The Method of Shrink Wrapping For the Validated Solution of ODEs

MSU Report MSUHEP-20510

May 12, 2002

Kyoko Makino
Department of Physics
University of Illinois Urbana-Champaign
Martin Berz
Department of Physics and Astronomy
Michigan State University

In this note, we outline one method to perform shrink wrapping to eliminate the remainder term of the Taylor model integration of ODEs. We point out that there are many variants of this approach, and while the one shown here is perhaps the simplest one to understand, it is not necessarily the optimal choice. More about other possibilities at the end of the section.

After the n-th step of the integration, the region occupied by the final variables is given by the set

$$A = \left\{ \bigcup_{x^{(i)} \in B} \mathcal{M}(x^{(i)}) \right\} + I$$

where $x^{(i)}$ are the initial variables, B is the original box of initial conditions, $\mathcal{M}$ is the polynomial part of the Taylor model, and I is the interval remainder bound; the sum is the conventional sum of sets. In the case of the COSY-VI integration, the map $\mathcal{M}$ is scaled such that the original box B is unity, i.e. $B = [-1, 1]^n$. The remainder bound I accounts for the local approximation error of the expansion in time carried out in the n-th step as well as floating point errors and potentially other accumulated errors from previous steps; it is usually very small. The purpose of shrink wrapping is to "absorb" the small remainder interval into a set very similar to the first part via

$$A \subset A^* = \left\{ \bigcup_{x^{(i)} \in B} \mathcal{M}^*(x^{(i)}) \right\} + I^*$$

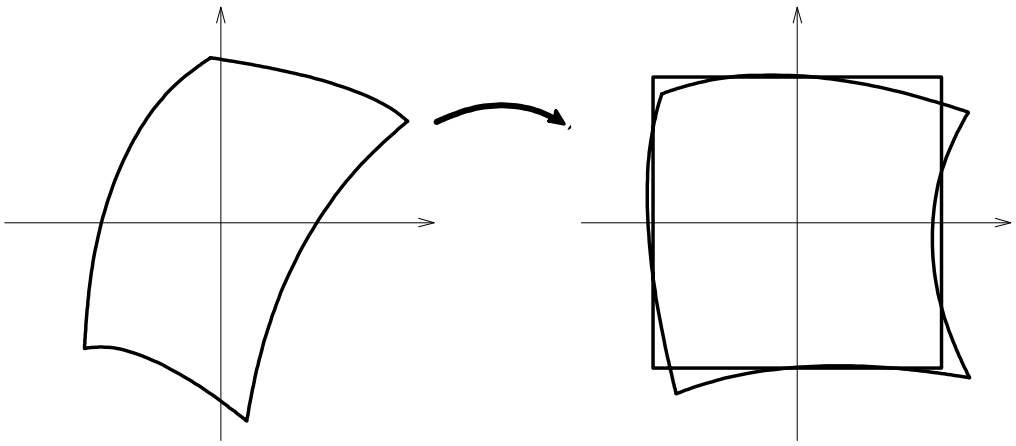
Where $\mathcal{M}^*$ is a slightly modified polynomial, and I^* is a significantly reduced interval of the size of machine precision.

As the first step, we extract the linear part M of $\mathcal{M}$ and determine a floating point approximation $\bar{M}^{-1}$ of its inverse. In case the ODEs admit unique solutions, as is typically the case for the problems at hand, the attempt to numerically invert the linear map M will likely succeed.

After the approximate inverse $\bar{M}^{-1}$ has been determined, we apply it from the left to the Taylor model $\mathcal{M}(x^{(i)}) + I$ that describes the current flow. As a result, the linear part of the resulting Taylor model is near identity. We write this new 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 $d \cdot [-1, 1]^n$, where d is a small number.



Definition

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 $d \cdot [-1, 1]^n$. We set

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

We call a map $\mathcal{M}$ shrinkable if $(1-nt) > 0$ and $(1 - s) > 0$; both of which are assured if $\mathcal{S}$ (and since it is a polynomial, hence also its derivative) is sufficiently small. Then we define q, the so-called shrink wrapping factor, as

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

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

(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, 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 estimate relating image distances to origin distances.

