Round-off induced instabilities and some unusual arithmetics

Michael Blank *

C.N.R.S., Observatoire de Nice, BP 229,
06304 Nice Cedex 4, France, e-mail: blank@obs-nice.fr

February 14, 1995

Abstract

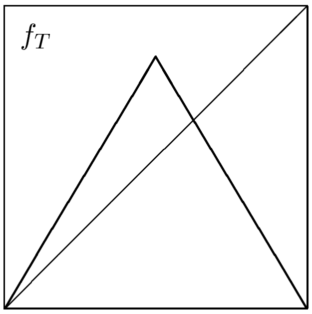
We consider some unusual phenomena that appear in long consecutive calculations due to a new sort of round-off errors of Pentium floating point division bug type. Mathematical model of these round-off errors is proposed and new phenomena in numerical simulations of dynamical systems (differential equations) and other numerical procedures are discussed. It is shown also that these new phenomena could appear not only with Pentium floating point division bugs, but in a more general situation.

1 Introduction

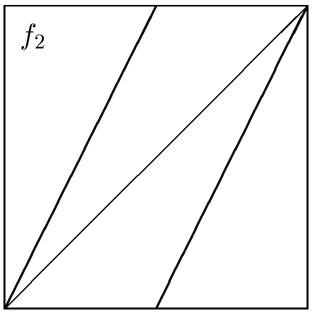
When solving differential equations or other dynamical systems on a computer, we inevitably run into perturbations caused by various types of discretizations. The first of them is time discretization, that is, the transition from differential equations or flows to difference equations or maps. The second type is space discretization, that is, the transition from continuous phase space to a discrete lattice. In numerical simulations these discretizations correspond to round-off errors. Recently discovered Pentium floating point division bug attracted a new attention to the well known problem of round-off errors and gave a possibility to take a fresh look to the problem.

A review of unusual phenomena that appear under the action of round-offs was published recently in [2]. The model of round-off errors, proposed in [2, 3, 4], is the following. Consider a difference equation $x_{n+1} = f(x_n)$, acting on some compact subset $X \subset \mathbb{R}^d$ (for example, on the unit interval $X = [0,1]$). As an example of type of maps under consideration, one could have in mind a so called "tent-map" - one-dimensional ($d = 1$) map $f_T : [0,1] \to [0,1]$, defined by the relation $f_T(x) = a(1/2 - |1/2 - x|)$ (see Fig. 1.a), or another one-dimensional map - a dyadic map $f_2(x) = \{2x\}$ (see Fig. 1.b). Here $\{.\}$ means the fractional part of a number.

Figure 1.a



1.b



*On leave from Russian Academy of Sciences, Inst. for Information Transmission Problems, Ermolovoy Str. 19, 101447, Moscow, Russia.

What do we mean by space discretization of such a system? Actually, in computer simulations we do not have a continuous compact phase space $X \subset \mathbb{R}^d$, but an ordered discrete lattice of points $X_\varepsilon \subset X$, and after the calculation of the next point x_{n+1} we in fact do not have the precise value of $f(x_n)$ but only the nearest point to it in our numerical lattice. The parameter ε (discretization diameter) corresponds to the maximal distance between neighboring points of this lattice. Let us denote the operator that associates to each point $x \in X$ its nearest point on the lattice $x_\varepsilon \in X_\varepsilon$ by $D_\varepsilon : X \to X_\varepsilon$ (if there are several such points we choose the point with the minimal index). Then the discretized map is $f_\varepsilon = D_\varepsilon \circ f$ and the discretized system is a pair $(f_\varepsilon, X_\varepsilon)$.

From the first sight, it seems, that this model is the most general, as possible, especially, if we take into account that the lattice X_ε needs not to be uniform. However any model of this type preserves the order in the following sense. In the one-dimensional case ($d = 1$) for any two points $x \leq y \in \mathbb{R}^1$ the natural order preserves under discretization: $D_\varepsilon(x) \leq D_\varepsilon(y)$. For uniform d-dimensional discretizations (or independent arbitrary discretizations on each coordinate) for any $i = 1, 2, \ldots, d$ the inequality $x_i \leq y_i$ implies $(D_\varepsilon(x))_i \leq (D_\varepsilon(y))_i$, where x_i is the i-th coordinate of a vector $x \in \mathbb{R}^d$. As it will be seen later this consideration could be crucial in some computer simulations.

