# The wrapping effect, ellipsoid arithmetic, stability and confidence regions

*dedicated to Prof. U. Kulisch on the occasion of his 60th birthday*

Arnold Neumaier

*Institut für Angewandte Mathematik*
*Universität Freiburg*
*Hermann-Herder-Str. 10*
*D-7800 Freiburg*
*Germany*

*e-mail: neum@indi4.mathematik.uni-freiburg.de*

**Abstract.** The wrapping effect is one of the main reasons that the application of interval arithmetic to the enclosure of dynamical systems is difficult. In this paper the source of wrapping is analyzed algebraically and geometrically. A new method for reducing the wrapping effect is proposed, based on an interval ellipsoid arithmetic.

Applications are given to the verification of stability regions for nonlinear discrete dynamical systems and to the computation of rigorous confidence regions for nonlinear functions of normally distributed random vectors.

**Zusammenfassung.** Der Verpackungseffekt ist eine der Hauptursachen dafür, daß die Anwendung von Intervallverfahren auf die Einschließung dynamischer Systeme schwierig ist. In dieser Arbeit wird dieser Effekt algebraisch und geometrisch analysiert. Um den Verpackungseffekt zu reduzieren, wird eine neue Methode vorgestellt, die auf einer Intervall-Ellipsoidarithmetik basiert.

Als Anwendungen werden die Verifikation von Stabilitätsbereichen nichtlinearer diskreter dynamischer Systeme und die Berechnung von rigorosen Konfidenzbereichen für nichtlineare Funktionen normalverteilter Zufallsvariablen skizziert.

# 1 Introduction

Consider the simple triangular linear system $Ax = b$, where

$$A = \begin{pmatrix} 1 & & & & & \\ 1 & 1 & & & 0 & \\ 1 & 1 & 1 & & & \\ & 1 & 1 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & 1 & 1 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} \beta \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{1}$$

Clearly

$$x_1 = \beta, \qquad x_2 = -\beta, \tag{2}$$

$$x_l = -x_{l-1} - x_{l-2} \quad \text{for} \quad l > 2, \tag{3}$$

so that

$$x_l = \begin{cases} -\beta & \text{if } l = 3k-1, \\ 0 & \text{if } l = 3k, \\ \beta & \text{if } l = 3k+1. \end{cases}$$

Suppose we only know that $\beta \in [-\epsilon, \epsilon]$, If we use naive interval arithmetic to solve the triangular system we get

$$x_1 \in [-\epsilon, \epsilon], \qquad x_2 = -x_1 \in [-\epsilon, \epsilon],$$
$$x_3 = -x_1 - x_2 \in -[-\epsilon, \epsilon] - [-\epsilon, \epsilon] = [-2\epsilon, 2\epsilon],$$

and inductively

$$x_l = [-a_l\epsilon, a_l\epsilon],$$

with the Fibonacci sequence $a_1 = a_2 = 1$, $a_{l+1} = a_l + a_{l-1}$, hence $a_l >$ const. $\cdot 1.618^l$. The interval bounds grow exponentially, and compared with the optimal bounds

$$x_l \in \begin{cases} [0, 0] & \text{if } l = 3k, \\ [-\epsilon, \epsilon] & \text{otherwise,} \end{cases}$$

the overestimation is excessive for large $n$. The same kind of overestimation persists when (1) is replaced by a system with wider bands.

The reason for the overestimation is the dependence of the $x_l$ on the *same* vagely known number $\beta$, while interval arithmetic — due to its memory-less nature — assumes that all the $x_l$ vary independently over their enclosing interval.

Of course, there are interval methods which are more reliable, and in this case even optimal, namely those which explicitly precondition the system (1) by an approximate (midpoint) inverse. However, the inverse of $A$ is a full lower triangular matrix which means that work for forming and solving the preconditioned system is of order $O(n^2)$, which is an order of magnitude larger than the work for solving (1) approximately (and in our case exactly since the matrix entries are so simple). For bounded systems arising in the solution of time-dependent problems over many time steps, this increase in work needed for realistic enclosures limits the scope of current interval methods. Thus it is very important to understand this overestimation problem in detail, and to device ways of reducing the complexity while still obtaining reasonable bounds. In this paper we only consider methods which conserve the time-like recurrent structure of banded equations. Other methods (Gambill and Skeel [5], Alvarado [1], [2]) which use divide and conquer strategies will not be discussed here.

We obtain an intuitive geometric interpretation of the mechanism underlying the overestimation if we rewrite the difference equation (3) as

$$\begin{pmatrix} x_l \\ x_{l+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} x_{l-1} \\ x_l \end{pmatrix}. \tag{4}$$

If we change notation we can view this as the particular case

$$A = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad x_l \in \mathbb{R}^2 \tag{5}$$

of the discrete dynamical system

$$x_{l+1} = Ax_l + b. \tag{6}$$

Geometrically, (6) describes an affine transformation of $x_l$ to $x_{l+1}$. The set of allowed positions of the initial value $x_0 = \binom{\beta}{-\beta}$ $(\beta \in [-\epsilon, \epsilon])$ — corresponding to $\binom{x_1}{x_2}$ in (2) — is a line segment, and, as affine transforms of $x_0$, all $x_l$ are line segments, too.

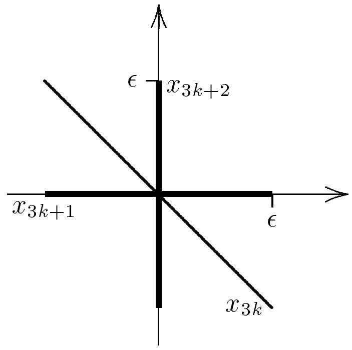


**Figure 1**: Optimal solution sets

In the first step, the effect of interval arithmetic is the replacement of the initial line segment by the smallest box $\boldsymbol{x}_0 = \begin{pmatrix} [-\epsilon,\epsilon] \\ [-\epsilon,\epsilon] \end{pmatrix}$ containing it. In the next step, this box is transformed into a parallelogram, and the smallest box containing it is $\boldsymbol{x}_1 = \begin{pmatrix} [-\epsilon,\epsilon] \\ [-2\epsilon,2\epsilon] \end{pmatrix}$. In the next step, $\boldsymbol{x}_1$transforms into another parallelogram whose interval enclosure is the box $\boldsymbol{x}_2 = \begin{pmatrix} [-2\epsilon,2\epsilon] \\ [-3\epsilon,3\epsilon] \end{pmatrix}$, etc. One sees that the tiny precious birthday present $x_l$ is wrapped into layer after layer of wrapping paper until a very conspicuous present $\boldsymbol{x}_l$ results, whose size has no longer anything to do with its contents. This is the so-called *wrapping effect*, already observed in the early days of interval calculations (Moore [16], [17]).

