International Journal of Differential Equations and Applications

Volume 0 No. 0 2006, 1-29

SUPPRESSION OF THE WRAPPING EFFECT BY TAYLOR MODEL- BASED VERIFIED INTEGRATORS: LONG-TERM STABILIZATION BY SHRINK WRAPPING

Martin Berz[§], Kyoko Makino

Department of Physics and Astronomy
Michigan State University
East Lansing, MI 48824, USA
e-mails: berz@msu.edu, makino@msu.edu

Abstract: The verified integration of ranges of initial conditons through ODEs faces two major challenges, namely the precise representation of the flow over the short term, and the avoidance of unfavorable buildup of errors in the long term. We discuss the method of shrink wrapping for meeting the second of these challenges within the framework of Taylor model methods. Illustrative examples of the performance of the method and comparisons to other approaches are given.

AMS Subject Classification: 65L05, 65G20, 34-04, 41A58
Key Words: Differential equations, ODE, Initial value problem, IVP, Taylor integration, Verification, Rigorous computation, Wrapping effect, Preconditioning, QR method, Taylor model, Interval method

1. Introduction

When utilizing Taylor model-based methods for the verified integration of ODEs [1], [3], [6], the dependency on initial condition is carried through the

[§] Correspondence author

whole integration process. This controls the bulk of the dependency problem arising in each integration step very efficiently and hence the main source of the wrapping effect is eliminated to order $n+1$ for the single step.

On the practical side, the inclusion requirement asserting existence of a solution reduces to a mere inclusion of the remainder intervals, and different from conventional methods based on two separate algorithms for initial validation by an Euler step and subsequent higher order execution, the entire steps is performed in one algorithm. There is also no need to utilize additional ODEs for derivatives with respect to initial conditions. Finally, the direct availability of the antiderivation on Taylor models allows to treat the Picard operator like any other function, avoiding the need to explicitly bound error terms of integration formulas and leading to a rather straightforward verified fixed point problem.

The results of the methods developed in [6], [1], [3] can be summarized in the following theorem.

Theorem 1. ***(Continuous Dynamical System with Taylor Models)*** *Let $P+I$ be an n-dimensional Taylor model describing the flow of the ODE at the time t; i.e. for all initial conditions x_0 in the original domain region $B \subset R^n$, we have*

$$x(x_0,t) \in I + \bigcup_{x_0 \in B} P(x_0).$$

Let $P^*(x_0,t)$ be the invariant polynomial depending on x_0 and t obtained in [1], and assume that the self-inclusion step of the Picard Operator mapping described there is satisfied over the interval $[t, t+\Delta t]$ by the remainder bound I^*. Then for all $x_0 \in B$, we have

$$x(x_0, t+\Delta t) \in I^* + \bigcup_{x_0 \in B} P^*(x_0, t+\Delta t).$$

Furthermore, if even $x\left(x_0,t\right) \in P\left(x_0\right) + I$, then $x\left(x_0, t+\Delta t\right) \in P^*(x_0, t+\Delta t) + I^*$.

By induction over the individual steps, we obtain a relationship between initial conditions and final conditions at time t. Thus formally, the continuous case is made equivalent to the discrete case, for which the respective property follows immediately from the respective enclosure properties of Taylor models, as described for example in [5].

Theorem 2. ***(Discrete Dynamical System with Taylor Models)*** *Let $P+I$ be an n-dimensional Taylor model describing the flow of the discrete* *dynamical system* $x_{n+1} = f(x_n, n)$*, i.e. for all initial conditions* x_0 *in the original domain region* $B \subset R^n$*, we have*

$$x_n(x_0) \in I + \bigcup_{x_0 \in B} P(x_0).$$

Let $P^* + I^*$ *be the Taylor model evaluation of* $f(P + I, n)$*. Then for all* $x_0 \in B$*, we have*

$$x_{n+1}(x_0) \in I^* + \bigcup_{x_0 \in B} P^*(x_0).$$

Furthermore, if even $x_n(x_0, t) \in P(x_0) + I$*, then* $x_{n+1}(x_0) \in P^*(x_0) + I^*$*.*

The two theorems thus allows the verified study of continuous and discrete dynamical systems, provided that the Taylor model arithmetic is performed in a verified manner. In the case of the implementation in COSY, all errors in the floating point coefficients are fully accounted for [5], [9].

2. The Shrink Wrapping Approach

In this section, we address one method to control the long-term growth of integration errors. As we saw in the last section, for a fixed time t of interest, the errors appearing in the remainder interval can at least in principle be kept as small as desired. However, for large values of the time t, the approach used there may become computationally impractical because the compounding of errors can be rapid, and so it is desirable to develop schemes that limit the error growth as a function of time for a fixed expansion order and computational accuracy. The shrink wrapping method[7] is one approach for this purpose. It is based on the idea of enclosing the remainder error including floating point errors and errors due to the finite order in time within the range of the polynomial part of the Taylor model. By doing so, the remainder error ceases to be an interval, and instead is transformed into a variable that is retained explicitly up to the order of the Taylor model.

While in the linear case, this problem reduces to mere linear algebra, in the nonlinear case the situation is more involved, as the present nonlinear terms should not be also simply lumped into the linear parts at the same time; so the task requires to absorb the interval into a nonlinear structure, and we refer to it as shrink wrapping. In the following, we present one method to perform shrink wrapping; we point out that there are many variants of this approach, and while the one shown here is one of the simpler ones to outline, it is not necessarily the optimal choice for given problems.

As discussed in the introduction, after the kth step of the integration, the region occupied by the final variables is given by the set

$$A = I_0 + \bigcup_{x_0 \in B} \mathcal{M}_0(x_0), \tag{1}$$

where x_0 are the initial variables, B is the original box of initial conditions, $\mathcal{M}_0$ is the polynomial part of the Taylor model, and I_0 is the remainder bound interval; the sum is the conventional sum of sets. In the case of the COSY-VI integration, the map $\mathcal{M}_0$ can be scaled such that the original box B is unity, i.e. $B = [-1,1]^v$. We assume this to be the case for the rest of the discussion. The remainder bound interval I_0 accounts for the local approximation error of the expansion in time carried out in the kth step as well as floating point errors and potentially other accumulated errors from previous steps; it is usually very small. As stated earlier, the purpose of shrink wrapping is to "absorb" the small remainder interval into a set very similar to the second part of the right hand side in eq. (1) via

$$A \subset A^* = I_0^* + \bigcup_{x_0 \in B} \mathcal{M}_0^*(x_0),$$

where $\mathcal{M}_0^*$ is a slightly modified polynomial, and I_0^* is a significantly reduced interval of the size of machine precision.

As the first step, we extract the constant part a_0 and linear part $M_0 \cdot x$ of $\mathcal{M}_0$ and determine a floating point approximation $\bar{M}_0^{-1}$ of the inverse of M_0. In case the ODEs admit unique solutions, as is typically the case for the problems at hand, also the linear part of the flow is invertible. Within a floating point environment, thus the attempt to invert the linear transformation M_0 will likely succeed as long as the linear transformation is sufficiently well-conditioned. If this is not the case, additional steps may be necessary, which will be discussed in some detail below.

After the approximate inverse $\bar{M}_0^{-1}$ has been determined, we apply the linear transformation $\bar{M}_0^{-1} \cdot (x - a_0)$ from the left to the Taylor model $\mathcal{M}_0(x_0) + I_0$ that describes the current flow. As a result, the constant part of the resulting Taylor model now vanishes, and its linear part is near identity. We write the resulting Taylor model as

$$\mathcal{M} + I = \mathcal{I} + \mathcal{S} + I,$$

where $\mathcal{I}$ is the identity, and the function $\mathcal{S}$ contains the nonlinear parts of the resulting Taylor model as well as some small linear corrections due to the error in inversion. We include I into the interval box $d \cdot [-1,1]^v$, where d is a small number.

