JOURNAL OF COMPUTATIONAL PHYSICS 8, 449–463 (1971)

Experimental Studies of the Numerical Stability of the Gravitational n-Body Problem

R. H. MILLER*

Department of Astronomy and Astrophysics, Committee on Information Sciences, Institute for Computer Research, The University of Chicago, Chicago, Illinois 60637

Received March 23, 1971

The gravitational n-body problem is remarkably unstable numerically, in spite of the fact that near constancy of the energy and angular momentum makes it appear that the calculation should be reliable. In an attempt to understand why this should be so, the properties of numerical solutions were explored through a perturbation technique in which equations of motion for the differences between two systems are integrated numerically. Results of a series of experiments show that the differences between computed systems and exact solutions to the differential equations tend to lie near the hypersurface on which the first integrals are conserved and preferentially along the trajectory through the Γ-space. The numerical instability remains even when conservation of the first integrals is improved by several orders of magnitude through a partial iterative refinement method. Implications of these results concerning the utility of computations of the gravitational n-body problem are discussed. It may be possible to treat numerical effects as a kind of relaxation process, in which the relaxation time is identified with the characteristic e-folding time of error growth.

I. INTRODUCTION

The gravitational n-body problem has a long and distinguished history. The case $n = 2$ is the well-known Kepler problem, while $n = 3$ is the famous "three-body problem." Larger values of n may be expected to lead to even more intractible problems. Computer experiments to study the evolution from some (possibly arbitrary but otherwise well defined) initial state as an initial-value problem seem to promise a useful approach. This has been the case with problems in celestial mechanics and with the orbits of artificial satellites. The results, when the method is applied to stellar dynamical problems (simulation of star clusters, galaxies, etc.) have not been nearly as satisfactory.

* Visiting Astronomer, Kitt Peak National Observatory, which is operated by the Association of Universities for Research in Astronomy, Inc., under contract with the National Science Foundation.

Among the various kinds of simulations of star clusters, that which looks most promising—an attempt to integrate the n-body equations of motion as precisely as computational methods allow—has led to unusual difficulties. These are more of numerical character than due to inadequate sampling of the parameter space represented by the choice of initial conditions. The first integrals of motion (energy, angular momentum, etc.) are well conserved, but other quantities are remarkably difficult to compute reliably [1]. This led to suggestions of special methods to test the reliability of computed results [2]. The most dramatic confirmation of the difficulties is contained in Lecar's comparative study of calculations carried out by different workers from an identical set of initial conditions [3]. Lecar's comparison is dramatic because it shows that even simple quantities which one might expect to be reliably computed are surprisingly unreliable. Not only do different numbers of particles escape from the clusters computed by various investigators, but even the identity of the escaping particles is different. Even so simple a parameter as the radius of a sphere containing half the particles varied by as much as a factor of two among the various calculations. Similar difficulties were encountered in the intercomparison of other test quantities. More recent studies indicate that the same program deck, compiled and executed on different computers, shows similar differences in detailed comparisons [4]. The test of [1] is more stringent still: The same program in the same computer with sets of data differing only at the roundoff level soon leads to markedly differing systems. It is not as dramatic as Lecar's study because the tests for differences deal with subtler quantities.

Under the circumstances, it is difficult to distinguish real physical effects from artifacts of the computations. Attempts to study the properties of physical systems must be based on some understanding of purely computational effects. This is not a qualitative difference from other kinds of computations, but the gravitational n-body problem presents a situation in which the need for caution and understanding is more apparent and urgent than with most calculations.

Some investigations into the nature of the computational effects are reported in this paper. While the picture is nowhere near complete, some useful insights are afforded that permit a preliminary discussion of the kinds of physical quantities that may be reliably computed. More importantly, a time-scale characteristic of the calculation emerges such that physical quantities that develop more slowly than this time scale may not be reliably computed. This finally leads to a discussion of the probable utility of gravitational n-body calculations. The arguments are largely based on experimental studies of the computations themselves.

In the studies reported earlier [1], the numerical stability of gravitational n-body calculations, considered as an initial-value problem, was investigated experimentally. The experimental technique made use of two systems, initially very similar, each integrated as a distinct problem. Calling the coordinates and momenta of

the two systems $q_i^{(2)}$, $q_i^{(1)}$, $p_i^{(2)}$, $p_i^{(1)}$ respectively ($i = 1, 2, \ldots, 3n$ for n particles), a normed separation

$$\Delta^2 = \sum_i \left\{(q_i^{(2)} - q_i^{(1)})^2 + \frac{T^2}{m_i^2}(p_i^{(2)} - p_i^{(1)})^2\right\} \tag{1}$$

was computed and served as a measure of the amount by which the two systems diverged. The separation, Δ, showed a characteristic exponential growth as a function of time, with superposed large "spikes" [1, 2]. The rapid separation of the representative points in the phase space was argued to be characteristic of the physics of the problem as expressed in the form of the differential equations and not to be peculiar to numerical representations in computers.