The Pentium floating point division bug, according to [9] and also to Cleve Moler and Thomas Nicely reports consists in the following. Double precision divisions involving operands with certain bit patterns can produce incorrect results. The most dramatic example was given by by Tim Coe of Vitesse Semiconductor: $x = 4195835,\ y = 3145727,\ z = x - (x/y) * y$. With exact computation, z would be zero, but, on the Pentium, $z = 256$. According to [9] such the bit patterns that could produce pathological results are very rare, and therefore it is quite improbable to find something of this type in calculations with a short number of consecutive steps. On the contrary, dealing with numerical simulations of dynamical systems or partial differential equations, one calculates very long trajectories of the system, i.e. consecutive iterations of the map f. Therefore the probability to run into the point with a "bad" bit pattern during the calculation of a long trajectory with the initial point chosen at random, could be very large, even if a percentage of "bad" bit patterns is very small. Actually exactly in such a kind of calculations the considered bug was discovered firstly by T. Nicely [7], who investigated asymptotic properties of the sum of reciprocals to prime numbers. It not known yet how many operands cause the Pentium's floating point division to fail (T. Nicely estimates the probability of this event as 10^{-9} [7]), or even what operands produce the largest relative error.

Remark, that the hardware errors of this type are not quite new. Consider, for example, well known errors in the implementations of trigonometric functions, random number generators, etc, on various computers. Consider now in detail how a function is calculated on a computer. Let, for example, $f(x) = \{\pi x\}$. Then in a computer implementation of this function we shall obtain $f_\varepsilon(x) = \{D_\varepsilon(D_\varepsilon(\pi)\ \ D_\varepsilon(x))\}$, rather than $D_\varepsilon(\{\pi x\})$. The difference is that one should consider the round-off error on every elementary step of the calculation. Clearly, when ε is small enough, the difference between these two values is also small. Formally, denoting the new discretization by $\mathcal{D}_\varepsilon$, one could write the discretized system as a pair $(\mathcal{D}_\varepsilon f, X_\varepsilon)$, however now $\mathcal{D}_\varepsilon f$ is not a composition of the discretization $\mathcal{D}_\varepsilon$ and the map f, but a more complex procedure, depending on the decomposition of the map into elementary exact calculated procedures and on the hardware errors of Pentium flaw type. The assumption that the number of "bad" points is small enough simply means that the proportion of the points in the numerical lattice such that $|D_\varepsilon \circ f(x) - \mathcal{D}_\varepsilon f(x)| \sim \varepsilon$ is close to 1.

One could tell that the only difference between D_ε-discretizations and $\mathcal{D}_\varepsilon$-discretizations consists in the fact, that the round-off error is larger in the latter case and there exists another lattice (with larger ε) corresponding to the $\mathcal{D}_\varepsilon$-discretization. However this is not so. Indeed, even if the magnitude of errors due to $\mathcal{D}_\varepsilon$-discretization is small enough, say of order ε, the operator $\mathcal{D}_\varepsilon$ may not preserve order, which as we show further leads to new phenomena in the behavior of discretized systems or numerical procedures.

It could be surprising, but the exact calculation of the usual round-off error on the every elementary step could bring to the absence of the order preserving property. Consider an example. Let $x = 1$, $y = 1 + 3\varepsilon/5$, $u = 1 - \varepsilon/3$, $v = 1 + 2\varepsilon/5$. Then $x/y > u/v$, while $\mathcal{D}_\varepsilon(x/y) = D_\varepsilon(x)/D_\varepsilon(y) = 1/2 < 1 = D_\varepsilon(u)/D_\varepsilon(v) = \mathcal{D}_\varepsilon(u/v)$ for the uniform D_ε-discretization with any $\varepsilon > 0$. Clearly such situations could not appear with other arithmetical operations, and even with the division they are very rare, exactly as in the case of the Pentiom flaw.

There are two typical situations in which very rare errors could lead to a visible result in numerical simulations. One of them is the calculation of very long trajectories or consecutive statistics, when the probability to run into the mistake becomes large. The second situation corresponds to the fact that this type of perturbations does not preserve order, and, thus, when order is important, even fine enough $\mathcal{D}_\varepsilon$-perturbations could alter the result not only quantitatively, but qualitatively.

