Taylor Models and the Validated ODE Integration

Martin Berz and Kyoko Makino

Department of Physics and Astronomy
Michigan State University

Department of Physics
University of Illinois at Urbana-Champaign

Introduction

Taylor model (TM) methods were originally developed for a practical problem from nonlinear dynamics, range bounding of normal form defect functions.

- Functions consist of code lists of 10^4 to 10^5 terms
- Have about the worst imaginable cancellation problem
- Are obtained via validated integration of large initial condition boxes.

Originally nearly universally considered intractable by the community. But ... a small challenge goes a long way towards generating new ideas!

Idea: represent all functional dependencies as a pair of a polynomial P and a remainder bound I, introduce arithmetic, and a new ODE solver. Obtain the following properties:

- The ability to provide enclosures of any function given by a finite computer code list by a Taylor polynomial and a remainder bound with a sharpness that scales with order $(n+1)$ of the width of the domain.
- The ability to alleviate the dependency problem in the calculation.
- The ability to scale favorable to higher dimensional problems.

Definitions - Taylor Models and Operations

We begin with a review of the definitions of the basic operations.

Definition (Taylor Model) Let $f : D \subset R^v \to R$ be a function that is $(n+1)$ times continuously partially differentiable on an open set containing the domain v-dimensional domain D. Let x_0 be a point in D and P the n-th order Taylor polynomial of f around x_0. Let I be an interval such that

$$f(x) \in P(x - x_0) + I \text{ for all } x \in D.$$

Then we call the pair (P, I) an n-th order Taylor model of f around x_0 on D.

Definition (Addition and Multiplication) Let $T_{1,2} = (P_{1,2}, I_{1,2})$ be n-th order Taylor models around x_0 over the domain D. We define

$$T_1 + T_2 = (P_1 + P_2, I_1 + I_2)$$
$$T_1 \cdot T_2 = (P_{1\cdot 2}, I_{1\cdot 2})$$

where $P_{1\cdot 2}$ is the part of the polynomial $P_1 \cdot P_2$ up to order n and

$$I_{1\cdot 2} = B(P_e) + B(P_1) \cdot I_2 + B(P_2) \cdot I_1 + I_1 \cdot I_2$$

where P_e is the part of the polynomial $P_1 \cdot P_2$ of orders $(n+1)$ to $2n$, and $B(P)$ denotes a bound of P on the domain D. We demand that $B(P)$ is at least as sharp as direct interval evaluation of $P(x - x_0)$ on D.

Definitions - Taylor Model Intrinsics

Definition (Intrinsic Functions of Taylor Models) Let $T = (P, I)$ be a Taylor model of order n over the v-dimensional domain $D = [a, b]$ around the point x_0. We define intrinsic functions for the Taylor models by performing various manipulations that will allow the computation of Taylor models for the intrinsics from those of the arguments. In the following, let $f(x) \in P(x - x_0) + I$ be any function in the Taylor model, and let $c_f = f(x_0)$, and $\bar{f}$ be defined by $\bar{f}(x) = f(x) - c_f$. Likewise we define $\bar{P}$ by $\bar{P}(x - x_0) = P(x - x_0) - c_f$, so that $(\bar{P}, I)$ is a Taylor model for $\bar{f}$. For the various intrinsics, we proceed as follows.

Exponential. We first write

$$\begin{aligned}\exp(f(x)) &= \exp\left(c_f + \bar{f}(x)\right) = \exp(c_f) \cdot \exp\left(\bar{f}(x)\right) \\ &= \exp(c_f) \cdot \left\{1 + \bar{f}(x) + \frac{1}{2!}(\bar{f}(x))^2 + \cdots + \frac{1}{k!}(\bar{f}(x))^k \right. \\ &\qquad \left. + \frac{1}{(k+1)!}(\bar{f}(x))^{k+1} \exp\left(\theta \cdot \bar{f}(x)\right)\right\},\end{aligned}$$

where $0 < \theta < 1$.

Definitions - Taylor Model Exponential, cont.

Taking $k \geq n$, the part

$$\exp(c_f) \cdot \left\{ 1 + \bar{f}(x) + \frac{1}{2!}(\bar{f}(x))^2 + \cdots + \frac{1}{n!}(\bar{f}(x))^n \right\}$$

is merely a polynomial of $\bar{f}$, of which we can obtain the Taylor model via Taylor model addition and multiplication. The remainder part of $\exp(f(x))$, the expression

$$\exp(c_f) \cdot \left\{ \frac{1}{(n+1)!}(\bar{f}(x))^{n+1} + \cdots + \frac{1}{(k+1)!}(\bar{f}(x))^{k+1} \exp\left(\theta \cdot \bar{f}(x)\right) \right\},$$

will be bounded by an interval. First observe that since the Taylor polynomial of $\bar{f}$ does not have a constant part, the $(n+1)$-st through $(k+1)$-st powers of the Taylor model $(\bar{P}, I)$ of $\bar{f}$ will have vanishing polynomial part, and thus so does the entire remainder part. The remainder bound interval for the Lagrange remainder term

Definitions - Taylor Model Exponential, cont.

$$\exp(c_f)\frac{1}{(k+1)!}(\bar{f}(x))^{k+1}\exp\left(\theta\cdot\bar{f}(x)\right)$$

can be estimated because, for any $x \in D$, $\bar{P}(x-x_0) \in B(\bar{P})$, and $0 < \theta < 1$, and so

$$(\bar{f}(x))^{k+1}\exp\left(\theta\cdot\bar{f}(x)\right) \in \left(B(\bar{P})+I\right)^{k+1} \times \exp\left([0,1]\cdot(B(\bar{P})+I)\right).$$

The evaluation of the "exp" term is mere standard interval arithmetic. In the actual implementation, one may choose $k = n$ for simplicity, but it is not a priori clear which value of k would yield the sharpest enclosures.

Definitions - Taylor Model Arc Sine

Arcsine. Under the condition $\forall x \in D,\ B(P(x-x_0)+I) \subset (-1,1)$, using an addition formula for the arcsine, we re-write

$$\arcsin(f(x)) = \arcsin(c_f) + \arcsin\left(f(x)\cdot\sqrt{1-c_f^2} - c_f\cdot\sqrt{1-(f(x))^2}\right).$$

Utilizing that

$$g(x) \equiv f(x)\cdot\sqrt{1-c_f^2} - c_f\cdot\sqrt{1-(f(x))^2}$$

does not have a constant part, we have

$$\begin{aligned}\arcsin(g(x)) = g(x) + \frac{1}{3!}(g(x))^3 + \frac{3^2}{5!}(g(x))^5 + \frac{3^2\cdot 5^2}{7!}(g(x))^7 \\ + \cdots + \frac{1}{(k+1)!}(g(x))^{k+1}\cdot\arcsin^{(k+1)}(\theta\cdot g(x)),\end{aligned}$$

where

$$\arcsin'(a) = 1/\sqrt{1-a^2}, \qquad \arcsin''(a) = a/(1-a^2)^{3/2},$$
$$\arcsin^{(3)}(a) = (1+2a^2)/(1-a^2)^{5/2}, ...$$

Definitions - Taylor Model Arc Sine, Antiderivation

A recursive formula for the higher order derivatives of arcsin

$$\arcsin^{(k+2)}(a) = \frac{1}{1-a^2}\{(2k+1)a\arcsin^{(k+1)}(a) + k^2\arcsin^{(k)}(a)\}$$

is useful. Then, evaluating in Taylor model arithmetic yields the desired result, where again the terms involving θ only produce interval contributions.

Antiderivation. We note that a Taylor model for the integral with respect to variable i of a function f can be obtained from the Taylor model (P, I) of the function by merely integrating the part P_{n-1} of order up to $n-1$ of the polynomial, and bounding the n-th order into the new remainder bound. Specifically, we have

$$\partial_i^{-1}(P, I) = \left(\int_0^{x_i} P_{n-1}(x)dx_i\,,\ (B(P - P_{n-1}) + I) \cdot (b_i - a_i)\right).$$

Thus, given a Taylor model for a function f, the Taylor model intrinsic functions produce a Taylor models for the composition of the respective intrinsic with f. Furthermore, we have the following result.

TM Scaling Theorem

Theorem (Scaling Theorem) Let f, $g \in C^{n+1}(D)$ and $(P_{f,h}, I_{f,h})$ and $(P_{g,h}, I_{g,h})$ be n-th order Taylor models for f and g around x_h on $x_h + [-h, h]^v \subset D$. Let the remainder bounds $I_{f,h}$ and $I_{g,h}$ satisfy $I_{f,h} = O(h^{n+1})$ and $I_{g,h} = O(h^{n+1})$. Then the Taylor models $(P_{f+g}, I_{f+g,h})$ and $(P_{f\cdot g}, I_{f\cdot g,h})$ for the sum and products of f and g obtained via addition and multiplication of Taylor models satisfy

$$I_{f+g,h} = O(h^{n+1}), \text{ and } I_{f\cdot g,h} = O(h^{n+1}).$$

Furthermore, let s be any of the intrinsic functions defined above, then the Taylor model $(P_{s(f)}, I_{s(f),h})$ for $s(f)$ obtained by the above definition satisfies

$$I_{s(f),h} = O(h^{n+1}).$$

We say the Taylor model arithmetic has the $(n+1)$-st order scaling property.

Proof. The proof for the binary operations follows directly from the definition of the remainder bounds for the binaries. Similarly, the proof for the intrinsics follows because all intrinsics are composed of binary operations as well as an additional interval, the width of which scales at least with the $(n+1)$-st power of a bound B of a function that scales at least linearly with h.

Fundamental Theorem of TM Arithmetic

The scaling theorem states that a given function f can be approximated by P with an error that scales with order $(n+1)$. Common mathematical jargon. But in interval community, a related but different meaning of scaling exists, namely the behavior of the overestimation of a given method to determine the range of a function.

Theorem (FTTMA, Fundamental Theorem of TM Arithmetic) Let the function $f : R^v \to R^v$ be described by a multivariate Taylor model $P_f + I_f$ over the domain $D \subset R^v$. Let the function $g : R^v \to R$ be given by a code list comprised of finitely many elementary operations and intrinsic functions, and let g be defined over the range of the Taylor model $P_f, +I_f$. Let $P + I$ be the Taylor model obtained by executing the code list for g, beginning with the Taylor model $P_f + I_f$. Then $P + I$ is a Taylor model for $g \circ f$.

Furthermore, if the Taylor model of f has the $(n+1)$-st order scaling property, so does the resulting Taylor model for g.

Proof. Induction over code list.

Example: Consider f with $f(x) = \sin^2(\exp(x+1)) + \cos^2(\exp(x+1))$. We know $f(x) = 1$, but validated methods don't.

Implementation of TM Arithmetic

Validated Implementation of TM Arithmetic exists. The following points are important

- Strict requirements for **underlying FP arithmetic**
- Taylor models require cutoff threshold (**garbage collection**)
- Coefficients remain FP, not intervals
- Package quite **extensively tested** by Corliss et al.

For practical considerations, the following is important:

- Need **sparsity** support
- Need efficient coefficient **addressing** scheme
- About $50,000$ lines of code
- **Language Independent** Platform, coexistence in F77, C, F90, C++

TM Enclosure Theorem

Theorem (Taylor Model Enclosure Theorem) Let the function $f : R^v \to R^v$ be contained within $P_f + I_f$ over the domain $D \subset R^v$. Let the function $g : R^v \to R$ be given by a code list comprised of finitely many elementary operations and intrinsic functions, and let g be defined over the range of an enclosure of $P_f, +I_f$. Let $P + I$ be the result obtained by executing the code list for g in admissible FP Taylor model arithmetic, beginning with the Taylor model $P_f + I_f$. Then $P + I$ is an enclosure for $g \circ f$ over D.

Proof The proof follows by induction over the code list of g from the elementary properties of the Taylor model arithmetic.

Apparently the presence of the floating point errors entails that P is not precisely the **Taylor polynomial**. In a similar fashion, also the **scaling property** of the remainder bound in a rigorous sense is lost. However, these properties of Taylor models are **retained in an approximate fashion**.

Important TM Algorithms

- **Range Bounding** (Evaluate f as TM, bound polynomial, add remainder bound)
- **Quadrature** (Evaluate f as TM, integrate polynomial and remainder bound)
- **Implicit Equations** (Obtain TMs for implicit solutions of TM equations)
- **Superconvergent** Interval Newton Method (Application of Implicit Equations)
- **ODEs** (Obtain TMs describing dependence of final coordinates on initial coordinates)
- **Implicit ODEs and DAEs**
- **Complex Arithmetic** (Describe complex ranges as two-dimensional TMs)

Examples for TM Bounding - Method

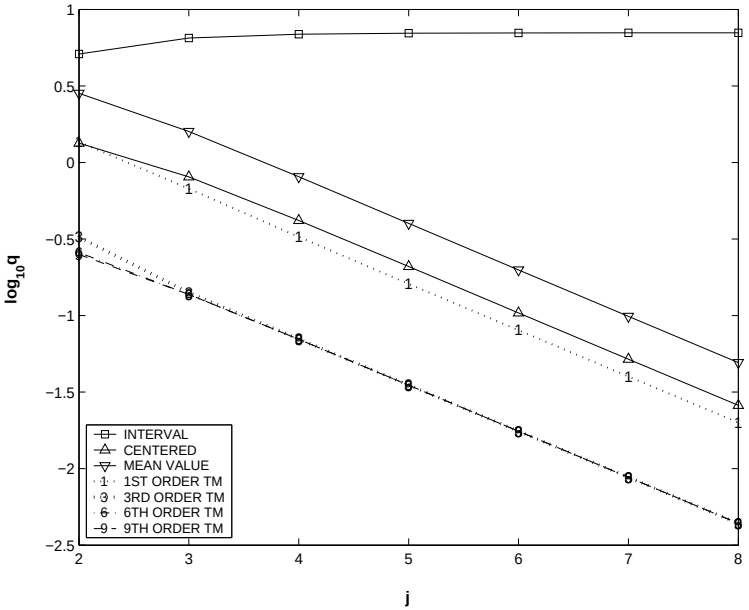
- Comparison of bounding with Taylor Model (**TM),** Interval (**I**), Centered Form (**CF**), Mean Value Form (**MF**)
- Used **COSY INFINITY** for TMs
- Used **INTLAB** Version 3.1 under Matlab Version 6 for others
- Study $x_0 + [-2^{-j}, +2^{-j}]$ for different values of j
- Estimate "true" range either by hand, or rastering with many points
- Determine **relative overestimation q** for each method
- Determine **empirical approximation order EAO** for each method

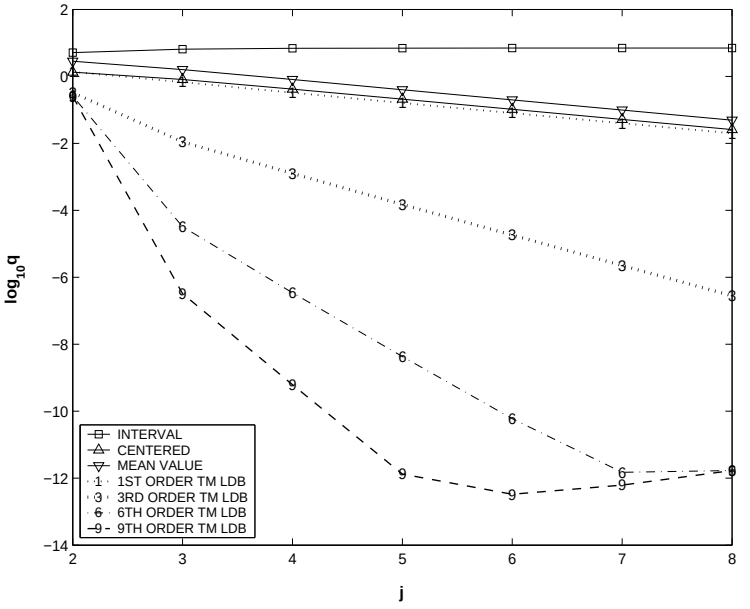
A Simple 3D Test Function

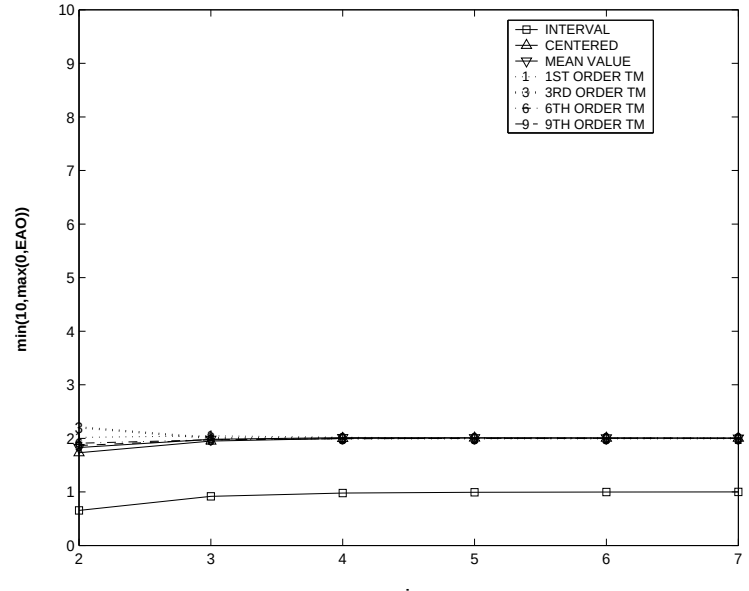
- Rather **innocent** function in three variables
- Can show dependency suppression even in simple cases
- To increase dependency, also consider function f_2

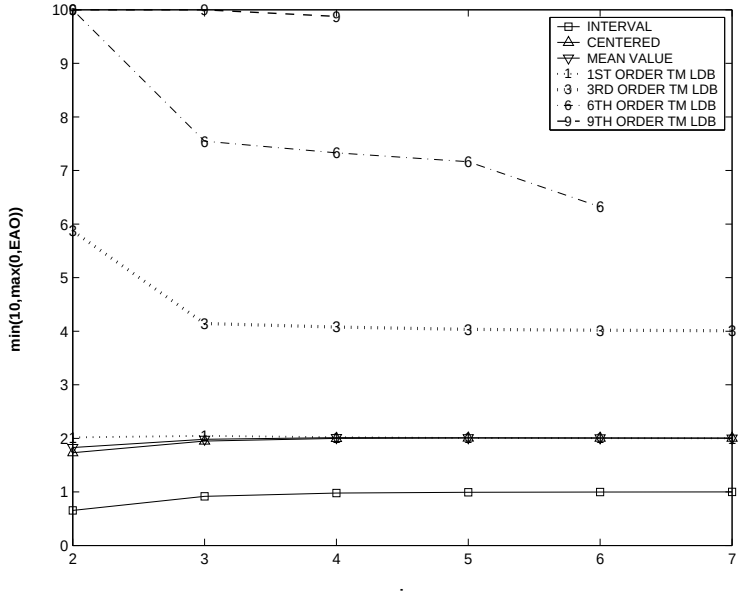
$$
\begin{aligned}
f_1(x,y,z) = {} & \frac{4\tan(3y)}{3x + x\sqrt{\frac{6x}{-7(x-8)}}} - 120 - 2x - 7z(1+2y) \\
& - \sinh\left(0.5 + \frac{6y}{8y+7}\right) + \frac{(3y+13)^2}{3z} \\
& - 20z(2z-5) + \frac{5x\tanh(0.9z)}{\sqrt{5y}} - 20y\sin(3z),
\end{aligned}
$$

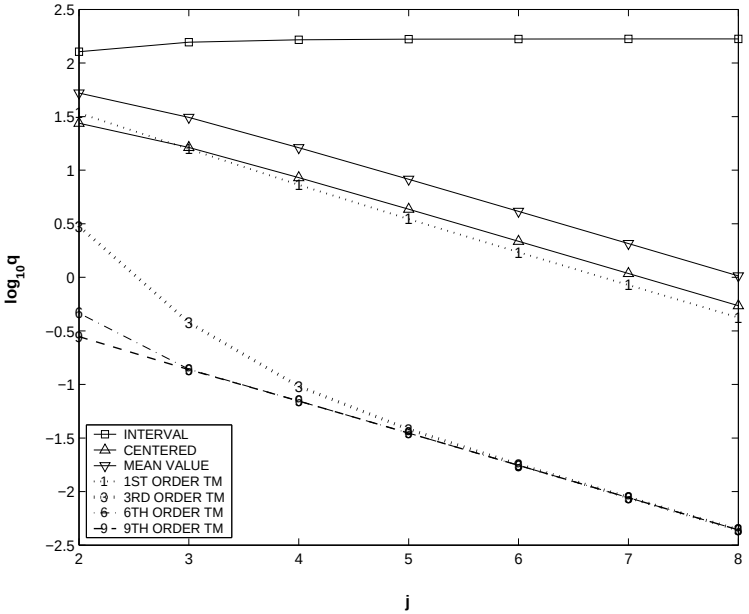
$$
f_2(x,y,z) = f_1(x,y,z) + \sum_{j=1}^{10} \left(f_1(x,y,z) - f_1(x,y,z)\right),
$$

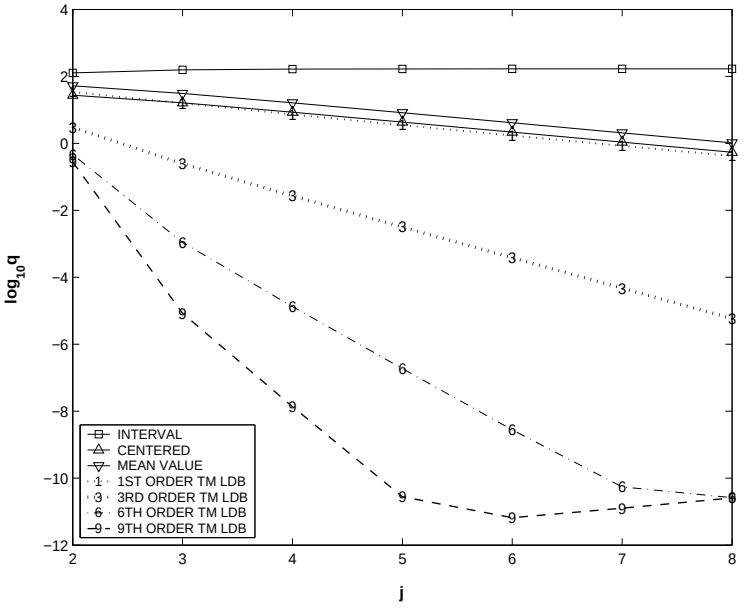

log10q
j
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM
3RD ORDER TM
6TH ORDER TM
9TH ORDER TM

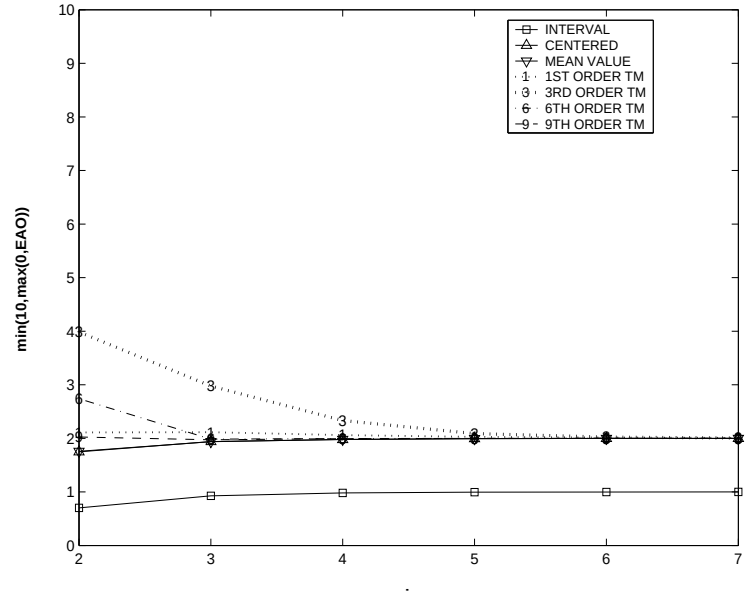

log10q
j
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM LDB
3RD ORDER TM LDB
6TH ORDER TM LDB
9TH ORDER TM LDB

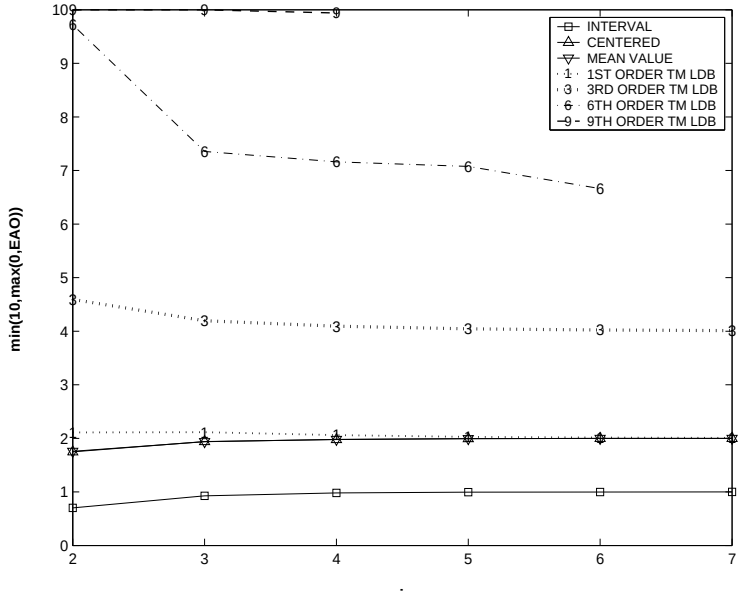

min(10,max(0,EAO))
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM
3RD ORDER TM
6TH ORDER TM
9TH ORDER TM


min(10,max(0,EAO))
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM LDB
3RD ORDER TM LDB
6TH ORDER TM LDB
9TH ORDER TM LDB


log10q
j
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM
3RD ORDER TM
6TH ORDER TM
9TH ORDER TM


log10q
j
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM LDB
3RD ORDER TM LDB
6TH ORDER TM LDB
9TH ORDER TM LDB


min(10,max(0,EAO))
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM
3RD ORDER TM
6TH ORDER TM
9TH ORDER TM


min(10,max(0,EAO))
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM LDB
3RD ORDER TM LDB
6TH ORDER TM LDB
9TH ORDER TM LDB

Gritton's Function

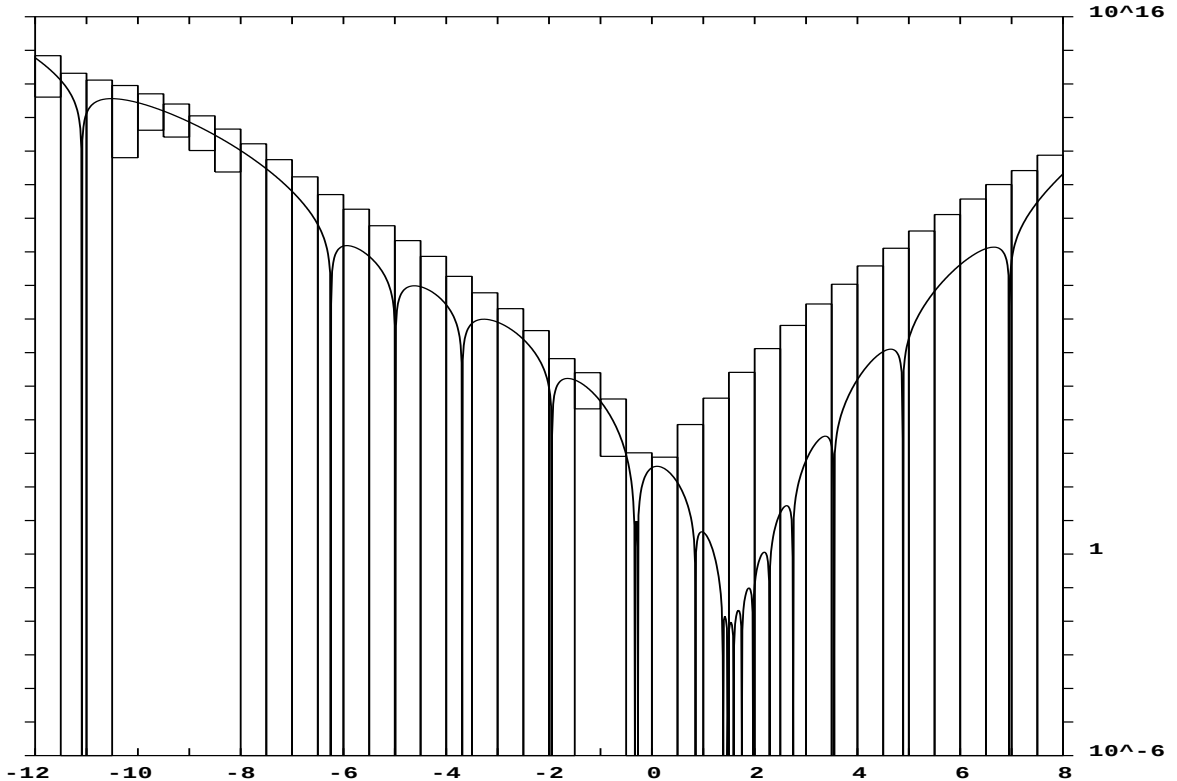
- Small function values, but very **strong cancellation problem**
- "Simply" a polynomial of 18 th order, has 18 zeros
- Particularly difficult near $x = 1.4$
- Consider behavior around $x_0 = 2$ and $x_0 = 1.4$

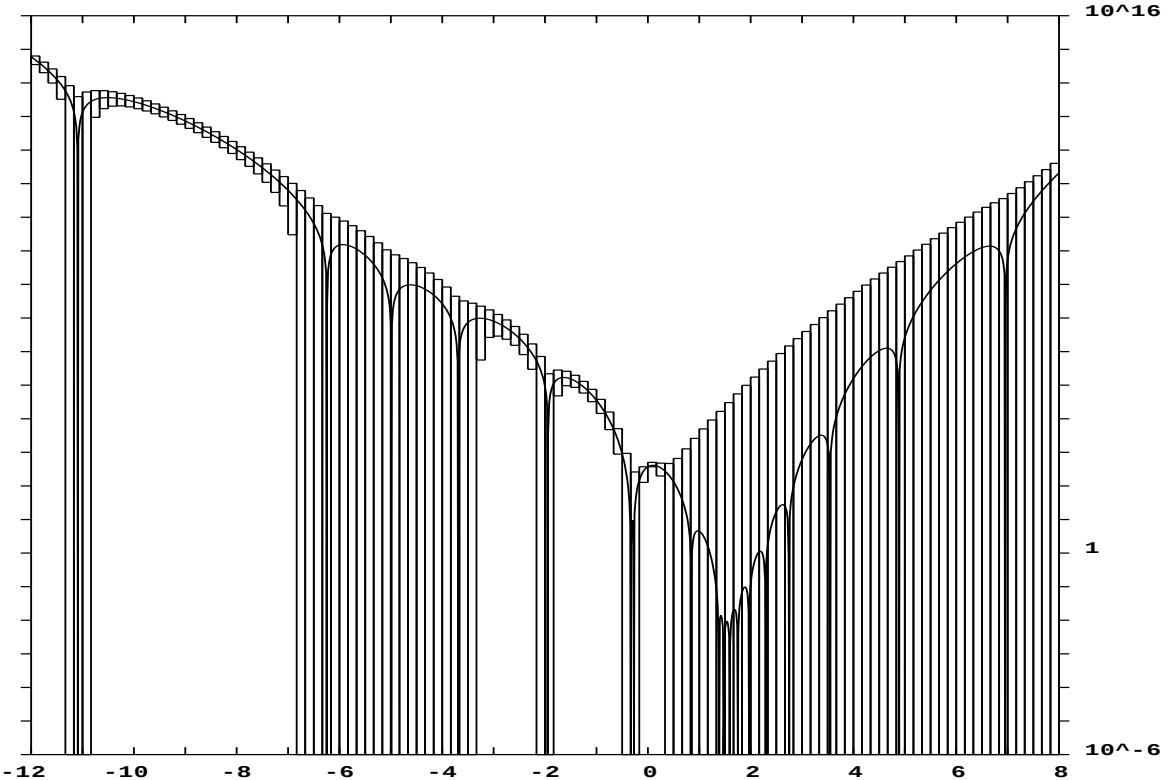
$$
\begin{aligned}
f_3(x) = -371.9362500 + x \cdot (-791.2465656 + x \cdot (4044.944143 \\
+ x \cdot (978.1375167 + x \cdot (-16547.89280 + x \cdot (22140.72827 \\
+ x \cdot (-9326.549359 + x \cdot (-3518.536872 + x \cdot (4782.532296 \\
+ x \cdot (-1281.479440 + x \cdot (-283.4435875 + x \cdot (202.6270915 \\
+ x \cdot (-16.17913459 + x \cdot (-8.883039020 + x \cdot (1.575580173 \\
+ x \cdot (1245990848 + x \cdot (-0.03589148622 + x \cdot (-0.0001951095576 \\
+ x \cdot (0.0002274682229)))))))))))))))))).
\end{aligned}
$$

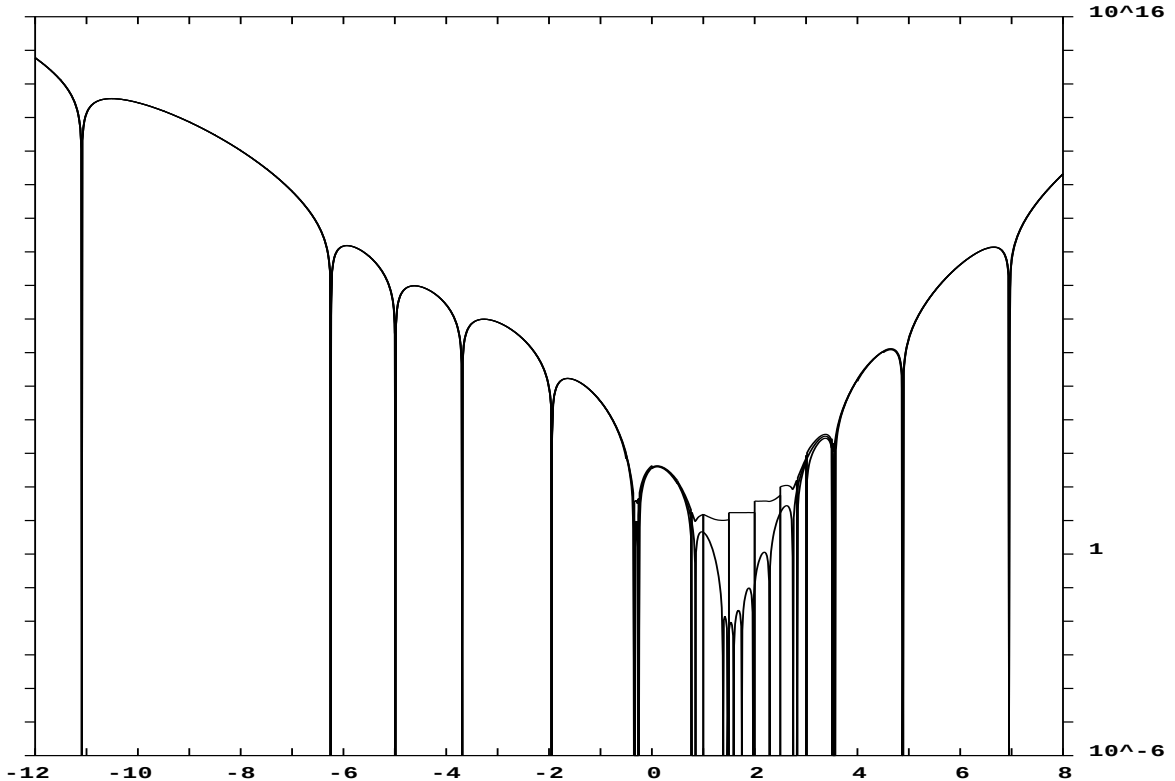
Gritton's Function - Visualization

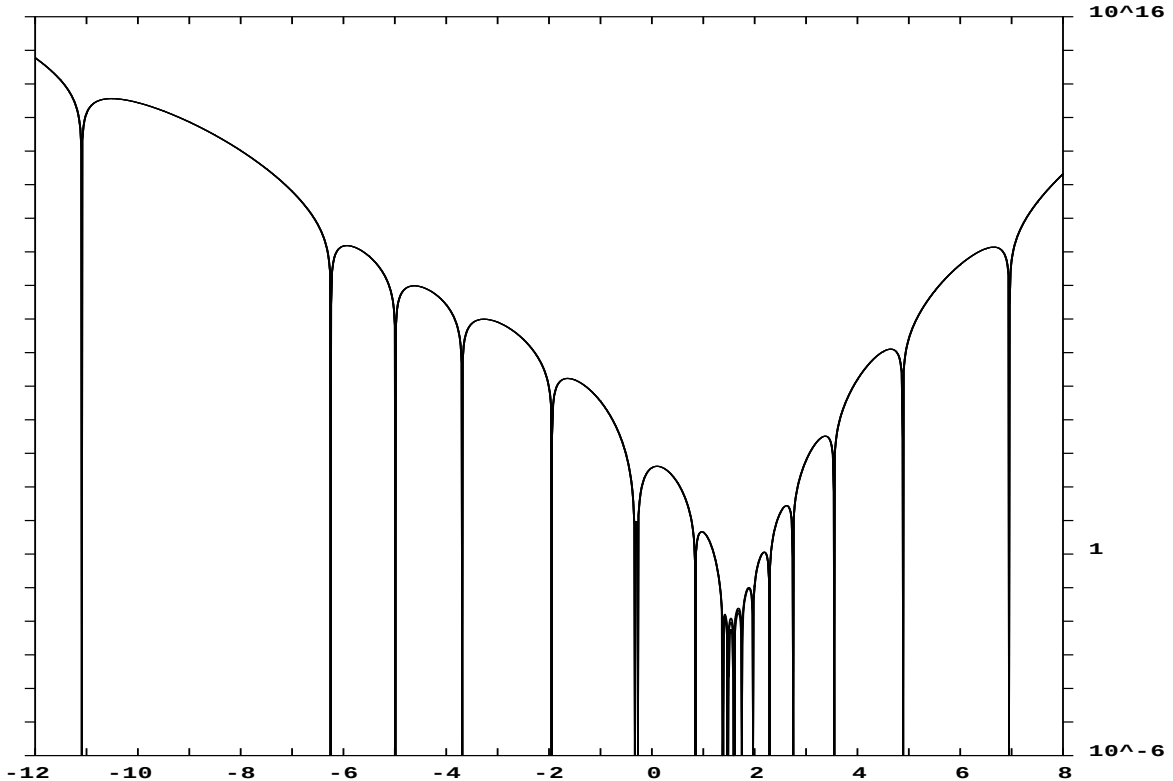
To visualize the behavior of Gritton's function is difficult. We choose the following method

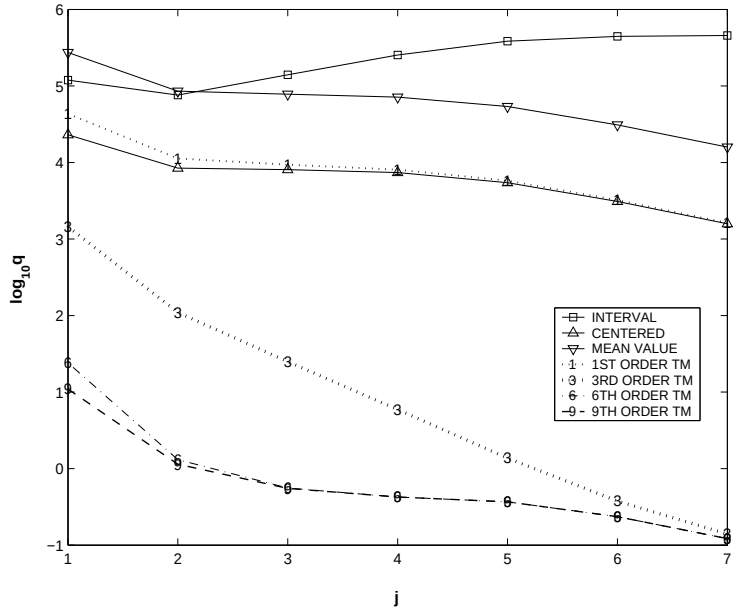
- showing the function values logarithmically
- show interval enclosures by 40 and 120 subdivisions
- show TM enclosures by 40 TMs of fourth order and eigth order
- The 40 eigth order TMs can separate all 18 zeros of the function

