THE ASTRONOMICAL JOURNAL VOLUME 69, NUMBER 1 FEBRUARY 1964

The Applicability of the Third Integral Of Motion: Some Numerical Experiments

MICHEL HÉNON* AND CARL HEILES

Princeton University Observatory, Princeton, New Jersey

(Received 7 August 1963)

The problem of the existence of a third isolating integral of motion in an axisymmetric potential is investigated by numerical experiments. It is found that the third integral exists for only a limited range of initial conditions.

1. INTRODUCTION

THERE has recently been a renewal of interest in the question of the existence of the third integral of galactic motion (Contopoulos 1957, 1958, 1960, 1963; Barbanis 1962; van de Hulst 1962, 1963; Ollongren 1962). A thorough review of the problem can be found in Ollongren's work, and we summarize it briefly here. We suppose that the gravitational potential of a galaxy is time-independent and has an axis of symmetry. In a system of cylindrical coordinates R, θ, z, this potential is then a given function $U_g(R, z)$. We are interested in the motion of a star in such a potential. In particular we ask: what part of the 6-dimensional phase space $(R, \theta, z, \dot{R}, \dot{\theta}, \dot{z})$ will be filled by the trajectory of the star if we follow it for a very long time, corresponding to many revolutions within the galaxy?

Since the phase space is six-dimensional, there must exist five independent conservative integrals of the motion; that is, five independent functions

$$I_j(R, \theta, z, \dot{R}, \dot{\theta}, \dot{z}) \qquad (j=1 \text{ to } 5),$$

which are constant along any trajectory. Conversely, a trajectory in phase space is determined by the five equations

$$I_j=C_j \qquad (j=1 \text{ to } 5), \tag{1}$$

where the C_j are five constants. Each equation represents a hypersurface in the phase space, and the trajectory is the intersection of the five hypersurfaces.

But each integral I_j can be isolating or nonisolating (for definition, see Wintner 1947; Lynden-Bell 1962; Ollongren 1962). A nonisolating integral is such that the corresponding hypersurface consists of an infinity of sheets which usually fill the phase space densely, so that for practical purposes the condition $I_j=C_j$ does not give any information and is equivalent to no condition at all. Thus from the physical point of view (as distinct from the mathematical one), nonisolating integrals have no significance. For that reason, isolating integrals are usually called simply "integrals," and the nonisolating integrals are ignored.

In the present case, two isolating integrals are known:

$$I_1=U_g(R,z)+\tfrac{1}{2}(\dot{R}^2+R^2\dot{\theta}^2+\dot{z}^2), \tag{2}$$

$$I_2=R^2\dot{\theta}. \tag{3}$$

They are the total energy and the angular momentum per unit mass of the star around the z axis. It can be shown that two of the other integrals, for example I_4 and I_5, are generally nonisolating. The problem is then: what is the nature of the last integral, I_3?

For many years, it was assumed that I_3 is nonisolating (see, for example, Jeans 1915, 1919; Lindblad 1933; Smart 1938; Van der Pahlen 1947; Lindblad 1959), on the ground that no third integral expressible in analytical form like I_1 and I_2 had been discovered, despite many efforts. But this assumption, as has been often remarked, is in conflict with the observed distribution of stellar velocities near the sun; for it implies that the dispersion of velocities should be the same in the direction of the galactic center and in the direction perpendicular to the galactic plane, whereas the observed dispersions have approximately a 2:1 ratio. More recently, a number of galactic orbits have been computed numerically (Contopoulos 1958, 1963; Ollongren 1962). Quite unexpectedly, all these orbits behaved as if they had not 2, but 3 isolating integrals. As a result, there was some change of opinion on the subject. Attempts were made to prove theoretically the existence of a third integral (see Contopoulos 1963).

In the present paper, we approach the problem again by numerical computations; but, in order to have more freedom of experimentation, we forget momentarily the astronomical origin of the problem and consider it in its general form: does an axisymmetrical potential admit a third isolating integral of motion? Thus, we allow the potential U_g to be an arbitrary function of R and z, not necessarily representing an actual galactic potential.

2. REDUCTION TO A SIMPLER FORM

As is easily seen, if we introduce the function

$$U(R,z)=U_g(R,z)+C_2^2/2R^2, \tag{4}$$

* Present address: Institut d' Astrophysique, Paris.