This problem is further discussed in this paper. An equation of motion for the difference vector between the two systems is obtained. The equation is valid in the sense of a perturbation and looks as if it might possess exponentially growing solutions. Such exponentially growing solutions imply instability. However, the coefficients in the equation vary with the time as the reference system evolves in a manner which makes it difficult to prove whether or not there are exponentially growing solutions. In addition to the development of the perturbation equations, further experimental results are presented which show the manner in which numerical effects interact with the "unstable" system represented by the differential equations in computer experiments.

A method of partial iterative refinements that keeps the first integrals of motion very nearly constant is described in a companion paper [5]. The arguments of that paper are not based on comparison of two complete calculations as are those of this paper; those arguments only require comparison of the present computed values of the first integrals with their original values. However, the partial iterative refinement method was developed to investigate the effect of tight control of the first integrals on the numerical stability of the gravitational n-body calculation; the conclusions of this paper are strengthened by the results of studies using the method of partial iterative refinements. Although the gravitational n-body calculation is so "unstable" that the method is not sufficient to stabilize it, the method of partial iterative refinements should be useful with other kinds of calculations, where it might suffice to stabilize an otherwise unstable numerical calculation. Partial iterative refinements are principally useful with the gravitational n-body problem as a diagnostic tool.

II. Formulation

The problem may be considered in the $6n$-dimensional Γ-space. Let $p^{(1)}$, $q^{(1)}$ and $p^{(2)}$, $q^{(2)}$ represent two points in the Γ-space. These are not to be regarded as

representatives of ensembles. Let $\delta p = p^{(2)} - p^{(1)}$, $\delta q = q^{(2)} - q^{(1)}$ be "small" quantities. The motion is governed by a Hamiltonian $H(p, q)$, through the usual canonical equations of motion. In the gravitational n-body problem, the Hamiltonian H, which is a function of position in the Γ-space, is smooth almost everywhere. Equations of motion for δp and δq can be constructed:

$$\begin{aligned}\delta\dot{q} &= \dot{q}^{(2)} - \dot{q}^{(1)} = \frac{\partial H}{\partial p}\bigg|_{p,q^{(2)}} - \frac{\partial H}{\partial p}\bigg|_{p,q^{(1)}} \\ &= \frac{\partial H}{\partial p} + \frac{\partial^2 H}{\partial p^2}\,\delta p + \frac{\partial^2 H}{\partial p\,\partial q}\,\delta q + \cdots - \frac{\partial H}{\partial p}, \end{aligned} \tag{2}$$

$$\begin{aligned}\delta\dot{p} &= \dot{p}^{(2)} - \dot{p}^{(1)} = -\frac{\partial H}{\partial q}\bigg|_{p,q^{(2)}} + \frac{\partial H}{\partial q}\bigg|_{p,q^{(1)}} \\ &= -\frac{\partial H}{\partial q} - \frac{\partial^2 H}{\partial q\,\partial p}\,\delta p - \frac{\partial^2 H}{\partial q^2}\,\delta q - \cdots + \frac{\partial H}{\partial q}. \end{aligned} \tag{3}$$

The derivatives in the second lines of Eqs. (2) and (3) are to be evaluated at the point $p^{(1)}$, $q^{(1)}$. If all terms were retained, these equations would be exact to the radius of convergence of the Taylor series expansions, or, equivalently, as long as δp, δq do not extend to a singularity of H. The linearized form of these equations is:

$$\delta\dot{q} = \frac{\partial^2 H}{\partial q\,\partial p}\,\delta q + \frac{\partial^2 H}{\partial p^2}\,\delta p = \frac{\partial\dot{q}}{\partial q}\,\delta q + \frac{\partial\dot{q}}{\partial p}\,\delta p \tag{4}$$

and

$$\delta\dot{p} = -\frac{\partial^2 H}{\partial q^2}\,\delta q - \frac{\partial^2 H}{\partial p\,\partial q}\,\delta p = \frac{\partial\dot{p}}{\partial q}\,\delta q + \frac{\partial\dot{p}}{\partial p}\,\delta p, \tag{5}$$

where, again, the derivatives are to be evaluated at $p^{(1)}$, $q^{(1)}$, which is the position of the first system in the phase space— the actual values to be inserted depend on the details of evolution of the unperturbed system. The second form, obtained by inserting the canonical equations, shows that the equations are just what one would expect them to be.