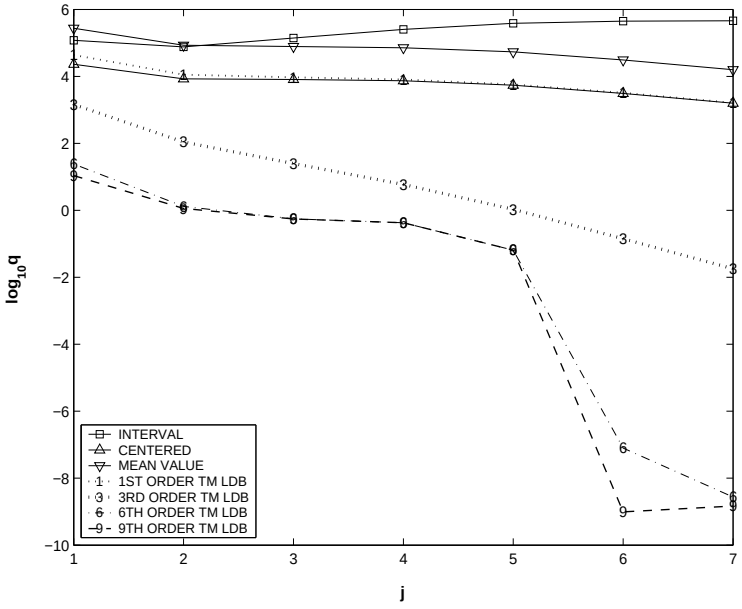

10^16
1
10^-6
-12
-10
-8
-6
-4
-2
0
2
4
6
8

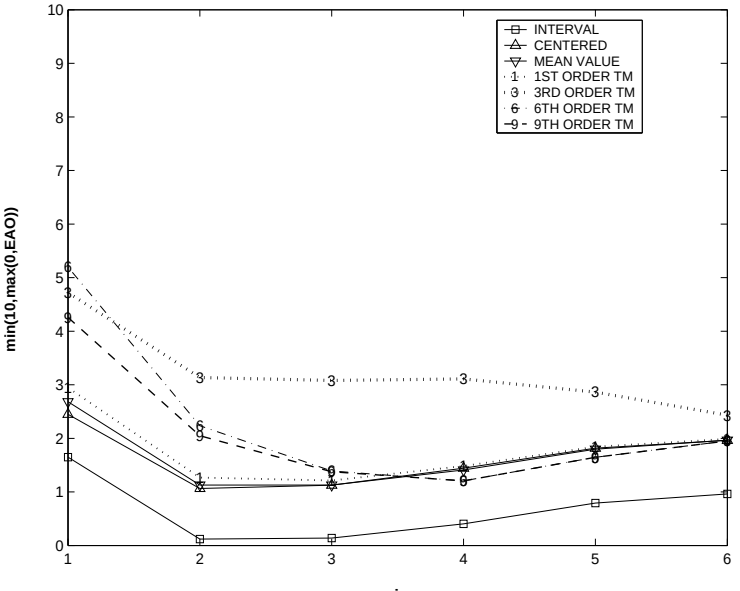

10^16
1
10^-6
-12
-10
-8
-6
-4
-2
0
2
4
6
8

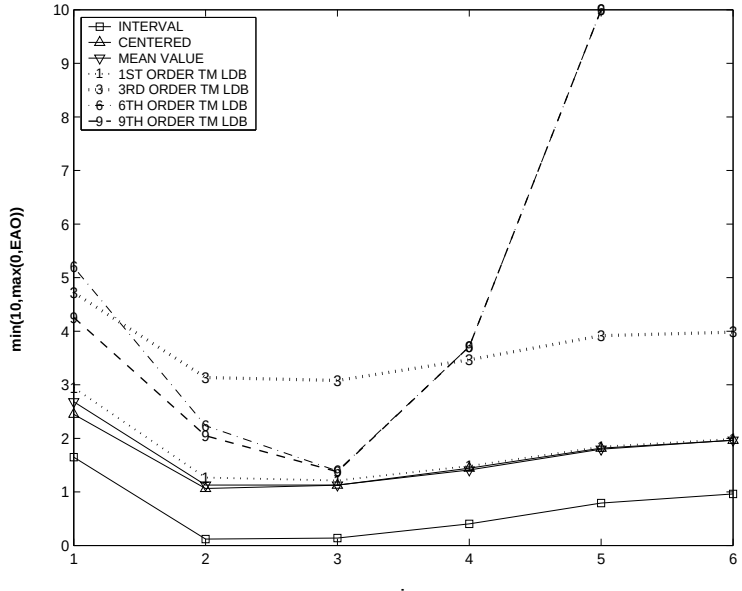

10^16
1
10^-6
-12
-10
-8
-6
-4
-2
0
2
4
6
8

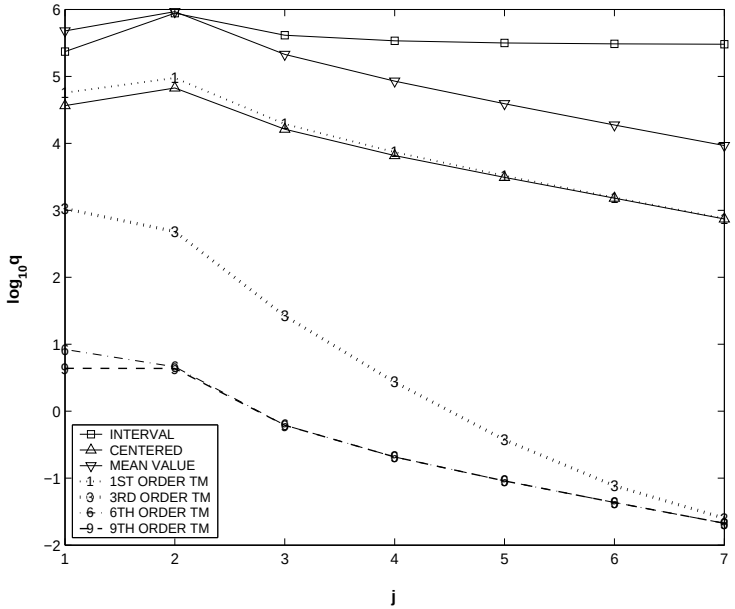

10^16
1
10^-6
-12
-10
-8
-6
-4
-2
0
2
4
6
8

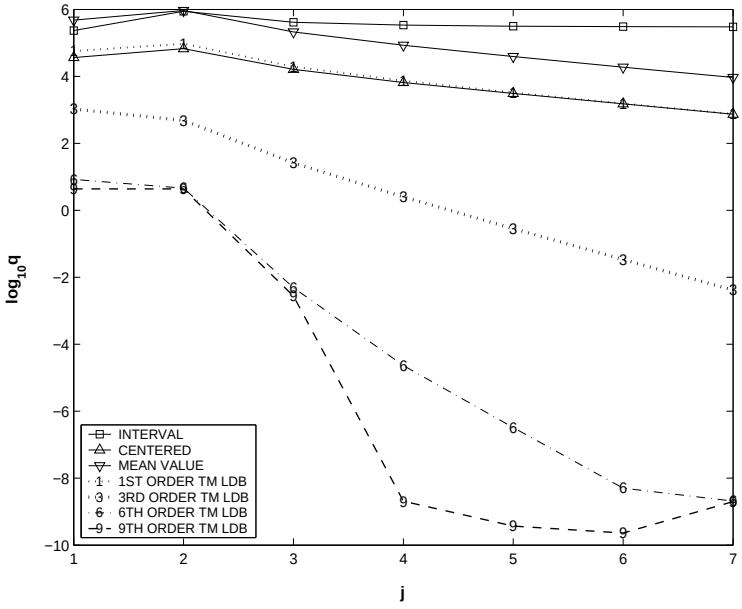

log10 q
j
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM
3RD ORDER TM
6TH ORDER TM
9TH ORDER TM

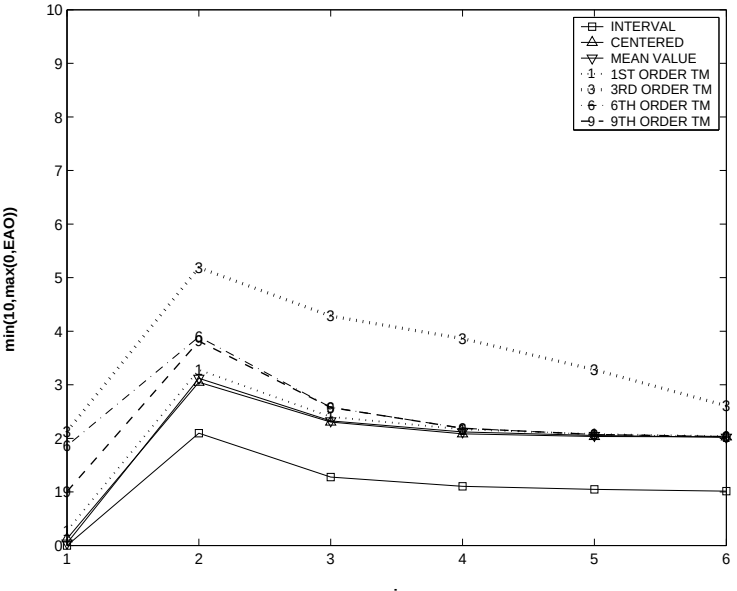

log10 q
j
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM LDB
3RD ORDER TM LDB
6TH ORDER TM LDB
9TH ORDER TM LDB

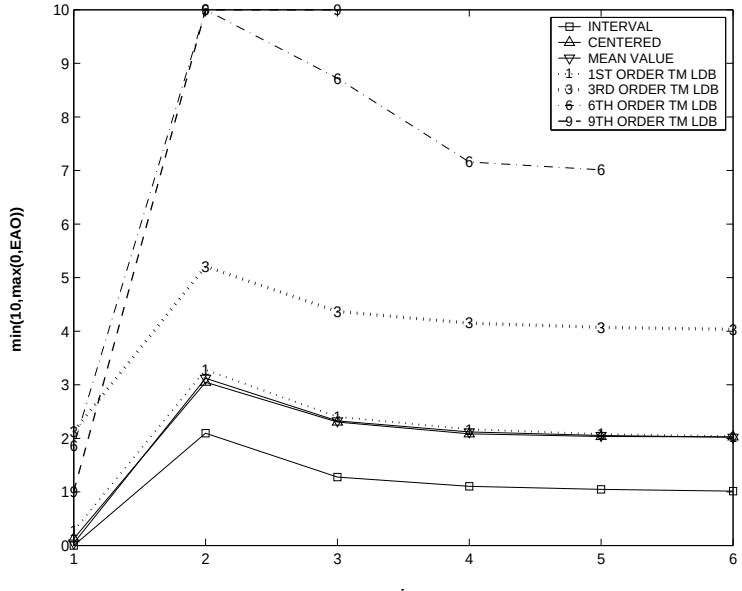

min(10,max(0,EAO))
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM
3RD ORDER TM
6TH ORDER TM
9TH ORDER TM


min(10,max(0,EAO))
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM LDB
3RD ORDER TM LDB
6TH ORDER TM LDB
9TH ORDER TM LDB


log10 q
j
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM
3RD ORDER TM
6TH ORDER TM
9TH ORDER TM


log10 q
j
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM LDB
3RD ORDER TM LDB
6TH ORDER TM LDB
9TH ORDER TM LDB


min(10,max(0,EAO))
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM
3RD ORDER TM
6TH ORDER TM
9TH ORDER TM

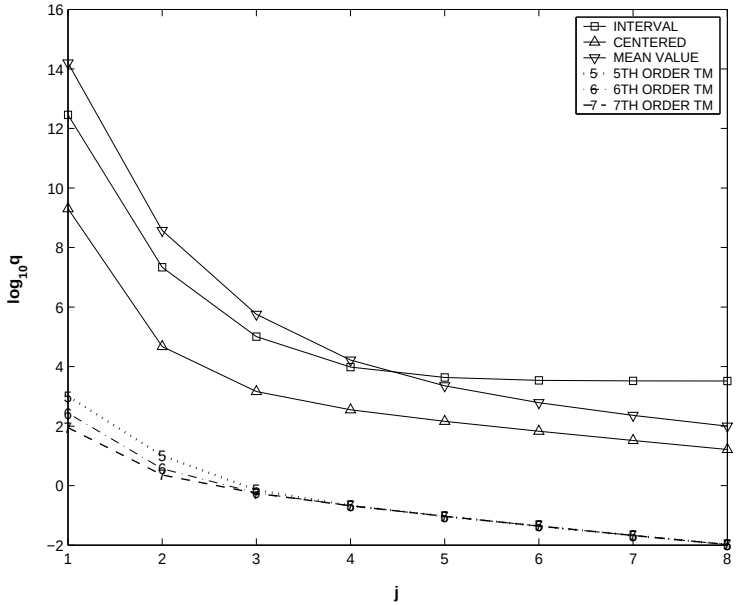

min(10,max(0,EAO))
INTERVAL
CENTERED
MEAN VALUE
1ST ORDER TM LDB
3RD ORDER TM LDB
6TH ORDER TM LDB
9TH ORDER TM LDB

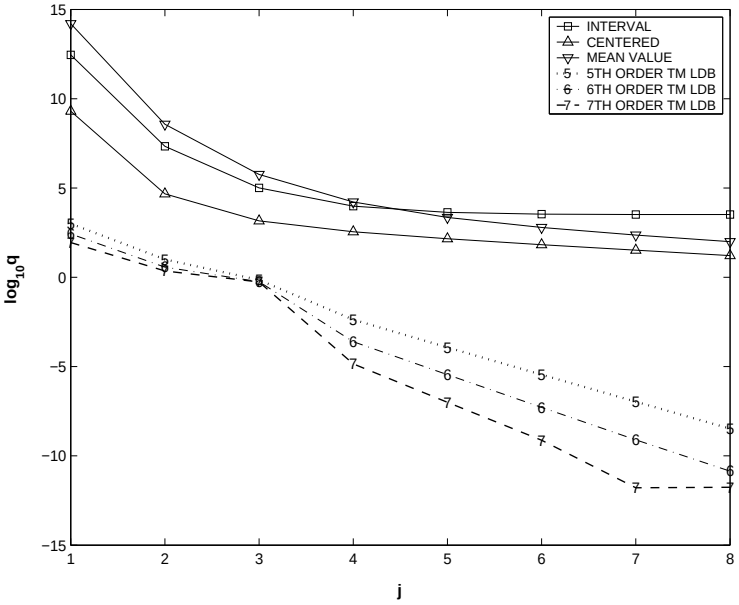
The Normal Form Defect Function

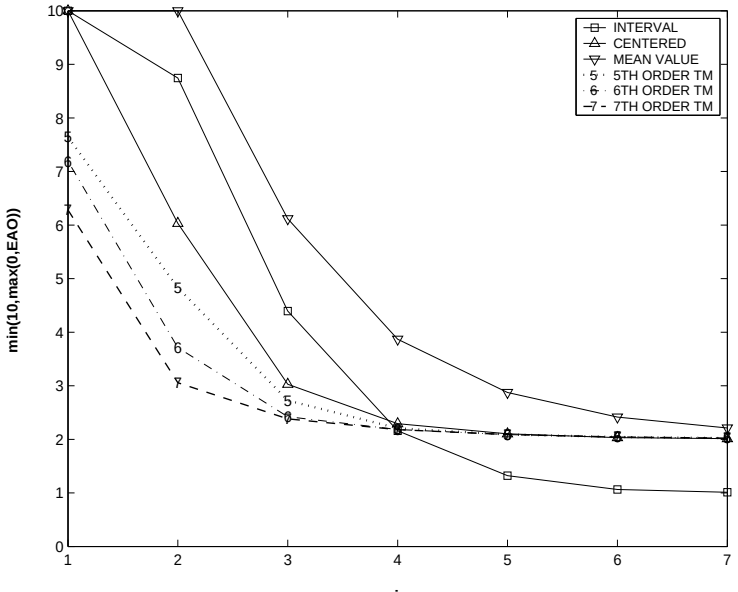
- **Extreme cancellation**; one of the reasons TM methods were invented
- Six-dimensional problem from dynamical systems theory
- Describes invariance defects of a particle accelerator
- Essentially composition of three tenth order polynomials
- The function vanishes identically to order ten
- Study for $a \cdot (1,1,1,1,1,1)$ for $a = .1$ and $a = .2$
- Interesting **Speed observation**: on same machine,
  * one CF in INTLAB takes 45 minutes
  * one TM of order 7 takes 10 seconds

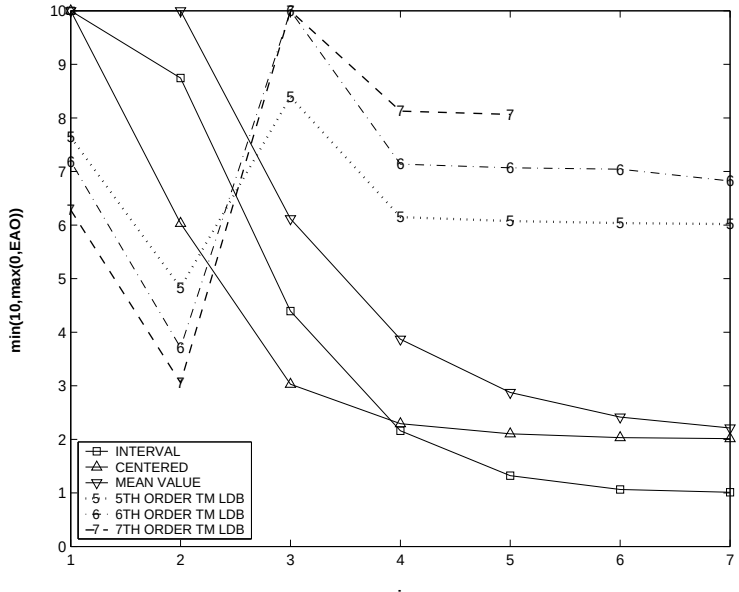
$$f_4(x_1,..,x_6) = \sum_{i=1}^{3} \left(\sqrt{y_{2i-1}^2 + y_{2i}^2} - \sqrt{x_{2i-1}^2 + x_{2i}^2} \right)^2$$

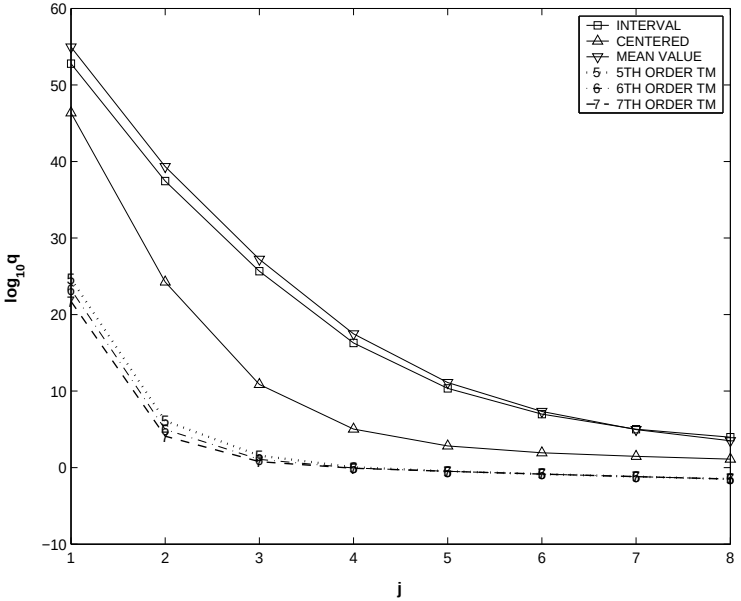
$$\text{where } \vec{y} = \vec{P}_1 \left(\vec{P}_2 \left(\vec{P}_3(\vec{x}) \right) \right)$$

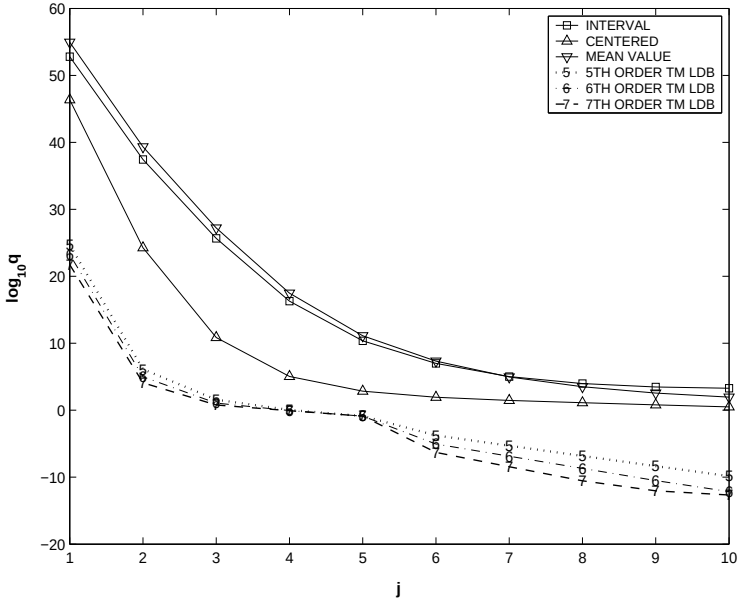

INTERVAL
CENTERED
MEAN VALUE
5TH ORDER TM
6TH ORDER TM
7TH ORDER TM
log10q
j

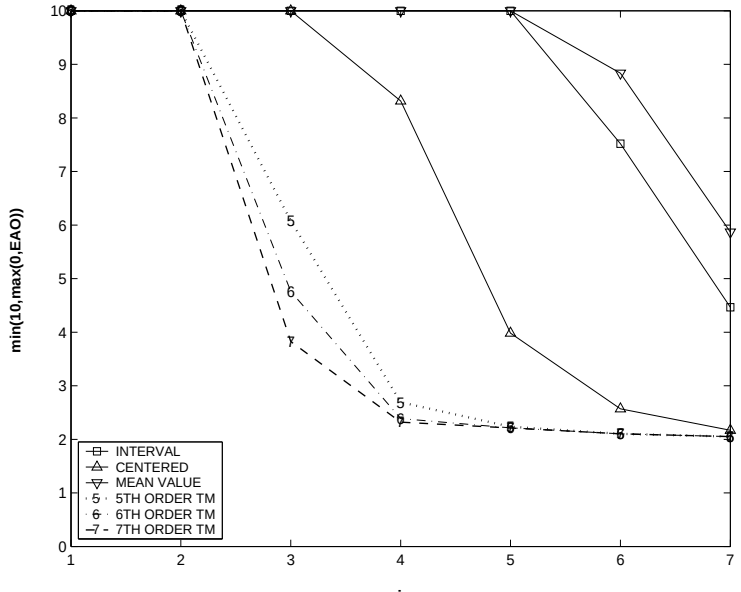

INTERVAL
CENTERED
MEAN VALUE
5TH ORDER TM LDB
6TH ORDER TM LDB
7TH ORDER TM LDB
log10q
j

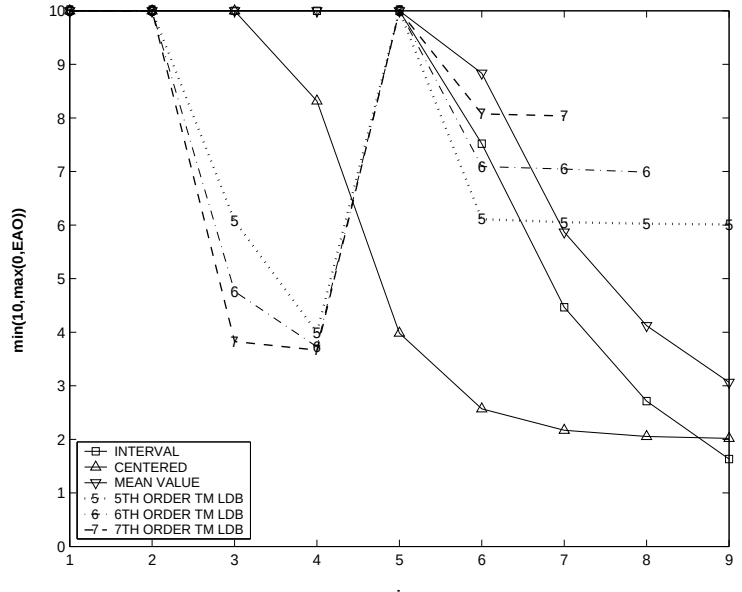

INTERVAL
CENTERED
MEAN VALUE
5TH ORDER TM
6TH ORDER TM
7TH ORDER TM
min(10,max(0,EAO))


INTERVAL
CENTERED
MEAN VALUE
5TH ORDER TM LDB
6TH ORDER TM LDB
7TH ORDER TM LDB
min(10,max(0,EAO))


INTERVAL
CENTERED
MEAN VALUE
5TH ORDER TM
6TH ORDER TM
7TH ORDER TM
log10 q
j


INTERVAL
CENTERED
MEAN VALUE
5TH ORDER TM LDB
6TH ORDER TM LDB
7TH ORDER TM LDB
log10 q
j


INTERVAL
CENTERED
MEAN VALUE
5TH ORDER TM
6TH ORDER TM
7TH ORDER TM
min(10,max(0,EAO))


INTERVAL
CENTERED
MEAN VALUE
5TH ORDER TM LDB
6TH ORDER TM LDB
7TH ORDER TM LDB
min(10,max(0,EAO))

Remainder Bounds from Interval AD

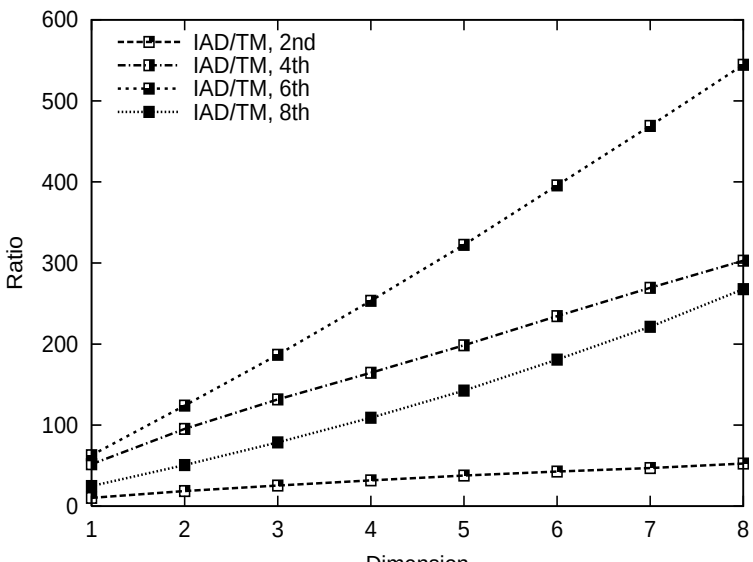
Use of AD has long history in interval analysis, goes back to Moore. One application: determine remainder bounds for Taylor expansion.

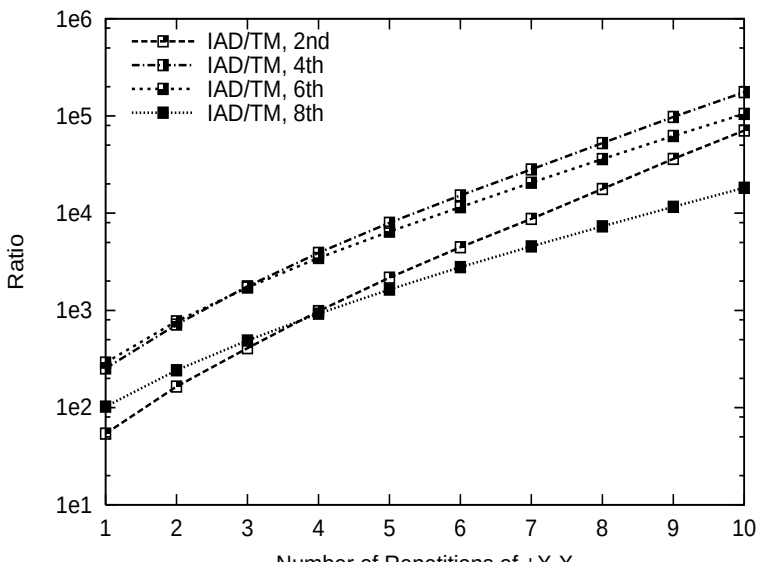
- Set up code list of function
- Evaluate code list with point initial condition and high-order AD
- Evaluate code list with interval initial condition to get bound for remainder bound

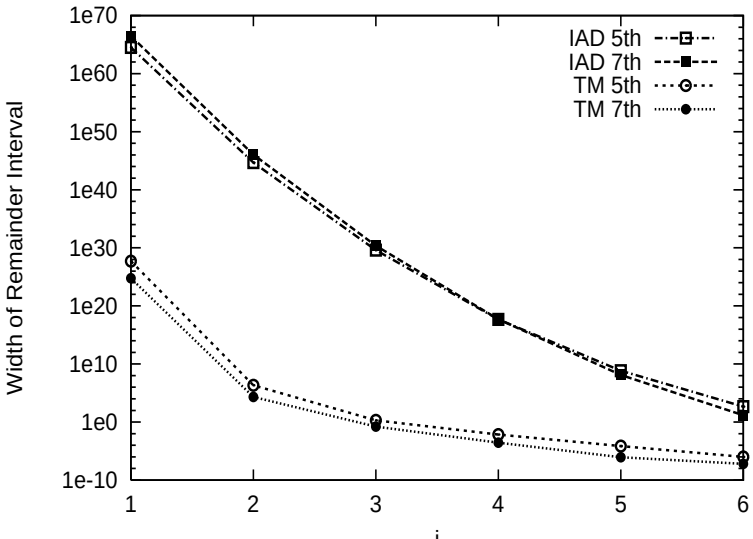
Practical limitation: the code for remainder bound will have **more dependency** than original function. So remainder bounds often have strong overestimation.

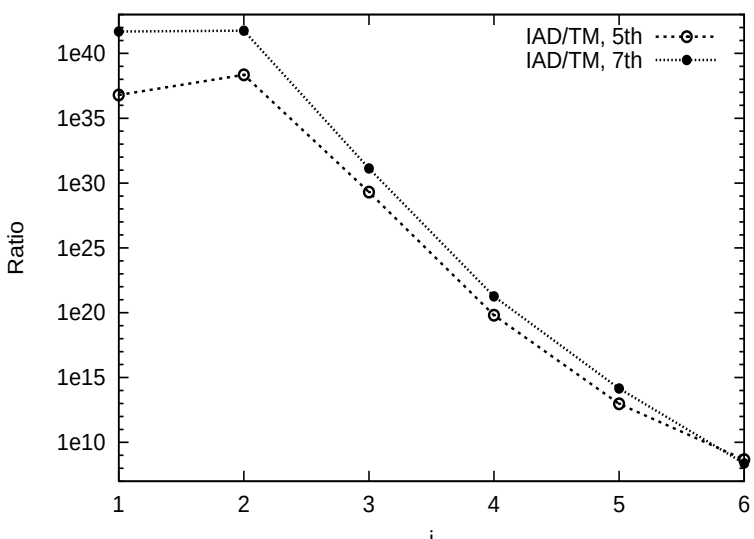
Compare to Taylor Model: Contributions to remainder bound are calculated only from currently accumulated polynomial. This has **less dependency** than original function. Example based on Gritton function:

$$f_7(\vec{x}) = G(2 + \sum_{i=1}^{v} x_i)$$
$$f_8(x) = G(2 + x + \sum_{i=1}^{m} (x - x)).$$


IAD/TM, 2nd
IAD/TM, 4th
IAD/TM, 6th
IAD/TM, 8th
Ratio
Dimension
0
100
200
300
400
500
600
1
2
3
4
5
6
7
8


IAD/TM, 2nd
IAD/TM, 4th
IAD/TM, 6th
IAD/TM, 8th
Ratio
Number of Repetitions of +X-X
1e1
1e2
1e3
1e4
1e5
1e6
1
2
3
4
5
6
7
8
9
10


Width of Remainder Interval
1e70
1e60
1e50
1e40
1e30
1e20
1e10
1e0
1e-10
1
2
3
4
5
6
j
IAD 5th
IAD 7th
TM 5th
TM 7th


Ratio
1e40
1e35
1e30
1e25
1e20
1e15
1e10
1
2
3
4
5
6
j
IAD/TM, 5th
IAD/TM, 7th

The Operator ∂^{-1}on Taylor Models

Let (P_n, I_n) be an n-th order Taylor model of f. From this we can obtain a Taylor model for the indefinite integral $\partial_i^{-1} f = \int f \; dx_i'$ with respect to variable x_i.

Taylor polynomial part: $\int_0^{x_i} P_{n-1} dx_i'$,

Remainder Bound: $(B(P_n - P_{n-1}) + I_n) \cdot B(x_i)$, where $B(P)$ is a polynomial bound.

So define the operator ∂_i^{-1}on space of Taylor models as

$$\begin{aligned} &\partial_i^{-1}(P_n, I_n) \\ &= \left(\int_0^{x_i} P_{n-1} dx_i' \;\; , \;\; (B(P_n - P_{n-1}) + I_n) \cdot B(x_i) \right) \end{aligned}$$

Taylor Models for the Flow

Goal: Determine a Taylor model, consisting of a Taylor Polynomial and an interval bound for the remainder, for the flow of the differential equation

$$\frac{d}{dt}\vec{r}(t) = \vec{F}(\vec{r}(t), t)$$

where $\vec{F}$ is sufficiently differentiable. The Remainder Bound should be fully rigorous for all initial conditions $\vec{r}_0$ and times t that satisfy

$$\vec{r}_0 \in [\vec{r}_{01}, \vec{r}_{02}] = \vec{B}$$
$$t \in [t_0, t_1].$$

In particular, $\vec{r}_0$ itself may be a Taylor model, as long as its range is known to lie in $\vec{B}$.

The Use of Schauder's Theorem

Re-write differential equation as integral equation

$$\vec{r}(t) = \vec{r}_0 + \int_{t_0}^{t} \vec{F}(\vec{r}(t'), t')\, dt'.$$

Now introduce the operator

$$A : \vec{C}^0[t_0, t_1] \to \vec{C}^0[t_0, t_1]$$

on space of continuous functions via

$$A\left(\vec{f}\right)(t) = \vec{r}_0 + \int_{t_0}^{t} \vec{F}\left(\vec{f}(t'), t'\right)\, dt'.$$

Then the solution of ODE is transformed to a fixed-point problem on space of continuous functions

$$\vec{r} = A(\vec{r}).$$

Theorem (Schauder): *Let A be a continuous operator on the Banach Space X. Let $M \subset X$ be compact and convex, and let $A(M) \subset M$. Then A has a fixed point in M, i.e. there is an $\vec{r} \in M$ such that $A(\vec{r}) = \vec{r}$.*

Satisfying Requirements of the Schauder Theorem

Here, $X = \vec{C}^0[t_0, t_1]$, Banach space of continuous functions on $[t_0, t_1]$, equipped with maximum norm. The integral operator A is continuous. The strategy to apply Schauder's Fixed Point Theorem consists of the following steps:

1. Determine family Y of subsets of X, the Schauder Candidate Sets. Each set in Y should be compact and convex, it should be contained in suitable Taylor model, and its image under A should be in Y.
2. Using RDA, determine initial set $M_0 \in Y$ that satisfies $A(M_0) \subset M_0$. Then last requirement of Schauder is satisfied, and M_0 contains solution.
3. Iteratively generate $M_i = A(M_{i-1})$. Each M_i also satisfies $A(M_i) \subset M_{i,}$ and we have $M_1 \supset M_2 \supset ...$ Continue until size stabilizes sufficiently.

Schauder Candidate Sets

As first step, it is necessary to establish a family of sets Y from which to draw candidates for M_0. Let $(\vec{P} + \vec{I})$ be a Taylor model depending on time as well as the initial condition $\vec{r}_0$. Then define the associated set $M_{\vec{P}+\vec{I}}$ as follows:

$$\begin{aligned} M_{\vec{P}+\vec{I}} &\subset \vec{C}^0[t_0, t_1]; \text{ and for } \vec{r} \in M_{\vec{P}+\vec{I}}: \\ \vec{r}(t_0) &= \vec{r}_0 \\ \vec{r}(t) &\in \vec{P} + \vec{I}\ \forall t \in [t_0, t_1]\ \forall \vec{r}_0 \\ |\vec{r}(t') - \vec{r}(t'')| &\leq k|t' - t''|\ \forall t', t'' \in [t_0, t_1]\ \forall \vec{r}_0 \end{aligned}$$

In the last condition, k is a bound for $\vec{F}$, which exists because $\vec{F}$ is continuous and the solutions can cover only finite range over interval $[t_0, t_1]$. The last condition means that all $\vec{r} \in M_{\vec{P}+\vec{I}}$ are uniformly Lipschitz with constant k. Define the candidate set Y as

$$Y = \bigcup_{\vec{P}+\vec{I}} M_{\vec{P}+\vec{I}}$$

Convexity, Compactness, Invariance of Candidate Sets

Let $M \in Y$. Then M is convex, because

$$\vec{x}_1, \vec{x}_2 \in M \Rightarrow$$
$$\alpha\vec{x}_1 + (1-\alpha)\vec{x}_2 \in M \ \forall\alpha \in [0,1]$$

Furthermore, M is compact, i.e. any sequence in M has a clusterpoint in M. To see this, let $(\vec{x}_n)$ be a sequence of functions in M.Then by definition of M, $(\vec{x}_n)$ is uniformly Lipschitz, and thus uniformly equicontinuous. $(\vec{x}_n)$ is also uniformly bounded, and hence according to the Ascoli-Arzela Theorem, has a uniformly convergent subsequence. Since the $\vec{x}_n$ are continuous, so is the limit $\vec{x}^*$ of this subsequence, and since M is closed, the limit $\vec{x}^*$ is in M.

Finally, A maps Y into itself, and the uniform Lipschitzness follows because $\vec{F}$ is bounded by k.

Satisfying Inclusion with Taylor Models

The only remaining requirements for Schauder's theorem is to find a Taylor model $\vec{P} + \vec{I}$ such that

$$A(\vec{P} + \vec{I}) \subset \vec{P} + \vec{I}.$$

But this condition can be checked with Taylor Models.

To succeed with inclusion requirement depends on finding suitable choice for $\vec{P}$ and $\vec{I}$. Furthermore, it is desirable that $\vec{I}$ be tight.

Both benefit from the choice of a polynomial $\vec{P}$ that is already "close" to the true solution of the ODE.

The Polynomial of the Self-Including Set

Attempt sets M^* of the form

$$M^* = M_{\vec{P}^*+\vec{I}^*} \text{ where}$$
$$\vec{P}^* = \mathcal{M}_n(\vec{r}_0, t),$$

the n-th order Taylor expansion of the flow of the ODE. It is to be expected that $\vec{I}^*$ can be chosen smaller and smaller as order n of $\vec{P}^*$ increases.

This requires knowledge of nth order flow $\mathcal{M}_n(\vec{r}_0, t)$, including time dependence. It can be obtained by iterating in polynomial arithmetic, or Taylor models without treatment of a remainder. To this end, one chooses an initial function $\mathcal{M}_n^{(0)}(\vec{r}, t) = \mathcal{I}$, where $\mathcal{I}$ is the identity function, and then iteratively determines

$$\mathcal{M}_n^{(k+1)} =_n A(\mathcal{M}_n^{(k)}).$$

This process converges to the exact result $\mathcal{M}_n$ in exactly n steps.

The Remainder of the Self-Including Set

Now try to find $\vec{I}^*$ such that

$$A(\mathcal{M}_n + \vec{I}^*) \subset \mathcal{M}_n + \vec{I}^*,$$

the Schauder inclusion requirement. Suitable choice for $\vec{I}^*$ requires experimenting, but is greatly simplified by the observation

$$\vec{I}^* \supset \vec{I}^{(0)} = A(\mathcal{M}_n(\vec{r}, t) + [\vec{0}, \vec{0}]) - \mathcal{M}_n(\vec{r}, t).$$

Evaluating the right hand side in RDA yields a lower bound for $\vec{I}^*$, and a benchmark for the size to be expected. Now iteratively try

$$\vec{I}^{(k)} = 2^k \cdot \vec{I}^{(0)},$$

until computational inclusion is found, i.e.

$$A(\mathcal{M}_n(\vec{r}, t) + \vec{I}^{(k)}) \subset \mathcal{M}_n(\vec{r}, t) + \vec{I}^{(k)}.$$

Iterative Refinement of the Self-Including Set

Once computational inclusion has been determined, solution of ODE is known to be contained in the Taylor model $\mathcal{M}_n(\vec{r},t) + \vec{I}^{(k)}$. Set $\vec{I}_{(1)} = \vec{I}^{(k)}$; since the solution is a fixed point of A, it is even contained in

$$A^k(\mathcal{M}_n(\vec{r},t) + \vec{I}_{(1)}) \text{ for all } k.$$

Furthermore, the iterates of A are shrinking in size, i.e.

$$A^k(\mathcal{M}_n(\vec{r},t) + \vec{I}_{(1)}) \subset A^{k-1}(\mathcal{M}_n(\vec{r},t) + \vec{I}_{(1)}) \; \forall k$$

So the width of the remainder bound of the flow can be decreased by iteratively determining

$$\mathcal{M}_n(\vec{r},t) + \vec{I}_{(k)} = A(\mathcal{M}_n(\vec{r},t) + \vec{I}_{(k-1)}),$$

until no further significant decrease in size is achieved. As a result,

$$\mathcal{M}_n(\vec{r},t) + \vec{I}_{(k)}$$

is the desired sharp inclusion of the flow of the original ODE.