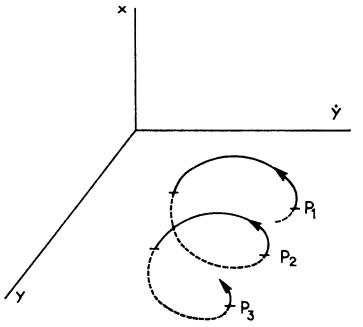

FIG. 1. Definition of the points P_i: $\dot{x}>0$, $x=0$.

where C_2 is the constant value of the angular momentum (3), the equations of motion in R and z become

$$R=-\partial U/\partial R, \qquad z=-\partial U/\partial z. \tag{5}$$

This shows that the problem considered is completely equivalent to the problem of the motion of a particle in a plane in an arbitrary potential U. We shall adopt from now on this new formulation and substitute x and y for R and z. The phase space $(x,y,\dot{x},\dot{y})$ has now four dimensions, and there must exist three independent conservative integrals of the motion. One of them is known and is isolating:

$$I_1=U(x,y)+\tfrac{1}{2}(\dot{x}^2+\dot{y}^2). \tag{6}$$

It is the total energy of the star divided by its mass, as before. There is no integral of angular momentum, because the potential U has no symmetry in general. It can be shown that one of the integrals, say I_3, is generally nonisolating, and the problem is now: what is the nature of the *second* integral I_2?

Because of the existence of the energy integral (6), it is sufficient to know three coordinates of the star in the phase space, such as: x, y, $\dot{y}$; the fourth coordinate $\dot{x}$ can then be obtained from

$$U(x,y)+\tfrac{1}{2}(\dot{x}^2+\dot{y}^2)=E, \tag{7}$$

if we know the energy E. Consequently, we can plot the trajectory in a three-dimensional space $(x,y,\dot{y})$ (see Fig. 1). The value of $\dot{x}^2$ found from (7) should be nonnegative, hence the condition

$$U(x,y)+\tfrac{1}{2}\dot{y}^2\leqslant E \tag{8}$$

which normally defines a bounded volume.

If there is no other isolating integral, the trajectory will fill the volume defined by (8), and we shall call it *ergodic*. If there is a second isolating integral, the trajectory will, instead, lie on a surface, whose equation is found by elimination of $\dot{x}$ between (7) and $I_2=C_2$.

Let us consider the successive intersections of the trajectory with the plane $x=0$, in the upward direction; that is, the successive points P_1, P_2, $\cdots$ of the trajectory which lie in the $(y,\dot{y})$ plane and satisfy

$$x=0, \quad \dot{x}>0. \tag{9}$$

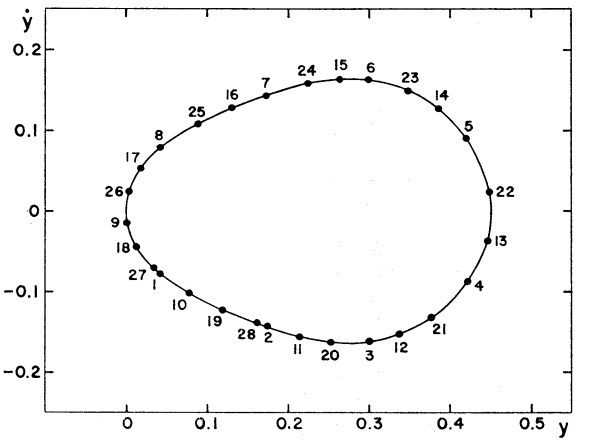

FIG. 3. A typical set of points P_i; $E=0.08333$.

If we follow the trajectory for an infinite time, there will be in general an infinite sequence of points P_i. If there is no second isolating integral, these points will fill an area, which is the intersection of the volume (8) with the plane $x=0$:

$$U(0,y)+\tfrac{1}{2}\dot{y}^2\leqslant E. \tag{10}$$

But if there is a second isolating integral, the points P_i will lie on a curve. Thus we get a simple criterion for the existence of the second integral: it is sufficient to compute a number of points P_i, plot them in the $(y,\dot{y})$ plane and see whether they lie on a curve or not. This method will be used in what follows.

