 0$ $q(0)$ is on the x axis and $\dot{q}(0)$ is orthogonal to $q(0)$. That is $x(0)$ is at maximal distance from the origin, $y(0) = 0$, $\dot{x}(0) = 0$, $\dot{y}(0) \geq 0$.

It is possible to reduce the dimension of the system to be integrated, by using the center of mass condition. I have not used this reduction. Rather the preservation of the first integrals has been used as a test. After all, faced to a problem with, say, 100 bodies, the reduction of the size of the system from 400 to 396 is not very important.

As mentioned in the variational part, it is not necessary to use $\varphi_{2\pi}$. The choreographic character of the solution allows to reduce (9) to

$$G(w) := \varphi_{2\pi/N}(w) - \sigma(w) = 0, \tag{10}$$

where $\sigma(w) = (w_{N-1}, w_0, w_1, \ldots, w_{N-2})^T$ is obtained from w by (right) cyclic permutation. This saves computations and errors. We shall represent σ by a permutation matrix with the same name.

To solve (10) we use Newton's method. This requires the evaluation of $\varphi_{2\pi/N}(w)$, that is, the numerical integration of (1) and the computation of the differential $DG(w) = D\varphi_{2\pi/N}(w) - \sigma$.

4.2 Numerical integration of ODE

It is suitable to obtain solutions of (1) with small error. A convenient procedure is to use Taylor's method. For simplicity of notation let us consider the general autonomous, analytic (assuming no collisions) differential equation

$$\dot{w} = F(w). \tag{11}$$

Assuming $w(t)$ known we want to compute the coefficients of the expansion

$$w(t+h) = w(t) + \Sigma_{k\geq 1} a_k h^k. \tag{12}$$

Then we have to select a suitable value of h and the number of terms to be kept, so that the relative errors are of the order of magnitude of the rounding errors with the current accuracy (or tolerance) used. Using this scheme the computations are done in the most efficient way.

The algorithm producing the components of $F(w)$ from the components of w uses, in general, the arithmetic operations, logarithms and exponentials (eventually trigonometric functions). The expansions for all the intermediate variables can be obtained recurrently by repeated application of Leibniz rule. Note that it is essential to produce the *numeric* coefficients in the expansions, rather than long *analytic* expressions for each of them. It is obvious than the computational cost to reach some given order, K, is $O(K^2)$. The appendix in Simó 2001a contains some details and a proof about the optimal value of h (from the point of view of efficiency), which has to be taken equal to the current radius of convergence of (12) divided by e^2. The current radius of convergence can be estimated from the experimental behaviour of the coefficients in (12). This gives also the suitable order to truncate (12). For small enough tolerances a "typical" optimal order is $K_{\rm opt} \approx \log(1/\varepsilon)/2$, where ε denotes the current tolerance.

In Jorba–Zou 2001 there is an explicit procedure for the automatic implementation of Taylor's method for arbitrary analytic F.

Note that the method is the most suitable one in case we are interested in very small rounding errors (say 10^{-100} leading to $K_{\rm opt} \approx 115$). This will be relevant later on.

4.3 Variational equations and stability

To compute $DG(w)$, required to do the successive Newton iterates $w^{(m+1)} = w^{(m)} - (DG(w^{(m)}))^{-1}G(w^{(m)})$, we have to integrate the first order variational equations associated to (11), that is

$$\frac{d}{dt}D_{w(0)}\varphi_t(w(0)) = D_wF(\varphi_t(w(0)))D_{w(0)}\varphi_t(w(0)), \quad D_{w(0)}\varphi_0(w(0)) = Id. \tag{13}$$

Furthermore, having computed the monodromy matrix $D_{w(0)}\varphi_{2\pi}(w(0))$ we can study the stability of the periodic solution found so far. Note that only the "partial matrix" $D_{w(0)}\varphi_{2\pi/N}(w(0))$ is available, but we can recover easily the full period matrix by composing with copies of the same matrix shifting, successively, rows and columns by $\sigma, \sigma^2, \ldots$ etc.