The Volterra Equation

Describe dynamics of two conflicting populations

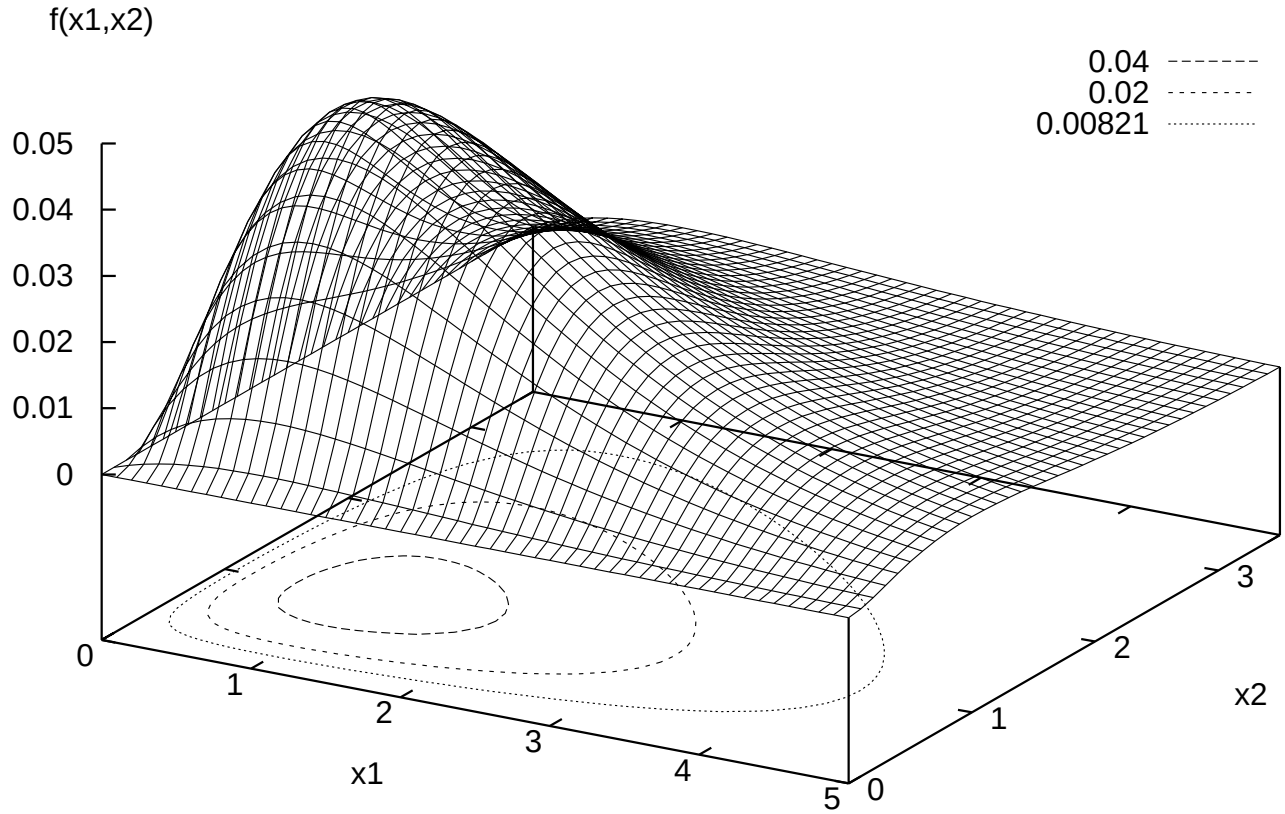
$$\frac{dx_1}{dt} = 2x_1(1 - x_2), \quad \frac{dx_2}{dt} = -x_2(1 - x_1)$$

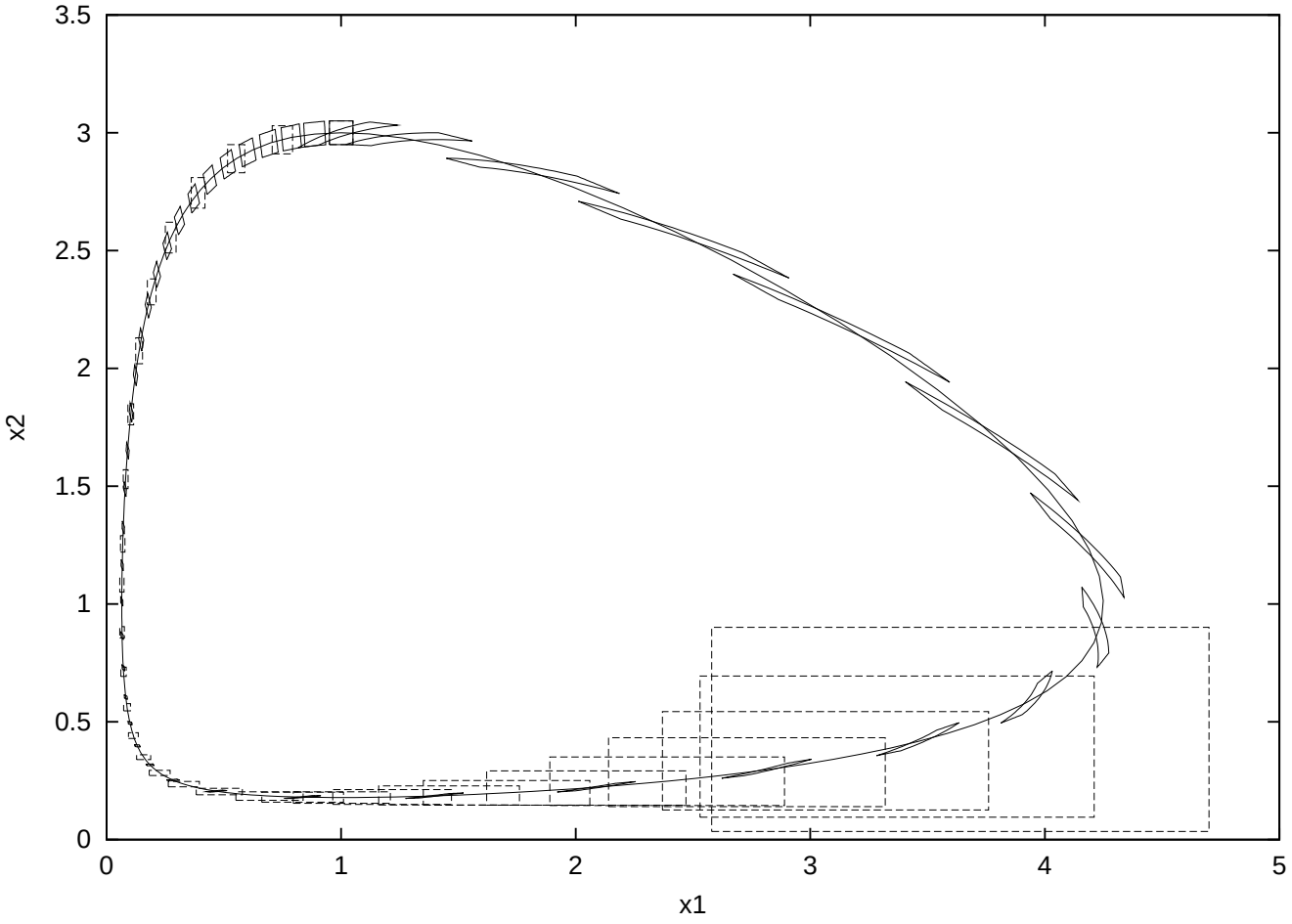
Interested in initial condition

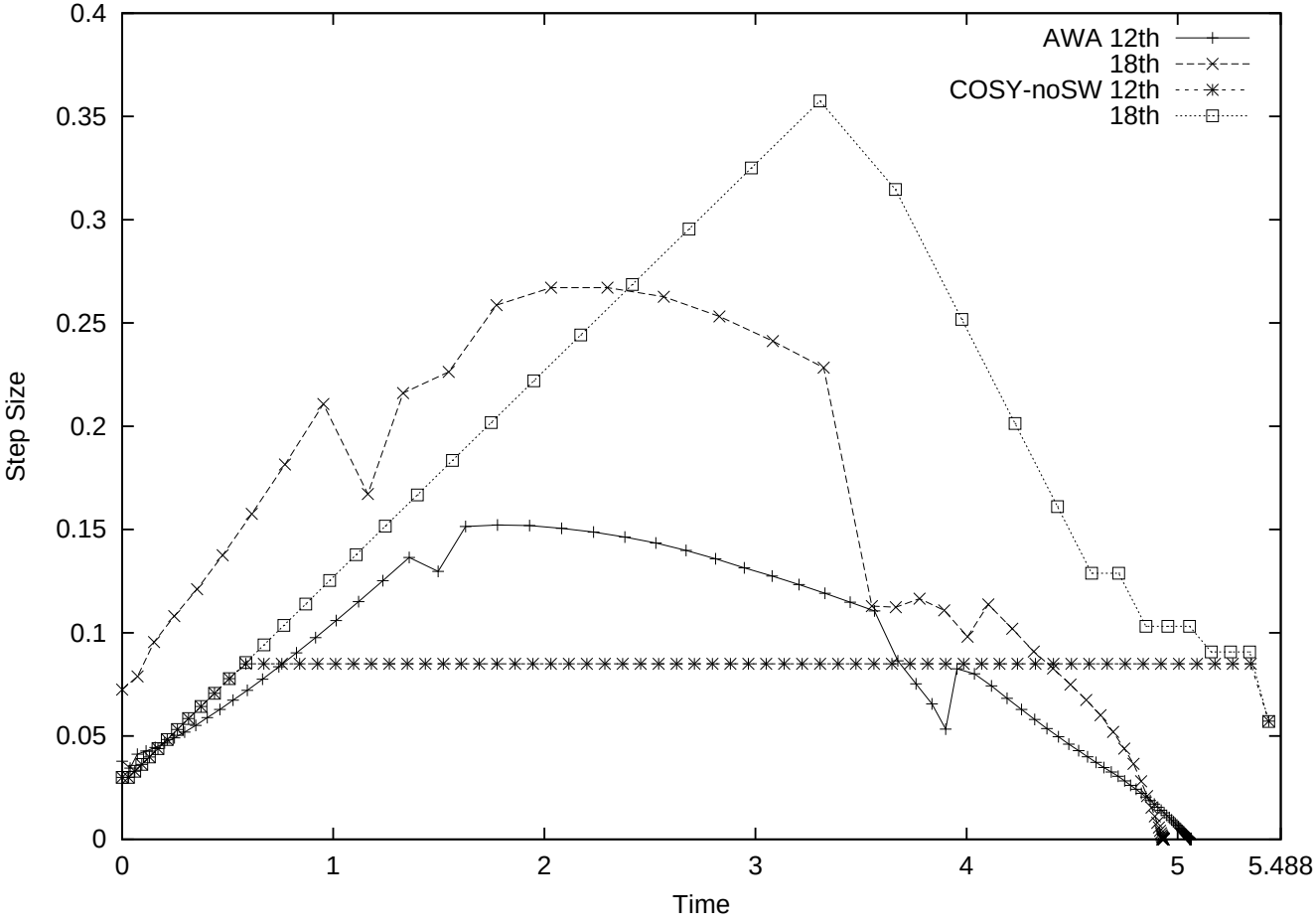
$$x_{01} \in 1 + [-0.05, 0.05], \quad x_{02} \in 3 + [-0.05, 0.05] \quad \text{at } t = 0.$$

Satisfies constraint condition

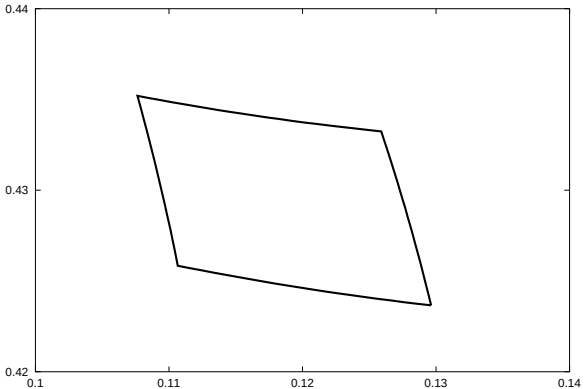
$$C(x_1, x_2) = x_1 x_2^2 e^{-x_1 - 2x_2} = \text{Constant}$$


f(x1,x2)
0.04
0.02
0.00821
0.05
0.04
0.03
0.02
0.01
0
0
1
2
3
4
5
x1
0
1
2
3
x2

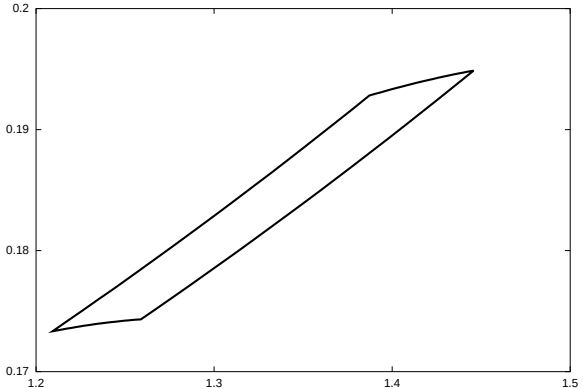

3.5
3
2.5
2
1.5
1
0.5
0
x2
0
1
2
3
4
5
x1


AWA 12th
18th
COSY-noSW 12th
18th
Step Size
Time
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0
1
2
3
4
5
5.488

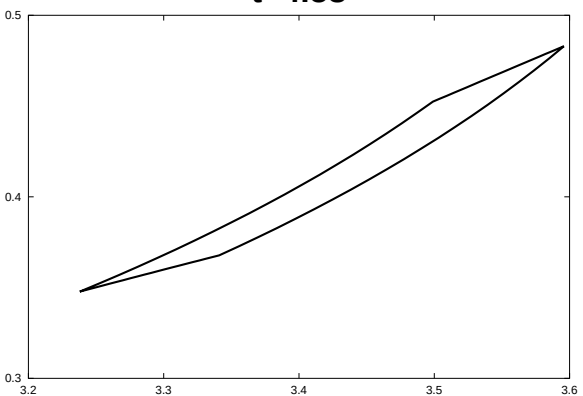
t=2.33



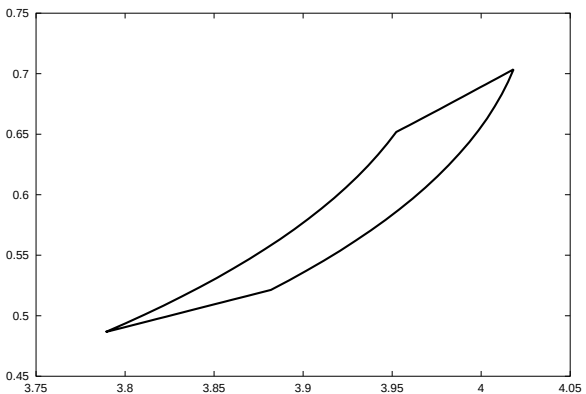
t=3.94



t=4.58

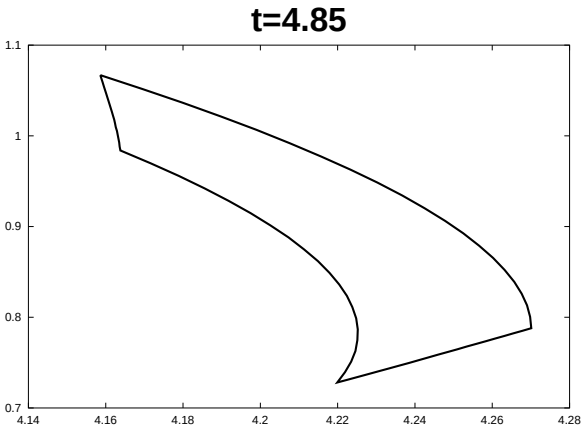


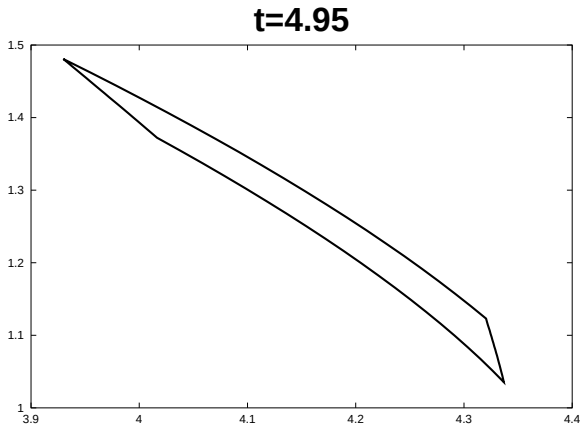
t=4.71

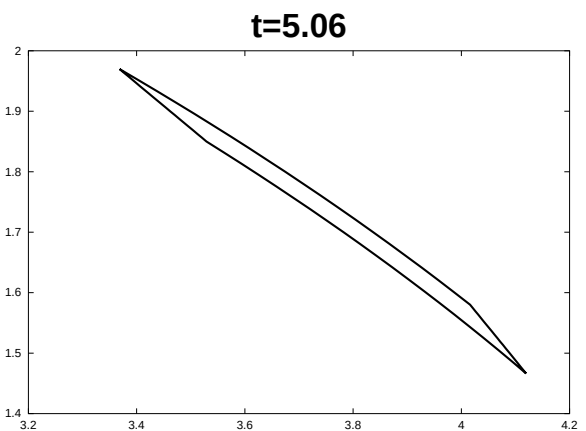


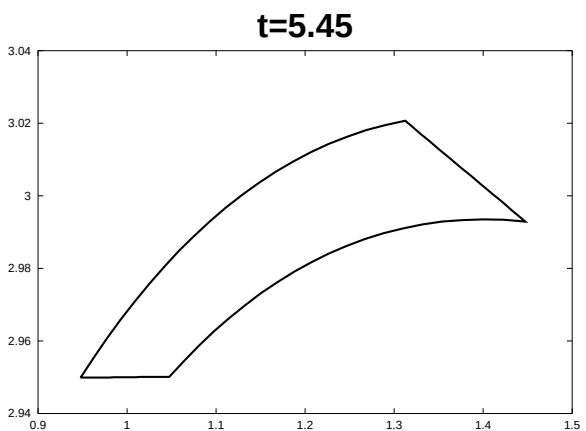
t=4.85

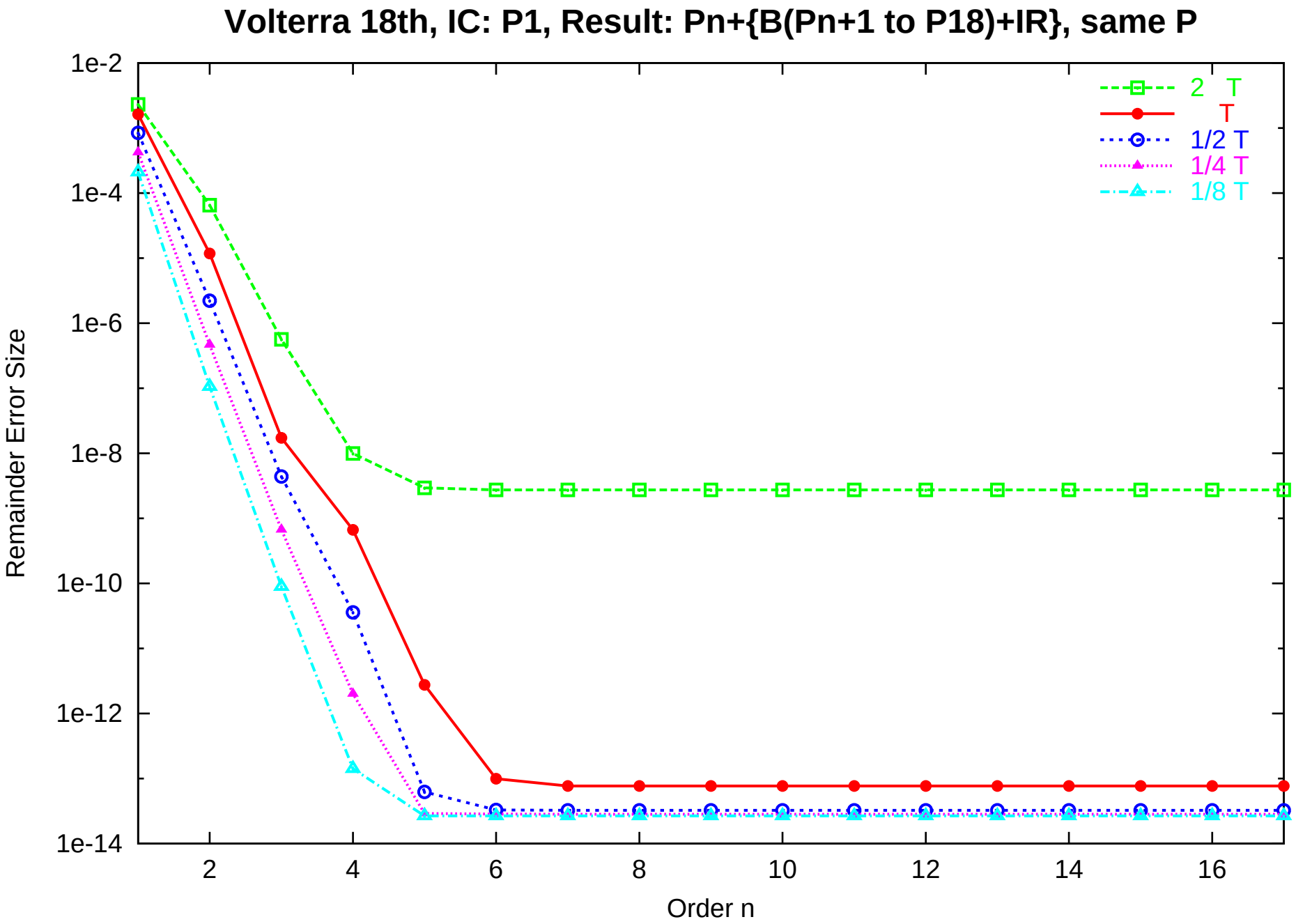
t=4.95

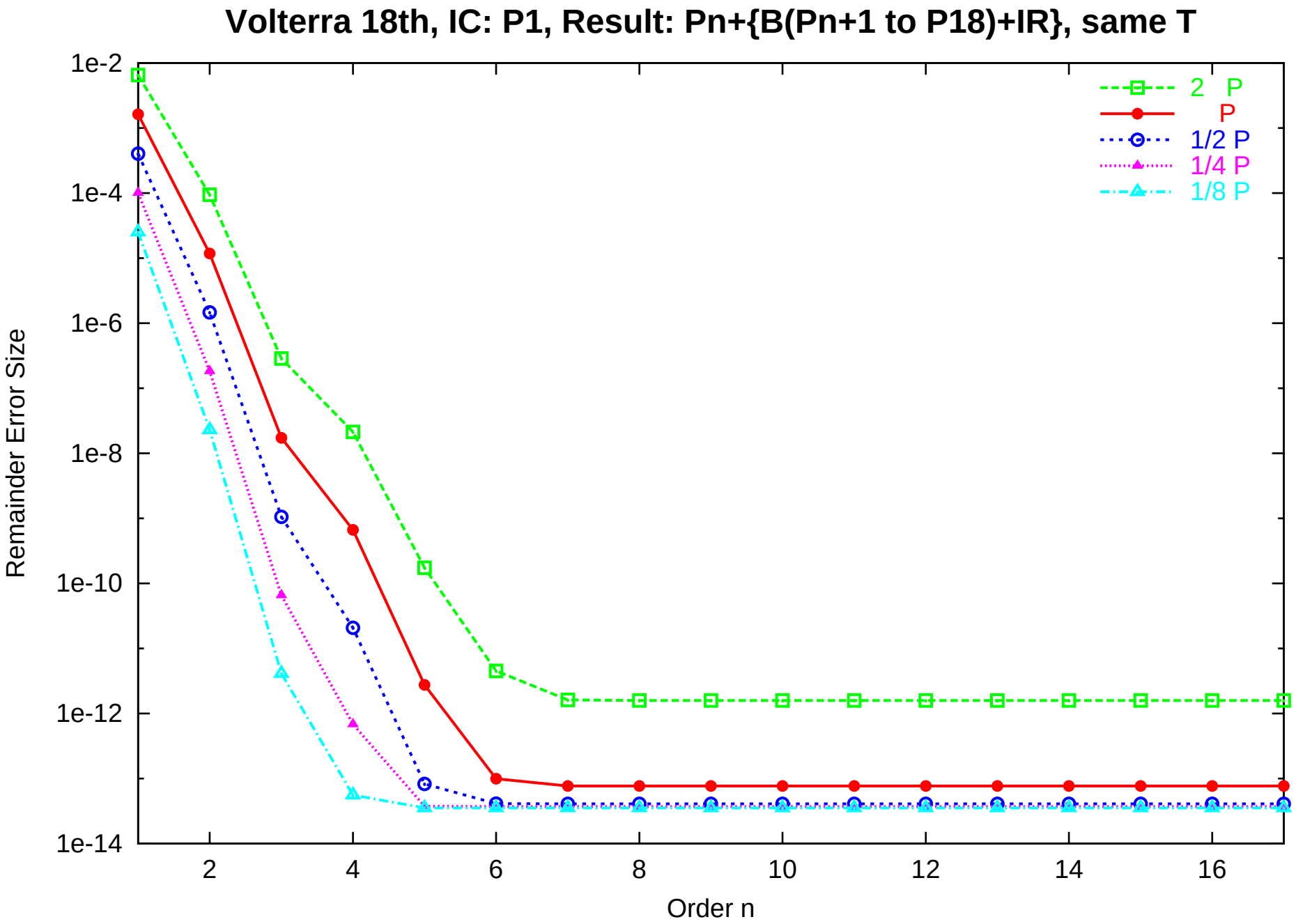

t=4.85
1.1
1
0.9
0.8
0.7
4.14
4.16
4.18
4.2
4.22
4.24
4.26
4.28


t=4.95
1.5
1.4
1.3
1.2
1.1
1
3.9
4
4.1
4.2
4.3
4.4


t=5.06
2
1.9
1.8
1.7
1.6
1.5
1.4
3.2
3.4
3.6
3.8
4
4.2


t=5.45
3.04
3.02
3
2.98
2.96
2.94
0.9
1
1.1
1.2
1.3
1.4
1.5


Volterra 18th, IC: P1, Result: Pn+{B(Pn+1 to P18)+IR}, same P
Remainder Error Size
1e-2
1e-4
1e-6
1e-8
1e-10
1e-12
1e-14
2
4
6
8
10
12
14
16
Order n
2 T
T
1/2 T
1/4 T
1/8 T


Volterra 18th, IC: P1, Result: Pn+{B(Pn+1 to P18)+IR}, same T
Remainder Error Size
1e-2
1e-4
1e-6
1e-8
1e-10
1e-12
1e-14
2
4
6
8
10
12
14
16
Order n
2 P
P
1/2 P
1/4 P
1/8 P

Shrink Wrapping I

A method to remove the remainder bound of a Taylor model by increasing the polynomial part.

After the kth step of the integration, the region occupied by the final variables is given by

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

where $\vec{x}_0$ are the initial variables, $\vec{B}$ is the original box of initial conditions, $\mathcal{M}_0$ is the polynomial part of the Taylor model, and $\vec{I_0}$ is the remainder bound interval. $\mathcal{M}_0$ is scaled such that the original box $\vec{B}$ is unity, i.e. $\vec{B} = [-1, 1]^v$. $\vec{I_0}$ 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. Try to "absorb" the small remainder interval into a set very similar to the first part via

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

where $\mathcal{M}_0^*$ is a slightly modified polynomial, and $\vec{I}_0^*$ is significantly reduced

Shrink Wrapping II

First, extract the constant part $\vec{a}_0$ and linear part $\hat{M}_0 \cdot \vec{x}$ of $\mathcal{M}_0$ and determine a floating point approximation $\bar{M}_0^{-1}$ of $\hat{M}_0$. If ODEs admits unique solutions, attempting to invert the linear transformation $\hat{M}_0$ in a floating point environement will very likely succeed.

After approximate inverse $\bar{M}_0^{-1}$ has been determined, apply linear transformation $\bar{M}_0^{-1} \cdot (\vec{x} - \vec{a}_0)$ from the left to the Taylor model $\mathcal{M}_0(\vec{x}_0) + \vec{I}_0$ 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

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

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 $\vec{I}$ into the interval box $d \cdot [-1, 1]^v$, where d is a small number.

Shrink Wrapping III

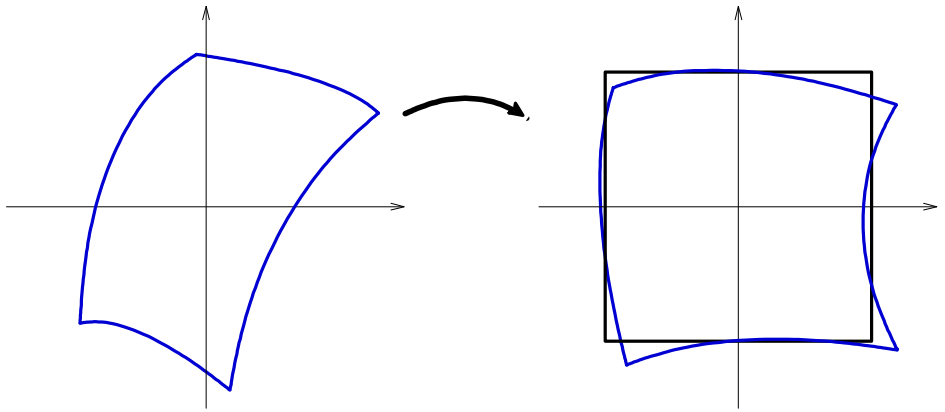


Figure 1: The region described by the Taylor model $\mathcal{M}_0 + \vec{I}_0$ is transformed to be normalized as $\mathcal{I} + \mathcal{S} + \vec{I}$, where $\mathcal{I}$ is the identity.

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

$$s \geq |\mathcal{S}_i(\vec{x})| \ \forall \ \vec{x} \in B, \ 1 \leq i \leq v,$$

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

We call a map $\mathcal{M}$ shrinkable if $(1 - vt) > 0$ and $(1 - s) > 0$;

Shrink Wrapping IV

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 (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 $q \approx 1 + 10^{-12}$.

Shrink Wrapping V

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

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

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

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

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

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

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

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 (1) dominates the right term for any set of c_j, and we thus have

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

which completes the proof.

Shrink Wrapping VI

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

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

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

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

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 $R = \vec{I} + \bigcup_{\vec{x} \in \vec{B}} \mathcal{M}(\vec{x})$ can extend beyond the surface in direction i at the surface point $\vec{y} = \mathcal{M}(x_1, ..., x_{i-1}, 1, x_{i+1}, ..., x_v)$.

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

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

We shall now find a bound for $r_i(\vec{y})$. 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 $\vec{\bar{y}}$ in the definition of $r_i(\vec{y})$ entails that in the $(v-1)$ dimensional subspace, $\left\| \vec{\bar{y}}^{(i)} - \vec{y}^{(i)} \right\| \leq d$. Let $\vec{\bar{x}}^{(i)}$ and $\vec{x}^{(i)}$ be the $(v-1)$ di-

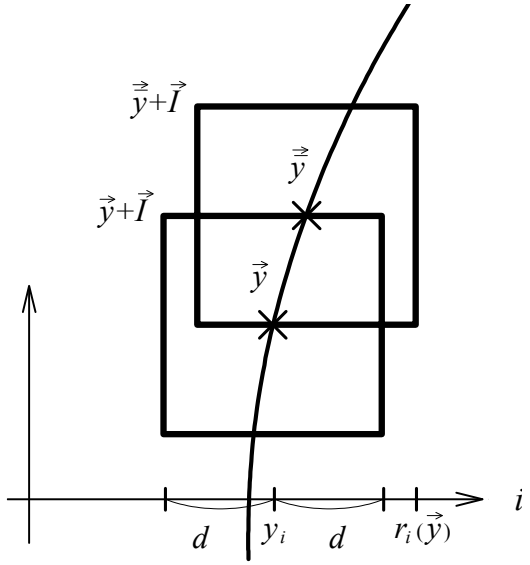


Figure 2: At the point y_i, the set $R = \vec{I} + \bigcup_{\vec{x}\in\vec{B}} \mathcal{M}(\vec{x})$ can extend to $r_i(\vec{y})$.

mensional pre-images of $\bar{\vec{y}}^{(i)}$ and $\vec{y}^{(i)}$, respectively; because $\left\|\bar{\vec{y}}^{(i)} - \vec{y}^{(i)}\right\| \leq d$, we have according to the above lemma that

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

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 $r_i(\vec{y})$ via

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

We now invoke the first statement of the lemma for the case of $\bar{\vec{x}}$, $\vec{x}$ 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 $|\mathcal{M}_i(\bar{\vec{x}}) - \mathcal{M}_i(\vec{x})| \leq (v-1)t \cdot d/(1-(v-1)t)$, and altogether

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

We observe that the second term in the last expression is independent of i. Hence we have shown that the "band" around $\bigcup_{\vec{x} \in \vec{B}} \mathcal{M}(\vec{x})$ generated by

the addition of $\vec{I}$ 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

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

as advertised.

Shrink Wrapping VII

Let us consider the practical limiations of the method; apparently the measures of the nonlinearities s and t must not become too large

Remark (Limitations of shrink wrapping) Apparently the shrink wrap method discussed above has the following limitations

Remark 1 *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 happen 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 (Blunting of an Ill-Conditioned Matrix)

Let $\hat{A}$ be a regular matrix that is potentially ill-conditioned and $\vec{q} = (q_1, ...q_n)$ a vector with $q_i > 0$. Arrange the column vectors $\vec{a}_i$ of $\hat{A}$ by size.

Let $\vec{e}_i$ be the familiar orthonormal vectors obtained through the Gram-Schmidt procedure, i.e.

$$\vec{e}_i = \frac{\vec{a}_i - \sum_{k=1}^{i-1} \vec{e}_k \ (\vec{a}_i \cdot \vec{e}_k)}{\left| \vec{a}_i - \sum_{k=1}^{i-1} \vec{e}_k \ (\vec{a}_i \cdot \vec{e}_k) \right|}.$$

We form vectors $\vec{b}_i$ via

$$\vec{b}_i = \vec{a}_i + q_i \vec{e}_i$$

and assemble them columnwise into the matrix $\hat{B}$. We call $\hat{B}$ the $\vec{q}$-blunted matrix belonging to $\hat{A}$

Proposition (Regularity of the Blunted Matrix) The $\vec{b}_i$ are linearly independent and thus $\hat{B}$ is regular.

Proof. By induction. Apparently $\vec{b}_1$ is linearly independent. Assume now that $\vec{b}_1, ..., \vec{b}_{i-1}$ are linearly independent. We first observe that for each i, the vector $\vec{b}_i$ is a linear combination of the $\vec{a}_k$ for $k = 1, ..., i$ and thus also of the $\vec{e}_k$ for $k = 1, ..., i$, since both the $\vec{a}_k$ and the $\vec{e}_k$ span the same space. Now assume $\vec{b}_i$ is linearly dependent on $\vec{b}_1, ..., \vec{b}_{i-1}$; then it is also linearly

dependent on $\vec{e}_1, ..., \vec{e}_{i-1}$, i.e. there are $\lambda_1, ..., \lambda_{i-1}$ such that

$$\vec{b}_i = \sum_{k=1}^{i-1} \lambda_k \vec{e}_k.$$

But because $\vec{b}_i = \vec{a}_i + q_i \vec{e}_i$, we have

$$\vec{a}_i \left(1 + \frac{q_i}{\left| \vec{a}_i - \sum_{k=1}^{i-1} \vec{e}_k \left(\vec{a}_i \cdot \vec{e}_k \right) \right|} \right) = \sum_{k=1}^{i-1} \left(\lambda_k + \vec{a}_i \cdot \vec{e}_k \right) \ \vec{e}_k$$

Since by requirement, $q_i > 0$, the factor of $\vec{a}_i$ is nonzero, and we have a contradiction to the linear independence of $\vec{a}_i$ from $\vec{e}_1, ..., \vec{e}_{i-1}$. Thus $\vec{b}_1, ..., \vec{b}_i$ are linearly independent.

Intuitively, of course, the effect of blunting is that each vector $\vec{b}_i$ is being "pulled away" from the space spanned by the previous vectors $\vec{b}_1, ..., \vec{b}_{i-1}$, and more strongly so if q_i becomes bigger and bigger. In fact, we have the following result: .

Long-Term Behavior - Floating Point Case

Consider very simple two-state dynamical system:

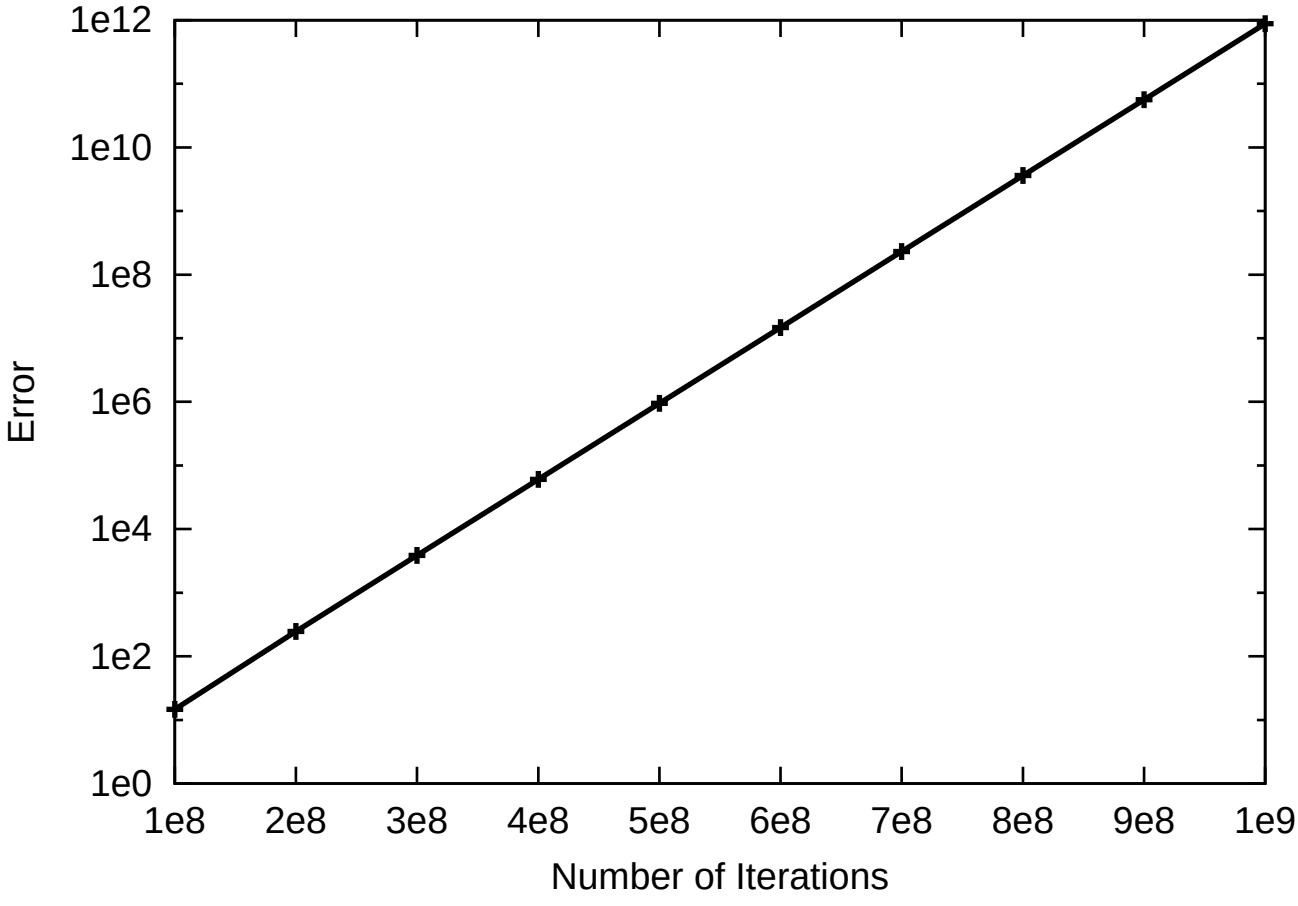
$$x_{n+1} = a \cdot x_n$$
$$x_{n+2} = (1/a) \cdot x_{n+1}$$

with initial condition $x_0 = 1$. Study the behavior for specific choices of a in both single and double precision arithmetic on

- F77 compiler by DEC, now distributed as f77 Digital Visual Fortran Version 5.0 as part of Microsoft Fortran PowerStation
- G77 compiler distributed by GNU; we specifically tested Version V0.5.24.

Choose $a_1 = 3$ for single precision, $a_2 = 0.9999999901608054$ for double precision

In both cases, we observe exponential growth of the error!


1e12
1e10
1e8
1e6
1e4
1e2
1e0
Error
1e8
2e8
3e8
4e8
5e8
6e8
7e8
8e8
9e8
1e9
Number of Iterations

Long-Term Behavior - Validated Case

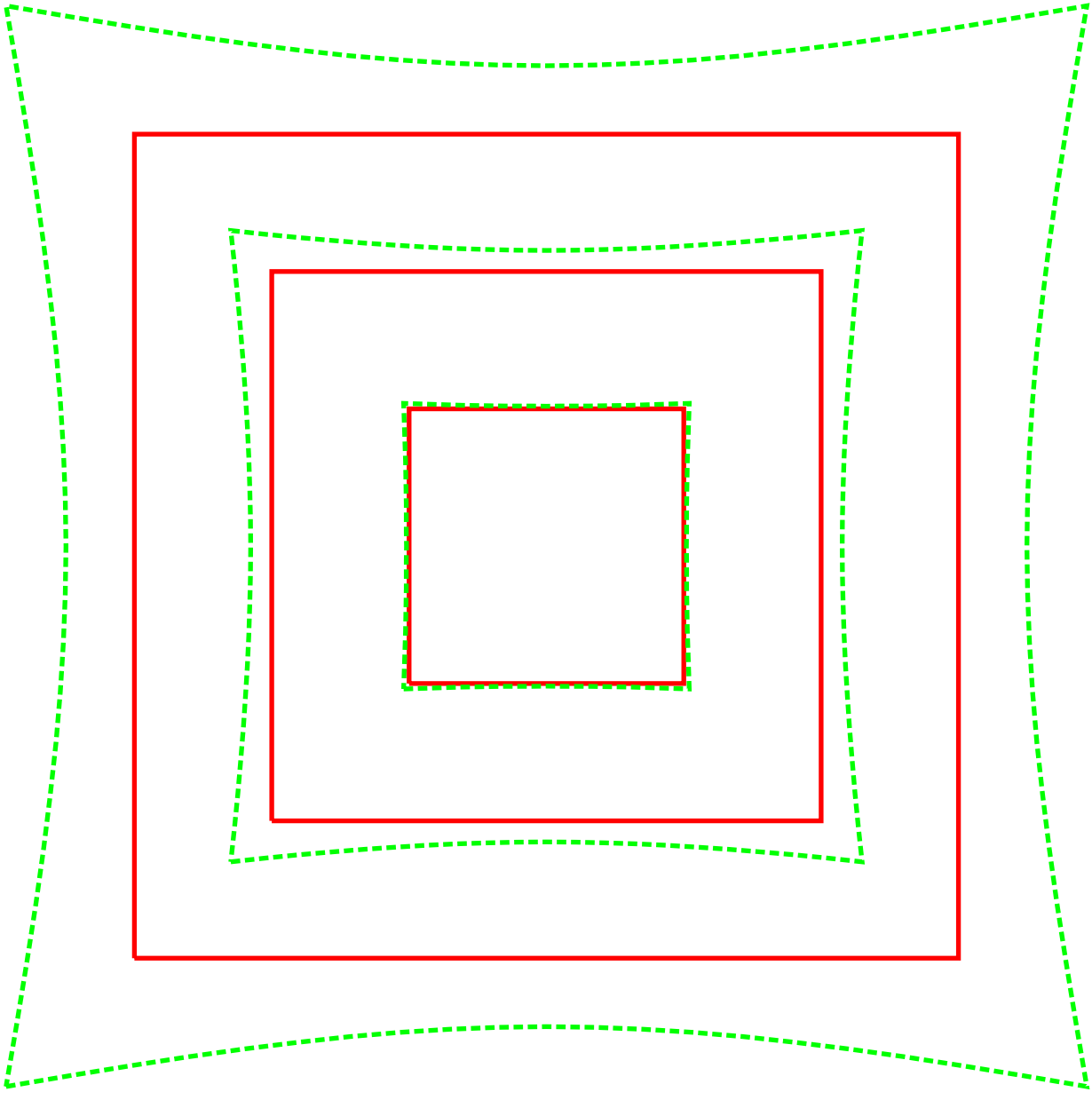
Consider very simple two-state dynamical system:

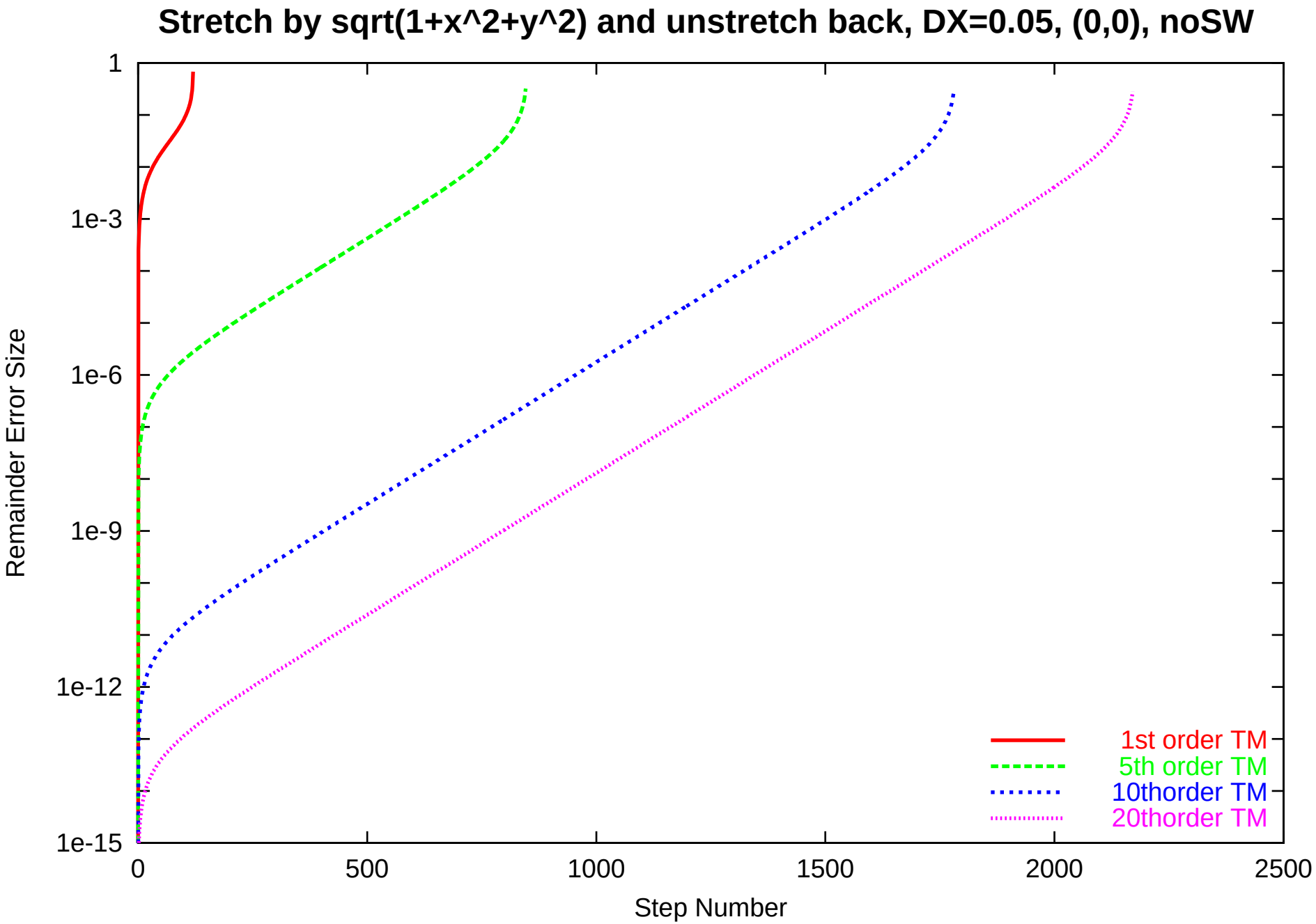
$$x_{n+1} = x_n \cdot \sqrt{1 + x_n^2 + y_n^2} \text{ and } y_{n+1} = y_n \cdot \sqrt{1 + x_n^2 + y_n^2}$$