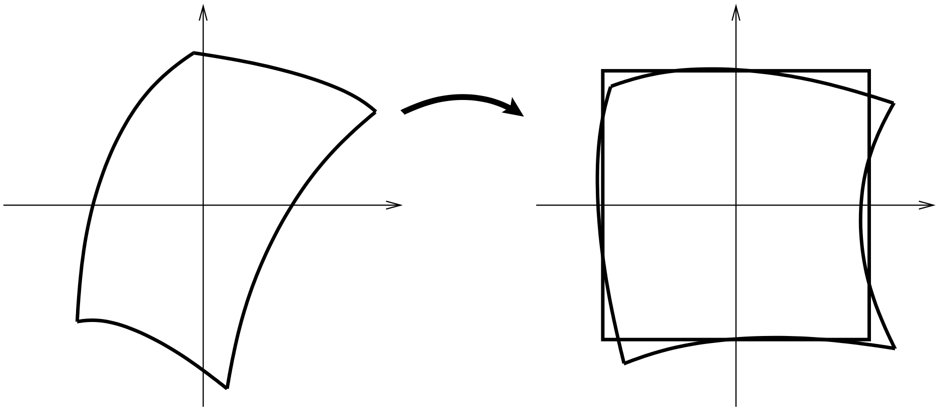

Figure 1: The region described by the Taylor model $\mathcal{M}_0 + I_0$ is transformed to be normalized as $\mathcal{I} + \mathcal{S} + I$, where $\mathcal{I}$ is the identity.

Definition 3. Let $\mathcal{M} = \mathcal{I} + \mathcal{S} + I$, where $\mathcal{S}$ is a polynomial and I is a small interval. We include I into the interval box $d \cdot [-1, 1]^v$. We pick numbers s and t satisfying

$$s \geq |\mathcal{S}_i(x)| \ \forall\ x \in B,\ 1 \leq i \leq v,$$
$$t \geq \left| \frac{\partial \mathcal{S}_i(x)}{\partial x_j} \right| \ \forall\ x \in B,\ 1 \leq i, j \leq v.$$

We call a map $\mathcal{M}$ shrinkable if $(1 - vt) > 0$ and $(1 - s) > 0$; both of which can be achieved if $\mathcal{S}$ (and since it is a polynomial, also its derivative) is sufficiently small in magnitude. Then we define q, the so-called shrink wrap factor, as

$$q = 1 + d \cdot \frac{1}{(1 - (v-1)t) \cdot (1 - s)}.$$

The bounds s and t for the polynomials $\mathcal{S}_i$ and $\partial \mathcal{S}_i / \partial x_j$ can be computed by interval evaluation. The factor q will prove to be a factor by which the Taylor polynomial $\mathcal{I}+\mathcal{S}$ has to be multiplied in order to absorb the remainder bound interval.

Remark 4. (Typical values for q) To put the various numbers in perspective, in the case of the verified integration of the Asteroid 1997 XF11, we typically have $d = 10^{-7}$, $s = 10^{-4}$, $t = 10^{-4}$, and thus $q \approx 1 + 10^{-7}$. It is interesting to note that the values for s and t are determined by the nonlinearity in the problem at hand, while in the absence of "noise" terms in the ODEs described by intervals, the value of d is determined mostly by the accuracy of the arithmetic. Rough estimates of the expected performance in quadruple precision arithmetic indicate that with an accompanying decrease in step size, if desired d can be decreased below 10^{-12}, resulting in $q \approx 1 + 10^{-12}$.

In order to proceed, we need some estimates relating image distances to origin distances.

Lemma 5. Let $\mathcal{M}$ be a map as above, let $\|\cdot\|$ denote the max norm, and let $(1 - vt) > 0$. Then we have

$$\begin{aligned} |\mathcal{M}_i(\bar{x}) - \mathcal{M}_i(x)| &\leq \sum_j |\delta_{i,j} + t| \;\; |\bar{x}_j - x_j| \,, \\ \|\mathcal{M}(\bar{x}) - \mathcal{M}(x)\| &\leq (1 + vt) \cdot \|\bar{x} - x\| \,, \text{ and} \\ \|\mathcal{M}(\bar{x}) - \mathcal{M}(x)\| &\geq (1 - vt) \cdot \|\bar{x} - x\| \,. \end{aligned}$$

where $\delta_{i,j}$ denotes the Kronecker delta.

Proof. For the proof of the first assertion, we observe that all $(v - 1)$ partials of $\partial \mathcal{M}_i / \partial x_j$ for $j \neq i$ are bounded in magnitude by t, while $\partial \mathcal{M}_i / \partial x_i$ is bounded in magnitude by $1 + t$; thus the first statement follows from the intermediate value theorem. For the second assertion, we trivially observe

$$\begin{aligned} \|\mathcal{M}(\bar{x}) - \mathcal{M}(x)\| &= \max_i |\mathcal{M}_i(\bar{x}) - \mathcal{M}_i(x)| \\ &\leq \max_i \sum_j |\delta_{i,j} + t| \;\; |\bar{x}_j - x_j| \\ &\leq (1 + vt) \;\; \|\bar{x} - x\| \,. \end{aligned}$$

For the proof of the third assertion, which is more involved, let k be such that $\|\bar{x} - x\| = |\bar{x}_k - x_k|$, and wlog let $\bar{x}_k - x_k > 0$. Then we have

$$\begin{aligned} \|\mathcal{M}(\bar{x}) - \mathcal{M}(x)\| &= \max_i |\mathcal{M}_i(\bar{x}) - \mathcal{M}_i(x)| \\ &\geq |\mathcal{M}_k(\bar{x}) - \mathcal{M}_k(x)| \\ &= \left| (1 + c_k)(\bar{x}_k - x_k) + \sum_{j \neq k} c_j (\bar{x}_j - x_j) \right| \qquad (2) \end{aligned}$$

for some set of c_j with $|c_j| \leq t \;\forall j = 1, ..., v$, according to the mean value theorem. Now observe that for any such set of c_j,

$$\begin{aligned}\left|\sum_{j\neq k} c_j(\bar{x}_j - x_j)\right| &\leq \sum_{j\neq k} |c_j| \ \ |\bar{x}_j - x_j| \leq \left(\sum_{j\neq k} |c_j|\right) \ |\bar{x}_k - x_k| \\ &\leq (v-1) \ t \ |\bar{x}_k - x_k| \\ &\leq (1-t) \ |\bar{x}_k - x_k| \leq (1+c_k)\,(\bar{x}_k - x_k)\,.\end{aligned}$$

Hence the left term in the right hand absolute value in (2) dominates the right term for any set of c_j, and we thus have

$$\begin{aligned}&\left|(1+c_k)(\bar{x}_k - x_k) + \sum_{j\neq k} c_j(\bar{x}_j - x_j)\right| \\ &\geq (1-t)(\bar{x}_k - x_k) - \sum_{j\neq k} t \ \ |\bar{x}_j - x_j| \\ &\geq (1-t)(\bar{x}_k - x_k) - (v-1) \ t \ (\bar{x}_k - x_k) \\ &= (1-vt)(\bar{x}_k - x_k) = (1-vt) \ \ \|\bar{x} - x\|\,,\end{aligned}$$

which completes the proof. □

Theorem 6. *(Shrink Wrapping) Let $\mathcal{M} = \mathcal{I} + \mathcal{S}(x)$, where $\mathcal{I}$ is the identity. Let $I = d \cdot [-1,1]^v$, and*

$$R = I + \bigcup_{x\in B} \mathcal{M}(x)$$

be the set sum of the interval $I = [-d,d]^v$ and the range of $\mathcal{M}$ over the original domain box B. Let q be the shrink wrap factor of $\mathcal{M}$; then we have

$$R \subset \bigcup_{x\in B} (q\mathcal{M})(x),$$

and hence multiplying $\mathcal{M}$ with the number q allows to set the remainder bound to zero.

Proof. Let $1 \leq i \leq v$ be given. We note that because $\partial\mathcal{M}_i/\partial x_i > 1-t > 0$, $\mathcal{M}_i$ increases monotonically with x_i. Consider now the $(v-1)$ dimensional surface set $(x_1, ..., x_v)$ with $x_i = 1$ fixed. Pick a set of $x_j \in [-1,1]$, $j \neq i$. We want to study how far the set $R = I + \bigcup_{x\in B} \mathcal{M}(x)$ can extend beyond the surface in direction i at the surface point $y = \mathcal{M}(x_1, ..., x_{i-1}, 1, x_{i+1}, ..., x_v)$.

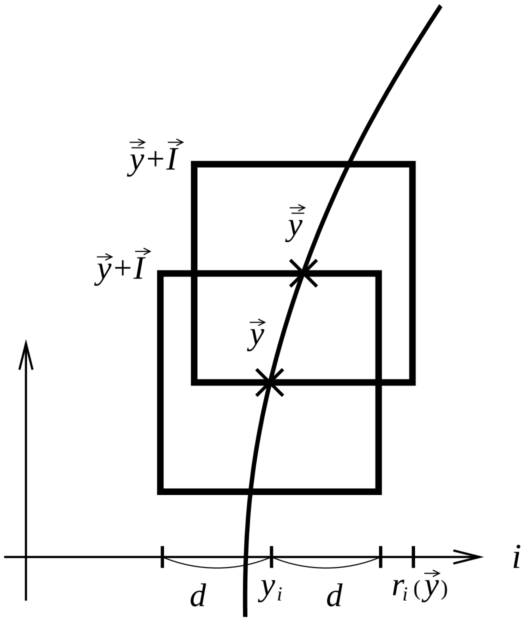

Figure 2: At the point y_i, the set $R = I + \bigcup_{x \in B} \mathcal{M}(x)$ can extend to $r_i(y)$.

