Computing 61, 47–67 (1998)

Computing
© Springer-Verlag 1998
Printed in Austria

Rigorously Computed Orbits of Dynamical Systems without the Wrapping Effect

W. Kühn, Berlin

Received March 1, 1997; revised January 15, 1998

Abstract

A new method for rigorously computing orbits of discrete dynamical systems is introduced. High order zonotope enclosures of the orbit are computed, using only matrix algebra. The wrapping effect can be made arbitrarily small by choosing the order high enough. The method is easy to implement and especially suited for parallel computing. It is compared to other well known strategies, and several examples are given.

AMS Subject Classifications: 65G10, 39A10, 65L05.

Key words: Dynamical systems, wrapping effect, zonotopes.

1. Introduction

For given initial set Ω_0 in $\mathbf{R}^d$ and maps $f_n : \mathbf{R}^d \to \mathbf{R}^d$, consider the discrete dynamical system

$$\Omega_n = f_n(\Omega_{n-1}) \tag{1}$$

for stages $n = 1,2,\ldots$ Systems of this kind are frequently studied in many branches of mathematics. One example is the time discretization of ordinary differential equations. There, f_n is the time-h_n map with $f_n(x) = \Phi(x, h_n)$, where $\Phi(x, \cdot)$ is the solution for the initial value problem $u' = g(u)$ and $u(0) = x$, and $\{h_n\}$ is a sequence of stepsizes.

The goal is to construct supersets (enclosures) for the orbit of (1) in such a way that the overestimation is kept small. System (1) is not immediately suitable for iteration on a computer, and it will be unavoidable to overestimate the range $f_n(\Omega_{n-1})$ at every stage n by wrapping it into a superset both feasible to construct and to represent on a computer. Because the overestimation of a wrapped set is proportional to its radius, a spurious growth of the enclosures can result if the composition of wrapping and mapping is now iterated. This **wrapping effect** can be completely unrelated to the stability properties of the system, and even stable systems are shown to exhibit exponentially fast growing enclosures (see below). Such over pessimistic enclosures, although rigorous, are useless for practical purposes.

There is an extensive literature on the wrapping effect, mostly in relation with validated numerical solutions for the initial value problem of ODEs. A nice review on this topic is given in [13], and two bibliographies are contained in [2] and [3]. The wrapping effect was first observed with the advent of interval methods in the early 1960s. Subsequently, interval methods gained a stigma of exponential overestimation in conjunction with dynamical systems.

A successful algorithm for computing enclosures of (1) has to address several issues. It should avoid exponential overestimation altogether in systems which are sufficiently stable, and a guaranteed (under mild assumptions) bound of subexponential order on the overestimation is certainly desirable. However, a good practical algorithm is not determined by a good asymptotic bound but by a convincing performance on average problems. Even a mere linear overestimation in n can be unacceptable, for example in the long term integration of hamiltonian systems. One can therefore wish to have a "variable precision" algorithm, with the understanding that higher precision implies higher complexity (storage and computational requirements).

A common strategy to construct enclosures is to replace (1) by a dynamical system suitable for iteration on a computer (marching method). A collection of sets representable on a computer (polytopes, ellipsoids, etc.) and maps F_n operating on this collection are chosen such that $f_n(\Theta) \subseteq F_n(\Theta)$. Such maps are also called extensions of f_n over the particular collection. The iterates of the dynamical system

$$\Theta_n = F_n(\Theta_{n-1}), \tag{2}$$

are then indeed enclosures for the orbit of (1), provided that $\Omega_0 \subseteq \Theta_0$.

Let us illustrate how the wrapping effect can effect the marching strategy. Consider the collection of all balls B_r with respect to a fixed norm $\|\cdot\|$, and define extensions on this collection "$F_n(B_r)$ is the smallest ball containing $f_n(B_r)$". If we assume that the maps $f_n = T$ are linear, then, by definition, $\operatorname{rad}(F_n(B_r)) = \|T\| r$ (the radius $\operatorname{rad}\Omega$ of a compact set Ω is the radius of the smallest ball containing Ω). Therefore, if $\Theta_0 = \Omega_0 = B_{r_0}$, then

$$\operatorname{rad}\Theta_n = \operatorname{rad}\left(F_n \circ \cdots \circ F_1(B_{r_0})\right) = \|T\|^n r_0,$$

whereas the subsets Ω_n obey the much better estimate

$$\operatorname{rad}\Omega_n \leq \|T^{(n)}\| r_0.$$

The factor by which the radii of the enclosures Θ_n exceed the radii of Ω_n is $\|T\|^n / \|T^{(n)}\|$, and this factor can grow exponentially. As a simple example, let $\|\cdot\|$ be the maximum norm and $T = R_{45^\circ}$ be the rotation by 45 degree about the origin, in which case $\operatorname{rad}\Theta_n = \sqrt{2}^n \operatorname{rad}\Omega_n$. The balls with respect to weighted maximum norms, boxes whose sides are parallel to the coordinate axis, play a dominant role in enclosure methods and are called interval vectors.

Surprisingly, the wrapping effect persisted all previous attempts to find marching methods which replace the simple interval vectors by better and more complex enclosures. In fact, this work was prompted by the discovery of a simple example (see Sections 4 and 9) for which the popular QR-method in [9] fails exponentially. The QR-method, which can be viewed as a preconditioned interval vector method, computes parallelotope enclosures. It is very effective for pure rotations, but yields linear overestimations for average problems. The main potential of the QR-method is that it can easily be combined with other marching methods to filter out the effect of rotations.

Besides parallelotope enclosures also polytopes [4] and ellipsoids [6], [12] are used in the literature. The construction of these enclosures is an intrinsic geometrical problem that is rather involved and hard to attack. For example, the Simplex Method is used in [4].

We present a marching method (subsequently called the cascade reduction) which beats the wrapping effect effectively and efficiently by constructing high order zonotope enclosures of the orbit. A zonotope is the Minowski sum of straight line segments. In this paper, the number of line segments is always a multiple m of d, the dimension of the space. The integer m is a measure for the geometrical complexity of the zonotopes. It can be chosen freely and is a **performance parameter** for the method. Zonotopes are relevant in such diverse areas as measure theory, computational geometry and convexity, but to the best of our knowledge, zonotopes have not yet been employed in numerical analysis. A brief introduction is therefore given in Section 2.

Some highlighted properties of the cascade reduction are

- approximations overestimate only sub-exponentially in the stage index n by an order of $\exp(cn^{1/m})$;
- the computation time needed for each stage is proportional to m on a single processor computer;
- the computation time needed for each stage is proportional to $c_1 + \log_2 m$ on a computer with more than m processors;
- the required storage is proportional to m;
- m can be chosen freely in order to balance between desired precision on one hand and execution time on the other; $3 \leq m \leq 15$ is typical;
- the relevant computations consist of simple matrix operations like products and row sums.

A **Java** implementation of the cascade reduction that runs in any applet-enabled browser is available at *http://www.zib.de/kuehn.*

The only non-marching method known to the author [5] uses the fact that the overestimation introduced by evaluating an interval matrix chain product using logarithmic multiplication (pair ordering) is polynomial in n for asymptotically small matrix width. The algorithm works for nonlinear systems, but it is not clear how the polynomial bound applies to the general case. It uses a stack of $\log_2(n)$ interval matrices and can thus not be described as a finite dimensional system. Contrary to marching methods, and somewhat counterintuitive, it cannot be viewed as successively mapping forward a set in time n. Logarithmic mulitplication generally gives good results for small enclosure radii, for which it compares to the cascade reduction with $m \approx \log_2(n)$. However, systems with medium or large enclosure radii ($> 10^{-7}$) are not handled well, which is probably due to the interval matrix multiplication used, Section 11.