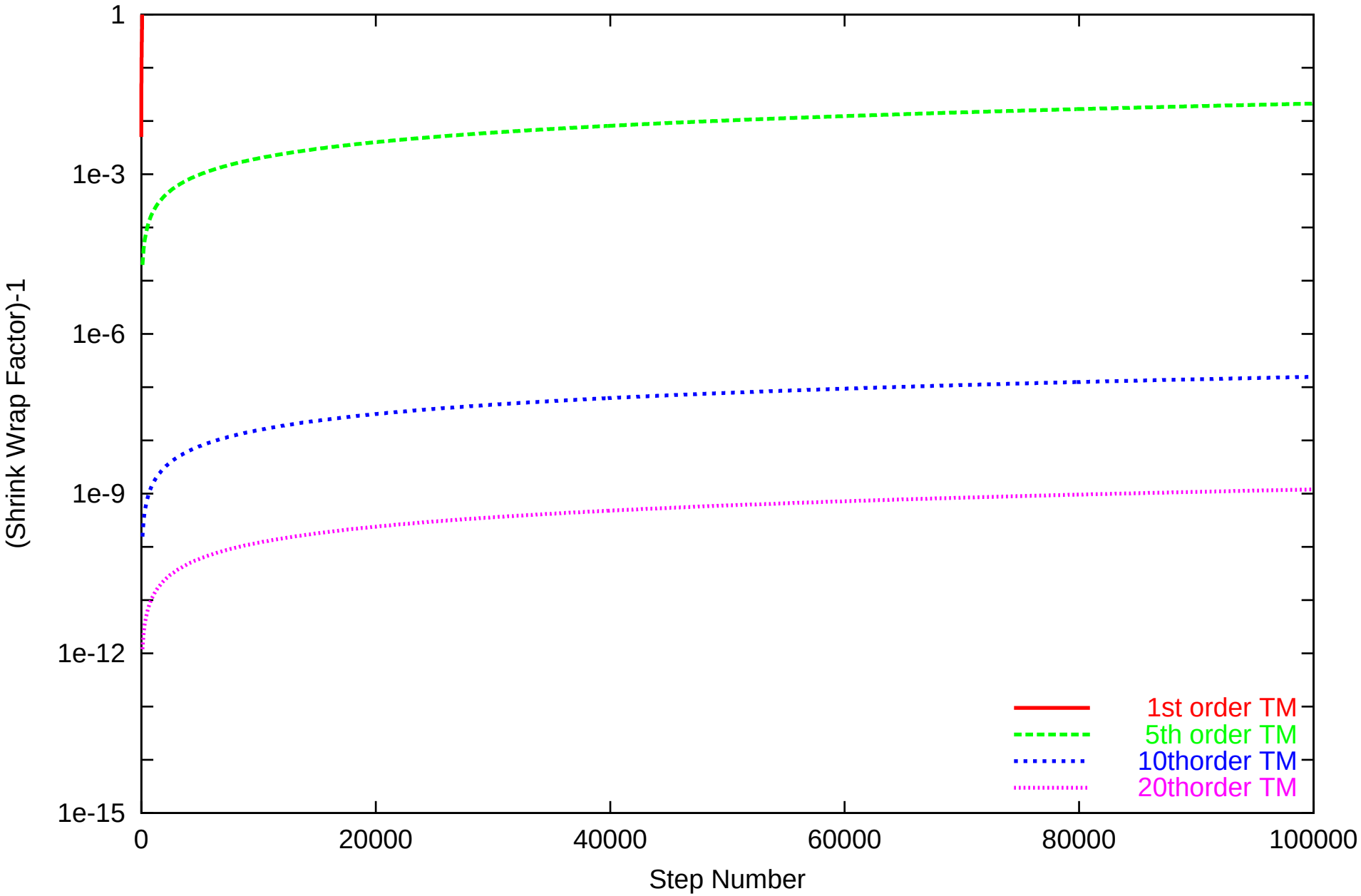
$$x_{n+2} = x_{n+1} \cdot \sqrt{\frac{2}{1 + \sqrt{1 + 4(x_{n+1}^2 + y_{n+1}^2)}}} \text{ and}$$

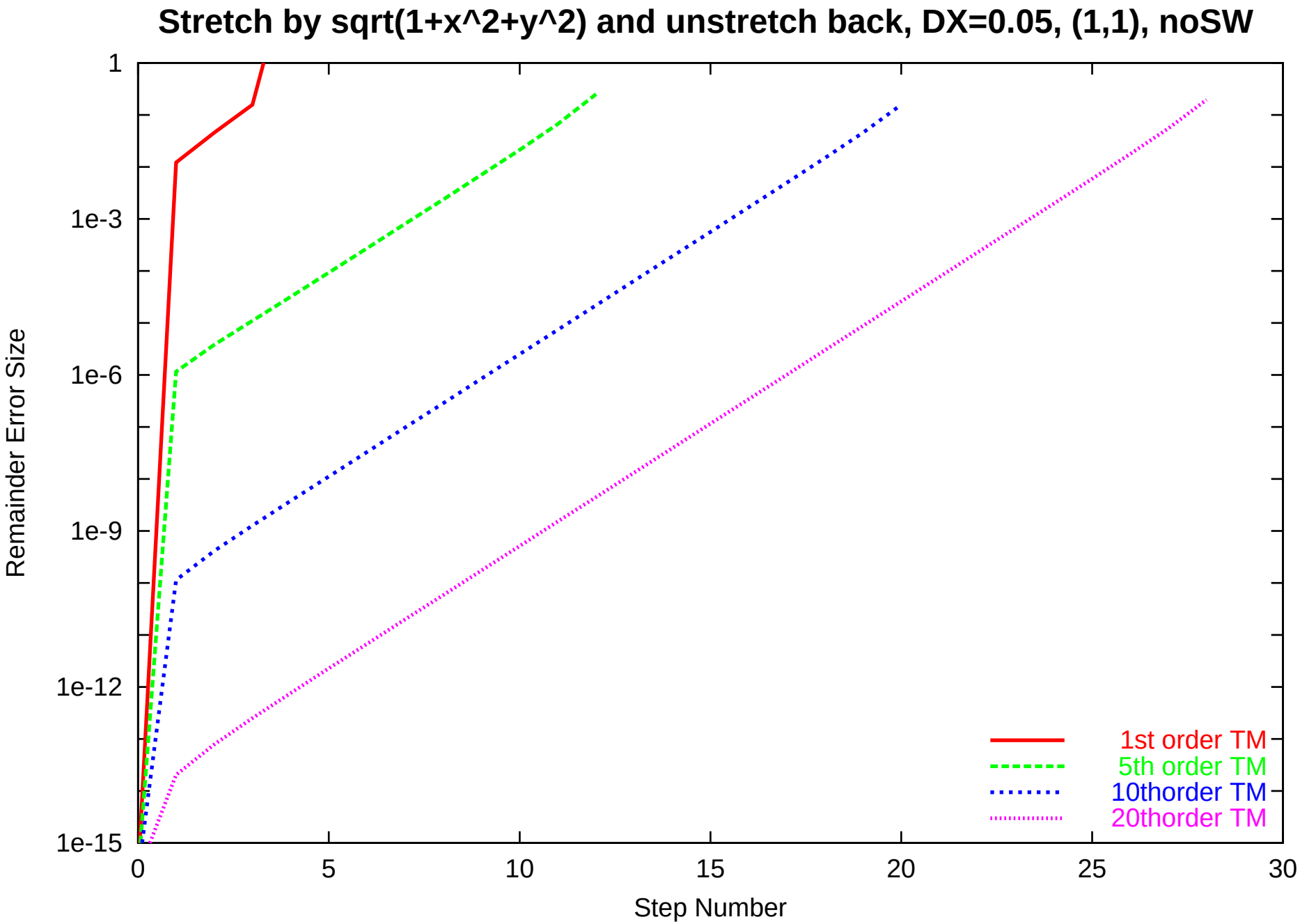
$$y_{n+2} = y_{n+1} \cdot \sqrt{\frac{2}{1 + \sqrt{1 + 4(x_{n+1}^2 + y_{n+1}^2)}}}.$$

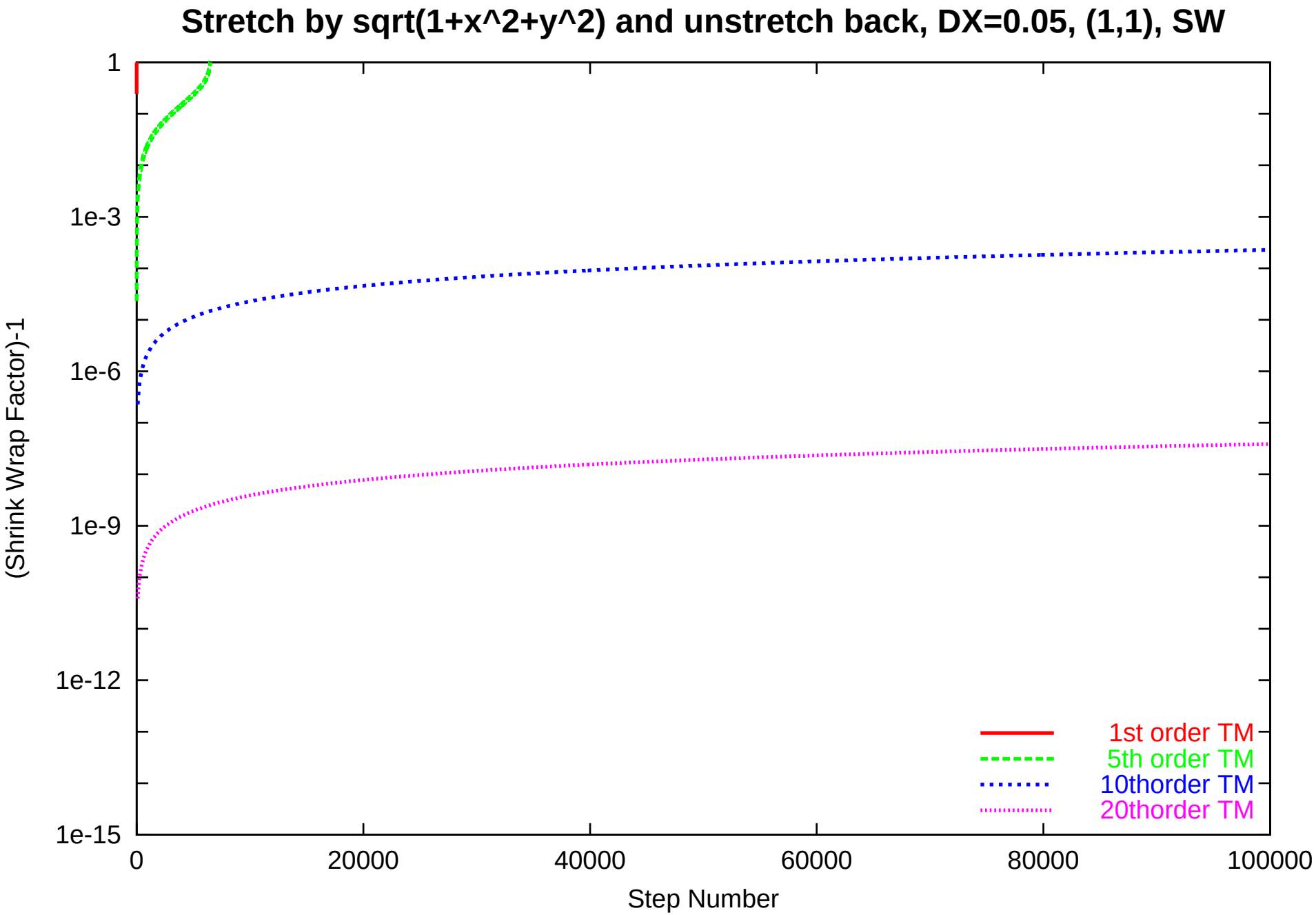
Simple arithmetic shows that, also here we have $(x_{n+2}, y_{n+2}) = (x_n, y_n)$.




Stretch by sqrt(1+x^2+y^2) and unstretch back, DX=0.05, (0,0), noSW
Remainder Error Size
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
500
1000
1500
2000
2500
Step Number
1st order TM
5th order TM
10thorder TM
20thorder TM


Stretch by sqrt(1+x^2+y^2) and unstretch back, DX=0.05, (0,0), SW
(Shrink Wrap Factor)-1
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
20000
40000
60000
80000
100000
Step Number
1st order TM
5th order TM
10thorder TM
20thorder TM


Stretch by sqrt(1+x^2+y^2) and unstretch back, DX=0.05, (1,1), noSW
Remainder Error Size
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
5
10
15
20
25
30
Step Number
1st order TM
5th order TM
10thorder TM
20thorder TM


Stretch by sqrt(1+x^2+y^2) and unstretch back, DX=0.05, (1,1), SW
(Shrink Wrap Factor)-1
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
20000
40000
60000
80000
100000
Step Number
1st order TM
5th order TM
10thorder TM
20thorder TM

Preconditioning the Flow

Idea: write the Taylor model of the solution as a composition of two Taylor models $(P_l + I_l)$ and $(P_r + I_r)$, and then choose $P_l + I_l$ in such a way that in each step, the operations appearing on I_r are minimized. At the same time, I_l will be chosen as small as possible. Can be viewed as a coordinate transformation.

In the factorization, we impose that $(P_l + I_l)$ is normalized such that each of its components has a range in $[-1, 1]$, and even near the boundaries.

Definition (Preconditioning the Flow) Let $(P + I)$ be a Taylor model. We say that $(P_l + I_l), (P_r + I_r)$ is a factorization of $(P + I)$ if $B(P_r + I_r) \in [-1, 1]$ and

$$(P + I) \in (P_l + I_l) \circ (P_r + I_r) \text{ for all } x \in B$$

where B is the domain of the Taylor model $(P_r + I_r)$.

The composition of the Taylor models is here to be understood as insertion of the Taylor model $P_r + I_r$ into the polynomial P_l via Taylor model addition and multiplication and subsequent addition of the remainder bound I_l. For the study of the solutions of ODEs, the following result is important

Preconditioning the Flow II

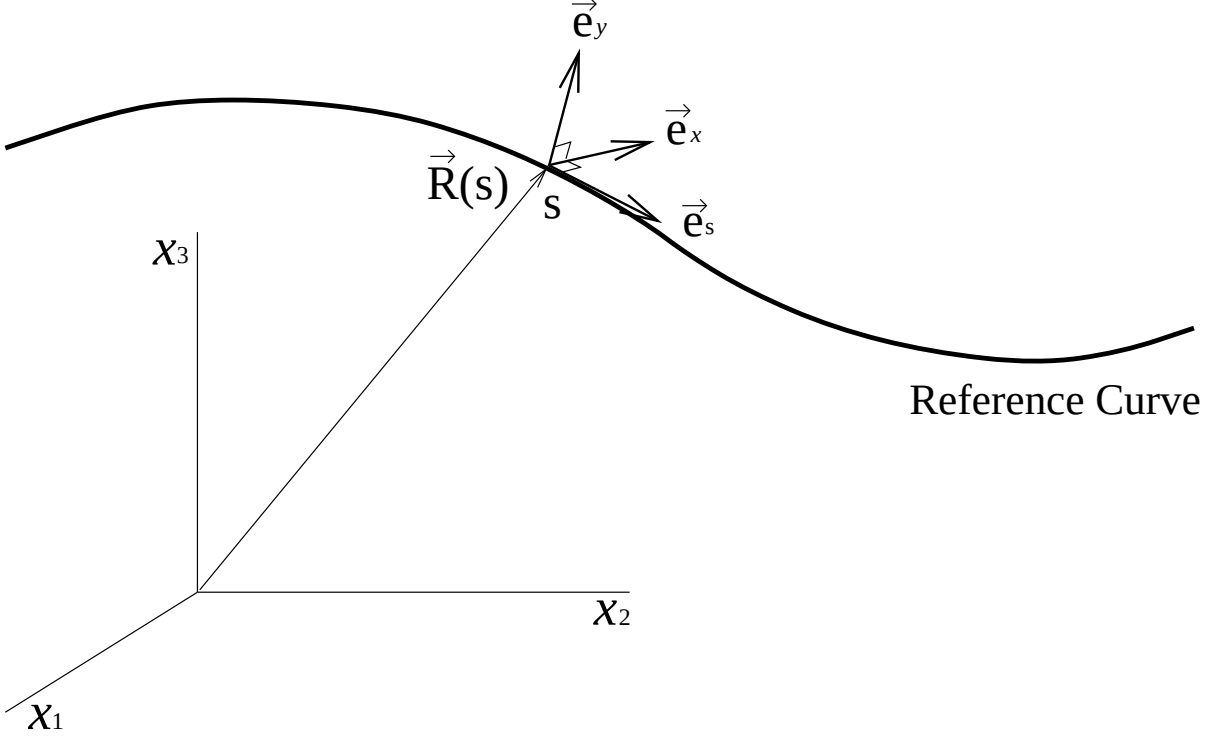
Proposition Let $(P_{l,n} + I_{l,n}) \circ (P_{r.n} + I_{r,n})$ be a factored Taylor model that encloses the flow of the ODE at time t_n. Let $(P^*_{l,n+1}, I^*_{l,n+1})$ be the result of integrating $(P_{l,n} + I_{l,n})$ from t_n to t_{n+1}. Then

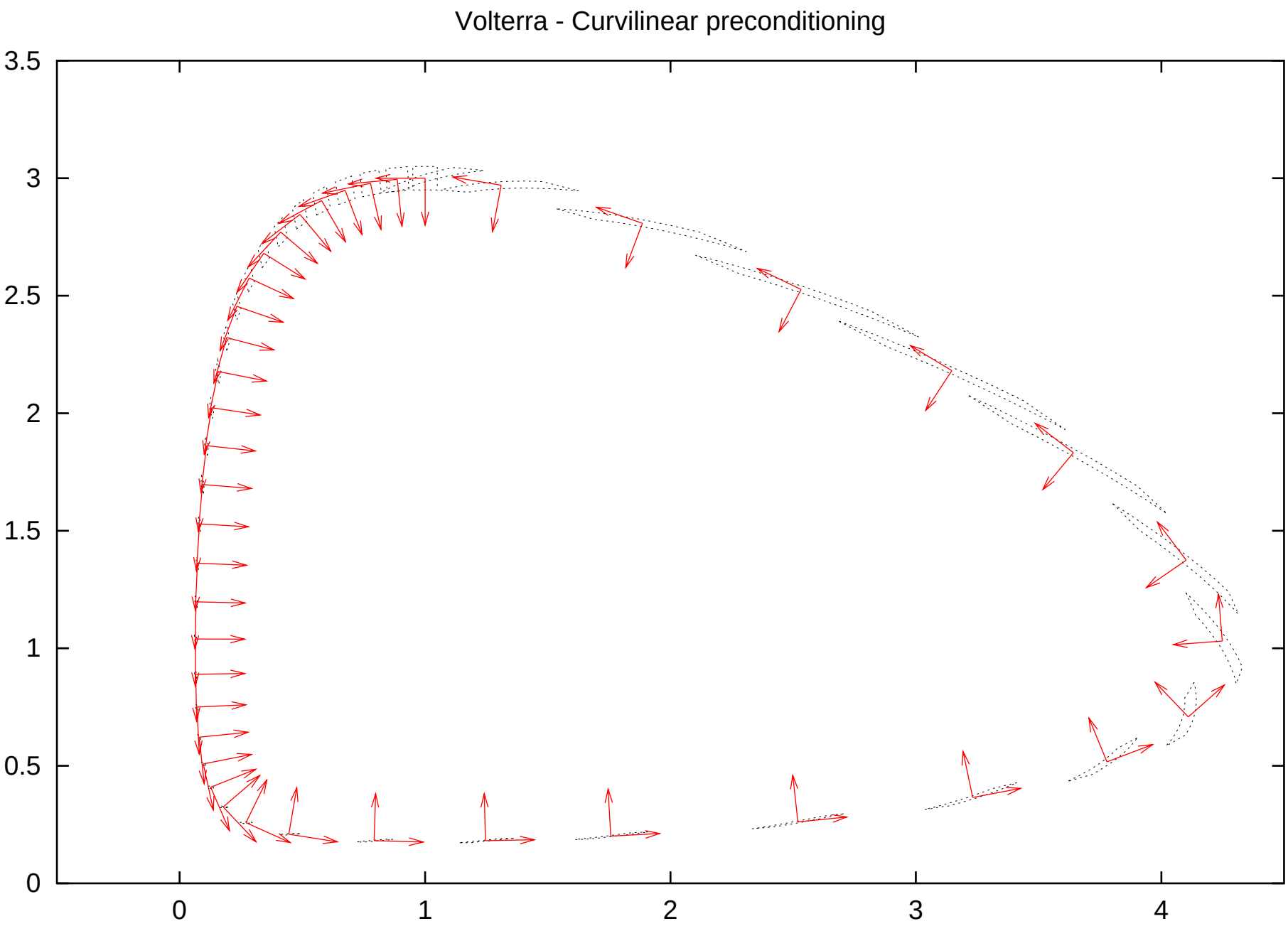
$$(P^*_{l,n+1}, I^*_{l,n+1}) \circ (P_{r.n} + I_{r,n})$$

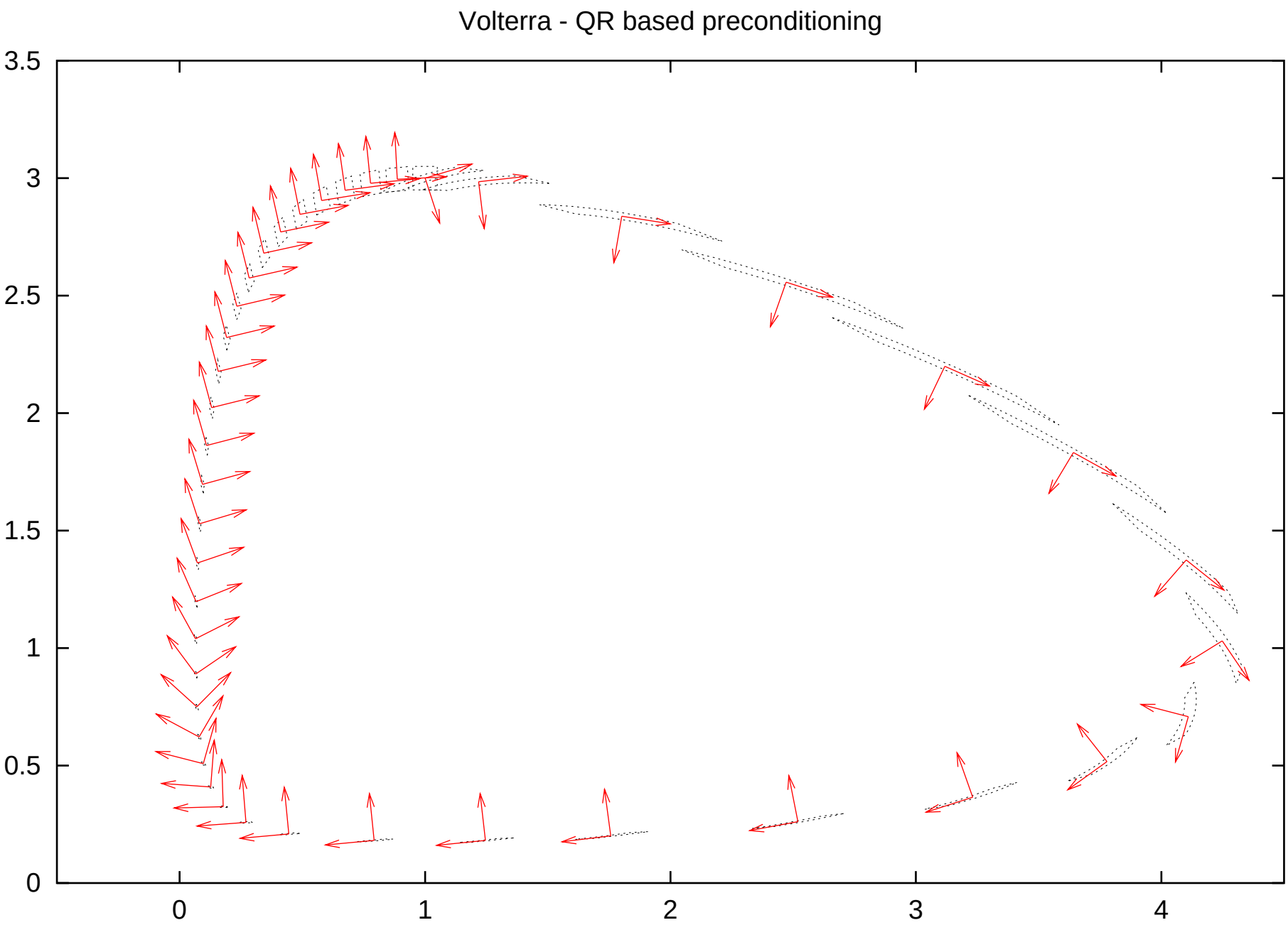
Definition (Curvilinear Preconditioning) Let $x^{(m)} = f(x, x', ...x^{(m-1)}, t)$ be an m-th order ODE in n variables. Let $x_r(t)$ be a solution of the ODE and $x'_r(t)$, ..., $x_r^{(k)}(t)$ its first k time derivatives. Let $\vec{e}_1, ..., \vec{e}_l$ be the l unit vectors not in the span of $x'_r(t)$, ..., $x_r^{(k)}(t)$, sorted by distance from the span. Then we call the Gram-Schmidt orthonormalization of the set $(x'_r(t)$, ..., $x_r^{(k)}(t), \vec{e}_1, ..., \vec{e}_l)$ the curvilinear basis of depth k.

Curvilinear coordinates have long history. Study of solar system, Beam Physics, ...

Example (Curvilinear Solar System and Particle Accelerators) In this case, $n = 3$, and one usually chooses $k = 2$. The first basis vector points in the direction of motion of the reference orbit. The second vector is perpendicular to it and points approximately to the sun or the center of the accelerator. The third vector is chosen perpendicular to the plane of the previous two.


$\vec{e}_y$
$\vec{e}_x$
$\vec{R}(s)$
s
$\vec{e}_s$
x_3
x_2
x_1
Reference Curve


Volterra - Curvilinear preconditioning
3.5
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4


Volterra - QR based preconditioning
3.5
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4

Preconditioning the Flow III

Theorem (Curvilinear Coordinates for Autonomous Linear Systems) Let $x' = \hat{A} \cdot x$ be an n-dimensional linear system that has n distrinct nonzero eigenvalues λ_i with eigenvectors a_i. Let B be a box with nonzero volume, and $x_r = \sum_{i=1}^{n} X_i a_i \in B$ such that $X_i \neq 0$ for all $i = 1, ..., n$. Then the derivatives of $x_r^{(i)}$, $i = 1, ..., n$, are linearly independent, and hence the depth n curvilinear coordinates are obtained by applying the Gram-Schmidt procedure to the derivatives $x_r^{(i)}$, $i = 1, ..., n$.

Proof. The motion of the reference point x_r as a function of time is apparently given by

$$x_r(t) = \sum_{i=1}^{n} X_i \cdot a_i \cdot \exp(\lambda_i t)$$

so that the jth derivative assumes the form

$$x_r^{(j)}(t) = \sum_{i=1}^{n} X_i \cdot a_i \cdot \lambda_i^j \exp(\lambda_i t).$$

We now consider the determinant of the matrix of coefficients in the basis

a_i, and observe

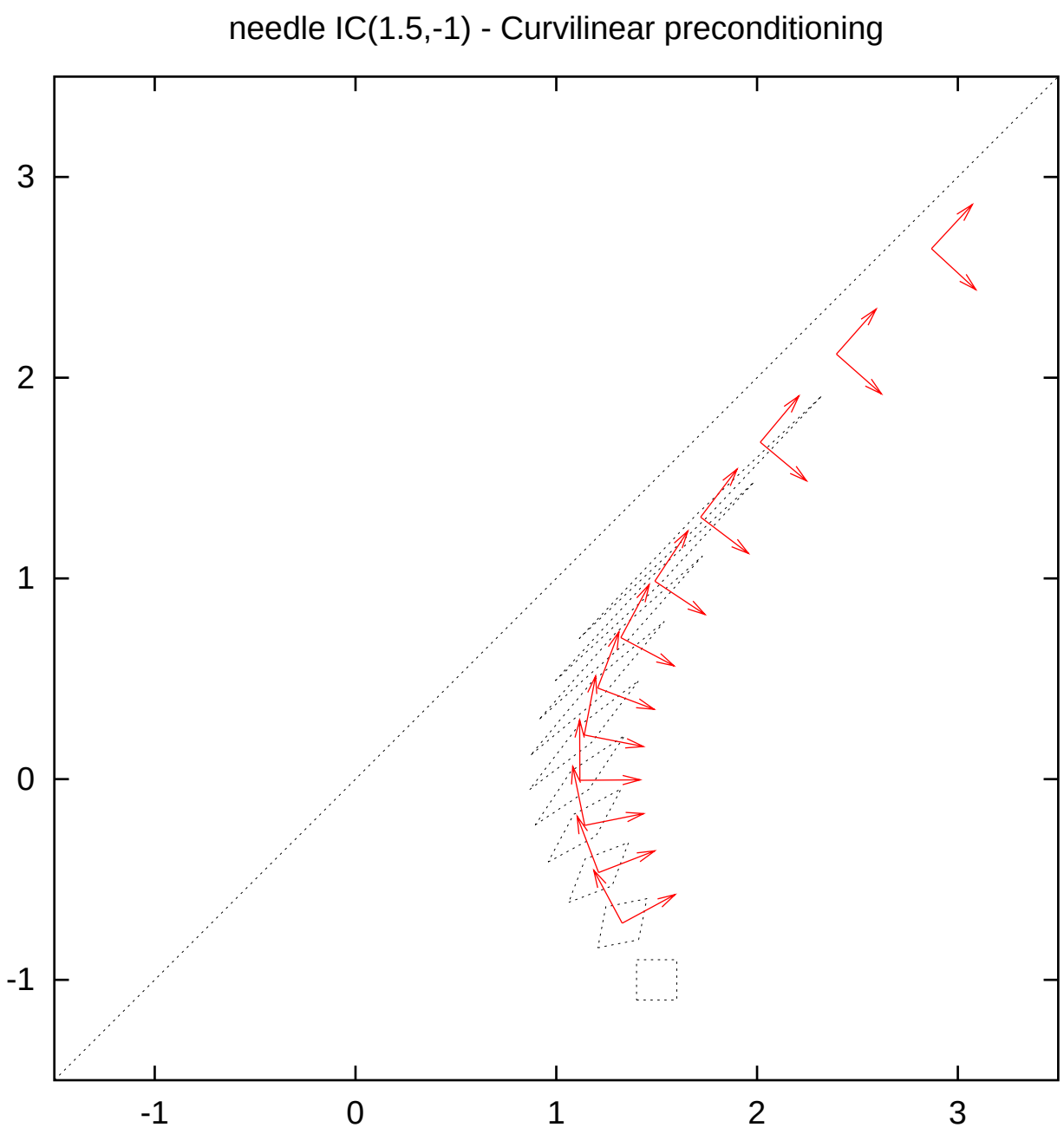
$$\det \begin{pmatrix} X_1\lambda_1 & X_1\lambda_1^2 & & X_1\lambda_1^n \\ X_2\lambda_2 & X_2\lambda_2^2 & & X_2\lambda_2^n \\ & & \ddots & \\ X_n\lambda_n & X_n\lambda_n^2 & & X_n\lambda_n^n \end{pmatrix}$$

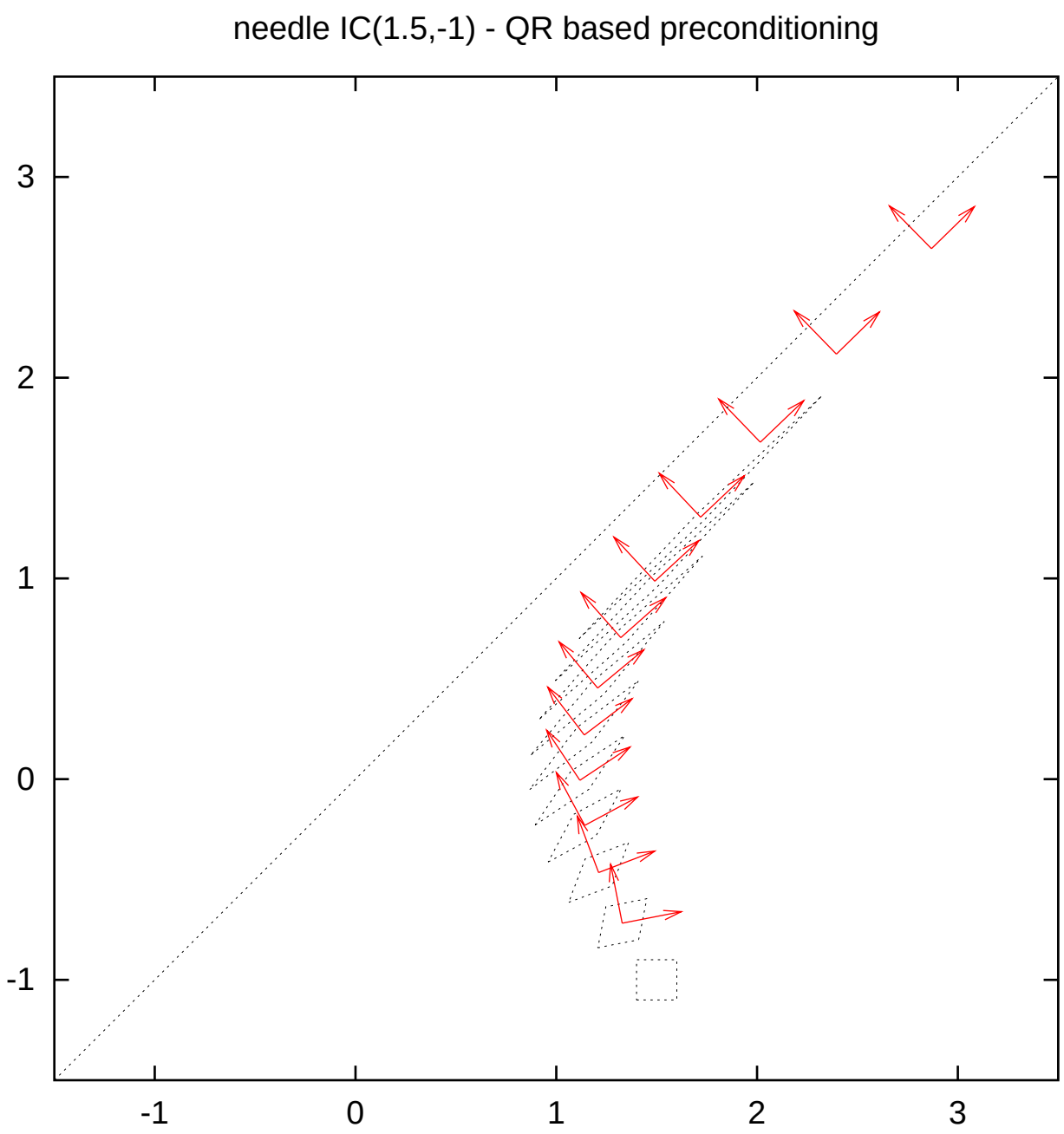
$$= \prod_{i=1}^{n} (\lambda_i X_i)^n \cdot \det \begin{pmatrix} 1 & \lambda_1^1 & & \lambda_1^{n-1} \\ 1 & \lambda_2^1 & & \lambda_2^{n-1} \\ & & \ddots & \\ 1 & \lambda_n^1 & & \lambda_n^{n-1} \end{pmatrix} = \prod_{i=1}^{n} (\lambda_i X_i)^n \prod_{i>j} (\lambda_i - \lambda_j) \neq 0$$

because of the well-known property of the Vandermonde matrix.

Definition (Natural Coordinate System for Linear System) Let $x' = \hat{A} \cdot x$ be an n-dimensional linear system that has n distrinct real eigenvalues $\lambda_1 > \lambda_2 > ... > \lambda_n$ with eigenvectors $a_1, ..., a_n$. We define the normal basis (b_i) of the system to be the result of applying the Gram-Schmidt orthonormalization procedure to the vectors $a_1, ..., a_n$, i.e. the result of the recursive computation

$$b_i = \frac{a_i - \sum_{j=1}^{i-1} b_j \cdot (a_i \cdot b_j)}{\left| a_i - \sum_{j=1}^{i-1} b_j \cdot (a_i \cdot b_j) \right|}.$$


needle IC(1.5,-1) - Curvilinear preconditioning
3
2
1
0
-1
-1
0
1
2
3


needle IC(1.5,-1) - QR based preconditioning
3
2
1
0
-1
-1
0
1
2
3

The Natural Coordinate System has the following important property:

Proposition (Curvilinear Coordinates for Autonomous Linear Systems) Let $x' = \hat{A} \cdot x$ be an n-dimensional linear system that has n distrinct real eigenvalues λ_i with eigenvectors a_i. Let b_i be the natural coordinate system of the linear system. Let B be a box with nonzero volume, and $x_r = \sum_{i=1}^{n} X_i a_i \in B$ such that $X_i \neq 0$. If x_r is used as the reference orbit to define the curvilinear coordinates c_i, then the curvilinear coordinates converge to the natural coordinates, i.e. we have

$$c_i \to b_i \text{ for all } i \text{ as } t \to \infty.$$

Remark: Variations are possible to treat the case of multiple eigenvalues.

A Muon Cooling Ring

Example from Beam Physics: Simple model of muon cooling ring, using curvilinear preconditioning.

Simultaneous damping via matter, and azimuthal accelerations. Equations of motion:

$$\begin{aligned}
\dot{x} &= p_x \\
\dot{y} &= p_y \\
\dot{p}_x &= p_y - \frac{\alpha}{\sqrt{p_x^2 + p_y^2}} \cdot p_x + \frac{\alpha}{\sqrt{x^2 + y^2}} \cdot y \\
\dot{p}_y &= -p_x - \frac{\alpha}{\sqrt{p_x^2 + p_y^2}} \cdot p_y - \frac{\alpha}{\sqrt{x^2 + y^2}} \cdot x
\end{aligned}$$

Has invariant solution

$$(x, y, p_x, p_y)_I(t) = (\cos t, -\sin t, -\sin t, -\cos t),$$

ODE asymptotically approach circular motion of the form

$$(x, y, p_x, p_y)_a(t) = (\cos(t - \phi), -\sin(t - \phi), -\sin(t - \phi), -\cos(t - \phi)),$$

where ϕ is a characteristic angle for each particle.

