JOURNAL OF COMPUTATIONAL PHYSICS 2, 1-7 (1967)

An Experimental Method for Testing Numerical Stability in Initial-Value Problems[1]

RICHARD H. MILLER

Department of Astronomy, Institute for Computer Research and Committee on Information Sciences,

The University of Chicago, Chicago, Illinois 60637

ABSTRACT

The numerical stability of evolutionary calculations can be tested experimentally by running two calculations in parallel, starting from initial conditions that are very similar, and monitoring the differences between the systems. An integration of an n-body system with gravitational interactions provides an example that is useful for illustrative purposes. Some features of the method are described, followed by a discussion of some considerations in its application.

I. INTRODUCTION

An initial-value problem can describe the evolution of a physical system from some specified initial state according to a set of governing equations. The equations may or may not include boundary conditions, but there is no eigenvalue character resulting from forcing some specified terminal state. When an initial-value problem is run on a computer, the results may appear plausible even if they are unreliable because of some unrecognized numerical instability. Analytic methods may not be available for studying numerical stability; the techniques of numerical analysis are usually applied to the much simpler processes out of which larger calculations are built. Some aspects of a large problem are not tractable in the usual numerical analysis study; the "errors" in inputs to a certain numerical process inside a large calculation are not independent, for example. Initial-value problems do not share the inherent stability of some kinds of calculations: self-consistent methods, for example, are intrinsically stable if they possess unique solutions. Theoretical studies of the stability of initial value problems have been made, for example, by Lax and Richtmyer [1].

[1] This work was supported by the Atomic Energy Commission under Contract AT(11-1)—614.

581/2/1-1

A suitably designed experimental approach can provide a method for studying the numerical stability of large calculations even when a theoretical approach is not available. An experimental method that should be applicable to a wide range of initial value problems is described in this paper (Sec. 2). It provided a valuable test in a gravitational n-body integration (Sec. 3), where it disclosed a surprisingly unstable situation. This example illustrates many of the features of the experimental method. Generalizations to other problems are discussed, along with suggestions concerning the kinds of problems for which the method might be useful.

II. Parallel Calculations

Similar systems are constructed in the computer memory and evolve together, while the differences among them are monitored. To the extent that the systems are quite similar, subtle differences can be noted. The calculation may be run as long as the comparisons are meaningful. The same copy of the program can be designed to operate on all sets of data, one after another. If variable time-steps are used, some care is required to assure that all systems use the same values of the time-variable; this is most easily done by always using the shortest time-step.

"Similar" systems may be obtained by starting from one system, which is copied to form the other systems. When the copies are made, small perturbations can be introduced. Ideally, these perturbations should be of known character and larger than roundoff or integration errors in one time-step, but small enough that the systems are recognizable as having come from a common origin. The stage of the calculation at which the copies are made can be chosen for convenience; for example, if there is some special starting process, to build up a "history" for an integration method that makes use of previous states, copies might be made after switching over to the usual (running) integration routine. The extent to which the present state is not properly attainable from the "history" is then part of the perturbation.

This method is conceptually similar to "interval arithmetic." It differs by extending over the entire calculation and by not requiring well-behaved or monotonic input/output relationships.

Other means for checking initial-value calculations can easily be devised but the parallel calculation seems better. Time-reversible systems might be run forward for a while, then reversed and run back to the starting point where the final state can be compared with the initial state. Time-reversal is not as effective as the parallel calculations because of errors in reversing the system. The parallel calculation usually requires less modification to a running program for test purposes, and may be used with systems that are not time-reversible.

III. Example

This method was devised to check an n-body integration in stellar dynamics that failed to time-reverse [2]. In this case the state of a system is given by the location of its representative point in the 6n-dimensional phase space. Two systems are described by two points in the same space. If the systems are similar, the points are very near each other; a convenient measure of the departure of one system from the other is the separation of the phase points:

$$\Delta^2 = \sum_{i=1}^{n} \{(x_i^{(2)} - x_i^{(1)})^2 + T^2(u_i^{(2)} - u_i^{(1)})^2\}, \tag{1}$$

where T is a dimensional factor introduced to make all coordinates and velocities equivalent, the x's are configuration space coordinates, the u's are velocities, the superscripts refer to corresponding particles in the two systems, and the sum extends over all particles. (In gravitational problems with nonvarying masses, it is not necessary to distinguish between velocities and momenta.)