Let y_i be the i-th component of y. The i-th components of the set $y + I$ apparently extends beyond y_i by d. However, it is obvious that R can extend further than that beyond y_i. In fact, for any other $\bar{y}$ with $|\bar{y}_j - y_j| \leq d$ for $j \neq i$, there are points in $\bar{y} + I$ with all but the i-th component equal to those of y. On the other hand, any $\bar{y}$ with $|\bar{y}_j - y_j| > d$ for some $j \neq i$ can not have a point in $\bar{y} + I$ with all but the i-th component matching those of y. So at the point y_i, the set R can extend to

$$r_i(y) = d + \sup_{\{\bar{y} |\ |\bar{y}_j - y_j| \leq d\ (j \neq i)\}} \bar{y}_i.$$

We shall now find a bound for $r_i(y)$. First we observe that because of the monotonicity of $\mathcal{M}_i$, we can restrict the search to the case with $x_i = 1$. We now project to an $(v-1)$ dimensional subspace by fixing $x_i = 1$ and by removing the i-th component $\mathcal{M}_i$. We denote the resulting map by $\mathcal{M}^{(i)}$, and similarly denote all $(v-1)$ dimensional variables with the superscript "(i)".

We observe that with the function $\mathcal{M}$, also the function $\mathcal{M}^{(i)}$ is shrinkable according to the definition, with factors s and t inherited from $\mathcal{M}$. Apparently the condition on $\bar{y}$ in the definition of $r_i(y)$ entails that in the $(v-1)$

dimensional subspace, $\left\|\bar{y}^{(i)} - y^{(i)}\right\| \leq d$. Let $\bar{x}^{(i)}$ and $x^{(i)}$ be the $(v-1)$ dimensional pre-images of $\bar{y}^{(i)}$ and $y^{(i)}$, respectively; because $\left\|\bar{y}^{(i)} - y^{(i)}\right\| \leq d$, we have according to the above lemma that

$$\left\|\bar{x}^{(i)} - x^{(i)}\right\| \leq \frac{d}{1-(v-1)t},$$

which entails that also in the original space we have $|\bar{x}_j - x_j| \leq d/(1-(v-1)t)$ for $j \neq i$. Hence we can bound $r_i(y)$ via

$$r_i(y) \leq d + \sup_{\substack{\{\bar{x} \mid \, |\bar{x}_j - x_j| \leq d/(1-(v-1)t) \\ (j \neq i), \; x_i = \bar{x}_i = 1\}}} \mathcal{M}_i(\bar{x}).$$

We now invoke the first statement of the lemma for the case of $\bar{x}$, x satisfying $|\bar{x}_j - x_j| \leq d/(1-(v-1)t)$ $(j \neq i)$, $x_i = \bar{x}_i = 1$. The last condition implies that the term involving $(\delta_{i,j} + t)$ does not contribute, and we thus have $|\mathcal{M}_i(\bar{x}) - \mathcal{M}_i(x)| \leq (v-1)t \cdot d/(1-(v-1)t)$, and altogether

$$\begin{aligned} r_i(y) &\leq y_i + d + \frac{d \cdot (v-1)t}{1-(v-1)t} \\ &= y_i + d \cdot \frac{1}{1-(v-1)t}. \end{aligned}$$

We observe that the second term in the last expression is independent of i. Hence we have shown that the "band" around $\bigcup_{x \in B} \mathcal{M}(x)$ generated by the addition of I never extends more than $d/(1-(v-1)t)$ in any direction.

To complete the proof, we observe that because of the bound s on $\mathcal{S}$, the box $(1-s)[-1,1]^v$ lies entirely in the range of $\mathcal{M}$. Thus multiplying the map $\mathcal{M}$ with any factor $q > 1$ entails that the edges of the box $(1-s)[-1,1]^v$ move out by the amount $(1-s)(q-1)$ in all directions. Since the box is entirely inside the range of $\mathcal{M}$, this also means that the border of the range of $\mathcal{M}$ moves out by at least the same amount in any direction i. Thus choosing q as

$$q = 1 + d \cdot \frac{1}{(1-(v-1)t) \cdot (1-s)}$$

assures that

$$\bigcup_{x \in B} (q\mathcal{M}) \supset R$$

as claimed. □

Remark 7. (Shrink Wrapping and Complex Arithmetic)
Taylor models have also been successfully used to perform operations in the complex plane. To this end, one merely identifies complex functions as functions from R^2 into R^2 and observes that analyticity entails infinite partial differentiability of the component functions. Thus complex analytic functions can be described as pairs of Taylor models in two variables, and the rules for Taylor model arithmetic can be applied to the component functions. Apparently the geometric properties of the resulting ranges of the Taylor models are analogous to the situation of the flows of ODEs above; and in a similar way it is thus possible to absorb the remainder term into the polynomial part of the Taylor model.

Let us consider the practical limitations of the method:

Remark 8. (Limitations of Shrink Wrapping) Apparently the shrink wrap method discussed above has the following limitations

1. The measures of nonlinearities s and t must not become too large
2. The application of the inverse of the linear part should not lead to large increases in the size of remainder bounds.

Apparently the first requirement limits the domain size that can be covered by the Taylor model, and it will thus be relevant only in extreme cases. Furthermore, in practice the case of s and t becoming large is connected to also having accumulated a large remainder bound, since the remainder bounds are calculated from the bounds of the various orders of s. In the light of this, not much additional harm is done by removing the offending s into the remainder bound and create a linearized Taylor model.

Definition 9. (Linearized Taylor Model) Let $M_0 \cdot x + \mathcal{S} + I$ be a Taylor model with nonlinear part $\mathcal{S}$, and let the components of S be bounded by $s = (s_i)$. We call

$$M_0 \cdot x + I + s \cdot [-1, 1]$$

the linearized Taylor model of $M_0 \cdot x + \mathcal{S} + I$.

The overestimation generated by the application of the inverse of the linear part is apparently directly connected to the condition number of the linear part M_0.

Definition 10. (Blunting of an Ill-Conditioned Matrix) Let A be a regular nxn matrix that is potentially ill-conditioned and $q = (q_1, ...q_n)$

be a vector with $q_i > 0$. Arrange the column vectors a_i of A by Euclidean length. Let e_i be the familiar orthonormal vectors obtained through the Gram-Schmidt procedure, i.e.

$$e_i = \frac{a_i - \sum\limits_{k=1}^{i-1} e_k \ (a_i \cdot e_k)}{\left| a_i - \sum\limits_{k=1}^{i-1} e_k \ (a_i \cdot e_k) \right|}.$$

We form vectors b_i via

$$b_i = a_i + q_i e_i$$

and assemble them columnwise into the matrix B. We call B the q-blunted matrix belonging to A.

Proposition 11. (Regularity of the Blunted Matrix) The b_i are linearly independent and thus B is regular.

Proof. By induction. Apparently b_1 is linearly independent. Assume now that $b_1, ..., b_{i-1}$ are linearly independent. We first observe that for each i, the vector b_i is by virtue of its definition a linear combination of the a_k for $k = 1, ..., i$ and thus also of the e_k for $k = 1, ..., i$, since both sets of vectors span the same space. Now suppose b_i is linearly dependent on $b_1, ..., b_{i-1}$; then it is also linearly dependent on $e_1, ..., e_{i-1}$, and in particular we must have $b_i \cdot e_i = 0$. Observe that we have $(a_i)^2 = \sum_{k=1}^{n} (a_i \cdot e_k)^2$ by virtue of the fact that the vectors e_k form an orthonormal basis. Using this, we obtain from the definition of b_i that

$$\begin{aligned} b_i \cdot e_i &= a_i \cdot e_i + q_i \\ &= \frac{(a_i)^2 - \sum\limits_{k=1}^{i-1} (a_i \cdot e_k) \ (a_i \cdot e_k)}{\left| a_i - \sum\limits_{k=1}^{i-1} e_k \ (a_i \cdot e_k) \right|} + q_i \\ &= \frac{\sum\limits_{k=i}^{n} (a_i \cdot e_k) \ ^2}{\left| a_i - \sum\limits_{k=1}^{i-1} e_k \ (a_i \cdot e_k) \right|} + q_i > 0, \end{aligned}$$

which represents a contradiction to $b_i \cdot e_i = 0$; thus $b_1, ..., b_i$ are linearly independent, which completes the induction step. □

Remark 12. (Effect of Blunting) The intuitive effect of the blunting is that b_1, and thus the dominating direction, which determines asymptotic behavior, remains unchanged. Smaller b_i are being "pulled away" from earlier ones in the direction of e_i, i.e. away from the space spanned by the previous vectors $b_1, ..., b_{i-1}$. Since $b_i \cdot e_i \geq q_i$, the "pulling" is stronger for larger choices of q_i. Thus larger choices for q_i lead to a matrix that has more favorable condition number.