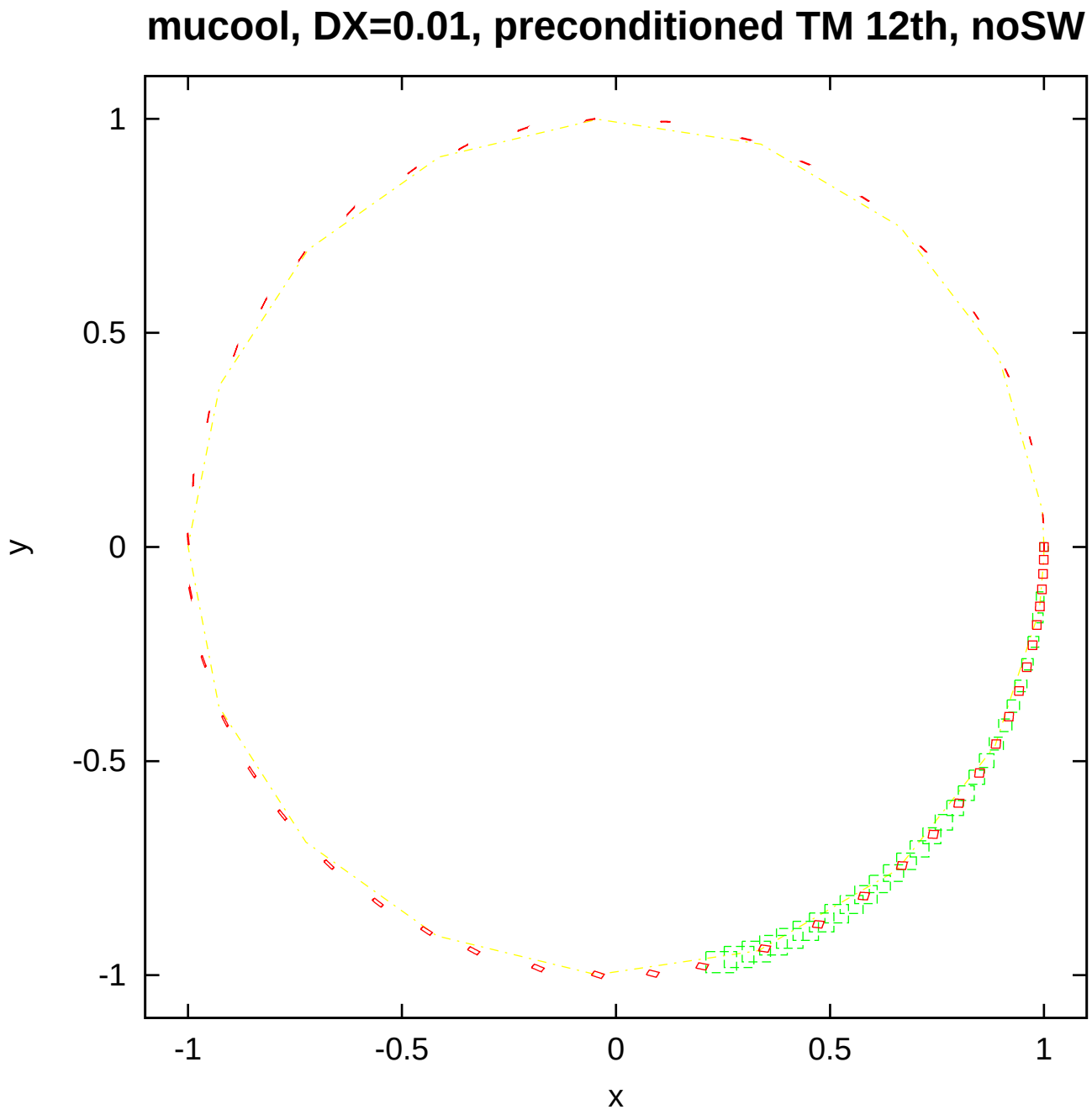
A Muon Cooling Ring - Results

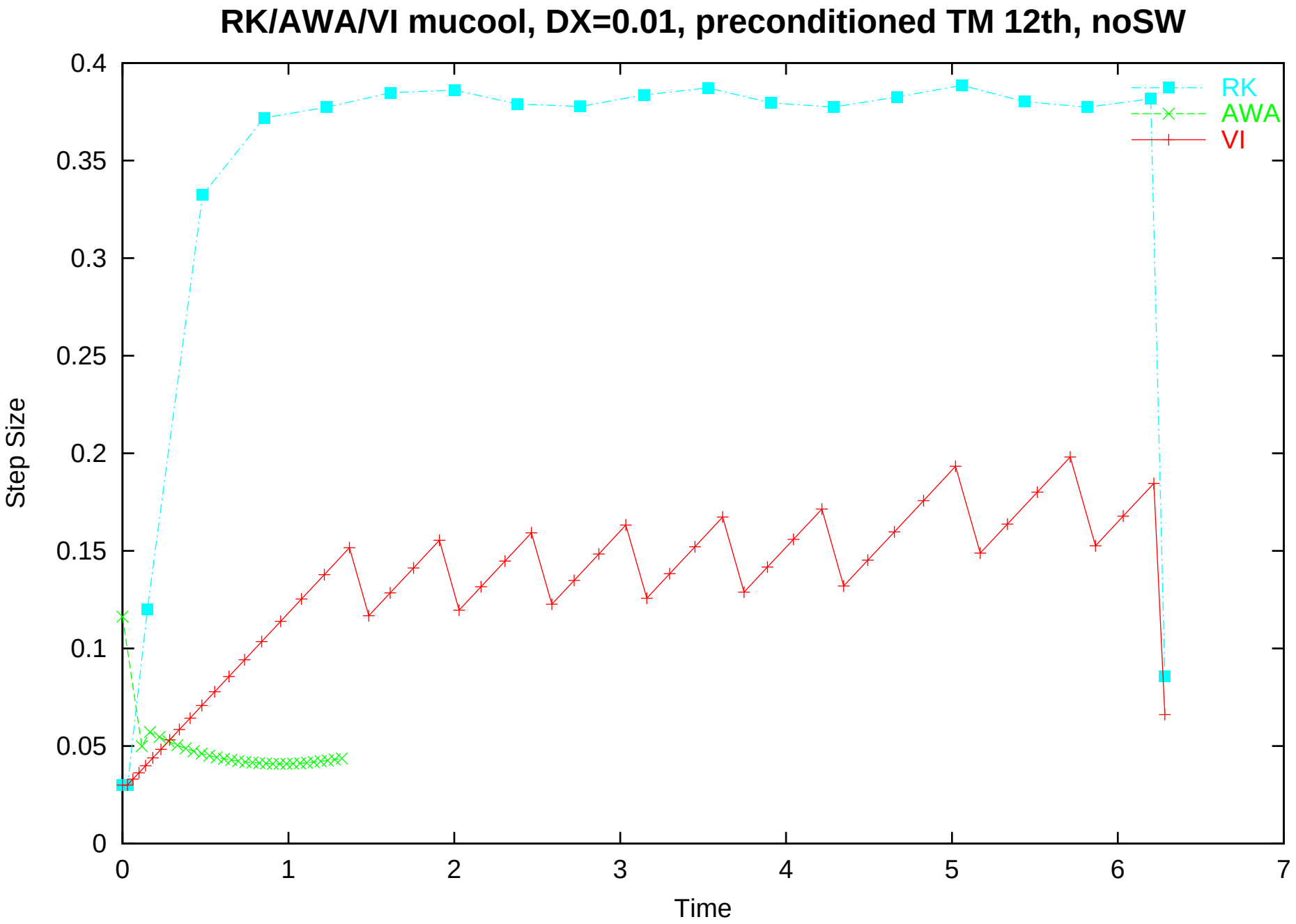
1. Need to treat a large box of $[-10^{-2}, 10^{-2}]^4$
2. Because of damping action towards the invariant limit cycle, the linear part of the motion is more and more ill-conditioned.

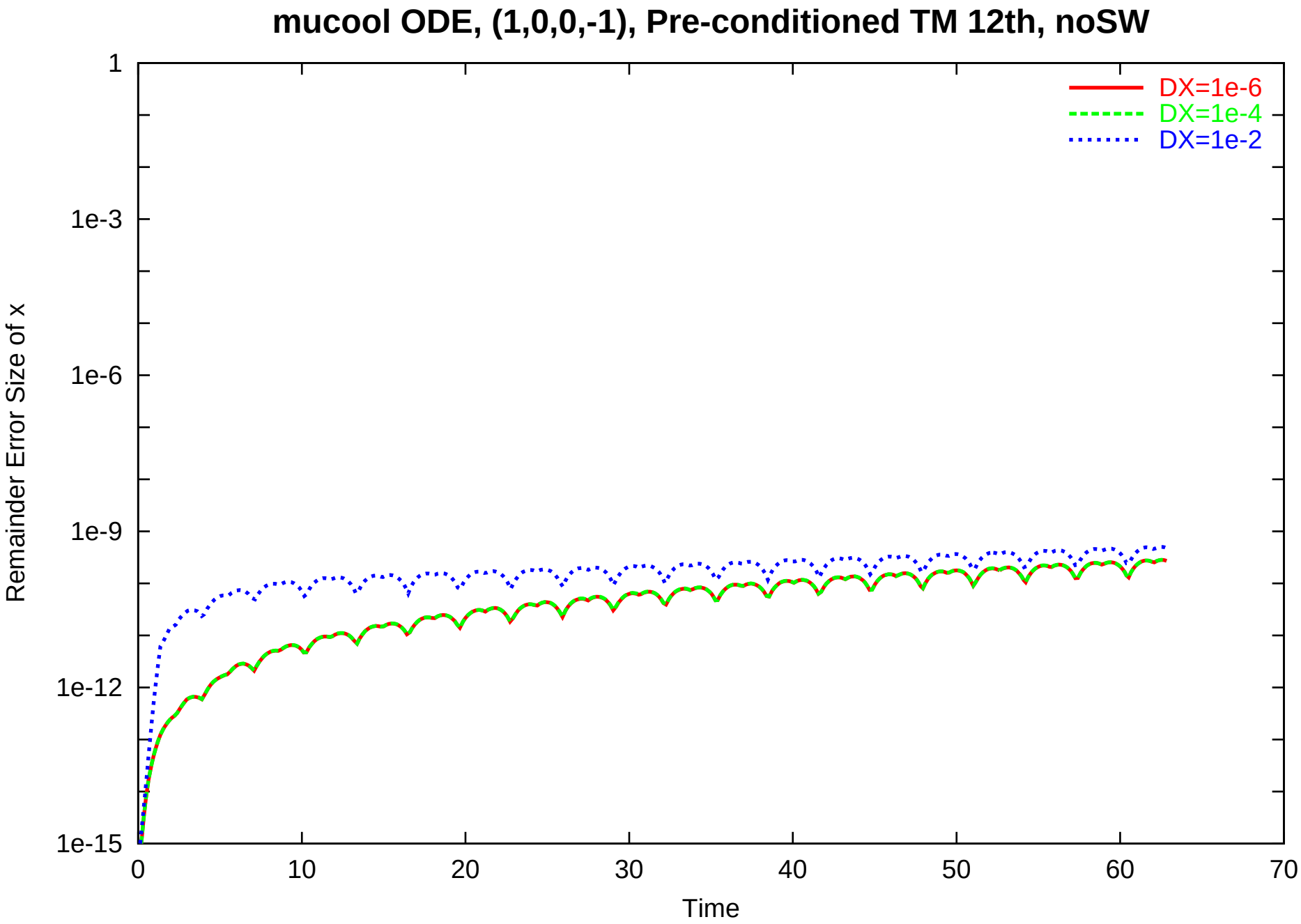
COSY easily integrates 10 cycles for $d = 10^{-2}$ with curvilinear preconditioning and QR preconditioning. AWA (method 4) behaves as follows:

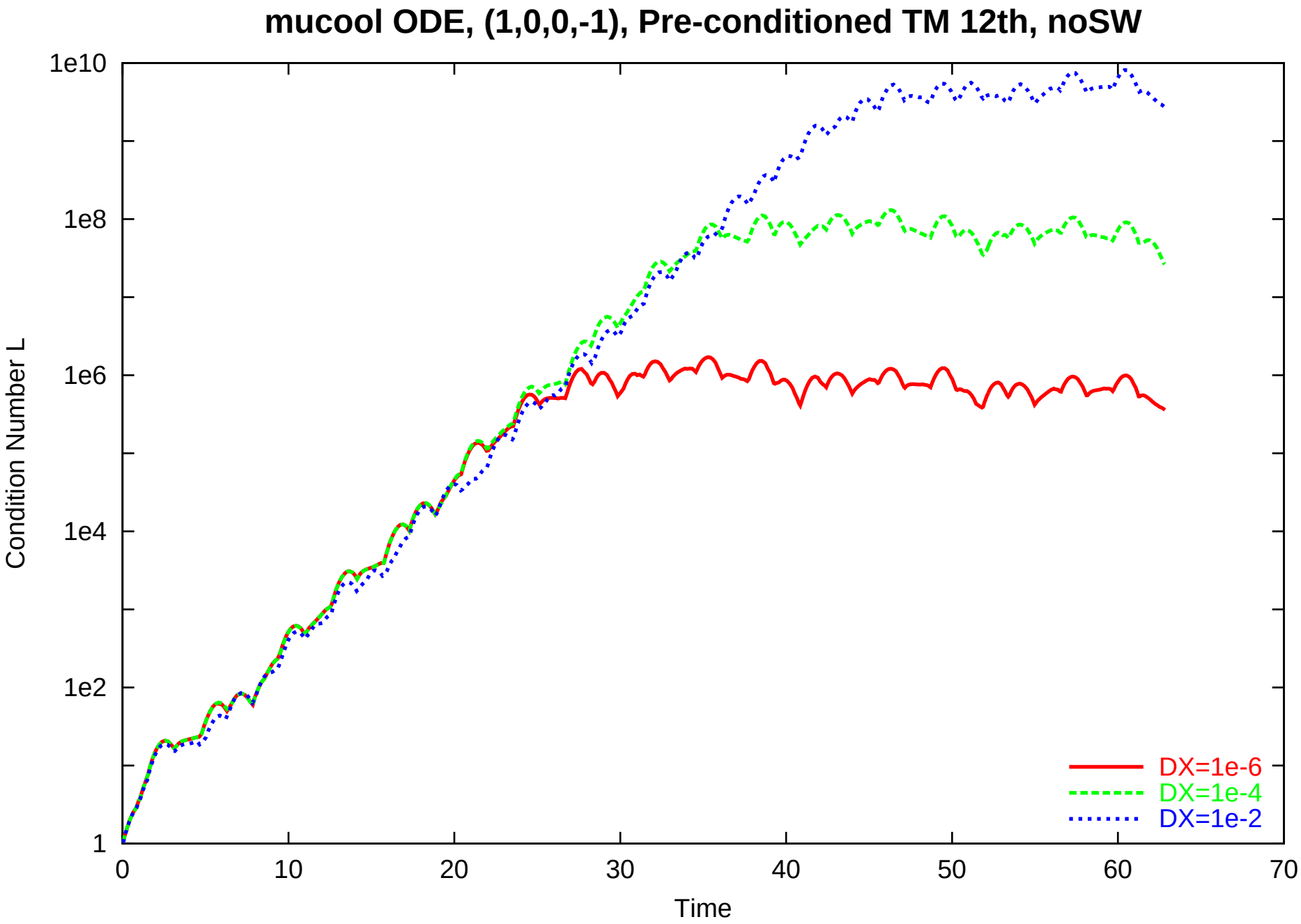
d	**Cycles**
10^{-2}	0.22
10^{-3}	1.25
10^{-4}	9.5

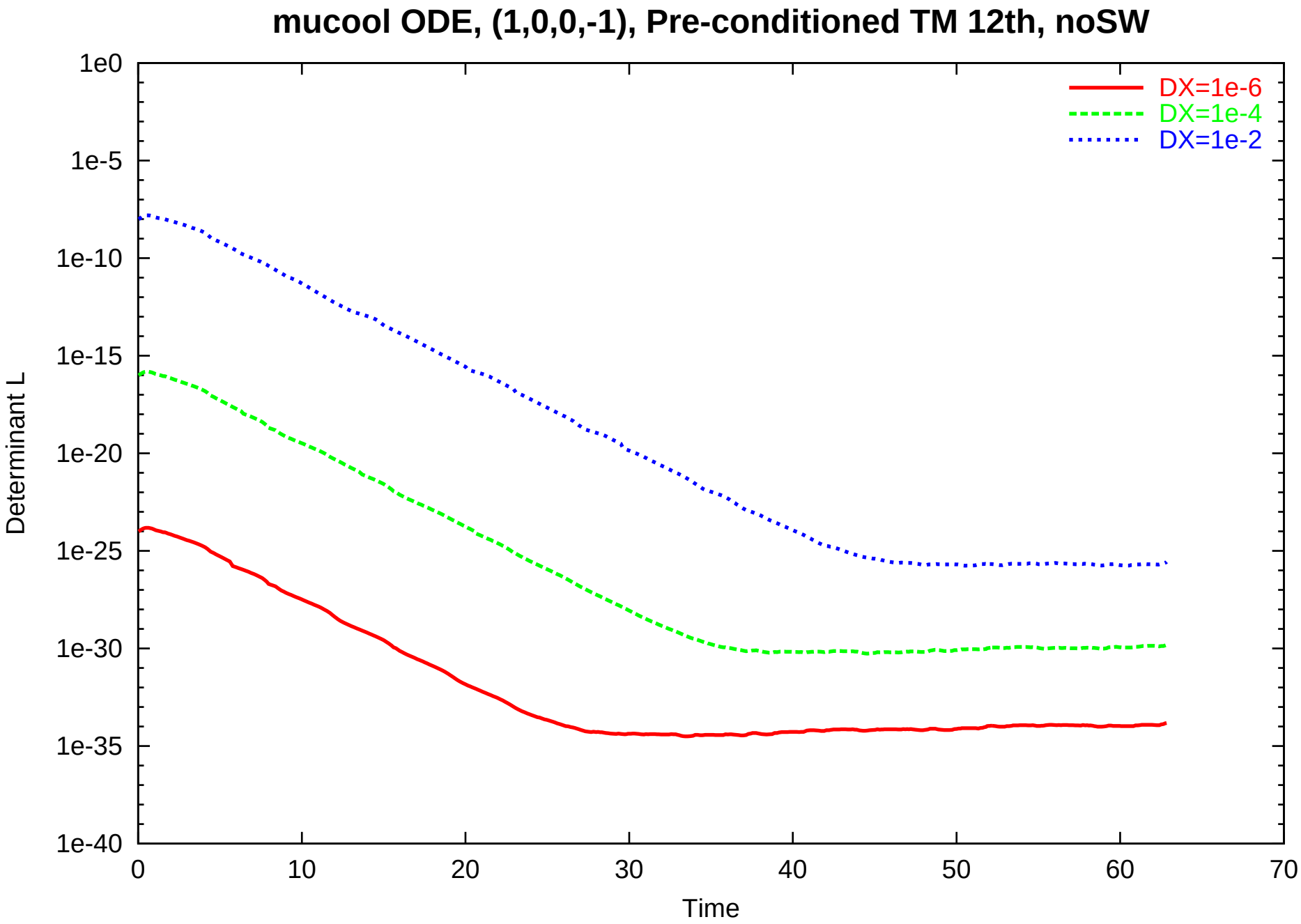
Thus, trying to simulate the system with AWA requires $> (10^2)^4 = 10^8$ subdivisions of the box that COSY can transport in one piece.


mucool, DX=0.01, preconditioned TM 12th, noSW
1
0.5
0
-0.5
-1
y
-1
-0.5
0
0.5
1
x


RK/AWA/VI mucool, DX=0.01, preconditioned TM 12th, noSW
Step Size
Time
RK
AWA
VI
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0
1
2
3
4
5
6
7


mucool ODE, (1,0,0,-1), Pre-conditioned TM 12th, noSW
Remainder Error Size of x
Time
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
10
20
30
40
50
60
70
DX=1e-6
DX=1e-4
DX=1e-2


mucool ODE, (1,0,0,-1), Pre-conditioned TM 12th, noSW
Condition Number L
1e10
1e8
1e6
1e4
1e2
1
0
10
20
30
40
50
60
70
Time
DX=1e-6
DX=1e-4
DX=1e-2


mucool ODE, (1,0,0,-1), Pre-conditioned TM 12th, noSW
Determinant L
Time
1e0
1e-5
1e-10
1e-15
1e-20
1e-25
1e-30
1e-35
1e-40
0
10
20
30
40
50
60
70
DX=1e-6
DX=1e-4
DX=1e-2

A 2D Discrete Kepler Problem

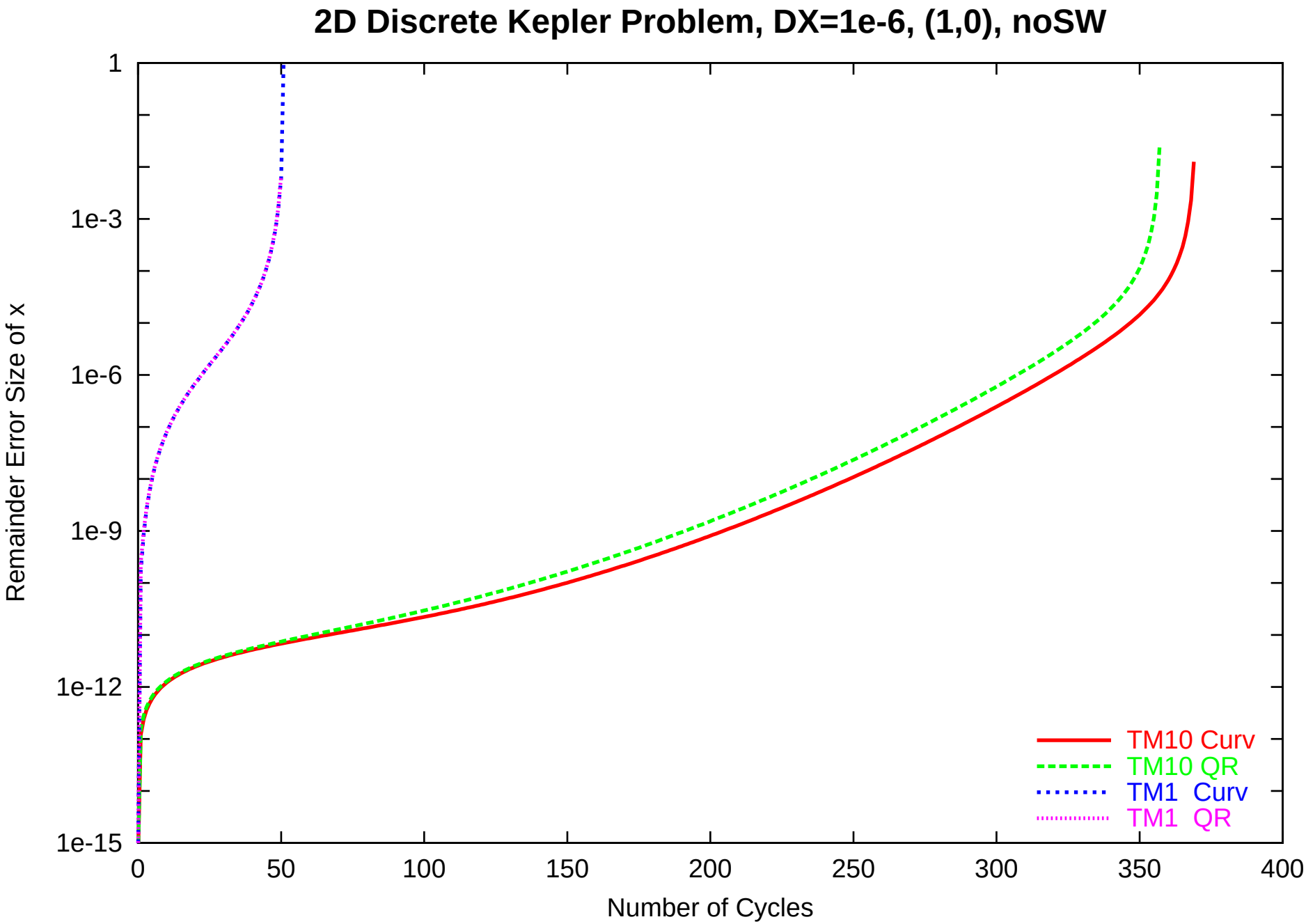
Dynamics of circular Kepler orbits around central mass. Period T and large semi-major axis are related via $T^2 = k \cdot a^3$. So $\omega = 2\pi/T = 2\pi \cdot r^{-3/2}$, and thus after Δt we have

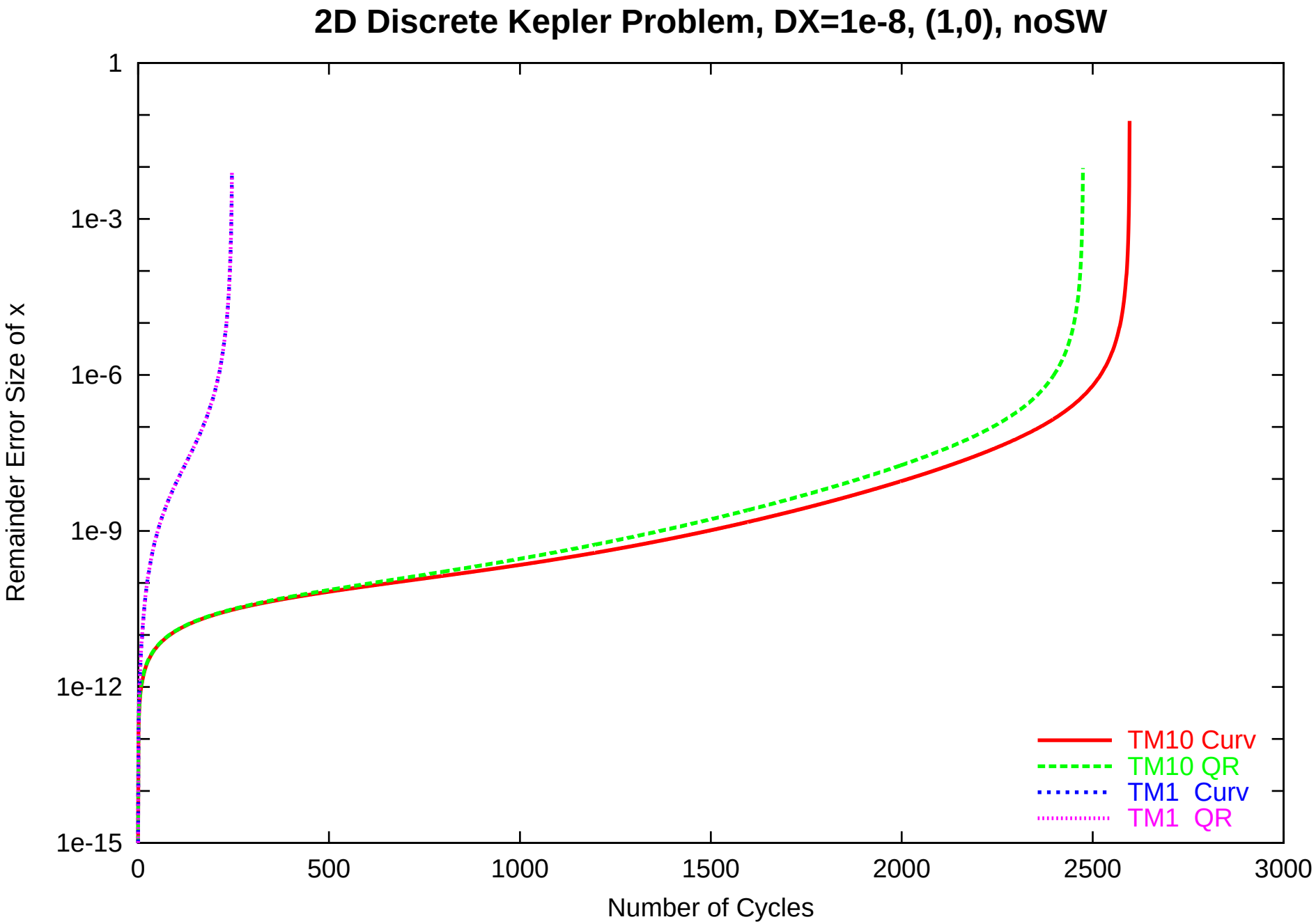
$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} \cos\Delta\phi & \sin\Delta\phi \\ -\sin\Delta\phi & \cos\Delta\phi \end{pmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

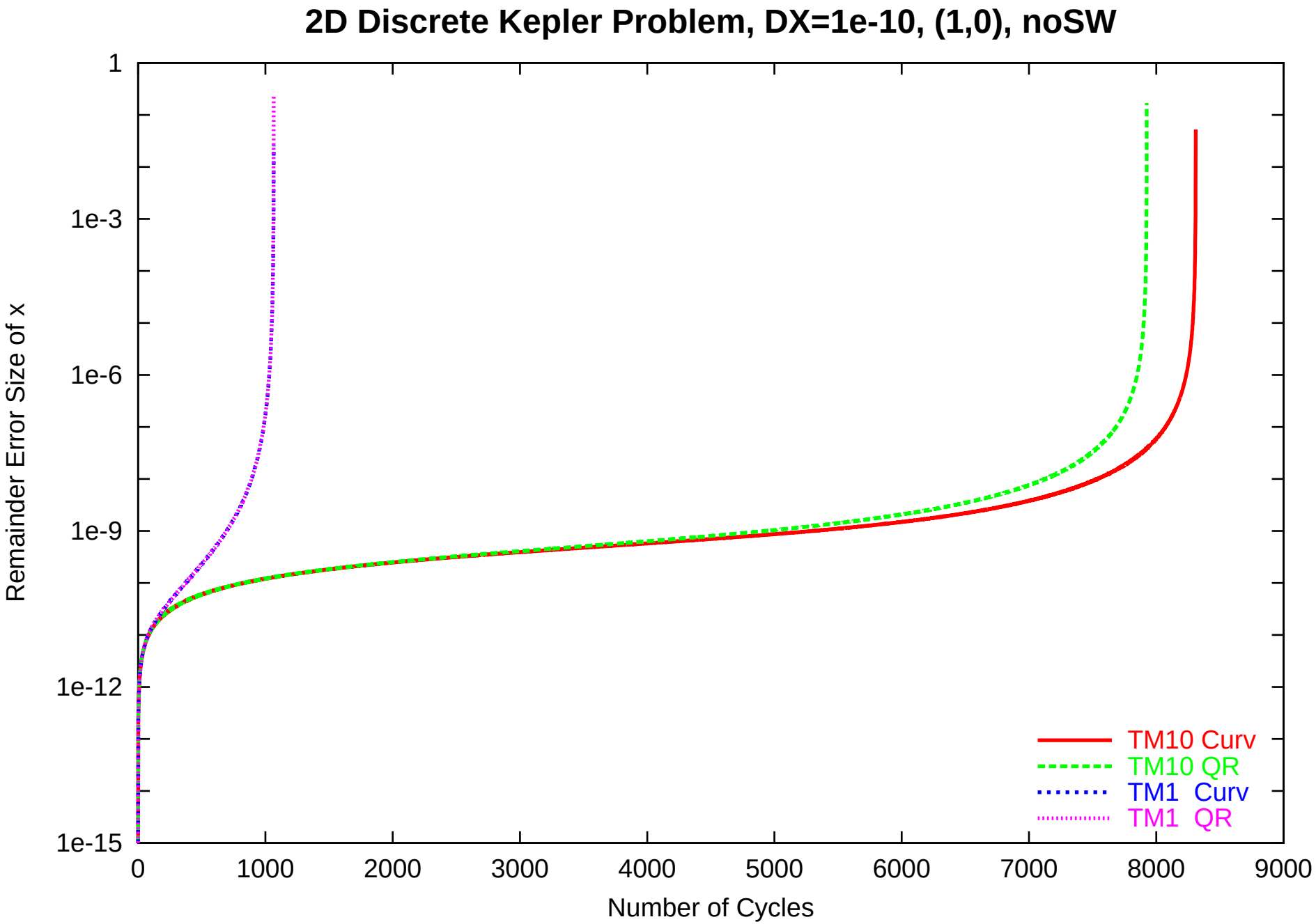
$$\text{where } \Delta\phi = \frac{2\pi\Delta t}{(x^2+y^2)^{3/4}}.$$

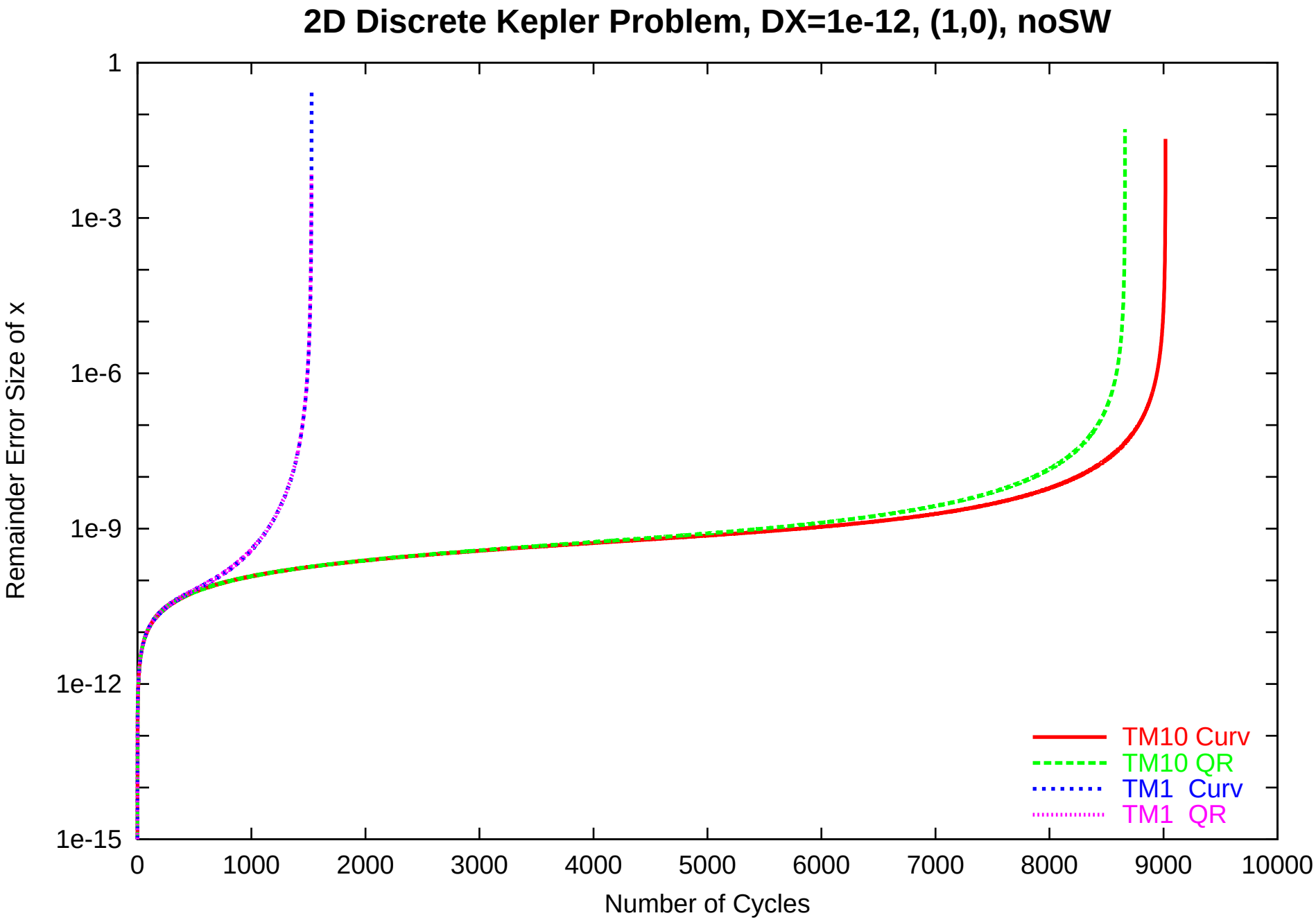
Characteristic of general Kepler problem: as time progresses, larger and larger lag in angle fr different r, resulting in shearing. Circular form makes Taylor expansion of final in terms of initial coordinates ultimately impossible. Thus, any Taylor method will eventually have to fail. The question is, how soon!

Also interesting: estimate growth rate of remainder bounds. For smallest d, have increase of $9 \cdot 10^{-9}$ over $5,000$ revolutions or $40,000$ iterations. This corresponds to about $2 \cdot 10^{-13}$ per map iteration. This is near floating point limit!


2D Discrete Kepler Problem, DX=1e-6, (1,0), noSW
Remainder Error Size of x
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
50
100
150
200
250
300
350
400
Number of Cycles
TM10 Curv
TM10 QR
TM1 Curv
TM1 QR


2D Discrete Kepler Problem, DX=1e-8, (1,0), noSW
Remainder Error Size of x
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
500
1000
1500
2000
2500
3000
Number of Cycles
TM10 Curv
TM10 QR
TM1 Curv
TM1 QR


2D Discrete Kepler Problem, DX=1e-10, (1,0), noSW
Remainder Error Size of x
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Number of Cycles
TM10 Curv
TM10 QR
TM1 Curv
TM1 QR


2D Discrete Kepler Problem, DX=1e-12, (1,0), noSW
Remainder Error Size of x
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Number of Cycles
TM10 Curv
TM10 QR
TM1 Curv
TM1 QR

The Henon Map

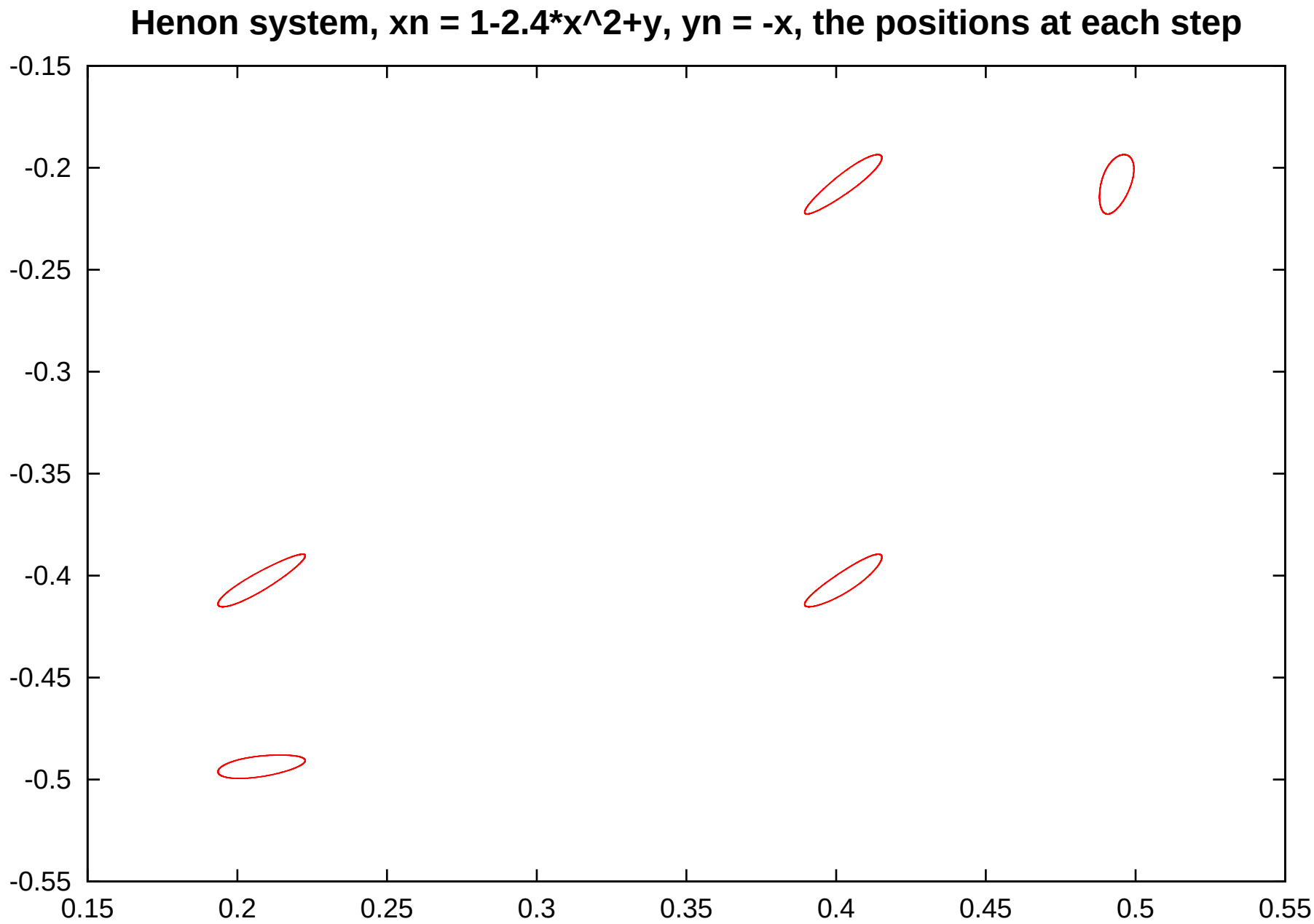
Henon Map: frequently used elementary example that exhibits many of the well-known effects of nonlinear dynamics, including chaos, periodic fixed points, islands and symplectic motion. The dynamics is two-dimensional, and given by

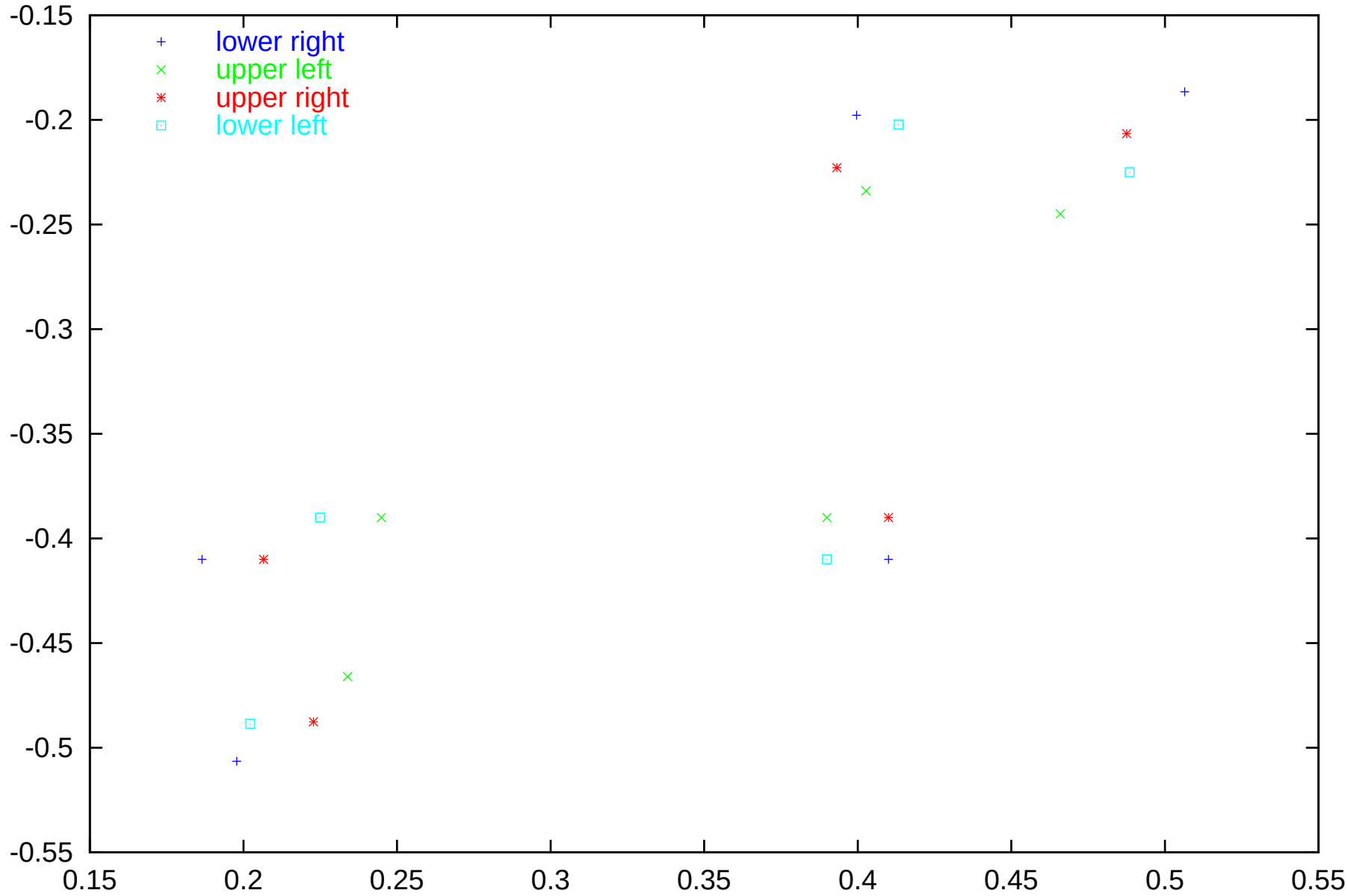
$$x_{n+1} = 1 - \alpha x_n^2 + y_n$$
$$y_{n+1} = \beta x_n.$$