The passage from a point P_i to the next one P_{i+1} can be considered as a *mapping*. This mapping is completely defined when the potential $U(x,y)$ and the energy E are given. [For, suppose that a point P_i is given. It defines y and $\dot{y}$; x is zero; and $\dot{x}$ is found from (7). Starting from

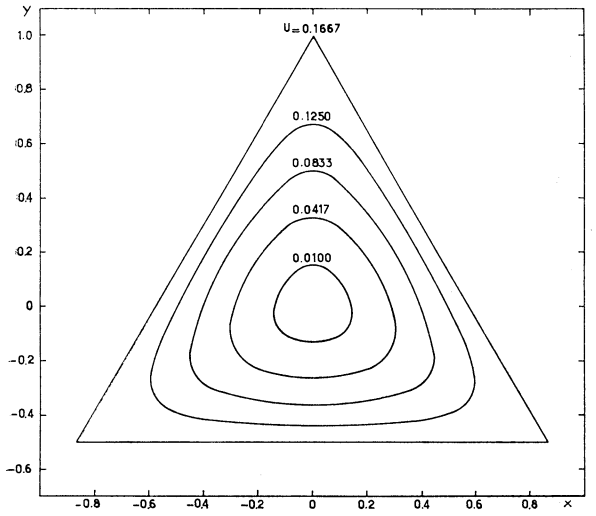

FIG. 2. Equipotential lines of (11).

these four initial values, the trajectory can be integrated to the next point satisfying (9), which is P_{i+1}.] It can also be shown that the mapping is area-preserving [see, e.g., Birkhoff (1927, p. 152); and see Moser (1962) for an important theorem concerning such mappings].

3. RESULTS

After some trials, the following potential was chosen for study:

$$U(x,y)=\tfrac{1}{2}(x^2+y^2+2x^2y-\tfrac{2}{3}y^3) \qquad (11)$$

because: (1) it is analytically simple; this makes the computation of the trajectory easy; (2) at the same time, it is sufficiently complicated to give trajectories which are far from trivial, as will be seen below. It seems probable that the potential (11) is a typical representative of the general case, and that nothing would be fundamentally changed by the addition of higher-order terms.

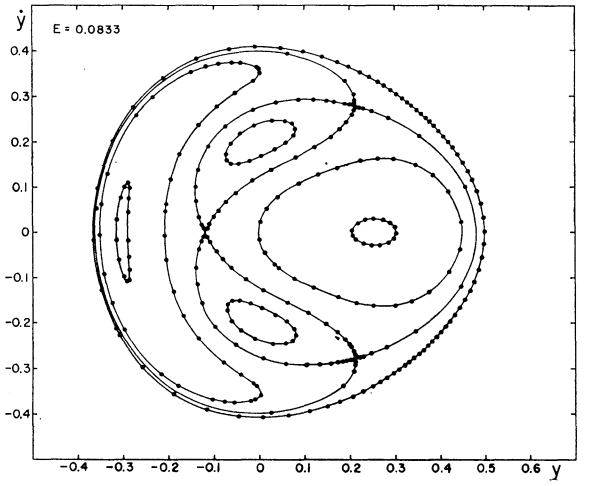

FIG. 4. Results for $E=0.08333$.

Figure 2 shows the equipotential lines. Near the center they tend to be circles; farther out they become elongated in three directions. The particular equipotential $U=\frac{1}{6}$ consists of three straight lines, forming an equilateral triangle.

A number of orbits were computed by numerical integration of the equations of motion:

$$\dot{x}=-\partial U/\partial x=-x-2xy,$$
$$\dot{y}=-\partial U/\partial y=-y-x^2+y^2. \qquad (12)$$

As a check, some of the orbits were computed independently by each of us, using different computers (CDC 1604 and IBM 7090) and different integration schemes (Adams and Runge–Kutta). The following results were obtained using the Runge–Kutta method; during the numerical integration the energy was observed to decrease very slightly ($<|0.00003|$ for 150 orbits).