The problem, apart from the parallel calculation feature, was formulated according to the recipes of von Hoerner [3]. In this formulation, the integration is carried out for n particles of equal mass in Cartesian coordinates. Variable time steps based on the closest pair of particles are used to retain the accuracy of integration for close encounters while permitting coarser steps to be used when all the particles are well separated. The same time step is used for all particles. A second-order predictor–corrector integration method is used in which an attempt is made to retain reasonable accuracy without requiring re-evaluation of the forces more than once per integration step. The first ten integrals of the motion were used as controls, and were constant to within the same limits as von Hoerner reported [3]. A better integration procedure has since been devised [4]; the usefulness of the experimental method reported here is independent of the methods of handling the main problem.

The system was started from some set of initial conditions that was externally supplied. After enough first-order integration steps to build up the history required for the predictor–corrector method, the second copy of the system was made as part of the changeover to the normal second-order integration procedure. In copying, a perturbation was introduced as a small change in one velocity component of one particle.

Plots of $\ln \Delta$ against time are shown in Fig. 1 for two different initial conditions. The track shows a steady increase of $\ln \Delta$ with time, but has a series of spikes superposed. The linear trend of the minima between the spikes illustrates an exponential increase of the separation of phase points with time. The general trend may be followed from $\ln \Delta \approx -20$ to $\ln \Delta \approx 0$; the latter value corresponds

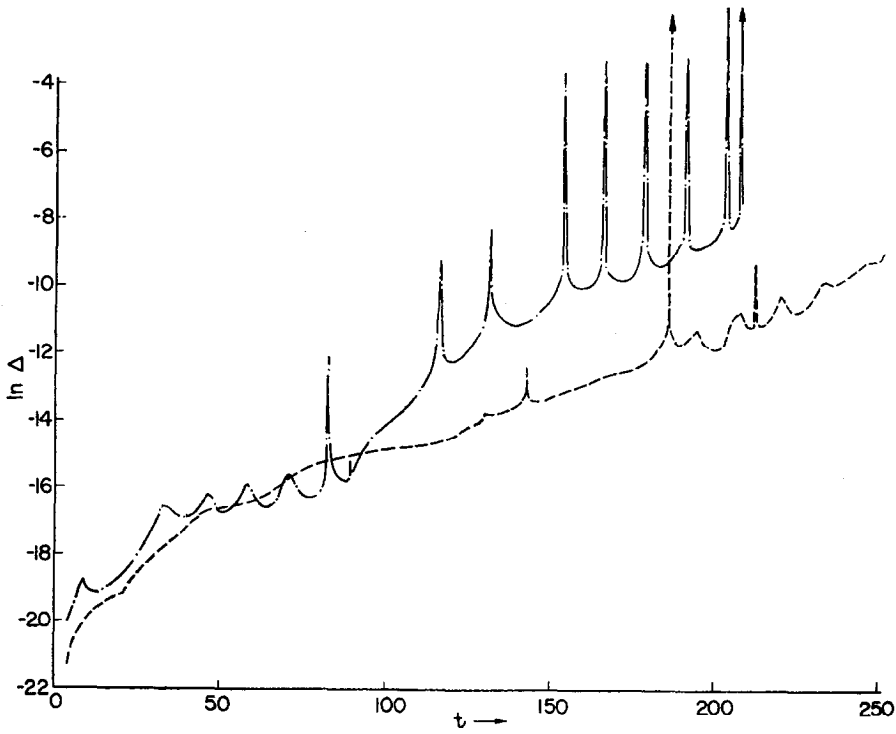

FIG. 1. Plot of $\ln \Delta$ vs. time for an 8-body system. Two different starting conditions are shown. The repeated equidistant spikes on the solid curve are caused by a long-lived binary pair that is destroyed by an encounter of one of its members with a third star at the last spike of that curve.

roughly to the mean particle separation or to the r.m.s. particle velocity. Beyond this, the system will blow up. This long track illustrates the sensitivity of parallel calculations. The spikes are of impressive amplitude—about a factor of 1000 or more over the prevailing level. The spikes are associated with close collisions; the spike results because the two systems enter the close collision at slightly different phases. The recovery afterward is especially dramatic.