Also, logarithmic multiplication behaves like a high order bandpass filter, favoring overestimation for rotations in small bands centered at 30° or dyadic multiples thereof, Section 10. This artifact can potentially hinder the use of logarithmic multiplication in adaptive discretization where the error radius is taken as a measure for the discretization error (and should therefore not correlate to the angle of rotation).

Let us conclude the introduction with two notes. First, we have to specify a measure for the overestimation between two sets. Probably the truest measure in our context, and used in this paper, is the Hausdorff distance. For compact sets $\Omega \subseteq \Theta$ it is defined by

$$\operatorname{dist}(\Omega,\Theta) = \min_{\text{all balls } B} \{2\operatorname{rad} B : \Theta \subseteq \Omega + B\}.$$

Volume measures seem to be too indiscriminating with respect to slivers, which can have big radii but arbitrarily small volume. The second note concerns the nature of the maps f_n in (1). The only restriction shall be that these are approximated well by linear maps in the sense that the range $f_n(\Omega)$ is close to an affine image of Ω. In particular, if f_n is C^1, then our analysis is essentially local in nature (see [1] for a non-local treatment).

2. Zonotopes and their Representation

Zonotopes are a special class of polytopes.

Definition 1 *Let A be a $d \times k$ matrix with column vectors A_j. Then the set*

$$\Diamond A := \left\{ \sum_{j=1}^{k} s_j A_j : |s_j| \leq 1 \right\} \tag{3}$$

*is called a k-**zonotope**, or a zonotope of **order** k. Let Z_k be the collection of all k-zonotopes. The collection $Z = \bigcup_{k=1}^{\infty} Z_k$ of all k-zonotopes is the set of all zonotopes.*

In the literature, "d-zonotope" is used to refer to any zonotope in d-space. In this paper, however, the number preceding "zonotope" always refers to the number of summands in (3). A review on zonotopes can be found in [15]. See [16] for an up-to-date treatment. The columns of A are not required to be independent. Geometrically, the zonotope in (3) is the Minkowski sum of the k centered straight line segments $\overline{-A_j, A_j}$. This immediately implies that $\Diamond A$ is a centered set (recall that a set Ω is centered if $\Omega = -\Omega$). The set $a + \Diamond A$ is the translate of $\Diamond A$ by a and is centered at point a.

Graphically, a zonotope is constructed as follows. Sweep the origin along the segment $\overline{-A_1, A_1}$, creating a line segment. Then sweep that line segment along $\overline{-A_2, A_2}$ to create a parallelepiped. Then sweep this parallelepiped along $\overline{-A_3, A_3}$, and so on. An illustration of this process for a (3)-zonotope in $\mathbf{R}^2$ can be found in Fig. 1, and Fig. 2 shows a (12)-zonotope in $\mathbf{R}^3$.

If A and B are identical up to a permutation of their column vectors, then $\Diamond A = \Diamond B$. The unit cube is $\Diamond I$, where I is the identity matrix. If the $d \times d$ matrix A is (i) invertible, (ii) orthogonal or (iii) diagonal, then $\Diamond A$ is (i) a parallelotope, (ii) a cube or (iii) an interval vector in $\mathbf{R}^d$.

We need several preparatory remarks to state the next lemma. For a given $d \times k$ matrix A, define the diagonal (row sum) matrix rs(A) by

$$(\mathrm{rs}(A))_{ii} = \sum_{j=1}^{k} |A_{ij}|.$$

Let $\leq$ be the partial order defined by $A \leq B$ if and only if $A_{ij} \leq B_{ij}$ for all entries (i, j). For the rest of the paper, $\|\cdot\|$ always denotes the maximum norm in $\mathbf{R}^d$ or its induced matrix norm. We also use block matrices of the form $A = [A^1, \ldots, A^m]$, where A_k is the k-th block. If all A_k are square, we say that A is an m-block matrix. The radius rad$\,\Omega$ of a compact set Ω is the radius of the smallest ball containing Ω.

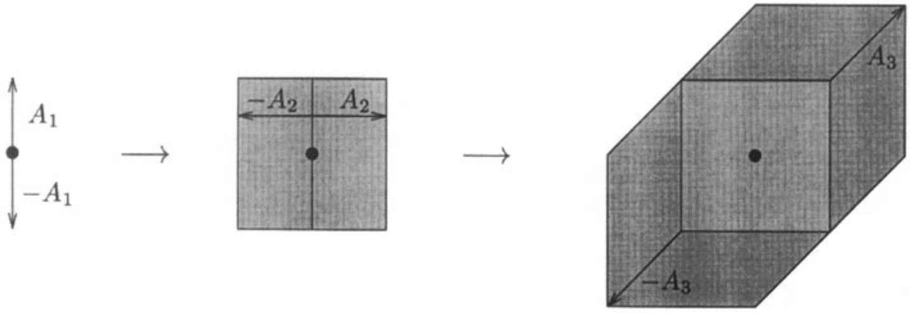

Figure 1. Constructing a 3-zonotope in $\mathbf{R}^2$

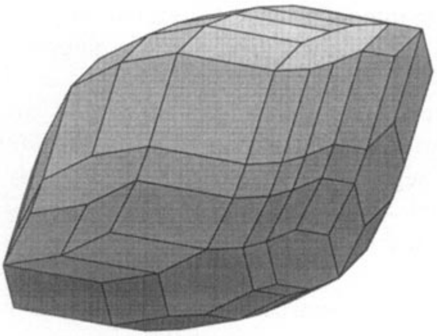

Figure 2. A typical 12-zonotope in $\mathbf{R}^3$, constructed in Section 12 as a solution enclosure for the Lorenz system. It is an optical illusion if the body seems to be non-convex

Lemma 2. *Let A and B be matrices with d rows and T be a matrix with d columns. Then*

(1) $\diamond A \subseteq \diamond \mathrm{rs}(A)$;
(2) $T\diamond A = \diamond(TA)$;
(3) $\diamond A + \diamond B = \diamond[A, B]$;
(4) *if* $\mathrm{rs}(A) \le \mathrm{rs}(B)$, *then* $\diamond \mathrm{rs}(A) \subseteq \diamond \mathrm{rs}(B)$;
(5) $\diamond \mathrm{rs}(A) + \diamond \mathrm{rs}(B) = \diamond(\mathrm{rs}(A) + \mathrm{rs}(B))$;
(6) $\diamond(A + B) \subseteq \diamond A + \diamond B$;
(7) $\mathrm{rad}\diamond A = \|A\|$.

Proof: Claims (2)–(6) are trivial. For (1) observe that if $x \in \diamond A$, then $x_i = \Sigma_j s_j A_{ij}$ with $\|s_j\| \le 1$ by Definition 1. Therefore $|x_i| \le \Sigma_i |A_{ij}|$, which implies $x \in \diamond \mathrm{rs}(A)$. Also

$$\mathrm{rad}\diamond A = \max_{x \in \diamond A} \|x\| = \max_{s_i \in [-1,1]} \|\sum_i s_i A_i\| = \max_{\|s\|=1} \|As\| = \|A\|,$$

which proves (7). qed.

Remarks 3.

- One can easily see that $\diamond \mathrm{rs}(A)$ is the smallest centered interval vector that contains $\diamond A$ and is called the interval hull of $\diamond A$.
- Items (2) and (3) imply that the collection Z of all zonotopes is invariant under addition and linear maps. In fact, it is easy to show that Z is the smallest (with respect to inclusion) of all such invariant collections.