Algorithm 13. (Pre-Conditioning of Shrink Wrapping) Let M_0 be the linear part of the Taylor model to be shrink wrapped. Subject M_0 to the blunting algorithm just described before attempting to compute its inverse. As a result, M_0 is less ill-conditioned, its approximate inverse $\bar{M}_0^{-1}$ is determined more easily, and is itself less ill-conditioned. As discussed in the main algorithm, the defect of applying $\bar{M}_0^{-1}$ to M_0 is moved to the remainder bound. Next, determine if the Taylor model is shrinkable as defined in 3. If it is not, or if the shrink wrap factor q exceeds a pre-specified threshold $q_{\max}$, bound the nonlinear part into the remainder bound. The result is a shrinkable Taylor model.

Remark 14. (Shrink Wrapping for Linear Systems) When applied to linear systems, the shrink wrapping with blunting limits the overestimation due to the conditioning of matrix when transforming the error interval to the new coordinate system. At the same time, the leading direction remains unchanged, and thus there is no error introduced that scales with the length of the leading direction, which determines the asymptotic error. On the other hand, the naive shrink wrapping method without blunting behaves like the well-known parallelepiped method.

Apparently the trade-off of blunting the linear part lies in an increase in the size of the remainder bound that then has to be absorbed into the Taylor model. However, this increase is not affected by the size of the dominating vector, since it remains unaffected by the blunting algorithm. Thus in studies of asymptotic behavior where the other directions become exponentially smaller compared to the dominating direction, the effect of blunting will become exponentially less significant. Since this requires sending the remainder bound through the inverse, which produces an overestimation increasing with condition number, it is expected that a moderate amount of blunting and the corresponding decrease in condition number will overall lead to a smaller shrink wrap factor. Furthermore, we observe that the less ill-conditioned inverse that results from blunting will also lead to smaller nonlinear terms, which leads to a more favorable shrink wrap factor, or may even prevent the breakdown of shrinking and the need to absorb the nonlinearities into the remainder bound.

More specifically, the larger the size of the remainder bound relative to the size of the range into which it is to be absorbed, the larger the blunting factor should be chosen, since the more important overestimation by application of the inverse becomes, while the less important the additional contributions from packing the original matrix in the blunted matrix becomes. A large ensemble of examples for the use of shrink wrapping under blunting will be studied in the next section.

In a practical environment, one may even use trial and error or other heuristics to determine suitable blunting parameters. Also, much further theoretical thought could be spent on the question of the optimal enclosure of one parallelepiped (the remainder interval) in another (the linear part). For example, one could attempt to find a "minimal" parallelepiped to do that; part of the problem is specifying the meaning of "minimal". One could think of minimizing volume, which would lead to a constrained nonlinear optimization problem. One may also think of minimizing the lengths of the vectors, which may lead to a linear programming problem.

The trade-off between these two cases seems far from obvious; first, both cases require the choice of a coordinate system that is somehow "natural" for the system, since both volume and coordinate lengths are affected by such a choice of coordinates. Furthermore, while small volume may have obvious immediate appeal, especially in the case of nonlinear systems, it may be more desirable to operate with less "extended" objects, which may reduce subsequent nonlinear effects. Finally, if the system under consideration exhibits a particular symmetry like energy conservation or symplecticity, emphasis may be placed on the satisfaction of these symmetries. Altogether, although of course all arguments remain verified in our setting, the efficiency of the method is greatly affected by heuristic choices, in much the same way as in conventional numerical integration.

Definition 15. (Parameterizing of Remainder Bounds) Let (P, I) be a Taylor model describing a function $f : D \subset R^n \to R^m$ We introduce a new polynomial $P^* : (D \times I) \subset R^{n+m} \to R^m$ via

$$P^*(x, t) = P(x) + t \text{ on } D^* = D \times I.$$

The Taylor model $(P^*, [0, 0])$ is called a parameter-extended Taylor model of f.

We have the following immediate result.

Proposition 16. **(Enclosure Property)** For all $x \in D$, we have $f(x) \in P^*(x, I) + [0, 0]$

What may appear as a simple mathematical slight of hand actually has important consequences, since for subsequent steps of the integration, we have uniquely represented f by only the Taylor model $(P^*, [0, 0])$ in a higher dimensional space that has no remainder bound. We may thus proceed with subsequent operations in Taylor model arithmetic with the parameter extended Taylor model $(P^*, [0, 0])$ instead of the Taylor model (P, I). The consequence is that in later steps, what was originally the interval I and is thus subject to the cancellation and wrapping problems, is now the variable t, which can be carried through all occurring Taylor model operations.

3. Example: Long-Term Error Growth of Floating-Point Operations

The long-term numerical study of differential equations and dynamical systems in a computer environment operating with fixed precision is frequently characterized by an exponential growth of the error. We first observe the important point that this fact is intimately tied to the use of arithmetic of finite precision, and does not merely appear in verified methods. We also observe that this effect is independent of the well-known and frequently studied phenomenon of chaos, which is characterized by exponential growth of errors in initial conditions in the true system.

To illustrate this phenomenon, let us consider the perhaps simplest conceivable discrete dynamical system, which merely oscillates between two states as

$$x_{n+1} = \begin{cases} a \cdot x_n, & n \text{ even} \\ (1/a) \cdot x_n, & n \text{ odd} \end{cases} \tag{3}$$

with initial condition $x_0 = 1$. We study the behavior for specific choices of a in both single and double precision arithmetic on two commonly used compilers, the f77 compiler by DEC, which is now distributed as f77 Digital Visual Fortran Version 5.0 as part of Microsoft Fortran PowerStation, as well as the g77 compiler distributed by GNU; we specifically tested Version V0.5.24. All tests were executed in the Cygwin Unix environment in Windows 2000 and run on a Pentium III processor, and no changes to default rounding modes were made.

Specifically, we chose $a_1 = 3$ in the single precision mode, while in the double precision mode we chose $a_2 = 0.9999999901608054$ (digits generated

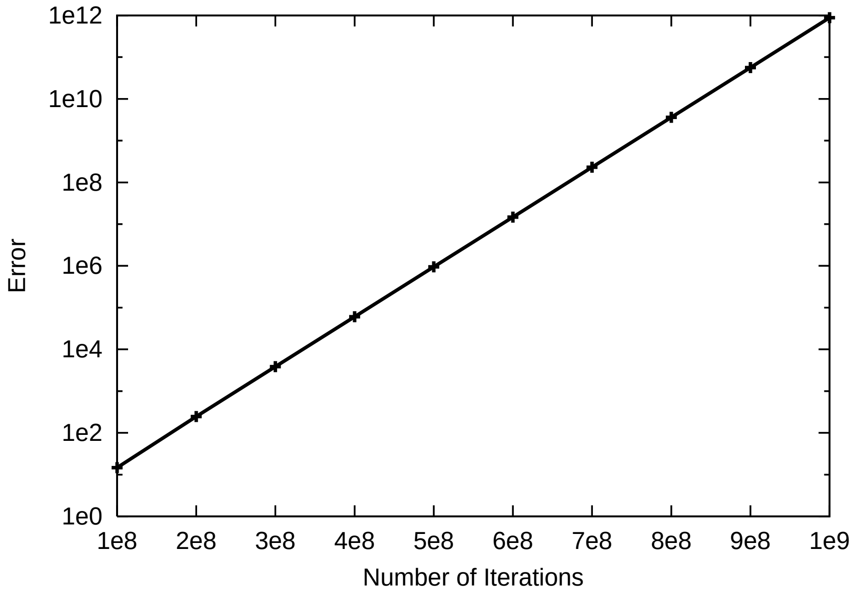

Figure 3: Arithmetic error observed in the computation of $x_{n+1} = (1/3) \cdot x_n$, $x_{n+2} = 3 \cdot x_{n+1}$, with $x_1 = 1$, for various values of n.

by FORTRAN output)

Figure 3 shows the result for the case of single precision computation using f77 with default compiler settings, revealing an exponential growth of the error that after merely 10^9 iterations reaches the value of 10^{12}. The error growth per iteration corresponds to approximately $1 + 1.2 \cdot 10^{-8}$, and hence represents an average increment near the last significant bit.