In general, the behavior of the iteration (6) depends on the spectrum of $A$. If all eigenvalues of $A$ have absolute values $< 1$, the $x_l$ tend to a limit point $x^*$ which is the solution of the linear system $x^* = Ax^* + b$. If some eigenvalues have absolute values $> 1$ then, for most starting points, the iteration will diverge, $\|x_l\| \to \infty$ for $l \to \infty$.

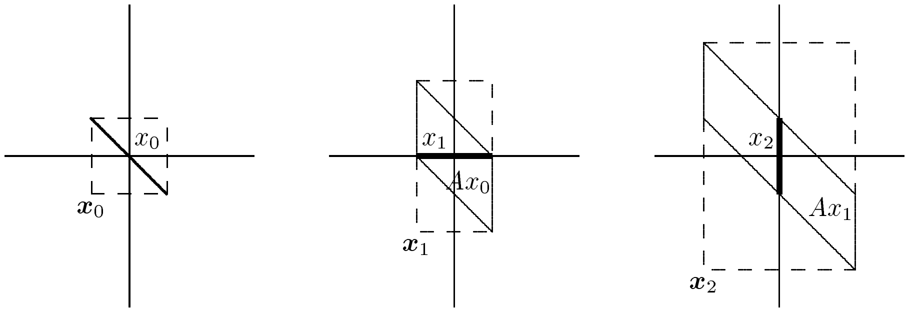


**Figure 2**: The wrapping effect

The volume of a set changes under the transformation (6) by a factor of $|\det A|$, the product of the absolute values of the eigenvalues of $A$. Therefore we can assess the amount of overestimation per step by monitoring the *overestimation factor*

$$q = \operatorname{vol} \boldsymbol{x}_{l+1} / |\det A| \operatorname{vol} \boldsymbol{x}_l, \tag{7}$$

where $\operatorname{vol}[\underline{x}, \bar{x}] = (\bar{x}_1 - \underline{x}_1) \cdots (\bar{x}_n - \underline{x}_n)$ is the volume of a box $\boldsymbol{x} = [\underline{x}, \bar{x}]$ in $\mathbb{R}^n$.

## 2 Simplices, parallelepipeds, hyperoctahedra and ellipsoids

It is clear that interval boxes are not sufficiently variable in shape to model all affine transforms of an initial box; to eliminate the wrapping effect for the case of the iteration (6) or any iteration

$$x_{l+1} = A_l x_l + b_l \tag{8}$$

we need to use a class of enclosure sets which is closed under affine transformations and still simple enough to be described with little effort.

A natural class of such sets is the class of all polytopes, but working with general polytopes is time-consuming. The simplest polytopes, the simplices, work and have indeed be used for enclosures (Conradt [3], Jansson [10], Rump [22]). Parallelepipeds are another natural affinely closed class of polytopes, and have been used in all recent enclosure methods for ordinary differential equations (Eijgenraam [4], Lohner [14], [15]); their special advantage is that the boxes belongs to this class, hence they can be handled easily by interval methods. Hyperoctahedra form another simple affinely closed class, but have received no attention so far. See also the survey article Nickel [20].

A very important affinely closed class consists of the set of (hyper-)ellipsoids. Since ellipsoids arise naturally in many applications, notably as confidence regions of stochastic variables, and since they are invariant under a much bigger symmetry group than the other classes mentioned, they have received considerable attention around 1968 (Kahan [11], Jackson [8], [9]). But in the absense of a simple way of calculating with ellipsoids — the Minkowski sum of two ellipsoids has a complicated shape and is expensive to enclose by another ellipsoid — they have not been used in actual codes for rigorous computation. However, as pointed out by a referee, *approximate* calculations with ellipsoids have found successful applications to several problems in linear control theory, see Kurzhanski & Vályi [13], Ovseevich & Chernousko [21] and the references there. In particular, the discrete case of the *reachable set problem* amounts to finding enclosures of (8) where $A_l$ is known exactly and $x_0$ and the $b_l$ vary in specified sets.

As we shall see, the use of ellipsoids *in combination with* intervals allows the efficient handling of ellipsoids and makes them again an interesting competitor for the best representation of multivariate enclosure sets.

In this section we represent ellipsoids in a form which, for different norms, would give other families of affine closed classes of convex sets which can be easily represented on a computer; parallelepipeds and hyperoctahedra become cases of this concept when the 2-norm is replaced by the $\infty$-norm or the 1-norm, respectively. We show how the chosen representation allows a very simple implementation of the iteration (8) without overestimation. But there are problems of numerical stability which often lead to a blow-up of the enclosures when implemented in (outward rounded) finite precision arithmetic. The next section is therefore devoted to the derivation of a stable version of the enclosure method.

In the following, we use freely interval arithmetical extensions of vector and matrix operations (cf. NEUMAIER [19]), writing interval quantities in boldface types. $\|\cdot\|$ denotes the Euclidian vector and matrix norm.

An *ellipsoid* is a set of the form

$$E(z, L, r) := \{z + L\xi \mid \xi \in \mathbb{R}^n, \|\xi\| \le r\}, \tag{9}$$

with fixed $z \in \mathbb{R}^n$, $L \in \mathbb{R}^{n\times n}$ (lower triangular), $r \in \mathbb{R}_+$. Thus ellipsoids are the images of a ball $\{\xi \in \mathbb{R}^n \mid \|\xi\| \le r\}$ under affine mappings $\xi \to z + L\xi$. Since $\|\xi\|$ is invariant under orthogonal transformations, lower triangular matrices are sufficient to represent all ellipsoids (including degenerate cases); for other norms, $L$ would have to be unrestricted.