The map $A \mapsto \mathrm{rs}(A)$ **reduces** any matrix A to a square (diagonal) matrix. More generally we have

Definition 4. *A map $\mathscr{R}$ from the space of all d-row matrices into itself is called a reduction if $\Diamond A \subseteq \Diamond \mathscr{R}A$.*

Note that if $\mathscr{R}$ is a reduction, then so is the map $[A, B] \mapsto [\mathscr{R}A, B]$. In particular, the map $[A, B] \mapsto [\mathrm{rs}(A), B]$ is a reduction, and it is the most general reduction used in this paper.

3. Zonotope Extensions

The following Lemma shows one way to implement the dynamical system (2) on the space of all zonotopes by constructing zonotope extensions for the maps f_n.

Lemma 5. *Let y_n and T_n be a sequence of vectors and square matrices, respectively. For given reduction $\mathscr{R}$ and initial A_0, let A_n be the orbit of the dynamical system*

$$A_n = \mathscr{R}[E_n, T_n A_{n-1}], \tag{4}$$

where E_n is a sequence of diagonal matrices such that

$$|(E_n)_{ii}| \geq |(y_n - f_n(y_{n-1}))_i| + \sum_{j=1}^{d} (|f'_n(y_{n-1} + \xi_{n-1}) - T_n)_{ij}(\zeta_{n-1})_j| \tag{5}$$

for all ξ_{n-1} and ζ_{n-1} in $\Diamond A_{n-1}$. Then

$$f_n(y_{n-1} + \Diamond A_{n-1}) \subseteq y_n + \Diamond A_n. \tag{6}$$

The proof consists of a simple application of the Mean Value Theorem. Note that for $\mathscr{R} = I$, the enclosure in (6) is a zonotope centered at y_n and of order d higher than the zonotope $\Diamond A_{n-1}$: it is the affine image of $\Diamond A_{n-1}$ plus an inflating error term $\Diamond E_n$, see Fig. 3. An inspection of (5) suggest to choose $y_n \approx f_n(y_{n-1})$ and $T_n \approx f'_n(y_{n-1})$. If (5) is evaluated using an interval matrix enclosure for the range of f'_n, then a good choice for T_n is probably the midpoint of this matrix. Also, instead of using the derivative, one can construct slope matrices which yield better approximations and do not need differentiability, see [11] and Section 11 for an example.

The issue of this paper is how much the "reduced" orbit $\Diamond A_n$ in (4) overestimates the unreduced orbit $\Diamond \hat{A}_n$ defined by

$$\hat{A}_n = \left[E_n, T_n \hat{A}_{n-1}\right], \tag{7}$$

i.e., with the identity as reduction. We will not consider the overestimation of the original orbit Ω_n in (1) by either $\Diamond A_n$ or $\Diamond \hat{A}_n$. The reason is that we do not assume any control over the error terms E_n (these are determined by the nonlinearity, the rounding errors and the linearization procedure as in Lemma 5), but rather want to study appropriate reductions $\mathscr{R}$. It is easy to see that the use of a reduction $\mathscr{R} \neq I$ is mandatory for practical purposes. Suppose that the initial term $\hat{A}_0$ and all error terms E_n are square matrices (usually, E_n will even be diagonal as in Lemma 5). Then $\hat{A}_n$ is an $(n+1)$-block matrix, which to

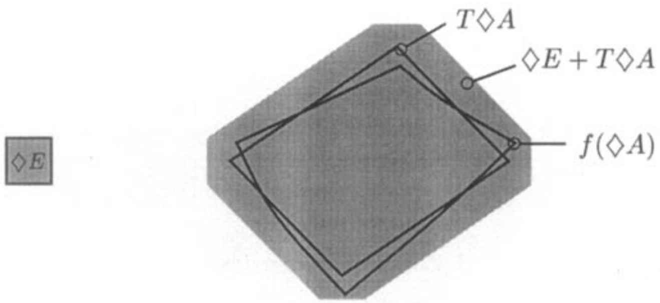

Figure 3. Inflated linear image $\Diamond E + T\Diamond A$ is enclosure of non-linear image $f(\Diamond A)$

compute amounts to a cost of order n^2 and a storage of order n. This is acceptable only for a small number n of stages, and the unreduced system (7) has no practical relevance.

4. Heuristic for a Good Reduction

Suppose that for fixed integer m, the reduction $\mathscr{R}$ maps $(m+1)$-block matrices to m-block matrices. Then the orbit A_n in (4) consists of m-block matrices, provided the initial matrix A_0 is an m-block matrix.

First consider the case $m = 1$, for which a simple reduction is

$$\mathscr{R}[E, TA] := \operatorname{rs}(E) + \operatorname{rs}(TA), \tag{8}$$

the reduction to diagonal form. Each iterate therefore represents an interval vector. Historically, a lot of effort was expended to find a good preconditioner S which yields a better reduction

$$\mathscr{R}[E, TA] := S\big(\operatorname{rs}(S^{-1}E) + \operatorname{rs}(S^{-1}TA)\big).$$

For example, [10] suggests $S = T^{-1}$, and [9] sets S equal to the Q-part of the QR-decomposition of T, with the columns ordered with respect to decreasing Euclidean length. The latter reduction, which we denote by $\mathscr{R}_{QR}$, is considered the best all purpose method. However, all these choices of S yield exponential overestimation in simple examples, see [14], [7] and Section 9.

Now consider the case $m = 2$. As stated in the Introduction, we should reduce only small blocks. Reducing blocks proportional to the size of A itself results in an exponential overestimation, if this is done "too often".

Naively, the best choice seems to be to reduce always the smallest block, i.e.,

$$\mathscr{R}[E, TA^1, TA^2] := \begin{cases} \big[\operatorname{rs}([E, TA^1]), TA^2\big] & \text{for } \|TA^1\| < \|TA^2\|, \\ \big[TA^1, \operatorname{rs}([E, TA^2])\big] & \text{otherwise}. \end{cases}$$

This is a greedy reduction that favors local optimality. But since the two blocks are treated in a symmetric fashion, both blocks are reduced at every second stage on average, resulting in exponential growth.

The key idea for a better reduction is to make sure there is a small block to reduce at most of the stages, with an inevitable but only occasional reduction of a large block. We propose

$$\mathscr{R}_c[E, TA^1, TA^2] := \begin{cases} \left[\mathrm{rs}([E, TA^1]), TA^2\right] & \text{for } \|[E, TA^1]\| \le \|TA^2\|, \\ \left[0, \mathrm{rs}([E, TA^1, TA^2])\right] & \text{otherwise} \end{cases} \tag{9}$$

For example, suppose $T = E = I$ and $A = [0, 0]$ initially. Then the first 6 stages of $A \mapsto \mathscr{R}_c[E, TA]$ are given in Table 1.

After the first block is set to zero, the dynamical system spends many stages reducing the small first block only. Of course, this block grows until it is larger than the second block. At this point the whole matrix is reduced, the first block is set to zero and the whole procedure starts again.

5. The Cascade Reduction

There are many possible ways to generalize the reduction of the last section to more than two blocks. We choose

Definition 6 (Cascade Reduction). *Let $A = [A^1, \dots, A^m]$ be an m-block matrix and let*

$$D(l) = \Big[E, \underbrace{TA^1, \dots, TA^l}_{l}\Big].$$

Choose l to be the biggest integer in $[2, m]$ for which

$$\|D(l-1)\| > \|TA^l\|, \tag{10}$$

or $l = 1$ if such an integer does not exist. Then define

$$\mathscr{R}_c[E, TA] := \Big[\underbrace{0, \dots, 0}_{l-1}, \mathrm{rs}(D(l)), \underbrace{TA^{l+1}, \dots, TA^m}_{m-l}\Big]. \tag{11}$$

Example 7. For the $m = 3$ block case, let us again assume $T = E = I$ and $A = [0, 0, 0]$ initially. Then the first 11 iterates of $A \mapsto \mathscr{R}_c[E, TA]$ are given in Table 2.