Two reductions of this system of equations are useful. The first is simply notational: let ξ be written as a 2-component vector

$$\begin{pmatrix}\delta q \\ \delta p\end{pmatrix}$$

(it actually has $6n$ components). Then the perturbation equations can be written in the usual form for a (time-dependent) homogeneous system of equations:

$$\dot{\xi} = \mathbf{M}\xi, \tag{6}$$

where the elements of the matrix $\mathbf{M} = \mathbf{M}(t)$ can be read off from Eqs. (4) and (5). The time dependence enters through the unperturbed motion of one of the two systems being considered. The second reduction comes from noting that, in a canonical coordinate system based on cartesian coordinates, with forces that are not velocity dependent (the usual formulation for a computer calculation), the matrix $\mathbf{M}$ reduces to $3n \times 3n$ blocks:

$$\mathbf{M} = \begin{pmatrix} 0 & \frac{1}{m}\mathbf{I} \\ (\text{grad } \boldsymbol{F}) & 0 \end{pmatrix} \tag{7}$$

The $\mathbf{I}$ in the upper right-hand corner is a $3n \times 3n$ identity matrix if all particles have unit mass. Generalization to various particle masses represents no complication of principle, but obscures the notation, so the remainder of the discussion will be based on the equal-mass case. The element in the lower left-hand corner is the $3n \times 3n$ gradient of the forces. This term, readily derived from the total potential energy term in the Hamiltonian (with a slightly changed notation: let $x_i^{(\alpha)}$ represent the i-th component of the position vector of particle α, $\alpha = 1, 2, 3, \ldots n$, $i = 1, 2, 3$), acts between individual particles:

$$\begin{aligned} (\text{grad } \boldsymbol{F})_{ij}^{(\alpha\beta)} &= \\ &= \frac{\partial}{\partial x_i^{(\alpha)}} \left\{ \frac{Gm^{(\alpha)}m^{(\beta)}(x_j^{(\beta)} - x_j^{(\alpha)})}{[(x_k^{(\beta)} - x_k^{(\alpha)})^2]^{3/2}} \right\} \\ &= \frac{Gm^{(\alpha)}m^{(\beta)}}{[(x_k^{(\beta)} - x_k^{(\alpha)})^2]^{3/2}} \left\{ \delta_{ij} - 3 \frac{(x_i^{(\beta)} - x_i^{(\alpha)})(x_j^{(\beta)} - x_j^{(\alpha)})}{[(x_l^{(\beta)} - x_l^{(\alpha)})^2]} \right\}. \end{aligned} \tag{8}$$

This term has the expected tensor character—for the inverse square-law force appropriate to the gravitational problem, it looks like the field of a dipole ([6]), leading to a kind of "polarization" when the perturbed systems are considered. Note that (grad $\boldsymbol{F}$) is symmetric under index (i, j) and particle (α, β) interchange. The $\alpha = \beta$ elements can be obtained by summing over $\beta \neq \alpha$. With equal particle masses, the terms $m^{(\alpha)}m^{(\beta)}$ can be suppressed.

In Eq. (6), the elements of $\mathbf{M}$ are given functions of the time, even though they are not *a priori* known but depend on the actual trajectory of the unperturbed system through Γ. The system could be integrated directly if $\mathbf{M}$ and $d\mathbf{M}/dt$ commuted, but they do not. The system can be generalized to consider all solutions at once by integrating the matrix equation

$$\dot{\mathbf{Y}} = \mathbf{MY}; \qquad \mathbf{Y}(0) = \mathbf{I}. \tag{9}$$

Because the trace of **M** is zero, the determinant of **Y** is constant. This, of course, is a restatement of the Liouville theorem. But the constancy of the determinant of **Y** only means that growing solutions are balanced by decreasing solutions in such a way that the product of all solutions stays constant.

In an attempt to treat a system like Eq. (6) or Eq. (9), the obvious thing to try is to diagonalize **M**, if it can be done. But the diagonalizing transformation must itself be time dependent, so the notion of diagonalization is not likely to be useful. Since the trace of **M** is zero, all eigenvalues must sum to zero (note that these are eigenvalues of the matrix **M**, and do not imply that Eq. (6) is an eigenvalue problem). Thus, any nonzero real parts must contain at least some that are positive. But the existence of some positive real parts need not imply exponentially growing solutions: because the mode-structure (the columns of the diagonalizing transformation) is time dependent, a mode with a positive real part may later have a negative real part.