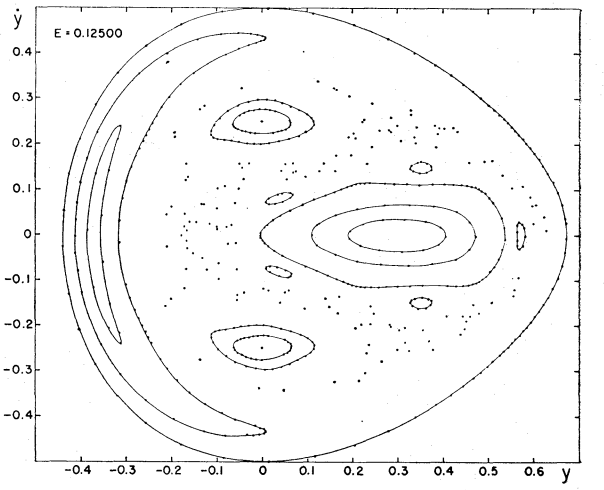

FIG. 5. Results for $E=0.12500$.

Figure 3 shows a set of points P_i for a typical trajectory. They seem to lie exactly on a curve. In fact, more points have been computed than those plotted here; after the 150th point there is still no perceptible deviation from a curve. It may be interesting to remark that the successive points P_1, P_2, $P_3\cdots$ (represented here by 1, 2, $3\cdots$) rotate regularly around the curve. The figure is topologically identical to one where the points P_i would lie on a circle of center O, the angle between OP_i and OP_{i+1} having a constant value α. This constant is not the same for different trajectories. In the case of Fig. 3, its approximate value is $\alpha=0.1143$ (taking one revolution as the unit). α is generally not rational, so that no point P_i will come back exactly on the initial point P_1, and the infinite set of the points P_i is dense everywhere on the curve. If α happens to be a rational number p/q, the point P_{q+1} will be identical with P_1 and the orbit is periodic.

Figure 4 shows the complete picture in the $(y,\dot{y})$ plane, for a given value of the energy: $E=\frac{1}{12}=0.08333$.

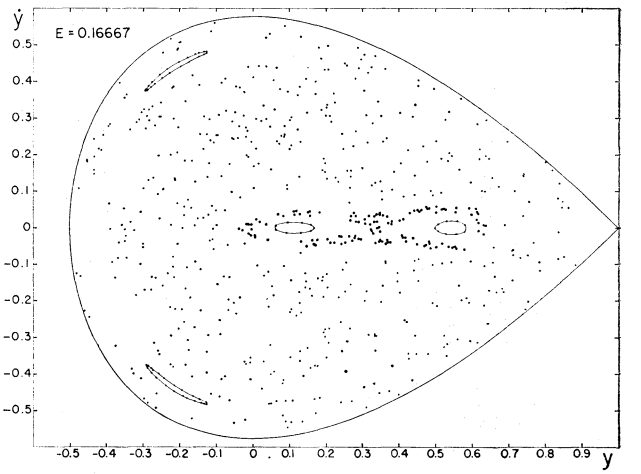

FIG. 6. Results for $E=0.16667$.

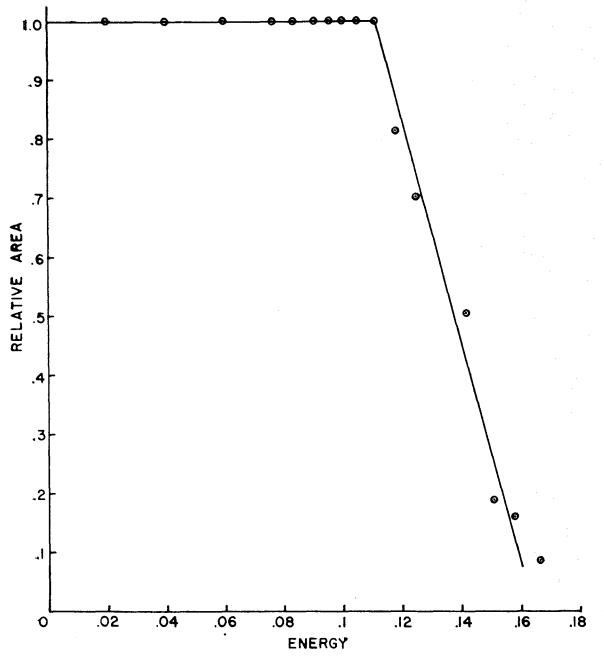

FIG. 7. Relative area covered by the curves as a function of energy, as computed by the method described in the text.

Each set of points linked by a curve corresponds to one computed trajectory. In fact, more trajectories and more points on each have been computed than shown on this picture. It appears that in every case, the points seem to lie exactly on a curve. These curves form a one-parameter family which fills completely the available area, defined by (10). (The boundary of this area is almost identical with the outer curve on Fig. 4.)