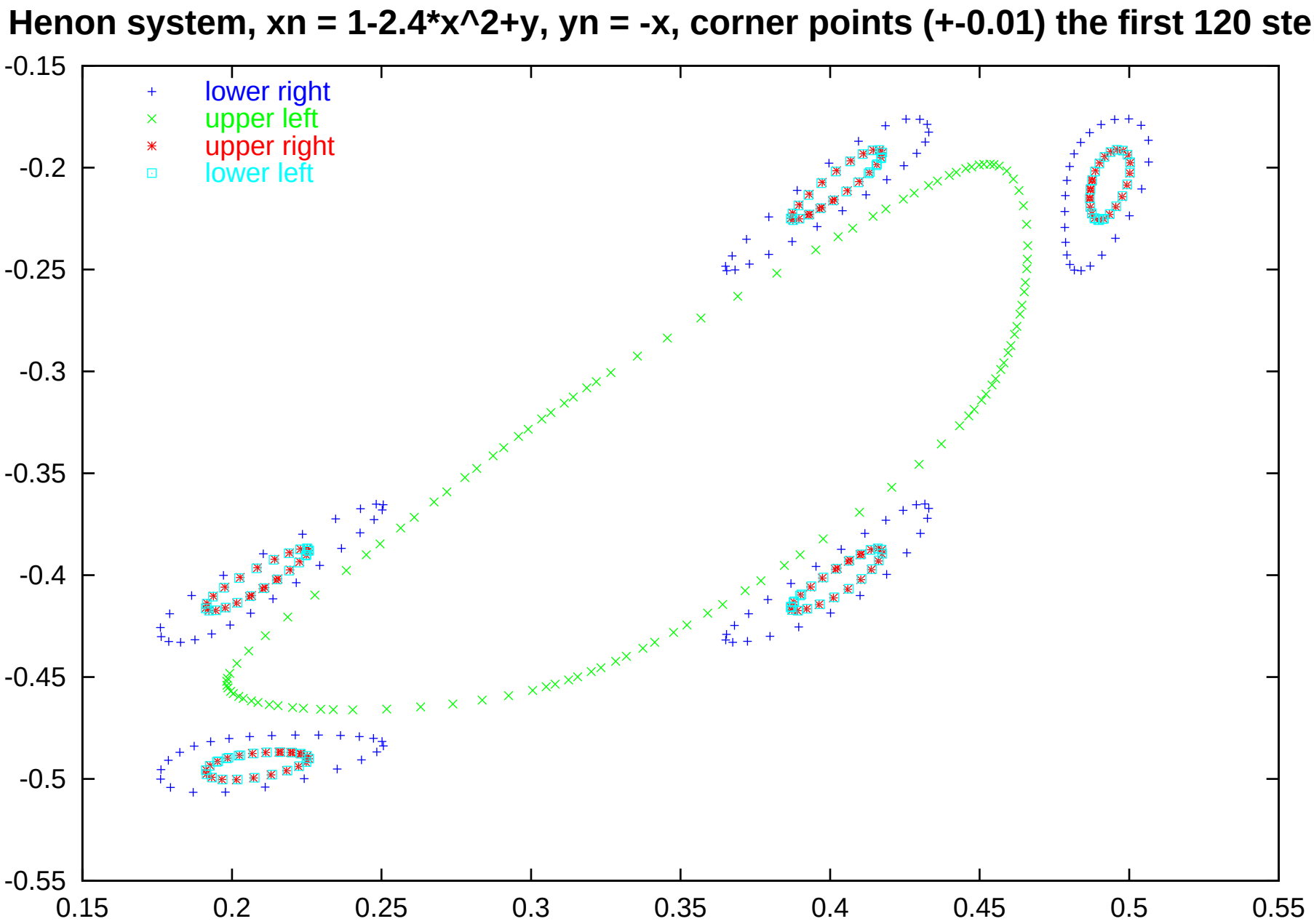
It can easily be seen that the motion is area preserving for $|\beta| = 1$.We consider

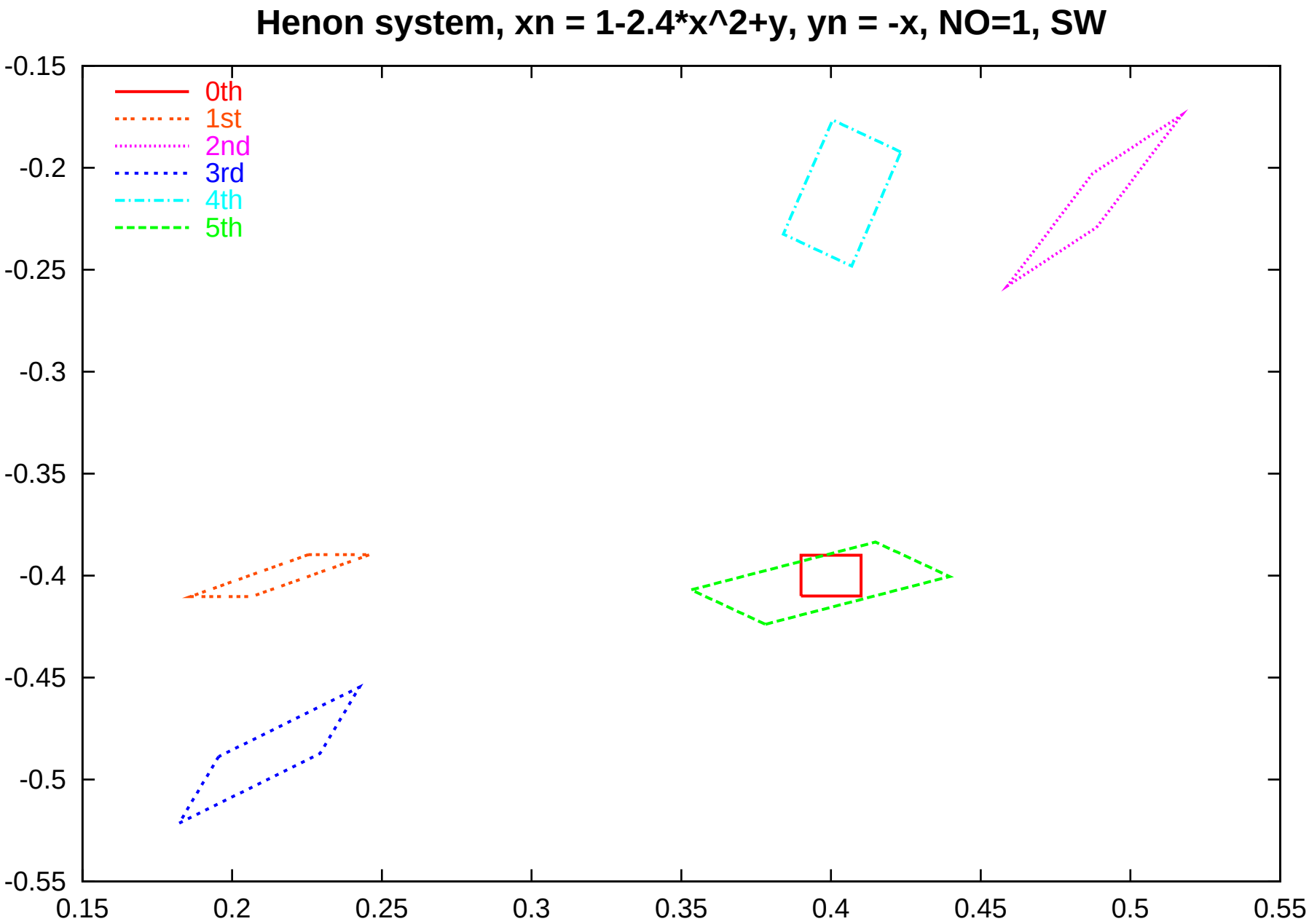
$$\alpha = 2.4 \text{ and } \beta = -1,$$

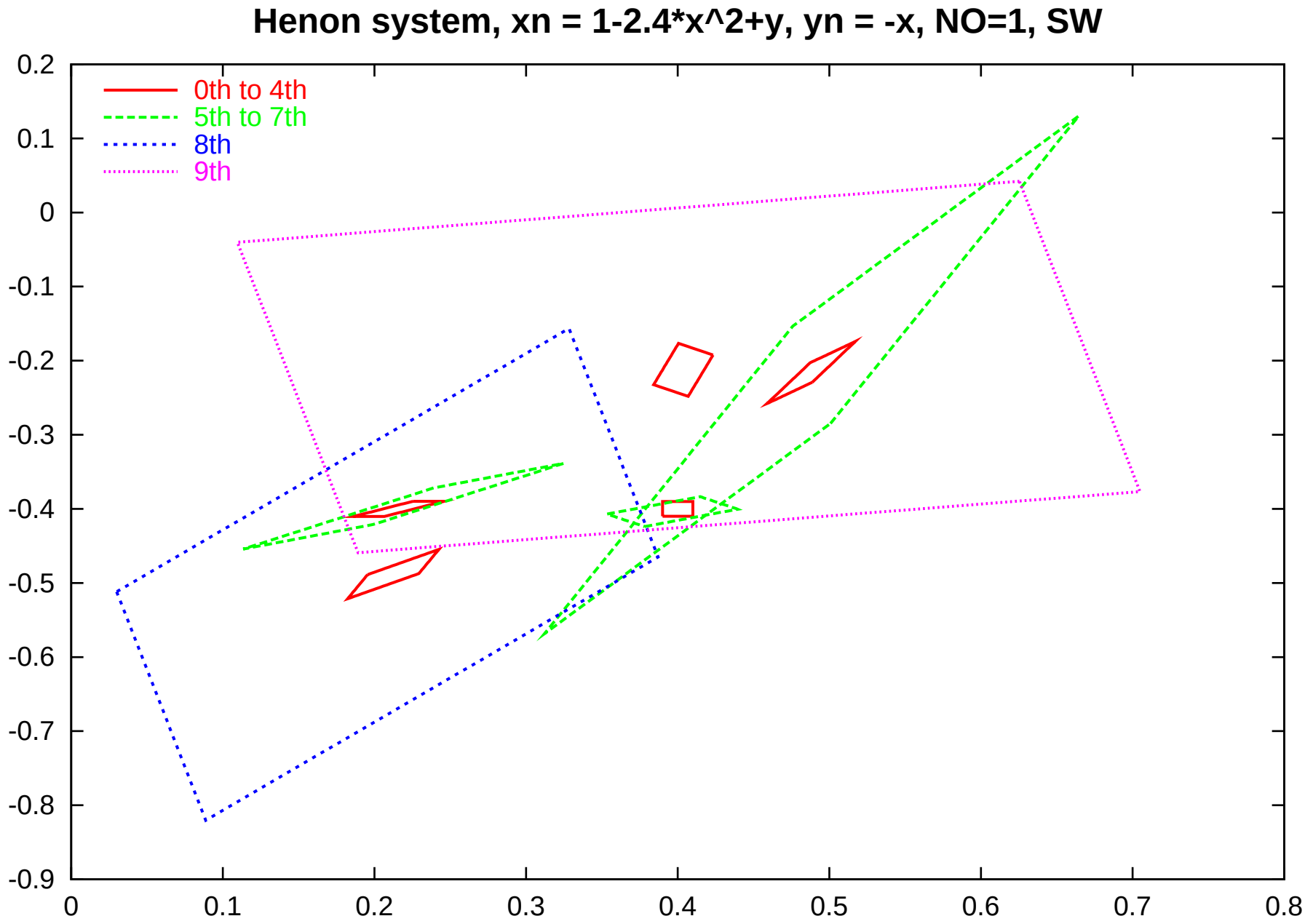
and concentrate on initial boxes of the from $(x_0, y_0) \in (0.4, -0.4) + [-d, d]^2$.

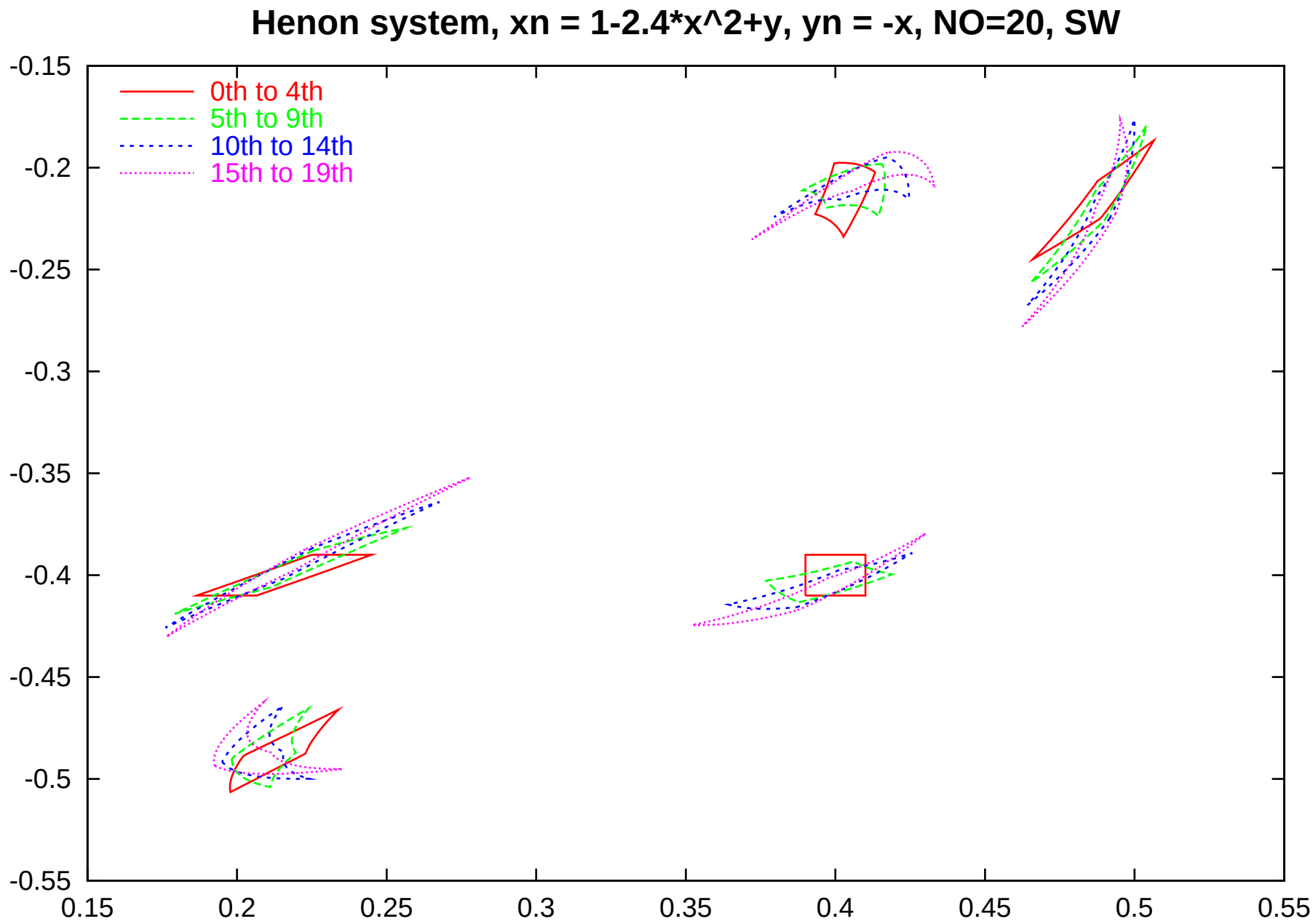

Henon system, xn = 1-2.4*x^2+y, yn = -x, the positions at each step
-0.15
-0.2
-0.25
-0.3
-0.35
-0.4
-0.45
-0.5
-0.55
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55

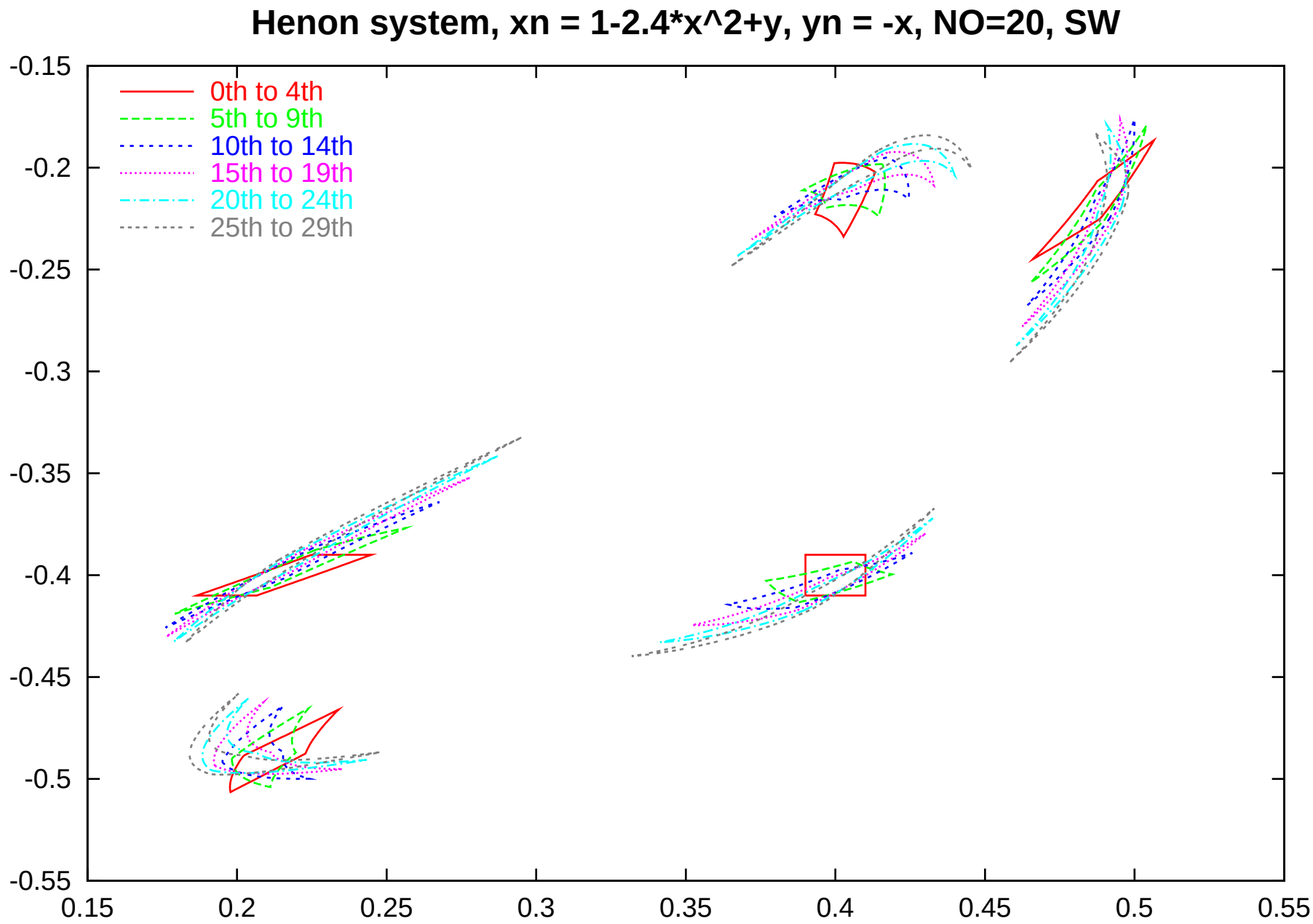

Henon system, xn = 1-2.4*x^2+y, yn = -x, corner points (+-0.01) the first 5 steps
lower right
upper left
upper right
lower left
-0.15
-0.2
-0.25
-0.3
-0.35
-0.4
-0.45
-0.5
-0.55
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55

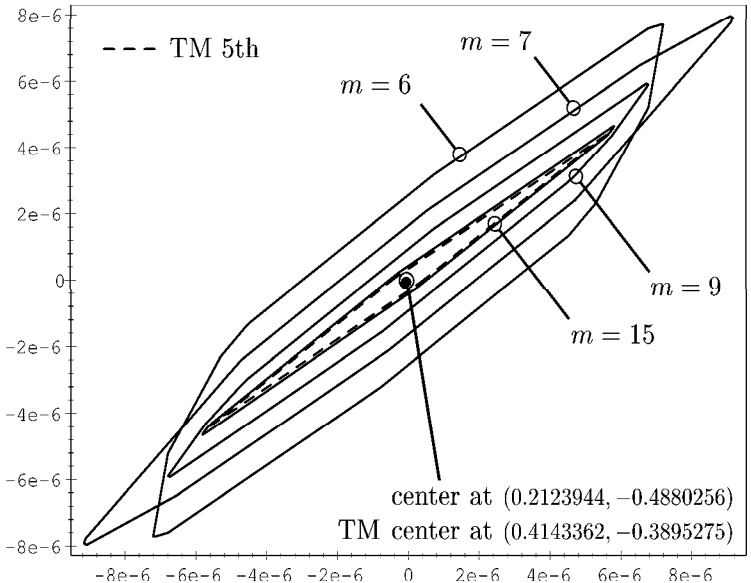

Henon system, xn = 1-2.4*x^2+y, yn = -x, corner points (+-0.01) the first 120 steps
lower right
upper left
upper right
lower left
-0.15
-0.2
-0.25
-0.3
-0.35
-0.4
-0.45
-0.5
-0.55
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55

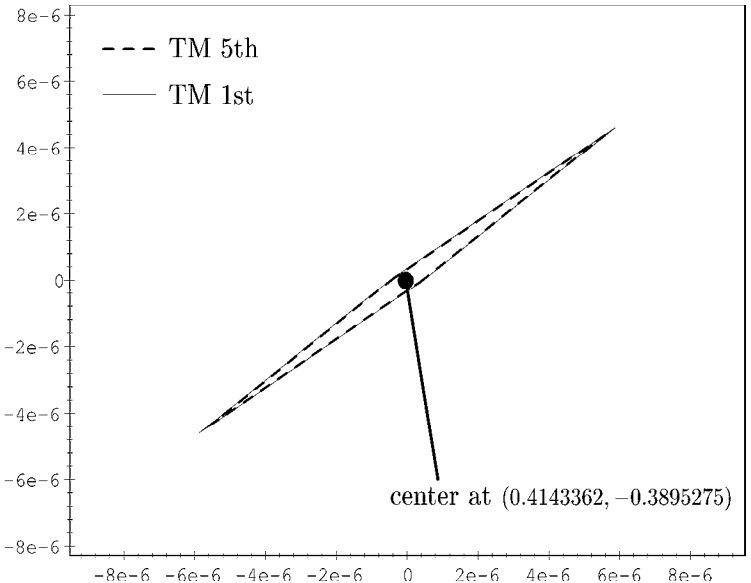

Henon system, xn = 1-2.4*x^2+y, yn = -x, NO=1, SW
0th
1st
2nd
3rd
4th
5th
-0.15
-0.2
-0.25
-0.3
-0.35
-0.4
-0.45
-0.5
-0.55
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55

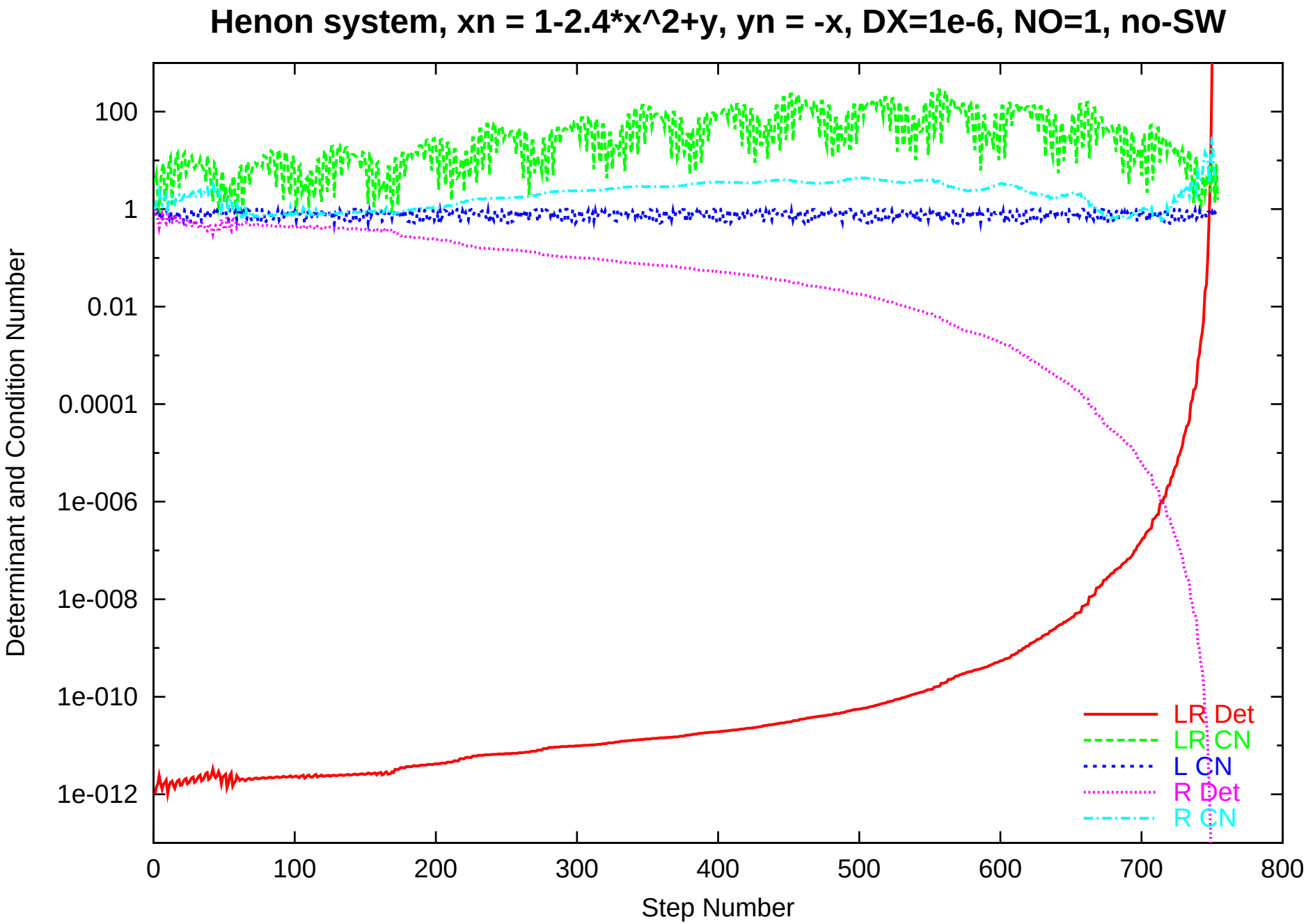

Henon system, xn = 1-2.4*x^2+y, yn = -x, NO=1, SW
0th to 4th
5th to 7th
8th
9th
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

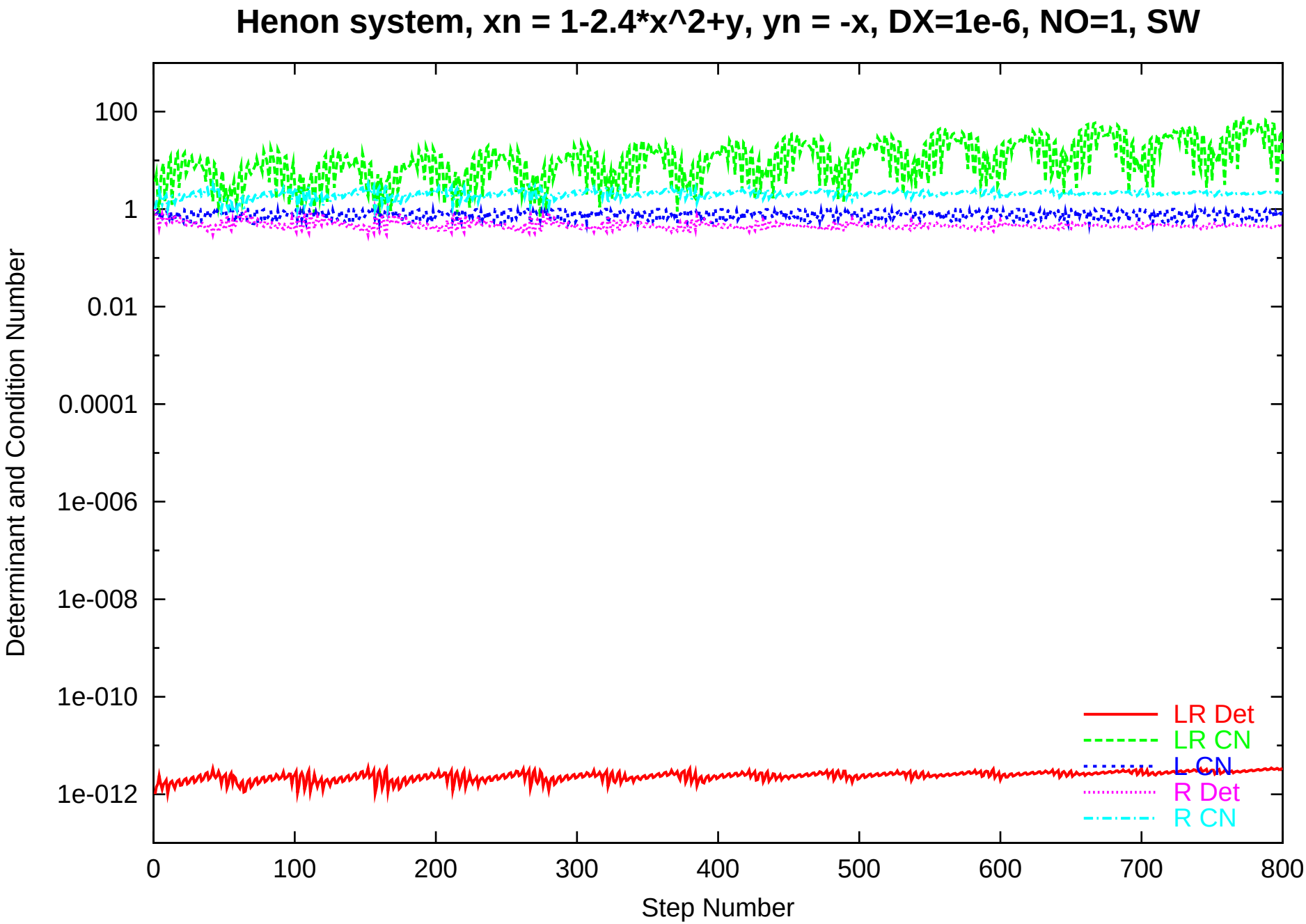

Henon system, xn = 1-2.4*x^2+y, yn = -x, NO=20, SW
0th to 4th
5th to 9th
10th to 14th
15th to 19th
-0.15
-0.2
-0.25
-0.3
-0.35
-0.4
-0.45
-0.5
-0.55
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55

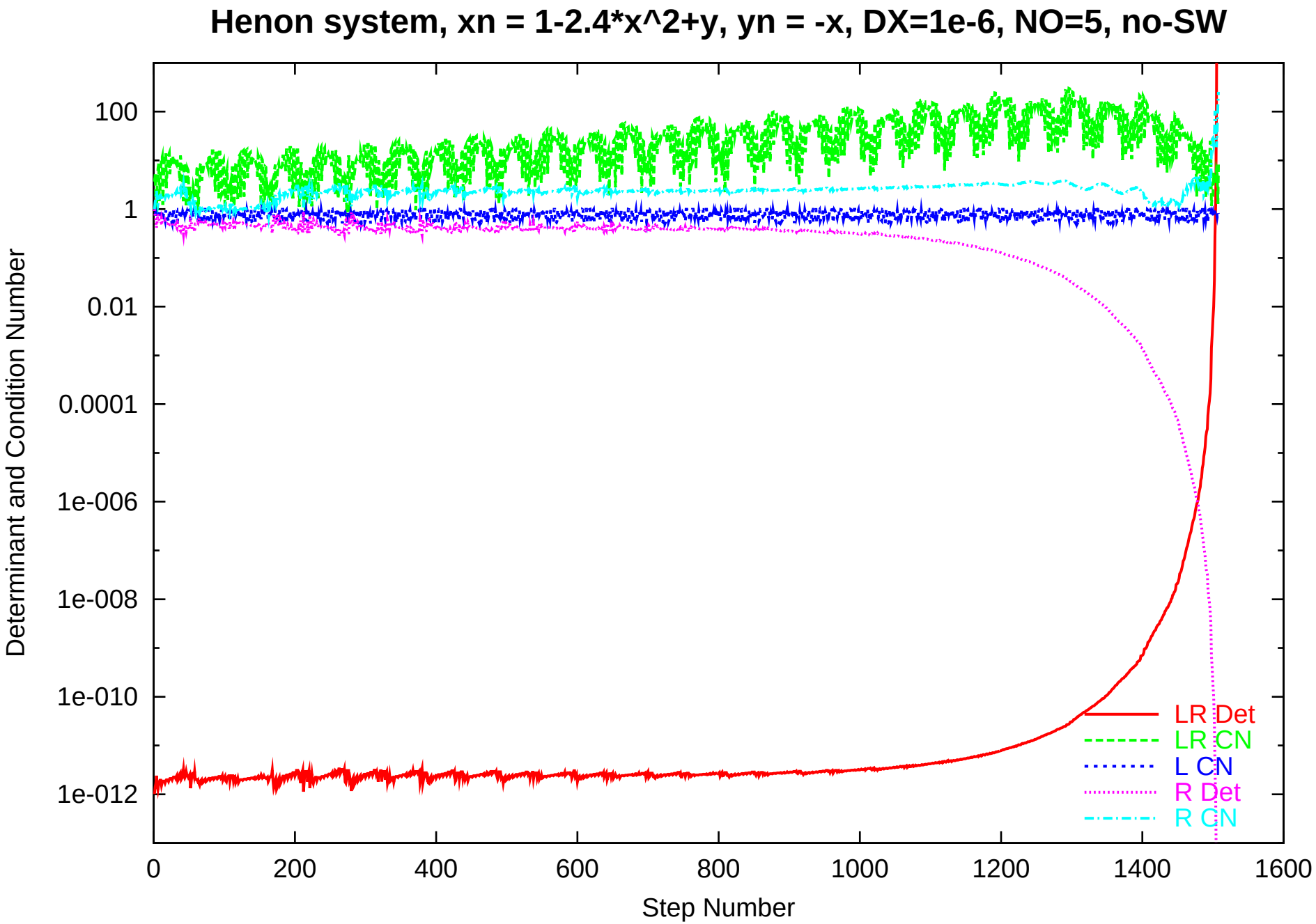

Henon system, xn = 1-2.4*x^2+y, yn = -x, NO=20, SW
0th to 4th
5th to 9th
10th to 14th
15th to 19th
20th to 24th
25th to 29th
-0.15
-0.2
-0.25
-0.3
-0.35
-0.4
-0.45
-0.5
-0.55
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55

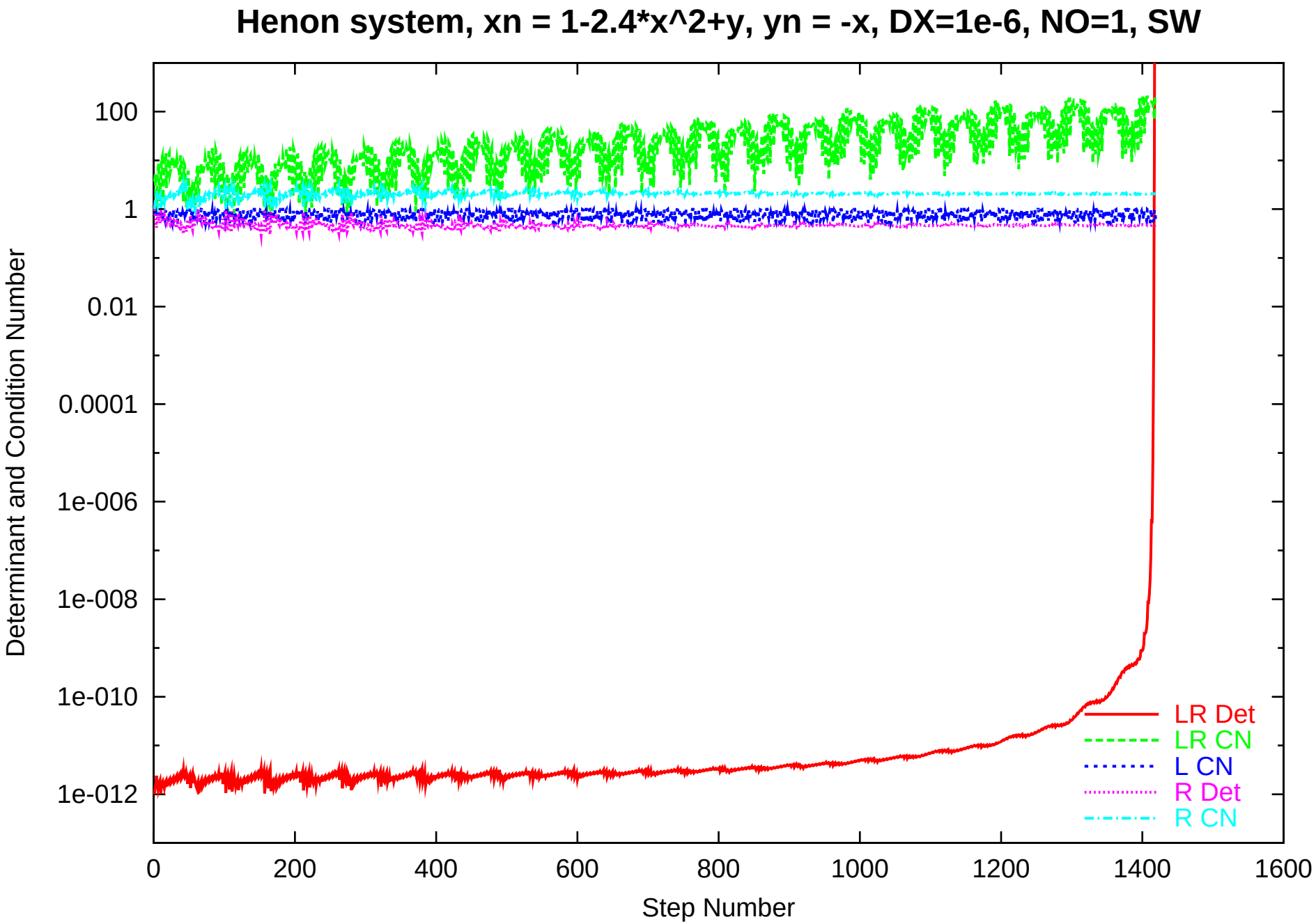

TM 5th
$m = 7$
$m = 6$
$m = 9$
$m = 15$
center at $(0.2123944, -0.4880256)$
TM center at $(0.4143362, -0.3895275)$
8e-6
6e-6
4e-6
2e-6
0
-2e-6
-4e-6
-6e-6
-8e-6
-8e-6 -6e-6 -4e-6 -2e-6 0 2e-6 4e-6 6e-6 8e-6

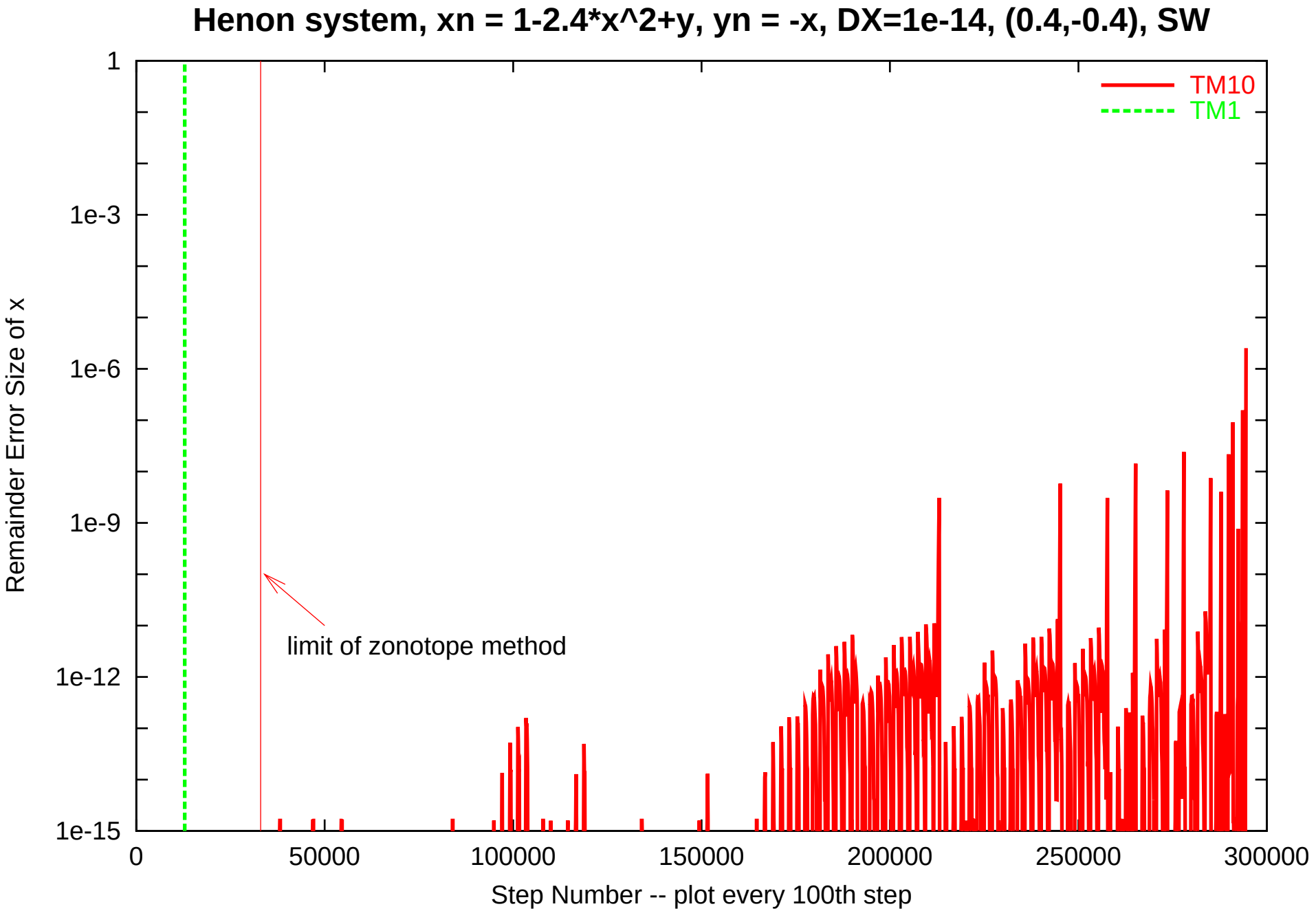

TM 5th
TM 1st
center at $(0.4143362, -0.3895275)$
8e-6
6e-6
4e-6
2e-6
0
-2e-6
-4e-6
-6e-6
-8e-6
-8e-6 -6e-6 -4e-6 -2e-6 0 2e-6 4e-6 6e-6 8e-6