**Proposition.** *Suppose* (8) *holds with nonsingular* $A_l$. *Then*

$$x_l \in E(z_l, L_l, r) \qquad \Longrightarrow \qquad x_{l+1} \in E(z_{l+1}, L_{l+1}, r), \tag{10}$$

*where*

$$z_{l+1} = A_l z_l + b_l, \qquad L_{l+1} = A_l L_l. \tag{11}$$

*Proof.* If $x_l = z_l + L_l\xi$ then $x_{l+1} = A_l(z_l + L_l\xi) + b_l = (A_l z_l + b_l) + (A_l L_l)\xi = z_{l+1} + L_{l+1}\xi$. $\square$

While this proposition seems to handle completely all linear dynamical systems, this only holds under the assumption of exact arithmetic. In finite precision arithmetic, one must account for the rounding errors in the formation of $z_{l+1}$ and $L_{l+1}$.

The simplest way to do this is to allow the $z_l$ and $L_l$ themselves to have interval components, and to compute (11) by outward rounding. But the recurrence for the $z_l$ is precisely the same as for the $x_l$, so the wrapping effect appears in the $z_l$ and magnifies rounding errors exponentially; and the same holds for each column of the $L_l$. Thus the proposition is useless for actual computation. To avoid excessive overestimation we must therefore keep $z_l$ and $L_l$ real, and account for the rounding errors made by increasing the ellipsoid radius $r$.

## 3 The enclosure of affine transforms of ellipsoids

We study the adaption of $r$ in a slightly more general setting motivated by the fact that in many realistic situations the entries of the $A_l$ are not precisely known and are allowed to vary in intervals. Thus we want to study the transformation of an ellipsoid under all affine mappings

$$x \to Ax + b \qquad \big(A \in \boldsymbol{A}, \quad b \in \boldsymbol{b}\big), \tag{12}$$

where $\boldsymbol{A}$ is an $n \times n$ interval matrix and $\boldsymbol{b}$ an $n$-dimensional interval vector.

**Theorem 1.** *Suppose that $x \in E(z, L, r)$. Select arbitrary $\bar{z} \in \mathbb{R}^n$ and nonsingular $\bar{L} \in \mathbb{R}^{n\times n}$, and let*

$$\tilde{r} := \|\bar{L}^{-1}(\boldsymbol{A}z + \boldsymbol{b} - \bar{z})\| + \|\bar{L}^{-1}\boldsymbol{A}L\| r. \tag{13}$$

*Then*

$$Ax + b \in E(\bar{z}, \bar{L}, \tilde{r}) \qquad \text{for all} \quad A \in \boldsymbol{A}, \quad b \in \boldsymbol{b}. \tag{14}$$

*Proof.* By assumption, $x = z + L\xi$ for some $\xi$ with $\|\xi\| \le r$. Therefore

$$\begin{aligned} Ax + b &= A(z + L\xi) + b = Az + b + AL\xi \\ &= \bar{z} + (Az + b - \bar{z} + AL\xi) = \bar{z} + \bar{L}\bar{\xi}, \end{aligned}$$

where

$$\bar{\xi} = \bar{L}^{-1}(Az + b - \bar{z}) + \bar{L}^{-1}AL\xi \in \bar{L}^{-1}(\boldsymbol{A}z + \boldsymbol{b} - \bar{z}) + (\bar{L}^{-1}\boldsymbol{A}L)\xi$$

satisfies

$$\|\bar{\xi}\| \le \|\bar{L}^{-1}(\boldsymbol{A}z + \boldsymbol{b} - \bar{z})\| + \|\bar{L}^{-1}\boldsymbol{A}L\| \, \|\xi\| \le \tilde{r}. \quad \square$$

Of course, if we want to have a *good* enclosure we must choose $\bar{z}$ and $\bar{L}$ properly. When $A$ and $b$ are precisely known (so that we can use ordinary arithmetic) we see that the choice $\bar{z} = Az + b$, $\bar{L} = AL$, $\bar{r} = r$ is consistent with (13) and recovers the previous proposition. In the presence of roundoff error or other uncertainties, this suggests the choice

$$\bar{z} = \mathrm{mid}(\boldsymbol{A}z + \boldsymbol{b}), \qquad \bar{L} = \mathrm{mid}(\boldsymbol{A}L), \tag{15}$$

where the interval expressions are computed with outward rounding to take correctly care of rounding errors. However, (13) shows that this will be disastrous when $\bar{L}$ is ill-conditioned since it blows up the radius $\bar{r}$ to a very large number. And, unfortunately, repeated iteration of (12) using (15) yields always, sooner or later, such ill-conditioned $L = (\mathrm{mid}\, A)^l L_0$, unless all eigenvalues of $\mathrm{mid}\, A$ have the same absolute value. Therefore, we must look for a regularized version which keeps $\bar{r}$ reasonably small.

Let us introduce some notation. For a vector $x$, we denote by $|x|$ the vector with components $|x_i|$. $A_{i\bullet}$ denotes the $i$-th row of a matrix $A$, and $\nu(A)$ denotes the *hybrid norm* of $A$ (Neumaier [18]), i.e. the vector with components $\nu_i(A) := \|A_{i\bullet}\|$. It is easy to see that

$$|Ax| \le \nu(A)\|x\|,$$

and

$$\|\nu(A)\| = \sqrt{\mathrm{tr}\, A^T A} = \|A\|_{\mathrm{F}} \quad (\ge \|A\|)$$

is the *Frobenius* (or *Schur*) norm of $A$. We write

$$\bar{z} := \mathrm{mid}(\boldsymbol{A}z + \boldsymbol{b}), \qquad d := |\boldsymbol{A}z + \boldsymbol{b} - \bar{z}|, \tag{16}$$

$$B := \mathrm{mid}(\boldsymbol{A}L), \qquad d' := \nu(\boldsymbol{A}L - B). \tag{17}$$