In the middle of the four small loops are four invariant points of the mapping (not represented on Fig. 4); they correspond to stable periodic orbits. The three intersections of curves are also invariant points, corresponding to unstable periodic orbits.

This picture seems, like the previous computations by Contopoulos (1958, 1963) and Ollongren (1962), convincing evidence of the existence of a second integral. But here comes the surprise. Figure 5 shows the same picture in the $(y,\dot{y})$ plane for a somewhat higher energy: $E=0.12500$. We still have a set of closed curves around each stable invariant point. But these curves no longer fill the whole area. All the isolated points on Fig. 5 correspond to one and the same trajectory, just as the points on one of the closed curves; but they behave in a completely different way. It is clearly impossible to draw any curve through them. They seem to be distributed at random, in an area left free between the closed curves. Most striking is the fact that this change of behavior seems to occur abruptly across some dividing line in the plane.

The picture is even more complicated than the above description would suggest. For example, the five little loops in the right of the diagram belong to the same trajectory; the successive points P_i jump from one loop to the next. Let us call this feature a *chain of islands*. Other such chains have been found in various parts of the diagram. The number q of the islands in a chain can apparently have any value. As a rule, the dimensions of the islands decrease very rapidly when q increases. Each chain is associated with a stable periodic orbit; the q islands surround the q points which correspond to that orbit. Note that each set of closed curves on Fig. 5 can be considered as a chain constituted by only one island; in both features no ergodic orbit seems to appear. The following properties are also suggested by our results:

(1) there is an infinite number of islands (and of chains);

(2) the set of all the islands is dense everywhere;

(3) but the islands do not cover the whole area since they become very small; there exists a "sea" between the islands and the ergodic trajectory is dense everywhere on the sea.

But, of course, mathematical proofs are needed to establish these points.

Figure 6 shows the situation for a still higher energy: $E=\frac{1}{6}=0.16667$. Again the picture changes drastically. All the isolated points correspond to one trajectory, and it is apparent that this "ergodic" trajectory covers almost the whole area. [The outer line on Fig. 6 is the limit given by (10).] Its random character is most strikingly seen when one plots the successive points; they jump from one part of the diagram to another without any apparent law. Two of the sets of closed curves of Fig. 5, those on the $\dot{y}$ axis, have now disappeared, presumably because their central invariant point has become unstable. The two other sets of closed curves have degenerated, each one into a chain of two small islands, successive points P_i jumping from one to the other. No other chain of islands has been found in Fig. 6; probably they still exist, but the dimensions of the islands are so small that finding them is difficult.

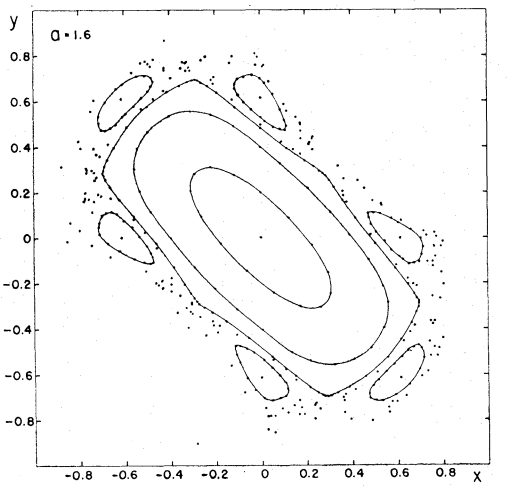

FIG. 8. The iterated mapping (14), for $a=1.6$.

The open circles in the middle of the diagram correspond to a trajectory of a new kind, intermediate between the closed curves and the ergodic behavior. They are approximately situated on an eight-shaped line, but with an important dispersion around it. The ultimate behavior of such an orbit is not known; perhaps the points will always remain in the vicinity of the same line, and fill an eight-shaped band; or perhaps they will after some time penetrate into the ergodic region. Some recent results, not shown here, are in favor of this last hypothesis.

A remarkable feature of Figs. 4 to 6 is the complete change in the picture over a moderate interval of the energy E. For $E=0.08333$, the area is completely covered with curves; for twice that value, the curves are almost completely replaced by an ergodic region. If, instead of the energy, one considers the amplitude of the motion indicated by the equipotential lines of Fig. 2, the change occurs on an even smaller interval.