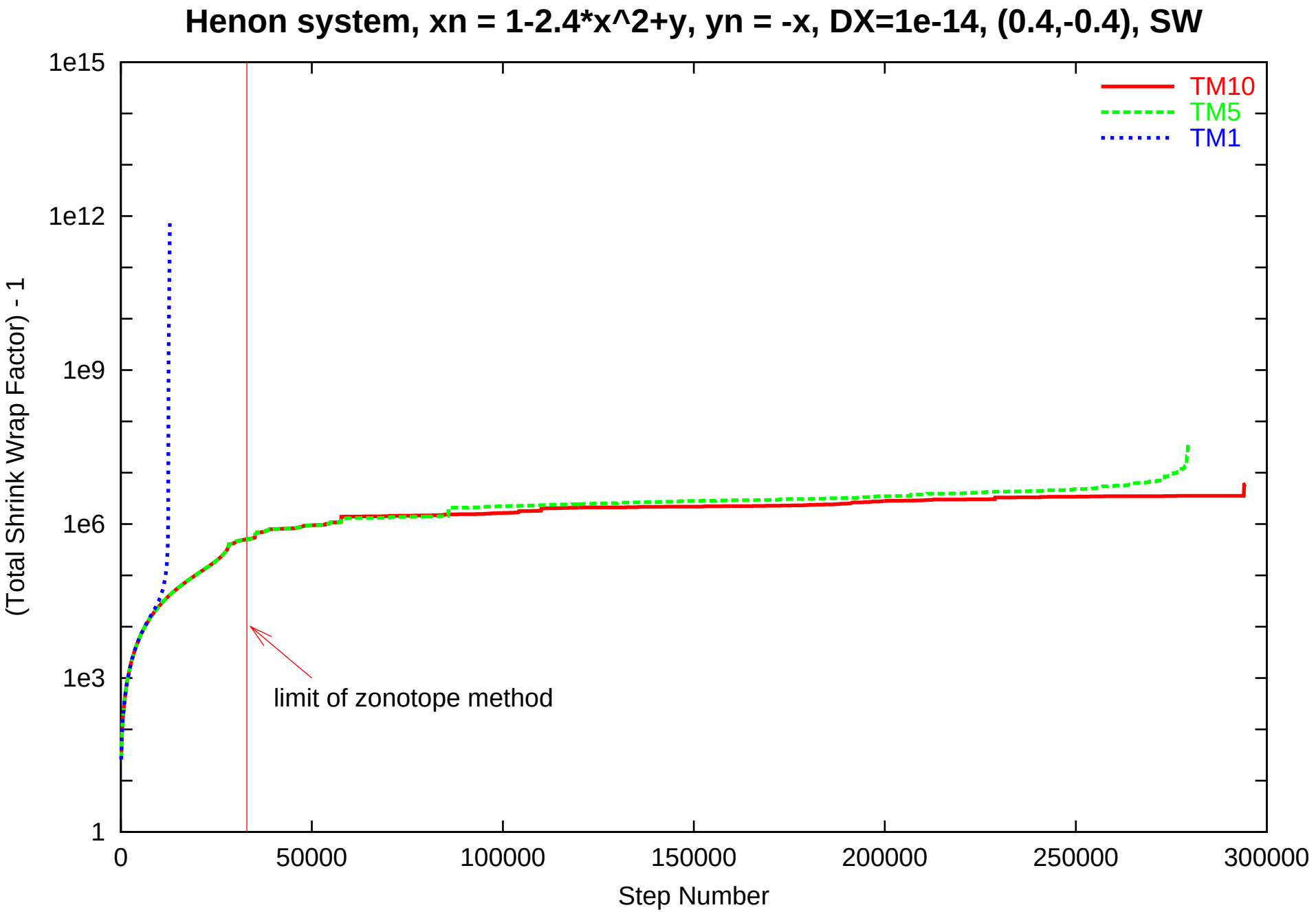

Henon system, xn = 1-2.4*x^2+y, yn = -x, DX=1e-6, NO=1, no-SW
Determinant and Condition Number
Step Number
100
1
0.01
0.0001
1e-006
1e-008
1e-010
1e-012
0
100
200
300
400
500
600
700
800
LR Det
LR CN
L CN
R Det
R CN

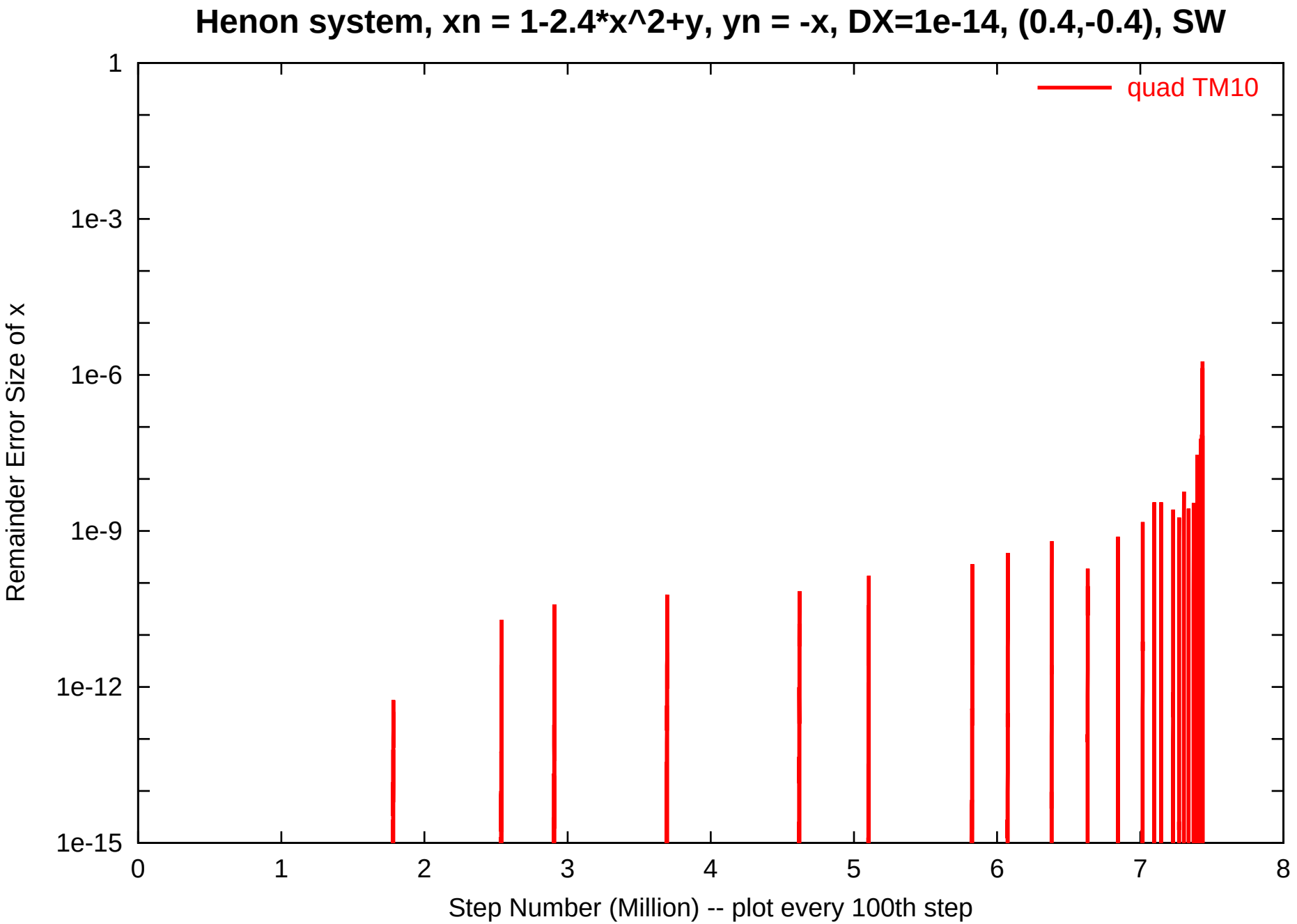

Henon system, xn = 1-2.4*x^2+y, yn = -x, DX=1e-6, NO=1, SW
Determinant and Condition Number
100
1
0.01
0.0001
1e-006
1e-008
1e-010
1e-012
0
100
200
300
400
500
600
700
800
Step Number
LR Det
LR CN
L CN
R Det
R CN

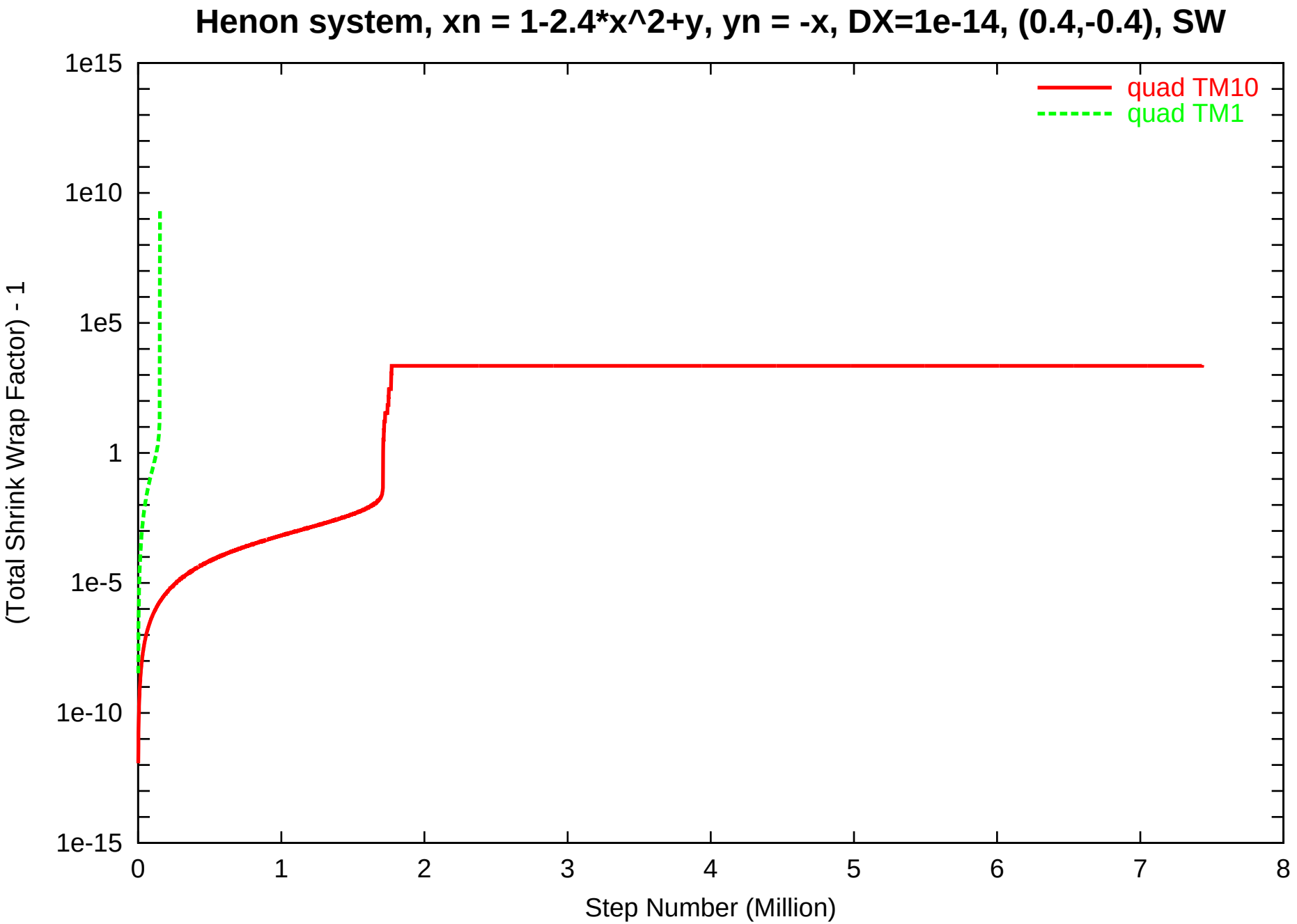

Henon system, xn = 1-2.4*x^2+y, yn = -x, DX=1e-6, NO=5, no-SW
Determinant and Condition Number
Step Number
100
1
0.01
0.0001
1e-006
1e-008
1e-010
1e-012
0
200
400
600
800
1000
1200
1400
1600
LR Det
LR CN
L CN
R Det
R CN


Henon system, xn = 1-2.4*x^2+y, yn = -x, DX=1e-6, NO=1, SW
Determinant and Condition Number
100
1
0.01
0.0001
1e-006
1e-008
1e-010
1e-012
0
200
400
600
800
1000
1200
1400
1600
Step Number
LR Det
LR CN
L CN
R Det
R CN


Henon system, xn = 1-2.4*x^2+y, yn = -x, DX=1e-14, (0.4,-0.4), SW
TM10
TM1
limit of zonotope method
Remainder Error Size of x
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
50000
100000
150000
200000
250000
300000
Step Number -- plot every 100th step


Henon system, xn = 1-2.4*x^2+y, yn = -x, DX=1e-14, (0.4,-0.4), SW
(Total Shrink Wrap Factor) - 1
1e15
1e12
1e9
1e6
1e3
1
0
50000
100000
150000
200000
250000
300000
Step Number
TM10
TM5
TM1
limit of zonotope method


Henon system, xn = 1-2.4*x^2+y, yn = -x, DX=1e-14, (0.4,-0.4), SW
quad TM10
Remainder Error Size of x
1
1e-3
1e-6
1e-9
1e-12
1e-15
0
1
2
3
4
5
6
7
8
Step Number (Million) -- plot every 100th step


Henon system, xn = 1-2.4*x^2+y, yn = -x, DX=1e-14, (0.4,-0.4), SW
quad TM10
quad TM1
(Total Shrink Wrap Factor) - 1
Step Number (Million)
1e15
1e10
1e5
1
1e-5
1e-10
1e-15
0
1
2
3
4
5
6
7
8

Random Matrices - Discrete

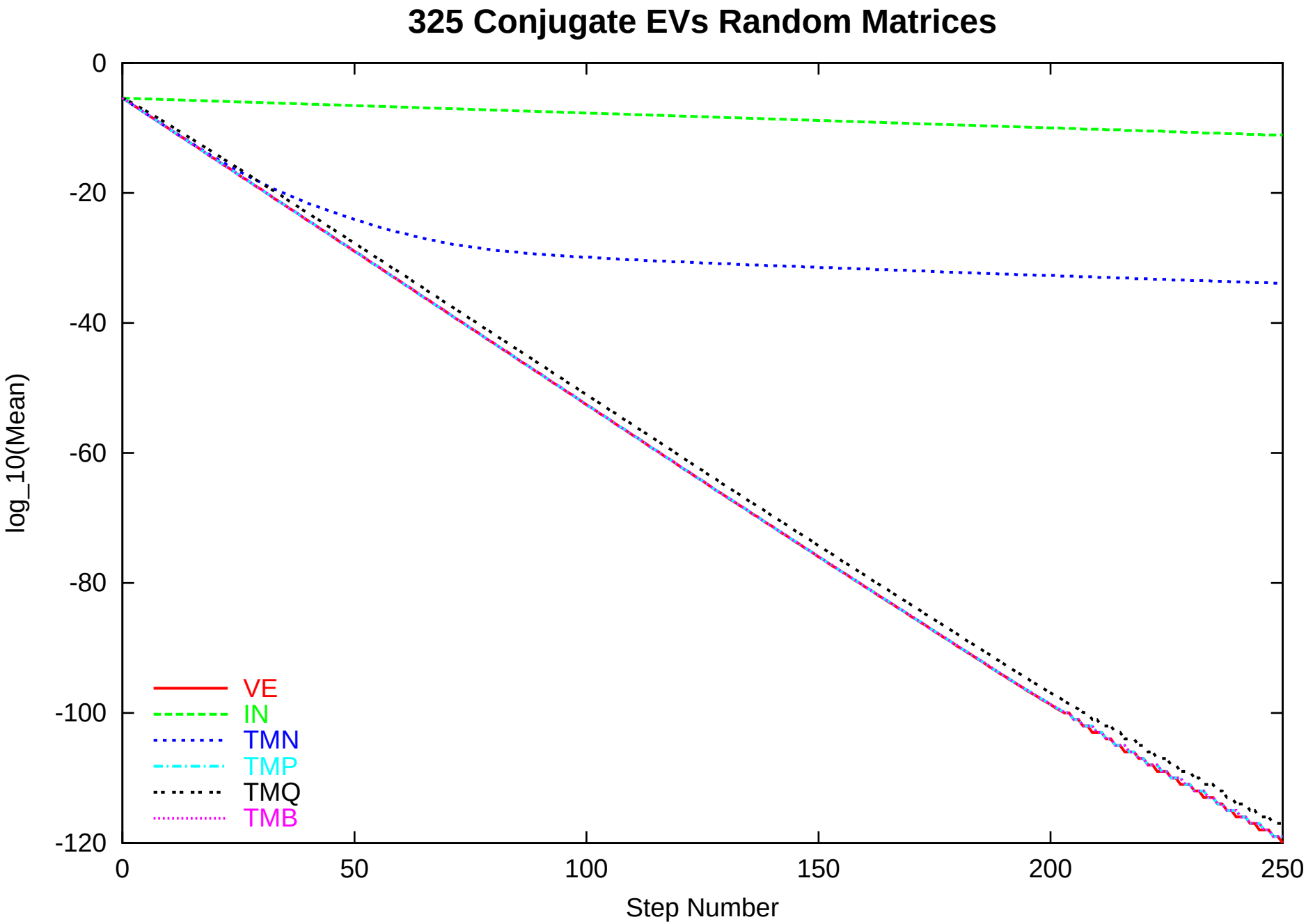
Select 1000 twodimensional random matrices with coefficients in $[-1, 1]$.Sort according to eigenvalues into seven sub-cases.

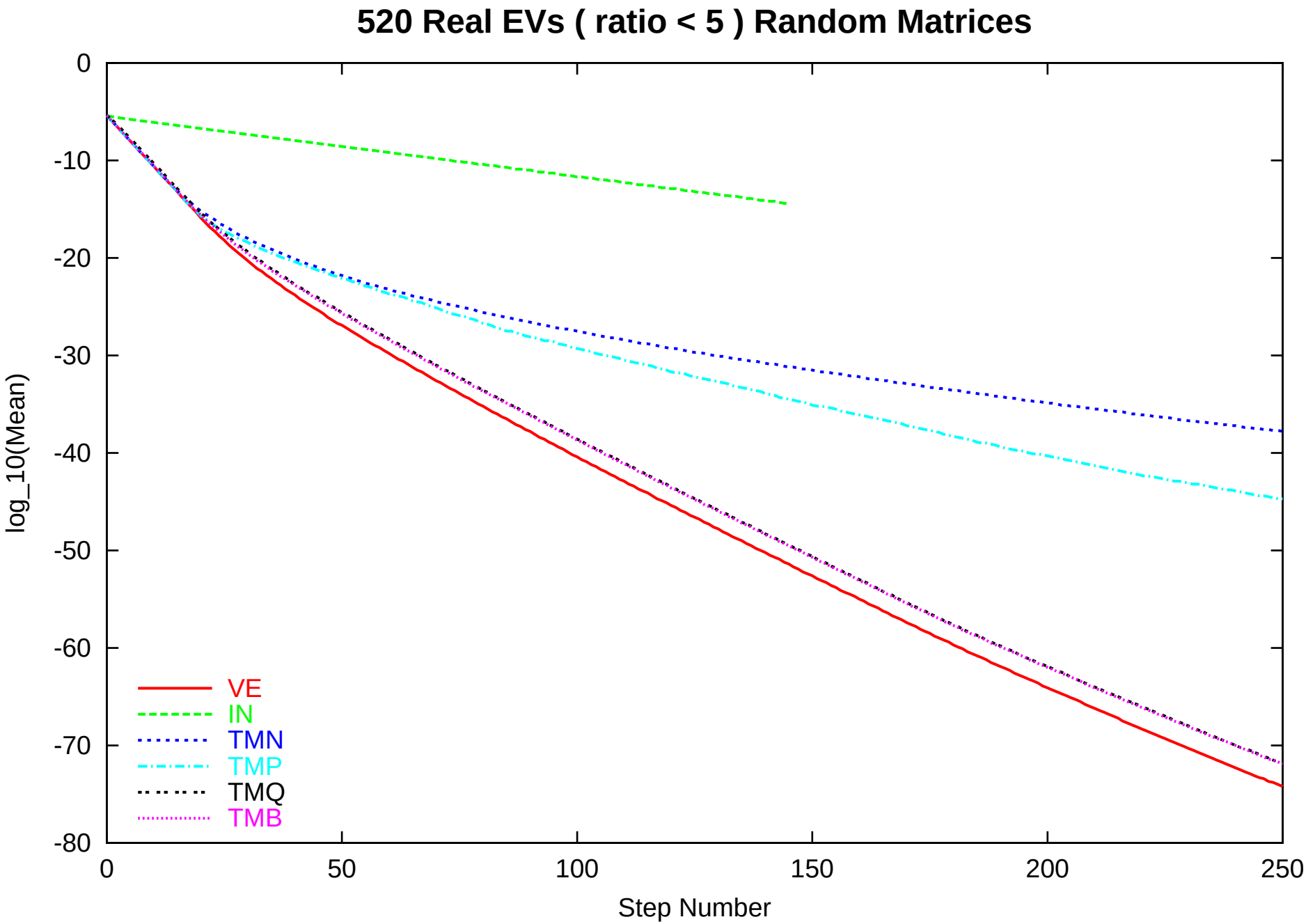
Perform iteration in the following ways:

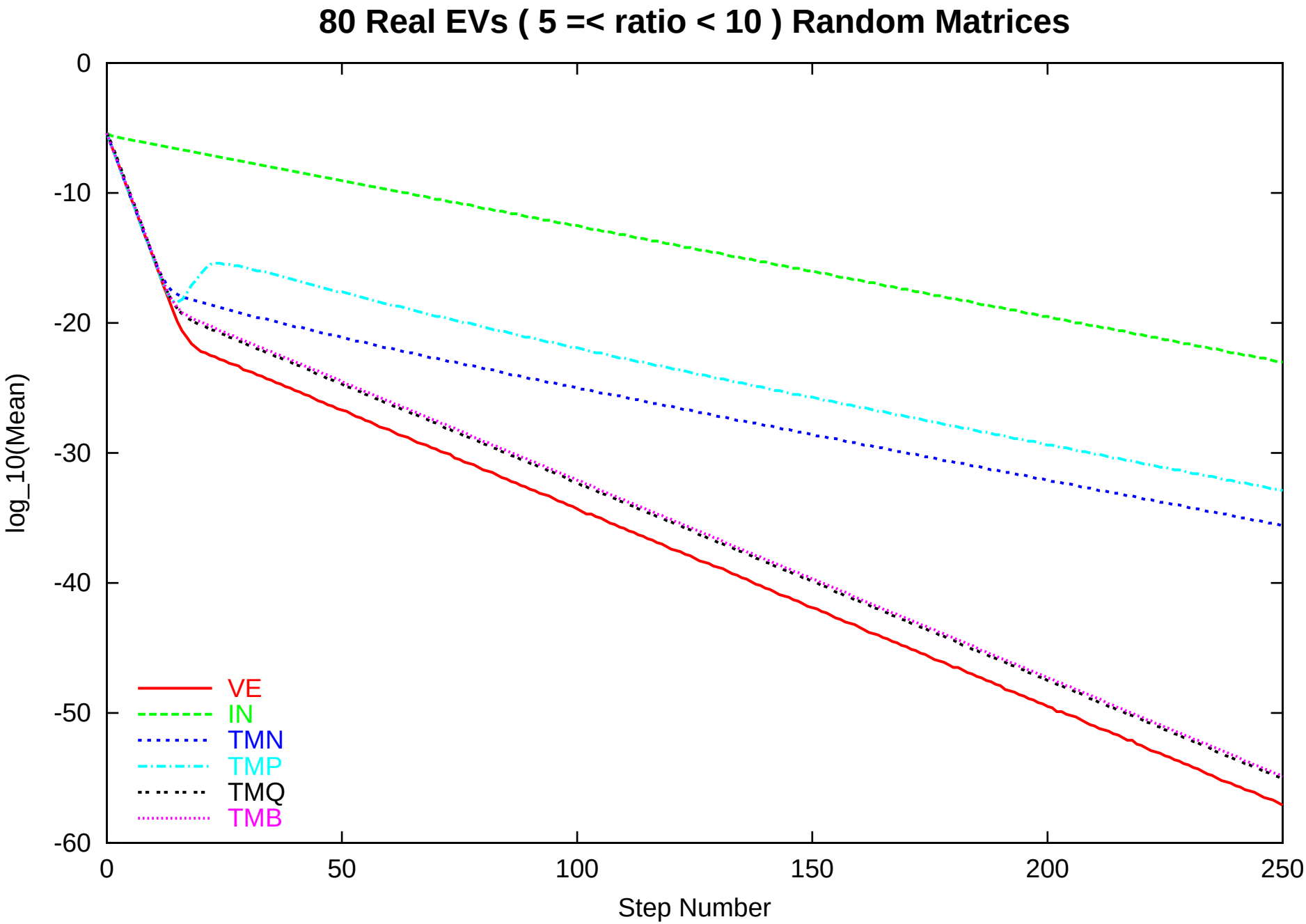
- Naive Interval
- Naive Taylormodel
- Parallelepiped-preconditioned Taylormodel
- QR-preconditioned Taylormodel
- Blunted preconditioned TM, various blunting factors
- Set of four floating point corner points for volume estimation

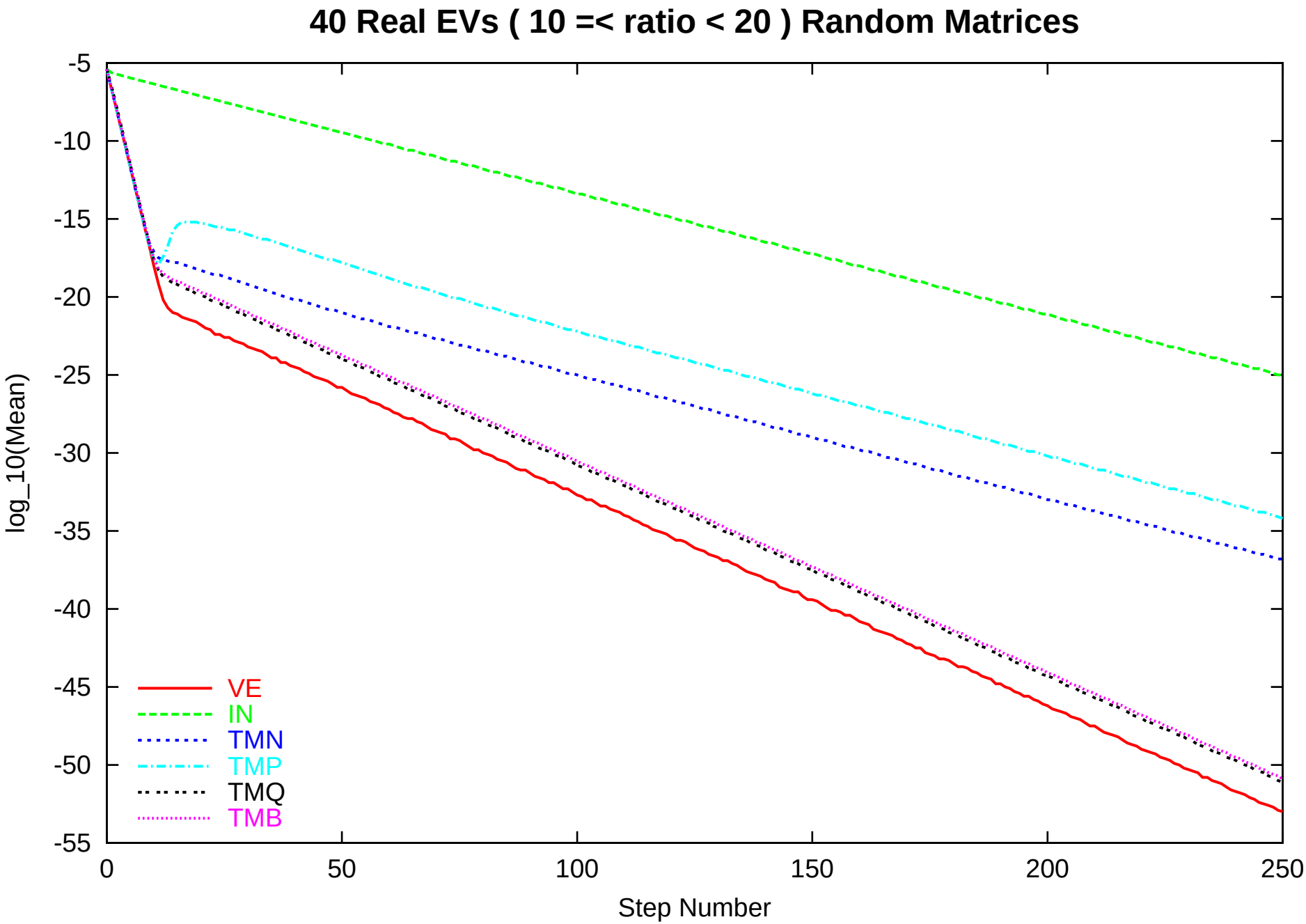
Perform the following tasks:

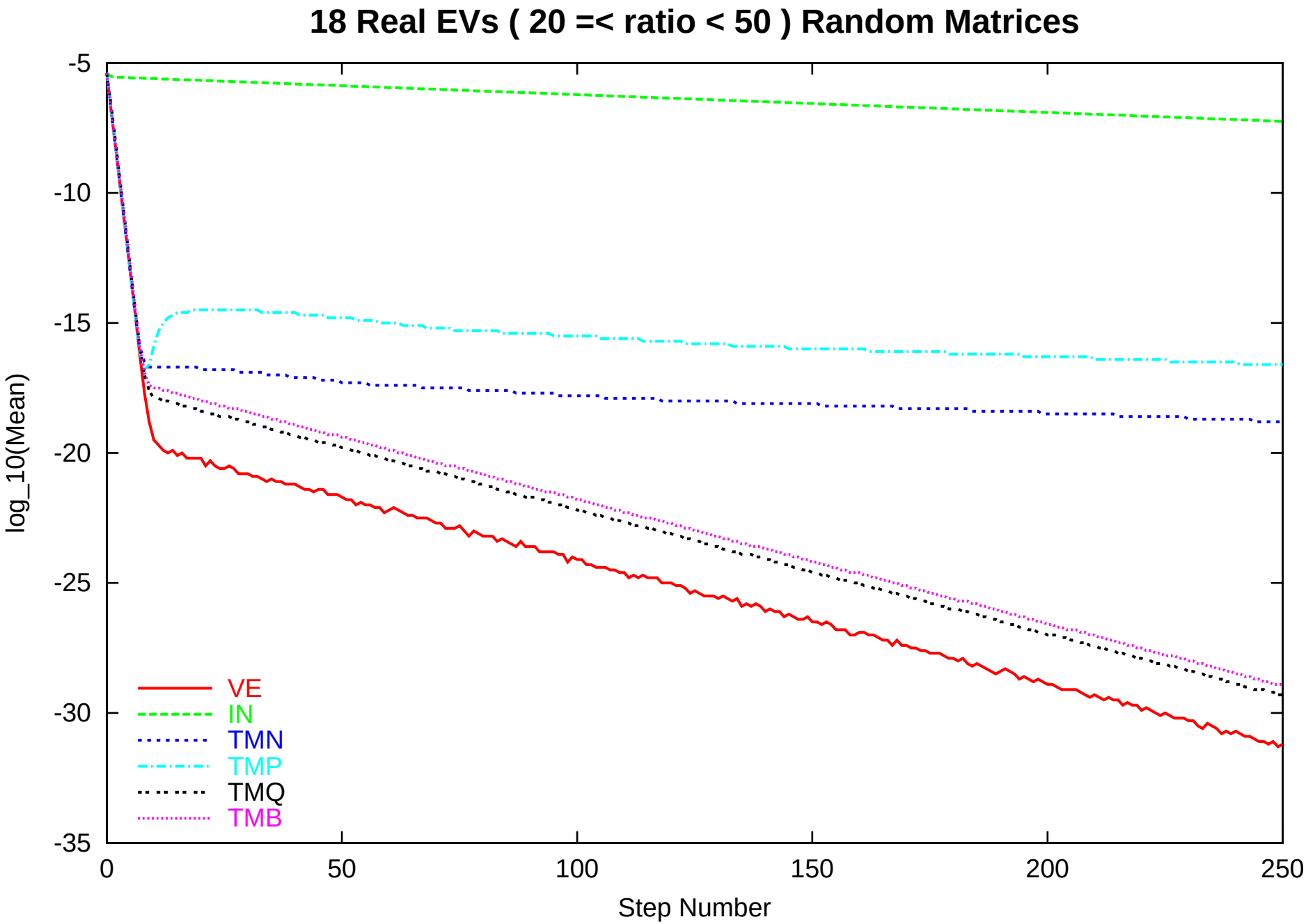
- Iterations through matrix
- Sets of iterations through matrix and its inverse


325 Conjugate EVs Random Matrices
log_10(Mean)
Step Number
0
-20
-40
-60
-80
-100
-120
0
50
100
150
200
250
VE
IN
TMN
TMP
TMQ
TMB


520 Real EVs (ratio < 5) Random Matrices
log_10(Mean)
Step Number
0
-10
-20
-30
-40
-50
-60
-70
-80
0
50
100
150
200
250
VE
IN
TMN
TMP
TMQ
TMB


80 Real EVs (5 =< ratio < 10) Random Matrices
log_10(Mean)
Step Number
0
-10
-20
-30
-40
-50
-60
0
50
100
150
200
250
VE
IN
TMN
TMP
TMQ
TMB


40 Real EVs (10 =< ratio < 20) Random Matrices
log_10(Mean)
Step Number
-5
-10
-15
-20
-25
-30
-35
-40
-45
-50
-55
0
50
100
150
200
250
VE
IN
TMN
TMP
TMQ
TMB


18 Real EVs (20 =< ratio < 50) Random Matrices
log_10(Mean)
Step Number
-5
-10
-15
-20
-25
-30
-35
0
50
100
150
200
250
VE
IN
TMN
TMP
TMQ
TMB

Random Matrices - Discrete

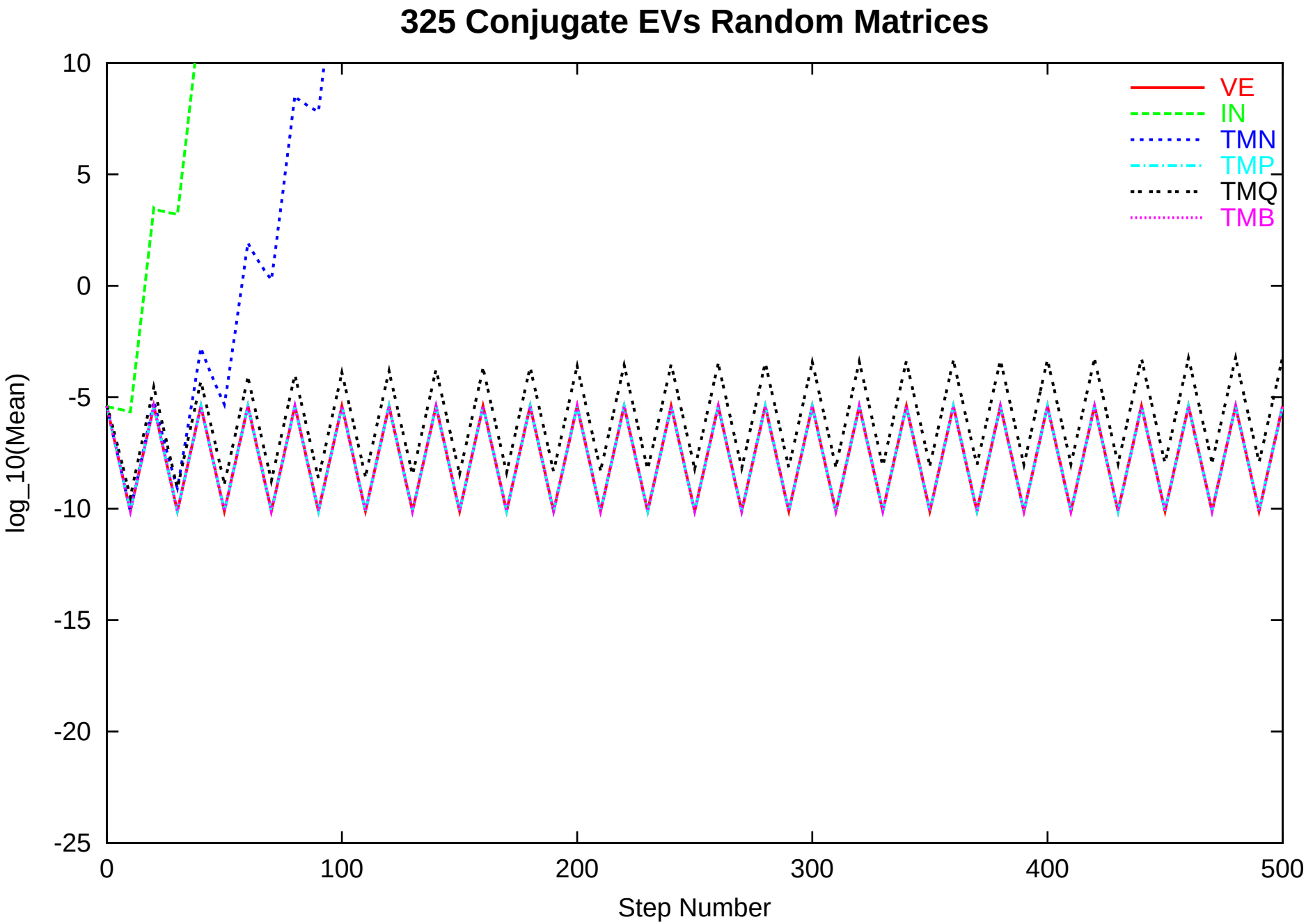
Select 1000 twodimensional random matrices with coefficients in $[-1, 1]$.Sort according to eigenvalues into seven sub-cases.

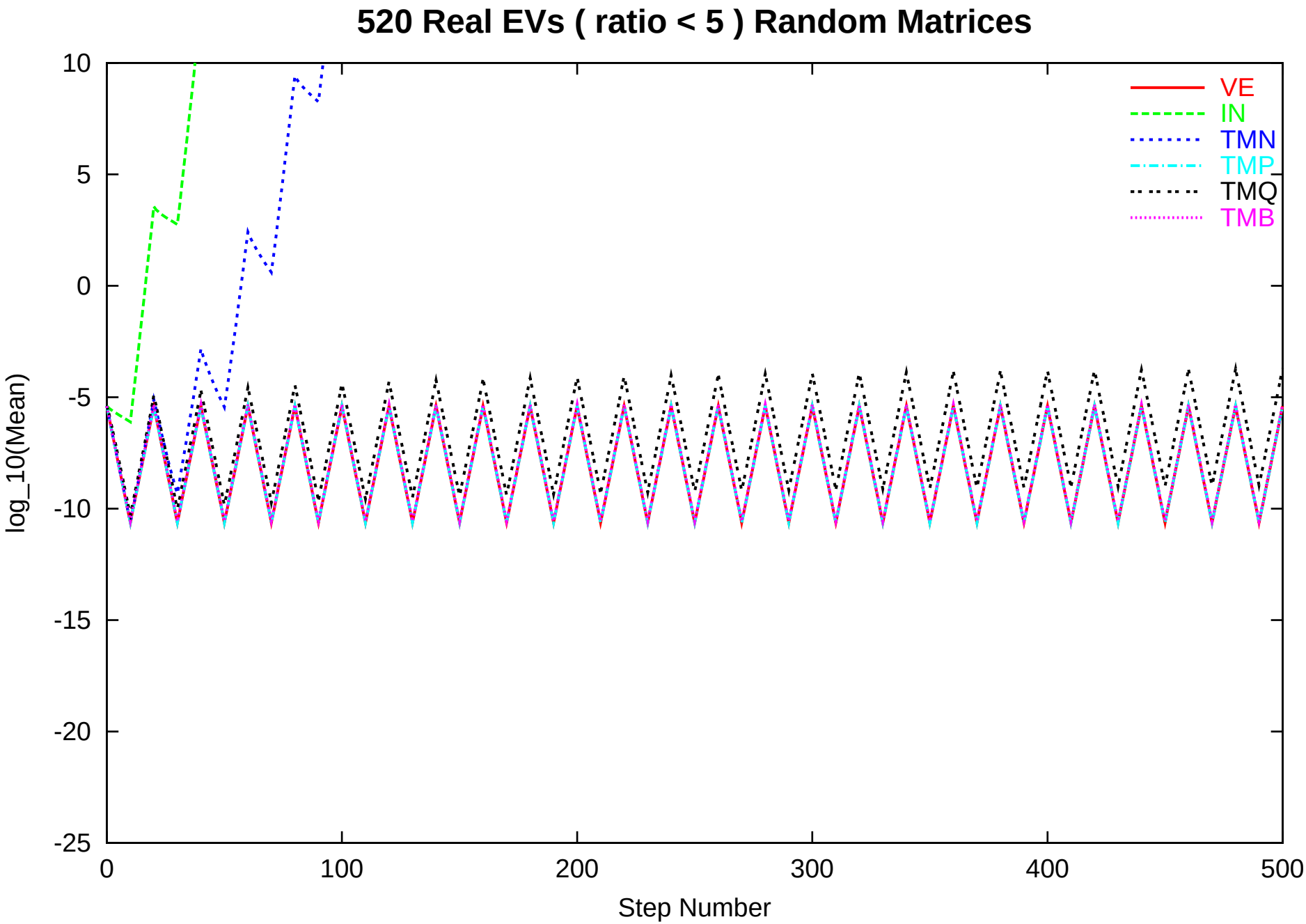
Perform iteration in the following ways:

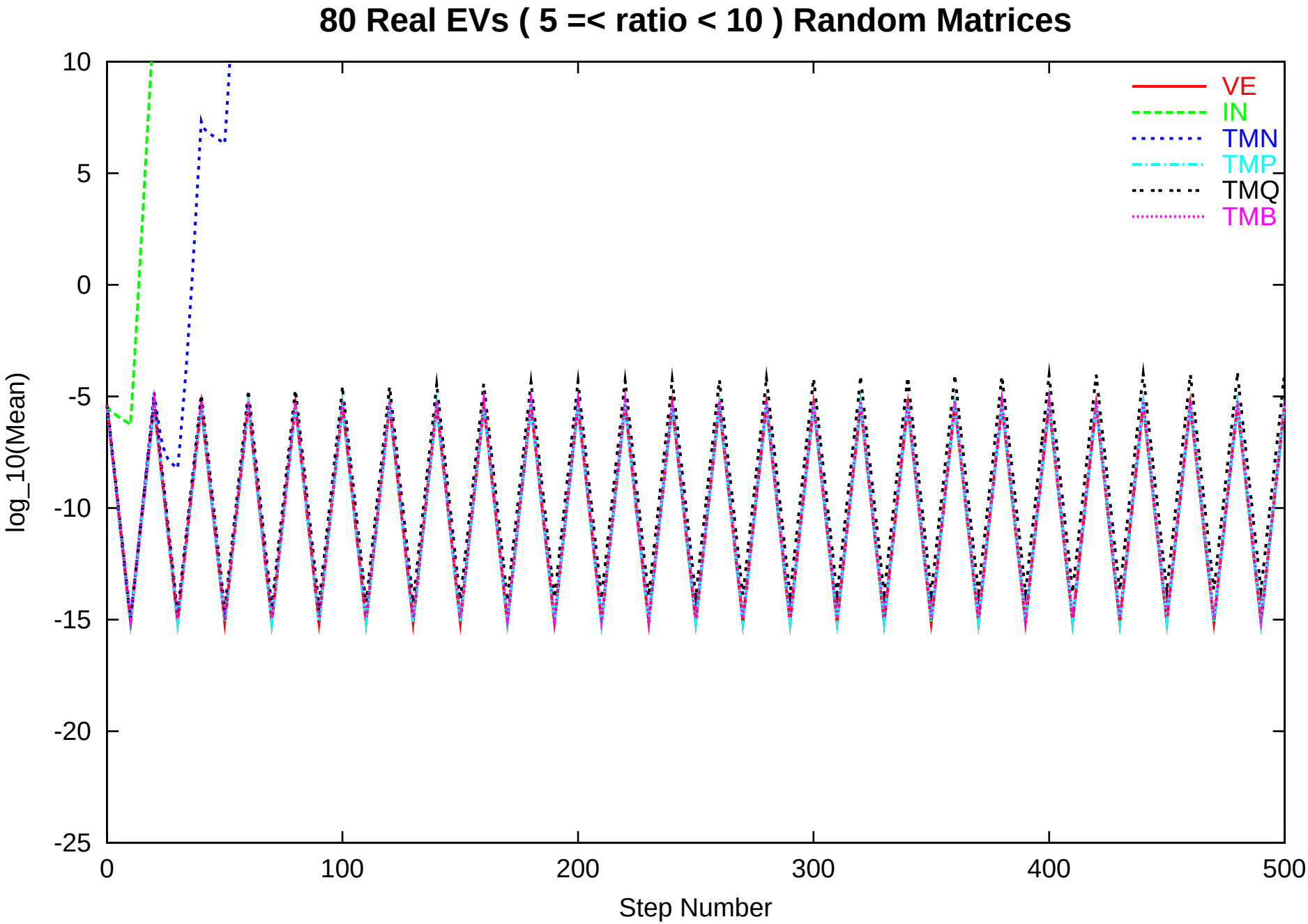
- Naive Interval
- Naive Taylormodel
- Parallelepiped-preconditioned Taylormodel
- QR-preconditioned Taylormodel
- Blunted preconditioned TM, various blunting factors
- Set of four floating point corner points for volume estimation