Remarks 8.

1. The order condition (10) has the effect that the m blocks A^k have an increasing norm in k. Because blocks are constantly “moved” to the right by virtue of (11), we can draw an analogy to a cascading waterfall in time reverse.

2. For $m = 1$ and $m = 2$, Definition 6 is identical to (8) and (9), respectively.
3. The case $l = m$ in Definition 6 corresponds to a **complete reduction** after which only the last block of $\mathscr{R}_c[E,TA]$ is non-zero. Let us make the simple but important observation that between two complete reductions we have $\mathscr{R}_c[E,TA] = [\mathscr{R}_c[E,TA^1,\dots,TA^{m-1}],TA^m]$. This means that in the cascade reduction of m blocks there are embedded reductions of $m-1$ blocks (self similarity), which is also evident by comparing Tables 1 and 2.
4. The cascade reduction applies to more general situations. The m blocks A^k need not represent d-zonotopes and the reduction $\operatorname{rs}(\cdot)$ need not represent the interval hull. For example, each A_k may represent an ellipsoid and $\operatorname{rs}(D(l))$ in (11) then represents an ellipsoid enclosure for the Minkowski sum of $l-1$ ellipsoids and one interval vector.

6. The Performance of the Cascade

Our goal is to show that for $m > 1$, $\Diamond A_n$ overestimates $\Diamond \hat{A}_n$ only sub-exponentially in n, where A_n and $\hat{A}_n$ are the orbits of (4) and (7), respectively, and $\mathscr{R} = \mathscr{R}_c$ in (4).

Theorem 9. *Suppose* $\|E_n\| \le \delta$ *and* $\|T_{n+k}T_{n+k-1}\cdots T_n\| \le M$ *for all* n, $k \ge 1$ *and some constants* δ *and* M. *Then there exist positive constants* c_1 *and* c_2 *(depending only on* m*) such that*

$$\operatorname{dist}\left(\Diamond \hat{A}_n, \Diamond A_n\right) \le \delta c_1 \exp\left(c_2 n^{1/m}\right). \tag{12}$$

Proof: The proof is by induction in m. Assume $M > 1$ and $\delta = 1$. To see that the latter is not a loss of generality, observe that $\mathscr{R}_c$ is homogeneous in the sense that $\mathscr{R}_c \lambda A = \lambda \mathscr{R}_c A$. We may also assume that $A_0 = \hat{A}_0 = 0$. This can always be achieved by shifting the sequences $\{E_n\}$ and $\{T_n\}$ to the right by one and setting $E_1 = A_0 = \hat{A}_0$ and $T_1 = I$.

Because $\operatorname{dist}(\Omega, \Theta) \le 2\operatorname{rad}\Theta$ for any subset Ω of Θ, it is sufficient to prove that (12) holds for $\|A_n\| = \operatorname{rad}\Diamond A_n$. We will repeatedly use the identities $\|\mathscr{R}_c A\| = \|A\|$ and $\|[A,B]\| = \|A\| + \|B\|$. Let $A_n = [A_n^1,\dots,A_n^m]$ be the orbit of (4) and note that if n is a **complete reduction stage** ($l = m$ in Definition 6), then $\|A_n^m\| = \|A_n\|$.

For $m = 1$, the recursion relation $\|A_n\| = \|\operatorname{rs}(E_n) + \operatorname{rs}(T_n A_{n-1})\| \le 1 + M\|A_{n-1}\|$ together with $A_0 = 0$ yields $\|A_n\| \le 1 + M + \cdots + M^{n-1} \le M^n/(M-1)$. This proves the induction start.

Table 1

n	A	n	A	n	A
0	$[0,0]$	2	$[I,I]$	4	$[I,3I]$
1	$[0,I]$	3	$[0,3I]$	5	$[2I,3I]$

Table 2

n	A	n	A	n	A
0	$[0, 0, 0]$	4	$[0, 0, 4I]$	8	$[I, 3I, 4I]$
1	$[0, 0, I]$	5	$[0, I, 4I]$	9	$[0, 0, 9I]$
2	$[0, I, I]$	6	$[I, I, 4I]$	10	$[0, I, 9I]$
3	$[I, I, I]$	7	$[0, 3I, 4I]$	11	$[I, I\ 9I]$

Now assume that $p < q$ are two successive complete reduction stages and let $\tilde{A}_n$ be the sub-matrix of A_n consisting of the first $m-1$ blocks, i.e., $\tilde{A}_n = [A_n^1, \dots, A_n^{m-1}]$ and $A_n = [\tilde{A}_n, A_n^m]$. Note that $A_p = [0, A_p^m]$. Because $\tilde{A}_p = 0$ and due to the self similarity of the cascade reduction by Remark 8, part (3), we can apply the induction hypothesis to the sequence $\{\tilde{A}_n\}_{n=p}^{q-1}$ to get

$$\|\tilde{A}_n\| \le c_1 c_2^{(n-p)^{1/(m-1)}} \qquad \text{for all} \qquad p \le n < q \tag{13}$$

and some positive constants c_1 and c_2, where we assume w.l.o.g. that $c_2 > 1$. By (10),

$$\|[E_n, T_n\tilde{A}_{n-1}]\| \begin{cases} \le \|T_n A_{n-1}^m\| & \text{for } p \le n < q, \\ > \|T_n A_{n-1}^m\| & \text{for } n = q, \end{cases} \tag{14}$$

and $A_n^m = T_n \cdots T_{p+1} A_p^m$ for $p < n < q$. The last result implies

$$\|T_n A_{n-1}^m\| \le M \|A_p^m\| = M\|A_p\| \qquad \text{for } p < n < q. \tag{15}$$

For all $p < n \le q$,

$$\|A_n\| = \|\mathscr{R}_c[E_n, T_n\tilde{A}_{n-1}, T_n A_{n-1}^m]\| = \|[E_n, T_n\tilde{A}_{n-1}]\| + \|T_n A_{n-1}^m\|.$$

Therefore, by (13) and (14), we have for $n = q$

$$\begin{aligned} \|A_q\| &< 2\|[E_q, T_q\tilde{A}_{q-1}]\| \le 2\left(1 + M\|\tilde{A}_{q-1}\|\right) \\ &\le 2\left(1 + Mc_1 c_2^{(q-p-1)^{1/(m-1)}}\right) \le c_3 c_2^{(q-p)^{1/(m-1)}} \end{aligned} \tag{16}$$

with $c_3 = 2(1 + Mc_1)$, and for $p < n < q$, (14) and (15) imply

$$\|A_n\| \le 2\|T_n A_{n-1}^m]\| \le 2M\|A_p\|. \tag{17}$$

Solving (16) for $q - p$ gives, provided that $\|A_q\| \ge c_3$,

$$q - p \ge d_1 \left(\ln \frac{\|A_q\|}{c_3}\right)^{m-1} \qquad \text{with } d_1 = (\ln c_2)^{1-m} \tag{18}$$