We remark that (13) has to be integrated simultaneously with (11). Taylor's method is again suitable for this integration. In the choreography problem and for large N the integration of the variational equations can be time–consuming. For 100 bodies we need to integrate, in all, around 160,000 equations. However, the columns in (13) evolve independently. This means that the task can be distributed to different processors. Each one of them can compute the contribution to (13) of a few columns. Despite all processors have to integrate (11), this results in important savings in computer time.

For further use we also display here the second order variational equation

$$\frac{d}{dt}D_{w(0),w(0)}\varphi_t(w(0)) = D_{ww}F(\varphi_t(w(0)))(V,V) + D_wF(\varphi_t(w(0)))D_{w(0),w(0)}\varphi_t(w(0)), \quad (14)$$

where V stands for $D_{w(0)}\varphi_t(w(0))$ and in (14) we set $D_{w(0),w(0)}\varphi_0(w(0)) = 0$, as initial condition.

4.4 Parallel shooting

One of the difficulties in finding same kinds of choreographies, especially if N becomes large, is the strong instability they show. Dominant eigenvalues of the monodromy matrix reach easily values like 10^{100} and larger. This means that initial (unavoidable) errors in the starting point or the errors introduced by the numerical methods increase in such a way that the computed orbit goes away from the true one.

But this is a common problem in boundary value problems and there are simple ways to prevent the growth of the errors. The most popular one is the *parallel shooting* method (also known as multiple shooting). See, e.g., Stoer–Bulirsch 1983.

For the moment we forget that we are interested in choreographies and assume we are faced to the problem of finding a 2π–periodic solution of an equation like (11) (still autonomous, but there is no need of that). We saw before that this can be solved by looking for zeros of (9). To cope with the strong instability we can introduce $P-1$ auxiliary intermediate values of t: $0 < t_1 < t_2 < \ldots < t_{P-1} < 2\pi$, either equispaced or not. Instead of looking for an initial value $w(0)$ we try to find also the values of w at the intermediate epochs $w(t_1)$, $w(t_2)$, $\ldots$, $w(t_{P-1})$ and impose the "matching and closing" conditions:

$$\begin{gathered}\varphi_{t_1}(w(0)) = w(t_1),\ \varphi_{t_2-t_1}(w(t_1)) = w(t_2),\ \ldots,\ \varphi_{t_{P-1}-t_{P-2}}(w(t_{P-2})) = w(t_{P-1}),\\ \varphi_{2\pi-t_{P-1}}(w(t_{P-1})) = w(0).\end{gathered} \quad (15)$$

As the time intervals are now shorter, the partial instabilities are less dramatic than in the previous approach.

Again (15) can be solved by Newton's method. The variational matrices associated to $\varphi_{t_j-t_{j-1}}$ are required. The system has larger dimension (if $w \in \mathbb{R}^n$ instead of dimension n we shall have nP). But the numerical errors in computing $\varphi_{t_j-t_{j-1}}$ are much less than before and the condition number of the linearized system is smaller.

Readers familiar with topics in Dynamical Systems can note that this idea is similar to the use of pseudo-orbits to obtain true orbits in a hyperbolic system by using the shadowing lemma. Readers familiar with spacecraft mission design can realize that the closing errors in the matching when we stop Newton's iterations play the same role as small mid course manoeuvres to correct errors coming from model, tracking and execution of previous manoeuvres.

In the present problem one has to modify slightly the conditions in (15). The values $t_1, \ldots, t_{P-1}$ belong to $(0, 2\pi/N)$ instead of $(0, 2\pi)$. The intermediate conditions have identical form, but the last condition should read $\varphi_{2\pi/N - t_{P-1}}(w(t_{P-1})) = w(2\pi/N) = \sigma w(0)$. The unknowns are the components of $w(0), w(t_1), \ldots, w(t_{P-1})$.

4.5 Continuation with respect to the exponent of the potential

The choreographies obtained by the previous procedures can be continued with respect to the exponent a. To this end we only need into account, in one of the formulations above, that a can be seen as a parameter. Using the version of (10) we have the condition $G(w, a) = 0$. If $w \in \mathbb{R}^n$ it is convenient to consider (w, a) as a new variable $z \in \mathbb{R}^{n+1}$ and not to make a distinction between the components of z.