From (13) and monotony, we get, for any nonsingular diagonal matrix $D$,

$$\begin{aligned}\tilde{r} &\le \|\bar{L}^{-1}D\|\,\|D^{-1}(\boldsymbol{A}z + \boldsymbol{b} - \bar{z})\| + \big(\|\bar{L}^{-1}B\| + \|\bar{L}^{-1}D\|\,\|D^{-1}(\boldsymbol{A}L - B)\|_{\mathrm{F}}\big)r\\ &\le \|\bar{L}^{-1}D\|\,\|D^{-1}d\| + \big(\|\bar{L}^{-1}B\| + \|\bar{L}^{-1}D\|\,\|D^{-1}d'\|\big)r,\end{aligned}$$

hence

$$\tilde{r} \le \|\bar{L}^{-1}B\|r + \|\bar{L}^{-1}D\|q \tag{18}$$

where

$$q = \|D^{-1}d\| + \|D^{-1}d'\|r. \tag{19}$$

In (18) and (19) all intervals are eliminated, making it more suitable for analysis. We want to choose $\bar{L}$ such that the volume of the new ellipsoid is small. Now this volume is proportional to $\tilde{r}^n \det \bar{L}$, and the next result shows how a nearly optimal $\bar{L}$ can be found. Since a change of $\bar{r}$ is equivalent to rescaling $L$ we aim at a radius $\bar{r} \approx 1$, thus having small $\bar{L}$ when $rL$ and hence $rB$ are small.

**Theorem 2.** *Suppose that*

$$r^2BB^T + q^2DD^T = \bar{L}\bar{L}^T, \tag{20}$$

$$\bar{r} = \|\bar{L}^{-1}B\|r + \|\bar{L}^{-1}D\|q. \tag{21}$$

*Then $\bar{r} \le 2$, and for arbitrary nonsingular $\tilde{L}$ we have*

$$\Big(\|\tilde{L}^{-1}B\|r + \|\tilde{L}^{-1}D\|q\Big)^n |\det \tilde{L}| \ge |\det \bar{L}|. \tag{22}$$

*In particular, choosing $\bar{L}$ by* (20) *implies optimality of $\bar{L}$ within a factor of* 2 *for the radius.*

*Proof.* We use the abbreviations $U = \bar{L}^{-1}B$, $V = \bar{L}^{-1}D$ to rewrite (20), (21) as

$$r^2UU^T + q^2VV^T = I,$$

$$\|U\|r + \|V\|q = \bar{r}.$$

With $W = \tilde{L}^{-1}\bar{L}$, the left hand side of (22) becomes

$$\Big(\|WU\|r + \|WV\|q\Big)^n |\det \bar{L}|/|\det W|.$$

Now

$$\|W(rU,qV)\|^2 = \|W(rU,qV)(rU,qV)^TW^T\| = \|WW^T\| = \|W\|^2,$$

hence

$$|\det W| \le \|W\|^n \le \|W(rU,qV)\|^n \le \Big(r\|(WU,0)\| + q\|(0,WV)\|\Big)^n = \Big(r\|WU\| + q\|WV\|\Big)^n$$

so the left hand side of (22) is $\ge |\det \bar{L}|$.

On the other hand,

$$\|U\| \le r^{-1}, \qquad \|V\| \le q^{-1},$$

hence $\bar{r} \le 2$. $\square$

Of course, this optimality result is based on the upper bound (18) which is not exact, but which is easily computable. The diagonal matrix $D$ can still be chosen freely, and its optimal choice is unsettled. However, a natural choice for $D$ comes from balancing the contributions of the components to $q$ in (19), and suggests that we take

$$D = \mathrm{Diag}(d_1 + d_1' r, \ldots, d_k + d_k' r). \tag{23}$$

(This forces $D^{-1}d + D^{-1}d'r = (1, \ldots, 1)^T$, leading to $\sqrt{n} \le q \le \sqrt{2n}$.) As one can see from (16) and (17), $D$ will have entries of the order of the radii of $\boldsymbol{A}$ and $\boldsymbol{b}$, so that the contributions in (20), (21) to $\bar{L}$ and $\bar{r}$ remain small as long as $B$ is well conditioned. Thus $\bar{L}\bar{L}^T \approx r^2 BB^T$ remains small when $rB$ was small and $Q = r\bar{L}^{-1}B$ satisfies $Q^TQ \approx I$ so that $\bar{r} \approx \|Q\| \approx 1$ and everything is stable.

The matrix $\bar{L}$ is determined by (20) only upto an orthogonal transformation, and is best chosen as a Cholesky factor of the left hand side of (20). One sees that because of the regularizing diagonal term in (20), $\bar{L}$ will be well-conditioned even when $B = \mathrm{mid}(\boldsymbol{A}L)$ is ill-conditioned or singular; so the instability mentioned earlier has been removed successfully. However, the formation of $\bar{L}$ by (20) directly is numerically unstable when $B$ is ill-conditioned, and we must proceed in a slightly different way. The first possibility is to add to the matrix $N = r^2BB^T + q^2DD^T$ extra diagonal terms to force it positive definite (e.g. $N_{ii} \leftarrow N_{ii}(1 + \epsilon^{1/2})$). Another possibility uses the fact that the matrix $M := (rB, qD) \in \mathbb{R}^{n\times 2n}$ satisfies $MM^T = r^2BB^T + q^2DD^T = \bar{L}\bar{L}^T$; hence we can obtain $\bar{L}$ from an $LQ$-factorization (= transposed $QR$ factorization) of $M$. Small diagonal entries in $\bar{L}$ due to roundoff can be corrected by replacing the diagonal entries $\bar{L}_{ii}$ with $\bar{L}_{ii} + \mathrm{sgn}\,\bar{L}_{ii} \cdot \eta \cdot \nu_i(M)$, where $\eta = n^{3/2}\epsilon$ and $\epsilon$ denotes the machine accuracy.