(note that $d_1 > 0$). Furthermore, $\|A_q\| = \|[E_q, T_q A_{q-1}]\| \le 1 + M2M\|A_p\|$, which implies, provided $\|A_p\| \ge 1$,

$$\|A_q\| \le \gamma\|A_p\| \qquad \text{with } \gamma = 3M^2. \tag{19}$$

Let us now assume that the sequence $\{\|A_n\|\}_{n=0}^{\infty}$ is unbounded, for otherwise there is nothing to prove. We claim that there is a strictly increasing sequence of complete reduction stages $n_k, k = 1, 2, \ldots$ such that

$$\|A_{n_k}\| \ge c_3 \qquad \text{for } k \ge 1 \tag{20}$$

and

$$\gamma^{k-1}\|A_{n_1}\| \le \|A_{n_k}\| \le \gamma^k\|A_{n_1}\| \qquad \text{for } k \ge 2. \tag{21}$$

If n_1 is chosen big enough such that (20) holds, then (20) also holds for all $k \ge 2$ by (21) (note that $\gamma > 3$). To show (21), assume that for some $k \ge 2$ no complete reduction stage n_k can be found for which (21) holds. Then there must be two successive complete reduction stages $p < q$ which skip (lie on opposite sides of) the interval $\|A_{n_1}\|[\gamma^{k-1}, \gamma^k]$, see Fig. 4. But then $\|A_p\| < \gamma^{k-1}\|A_{n_1}\|$ and $\gamma^k\|A_{n_1}\| < \|A_q\|$ imply $\|A_q\| > \gamma\|A_p\|$, contradicting (19).

Now suppose $n \ge n_2$ (the initial transient phase $n < n_2$ is insignificant) and let $r \ge 2$ be such that $n \in [n_r, n_{r+1}]$. Then

$$\begin{aligned}
n \ge n_r - n_1 &= \sum_{k=1}^{r-1} (n_{k+1} - n_k) \\
&\overset{\text{by (18) and (20)}}{\ge} \sum_{k=1}^{r-1} d_1\left(\ln\frac{\|A_{n_{k+1}}\|}{c_3}\right)^{m-1} \\
&\overset{\text{by (21)}}{\ge} \sum_{k=1}^{r-1} d_1\left(\ln\frac{\|A_{n_1}\|}{c_3} + k\ \ln\gamma\right)^{m-1} \\
&\overset{\text{by (20)}}{\ge} d_1(\ln\gamma)^{m-1}\sum_{k=1}^{r-1} k^{m-1} \\
&\ge d_1(\ln\gamma)^{m-1}\int_0^{r-1} x^{m-1}dx \\
&= \frac{d_1(\ln\gamma)^{m-1}}{m}(r-1)^m,
\end{aligned}$$

or, solved for r,

$$\left(\frac{mn}{d_1(\ln\gamma)^{m-1}}\right)^{1/m} + 1 \ge r.$$

Because of (17) and (21),

$$\|A_n\| \le 2M\|A_{n_r}\| \le 2M\gamma^r\|A_{n_1}\|,$$

which completes the proof. Note that neither d_1 nor n_1 depend on n. qed.

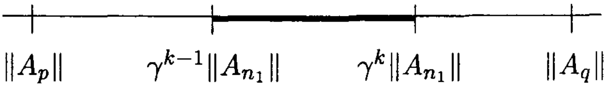

Figure 4

7. The Cascade Reduction for Finite Precision

The sum and the matrix products in Definition 6 cannot be computed with a finite precision arithmetic without committing rounding errors. The following lemma takes care of this by using quantities that are computable with finite precision.

Lemma 10. *Let A and l be as in Definition 6. Suppose $-R^j \le TA^j - C^j \le R^j$ with $R^j \ge 0$ for $j = 1, \ldots, l$ and $\mathrm{rs}(TA^j) \le \mathrm{rs}(M^j)$ for $j = l+1, \ldots, m$. If N is chosen such that*

$$\mathrm{rs}(E) + \sum_{j=1}^{l} \mathrm{rs}(M^j) + \sum_{j=l+1}^{m} \mathrm{rs}(R^j) \le \mathrm{rs}(N) \tag{22}$$

holds, then $\Diamond\mathcal{R}_c[E, TA] \subseteq \Diamond[0, \ldots, 0\ rs(N), C^{l+1}, \ldots, C^m]$.

Proof: First note that $\mathrm{rs}(TA^j - C^j) \le \mathrm{rs}(R^j)$. Then by Lemma 2 parts (2)–(5),

$$\begin{aligned}\Diamond\mathcal{R}_c[E, TA] &= \Diamond\left[\mathrm{rs}(E) + \sum_{j=1}^{l} \mathrm{rs}(TA^j), TA^{l+1}, \ldots, TA^m\right] \\ &\subseteq \Diamond\left[\mathrm{rs}(E) + \sum_{j=1}^{l} \mathrm{rs}(M^j) + \sum_{j=l+1}^{m} \mathrm{rs}(R^j), C^{l+1}, \ldots, C^m\right] \\ &\subseteq \Diamond\left[\mathrm{rs}(N), C^{l+1}, \ldots, C^m\right].\end{aligned}$$

qed.

One way to find the matrices R^j, C^j and M^j in the lemma is to compute the interval matrix product of T and A^j. This gives bounds on each entry e of TA^j in the form $e \in [\underline{e}, \bar{e}]$. Then set $c \approx \frac{1}{2}(\bar{e} - \underline{e})$ and $r = \max\{\bar{e} - c, c - \underline{e}\}$ to get the corresponding entries in C^j and R^j. As entries for M^j use $m = \max\{|\underline{e}|, |\bar{e}|\}$. See [10] or [11] for details. It is not known whether the result of Theorem 9 applies to the finite precision version in Lemma 10.

8. Parallel Computing and Sparse Maps

The most expensive computation in Lemma 10 is the estimation of the products TA^j for $j = 1, \ldots, m$, which is done by the interval matrix product. The sum in (22), which is the sum of $1 + dm$ vectors, is inexpensive. If the cost for the sum of two vectors is c_1 and the cost for the product of two matrices in c_2, then the total cost for Lemma 10 is $(1 + dm)c_1 + mc_2$ on a single processor machine. On a machine with more than m processors, all products can be performed in parallel with a cost of c_2. Also, each row sum in (22) can be performed in parallel with a cost of dc_1. To compute the sum of $\mathrm{rs}(E)$ and these m row sums, another $\log_2(m+1)$ additions are necessary, using $(m+1)/2$ processors. The total cost for a parallel computer is therefore $(d + \log_2(m+1))c_1 + c_2$. Considering that $c_1 \ll c_2$, this is almost independent of m. Numerical examples show that m typically ranges between $m = 3$ and $m = 15$. These are also typical values for the number of processors of multi-processor computers. The cascade reduction has not yet been tested on a parallel computer.

9. Example: A Periodic Sequence of Linear Maps

This section gives an example for which the QR-reduction of [9] (cf. Section 4) fails exponentially. In the $x-y$ plane, consider the shear along the x-axis by one unit, the dilation along the x-axis by two units, and the shear along the y-axis by one unit:

$$S_x = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{and} \quad S_y = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Define the periodic sequence of maps (with period 6)

$$\{T_n\}_{n=1}^{\infty} = \{S_x, S_x^{-1}, D, S_y, S_y^{-1}, D^{-1}, \ldots\}.$$

Note that this sequence satisfies the assumption of Theorem 9 because the composition of the first six elements of $\{T_n\}$ is the identity.