Then G is a map from an open set in $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$. If rank $D_z G$ is maximal $(= n)$ then $G(z) = 0$ defines locally a curve and we are at a regular point. The tangent to the curve is given by ker $D_z G$. Otherwise we are in presence of bifurcations. Note that a "turning point" along the curve where a reaches a local extremum is a regular point, in general.

As the tangent to the curve is available at every regular point, we can set up a differential equation for the curve. A convenient parameter is the arc length. It is convenient to refine the point, from time to time, to prevent from too large deviations from zero of the values of $G(z)$. This can be done by using a *modified Newton's method* where no variable is privileged (see Simó 1990).

The only additional thing we need are the derivatives of G with respect to a. To obtain them one has to add an extra variational equation, which accounts for the variation of the flow with respect to the parameter.

4.6 Computer assisted proofs

An interesting question (and not only for the present problem) is *how to turn the numerical computations into mathematical proofs.* This aspect has not yet been implemented, but we proceed to the description of the algorithm able to lead to a complete *computer assisted proof* of the existence of choreographies for the Newtonian potential and different values of N.

The first idea is to have some "explicit" version of the implicit function theorem. Different possibilities are available. For our purposes it is convenient to state the following version of Newton–Kantorovich theorem (see Isaacson–Keller 1966 and Stoer–Bulirsch 1983 for different versions). Assume we want to solve the equation $G(w) = 0$ by Newton's method starting at some initial point w_0. Then

Theorem 2 *Assume G is a $\mathcal{C}^2$ function on a ball $\mathcal{B}$ of radius 2α around a point w_0 and the following holds:*

- *The first Newton's correction $-(DG(w_0))^{-1}G(w_0)$ has norm bounded by α.*
- *The norm of $(DG(w_0))^{-1}$ is bounded by β.*

- *On $\mathcal{B}$ the norm of D^2G is bounded by γ.*

Then if $\alpha\beta\gamma < 1/2$ all the iterates of Newton's method fall in $\mathcal{B}$, there is a unique solution of $G(w) = 0$ in $\mathcal{B}$ and the convergence is quadratic. If $\alpha\beta\gamma = 1/2$ one has still a unique solution and convergence, but the convergence is only linear.

Note that the theorem is valid in Banach spaces (no need of finite dimension) and that other versions require only G to have Lipschitz first derivative.

To apply it to an existence proof of choreographies we consider as function G the reduced version appearing in (10) or, even more convenient, the modification introduced in (15) by using some intermediate time values for the parallel shooting. For concreteness let us denote as $G^*(W) = 0$ the current system of equations. We proceed with the following steps:

- First we produce a good estimate of the solution that we take as W_0. To this end it is suitable to work with high accuracy (say, 100 decimal digits).

- Starting at W_0 we integrate both the equations of motion and the first order variationals using interval arithmetics (based on the 100 decimal digits arithmetics used previously, for instance). We end up with $G^*(W_0)$ and $DG^*(W_0)$. Let δ be a bound of the components of $G^*(W_0)$ (now given as intervals). Let β be a bound derived from the interval containing the norm of $(DG^*(W_0))^{-1}$. The value of β is the one required by the theorem and $\delta\beta$ is a bound for α.

- The more delicate part is the estimate of γ. Values of the norm of $D^2G^*(W_0))$ can be obtained by using again interval arithmetics in the integration of (14), but what is needed now is a bound on a ball around W_0. To shorten notation in (14) we shall skip the subscripts which refer to the variable used for the differentials. Instead of D_{ww} we shall use D^2, for instance. The meaning and the variables used become clear from the context.