In order to study this transition in more detail, we have computed, for a number of values of E, the proportion of the total allowable area in the $(y,\dot{y})$ plane which is covered by curves. The following method was used to decide whether a given point P_1 belongs to a curve or to an ergodic orbit. A second initial point P_1' was taken very close to P_1 (usually at a distance 10^{-7}). Then a number (usually 25) of successive transforms of both P_1 and P_1' were computed. Experience had shown previously that if P_1 and P_1' are in a region occupied by curves, the distance P_iP_i' increases only slowly, about linearly, with i; but if P_1 and P_1' are in the ergodic region, the distance P_iP_i' increases rapidly, roughly exponentially. The quantity

$$\mu = \sum_{i=1}^{i=25} (\text{distance } P_iP_i')^2 \qquad (13)$$

was computed, and the point P_1, as well as its transforms, were considered as belonging to the ergodic region if $\mu > \mu_c$, to a curve if $\mu < \mu_c$; μ_c is a chosen constant. The values found for μ covered a very wide range, from about 10^{-12} to 10^{+1}; the criterion seems to be very sensitive, and the exact value chosen for μ_c is not of great importance. Here $\mu_c \cong 10^{-4}$.

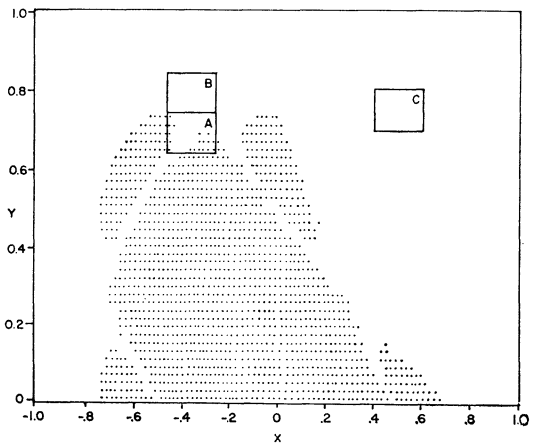

FIG. 9. All nonergodic points in upper half of Fig. 8; Grid size = 0.02.

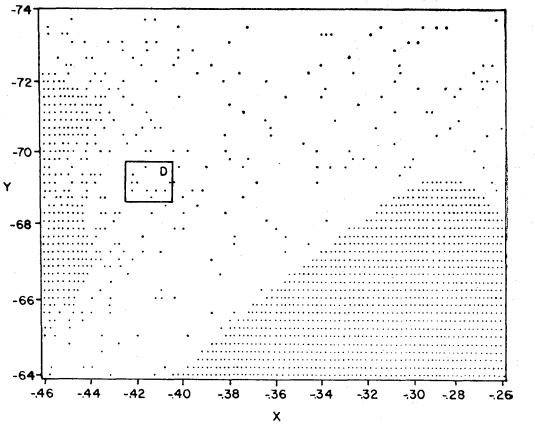

FIG. 10. Enlargement of area A; Grid size = 0.002.

Figure 7 shows the results. Up to a critical energy (about $E=0.11$) the curves cover the whole area; there is no ergodic orbit. For higher energies the area covered by curves shrinks very rapidly. Thus the situation could be very roughly described by saying that the second integral exists for orbits below a "critical energy," and does not exist for orbits above that energy.

$E=\frac{1}{6}$ is the *energy of escape* in the potential (11); for $E>\frac{1}{6}$, the equipotential lines open and the star can eventually escape to infinity, if the orbit is ergodic. The area in the $(y,\dot{y})$ plane becomes infinite and the relative area represented on Fig. 7 ceases to have meaning. No obvious connection exists between the critical energy and the energy of escape; in the present case the critical energy is less than the energy of escape. But results from computations with $U=\frac{1}{2}(x^2+y^2-x^2y^2)$, not shown here, indicate the opposite situation, as do the results of computations by Ollongren (1962) with an approximation to the Galactic potential. However, such a potential, derived from an actual three-dimensional potential, is dependent on the angular momentum assumed; so that more computations for other values of the angular momentum and higher energies are needed to establish the prevalence of the third integral in the Galaxy.