Figure 5 shows the norms of the first 1000 stages for the iterated system $A_n = \mathscr{R}_c[E_n, T_n A_{n-1}]$ with different values for m in the cascade reduction $\mathscr{R}_c$ and $A_0 = E_n = 10^{-10} I$ (recall that $\Diamond I$ is the centered square of radius 1). Note the jumps of the radii, most visible for $m = 3$ at stages $n = 400$, 550 and 800. At these stages, the biggest block A_n^3 is reduced. Also note that the lower envelope for $m = 10$ is the graph of a linear function with slope 10^{-10}. Because an inflating term of radius 10^{-10} is added in each stage, this enclosure cannot be improved. The zonotope enclosures for $m = 10$ therefore do not overestimate the orbit beyond graphical resolution.

To demonstrate how the cascade reduction operates on the individual blocks, we have also plotted in Fig. 6 the values of $\|A_n^1\|$ and $\|A_n^2\|$ for $m = 2$ through the first 100 stages.

The geometrical meaning of the QR-reduction $\mathscr{R}_{QR}$ in the planar case is as follows: if A is a 2×2 matrix then $\Diamond \mathscr{R}_{QR} A$ is the smallest rectangle containing $\Diamond A$ that has one of its longest edges aligned with one of $\Diamond A$'s longest edges. Figure 7 shows the first 6 iterates of $A_n = \mathscr{R}_{QR} T_n A_{n-1}$ and $A_0 = I$ (no error terms here). Because the unreduced sequence $\hat{A}_n = T_n \hat{A}_{n-1}$ is periodic with $A_6 = A_0 = I$, the set $\Diamond A_6$ overestimates $\Diamond \hat{A}_6$ by a factor of 2. Therefore, the radii of the enclosures produced by the $\mathscr{R}_{QR}$ reduction grows exponentially like $2^{n/6}$ in the number n of stages. For $n = 1000$, this corresponds to an increase in radius by a factor of 10^{50}, and the $\mathscr{R}_{QR}$ reduction fails dramatically.

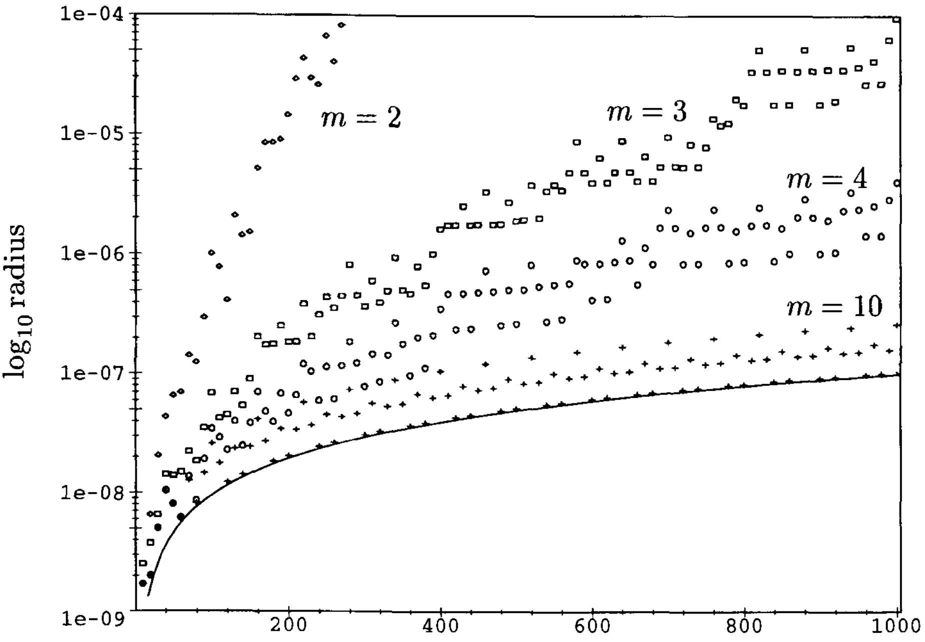

Figure 5. Norms of A_n (radii of enclosure) for different values of m

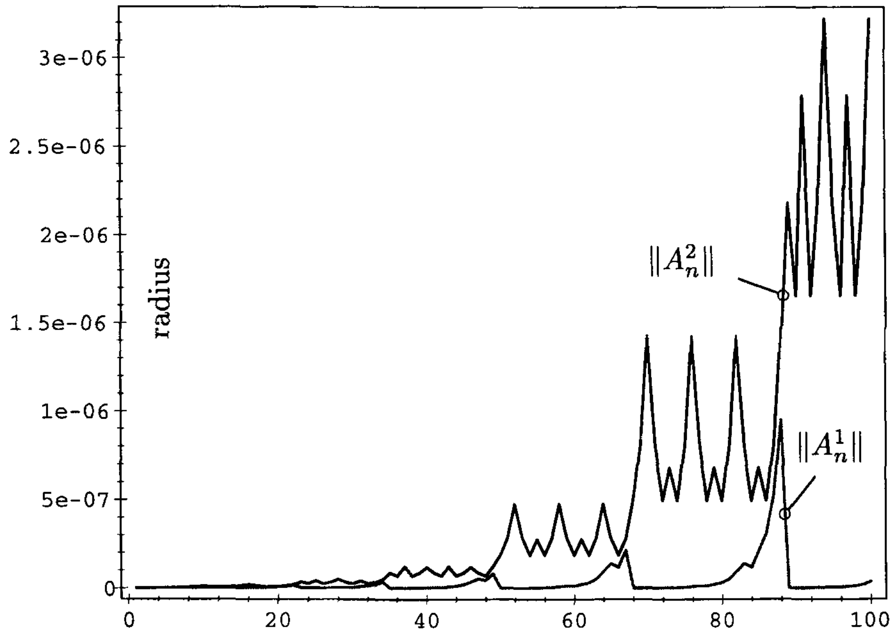

Figure 6. A_n^1 is "transferred" to A_n^2 and then reset to zero whenever it is equal to A_n^2 in magnitude

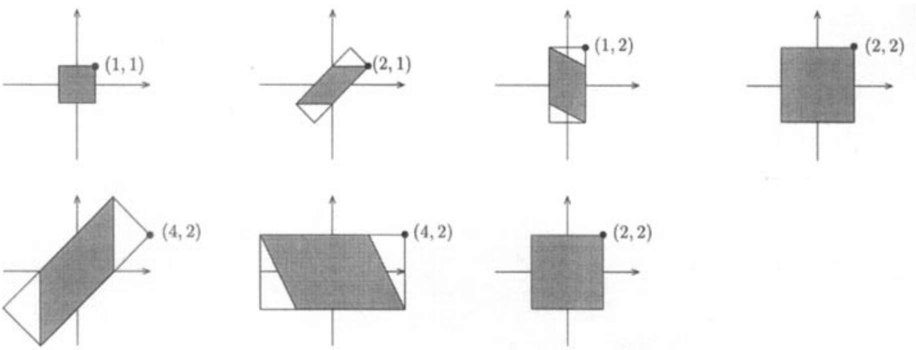

Figure 7. Exponential overestimation of the $\mathscr{R}_{QR}$ reduction. The unit square doubles after the first six iterates for the periodic sequence of maps in Section 9

10. Example: a Rotation by 30°

The filter characteristic of the logarithmic multiplication is nicely shown for the linear system $y_n = A_n y_{n-1}$ with $A_\mathrm{n} \in [0.97, 0.98]R_\theta$. Although the system is stable, logarithmic multiplication fails to pick up the exponential decay for $\theta = 30°$. The results of 5000 iterations with initial value of $y_0 = (1,0)$ and for angles 27°, 30° and 33° are shown in Fig. 8. We also compute the radii after 5000 iterations for angles between 1° and 44°, Fig. 9.

The cascade reduction has a much more uniform frequency response, whereas logarithmic multiplication behaves seemingly erratic. The reason for this peculiar response is that power of 2 multiples of 30 are far from integer multiples of 90 (hint: $1/3 = 0.\overline{01}$ has the binary expansion of smallest period), and similar arguments hold for angles near 15 and 7.5 degrees.