Several experiments were tried with this system. The computation was stopped during a close encounter (on top of a spike) to determine which particles were participating. The calculation could then be repeatedly restarted from the same initial conditions, but with the perturbation placed on one of the participants in the close encounter or on a particle that did not participate. In this way it was found that information propagates through the system very rapidly.

Another experiment was to modify the step size in the integration procedure. The nature of the tracks was unchanged.

Parallel calculations with the perturbation set to zero keep $\Delta = 0$ for all time. This is characteristic of computer calculations. The size of the perturbation was varied; in general it is desirable to keep the perturbation as small as possible but if it is too small, Δ fluctuates about some small value for a substantial time before the reported trends set in.

This calculation raised many questions concerning the nature of the physical system that is supposedly represented. The principal point at issue is the sense in which the calculation represents a physical system. The calculation takes on a Monte Carlo character with the random element generated by roundoff and integration errors. With an exponential growth, the error terms will ultimately dominate, and the current state of the system cannot be causally related to the initial state in the manner that the physical system would be related to its initial state. Subsidiary questions include the rate of information transfer through the system, the factors that govern the growth of Δ, and so on. A result like this stimulates further work. In the gravitational case, it was possible to reproduce some of these effects analytically [5], although it is unlikely that the effect would have been discovered without the computer experiments. Fuller discussions of the physical implications of these results appear elsewhere [2], [5].

IV. Generalizations

Some guides to the application of the parallel calculation method to other problems emerged in the treatment of this example.

A. Measure of Departure

The quantity Δ that was used in the example is very sensitive to departures of one system from another. It may be argued that it is too sensitive and that the quantities sought are functionals of the motion which may be much less sensitive to calculational errors than Δ. In the example of Section III, the conventional "first integrals of the motion" are well conserved; it is a consequence of the equations of motion that errors in these integrals are of higher order than the errors in coordinates. Other functionals may have intermediate behavior.

A safe and convenient procedure appears to be to subject the quantities of interest—the "results" of the calculation—to the test afforded by parallel calculations. However, the most sensitive indicator is useful to warn of a possibly dangerous situation. Other kinds of problems may not admit of as obvious a sensitive indicator as Δ of the example.

The inherent stability of the usual first integrals means that a calculation may contain unstable features even if it appears safe or plausible when they are used as indicators.

B. Magnitude and Character of Initial Perturbation

Ideally the initial perturbation should be large enough that a response will develop immediately and it should also be substantially larger than the errors of integration in one step or the errors of roundoff, but otherwise it should be as small as possible. Several trials may be required to find a good perturbation. Various perturbations should be tried stepping in different directions in the function space of the problem. Complicated perturbations that might be difficult to interpret should be avoided.

C. Interpretation

As in the example, it can be difficult to determine whether an unstable result indicates a purely numerical effect or a property of the physical system that is being represented. Usually, more experimentation is required to distinguish; parallel calculations provide additional measures of the behavior of the system under these experiments. A very sensitive indicator of departure is helpful. The checks with different integration step sizes, mentioned in Section III, are an example of this. Different kinds of experiments, which should alter the interplay between numerical and physical effects, can be tried.

Perhaps the most difficult point to answer, in the case of as strong an instability as that of Section III, is whether any result can have meaning. The best answer to this is to display some quantities, such as the first integrals of the motion, that behave as expected. Then it might be safe to infer that some things can be reliably calculated, even if some others cannot.

D. Application to Other Calculations

The method described here should be of some use with almost any calculation of evolutionary character. Stellar evolution calculations are an obvious example [6]; in effect, parallel calculations have almost been done because, in exploring the evolutionary history of stars of nearly the same mass, the initial conditions are quite similar. Comparisons using a sensitive measure have not been explicitly carried out, however.

References

1. P. D. Lax and R. D. Richtmyer, *Commun. Pure Appl. Math.* **9**, 267 (1956).
2. R. H. Miller, *Astrophys. J.* **140**, 250 (1964).
3. S. von Hoerner, *Z. Astrophys.* **50**, 184 (1960).
4. S. J. Aarseth, *Monthly Notices Roy. Astron. Soc.*, **126**, 133 (1963).
5. R. H. Miller, *Astrophys. J.* **146**, 831 (1966).
6. E. Hofmeister, R. Kippenhahn, and A. Weigert, *Meth. Comp. Phys.* **7** (to be published).