Performing the same experiment with $a_1 = 3$ in double precision arithmetic on either f77 or g77 did not produce any exponential growth of errors; however, performing a random search for values of a near 1 that might lead to exponential growth yielded the above a_2 within the first 10 tries, and many other values of a with a similar behavior have also been found quite easily. The empirically computed error growth factor per iteration is about $1 + 1.1 \cdot 10^{-16}$, again corresponding to an increment near the last significant bit.

Executing the simple dynamical system with interval arithmetic leads to exponentially inflating bounds, as is expected from interval methods; however, in a well-written interval environment that rounds by a minimally sufficient amount, the overestimation of the computed bounds tightly enclose the growing error. Thus in this case, the observed exponential growth of the interval results is not due to any artificial overinflation of the interval method, but rather to the unavoidable uncertainty of the results of the underlying floating point arithmetic.

4. Example: A Nonlinear Problem and Shrink Wrapping

Let us now study such two-state systems in the multidimensional nonlinear setting. First we observe that any errors that may occur lead to a more complicated geometric shape of the solution sets that have to be studied. While in the one-dimensional case, an interval can always tightly contain the results of all such overestimations, this no longer holds in the multidimensional case. As a simple example, consider the following two-state discrete dynamical system

$$x_{n+1} = x_n \cdot \sqrt{1 + x_n^2 + y_n^2} \quad \text{and} \quad y_{n+1} = y_n \cdot \sqrt{1 + x_n^2 + y_n^2}$$
$$x_{n+2} = x_{n+1} \cdot \sqrt{\frac{2}{1 + \sqrt{1 + 4(x_{n+1}^2 + y_{n+1}^2)}}} \text{ and}$$
$$y_{n+2} = y_{n+1} \cdot \sqrt{\frac{2}{1 + \sqrt{1 + 4(x_{n+1}^2 + y_{n+1}^2)}}}. \tag{4}$$

Simple arithmetic shows that, similar to the two-state system in eq. 3, also this transformation has the property that $(x_{n+2}, y_{n+2}) = (x_n, y_n)$. Considering the action of the system on the box $[-d, d]^2$, we see that the corner points $(\pm d, \pm d)$ are stretched out more than the axis intersection points $(\pm d, 0)$ and $(0, \pm d)$, which leads to a pincushion shape with four-fold symmetry after each odd step; the action on three centered squares is shown in figure 4. Attempting to represent this structure by an interval box, or for that matter any linear transformation of an interval box, will thus necessarily lead to a noticeable overestimation. On the other hand, representing the action of the iteration by a Taylor model will, for moderate values of d, be able to lead to a much more accurate representation. Finally, note that the linear transformations of the action of this system will always return to the identity after even numbers of iteration and is also rather well conditioned after odd iterations, so numerical difficulties due to conditioning do not arise

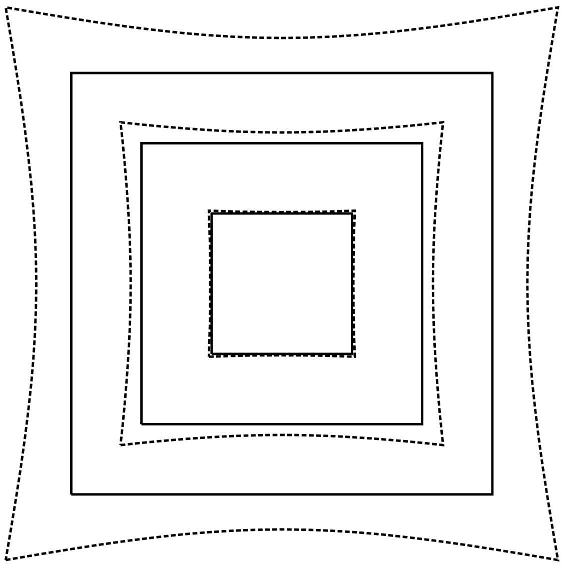

Figure 4: The action of the two-step nonlinear transformation. Squares are subjected to pincushion-shaped deformation and transformed back into themselves.

in this case. Thus the example represents a good test for a method to treat nonlinear effects.

The results of a simulation with Taylor models of various orders and with and without shrink wrapping are shown in figures 5 and 6 for the point $(0,0)+[-.05,.05]^2$ and in figure 7 and 8 for the point $(1,1)+[-.05,.05]^2$. Because after two steps the linear part is the identity, the problem allows to study the ability of the shrink wrap method to handle nonlinear effects, but without possible complications that may arise due to the conditioning of the linear part, which will be studied in other examples below.

Because the linear part represents the identity, shrink wrapping with first order Taylor model behaves exactly like the QR and PE methods, and so a useful comparison to these methods is possible. Apparently the use of shrink wrapping and higher order Taylor models leads to very extended stability; for example, Taylor models of order 20 lead to survival for 10^5 iterations with an accumulated error around 10^{-9}, while the lack of use of shrink wrapping or the use of linear methods leads to unacceptable errors in 100 or less iterations.

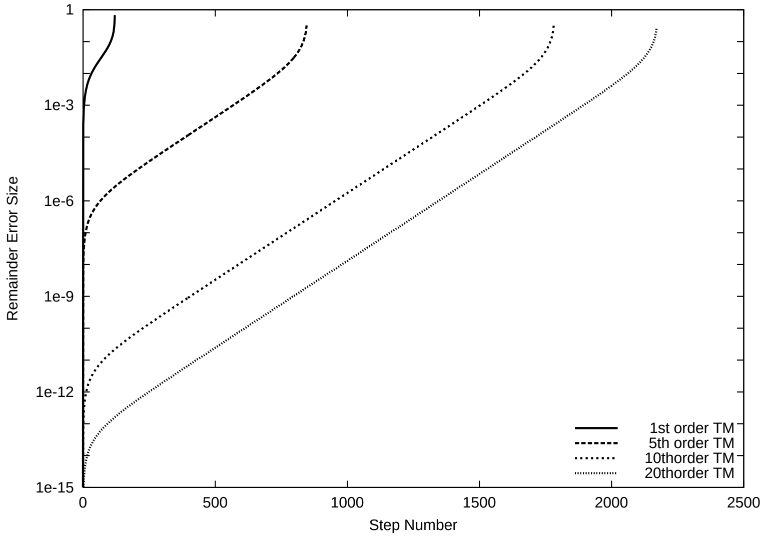

Figure 5: Discrete dynamics of the nonlinear stretch at $(0, 0) + [-.05, .05]^2$. Treatment by naive Taylor models. First order Taylor models without shrink wrapping behave like the linear PE, QR, or PEQR methods.

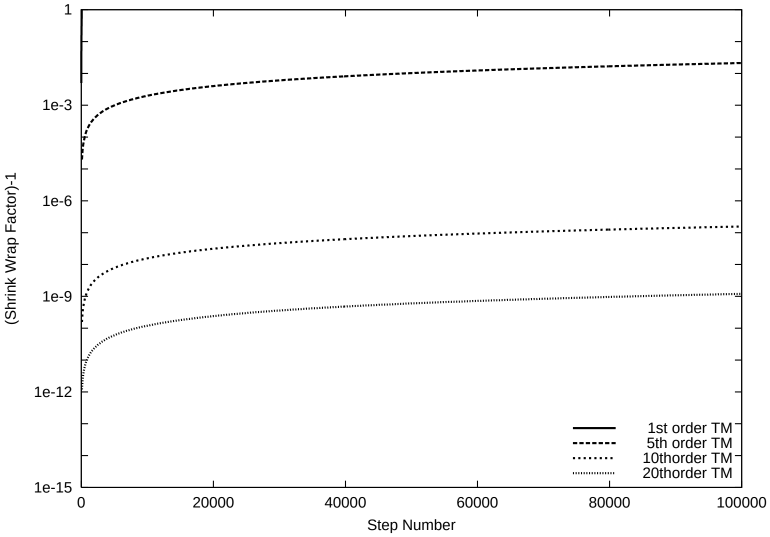

Figure 6: Discrete dynamics of the nonlinear stretch at $(0, 0) + [-.05, .05]^2$. Treatment by Taylor models with shrink wrapping.