If a system like Eq. (6) were not time dependent, it would have exponential solutions. Bounds for the admissible growth-rates of ξ might be established by inserting some kind of upper bound for elements of **M**. The exponentially growing solutions to the bounding equations are not likely to be useful bounds because they grow too rapidly. They must accommodate the large spikes of Refs. [1, 2].

The normal techniques for investigating the character of solutions to systems of equations like Eq. (6) or Eq. (9) do not seem to offer much help (see, e.g., Bellman [7]). The elements of **M** continue to vary with the time—they show no tendency to settle down to some static condition after a long time. Any periodicity is of the order of a Poincaré recurrence-time. Replacement of the (grad $\boldsymbol{F}$) part of **M** by averaged values will not help: (grad $\boldsymbol{F}$) is essentially $r_{\alpha\beta}^{-3}$ multiplied by a spherical harmonic of second order. The average over all directions is zero.[1] Because perturbation solutions must be terminated once the separation of the two systems becomes comparable with interparticle separations in one system, approximate solutions as $t \to \infty$ are of little interest.

An interesting feature that follows from the structure of **M** (Eq. 7) is that one-dimensional systems and noninteracting hard-sphere systems, which have (grad $\boldsymbol{F}$) = 0 except when two "particles" collide, yield a system that does not have exponentially growing perturbation solutions. These systems are stable.

An initial displacement along the trajectory is outstanding among the solutions that do not grow exponentially: $\delta q = \dot{q}\,\delta t$, $\delta p = \dot{p}\,\delta t$ for some small δt. The second form of Eqs. (4) and (5) reduce in this case to the usual expressions for $d(\delta q)/dt$ and $d(\delta p)/dt$ since δq and δp are functions of q and p but not explicitly of

[1] Technically, these terms cannot be averaged over all directions because of the constraints placed on the set of permissible interparticle separation vectors by the fact that there are many more such vectors than there are particles. The particle positions may be taken as independent for the averaging, however

time. Thus, in the mathematical solution, a perturbation initially along the trajectory will forever remain along the trajectory (if first-order expressions are valid).

Finally, the form taken by the expressions for the conservation of the usual first integrals of motion in this language is of some interest. Conservation of total momentum is typical: $P_i = \sum_\alpha p_i^{(\alpha)}$. The difference of total momentum of the two systems is the scalar product of the ($6n$-dimensional) gradient of total momentum and the displacement vector

$$\begin{pmatrix} \delta q \\ \delta p \end{pmatrix}.$$

This allows for the variation of the perpendicular separation of the surfaces on which the integrals are constant—where the surfaces are farther apart, larger displacements in the direction of the gradient are allowed. The spatial part of (grad P_i) is zero, and the momentum components are unity. From these, it is readily verified that (grad P_i) is orthogonal to the normal trajectory: a displacement along the trajectory is given by

$$\begin{pmatrix} p/m \\ \boldsymbol{F} \end{pmatrix} \delta t,$$

and the scalar product is $\delta t \sum_\alpha F_i^{(\alpha)} = 0$. With an arbitrary initial displacement,

$$\begin{pmatrix} \delta q \\ \delta p \end{pmatrix},$$

the initial momentum difference is

$$(0, 1)\begin{pmatrix} \delta q \\ \delta p \end{pmatrix} = \sum_\alpha \delta p_i^{(\alpha)},$$

but after a (small) time δt, the change in momentum difference is given by

$$(0, 1)\begin{pmatrix} \dfrac{\delta p}{m} \\ (\text{grad } \boldsymbol{F})\, \delta q \end{pmatrix} \delta t = \delta t \sum_\alpha \sum_{j,\beta} (\text{grad } \boldsymbol{F})_{ij}^{(\alpha\beta)}\, \delta q_j^{(\beta)} = \delta t \sum_\alpha \delta F_i^{(\alpha)} = 0. \quad (10)$$

But the result is just the variation in the summed forces, which is zero because the forces are taken as internally generated, with sums of zero. Expressions for the remaining integrals carry through similarly.

III. Numerical Experiments

The advantages of the present formulation, apart from the insight afforded, are principally numerical. Rather than integrating two systems independently, and differencing tham, as implied by Eq. (1), the differential equation for the difference vector is separately integrated. Normally, this will not have a computational speed advantage: the integration of the differential equations for the difference vector requires $(3n)^2$ operations per integration step, while separate integration of a second system requires about the same number. The quantities in (grad $\boldsymbol{F}$) may be computed at little extra cost as $\boldsymbol{F}$ itself is being computed. As more sophisticated integration schemes are used for the direct integration of the unperturbed system, the speed advantage will definitely tip in favor of separate integrations whose results are differenced. Typically, the perturbation equations are not integrated as carefully as are the n-body equations governing the unperturbed system.