4. STUDY OF A MAPPING

It has been remarked above that the whole problem can be reduced to the study of a plane mapping. As was suggested to us by Dr. Kruskal, one can then define an area-preserving mapping and study it directly, thus by-passing the lengthy integration of orbits. The advantage of this method is that the computation is much simpler and much faster (by a factor 1000 approximately), so that more examples and more points can be computed. The disadvantage is that we are now quite

far from the initial astronomical problem. Also, it is not obvious that an arbitrary area-preserving mapping corresponds to a possible dynamical situation. For these reasons, we give only a short account of the experiments made. The following mapping was studied:

$$X_{i+1}=X_i+a(Y_i-Y_i^3),$$
$$Y_{i+1}=Y_i-a(X_{i+1}-X_{i+1}^3), \qquad (14)$$

where a is a constant. The coordinates of P_i are named here X_i and Y_i.

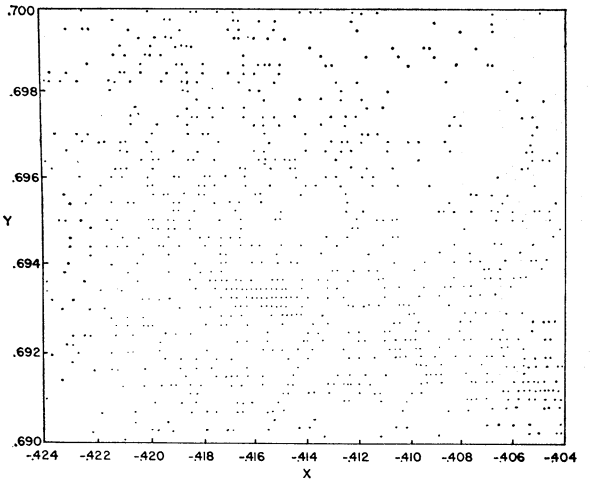

FIG. 11. Enlargement of area D; Grid size=0.0002.

Figure 8 shows the results for $a=1.6$. Each set of points linked by a curve is the set of the successive transforms of an initial point P_1 under the iterated mapping (14). The isolated points are also the successive transforms of a single initial point. The picture is quite similar to the right part of Fig. 5. There is a central region occupied by a set of simple closed curves which surround the stable invariant point $X=Y=0$; a chain of six islands (instead of five); and an outer "ergodic" region. Other chains of islands have been found here too. This similarity suggests that the problem of the area-preserving mapping is really identical with the dynamical problem of the third integral.

Up to 10^5 points have been computed for some of the curves, without any detectable deviation.

Figure 9 represents the upper half of Fig. 8, and Figs. 9–12 were produced in the following manner: initial points were chosen on a grid size indicated in the figure, throughout the whole area of the figure, and 1000 successive iterations of each initial point were computed. Experience has shown that iterations of points which produce an "ergodic orbit" are eventually mapped to infinity; furthermore, this divergence is quite rapid, due to the cubic terms in (14). Thus in Fig. 9 for example, if all 1000 points remained in the vicinity of the origin (this being practically expressed by $X^2+Y^2<100$) the position of the initial point was marked with a dot; otherwise, the position was left blank. The result is a replica of Fig. 8, the only difference being that Fig. 9 shows, to the scale of the grid, *all* initial points whose successive iterations lie on closed curves. Note that it is somewhat distorted, because the vertical and horizontal scales are not equal.

In order to investigate the mapping on a finer scale, we subdivide Fig. 9 into areas A, B, and C. Area A, ten times enlarged, is shown in Fig. 10. The most striking feature is the apparition of a multitude of small islands and tiny details, distributed in a random fashion. It can be remarked also that the boundary of the central region seems very sharp, whereas the boundary of the large island (on the left) is rather fuzzy. Area D of Fig. 10 was again enlarged ten times; see Fig. 11. Again a host of new details emerge. It seems very likely that this would go on indefinitely; with more magnification more details would appear, without end. These results support the hypotheses made above, namely, that there is an infinite number of islands and that their set is dense everywhere.

