A Letters Journal Exploring
the Frontiers of Physics

OFFPRINT

Relaxation to the asymptotic distribution of global errors due to round off

G. Turchetti, S. Vaienti and F. Zanlungo

EPL, **89** (2010) 40006

epl

A Letters Journal Exploring the Frontiers of Physics

February 2010

EPL, **89** (2010) 40006
doi: 10.1209/0295-5075/89/40006

www.epljournal.org

Relaxation to the asymptotic distribution of global errors due to round off

G. Turchetti[1], S. Vaienti[2] and F. Zanlungo[1(a)]

[1] *Department of Physics, University of Bologna - Bologna, Italy, EU*
[2] *UMR-6207 Centre de Physique Théorique, CNRS, Universités d'Aix-Marseille I, II, Université du Sud, Toulon-Var and FRUMAM - Toulon, France, EU*

received 16 November 2009; accepted in final form 1 February 2010
published online 5 March 2010

PACS 05.45.-a – Nonlinear dynamics and chaos
PACS 05.45.Pq – Numerical simulations of chaotic systems

Abstract – We propose an analysis of the effects introduced by finite accuracy and round-off arithmetic on discrete dynamical systems. We investigate, from a statistical viewpoint and using the tool of the decay of fidelity, the error of the numerical orbit with respect to the exact one. As a model we consider a random perturbation of the exact orbit with an additive noise, for which exact results can be obtained for some prototype maps. For regular anysocrounous maps the fidelity has a power law decay, whereas the decay is exponential if a random perturbation is introduced. For chaotic maps the decay is superexponential after an initial plateau and our method is suitable to identify the reliability threshold of numerical results, *i.e.* a number of iterations below which global errors can be ignored. The same behaviour is observed if a random perturbation is introduced.

Introduction. – Numerical computations for dynamical systems are affected by round-off errors due to the finite-accuracy representation of real numbers. Continuous-time systems, defined by initial-value ordinary differential equations, are replaced by iterated maps, defined by numerical-integration algorithms. Supposing accurate bounds for the global discretization error and estimates on the round-off error for the map are available, the discrepancy between the exact and the numerical orbit is controlled. For maps with an hyperbolic attractor the shadowing lemma assures the existence of an exact orbit close to a numerical one, provided that the local error (single iteration) is small enough. This lemma is of relevant theoretical importance since it gives information about the reliability of computations of average properties, but its non-constructive character (the initial point of the shadowing orbit is not known) limits its practical use [1,2]. We propose a direct comparison between the exact and the numerical orbit to determine the statistical distribution of the error when the initial condition spans its accessible range and the use of fidelity to determine the asymptotic distribution of errors and the way it is reached. This approach allows us to study the global error in a quantitative way. Due to the peculiar character of the round-off error, which depends on the machine architecture, only experimental results can be presented. Nevertheless, since it is customary to assume that the local truncation error is random, we analyze also the error induced by random perturbations, for which an analytical treatment can be provided by using the fidelity.

Additive noise. – We begin by recalling the main results obtained applying the fidelity to random perturbations of dynamical systems. We consider a map T defined on a phase space X which is a subset of $\mathbb{R}^d$, endowed with an invariant *physical* measure μ defined by

$$\lim_{n\to\infty}\int_X \Phi(T^n(x))\,\mathrm{d}m(x)=\int_X \Phi(x)\,\mathrm{d}\mu(x), \tag{1}$$

where m denotes the Lebesgue measure and Φ a continuous observable. Let us then consider a sequence of independent and identically distributed random variables ξ_i with values in the probability space Ξ and with probability density $\eta(\xi)$ such that $Tx+\varepsilon\,\xi$ still maps X into itself. The iteration of the map T is therefore replaced by a composition of maps chosen randomly close to it (note that T itself is recovered when $\varepsilon=0$): $T^n_\varepsilon(x)=(T+\varepsilon\xi_n)\circ(T+\varepsilon\xi_{n-1})\circ\ldots\circ(T+\varepsilon\xi_1)\,(x)$ and the stationary measure μ_ε of the process is defined by [3]

$$\lim_{n\to\infty}\int_{X,\Xi}\Psi(T^n_\varepsilon(x))\,\mathrm{d}m(x)\prod_i\eta(\xi_i)\,\mathrm{d}\xi_i=\int_X\Psi(x)\,\mathrm{d}\mu_\varepsilon(x). \tag{2}$$