- Assume we start the integration of (14) not at W_0 but at a point $W_1 \in \mathcal{B}$. The norm of $\xi := W_1 - W_0$ is bounded by 2α. Let $c(t)$ be the supremum of the norm of $D^2\varphi_t(W_1)$ for $W_1 \in \mathcal{B}$. Then $c(t)$ satisfies the differential inequality

  $$\frac{d}{dt}c(t) \leq \sup ||D^2F(\eta)||(\sup ||D\varphi_t(W_0) + c(t)(\xi)||)^2 + \sup ||DF(\eta)||c(t), \qquad (16)$$

  where

  $$c(0) = 0, \quad \eta = \varphi_t(W_0) + D\varphi_t(W_0)\xi + \frac{1}{2}c(t)(\xi,\xi)$$

  and the supremum in (16) is taken with respect to all ξ with $||\xi|| \leq 2\alpha$.

- We bound $||\eta - \varphi_t(W_0)||$ by $||D\varphi_t(W_0)||2\alpha + c(t)2\alpha^2$. This defines a tubular neighbourhood around $\varphi_t(W_0)$ where we have to bound $||D^2F||$ and $||DF||$ on that neighbourhood. The bounds are rather simple because they rely, essentially in bounds of the mutual distances $r_{i,j}$ between bodies. It is clear that the maximum radius of the neighbourhood goes to zero when $\alpha \to 0$.

- We proceed to the estimates by trial and error. Assume that the maximum radius of the tubular neighbourhood is bounded by ρ and let m_1, m_2 be upper bounds of $||DF||$ and $||D^2F||$, respectively. Note that estimates of m_1, m_2 when $\rho = 0$ are easy to obtain. Solve equation (16) (with interval arithmetics) by using these bounds and replacing $\leq$ by $=$. Denote as $\bar{c}(t)$ the variable, which will be a bound for $c(t)$. This gives

$$\frac{d}{dt}\bar{c}(t) = m_2(||D\varphi_t(W_0)|| + \rho\bar{c}(t))^2 + m_1\bar{c}(t), \quad \bar{c}(0) = 0. \tag{17}$$

- Check if with the value of $\bar{c}(t)$ at the end of the time interval, the assumption on ρ is satisfied. This has to be true for all time intervals used for the parallel shooting. If it holds, let $\bar{c}^*$ be the maximum of the bounds on $\bar{c}(t)$ at the end of the intervals. This is an upper bound for γ. Finally check the condition $\alpha\beta\gamma \leq 1/2$. If the assumption on ρ is not satisfied we can try different values for ρ. If none of them works we have to decrease α. The same happens if the condition on ρ holds but $\alpha\beta\gamma > 1/2$. To try to reduce α we can improve the estimate on W_0 by working with interval arithmetics with higher accuracy.

Hence, if we are not in presence of some bifurcation (i.e., if $(DG^*(W_0))^{-1}$ has bounded norm) and we are able to produce estimates of W_0 with increasing accuracy, we will obtain a complete proof of the existence of the choreography.

5 A sample of choreographies

Figures 2, 3 and 4 display several examples of Newtonian choreographies. They show part of the choreographies found for 4 bodies, 5 bodies and some additional examples. For $N = 4$ and $N = 5$ all the choreographies found have some symmetry (at least an axial symmetry).