In [2] a series of examples was considered when usual discretizations with arbitrary small diameters (corresponding to arbitrary small round-off errors) lead to a very drastic difference in behavior of the perturbed (discretized) systems with respect to the original ones. On the one hand rigorous conditions under which statistical (the roughest) properties of the system survive discretization were discussed there, and on another hand examples where, even in the limit of vanishing perturbation, a "localization" phenomenon takes place: trajectories which should normally be dense remain confined to a small number of points. In the present paper we shall discuss new phenomena that could appear under the action of $\mathcal{D}_\varepsilon$-discretizations to compare the behavior of dynamical systems or numerical procedures under D_ε- and $\mathcal{D}_\varepsilon$-discretizations, when round-off errors are small enough.

2 Periodic trajectories

Remember that an orbit or a trajectory of a map f is a sequence of points $\{x_k\}_1^\infty$ such that $x_{k+1} = f(x_k)$ for all k. For the case of a periodic orbit of period n such a sequence is formed by a repetition of a finite collection of different points $(x_1, x_2, \cdots, x_n)$ and therefore it is defined by any of these points (for example x_1) and by the period. We call such a point periodic with period n. In contrast with the original dynamical system, for every map f, the set of non-wandering points (that is the broadest nontrivial invariant set) of the discretized system coincides with the set of periodic points. Indeed, the phase space of the discretized map X_ε consists of only a finite number of points, and thus any trajectory will eventually hit into some periodic trajectory. Therefore, the investigation of the asymptotic properties of the perturbed system is reduced to an investigation of solely periodic orbits.

"Period multiplication" phenomenon, described in [2] (Theorem 1), consists in the emergence in a neighborhood of the original cycle of a new cycle, the period of which is a multiple of that of the parent cycle. In the 1-dimensional case ($d = 1$) only the period doubling could take place, while for $d > 1$ any multiple period could appear (Theorem 1). In a numerical investigation of a bifurcation structure of the system (especially in the well known Feigenbaum scenario of the transition to chaos by repeated period doubling) this leads to a very complex problem to distinguish the real bifurcations from "numerical" ones.

The general scheme of the period doubling of a stable fixed point is the following:

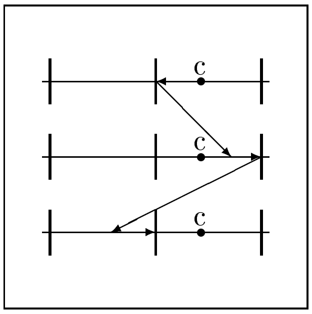


Figure 2.

The proof of this theorem is based on the order-preserving property for both the map f (locally around the periodic trajectory) and the discretization D_ε. Clearly, this is not the case with $\mathcal{D}_\varepsilon$-discretizations, and therefore even for $d = 1$ one could find period multiplication phenomena with arbitrary multipliers. Consider, for example, the situation above, when $\mathcal{D}_\varepsilon(c) \neq D_\varepsilon(c)$, where $c \in X$ is a stable fixed point of the map f, i.e., $f(c) = c$.

"Period multiplication" does not depend on the stability of the original periodic trajectory and in the $\mathcal{D}_\varepsilon$-discretizations case large period multipliers could appear even when the $\mathcal{D}_\varepsilon$-round-off error is of the same order as the usual round-off error. On the other hand, the $\mathcal{D}_\varepsilon$-round-off errors could prevent "period multiplication" even if all the conditions for its existence are satisfied.

The problem of the description of general effects of space discretization is of special interest in the case of chaotic dynamical systems (see, for example, [4] and referencies therein), because all their trajectories are unstable and therefore the influence of perturbations on individual trajectories may grow (exponentially) in time. A typical example is the "dyadic" map (see Fig 1.b). The iterates of almost all initial points are uniformly distributed and all features of developed chaos are visible. The dyadic map has periodic trajectories of all periods and the set of its periodic points is dense on the interval $[0, 1]$. More complex examples occur in the three body problem in celestial mechanics, the Hénon mapping $(x, y) \to (1 - ax^2 + y, bx)$ and others (see [4] and references therein for further examples). Questions about properties of discretized chaotic systems are an active area of mathematical investigation and many problems are still open. Practically, there are only a few results, showing that orbits of the discretized chaotic systems may have a connection to the original orbits. Among these results the main general one shows the connection of the orbit of weakly perturbed systems (with arbitrary type of perturbation) to the original orbit. This result (shadowing property) is due to D.V. Anosov [1], who showed that for a smooth hyperbolic system for each orbit of a weakly perturbed system there exists uniformly closed orbit of the original one.