If the wider matrix $A_n \in [0.96, 0.98]R_{30°}$ is chosen, then even a strong error amplification is observed for logarithmic multiplication while cascade reduction still performs adequately.

11. Example: the Hénon Map

The Hénon Map is defined as

$$f(x) = \begin{pmatrix} 1 - \alpha x_1^2 + x_2 \\ \beta x_1 \end{pmatrix}$$

It is area conserving $|\beta| = 1$. We compute enclosures for the orbit $\Omega_n = f(\Omega_{n-1})$.

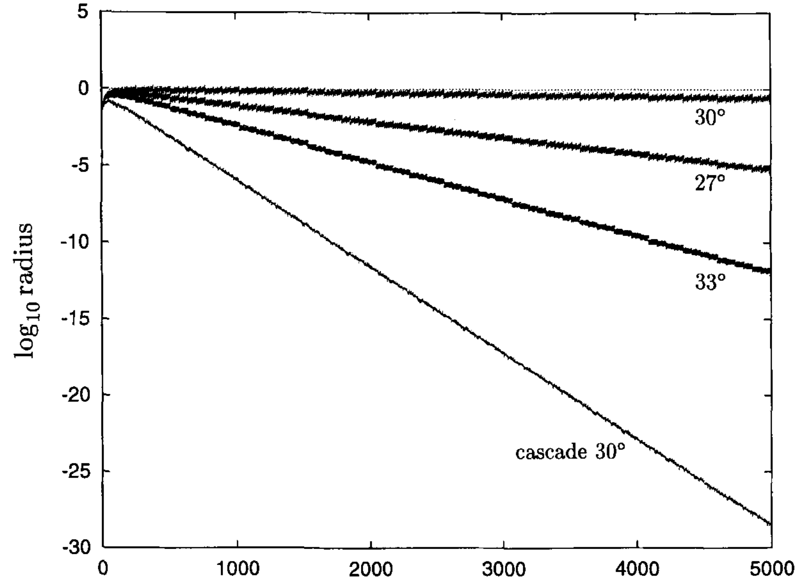

Figure 8. Filter characteristic of logarithmic multiplication (topmost three graphs). Errors are least suppressed for rotations through angles near 30°

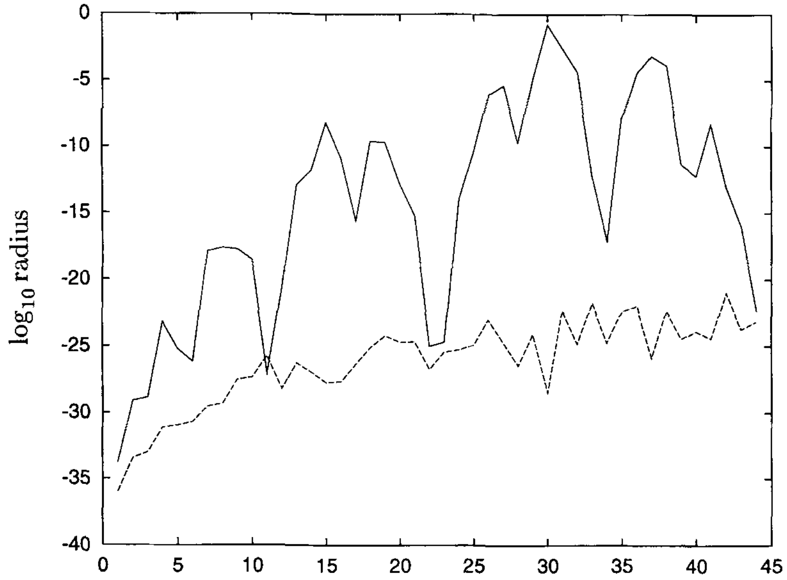

Figure 9. Frequency response of logarithmic multiplication (top) and cascade reduction (bottom)

Let $\Omega = \eta + \Sigma$ with $\Sigma \in \diamond\begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}$. Then

$$f(x) = f(\eta) + \begin{pmatrix} -2\alpha\eta_1 & 1 \\ \beta & 0 \end{pmatrix}(x - \eta) + \begin{pmatrix} -\alpha(x_1 - \eta_1)^2 \\ 0 \end{pmatrix}$$

implies

$$f(\Omega)-f(\eta)\subseteq\diamondsuit\begin{pmatrix}\alpha a_1^2\\ 0\end{pmatrix}+\begin{pmatrix}-2\alpha\eta_1 & 1\\ \beta & 0\end{pmatrix}\Sigma,$$

the right hand side being an inflated linear enclosure of $f(\Omega)-f(\eta)$. Note the quadratic dependence of the inflating term on the radius of Σ due to the non-linearity of f.

The map is iterated 500 times with parameters $\alpha=2.4$ and $\beta=-1$, and Ω_0 is the ball centered at $(0.4,-0.4)$ with radius 10^{-6}. Enclosures in the form of $(2m)$-zonotopes for the set $\Omega_{500}=f^{500}(\Omega_0)$ are shown in Fig. 10. Note that the enclosures are not monotone, i.e., the enclosure for $m=6$ does not contain the enclosure for $m=7$.

Figure 11 compares the enclosure radius of both the cascade reduction and logarithmic multiplication. The same parameters as in the previous figure are used, but this time we have fixed $m=10$ and set the initial radius to 10^{-14}. Similar results are obtained for $n<5000$ with both strategies, although logarithmic multiplication generally yields a much larger variance in the radius. Logarithmic multiplication is outperformed at radii exceeding 10^{-7}. Note that $m=10$ is approximately $\log_2(1000)$, the number of interval matrices which are allocated through the first 1000 stages of logarithmic multiplication. For $m=15$, the cascade reduction is able to execute more than 33000 iterations.

12. Example: The Lorenz System

Consider the time-continuous Lorenz system

$$\begin{aligned}u'_1&=6(-u_1+u_2),\\ u'_2&=u_1(-u_3+28)-u_2, &&\text{and}\qquad u(0)=u_0.\\ u'_3&=u_1u_2-8/3u_3.\end{aligned}$$

In [8] it is showed how to find inflated linear enclosures for the range of the time-h map

$$u_0\mapsto\Phi(u_0,h)\tag{23}$$

mentioned in the Introduction. To test the cascade reduction, (23) is iterated 933 times with a stepsize of $h=0.03$, corresponding to an integration over the time interval $[0,20]$. The initial set Ω_0 is a ball of diameter 10^{-6}, centered at the point $(6.8311,3.222,27.0)$ in phase space. This ball intersects a periodic solution of period $T\approx0.6899$ as shown in [8].

Figure 12 shows the radii of zonotope enclosures constructed for different values of m. Higher values of m do not improve the result. Note that the radius, even in the limit $m\to\infty$, has to increase, because the periodic solution is unstable.