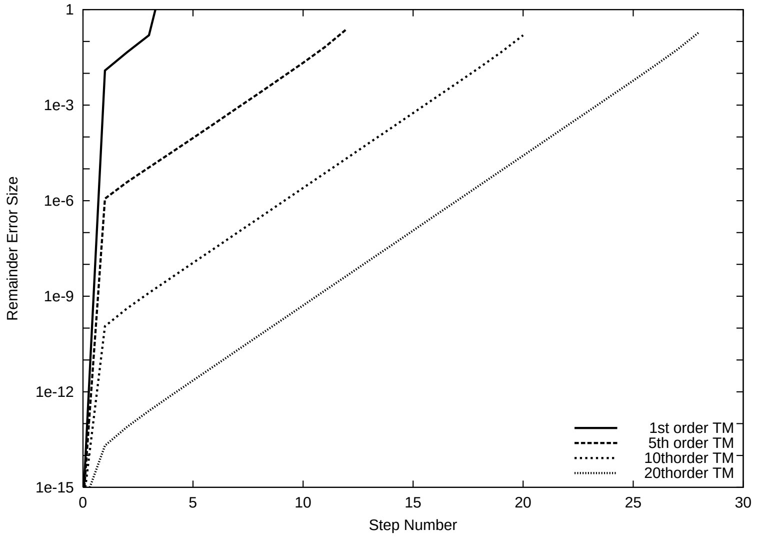

Figure 7: Discrete dynamics of the nonlinear stretch at $(1, 1) + [-.05, .05]^2$. Treatment by naive Taylor models.

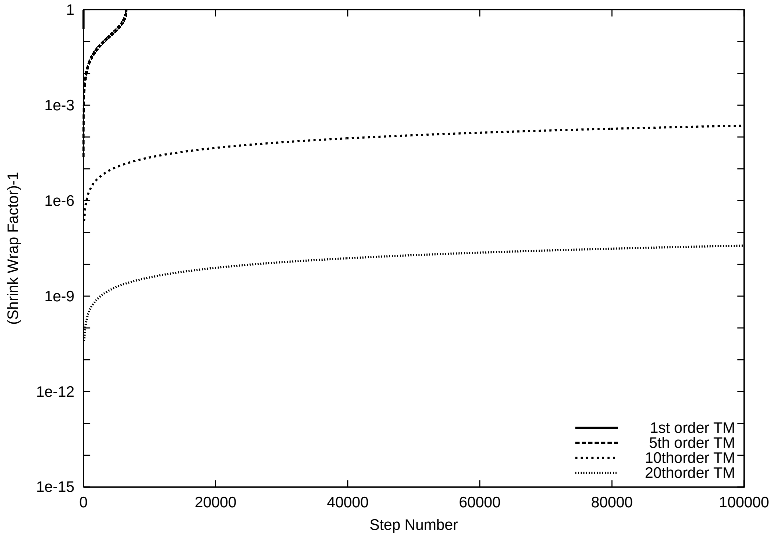

Figure 8: Discrete dynamics of the nonlinear stretch at $(1, 1) + [-.05, .05]^2$. Treatment by naive Taylor models (left) and Taylor models with shrink wrapping.

5. Examples: Linear Problems and Shrink Wrapping

While in the previous section, the emphasis was on the treatment of nonlinear effects in the absence of complications due to linear conditioning, in this section we will study the opposite: we will address linear problems that may become ill-conditioned and forgo the study of nonlinear effects. Because linear problems lead to a merely linear dependence on initial conditions, they thus allow a clear separation of the effects of the Taylor model methods that are due to the expansion in initial conditions and those of their asymptotic behavior. We focus our attention here on autonomous problems, the asymptotic behavior of which can apparently also be studied more efficiently with verified eigenvalue/eigenvector tools. The behavior of non-autonomous problems under shrink wrapping will be considered in a subsequent paper [4] within the wider context of a connection of shrink wrapping and preconditioning.

For the purpose of our study, we consider the behavior of the various methods by studying discrete dynamics of iteration through two-dimensional matrices. To minimize the influence of particular choice, we consider a collection of 1000 matrices with coefficients randomly chosen in the interval $[-1,1]$. The initial condition under study is chosen to be $(1,1)+d\cdot[-1,1]$ with a value of $d=10^{-3}$. Apparently the choice of the center point of the domain box is rather immaterial due to the randomness of the matrices; and because of linearity, the value of d is of importance only relative to the floor of precision of the floating precision environment.

We study the development of the area of enclosure as a measure of the sharpness of the method. We compare shrink wrapping without blunting (TMSP) and with blunting (TMSB). We chose the blunting factors q_i to be 10^{-3} times the length of the longest column vector of the linear matrix. In order to provide a frame of reference, we also study the performance of naive interval (IN) method as well as the naive Taylor model method (TMN); in the latter case, the area is estimated as the sum of the determinant of the linear part plus the area of the remainder bound interval box. In addition, in order to provide an assessment of the influence of the effects of the underlying floating point arithmetic, we also perform a non-verified tracking of the vectors of the four corner points $(1,1)+d\cdot(\pm1,\pm1)$ and determine the area of the linear structure spanned by the vectors; this method is referred to as the vector method (VE). Since this method is naturally inaccurate in particular for strongly elongated structures, we average over a large number of matrices to control statistical fluctuations.

In the first test, we study an autonomous problem for 500 iterations. Apparently in this case, the true solution of the problem shows an exponential shrinkage of the area by the product $|\lambda_1|\cdot\, |\lambda_2|$ of the magnitudes of the eigenvalues. For the purpose of analysis, we study two kinds of matrices; the category C_1 contains all matrices in which the eigenvalues form conjugate pairs, and the category C_2 consists of matrices for which the ratios $r = |\lambda_1|/|\lambda_2|$ of the eigenvalue λ_1 of larger magnitude to the one of smaller magnitude satisfies $1 \le r < 5$. Within the two categories, we calculate the average of the logarithm of the areas enclosed by the various methods as a function of the iteration number, which for the true dynamics would lead to a decrease along a straight line, the slope of which is given by the value $\log\left(|\lambda_1| \cdot |\lambda_2|\right)$.

Figures 9 and 10 show the results of the situation for categories C_1 and C_2, respectively. It is clearly visible that in the dynamics of C_1, the behavior is characterized by the expected linear decrease, and the blunted (TMSB) and parallelepiped (TMSP) methods show this behavior. Both of these methods very closely follow the non-verified result (VE). The behavior of the methods is in agreement with the theoretical results and practical examples found in [8]. On the other hand, the naive interval method (IN) as well as the naive Taylor model method (TMN) show a qualitatively different behavior; the interval method leads to a different slope, while over the short term the naive Taylor model method performs similar to the other methods until the size of the remainder bound becomes the dominating contribution, at which time its slope becomes similar to that of the interval method.

Studying the behavior of the class C_2 shows a similar pattern, except that now the TMSP behaves significantly worse over time because of the negative consequences of the condition number of the matrix, while TMSB still behaves similar to the non-verified VE case. More extensive studies of the linear behavior.

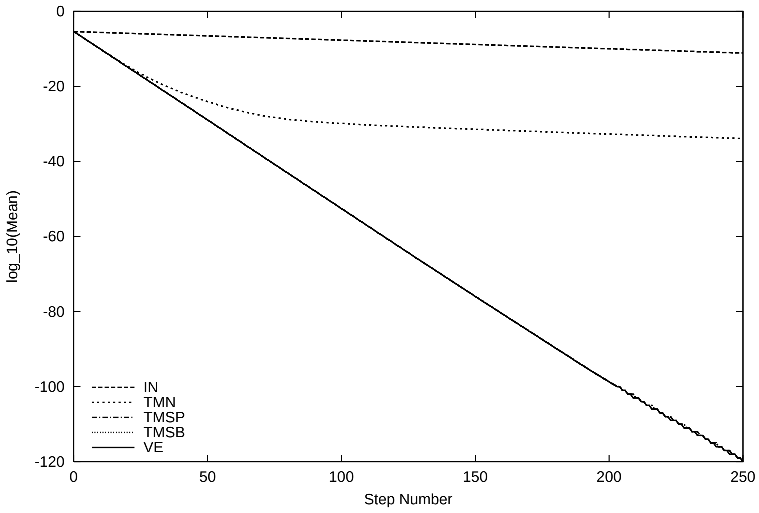

Figure 9: Areas predicted in the iteration through random 2 × 2 matrices with conjugate eigenvalues.