Collecting together the various formulas we find the following

**Algorithm.** *(Ellipsoid propagation algorithm)*

*Purpose*: Enclose the transformation of the ellipsoid $\{z + L\xi \mid \|\xi\| \le r\}$ ($L$ lower triangular)
by $x \to \boldsymbol{A}x + \boldsymbol{b}$ within the ellipsoid $\{\bar{z} + \bar{L}\bar{\xi} \mid \|\bar{\xi}\| \le \bar{r}\}$ ($\bar{L}$ lower triangular)

! The rounding mode for computing each left hand side is indicated by
! $\supseteq$ (outward), $\approx$ (approximate), $\ge$ (upwards)

$\bar{\boldsymbol{z}} \supseteq \boldsymbol{A}z + \boldsymbol{b}, \quad \bar{z} \approx \mathrm{mid}\,\bar{\boldsymbol{z}}, \quad d \ge |\bar{\boldsymbol{z}} - \bar{z}|$

$\boldsymbol{B} \supseteq \boldsymbol{A}L, \quad B \approx \mathrm{mid}\,\boldsymbol{B}, \quad d' \ge \nu\big(\boldsymbol{B} - B\big)$

$D \approx \mathrm{Diag}(d_1 + d_1'r, \ldots, d_n + d_n'r)$

$q \ge \|D^{-1}d\| + \|D^{-1}d'\|r$

$M \approx (rB, qD)$

$\bar{L}\bar{Q} \approx M$

For $i = 1, \ldots, n$, change $\bar{L}_{ii}$ to $\bar{L}_{ii} + \mathrm{sgn}\,\bar{L}_{ii} \cdot \eta \cdot \nu_i(M)$

Enclose the solution of $\bar{L}C = B$ by $\boldsymbol{C}$ and compute an upper bound $\gamma$
on the largest singular value of $\boldsymbol{C}$

Enclose the solution of $\bar{L}C = D$ by $\boldsymbol{C}$ and compute an upper bound $\delta$
on the largest singular value of $\boldsymbol{C}$

$\bar{r} \ge \gamma r + \delta q.$

# 4 Bounds for the range over an ellipsoid

The wrapping effect does not only occur for linear transformations, but even more when one transforms a set by a nonlinear transformation. In this case, even exact arithmetic does not allow optimal enclosures since nonlinear mappings generally distort the shapes of ellipsoids (or simplices, parallelepipeds and hyperoctahedra), and clearly the amount of unavoidable wrapping increases with the amount of nonlinearity present in the set. In particular, this implies that one will be able to obtain realistic enclosures of general nonlinear transformations only for sufficiently narrow sets (where nonlinearities contribute in the order of the squared diameter only). On the other hand, realistic enclosures of an image of a big set can be obtained only for mappings which are nearly linear.

We now show that the techniques of the previous section suffice to enclose nonlinear images of narrow ellipsoids by another ellipsoid. The key is the observation that for any $C^1$-function $F : \mathbb{R}^n \to \mathbb{R}^m$ (the dimensions need not be equal!) which is defined by a Lipschitz expression (a notion explained in NEUMAIER [19]), any bounded set $E \subseteq \mathbb{R}^n$ and all *centers* $z \in E$, one can find a *slope matrix* $\boldsymbol{A}$ such that we can represent $F$ as *centered form*

$$x \in E \qquad \Longrightarrow \qquad F(x) = F(z) + \tilde{A}(x-z) \quad \text{for some} \quad \tilde{A} \in \boldsymbol{A}. \tag{24}$$

Thus if we know an enclosure

$$\boldsymbol{b} \supseteq F(z) \tag{25}$$

of $F(z)$ which accounts for roundoff, we find that

$$F(x) \in \bigcup \{\boldsymbol{A}x + \boldsymbol{b} \mid x \in E\},$$

so that the theory of the previous section applies when $E$ is an ellipsoid.

The only difficulty is the computation of the slope matrix $\boldsymbol{A}$. An optimal computation of $\boldsymbol{A}$ using the full ellipsoid information seems difficult; therefore we compute $\boldsymbol{A}$ using an interval enclosure for the ellipsoid $E$ by a box $\boldsymbol{x}$; then recursive techniques (KRAWCZYK & NEUMAIER [12], NEUMAIER [19]) allow the computation of the slope matrix $\boldsymbol{A}$ for the box $\boldsymbol{x}$, and this is obviously also a slope matrix for the subset $E$ of $\boldsymbol{x}$. The optimal box is given by

**Theorem 3.** *The smallest box containing the ellipsoid $E(z, L, r)$ is*

$$\boldsymbol{x} := \Box E(z, L, r) = z + [-r, r]\nu(L). \tag{26}$$

*Proof.* The $i$-th component of $\boldsymbol{x}$ is given by the hull of all $x_i = (z + L\xi)_i$ with $\|\xi\| \le r$. Now, by the Cauchy-Schwarz inequality,

$$|x_i - z_i| = |(L\xi)_i| = |L_{i\bullet}\xi| \le \|L_{i\bullet}\|_2 \|\xi\|_2 = \nu_i(L)\|\xi\|_2 \le \nu_i(L) r,$$

and clearly the bound can be attained with either sign of $x_i - z_i$. Hence formula (26). □

In practice, when doing a sequence of nonlinear transformations, it is important to update the box containing the new ellipsoid $E(\bar{z}, \bar{L}, \bar{r})$ by

$$\bar{\boldsymbol{x}} = \big(\boldsymbol{A}(\boldsymbol{x} - z) + \boldsymbol{b}\big) \cap \big(\bar{z} + [-\bar{r}, \bar{r}]\nu(\bar{L})\big). \tag{27}$$

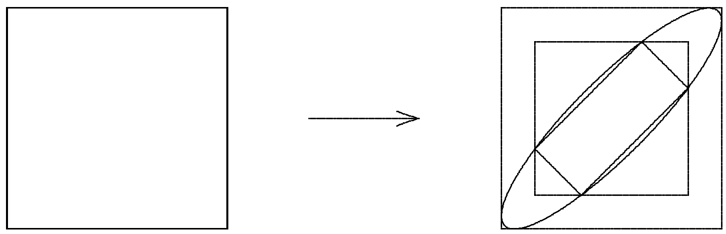

**Figure 3**: Updating the box by (27)

This often eliminates the ends of long and thin ellipsoids which (due to overestimation) no longer contain points of the (iterated) image of the original set; these ends would inflate the box (26) considerably.

This is particularly relevant when $\boldsymbol{A} \geq 0$, where it is well known that the formula $\bar{\boldsymbol{x}} = \boldsymbol{A}(\boldsymbol{x} - z) + \boldsymbol{b}$ gives optimal boxes (though large volume overestimation).