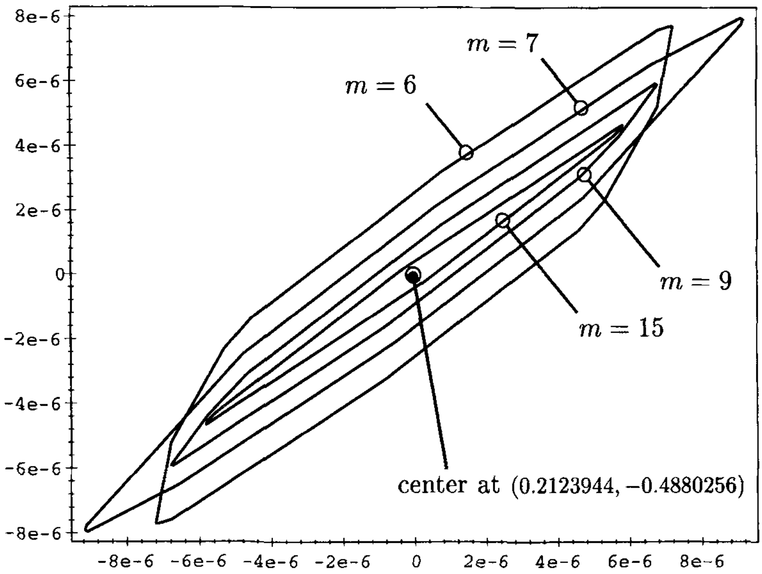

Figure 10. $(2m)$-zonotope enclosures of the Hénon Map $f^{500}(\Omega_0)$ for different values of m

13. Concluding Remarks and Outlook

The cascade reduction is efficient and effective. One of its striking features is the performance parameter m. By simply tuning this parameter, we produce

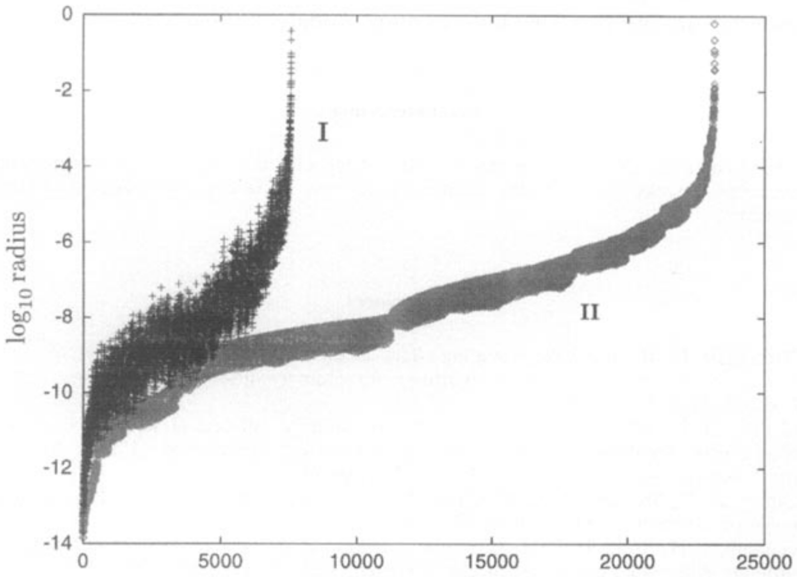

Figure 11. Enclosure radii for the Hénon Map with respect to (I) logarithmic multiplication and (II) the cascade reduction $m = 10$

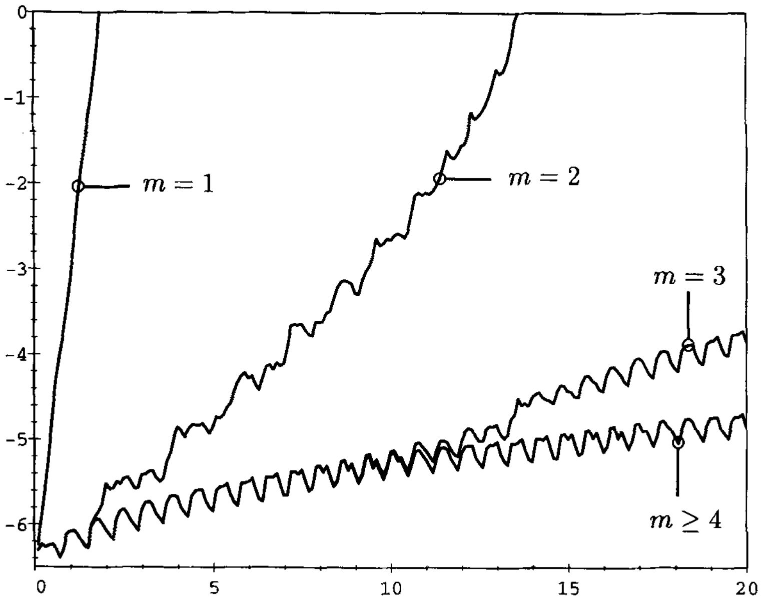

Figure 12. Radii of zonotope enclosures for the Lorenz system for different values of m

either very tight or very inexpensive approximations. Future research should be directed towards an adaptive m-control, in which m is changed during the iteration. It is clear that a too small m may yield too big overestimations. On the other hand, the quality of the approximations does not improve significantly after m is increased beyond a certain threshold. Therefore, an adaptive cascade reduction has to find this threshold and has to balance between cost and quality.

Acknowledgements

The author thanks Dr. Donald Estep and Dr. Ernst Adams for their help during the preparation of the manuscript. Thanks also go to the anonymous referee, in particular for suggesting the use of block matrices.

References

[1] Barbăroşie, C.: Reducing the wrapping effect. Computing *54*, 347–357 (1995).

[2] Corliss, G. F.: Survey of interval algorithms for ordinary differential equations. Appl. Math. Comput. *31*, 112–120 (1989).

[3] Corliss, G. F.: Guaranteed error bounds for ordinary differential equations. VI-th SERC Numerical Analysis Summer School, Leicester University, 1994. Available at ftp://interval.usl.edu/pub/interval_math/bibliographies/surv_ode.bib.

[4] Davey, D. P., Stewart, N. F.: Guaranteed error bounds for the initial value problem using polytope arithmetic. BIT *16*, 257–268 (1976).

[5] Gambill, T. N., Skeel, R. D.: Logarithmic reduction of the wrapping effect with application to ordinary differential equations. SIAM J. Numer. Anal. *25*, 153–162 (1988).

[6] Guderley, K. G., Keller, C. L.: A basic theorem in the computation of ellipsoidal error bounds. Numer. Math. *19*, 218–229 (1972).

[7] Jackson, L. W.: Interval arithmetic error-bounding algorithms. SIAM J. Numer. Anal. *12*, 223–238 (1975).

[8] Kuhn, W.: Rigorous and reasonable error bounds for the numerical solution of dynamical systems. PhD thesis, Georgia Institute of Technology, Atlanta, 1997.

[9] Lohner, R.: Einschließung der Lösung gewöhnlicher Anfangs- und Randwertaufgaben und Anwendungen. PhD thesis, University of Karlsruhe, *1988*. Dissertation.

[10] Moore, R. E.: Interval Analysis. Prentice-Hall, Englewood Cliffs, NJ, 1966.

[11] Neumaier, A.: Interval methods for systems of equations. Cambridge University Press, 1990.

[12] Neumaier, A.: The wrapping effect, ellipsoid arithmetic, stability and confidence regions. Computing [Suppl] *9*, 175–190 (1993).

[13] Nickel, K. L. E.: Using interval methods for the numerical solution of ODEs. Z. Angew. Math. Mech. *66*, 513–523 (1986).

[14] Rihm, R.: Interval methods for initial value problems in ODEs. In: Topics in validated computations (Herzberger, J., ed.). New York: Elsevier Science B.V., 1994.

[15] Schneider, R., Weil, W.: Zonoids and related topics. In: Convexity and its Applications (Gruber, P. M., Wills, J. M., eds.), pp. 296–317. Basel: Birkhauser, 1986.

[16] Ziegler, G. M.: Lectures on polytopes. Berlin Heidelberg, New York Tokyo, Springer, 1995.

W. Kühn
ZIB
Takustrasse 7
D-14195 Berlin-Dahlem
e-mail: kuehn@zib.de