The shadowing orbit may be untypical in the sense that the distribution of its points (which is the most important feature of a chaotic system) differs from the typical one, defined by the Sinai-Bowen-Ruelle measure [6]. Theorem 2 [2] shows that this situation may often happen in case of the usual arbitrary fine discretizations. Indeed, the "localization phenomenon", described in this theorem corresponds to a situation, when trajectories, which should normally be dense, remain confined to a small number of points. The simplest example, when this phenomenon takes place is the dyadic map under binary uniform discretizations ($\varepsilon = 2^{-n}$). Let $f(x) := \{2x - q/2^n\}$ and let $q \in \mathbb{Z}^1$. Then the uniformly 2^{-n}-discretized map has a fixed poit $q/2^n$ and under the action of the discretized map any point $x \in X_{2^{-n}}$ eventually hits to this point after not more than n iterations. However, in a small neighborhood of the point $q/2^n$ there are points $\beta \in X_{2^{-n}}$ such that $\min\{k : (D_{2^{-n}} \circ f)^k(\beta) = q/2^n\} = n$. Suppose now, that $\mathcal{D}_{2^{-n}}(q/2^n) = \beta$ and there are no $\mathcal{D}_\varepsilon$-errors during the first n iterates of the point β. This assumption is quite natural, because the "bad" points supposed to be very rare. Then the $\mathcal{D}_\varepsilon$-discretized system has a n-periodic trajectory, which certainly differs from the statement above.

As we mentioned in Section 1, in spite of the fact that $\mathcal{D}_\varepsilon$-round-off errors are very rare, their magnitude could be large enough (remember the example of Tim Coe) and thus Anosov's theorem about the shadowing property could not be applyed here. Actually this type of perturbations may be considered as a noise with very rare large fluctuations. A generalized shadowing property for perturbations small in average was considered in [5], where under the same assumptions as in Anosov's theorem the existence of shadowing in average trajectories was proved.

3 Neutral periodic trajectories

Consider now the influence of discretizations on systems, such that the trajectories are neither stable nor unstable, or such they have a local invariant component of this type. Systems of this type appear in various mathematical models of intermittent behavior (for example, a map $x \to \{x + x^2\}$) and therefore their investigation is very important. The simplest example of a map with neutral trajectories is a rotation $f^{(\alpha)}$ of points in the d-dimensional plane $\mathbb{R}^d$ by a fixed angle $\alpha \in \mathbb{R}^{d-1}$ around the origin. In this case the investigation of individual trajectories has no sense, and only the global properties like ergodicity, stability etc. are of interest. Once more, statistical approach is quite useful here for the case of usual space discretizations, and due to the same reason as before, it could not be applied for $\mathcal{D}_\varepsilon$-discretizations.

Numerical simulations of D_ε-discretizations of the two-dimensional irrational rotation in [2] shows (see Fig. 3.a) that the asymptotic trajectories look like thick clusters of quasy circles with gaps (without any trajectory there) between them. Typical numerical trajectory is filling a large part of such cluster, and for any trajectory the distance to the origin remains bounded.

The main question here is if there are numerical trajectories going to infinity or to the origin. Numerical investigations give a negative answer to this question (see Fig. 3.a). However, due to the thickness of numerical trajectories in the clusters, the appearance of "bad" points, such that $\mathcal{D}_\varepsilon(x) \neq x$, in these clusters becomes quite probable, and, thus, a trajectory could go very far from the origin, or could return to the origin.