Another trick is often important to guarantee reasonable results when the nonlinear transformation $F(x, \lambda)$ depends on a parameter vector $\lambda$ which varies in a box $\boldsymbol{\lambda}$ (or an ellipsoid). Treating the $\boldsymbol{\lambda}_i$ as interval constants in the centered form (24) often leads to significant wrapping, especially after a sequence of several transformations involving $\boldsymbol{\lambda}$. In this case, the correct way to treat these parameters is by extending the state vector $x$ to $x' = \binom{x}{\lambda}$ and enclosing it by higher-dimensional ellipsoids. While this incurs in the first step a volume overestimation factor of the unit ball volume, this factor will often have been gained after a few more steps, due to reduced wrapping. For wide intervals in $\boldsymbol{\lambda}$, however, this need no longer be the case, and one may have to resort to methods of global optimization (Hansen [6]).

**Examples.** We iterate (27) for a two-dimensional discrete dynamical system

$$x_{l+1} = A_l x_l + b_l \tag{28}$$

where $A_l \in \boldsymbol{A}$, $b_l \in \boldsymbol{b} = 10^{-12}\binom{[-1,1]}{[-1,1]}$, $x_0 \in \boldsymbol{x}_0 = \binom{[-1,1]}{[-1,1]}$, and various matrices $\boldsymbol{A}$. Thus we start with a square of side 2, and assume small vagueness in the vectors $b_l$ but (except in case 4) large vagueness in the coefficients of $A_l$.

In each case we list the side $\beta_l$ of a cube with the same volume as $\boldsymbol{x}_l$ and the side $\gamma_l$ of a cube with the same volume as the ellipsoid. For comparison we also iterate in naive interval arithmetic

$$\boldsymbol{y}_{l+1} = \boldsymbol{A}\boldsymbol{y}_l + \boldsymbol{b}_l,$$

starting with $\boldsymbol{y}_0 = \boldsymbol{x}_0$, and record the quotient $\beta_l'/\beta_l$, where $\beta_l'$ is the side of a cube with the same volume as $\boldsymbol{y}_l$. The computations were done with the CALCULUS system [7] of Siegfried Rump, whose help with the examples is gratefully acknowledged. On the machine used, $\epsilon = 10^{-17}$.

*Case* 1: $\qquad \boldsymbol{A} = \begin{pmatrix} \boldsymbol{p} & \boldsymbol{p} \\ -\boldsymbol{p} & \boldsymbol{p} \end{pmatrix}, \qquad \boldsymbol{p} = \left[\frac{4}{10}, \frac{5}{10}\right].$

| $l$ | $\beta_l$ | $\gamma_l$ | $\beta_l'/\beta_l$ |
|---|---|---|---|
| 10 | 1.33E − 01 | 6.64E − 02 | 1.51E + 01 |
| 20 | 6.22E − 03 | 3.11E − 03 | 3.22E + 02 |
| 30 | 2.92E − 04 | 1.46E − 04 | 6.86E + 03 |
| 40 | 1.37E − 05 | 6.84E − 06 | 1.47E + 05 |
| 50 | 6.42E − 07 | 3.21E − 07 | 3.12E + 06 |
| 60 | 3.01E − 08 | 1.51E − 08 | 6.65E + 07 |
| 70 | 1.43E − 09 | 7.11E − 10 | 1.41E + 09 |
| 80 | 7.69E − 11 | 3.85E − 11 | 2.61E + 10 |
| 90 | 1.39E − 11 | 6.92E − 12 | 1.45E + 11 |
| 100 | 1.09E − 11 | 5.44E − 12 | 1.84E + 11 |

The ellipsoids contract until about the size of rad $\boldsymbol{b}$; naive interval arithmetic does not contract.

*Case* 2: $\qquad \boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -1 & -\boldsymbol{p} \end{pmatrix}, \qquad \boldsymbol{p} = \left[1, \frac{10}{9}\right].$

| $l$ | $\beta_l$ | $\gamma_l$ | $\beta_l'/\beta_l$ |
|---|---|---|---|
| 10 | 1.21E + 01 | 5.79E + 00 | 3.02E + 01 |
| 20 | 5.60E + 01 | 2.42E + 01 | 1.31E + 03 |
| 30 | 2.05E + 02 | 9.79E + 01 | 7.19E + 04 |
| 40 | 7.09E + 02 | 3.69E + 02 | 4.18E + 06 |
| 50 | 2.69E + 03 | 1.35E + 03 | 2.21E + 08 |
| 60 | 1.09E + 04 | 5.02E + 03 | 1.11E + 10 |
| 70 | 3.99E + 04 | 1.89E + 04 | 6.03E + 11 |
| 80 | 1.36E + 05 | 6.86E + 04 | 3.56E + 13 |
| 90 | 4.85E + 05 | 2.45E + 05 | 2.01E + 15 |
| 100 | 1.83E + 06 | 8.76E + 05 | 1.08E + 17 |

All matrices in $A$ have determinant 1, so the dynamical system is volume preserving. However, the ellipsoid volume grows exponentially, with an average factor of $(8.76 \cdot 10^5)^{1/100} \approx 1.147$, but much less than the boxes in naive calculation.

*Case* 3: $\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 1 & \boldsymbol{p} \end{pmatrix}$, $\boldsymbol{p} = [1, \frac{10}{9}]$.

| $l$ | $\beta_l$ | $\gamma_l$ | $\beta'_l/\beta_l$ |
|---|---|---|---|
| 10 | 3.64E + 02 | 7.78E + 03 | 1.00E + 00 |
| 20 | 7.32E + 04 | 8.98E + 08 | 1.00E + 00 |
| 30 | 1.48E + 07 | 1.04E + 14 | 1.00E + 00 |
| 40 | 2.96E + 09 | 1.20E + 19 | 1.00E + 00 |
| 50 | 5.95E + 11 | 1.39E + 24 | 1.00E + 00 |
| 60 | 1.20E + 14 | 1.60E + 29 | 1.00E + 00 |
| 70 | 2.41E + 16 | 1.84E + 34 | 1.00E + 00 |

Again the system is volume preserving but we get boxes exploding with a factor $\approx 1.714$ per iteration and ellipsoids exploding even faster. The intersection in (27) is here very effective. Since $A \geq 0$, naive interval calculation gives optimal boxes (but not optimal volumes), explaining $\beta'_l = \beta_l$. In this example, the ellipsoids are clearly not useful.