When differencing two integrated systems, the differences obtained may not be much above the roundoff noise level at the start of a calculation—while with the perturbation formulation, the elements of the difference vector should be well above the roundoff level except in cases of extreme accidental cancellation. As a by-product of the better numerical accuracy afforded by the perturbation method, it is easier to introduce controlled initial difference vectors between the two systems, with confidence that the difference vector generated is the one actually being used rather than some other difference vector heavily conditioned by roundoff. This permits the use of initial difference vectors lying along the initial trajectory (of the unperturbed system), along the initial (grad E), and other interesting special cases.

The perturbation formulation shows that the growth of the difference vector is essentially independent of the accuracy with which the trajectory of the unperturbed system is known. The perturbation calculation is valid only if the region between the points representing the two systems is well behaved, a condition that can be assured only as long as the difference vector is much "smaller" than any difference vector that might be constructed from the unperturbed system by permuting two particles, for example. If this condition is met, the changes in (grad $\boldsymbol{F}$) in going from one system point to the other will induce only second-order changes in the growth of the difference vector. This conclusion has been confirmed by some experimental studies reported later (Section IV).

(a) Experiments Run

The experiments were all run using 32 particles. Initial conditions were generated approximately according to the virial theorem using pseudo-random numbers. Initial perturbation vectors, $\xi(0)$ were used that were arbitrary (only $x_1^{(1)} \neq 0$), were along the starting trajectory, or lay in the integral hypersurface but were orthogonal to the initial trajectory. The initial value of Δ^2 [Eq. (1)] was 10^{-30},

while mean near interparticle separations corresponded to $\Delta^2 \approx 1/10$. Integrations were carried out with or without a partial iterative refinement procedure that keeps the first integrals of motion tightly constrained to their initial values. Finally, each of these conditions was run with three different sets of initial conditions, to try to eliminate effects peculiar to a particular set of starting conditions.

The experiments consisted of monitoring Δ^2 and the direction cosines of ξ along the trajectory and along (grad E) as the system evolved. (Current values of the vector along the trajectory and that along (grad E) were used—not the initial values.) The direction cosines were computed in the obvious way from the scalar products of the vectors involved:

$$\cos \theta_E = (\xi, \text{grad } E)/(\sqrt{(\xi, \xi)}\ \sqrt{(\text{grad } E, \text{grad } E)}), \tag{11}$$

and show the fractional projection of ξ in the direction of interest without the confusion due to changing magnitudes of $\Delta^2 = (\xi, \xi)$. If the direction cosines behave randomly (i.e., if ξ takes on random directions) in $6n$-dimensional space, the expectation of $\cos \theta$ is zero and its variance is $1/(6n)$, or, for $n = 32$, the standard deviation of $\cos \theta$ should be about 0.07. For all the scalar products, T and m_i of Eq. (1) were taken as 1. As mentioned earlier, initial perturbation vectors $\xi(0)$ were chosen in certain interesting directions such that $\Delta^2(0) \approx 10^{-30}$. Typical calculations ran until Δ^2 was around 10^{-10}, but the summaries on which conclusions were based mostly stopped when Δ^2 reached 10^{-20}. The value of Δ^2 that might be reached by interchanging two particles that lay near each other might be about 1, so all experiments were restricted to perturbations enough smaller than the critical system dimensions that the first-order approximations should be valid.

No experiments have been tried in which the full matrix equation (9) was integrated. Part of the reason for this is the mere logistical problem of interpreting the behavior of a 192×192 matrix (or even the smaller matrices that result with fewer particles).

(b) Experimental Results

The results obtained were, in their overall properties, essentially identical with those reported earlier [1, 2]. Plots of $\ln \Delta^2$ vs time are essentially linear with large superimposed spikes in which Δ^2 may temporarily increase by factors of 10^4–10^5, only to recover. Because these plots are so similar to those shown in Refs. [1] and [2], they are not reproduced here. The e-folding time of the underlying exponential (in Δ^2) is about 1/10 of a crossing time (the time a "typical" particle requires to go a distance equal to the "diameter" of the cluster—it is a rather crude concept

usually defined from the virial theorem). The dependence of the *e*-folding time on particle number and integration step size were reported earlier [1] and were not further studied in the present investigation. The value obtained for the *e*-folding time was essentially independent of the direction of the initial perturbation and whether the partial iterative refinement procedure was being used.

There were differences among the plots of $\ln \Delta^2$ vs t for different initial perturbation vectors $\xi(0)$, even with the same initial positions and velocities for all the particles of the unperturbed system. However, as in the earlier experiments, the large "spikes" were present at the same time and at about the same amplitude irrespective of the nature of the initial perturbation. These "spikes" appear during a close encounter, and indicate that the two systems enter the close encounter with different phases.