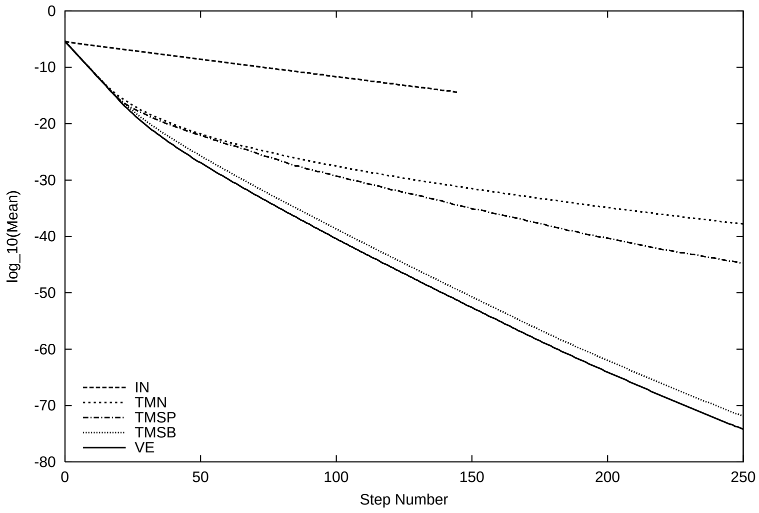

Figure 10: Areas predicted in the iteration through random 2 × 2 matrices with eigenvalues differing in magnitude by a factor of 1 to 5 for various enclosure methods.

6. Example: The Area-Preserving Henon Map

The discrete dynamics of the repeated application of the Henon map is a frequently used elementary example that exhibits many of the well-known effects of nonlinear dynamics, including chaos, periodic fixed points, islands and symplectic motion. The dynamics is two-dimensional, and given by

$$\begin{aligned} x_{n+1} &= 1 - \alpha x_n^2 + y_n \\ y_{n+1} &= \beta x_n. \end{aligned} \tag{5}$$

Since our study is focused on the prevention of overestimation of rigorous flow enclosures, it is advantageous to focus on area preserving cases so that artificial growth is not masked by the natural contractivity of the system. It can easily be seen that the motion is area preserving for $|\beta| = 1$.For our study, we borrow an example from the work of Kühn[2] illustrating the performance of the zonotope method and compare with TMs using shrink wrapping. We consider the dynamics for the special cases of $\alpha = 2.4$ and $\beta = -1$, and concentrate on initial boxes of the from $(x_0, y_0) \in (0.4, -0.4) + [-d, d]^2$. As an example to assess the dynamics, we consider the box with $d = 10^{-2}$ and study its evolution for a few turns.

Figure 11 shows the motion of the four corner points for five iterations and for 120 iterations. It becomes apparent that three of the corner points are trapped in a five-fold island structure, while one of them follows an ergodic curve inside the islands. This situation makes very long-term verified

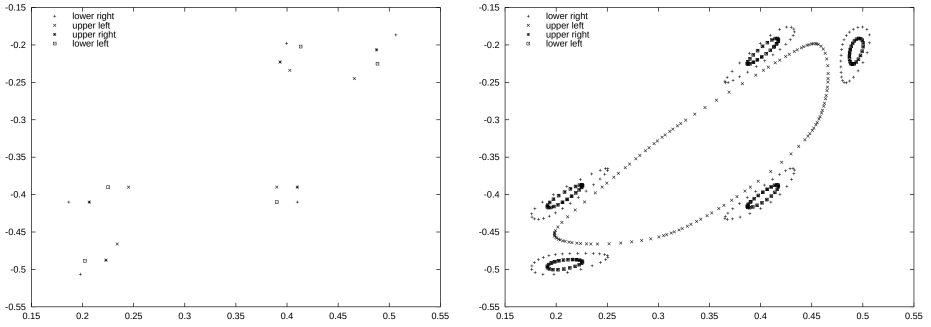

Figure 11: Iteration through the Henon map. Shown are the motion of the corner points of the box $(x_0, y_0) \in (0.4, -0.4) + [-d, d]^2$ for $d = 10^{-2}$ for five iterations (left) and for 120 iterations (right).

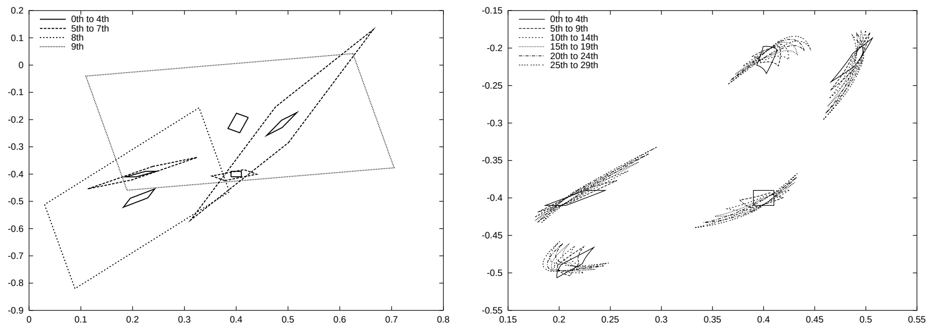

Figure 12: Dynamics through the Henon map for the box $(x_0, y_0) \in (0.4, -0.4) + [-d, d]$ for $d = 10^{-2}$ for nine turns with first order (left) and for 29 turns with tenth order (right) Taylor models with non-blunted shrink wrapping.

integration impossible since the transition region between the islands and the ergodic part is chaotic. As a first test, we study the dynamics of the box $(x_0, y_0) \in (0.4, -0.4) + [-d, d]^2$ for $d = 10^{-2}$ with first order Taylor models with shrink wrapping and compare with the results obtained by tenth order Taylor models with shrink wrapping; the results are shown in figure 12. It can be seen that the presence of the nonlinearities in the dynamics makes the size of the enclosures obtained by the linear method increase quickly. On the other hand, the higher order method can follow the details of the dynamics, including the “pulling apart” of the corner points rather well.

As a first example to study long term motion, we show the predicted inclusion after 500 iterations of the map for the case $d = 10^{-6}$. This choice of d entails that the entire box stays confined within the island structure, and is at least not subject to chaotic motion. Figure 13 shows the results obtained by the zonotope method, linear maps from $R^{m \cdot n}$ into R^n, for various numbers of of the parameter m and Taylor model methods of orders 1 and 5 using shrink wrapping. On the left, the results obtained by the Taylor model methods are overlaid on the respective results of the zonotope method; the picture was taken from [2]. For the purpose of better comparison, the TM results are also shown separately on the right. We see that the enclosure by the TM method is similarly accurate, and perhaps slightly sharper, than that of the zonotope method with $m = 15$. The right picture reveals that the TM method of order 5 produces a slightly sharper result than the TM

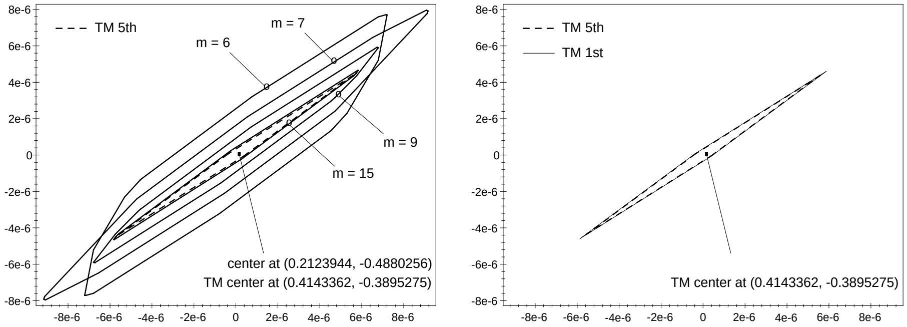

Figure 13: Validated enclosures after 500 iterations of the initial condition $(0.4, -0.4) + [-10^{-6}, 10^{-6}]^2$ through the Henon map. Shown are enclosures by the zonotope method of various values of the parameter m and by the TM methods of orders 1 and 5 using shrink wrapping (left). For better comparison, the results of the TM methods are also shown separately (right).

method of order 1.

In passing we note that the values for the center point reported for the zonotope method in [2] are incorrect; in fact, the values provided there agree to all digits shown to those after 3 iterations, but not even to one digit with those after 500 iterations, which because of the five-fold repetitive structure of the Henon map should be close to the starting point. However, because of the high degree of similarity of the $m = 15$ zonotope enclosure with that of the TM method after 500 iterations and the dissimilarity after 3 iterations, it appears very likely that the enclosure itself is indeed provided correctly.