[(a)] E-mail: francesco.zanlungo@gmail.com

We want to study the statistical properties of the error at the n-th iteration, defined as $\Delta^n_\varepsilon(x)=T^n(x)-T^n_\varepsilon(x)$. Let ρ^n_ϵ be the probability density of the random variable Δ^n_ϵ defined as $\mathbb{P}(\Delta^n_\epsilon\leqslant t)=\int_{-\infty}^{t}\mathrm{d}s\,\rho^n_\epsilon(s)\,\mathrm{d}s$. The expectation value is defined by $\mathbb{E}(f(\Delta^n_\epsilon))=\int f(\Delta^n_\epsilon)\,\mathrm{d}m(x)\prod\eta(\xi_i)\,\mathrm{d}\xi_i=\int_{-\infty}^{+\infty}f(s)\,\rho^n_\epsilon(s)\,\mathrm{d}s$. The probability density can be studied directly, through a Monte Carlo sampling over initial conditions x and random perturbations ξ, or indirectly, using the fidelity defined through the following integral:

$$F^n_\varepsilon=\int_{X,\Xi}\Phi(T^n(x))\,\Psi(T^n_\varepsilon(x))\,\mathrm{d}m(x)\prod_i\eta(\xi_i)\,\mathrm{d}\xi_i. \qquad (3)$$

Indeed the expectation value of $e^{iu\Delta_\epsilon}$, is just the fidelity $F_n(\epsilon)=\mathbb{E}(e^{iu\Delta^n_\epsilon})$ if we choose $\Phi(x)=e^{iu\,x}$ and $\Psi(x)=e^{-iu\,x}$. As a consequence since $\mathbb{E}(e^{iu\Delta^n_\epsilon})=\int e^{iut}\,\rho^n_\epsilon(t)\,\mathrm{d}t$ the probability density $\rho^n_\epsilon(t)$ is given by the inverse Fourier transform of the fidelity. For a large class of maps which mix exponentially fast, it can be shown that the fidelity converges to $\int_X\Phi(x)\,\mathrm{d}\mu(x)\int_X\Psi(x)\,\mathrm{d}\mu_\varepsilon(x)$, and the absolute value of the difference between the integral (3) and its limiting value in terms of the invariant and stationary measure will be called the *fidelity error* and denoted with δF^n_ε. Note that by the asymptotic characterization of the invariant and stationary measures, the fidelity error could be equivalently defined as

$$\begin{aligned}\delta F^n_\varepsilon\;=\;&F^n_\varepsilon-\int_X\Phi(T^n(x))\,\mathrm{d}m(x)\\&\times\int_{X,\Xi}\Psi(T^n_\varepsilon(x))\prod_i\eta(\xi_i)\,\mathrm{d}\xi_i\,\mathrm{d}m(x). \qquad (4)\end{aligned}$$

If the fidelity error converges to zero, then the fidelity converges to the product of the Fourier transforms of the measures $\int_X e^{iux}\,\mathrm{d}\mu(x)\int_X e^{-iux}\,\mathrm{d}\mu_\varepsilon(x)$ and its inverse Fourier transform is just ρ^∞_ε. The initial error distribution is the Dirac distribution $\rho^0(s)=\delta(s)$, whereas it can be shown that if the invariant and stationary measures are Lebesgue and the map is defined on the torus $\mathbb{T}^1$, then the asymptotic distribution is the triangular function $\rho_\infty(s)=(1-|s|)\vartheta(1-|s|)$ [4]. If we suppose the perturbation is not random, setting for instance all the ξ_i equal to 1, then $T^n_\varepsilon=(T+\varepsilon)^n$ and we are just comparing the iterates of two close deterministic maps. In this case we have not to integrate on the noise and the fidelity simply becomes

$$F'^n_\varepsilon=\int_X\mathrm{d}m(x)\Phi(T^n(x))\Psi(T^n_\varepsilon(x)) \qquad (5)$$

This is exactly the definition of classical fidelity which was proposed in [5] and which was modified in [6] with the addition of the noise. In fact it can be proved that the fidelity with noise (3) generally decays, while this is not the case for the version (5). However the latter is closer to the quantum fidelity which is defined in the following way. Let us suppose that $|\psi\rangle$ is an initial quantum state which evolves forward up to time t under the Hamiltonian H_0 and then backward for the same time t under the perturbed Hamiltonian $H_\varepsilon=H_0+\varepsilon V$, where V is a potential. The overlap of the initial state with its image $e^{iH_\varepsilon t}e^{-iH_0t}|\psi\rangle$ is quantified by the quantum fidelity defined as $f_q(t)=|\langle\psi|e^{iH_\varepsilon t}e^{-iH_0t}|\psi\rangle|^2$.