*Case* 4: $\boldsymbol{A} = \begin{pmatrix} 8.0 & -2.6 \\ 9.0 & -2.8 \end{pmatrix}$.

a) with direct computation of $\bar{L}$ from (20):

| $l$ | $\beta_l$ | $\gamma_l$ | $\beta'_l/\beta_l$ |
|---|---|---|---|
| 10 | 3.96E + 07 | 2.36E + 05 | 1.05E + 03 |
| 20 | 3.87E + 14 | 2.30E + 12 | 2.53E + 06 |
| 30 | 3.78E + 21 | 2.25E + 19 | 6.07E + 09 |
| 40 | 3.69E + 28 | 2.20E + 26 | 1.46E + 13 |
| 50 | 3.60E + 35 | 2.15E + 33 | 3.52E + 16 |

Again the system is volume preserving, but with eigenvalues 0.2 and 5 which cause the ellipsoids to flatten rapidly. The growth rate of the box volumes is about 2/3 of that of the naive calculation, a consequence of the diagonal needle shape of the ellipsoids.

b) with $\bar{L}$ computed as stated in the algorithm:

| $l$ | $\beta_l$ | $\gamma_l$ | $\beta'_l/\beta_l$ |
|---|---|---|---|
| 1 | $2.24\mathrm{E}+01$ | $1.42\mathrm{E}+00$ | $1.00\mathrm{E}+00$ |
| 2 | $1.30\mathrm{E}+02$ | $1.42\mathrm{E}+00$ | $1.88\mathrm{E}+00$ |
| 3 | $6.51\mathrm{E}+02$ | $1.42\mathrm{E}+00$ | $4.08\mathrm{E}+00$ |
| 4 | $3.26\mathrm{E}+03$ | $1.42\mathrm{E}+00$ | $8.89\mathrm{E}+00$ |
| 5 | $1.63\mathrm{E}+04$ | $1.42\mathrm{E}+00$ | $1.94\mathrm{E}+01$ |
| 6 | $8.14\mathrm{E}+04$ | $1.42\mathrm{E}+00$ | $4.22\mathrm{E}+01$ |
| 7 | $4.07\mathrm{E}+05$ | $1.42\mathrm{E}+00$ | $9.19\mathrm{E}+01$ |
| 8 | $2.04\mathrm{E}+06$ | $1.42\mathrm{E}+00$ | $2.00\mathrm{E}+02$ |
| 9 | $1.03\mathrm{E}+07$ | $1.43\mathrm{E}+00$ | $4.34\mathrm{E}+02$ |
| 10 | $5.42\mathrm{E}+07$ | $1.51\mathrm{E}+00$ | $8.92\mathrm{E}+02$ |
| 20 | $1.27\mathrm{E}+18$ | $1.28\mathrm{E}+10$ | $8.99\mathrm{E}+02$ |
| 30 | $2.97\mathrm{E}+28$ | $5.79\mathrm{E}+20$ | $8.99\mathrm{E}+02$ |

Here the ellipsoid volume hardly grows initially, but after a while the volumes grow at the same rate as in the naive calculation.

The rapid blow-up of the ellipsoids in case 4, where the ambiguity in (28) is only small, $O(10^{-12})$, suggests that the method proposed does not yet choose nearly optimal enclosing ellipsoids, so that there is further room for improvement.

# 5 Application: Stability regions

A discrete dynamical system

$$x_{l+1} = F(x_l) \tag{29}$$

is called *stable* in a region $E$ when, for any initial state $x_0 \in E$, the state vectors $x_l$ defined by (29) remains bounded for all $l$. A sufficient condition for this is — when $E$ is bounded — that

$$F(x) \in E \quad \text{for all} \quad x \in E. \tag{30}$$

If $F$ is also continuous, then Brouwer's fixed point theorem guarantees a fixed point of $F$ under these conditions.

Clearly we can verify (30) by the techniques of the present paper when $E = (z, L, r)$ is an ellipsoid. Since the image is to be enclosed by the same ellipsoid, the natural choice here is $\bar{z} = z$ and $\bar{L} = L$. Then, using (24) and (13), a sufficient condition for (30), hence for stability, is

$$\left\|L^{-1}(\boldsymbol{A}z + \boldsymbol{b} - z)\right\| + \|L^{-1}\boldsymbol{A}L\|r \le r. \tag{31}$$

Since an initial enclosure is here not given we are free to choose $z$ and $L$ arbitrarily, and we want to choose it to make the verification of (31) most likely. Since we know that $E$ must contain a fixed point, we choose $z$ as an approximation to such a fixed point; then $F(z) \approx z$ implies that the first norm in (31) will be small.

The second norm in (31) must be made $< 1$. Since in the limit $r \to 0$, where the ellipsoid shrinks to the point $z$, the slope matrix tends to $F'(z)$ (assuming $F$ to be differentiable), we have $F'(z) \in \boldsymbol{A}$. Thus we try to make $\|L^{-1}F'(z)L\|$ small, and in particular $< 1$.

Since the norm of a matrix is an upper bound for the spectral radius, and the latter is invariant under similarity transformations, $F'(z)$ must have spectral radius $\rho\big(F'(z)\big) < 1$, i.e., $z$ must be an attractive fixed point. In this case we can transform $F'(z)$ to a block-diagonal form by the modal matrix $S$ whose columns are eigenvectors of real eigenvalues, or real and complex part of eigenvectors to complex eigenvalues. (This assumes that $F'(z)$ is nondefective. In the nearly defective case one must consider higher-dimensional invariant subspaces.) The diagonal blocks of $S^{-1}F(z)S$ will have the form $(\lambda)$ for real eigenvalues and $\left(\begin{smallmatrix} \operatorname{Re}\lambda & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & \operatorname{Re}\lambda \end{smallmatrix}\right)$ for complex eigenvalues. Thus if we take for $L$ an approximation to $S$ (silently dropping the assumption of $L$ being lower triangular) we find that $L^{-1}F'(z)L$ and hence $L^{-1}\boldsymbol{A}L$ (for small $r$; the entries of $\boldsymbol{A}$ have radius $O(r)$) are approximately block-diagonal. Now let

$$\boldsymbol{C} \supseteq L^{-1}(\boldsymbol{A}L), \qquad C \approx \operatorname{mid}\boldsymbol{C}, \qquad R \geq |\boldsymbol{C} - C|. \tag{32}$$