Calculations started from identical positions and velocities for all particles and with the same perturbation vector $\xi(0)$, but run with and without the partial iterative refinements (of the 10 first integrals) had essentially identical plots of $\ln \Delta^2$ vs t initially. Soon (after about $\frac{1}{2}$ crossing time) the plots deviated noticeably, getting farther and farther apart as the calculation proceeded. The basic slope, which indicates the exponential growth rate, remains the same, however.

The direction cosines are particularly interesting. They indicate the direction of the difference vector relative to (grad E) and to the trajectory. The direction cosines, like $\ln \Delta^2$, show large fluctuations. Several features stand out.

The direction cosine along the trajectory is usually well outside the $\pm.07$ limit appropriate to a randomly-oriented vector. It frequently attains a value very near 1 (occasionally in excess of .9999). It swings slowly from positive values to negative values. It is largest when the system is undergoing a close collision, as might be expected. Even when the system is not undergoing a particularly close collision, the direction cosine is often $\pm$(0.2 to 0.6), well beyond the limits set by the variance. Even with initial perturbations along (grad E) (the initial projection along the trajectory is then zero), large projections along the trajectory develop very soon (tenths of a crossing time or less). The alternating sign of the projection onto the trajectory also occurs with the initial perturbation along the trajectory. The effect of the partial iterative refinement procedure is about the same with this projection as with $\ln \Delta^2$—essentially no difference until the two systems diverge from each other.

The direction cosine along (grad E) is always significantly less than the $\pm.07$ standard deviation expected from random orientations. Its magnitude seldom exceeds 0.01. Again, it slowly alternates in sign, and has values of about the same magnitude for all initial perturbation vectors. The cases with the initial perturbation along the trajectory, which have no initial projection along (grad E), behave like the others—a component along (grad E) develops within 0.1 crossing time or less, and thereafter the direction cosine along (grad E) neither grows nor diminishes,

but rather seems to have some kind of noisy slow fluctuation. The rate of fluctuation is tied to the frequency of close encounters. A system undergoing frequent close encounters changes the sign of both direction cosines more frequently.

IV. Discussion

The basic straight line in the plot of $\ln \Delta^2$ vs t indicates that the numerical solutions to Eq. (6) are dominated by exponentially growing solutions even though it is difficult to prove whether or not there should be solutions of that character to the differential equations. This behavior continues as long as perturbation solutions are valid.

The numerical calculations do things that the mathematical solution indicates that they should not do. For example, they promptly develop components along the gradients to the integrals.

The direction cosines indicate that the difference vector tends to lie principally along the trajectory, although its components in all directions in the hypersurface of first integrals [5] are probably large. Its components orthogonal to that hypersurface tend to be significantly smaller than the components of a randomly-oriented vector of equal length. However, as the difference vector grows with the passage of time, the relative size of components lying in and orthogonal to the integral hypersurface remain about the same.

The iterative refinement procedure made little change in the gravitational n-body calculation, by the measure of the rate at which representative points of two computed systems separate in the phase space. The growth of $\ln \Delta^2$ with t is essentially the same with and without the partial iterative refinements. Calculations started from identical initial conditions, one run with the partial iterative refinement, the other without, yielded plots of $\ln \Delta^2$ vs t that were initially indistinguishable, but with slowly increasing difference. By half a crossing time, the two plots are noticeably different, and by one crossing time, they are quite different. The basic slopes continue to be the same, but the "spikes" and other such detailed features are quite different. By this time, the value of Δ^2 typically has grown to about 10^{-20}. This is the experimental result obtained using the partial iterative refinements as a diagnostic tool.

The essential similarity of plots of $\ln \Delta^2$ vs t for otherwise identical runs differing only in the use (or not) of the partial iterative refinement procedure is to be expected. As indicated earlier, the growth of the difference vector is expected to be independent of the accuracy with which the trajectory of the unperturbed system is known. Because the difference vector does not swim around as much in the directions perpendicular to the integral hypersurface as it does within the hypersurface, the restriction imposed on the 32-particle system by forcing the

integrals to be conserved is effectively much less than a reduction from 192 to 182 dimensions—it is more like a reduction from 183 to 182 dimensions. Such a small reduction would be very difficult to find experimentally.