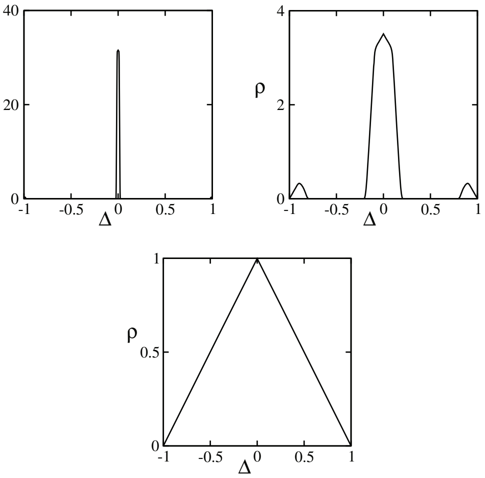

Fig. 1: ρ^n_ε for $3x$ mod 1, $\varepsilon=2^{-25}$. Top, left: $n=13$; top, right: $n=15$, bottom: $n=18$. Compare the transition times with the corresponding decay of fidelity in fig. 3, and note that the transition happens in correspondence with the threshold.

We first recall the main results obtained for randomly perturbed maps [4]. For two prototype systems such as the translations on the torus (a regular system) and the Bernoulli map (a chaotic system), analytical results are available also for the transient. Choosing the probability distribution of the random perturbations as $\eta(\xi)=\frac{1}{2}\chi_{[-1,1]}(\xi)$ we found that the fidelity for translations is given by [4]

$$F^n_\varepsilon=\sum_{k\in\mathbb{Z}}\Phi_k\Psi_{-k}S^n(k\varepsilon), \qquad S(x)=\frac{\sin(2\pi x)}{2\pi x}, \qquad (6)$$

where Φ_k and Ψ_k are the Fourier components of functions Φ and Ψ, whereas for the Bernoulli map $Tx=qx$ mod 1, with integer $q\geqslant 2$ we have

$$F^n_\varepsilon=\sum_{k\in\mathbb{Z}}\Phi_k\Psi_{-k}S_{n,q}(k\varepsilon), \qquad S_{n,q}(x)=\prod_{j=0}^{n-1}S(q^jx). \quad (7)$$

In the first case the decay of fidelity is exponential, with time scale ε^{-2}. In the second case we have a plateau of length $n_*\propto-\ln\varepsilon$, followed by an ε-independent super-exponential decay. Below the threshold n_* the error probability distribution can be approximated by a δ-function, and the perturbed system can be considered as equivalent to the unperturbed one. The asymptotic error distribution is the same for the two systems (since the physical and stationary measure coincide with Lebesgue), and results to be the triangular function (fig. 1).

Our numerical study of maps for which analytical results are not available (Hénon, Baker's, Intermittent, Logistic and Standard maps) [4] shows that the behavior of translations and of Bernoulli maps can be considered as a prototype of, respectively, regular and chaotic maps. In particular for all the studied chaotic maps it is possible to identify a threshold $n_* \propto -\ln\varepsilon$ below which the perturbed system can be considered as faithful to the unperturbed one. The threshold is followed by an ε-independent super-exponential decay.

Numerical noise. – Since real numbers have to be represented as strings of bits on a computer, to each discrete map T there corresponds a numerical map T_*. The action of the numerical map depends on the length of bit strings used to represent real numbers and on the details of round-off algebra, which are hardware dependent [7]. A preliminary analysis was carried out in [8] and some general results are stated in [9]. We can write, using the notation introduced for additive noise, $T_*x = T_\varepsilon x = Tx + \varepsilon\xi(x)$, where ξ now depends in a deterministic way on the initial condition and ε is a constant whose magnitude is of the order of the last significant bit used to represent x. The iterated map T_*^n can be written as T_ε^n using the previously introduced notation, but in this case the single-step errors ξ_i for $i = 1, \ldots, n$ will be n different functions of the initial condition x. Notice that T_* is defined by the round-off rules of the computer, and the notation T_*^n correspond to apply the map as defined by those rules n times. The single-step error $\varepsilon\xi_n = T_*^n x - T(T_*^{n-1}x)$ is introduced to compare the round-off results with the previous results for additive noise.