Area B, which is farther from the center, is represented on Fig. 12. The density of the islands is much smaller than in area A. Also a strong density gradient is apparent in the vertical direction. Area C, still farther out, was found to contain no dots at all to a grid size of 0.002 and therefore is not represented. Thus the density

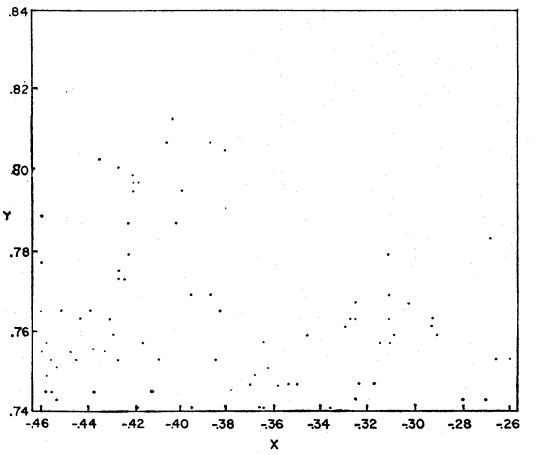

FIG. 12. Enlargement of area B; Grid size=0.002.

of the islands seems to decrease very rapidly as the distance from the central region increases.

5. CONCLUSIONS

We return now to the original three-dimensional problem. The above experiments indicate that the behavior of the orbits is in general quite complicated, and there seems to be no hope of a simple general answer, such as: (a) the third isolating integral always exists; or (b) the third isolating integral does not exist.

The true situation can perhaps be summarized as follows. Consider a given potential, and orbits with given angular momentum and energy. If the energy is small, it seems that a third isolating integral always exists. Perhaps it is only a quasi-integral; but then, to the accuracy of the computers, it is as good as a true integral. If the energy is higher than the critical energy, there are an infinite number of separated regions in the phase space where such a third integral still seems to exist. The space left free between these regions is the "ergodic region" where the third integral is nonisolating. If the energy is further increased, the proportion of allowable phase occupied by this ergodic region increases very rapidly and tends to be the whole space.

A number of questions are raised, for example: are the curves found here exactly or only approximately invariant? What is the topological nature of the set of all the islands? Is it possible to compute the curves directly from the potential, without integrating all the orbits? The ultimate answer to such questions should rest on rigorous mathematical proofs, not on numerical experiments; but the mathematical approach to the problem does not seem too easy.

Finally, it should be mentioned that the problem considered here belongs to the general family of the dynamical systems with two degrees of freedom, and thus is a close relative of the famous restricted three-body problem. Although we cannot attempt it here, a comparison of the two problems would certainly be most fruitful.

ACKNOWLEDGMENTS

Our thanks go to Drs. G. Contopoulos, H. C. van de Hulst, M. Kruskal, J. Moser, and M. Schwarzschild, for many stimulating discussions. One of us (M. Hénon) wants also to thank Princeton University for a one-year stay, during which this work was done; the other was supported by a William Charles Peyton Fellowship during this year.

REFERENCES

Barbanis, B. 1962, *Z. Astrophys.* **56**, 56.
Birkhoff, G. 1927, *Dynamical Systems* (American Mathematical Society, New York).
Contopoulos, G. 1957, *Stockholms Obs. Ann.* **19**, No. 10.
——. 1958, *ibid.* **20**, No. 5.
——. 1960, *Z. Astrophys.* **49**, 273.
——. 1963, *Astron. J.* **68**, 1.
Jeans, J. H. 1915, *Monthly Notices Roy. Astron. Soc.* **76**, 81.
——. 1919, *Problems of Cosmogony and Stellar Dynamics* (Cambridge University Press, New York), p. 233.
Lindblad, B. 1933, *Handbuch der Astrophysik* (Springer-Verlag, Berlin), Vol. V/2, p. 1038.
——. 1959, *Handbuch der Physik* (Springer-Verlag, Berlin), Vol. **53**, p. 28.
Lynden-Bell, D. 1962, *Monthly Notices Roy. Astron. Soc.* **124**, 1.
Moser, J. 1962, *Nachr. Akad. Wiss. Göotingen, Math. Phys. Kl.*, 1.
Ollongren, A. 1962, *Bull. Astron. Inst. Neth.* **16**, 241.
Smart, W. M. 1938, *Stellar Dynamics* (Cambridge University Press, New York), p. 338.
van de Hulst, H. C. 1962, *Bull. Astron. Inst. Neth.* **16**, 235.
——. 1963 (to be published).
van der Pahlen, E. 1947, *Einführung in die Dynamik von Sternsystemen* (Verlag Birkhäuser, Basel), p. 61.
Wintner, A. 1947, *The Analytical Foundations of Celestial Mechanics* (Princeton University Press, Princeton, New Jersey), p. 96.