A picture emerges in which the numerical difference vector distinguishes directions orthogonal to the integral hypersurface and further distinguishes the trajectory among those directions lying in the hypersurface. It prefers to go along the trajectory, but it makes only small excursions in directions orthogonal to the hypersurface, preferring to make up the remainder of its length in the hypersurface but in directions other than that of the trajectory. The volume in which it can swim about grows with the time, retaining its elongated disk-like shape. This volume is not subject to the Liouville theorem; this is another way in which the calculated system differs from the mathematical system.

At each integration step, roundoff and truncation errors introduce some noise which is an additional perturbation on the displacement vector. The "noise vector" may be thought of as being resolved into a basis which is constructed from the (instantaneous) eigenvectors of **M**. The components of growing eigenvectors grow—in this way, the fastest-growing eigenvector will soon dominate the system. With a changing eigenvector basis (due to the time dependence of **M**), the dominance will trade off from one eigenvector to another—but the numerical growth continues through this process. Whether the differential-equation system Eq. (6) has exponentially growing solutions or not, the numerical systems are experimentally found to have them.

Although the differential equations (6) possess as many decreasing solutions as growing ones, all the numerical examples found so far have only shown increasing solutions. This is reminiscent of the usual discussions of stability analysis, in which it is argued that just one unstable mode is sufficient to render the entire system unstable. The decomposition of roundoff and truncation noise is standard too. Note that it has not been asserted that this system is unstable. There are formal difficulties in comparing these results (and the prejudices stated here) to any of the criteria for stability. It seems likely, however, that the gravitational n-body calculation may be unstable in the sense that near any trajectory (in the Γ-space) there lies another trajectory that departs from the first in such a way that the "distance" between the two grows exponentially with the time until some kind of limiting process takes over.

What is surprising in the present case is the growth rate indicated by the e-folding times. If an upper bound to the growth rate is experimentally estimated from the rate necessary to accommodate the rising edge of the large "spikes," that growth rate is only about ten times the e-folding rate actually observed. The gravitational n-body calculation must be as close to unstable as any kind of calculation commonly undertaken. There are stable finite-difference approximations to it according to the usual criteria for the stability of initial-value calculations [2]. The gravitational

n-body calculation may therefore be of interest to computational theory as an example in which the interplay between the physical problem and numerical computation leads to an unusually large error growth rate.

Standish [8] has shown that the growth rate of the exponentially increasing difference vector can be reduced by modifying the force law to provide a near-cutoff to the interaction. He used a force derivable from a 2-body potential of the form:

$$\phi = (x^2 + y^2 + z^2 + a^2)^{-1/2}, \tag{12}$$

where x, y, and z represent the differences of the three coordinates for two particles in question. When the offset, a, is about the mean closest encounter distance or larger, a marked reduction in the growth rate results. There are still some "spikes," although they are not as large and lack the very steep edges. The ratio of slopes on the typical rising edge of a "spike" to the general growth in $\ln \Delta^2$ is about the same as noted above. Thus, both the experimental upper bound inferred from the slope of the rising edge of a "spike" and the experimental slope seem to be decreased by about the same amount, so the actual growth rate is still surprisingly close to the upper bound.

If two computed systems separate in this way, a computed system must depart from a physical system, or from a system that represents a correct integration of the differential equations, in the same way. The argument reduces to that used, for example, by Henrici [9] in studies of the stability of discrete-variable and finite-difference systems. The two systems of Section II are now the physical or mathematical system on the one hand and the computed system on the other.

The conclusions reached in this paper are independent of the particular canonical representation used in Eq. (7). Two systems cannot be prevented from diverging merely by looking at them from a different coordinate system. Stated differently, the physics cannot depend on the canonical representation used in deriving the results. This means that various ingenious transformations that might be introduced to regularize the treatment of close encounters, for example, cannot produce long-term improvements in the calculation.

The fact that the gravitational n-body calculation, even with the iterative refinement of the first integrals, departs exponentially from a physical trajectory, raises the question as to the value of such computations. Most attributes of the system must be determined through some kind of averaging process—either an ensemble average (many calculations, each started from a different, but somehow equivalent, initial condition) or a time average (same system, looked at at different times). Formally, the two averages are equivalent under the ergodic hypothesis. But it is not at all clear in what sense, if at all, the ergodic hypothesis applies to computed systems. Normally, time averages should extend over a time long compared to a characteristic time determined from the autocorrelation of the function being studied [10, 11]. The correlations die out in a fairly short time due to numerical

effects, according to the results of this paper. But does that assure the validity of time averages as approximations to ensemble averages? Or does it diminish its validity? The loss of even so fundamental an attribute of the differential-equation systems or physical systems as the Liouville theorem, which cannot hold for the numerical system, is quite disturbing.