Fidelity as previously introduced (3) included an integral over all the possible single-step error realizations ξ_i. Nevertheless, from a numerical point of view, integrals were performed with a Monte Carlo method, *i.e.* choosing N representative random vectors (x, ξ_i), a procedure that led to a relative error of order $N^{-1/2}$.

We suppose that if the deterministic $\xi_i(x)$ functional dependence of single-step errors on the initial condition is complex enough, the vector $(x, \xi_i(x))$ can be considered as equivalent to a random sequence. In the case of round-off errors, the Monte Carlo integral over initial conditions and noise performed for random perturbations is thus replaced by an integral over the only initial conditions.

Corresponding to this *ansatz*, whose validity we are going to verify, we can compute the fidelity error for a system perturbed with numerical noise as

$$F_*^n = \int_X \Phi(T^n x)\Psi(T_*^n x)\,\mathrm{d}m(x), \tag{8}$$

$$\delta F_*^n = F_*^n - \left(\int_X \Phi(T^n x)\,\mathrm{d}m(x)\right)\left(\int_X \Psi(T_*^n x)\,\mathrm{d}m(x)\right). \tag{9}$$

This definition requires the knowledge of the exact map T, which is in general not available. This problem can be solved by comparing the round-off map T_*, realized with a given precision, *i.e.* as a string of bits of a given length, with a map realized at an higher precision, $T_\dagger$, that we call the "reference" map. For example T_* could be a single-precision (8 digits) map, and $T_\dagger$ a double-precision (16 digits) map. The numerically computed fidelity will thus be

$$F_*^n = \int_X \Phi(T_\dagger^n x)\Psi(T_*^n x)\,\mathrm{d}m(x). \tag{10}$$

To check the relevance of these results, we can compare them with those obtained substituting $T_\dagger$ with $T_\ddagger$ (for example a map realized using 24 or 32 significant digits). If the results do not depend on the precision of the reference map, we can assume that they are equivalent to those that could be obtained if we had access to the exact map. We have applied this procedure to obtain the results shown in this paper. Typically, the results obtained using as reference map a double precision $T_\dagger$ or a 24 (32) digit map $T_\ddagger$ are equivalent below a given time scale. This time scale, that corresponds to the time scale under which $T_\dagger$ can be considered equivalent to $T_\ddagger$, as can be checked through a direct comparison between $T_\dagger$ and $T_\ddagger$, is considerably longer than the time scale at which the error probability distribution of T_* has reached its asymptotic form and thus the results of the single-double precision comparison can be considered equivalent to those that would be obtained by a comparison between a single precision and an exact map.

Another way to avoid the problems related to the inaccessibility of the exact map would be to rely, for invertible maps, on a different definition of fidelity as

$$\tilde{F}_*^n = \int_X \Phi(x)\Psi(T_*^n T_*^{-n} x)\,\mathrm{d}m(x), \tag{11}$$

where T_*^{-1} is the numerical realization of the inverse map, which is in general different from the inverse of T_*. We notice that the new definition (11) of fidelity would be equivalent to the original one (3) if we replace $\Psi(T_*^n T_*^{-n} x)$ with $\Psi(T_*^n T^{-n} x)$. This is not surprising, since it can be shown that in general the points $T_*^n T_*^{-n} x$ and $T_*^n T^{-n} x$ have a comparable distance from x. When we replace the first expression $(T_*^n T_*^{-n} x)$ with the second expression $(T_*^n T^{-n} x)$ in eq. (11) we are back to the standard definition of fidelity for the map T_*, so that (11) can be considered a sort of equivalent definition of fidelity for a map with round off. For the numerical realization of an invertible map as the standard map, the equivalence between the two definitions has been checked with good results, at least in the chaotic regime [9].