Consider now $\mathcal{D}_\varepsilon$-discretizations. This means that the elements of the rotation matrix are calculated with the precision ε and all the multiplications are done with the same precision. The result is shown on Figure 3.b. The difference becomes especially striking if one takes into account that all the points outside of the circle of radius 0.45 are going out of the circle of radius 10^6, thus the asymptotic absolute error is of order 10^6, which gives actually a positive answer to the question above. To explain this phenomenon consider a "toy" model of the rotation to the angle π on the one-dimensional integer lattice: $x \to -D_1(x/p)\,p$, where $p > 1$ is an integer. This map has a fixed point 0 and 2-periodic trajectories ($kp \longleftrightarrow -kp$, $k = 1, 2, \ldots$) with the distance p between them. Even more interesting features could be found in the integer version of the rotation to the angle 2π:

$$x \to x + D_1(x/p)\,p - D_1((x-1)/p)\,p - D_1(1/p)\,p \equiv x + D_1(x/p)\,p - D_1((x-1)/p)\,p.$$

In the exact calculation without the discretization this is the identical map. However for $p = 2q + 1$, $q \in \mathbb{Z}_+$, the discretized map preserves all the integer points of the real line except the points $x = kp + q + 1$, $k = 0, \pm 1, \pm 2, \ldots$ going to plus infinity.

4 Instability of sorting procedures

Order preserving property is especially important in numerical realizations of various sorting procedures. Let us start from the well known and very efficient procedure to find zeroes of a one-dimensional function. The scheme is the following. On the every step of the procedure the interval, where the zero is located is divided into two equal parts and, according to the value of the function in the boundary point, the left (right) subinterval is chosen. The procedure converges exponentially even in the presense of $\mathcal{D}_\varepsilon$-round-off errors, but the solution may be quite far from the real zero of the function. Indeed, if, during the long sequence of comparisons, we only once run into the $\mathcal{D}_\varepsilon$-round-off error, then all the consequent steps are wrong.

This situation is far from from the worst in the presense of $\mathcal{D}_\varepsilon$-round-off errors, because the error, while very probable, does not grow in time. The latter case becomes typical in consecutive sorting procedures. Modern sorting algorithms like "submassive merging", "priority queues" etc. (see [8] for details) use tree-like structures in sorting and therefore even the only one small $\mathcal{D}_\varepsilon$-round-off error could switch the procedure to another branch of the tree, and therefore to bring to a large sorting error.

Surprisingly, application of various sorting procedures becomes an efficient tool in modern mathematical physics. Recently an algorithm heavily based on sorting was proposed in [10] for numerical investigation of Burgers' equation with random initial data. Implementation of the algorithm leads to a consecutive sorting of a very large amount of data (about 10^7 points) in each step, thus, the appearance of $\mathcal{D}_\varepsilon$-round-off errors becomes unavoidable.

Acknowledgements

I would like to thank Uriel Frisch, who actually proposed the idea to investigate Pentium-type errors, and Michel Henon for helpfull discussions. This work was supported by the French Ministry of Higher Education.

References

[1] D. V. Anosov, Geodesic flows on closed Riemannian manifolds with negative curvature, *Trudi. Math. Inst. Russian Ac. of Sci*, **90**, 1967, pp. 210–283.

[2] M. L. Blank, Pathologies generated by round-off in dynamical systems, *Physica* D **78**:1-2, 1994, pp. 93–114.

[3] M. L. Blank, Ergodic properties of discretizations of dynamical systems, *Docl. Russian Ac. of Sci.*, **278**:4, 1984, pp. 779–782.

[4] M. L. Blank, Small perturbations of chaotic dynamical systems, *Uspekhi Matem. Nauk.*, **44**:6, 1989, pp. 3–28.

[5] M. L. Blank, Metric properties of ε-trajectories of dynamical systems with stochastic behavior, *Ergodic Theory and Dynamical Systems*, **8**:3, 1988, pp. 365–378.

[6] L. A. Bunimovich, Ya. G. Pesin, Ya. G. Sinai, M. V. Jacobson, Ergodic theory of smooth dynamical systems. Modern problems of mathematics. Fundamental trends, **2**, 1985, pp. 113–231.

[7] B. Cipra, How number theory got the best of the Pentium chip, *Science* **267**, 1995, p. 175.

[8] D. E. Knut, The art of computer programming. **3** Sorting and searching, Addison-Wesley, 1973.

[9] H. P. Sharangpani, M. L. Barton, Statistical analysis of floating point flaw in the Pentium^{TM} processor, Preprint, 1994.

[10] M. Vergassola, B. Dubrulle, U. Frisch, A. Noullez, Burgers' equation, devil's staircases and the mass distribution for large-scale structures, *Astron. & Astrophys.* **289**, 1994, pp. 325–356.