By construction, $C$ is nearly block-diagonal, and $R = O(r)$. The form of the diagonal blocks implies that $CC^T$ is approximately diagonal, with diagonal entries $\approx |\lambda|^2 < 1$. Thus $\|CC^T\|_\infty$ will approximately equal the square of the spectral radius of $F'(z)$, and since

$$\|L^{-1}\boldsymbol{A}L\| \leq \|\boldsymbol{C}\| \leq \|C\| + \|R\| \leq \sqrt{\|C^TC\|_\infty} + \sqrt{\operatorname{tr} R^TR}$$

we get the *stability condition*

$$\|L^{-1}(\boldsymbol{A}z + \boldsymbol{b} - z)\|_2 + \Big(\sqrt{\|C^TC\|_\infty} + \sqrt{\operatorname{tr} R^TR}\Big)r \leq r \tag{33}$$

as verifiable sufficient condition for stability in $E(z, L, r)$. Moreover, by our analysis, the left hand side is $\rho\big(F'(z)\big)r + O(r^2) < r$ for sufficiently small $r$, so that (33) verifies some stability region around any attractive fixed point. By checking (33) for various $r$ (using a bisection procedure in $[0, r]$) one can find the maximal radius in which stability can be verified.

# 6 Application: Confidence regions

Let $x$ be an $n$-dimensional random vector with mean $z$, covariance matrix $\Sigma$ and Gaussian distribution. We want to find a confidence region $\bar{E}$ for the transformed random vector

$$\bar{x} = F(x)$$

which contains $\bar{x}$ with specified probability $\alpha$ (or higher).

Since $(x-z)^T\Sigma^{-1}(x-z)$ is $\chi^2(n)$ distributed, one can compute a radius $r_\alpha$ such that

$$(x-z)^T\Sigma^{-1}(x-z) \leq r_\alpha^2 \qquad \text{with probability } \alpha. \tag{34}$$

If we use a Cholesky factorization $LL^T$ of $\Sigma$ and introduce $\xi = L^{-1}(x-z)$ we find $x = z + L\xi$ and $\xi^T\xi = (x-z)^TL^{-T}L^{-1}(x-z) = (x-z)^T\Sigma^{-1}(x-z)$. Hence (34) can be rewritten as

$$x \in E(z, L, r_\alpha) \qquad \text{with probability } \alpha. \tag{35}$$

Therefore, the image of $E(z,L,r_\alpha)$ under the transformation $F$ will contain $\bar{x}$ with probability $\alpha$, and the enclosing ellipsoid will therefore satisfy

$$\bar{x} \in \bar{E}(\bar{z},\bar{L},\bar{r}_\alpha) \qquad \text{with probability } \geq \alpha. \tag{36}$$

Thus we have a rigorous confidence region for $\bar{x}$ to the confidence level $\alpha$, and since the probability is at least $\alpha$, overestimation leads to an error on the safe side.

## References


1. F. L. Alvarado. Sparse W-matrix for interval arithmetic. Manuscript (1990).

2. F. L. Alvarado. Sparsity preservation in matrix interval solutions. Manuscript (1990).

3. J. Conradt. Ein Intervallverfahren zur Einschließung des Fehlers einer Näherungslösung bei Anfangswertaufgaben für Systeme von gewöhnlichen Differentialgleichungen. Freiburger Intervall Berichte 80/1, Inst f. Angew. Math., Univ. Freiburg (1980).

4. P. Eijgenraam. *The Solution of Initial Value Problems using Interval Arithmetic.* Math. Centre Tracts 144, Amsterdam (1981).

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

6. E. Hansen. *Global Optimization using Interval Analysis.* Marcel Dekker, New York (1992).

7. D. Husung and S. M. Rump. CALCULUS. In *Wissenschaftliches Rechnen mit Ergebnisverifikation* (U. Kulisch, ed.). Vieweg, Berlin (1989).

8. L. W. Jackson. A comparison of ellipsoidal and interval arithmetic error bounds, numerical solutions of nonlinear problems (notice). *SIAM Rev.* **11** (1969), 114.

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

10. C. Jansson. A geometric approach for computing a posteriori error bounds for the solution of a linear system. *Computing* **47** (1991), 1–9.

11. W. M. Kahan. Circumscribing an ellipsoid about the intersection of two ellipsoids. *Can. Math. Bull.* **11** (1968), 437–441.

12. R. Krawczyk and A. Neumaier. Interval slopes for rational functions and associated centered forms. *SIAM J. Numer. Anal.* **22** (1985), 604–616.

13. A. B. Kurzhanski and I. Vályi. Ellipsoidal techniques for dynamic systems: The problem of control synthesis. *Dynamics and Control* **1** (1991), 357–378.

14. R. Lohner. Enclosing the solutions of ordinary initial- and boundary-value problems. In *Computerarithmetic* (E. Kaucher, U. Kulisch, and Ch. Ullrich, eds.), pp. 255–286. Teubner, Stuttgart (1987).

15. R. Lohner. *Einschließung der Lösung gewöhnlicher Anfangs- und Randwertaufgaben und Anwendungen*. Dissertation, Univ. Karlsruhe (1988).

16. R. E. Moore. *Interval Arithmetic and Automatic Error Analysis in Digital Computing*. PhD thesis, Appl. Math. Statist. Lab. Rep. 25, Stanford University (1962).

17. R. E. Moore. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, NJ (1966).

18. A. Neumaier. Hybrid norms and bounds for overdetermined linear systems. *Linear Algebra Appl.*. To appear.

19. A. Neumaier. *Interval Methods for Systems of Equations*. Cambridge University Press (1990).

20. K. Nickel. Using interval methods for the numerical solution of ODE's. *Freiburger Intervall-Berichte* **83/10** (1983), 13–44.

21. A. I. Ovseevich and F. L. Chernousko. On optimal ellipsoids approximating reachable sets. *Problems of Control and Information Theory* **16** (1987), 125–134.

22. S. M. Rump. On the solution of interval linear systems. *Computing* **47** (1992), 337–353.