The details of the round-off process depend strongly on the architecture, nevertheless our studies show that some general rules about the error distribution can be stated [9]. For regular maps the behavior is significantly different from additive noise, showing that the integral over the only initial conditions is not equivalent for these systems to an integral also on the noise. For translations on the torus, $Tx = x + \omega \bmod 1$, which are the prototypes for integrable maps, we have found [9] that it is possible to write the global error $\Delta^n_*(x)=T^nx-T^n_*x$ as

$$\Delta^n_*(\omega,x)=\varepsilon\big(\bar{\xi}(\omega)\phi(n)+w_n(x)\big), \tag{12}$$

$$\phi(n)=n+\phi_0+\phi_1 n^{-1}+\dots .$$

Here ε is a constant that represents the last significant bit (2^{-25} for a real number on the torus represented in single precision), $\bar{\xi}$ is a constant that depends in a non-trivial (and machine-dependent) way on ω, while w_n is a bounded, periodic function dependent on the initial condition x, having zero mean once averaged over initial conditions. We thus have

$$\lim_{n\to\infty}\frac{\Delta^n_*(\omega,x)}{n}=\varepsilon\bar{\xi}. \tag{13}$$

We distinguish between isochronous maps, for which the frequency does not depend on initial conditions, and anisochronous maps, such as the skew map on the cylinder $x'=x+\omega(y)$, $y'=y$. In the former case fidelity does not decay since the system is basically equivalent to a deterministic $\omega+\varepsilon\bar{\xi}$ rotation [9]. In the latter case, discussed below, the fidelity has a power law decay. The period and the maximum value of w_n depend on the algorithmic realization of the map, for example for the most simple realization the period is of order 100 and the magnitude $\approx$10, while it has a period of order $\approx 10^5$ below which its variance grows linearly if it is realized as a 2D rotation.

For the anisochronous map with $\omega(y)=y$ (to which any map with monotonic $\omega(y)$ can be reduced), the numerically observed result can be analytically proved if we assume that $\bar{\xi}$, defined above, depends linearly on y. In this case, even though there is no integration over a random variable, the integration over y is equivalent to the integration over $\bar{\xi}$. Letting $T^n=x+ny$ and $T^n_\epsilon=x+ny+n\epsilon\bar{\xi}$ and integrating over x and $\bar{\xi}$ the fidelity is [9]

$$F^n_*=\sum_{k\in\mathbb{Z}}\Phi_k\Psi_{-k}\frac{\sin(2\pi nk\varepsilon)}{2\pi nk\varepsilon}. \tag{14}$$

This result corresponds to original definition of fidelity (5) but the equivalence of the integration over y and $\bar{\xi}$ causes the $1/n$ decay (fig. 2). In the same figure we show the decay law, obtained when the skew map is perturbed by an additive noise according to $x_n=x_{n-1}+y_{n-1}+\epsilon\xi^x_n$, $y_n=y_{n-1}+\epsilon\xi^y_n$. The explicit result in this case reads [9]

$$F^n_*=\sum_{k\in\mathbb{Z}}\Phi_k\Psi_{-k}\left(\frac{\sin(2\pi k\varepsilon)}{2\pi k\varepsilon}\right)^n\prod_{j=1}^{n}\frac{\sin(2\pi jk\varepsilon)}{2\pi jk\varepsilon}. \tag{15}$$

Maps very close to integrable exhibit almost the same behavior as the skew map described above.

Fidelity decay can be easily studied for chaotic numerical maps (for example Bernoulli, Hénon, Logistic, Intermittent, Baker's map and the Standard map with $K\gg 1$) and shows always a good qualitative agreement with the results obtained for additive noise, *i.e.* the presence of a threshold below which fidelity is constant and the error function is qualitatively a δ-function (its support is many orders of magnitude smaller than the size of the phase space), and thus the results of numerical computations can be considered quantitatively reliable [9]. Beyond this threshold, that we can call n_* and that grows as $-\ln\varepsilon$, *i.e.* linearly in the number of bits used to represent real numbers (fig. 3), the error distribution spreads quickly over the whole phase space, as can be checked using also a Monte Carlo sampling of the error distribution.

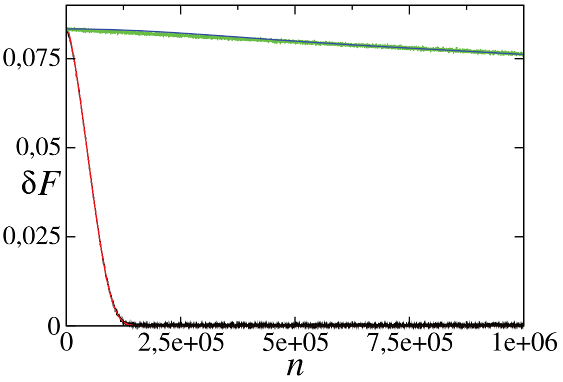

Fig. 2: (Colour on-line) Decay of fidelity for the skew map. Black: additive noise, compared with its analytical prediction eq. (15) (red); green: round-off noise compared with its analytical prediction eq. (14) (blue).