Lemma

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

$$\begin{aligned} &\|\mathcal{M}(\bar{x}) - \mathcal{M}(x)\| \leq (1+nt) \cdot \|\bar{x} - x\| \text{ and} \\ &\|\mathcal{M}(\bar{x}) - \mathcal{M}(x)\| \geq (1-nt) \cdot \|\bar{x} - x\| . \end{aligned}$$

Proof

For the proof of the first assertion, we trivially observe

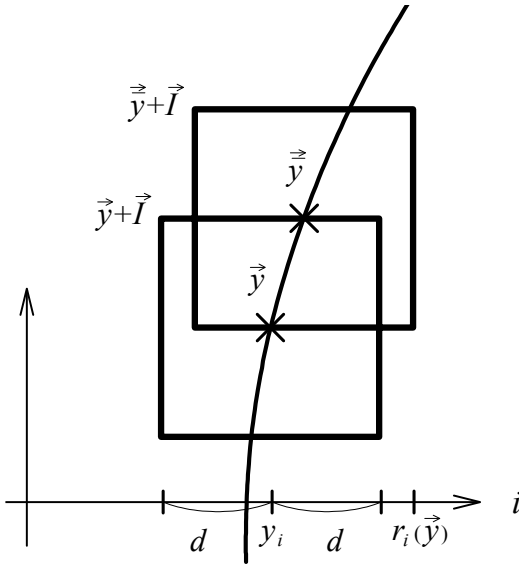
$$\begin{aligned} \|\mathcal{M}(\bar{x}) - \mathcal{M}(x)\| &= \max_i |\mathcal{M}_i(\bar{x}) - \mathcal{M}_i(x)| \\ &\leq \max_i \sum_j |\delta_{i,j} + t| \;\; |\bar{x}_j - x_j| \\ &\leq (1+nt) \;\; \|\bar{x} - x\| \end{aligned}$$

For the proof of the second 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

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

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

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



Hence the left term in the right hand absolute value in eq.(1) dominates the right term for any set of c_j, and we thus have

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

which completes the proof.

Theorem

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

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

be the set sum of the interval $I = [-d, d]^n$ and the range of $\mathcal{M}$ over the original domain box B. So R is the range enclosure of the flow of the ODE over the interval B provided by the Taylor model. Let q be the shrink wrap factor of $\mathcal{M}$; then we have

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

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

Proof

Let $1 \leq i \leq n$ 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 $(n-1)$ dimensional surface set $(\ x_1, ..., x_n)$ 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 $R = I + \bigcup_{x^{(i)} \in B} \mathcal{M}(x)$ can extend beyond the surface in direction i at the surface point $y = \mathcal{M}(x_1, ..., x_{i-1}, 1, x_{i+1}, ..., x_n)$.

Let y_i be the i-th component of y. The i-th components of the set $y + I$ apparently extend 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 the set R can extend beyond y_i at the point y by the amount

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

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 those cases with i-th component equal 1. We now project to the $(n-1)$ dimensional subspace with $x_i = 1$; for notational clarity, we denote the $(n-1)$ dimensional projection of $\mathcal{M}$ by $\mathcal{M}^{(i)}$, and similarly denote all $(n-1)$ 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 the factors s and t projecting down from $\mathcal{M}$. Apparently the condition on $\bar{y}$ in the definition of $r_i(y)$ entails that in the $(n-1)$ dimensional subspace, $\left\|\bar{y}^{(i)} - y^{(i)}\right\| \leq d$. Let $\bar{x}^{(i)}$ and $x^{(i)}$ be the $(n-1)$ 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

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

which entails that we can bound $r_i(y)$ in the $(n-1)$ dimensional space via $r_i(y) \leq d + \sup_{\left\|\bar{x}^{(i)} - x^{(i)}\right\| \leq d/(1-(n-1)t)} \mathcal{M}_i^{(i)}(\bar{x}^{(i)})$We now employ that $\left\|\mathcal{M}^{(i)}(\bar{x}^{(i)}) - \mathcal{M}^{(i)}(x^{(i)})\right\| \leq (1 + (n-1)t) \cdot \left\|\bar{x}^{(i)} - x^{(i)}\right\|$ and have

$$r_i(y) \leq y_i + d \cdot \frac{1 + (n-1)t}{1 - (n-1)t}\ .$$

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

To complete the proof, we observe that because of the bound s on $\mathcal{S}$, the box $(1-s)[-1, 1]^n$ lies entirely in the range R of the Taylor model. Thus multiplying the map $\mathcal{M}$ with any factor $q > 1$ entails that the edges of the box $(1-s)[-1, 1]^n$

move out by the amount $(1-s)(q-1)$ in all directions. Since the box is entirely inside R, this also means that the border of R moves out by at least the same amount in any direction i. Thus choosing q as

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

assures that

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

as advertised.