One obvions way to get ensemble averages is through the use of Monte Carlo calculations. It should not be difficult to construct an ensemble of systems with the desired values of all ten first integrals. Why should an integration give better results than Monte Carlo? Presumably, the integration is somehow closer to the physics than a Monte Carlo random selection of systems. The integration can, somehow, take account of the many-particle correlations that dominate n-body systems. These are difficult to include in Monte Carlo calculations, especially since their nature is a priori unknown. In particular, one hopes that the integration, imperfect as it is, properly accounts for the many-particle terms. But this is at most a pious hope. Any effects in the many-body correlations, that develop slowly compared to the e-folding time over which the numerical calculation may be regarded as being "locally close" to some real physical system, must be lost to the integration as well. The integration can most safely be regarded as a kind of Monte Carlo calculation in which correlation effects that evolve in times on the order of the e-folding time of the growth of Δ^2 (or faster) are properly taken into account.

It is sometimes argued that the departure of a computed system from the physical system it is supposed to mimic adequately takes account of real physical effects that are not built into the problem as formulated. In the gravitational n-body problem, for example, numerical effects might mimic the effect of the galaxy on the star cluster being studied, or of other irregularities in the background force field. That seems to be a dangerous attitude; it is certainly safer to have all error terms understood, with such effects being intentionally introduced in a controlled and understood manner. It is also frequently argued that the departures of the computed system from the physical system occur in such a way that the computed system "heads for more probable regions of the phase space." The same objections hold again—how does one know that a computed system that cannot follow the track of a physical system will populate the phase space with the same or even a similar probability distribution?

Certainly the numerical difficulties make it almost impossible to conduct carefully controlled experiments. For example, it would be difficult to design a numerical experiment to test the effect of surrounding each particle with a hard sphere in addition to its gravitational force field, because differences between systems with and without the hard sphere could easily be no greater than would appear between two systems, neither of which had the hard sphere interaction, but were somehow slightly disturbed to force them down different evolutionary tracks.

This does not mean that gravitational n-body calculations are useless. Many

valuable things can be learned from them. But the studies must be directed toward clear-cut qualitative effects, rather than toward subtle quantitative effects. And, preferably, effects once found should be confirmed either by other kinds of numerical experiments or by theory and/or experiment. Internal consistency experiments along the lines suggested in [2] are also useful. The results reported in this paper, for example, were confirmed by seeing that the same general effect was present in every run starting from various initial conditions and using a variety of initial perturbations.

One way of interpreting these results is by considering a "numerical relaxation time," to be identified with the e-folding time of Δ^2. This is admittedly unphysical, since its origin is precisely in the difference between computed and physical systems. But it may be possible, by considering that the computed system has this one extra relaxation process that the physical system lacks, to draw valid inferences about physical systems from the behavior of computed systems. It is implicit, in the notion of treating numerical effects like relaxation processes, that the irreversible results of roundoff and truncation errors, as amplified by the character of the differential equations, indeed drive the system toward more probable parts of the phase space. Although this is uncertain, as pointed out earlier, it accords with the inherently optimistic nature of the numerical experimenter, and this hopes of retrieving something from his computations.

ACKNOWLEDGMENTS

It is a pleasure to thank J. Hofslund for his help with the experimental results. This work was partly carried out at the Kitt Peak National Observatory, while the writer was there as Consulting Astronomer. The use of the computational facilities at Kitt Peak is gratefully acknowledged. Support from the Shirley Farr Fund was very helpful. This work was also partially supported by the U. S. Atomic Energy Commission under Contract No. AT(11-1)-2094.

REFERENCES

1. R. H. MILLER, *Astrophys. J.* **140** (1964), 250.
2. R. H. MILLER, *J. Computational Phys.* **2** (1967), 1.
3. M. LECAR, *Bull. Astron.* (3), **3** (1968), 91.
4. A. HAYLI, *in* "Proceedings of IAU Colloquium 10 on the Gravitational n-body Problem" (M. Lecar, Ed.), Reidel, Dordrecht, Holland, to appear.
5. R. H. MILLER, companion paper, this issue of *J. Computational Phys.*
6. R. H. MILLER, *Astrophys. J.* **146** (1966), 831.
7. R. E. BELLMAN, "Stability Theory of Differential Equations," Dover, New York, 1969.
8. E. M. STANDISH, Thesis, Yale University, 1968.
9. P. HENRICI, "Error Propagation for Difference Methods," John Wiley and Sons, New York, 1963.
10. J. L. LEBOWITZ, J. K. PERCUS, AND L. VERLET, *Phys. Rev.* **153** (1967), 250.
11. R. ZWANZIG AND N. K. AILAWADI, *Phys. Rev.* **182** (1969), 280.