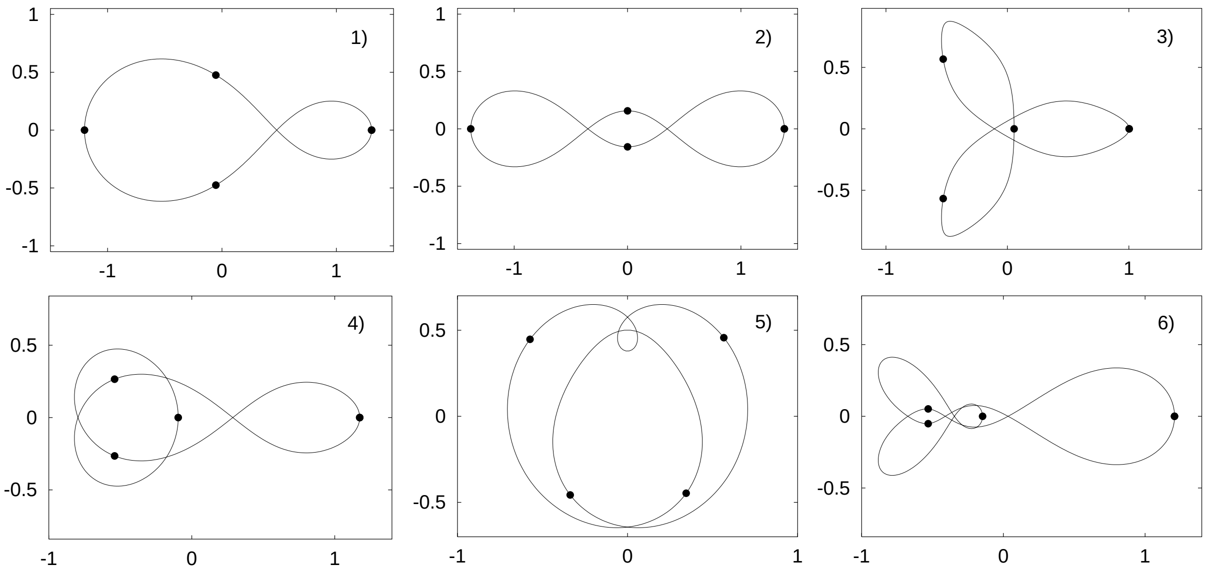

Figure 2. *Choreographies for four bodies under the Newtonian potential. Note that there is another choreography, also with $N = 4$, very similar to case number 5.*