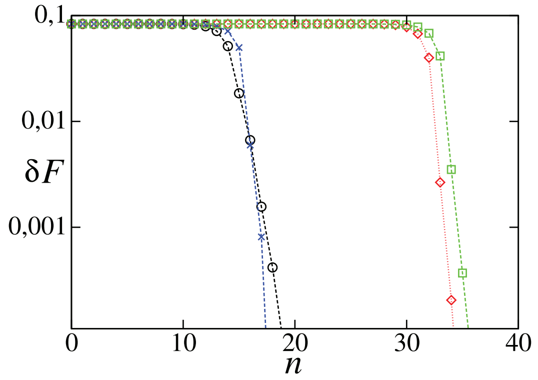

Fig. 3: (Colour on-line) Decay of fidelity for $3x \bmod 1$ represented in single precision (black, circles) and double precision (red, diamonds) compared to a reference map $T_\dagger$ using 32 digits; the results are compared to the decay of fidelity for random noise with $\varepsilon=2^{-25}$ (blue, crosses) and $\varepsilon=2^{-53}$ (green, squares), the value of the last significant bit of real numbers on the torus represented, respectively, as single- and double-precision floating points. Notice the slower decay for single precision after the threshold.

Nevertheless we cannot assume that the sequence ξ_i is always equivalent to a random one. Actually, for the map $3x \bmod 1$ we have found [9] that the global error can be described as

$$\Delta^n_*(x)=\sum_{i=0}^{n}3^{n-i}\varepsilon\xi_i(x), \tag{16}$$

where the initial-condition round off ξ_0 has a step-wise continuum spectrum distribution once sampled over the space of initial configurations, while the ξ_i with $i \geqslant 1$ have a discrete spectrum, which results to be almost completely reduced to zero for $i \geqslant 2$ (*i.e.* no relevant errors are made after the first iteration) [9]. This effect, which is probably due to the extremely simple algorithmic nature of the map, is reflected in the decay law that follows the threshold n_*, which is different from additive noise (compare the decay law for single precision and additive noise in fig. 3).
For the other, (slightly) more algorithmically complex maps, the sequence of error has a continuous spectrum and a period significantly longer than the threshold time scale [9]. For these systems the decay after the threshold is qualitatively equivalent to the additive noise one.

Conclusions. – We have used the results of a previous work on additive noise to study the effects of round-off on discrete dynamical systems. We have generalized the fidelity to the maps perturbed by "numerical noise" (*i.e.*, finite accuracy in numerical computations), and its decay allows us to analyze the probability distribution function of global errors with respect to the exact solution. For regular maps the behavior depends on the algorithmic realizations and on their character. For isochronous maps the fidelity error does not decay whereas for anisochronous maps it has a power law decay, in contrast with the exponential decay caused by additive noise. Chaotic systems with round-off and additive noise exhibit an almost equivalent behavior, *i.e.* it is possible to identify a threshold for a sharp transition from a δ-like error distribution (faithful numerical map) to the asymptotic error distribution. For chaotic numerical maps, below this threshold, which grows linearly as the number of bits used to represent real numbers, the numerical system can be considered as equivalent to the exact one.

* * *

GT acknowledges financial support from the PRIN 2007 project coordinated by D. Batani. FZ has been partially supported by the GDRE 224, INdAM CNRS GREFI-MEFI.

REFERENCES

[1] Grebogi C., Hammel S. and Yorke J., *J. Complex.*, **3** (1987) 136.
[2] Sauer T., *Phys. Rev. E*, **65** (2002) 036220.
[3] Kifer Y., *Ergodic Theory of Random Transformations* (Birkhäuser) 1986.
[4] Marie P., Turchetti G., Vaienti S. and Zanlungo F., *Chaos*, **19** (2009) 043118.
[5] Benenti G., Casati G. and Velbe G., *Phys. Rev. E*, **670** (2003) 055202.
[6] Liverani C., Marie Ph. and Vaienti S., *J Stat. Phys.*, **128** (2007) 1079.
[7] Knuth D., *The Art of Computer Programming,* Vol. **2** (Addison-Wesley) 1969.
[8] Turchetti G., in *CHAOS, Systèmes Dynamiques* (Hermann Éditeurs) 2007, p. 239.
[9] `www.physycom.unibo.it/zanlungo_web_dir/num.pdf`.