In order to study the behavior of the TM methods for long term problems, we iterate the map until failure occurs. In [2] it is reported that for the domain $(x_0, y_0) \in (0.4, -0.4) + [-d, d]^2$ with $d = 10^{-12}$, the $m = 15$ zonotope method succeeds for about $33,000$ iterations. We compare this behavior with the preconditioned TM method of order 5 with shrink wrapping and observe that the method can succeed to provide enclosures for about $280,000$ iterations for order 5 and slightly longer for order 10. The TM method of order 1 survives for about $20,000$ iterations. Figure 14 shows some results of these computations. On the left we show the size of the remainder bounds for each turn, which is nonzero if the shrink wrapping fails to be executed. The remainder terms are usually in the range of 10^{-12},

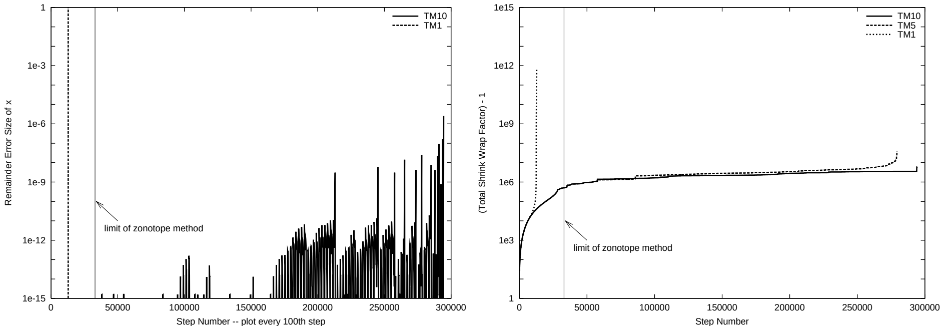

Figure 14: Dynamics in the Henon map for $(x_0, y_0) \in (0.4, -0.4) + [-d, d]^2$ with $d = 10^{-12}$. Shown are the remainder bounds (left) and total shrink wrap factors (right) for TMs of order 1, 5, and 10.

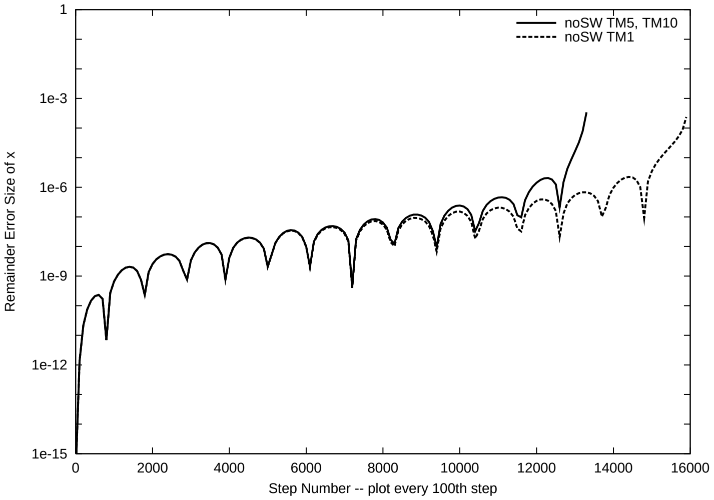

Figure 15: Dynamics in the Henon map for $(x_0, y_0) \in (0.4, -0.4) + [-d, d]^2$ with $d = 10^{-12}$ using naive Taylor models without shrink wrapping. Shown are the the remainder bounds for TMs of order 1, 5, and 10.

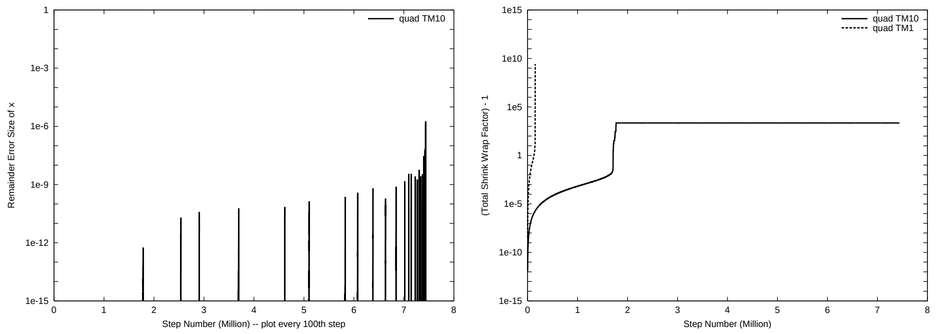

Figure 16: Non-verified dynamics in the Henon map for floating point errors similar to those in quadruple precision for $(x_0, y_0) \in (0.4, -0.4) + [-d, d]^2$ with $d = 10^{-12}$ Shown are the the remainder bounds (left) and shrink wrap factors (right) for TMs of order 1, 5, and 10.

but occasionally exceed 10^{-9}. The right shows the total accumulated shrink wrap factor, which is a measure of the inflation of the box. The seemingly large value of 10^6 is due to the fact that because of the proximity to the floating point floor, the initially small box size of 10^{-12} increases quickly. Approximately at the number of iterations at which the zonotope method fails to proceed, the shrink wrap factor stabilizes at about 10^6, leading to an overall box size of around 10^{-6}.

It is also interesting to study how much of an improvement shrink wrapping provides compared to iteration with naive Taylor models. Figure 15 shows the remainder bounds obtained in this approach, and it is apparent that failure now occurs much more rapidly at around $16,000$ iterations, about half as much as the zonotope method is able to succeed.

In order to assess the expected influence of double precision floating point error, we attempt to simulate the behavior in quadruple precision. Due to the absence of an arbitrary precision or quadruple precision implementation of our TM tools, we perform a non-verified experiment in which the floating point accuracy threshold ε_m that is used in the internal interval operations was artificially set to the 10^{-30}, a number typical for the use of quadruple precision arithmetic. While the resulting inclusions are of course not verified results since the actual accuracy remains at the level of 10^{-15} or so, the results provide a rather good estimate for the growth of errors that is to be expected in quadruple precision.

Repeating the study in this way, we observe that the survival time of the first order method now increases to a respectable $150,000$ iterations. But on the other hand, the higher order methods can now execute more than $7,500,000$ iterations, or about 50 times as much. A more detailed study of the results in figure 16 shows that beyond well over one million turns, the shrink wrap factor grows very moderately to about 10^{-6}, until just before 2 million turns, the first intermediate failures of shrink wrapping occur. At this point, the shrink wrap factor increases appreciably to reabsorb the remainder term a few iterations later, the map again becomes shrinkable. Overall it is clear that here the use of the higher order methods quite significantly improves performance, which seems to be limited mostly by floating point errors.

Acknowledgement

This work was supported by the US Department of Energy, the National Science Foundation, the Illinois Consortium for Accelerator Research and an Alfred P. Sloan Fellowship.

References

[1] M. Berz and K. Makino. Verified integration of ODEs and flows using differential algebraic methods on high-order Taylor models. *Reliable Computing*, 4(4):361–369, 1998.

[2] W. Kühn. Rigorously computed orbits of dynamical systems without the wrapping effect. *Computing*, 61:47–67, 1998.

[3] K. Makino. *Rigorous Analysis of Nonlinear Motion in Particle Accelerators*. PhD thesis, Michigan State University, East Lansing, Michigan, USA, 1998. Also MSUCL-1093.

[4] K. Makino. Suppression of the wrapping effect by Taylor model- based verified integrators: Long-term stabilization by preconditioning. *International Journal of Differential Equations and Applications*, 2006.

[5] K. Makino and M. Berz. Taylor models and other validated functional inclusion methods. *International Journal of Pure and Applied Mathematics*, 6,3:239–316, 2003.

[6] K. Makino and M. Berz. Suppression of the wrapping effect by Taylor model- based verified integrators: The single step. *International Journal of Pure and Applied Mathematics*, 2006.

[7] K. Makino and M. Berz. The method of shrink wrapping for the validated solution of ODEs. Technical Report MSUHEP-20510, Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, May 2002.

[8] N. S. Nedialkov and K. R. Jackson. A new perspective on the wrapping effect in interval methods for IVPs for ODEs. *Proc. SCAN2000, Kluwer*, 2001.

[9] N. Revol, K. Makino, and M. Berz. Taylor models and floating-point arithmetic: Proof that arithmetic operations are validated in COSY. *Journal of Logic and Algebraic Programming*, 64/1:135–154, 2004.