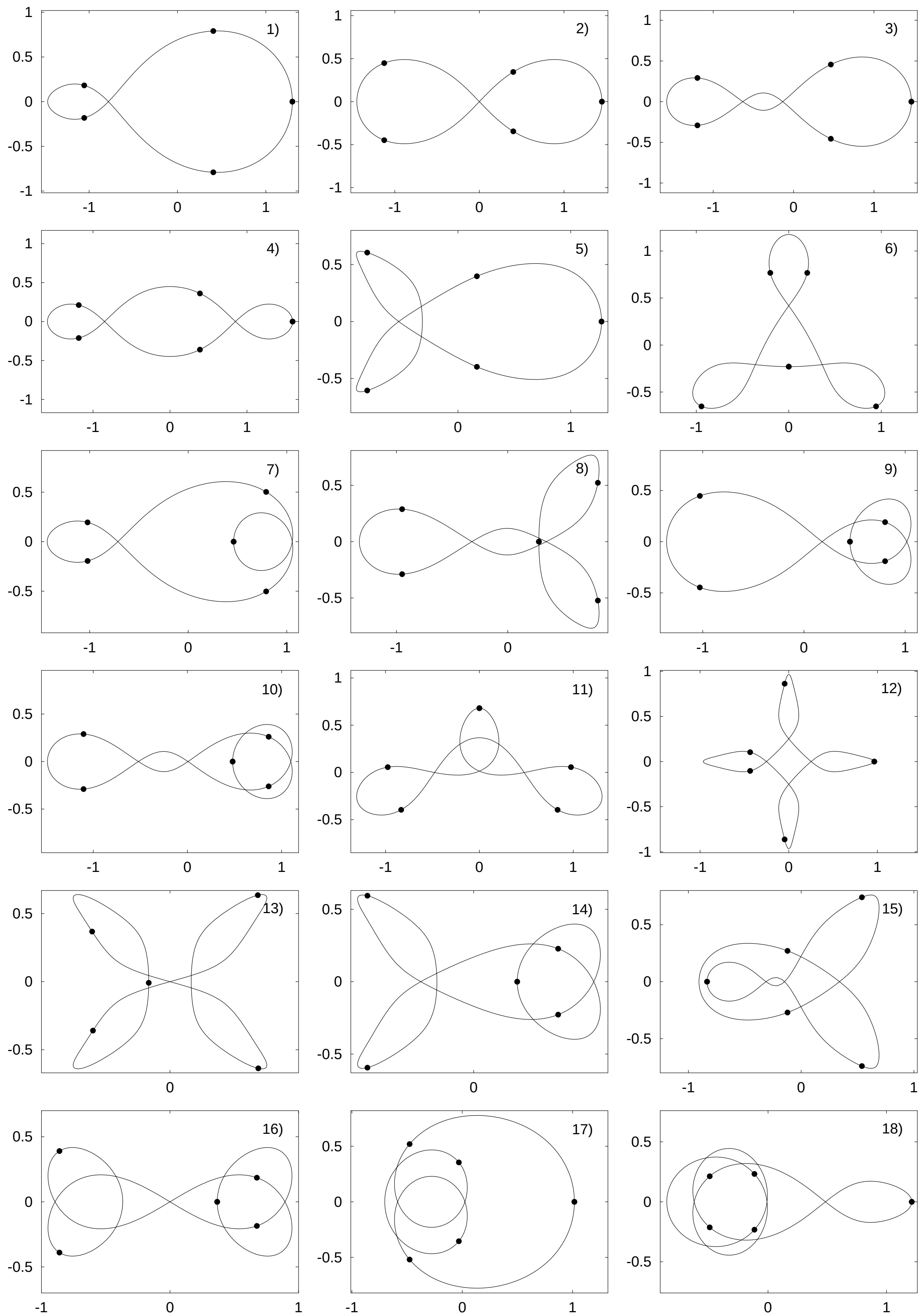

Figure 3. *Choreographies found for 5 bodies. The dots denote initial conditions. Note that all of them have, at least, one (geometrical) symmetry.*

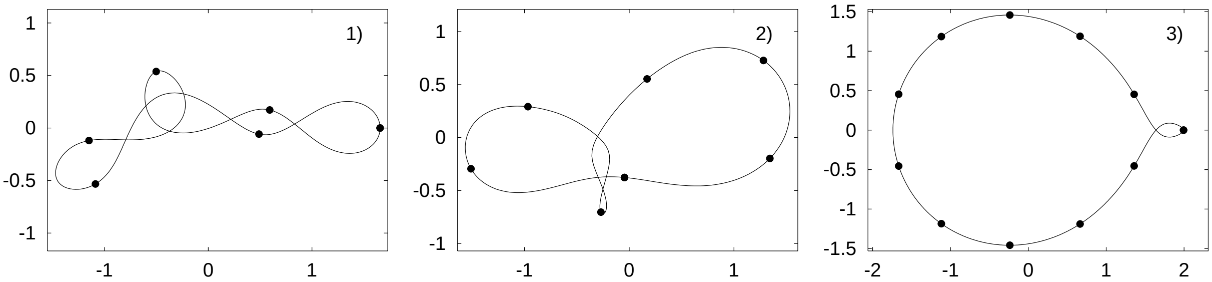

Figure 4. *Case 1: A choreography without any geometrical symmetry for $N = 6$. Case 2: Idem for $N = 7$. Case 3: A circle–like choreography with a small outer loop.*

On Figure 4 we show examples of choreographies without any symmetry. The lowest value of N for which this kind of examples has been found is $N = 6$.

The third choreography displayed in Figure 4 illustrates a different phenomenon. It contains 11 bodies and has a small loop with integer length $\bar{\ell} = 1$. Case 1 in Figures 2 and 3 shows similar choreographies with $N = 4, N = 5$. No problem has been found in finding this type of solutions (small outer loop) for all the tested values of N (up to 100).

However, we had no success when trying a similar pattern but with the small loop (of integer length 1) *inside* the large loop. Hence, it seems that not all the possible choreographies are realized for the Newtonian potential.

On the other hand there is no problem in finding N–body choreographies of circle–like type with a small loop inside for $a = 2$. Using continuation with respect to a it is possible to proceed up to values of a close to $a = 1$. But before reaching $a = 1$ a turning point appears and, along the family of choreographies of these type, a increases from the turning point on. Other turning points can appear later. If we denote by a_N the value of a at which the turning point appears the numerical experiments carried out up to now give a good evidence that $a_N \approx 1 + cN^{-2}$, for some $c > 0$. If this is the real behaviour we have the surprising fact that for any fixed $a > 1$ these choreographies exist from some N on, but never for $a = 1$.

We can return to the circle–like choreography with a small outer loop. As said they seem to exist for all N if $a = 1$. Using continuation again a turning point appears for some value $a*_N < 1$. But in this case one has $a*_N \approx 1 - d/N$ for some $d > 0$.

For $N = 4$ two different choreographies have been found on the same choreographic class. This has been obtained by continuation of case 5 in Figure 2. Trying to decrease a a turning point is found close to $a = 1$. The family of periodic solutions returns to $a = 1$ producing a similar, but definitely different, choreography.

The figure eight solution for $N = 3$ is in some sense quite exceptional. For the Newtonian potential is the only stable choreography found up to now. Certainly other choreographies have been found to be stable for some ranges of a, but not for $a = 1$. Furthermore is seems quite robust concerning continuation. It persists up to values of a beyond the logarithmic potential (which is the potential replacing the case $a = 0$). Among the other choreographies the only ones which persist until the logarithmic potential are also the figure eight ones with an odd N.

Finally, it is worth to mention that when N goes to infinity there are some shapes which seem to tend to a limit after scaling (eventually using different scalings in the x

and y directions). These are the figure eight solutions and the direct linear chains with N bodies and $N-1$ loops.

For further information see Chenciner et al. 2000 and Simó 2000.

6 Discussion

The existence of the choreographies and the variety of patterns they display comes as a surprise, especially in what concerns complicated or even asymmetric patterns. Every piece of the N–body problem we start to understand opens a new world of questions. The choreographies are, in some sense, the simplest examples after the central configurations. For both topics the existence and classification problems seem to be far from reach. In any case they will probably contribute to the advance of Celestial Mechanics due to the new ideas and methods to be developed in order to study them.

Acknowledgments

This work was born from the information that A. Chenciner, R. Montgomery and J. Gerver shared with me and has been largely influenced by many discussions on the topic. I am also indebted to R. Martínez and À. Jorba for many useful suggestions. A large part of the work was started during a sabbatical leave at the Institute de Mécanique Celeste (Observatoire de Paris), thanks to the support of CNRS. My gratitude to my colleagues at that institution for the hospitality and interest on the work. The parallel computing facilities of the UB Grup de Sistemes Dinàmics have been widely used. The support of grants BFM2000–805 (Spain), 2000SGR–27 (Catalonia) and INTAS 97–771 is also acknowledged.

References

[1] Chenciner A, 2000. Private communication

[2] Chenciner A and Montgomery R, 2000. A remarkable periodic solution of the three body problem in the case of equal masses, to appear in *Annals of Mathematics*

[3] Chenciner A, Gerver J, Montgomery R and Simó C, 2000. Simple Choreographic Motions of N Bodies: A Preliminary Study in *Geometry, Mechanics and Dynamics* to appear (New York: Springer)

[4] Isaacson E and Keller H B, 1996. *Analysis of Numerical Methods* (New York: John Wiley)

[5] Jorba À and Zou M, 2001. On the numerical integration of ODE by means of high-order Taylor methods. Preprint

[6] Lagrange J L, 1772. Essai sur le problème des trois corps *Œuvres*, Vol. **6** p 273

[7] Moore C, 1993. Braids in Classical Gravity *Physical Review Letters* **70** 3675–3679

[8] Poincaré H, 1896. Sur les solutions périodiques et le principe de moindre action *Comptes Rendus Acad. Sci. Paris* **123** 915–918

[9] Simó C, 1990. Analytical and numerical computation of invariant manifolds in *Modern Methods in Celestial Mechanics*, ed. D. Benest and C. Froeschlé 285–330 (Paris: Editions Frontières)

[10] Simó C, 2000. New families of Solutions in N–Body Problems in *Proceedings of the ECM 2000* to appear (Basel: Birkhäuser)

[11] Simó C, 2001a. Global Dynamics and Fast Indicators in *Global Analysis of Dynamical Systems*, ed. H W Broer, B Krauskopf and G Vegter (Bristol: IOP Publishing) to appear.

[12] Simó C, 2001b. Dynamical properties of the figure eight solution of the three-body problem in *Proceedings of the Chicago Conference on Celestial Mechanics dedicated to Don Saari* to appear

[13] Stoer J and Bulirsch R, 1983. *Introduction to Numerical Analysis* (Heidelberg: Springer, second printing)