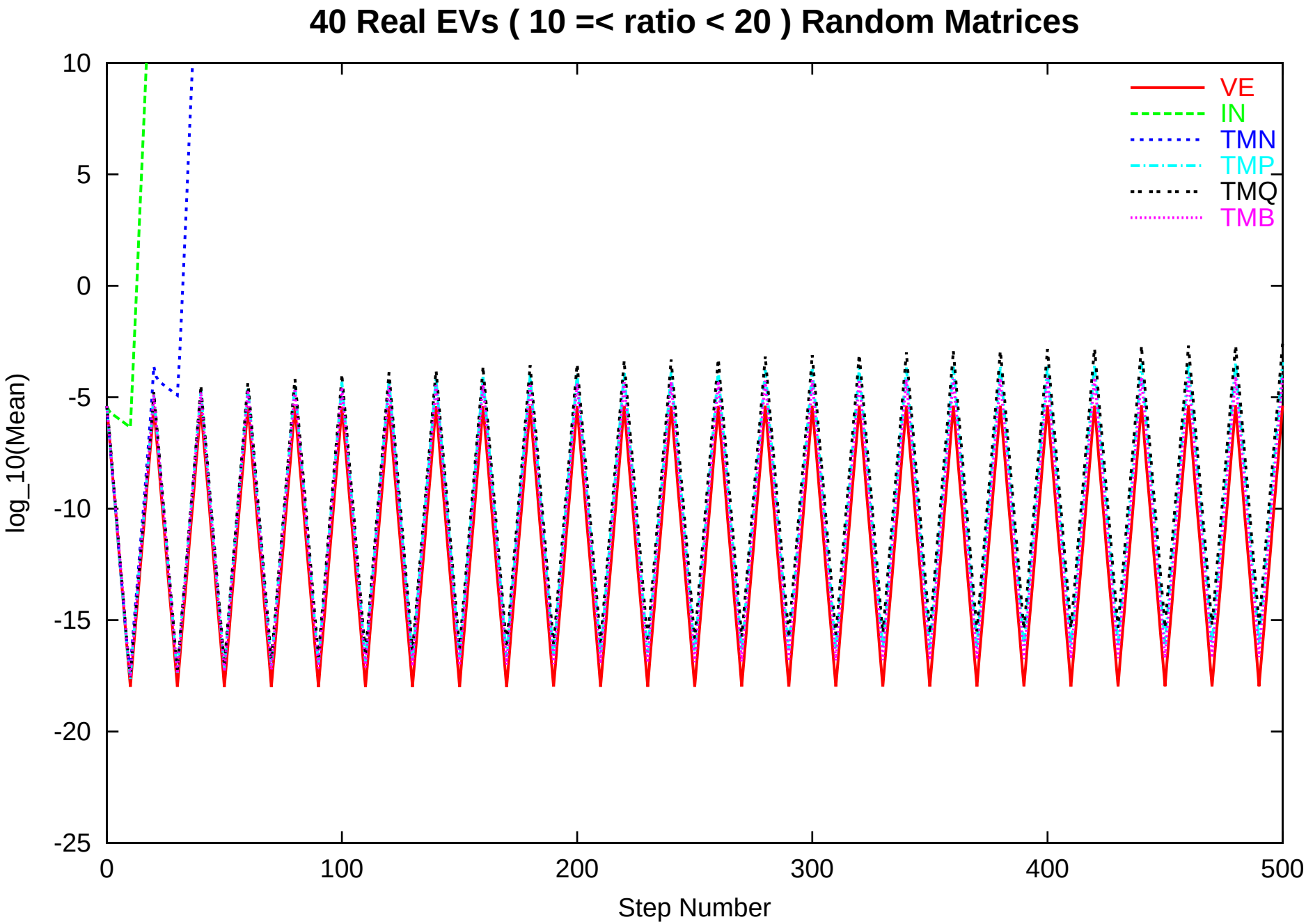
Perform the following tasks:

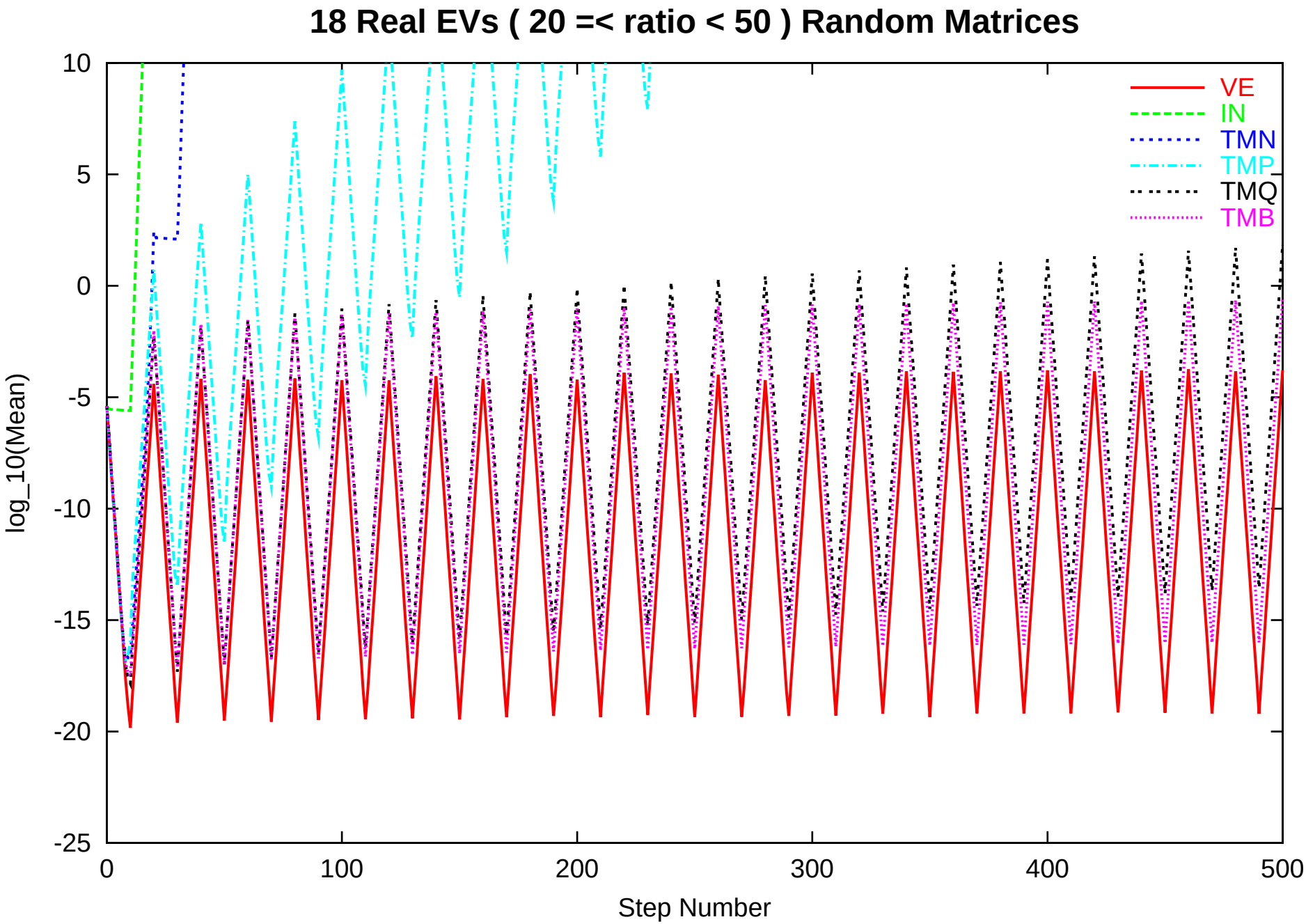
- Iterations through matrix
- Sets of iterations through matrix and its inverse

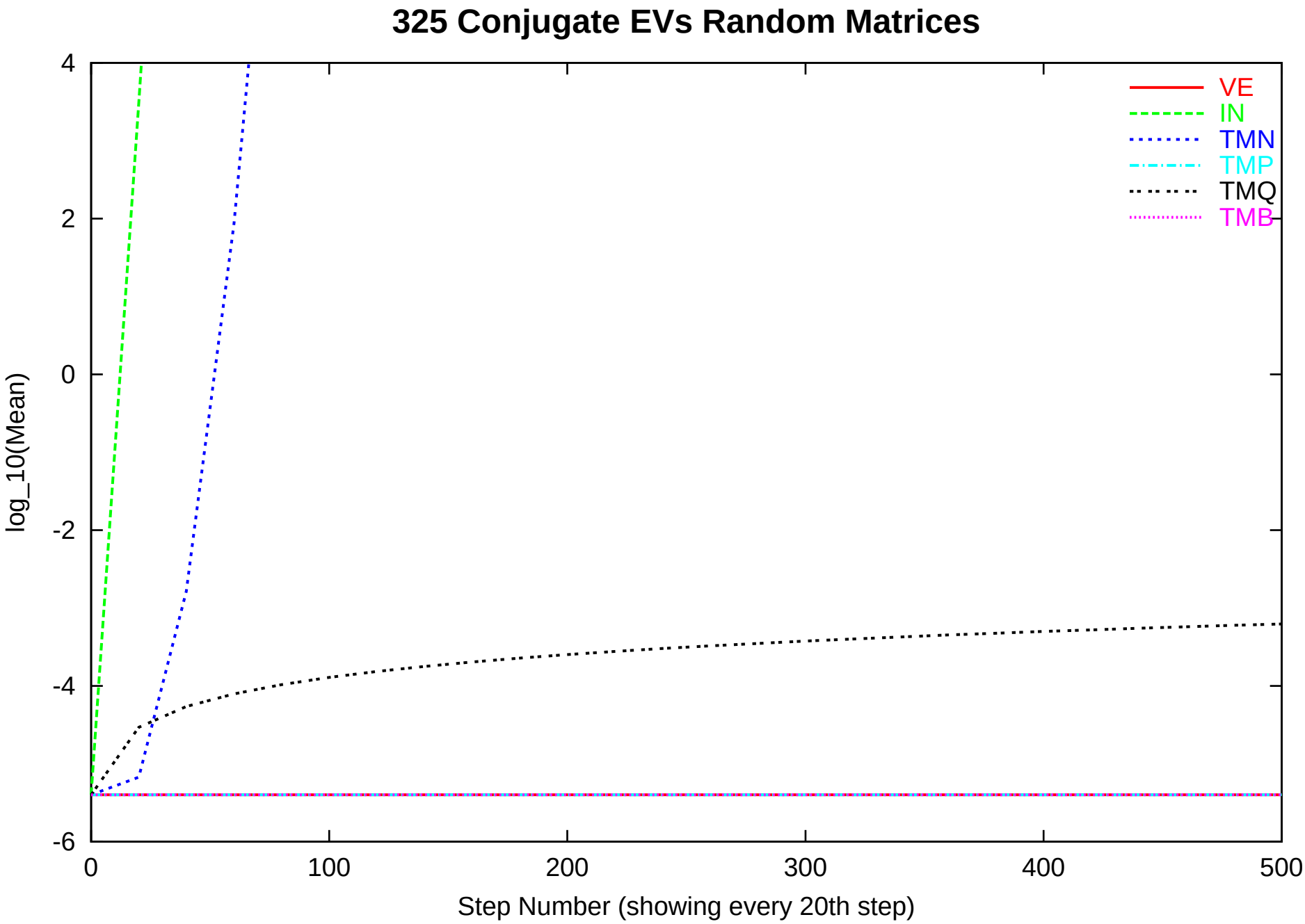

325 Conjugate EVs Random Matrices
log_10(Mean)
Step Number
10
5
0
-5
-10
-15
-20
-25
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB

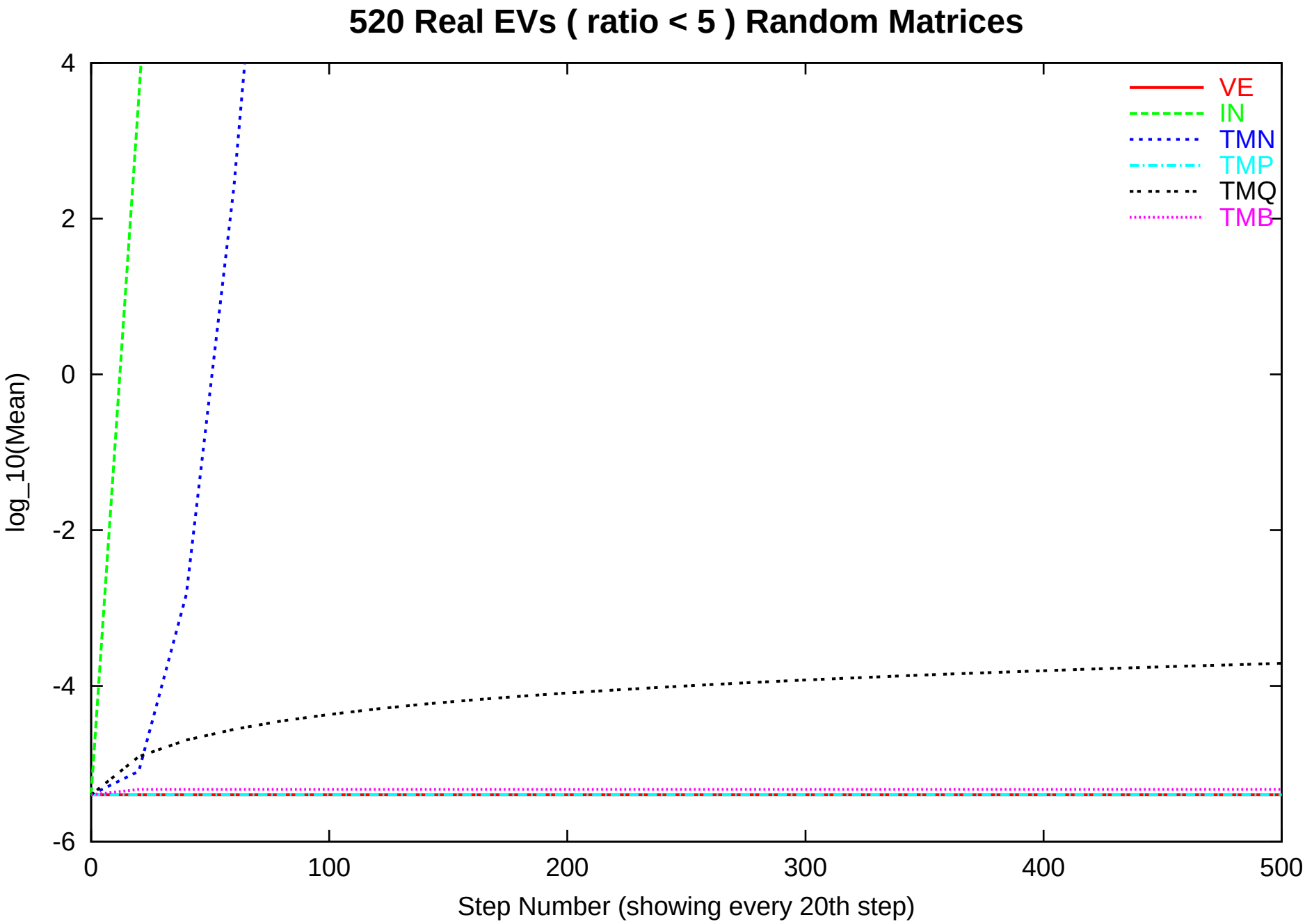

520 Real EVs (ratio < 5) Random Matrices
log_10(Mean)
Step Number
10
5
0
-5
-10
-15
-20
-25
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB

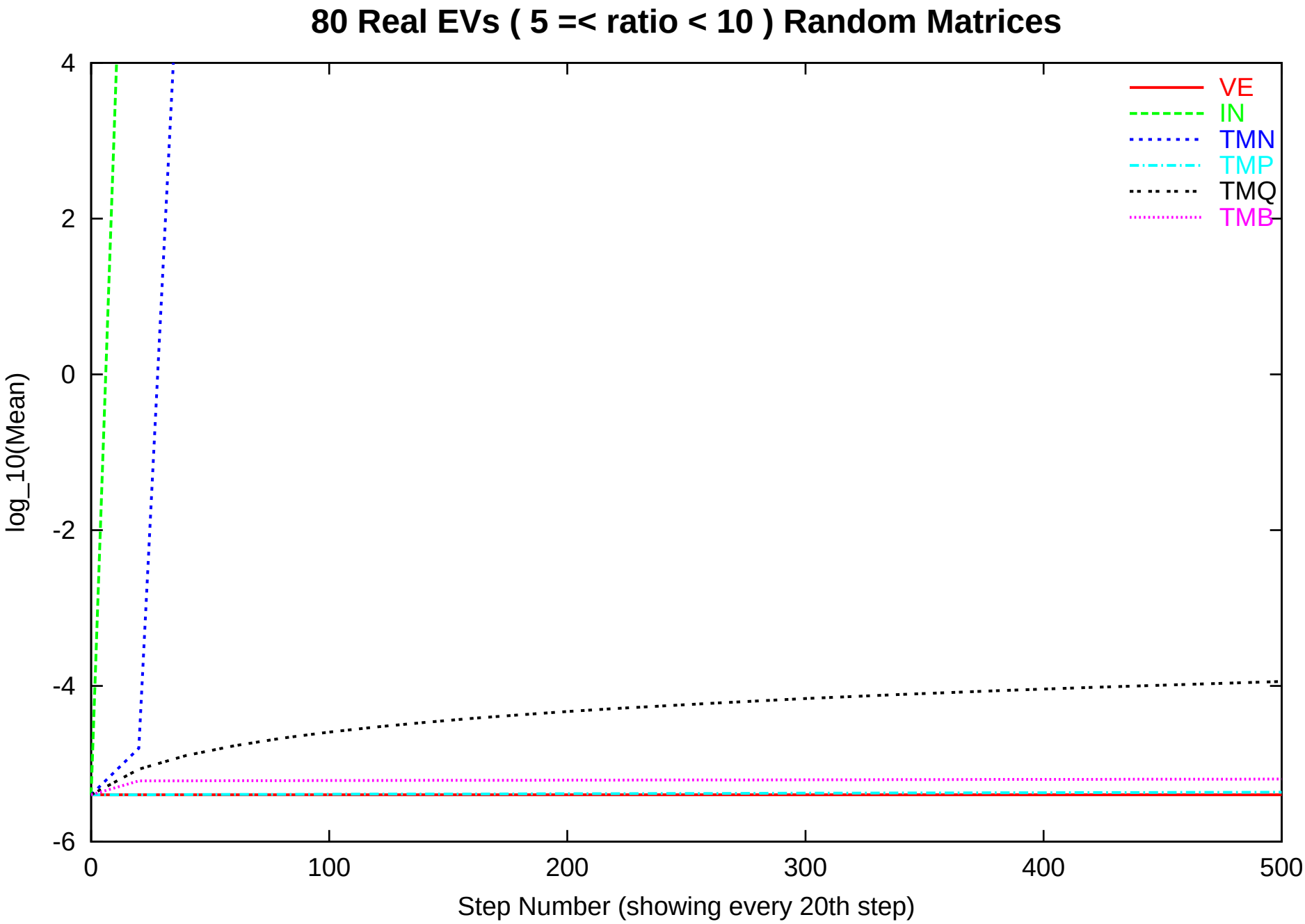

80 Real EVs (5 =< ratio < 10) Random Matrices
log_10(Mean)
Step Number
10
5
0
-5
-10
-15
-20
-25
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB

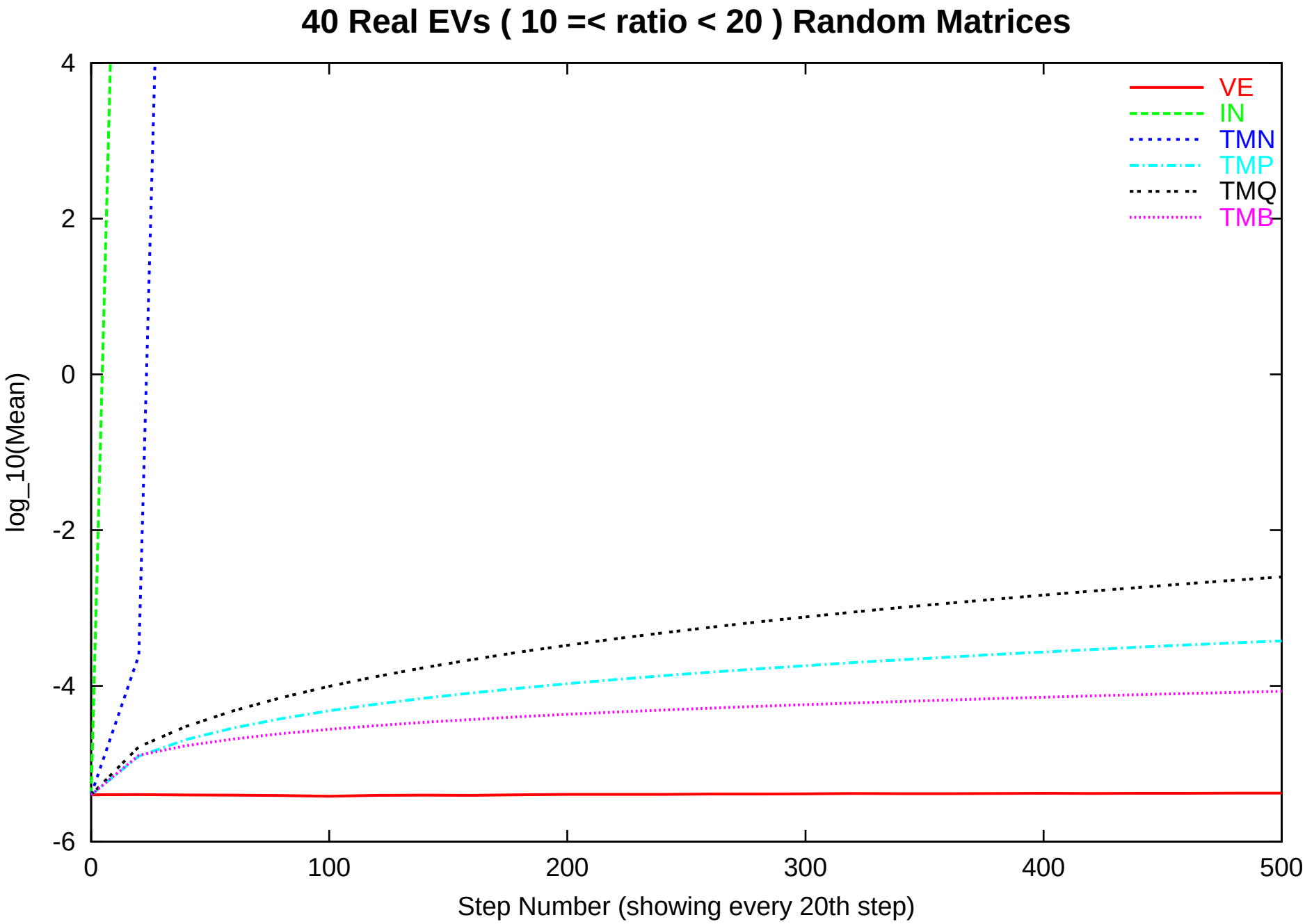

40 Real EVs (10 =< ratio < 20) Random Matrices
log_10(Mean)
Step Number
10
5
0
-5
-10
-15
-20
-25
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB

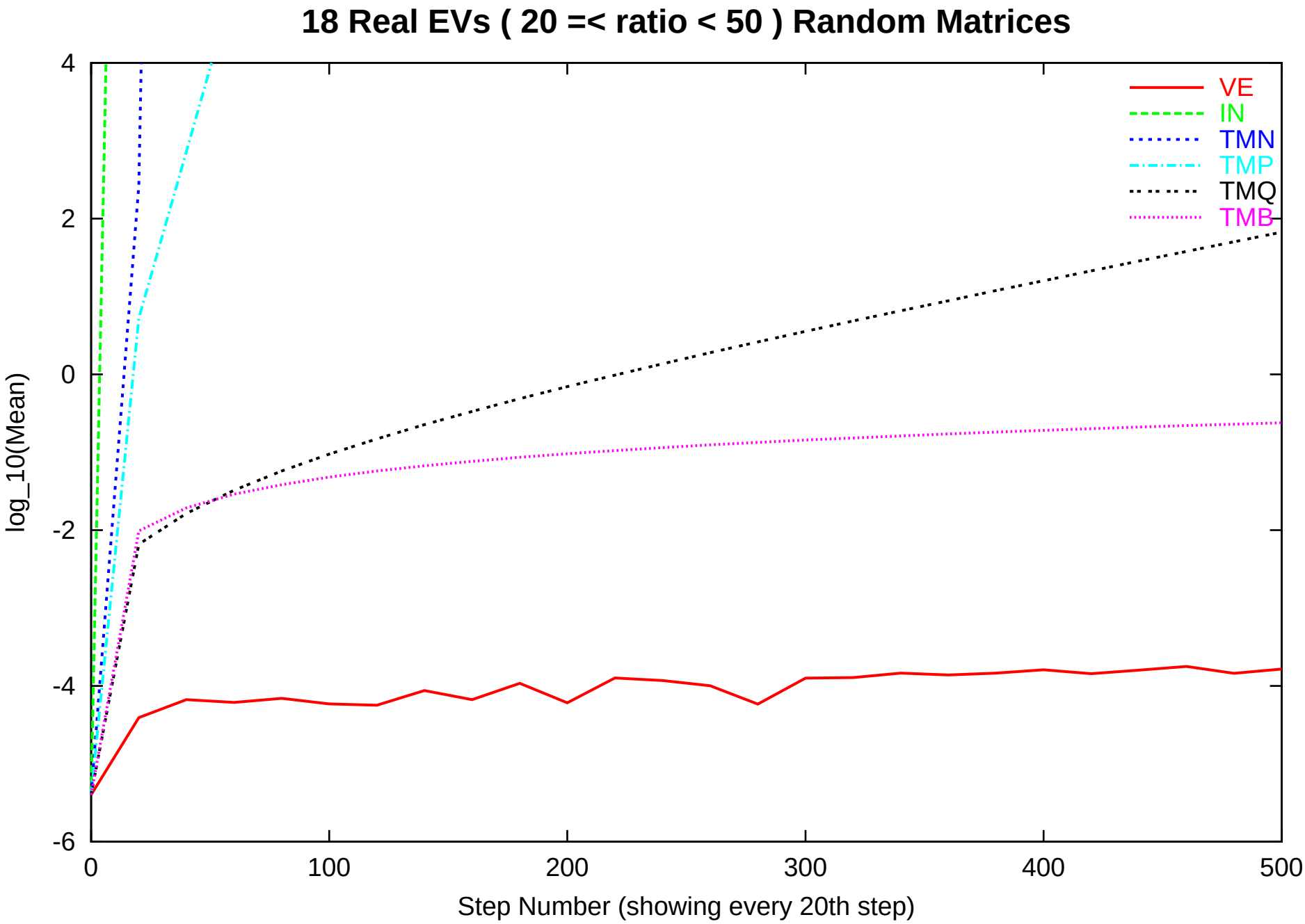

18 Real EVs (20 =< ratio < 50) Random Matrices
log_10(Mean)
Step Number
10
5
0
-5
-10
-15
-20
-25
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB

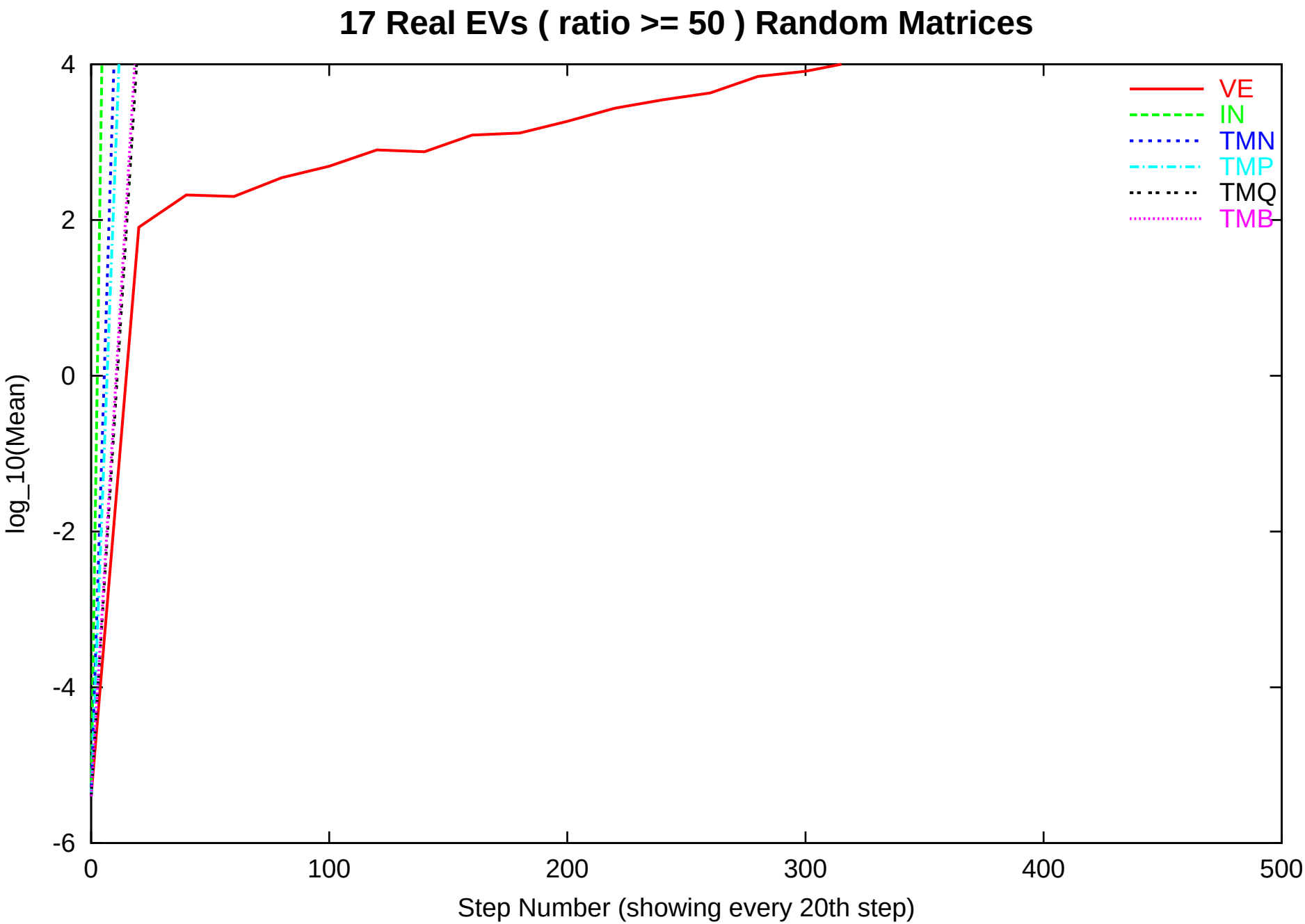

325 Conjugate EVs Random Matrices
log_10(Mean)
Step Number (showing every 20th step)
4
2
0
-2
-4
-6
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB

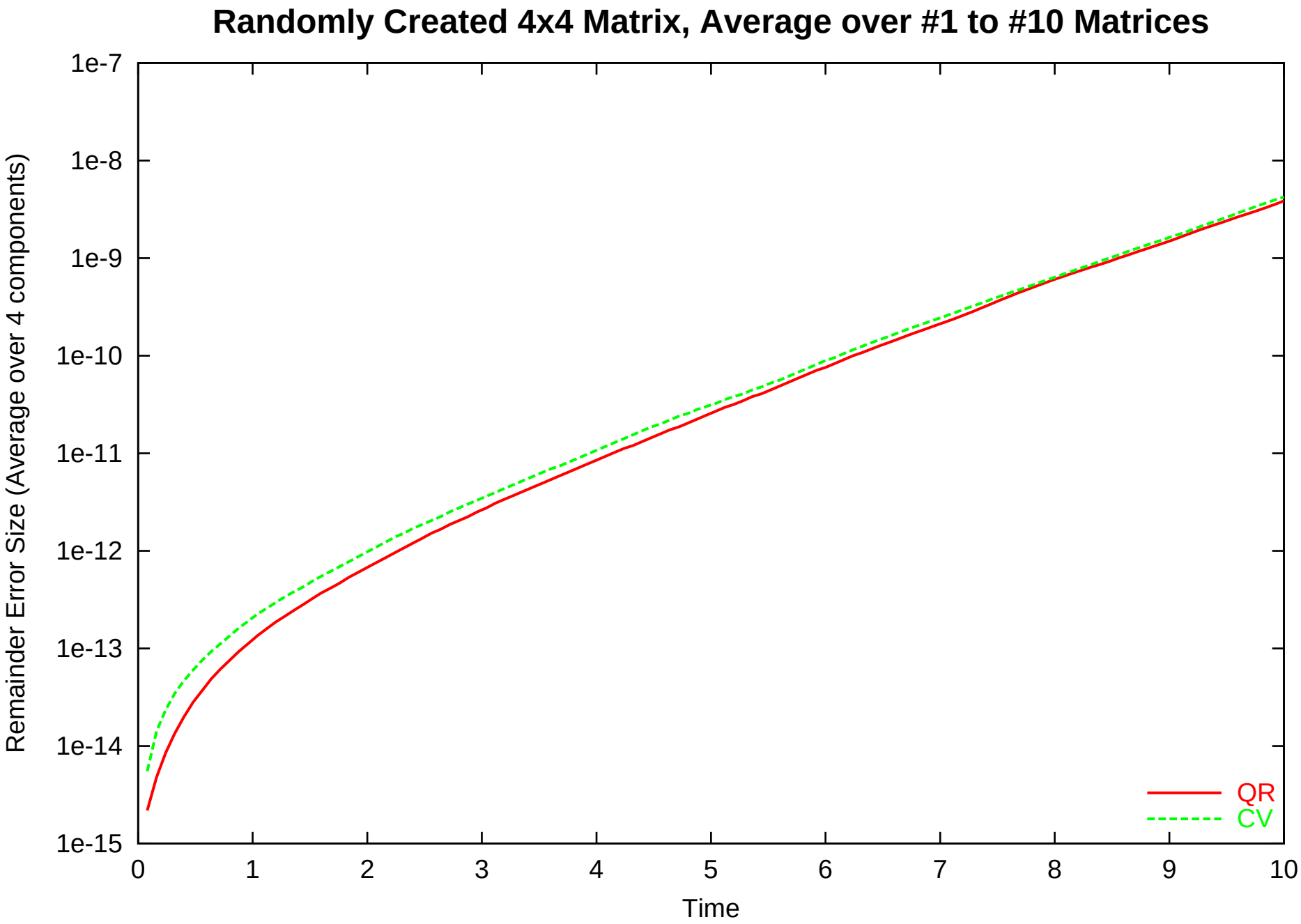

520 Real EVs (ratio < 5) Random Matrices
log_10(Mean)
Step Number (showing every 20th step)
4
2
0
-2
-4
-6
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB


80 Real EVs (5 =< ratio < 10) Random Matrices
log_10(Mean)
Step Number (showing every 20th step)
4
2
0
-2
-4
-6
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB


40 Real EVs (10 =< ratio < 20) Random Matrices
log_10(Mean)
Step Number (showing every 20th step)
4
2
0
-2
-4
-6
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB


18 Real EVs (20 =< ratio < 50) Random Matrices
log_10(Mean)
Step Number (showing every 20th step)
4
2
0
-2
-4
-6
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB


17 Real EVs (ratio >= 50) Random Matrices
log_10(Mean)
Step Number (showing every 20th step)
4
2
0
-2
-4
-6
0
100
200
300
400
500
VE
IN
TMN
TMP
TMQ
TMB


Randomly Created 4x4 Matrix, Average over #1 to #10 Matrices
Remainder Error Size (Average over 4 components)
1e-7
1e-8
1e-9
1e-10
1e-11
1e-12
1e-13
1e-14
1e-15
0
1
2
3
4
5
6
7
8
9
10
Time
QR
CV

Example Matrix - Continuous

Consider exampe random matrix

$$A_1 = \begin{pmatrix} +0.9564 & +0.2004 & +0.4826 & +0.8871 \\ -0.4922 & +0.5651 & -0.1474 & -0.7678 \\ -0.0269 & -0.8587 & -0.3785 & -0.6168 \\ -0.8271 & +0.2661 & -0.9380 & +0.5289 \end{pmatrix}$$

Approximate eigenvalues 0.3928, -0.3911, $1.005 \pm 0.8669i$. Center point of the initial domain box $(0.6446, 0, 0050, -0.2394, 0.4581)$, width 10^{-3}.

- Exponential rise from 10^{-11} at $t = 3$ to 10^{-7} at $t = 10$, corresponds to $10^{4/7} \approx 10^{.5715}$ per time unit
- Magnitude of complex eigenvalues is approximately 1.327, leading to $\exp(1.327) \approx 3.769 \approx 10^{0.5763}$ per time unit.
- Very close agreement between growth of error and growth of true solution
- Asymptotics same as with good non-validated integrator

Random Matrices - Continuous

Select 10 fourdimensional random matrices A with coefficients in $[-1, 1]$. Solve ODE

$$\frac{d}{dt}r = A \cdot r$$

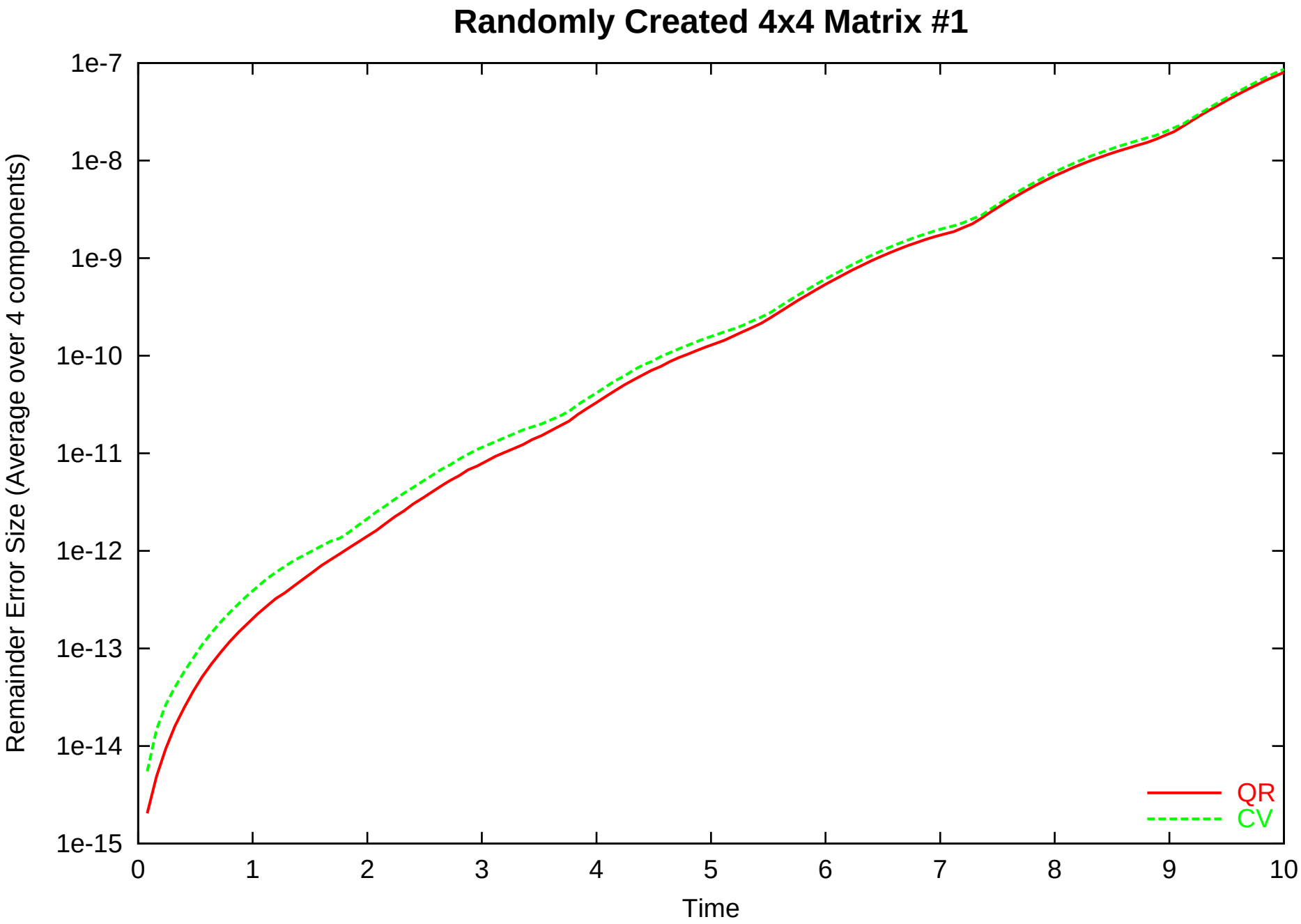
with random initial conditions.

Perform integration in the following ways:

- Curvilinear-Preconditioned Taylormodel
- QR-Preconditioned Taylormodel

Observe that

- CV and QR preconditiong have the same asymptotic behavior
- Both lead to error growth agreeing with growth along longest EV up to 1%.
- Thus, same error growth as in non-validated case.


Randomly Created 4x4 Matrix #1
Remainder Error Size (Average over 4 components)
1e-7
1e-8
1e-9
1e-10
1e-11
1e-12
1e-13
1e-14
1e-15
0
1
2
3
4
5
6
7
8
9
10
Time
QR
CV

Conclusions - Linear Problems

- In case of distinct real eigenvalues, **Curvilinear (CV)** and QR coordinate systems converge to same limit
- Thus, CV and QR preconditioning has **same asymptotic behavior**
- The asymptotic behavior is essentially that of a good non-validated integrator (Nedialkov - Jackson)
- For complex eigenvalues, CV and QR both lead to rotations, and are thus expected to behave the same

- **Blunting** method leaves eigenvectors with largest eigenvalues unchanged
- Longest direction(s) of blunted parallelepiped are **not affected**
- Only asymptotically non-dominating directions are affected
- Blunted method has essentially **same asymptotic behavior** as non-validated integrator

Collection of Random Matrices

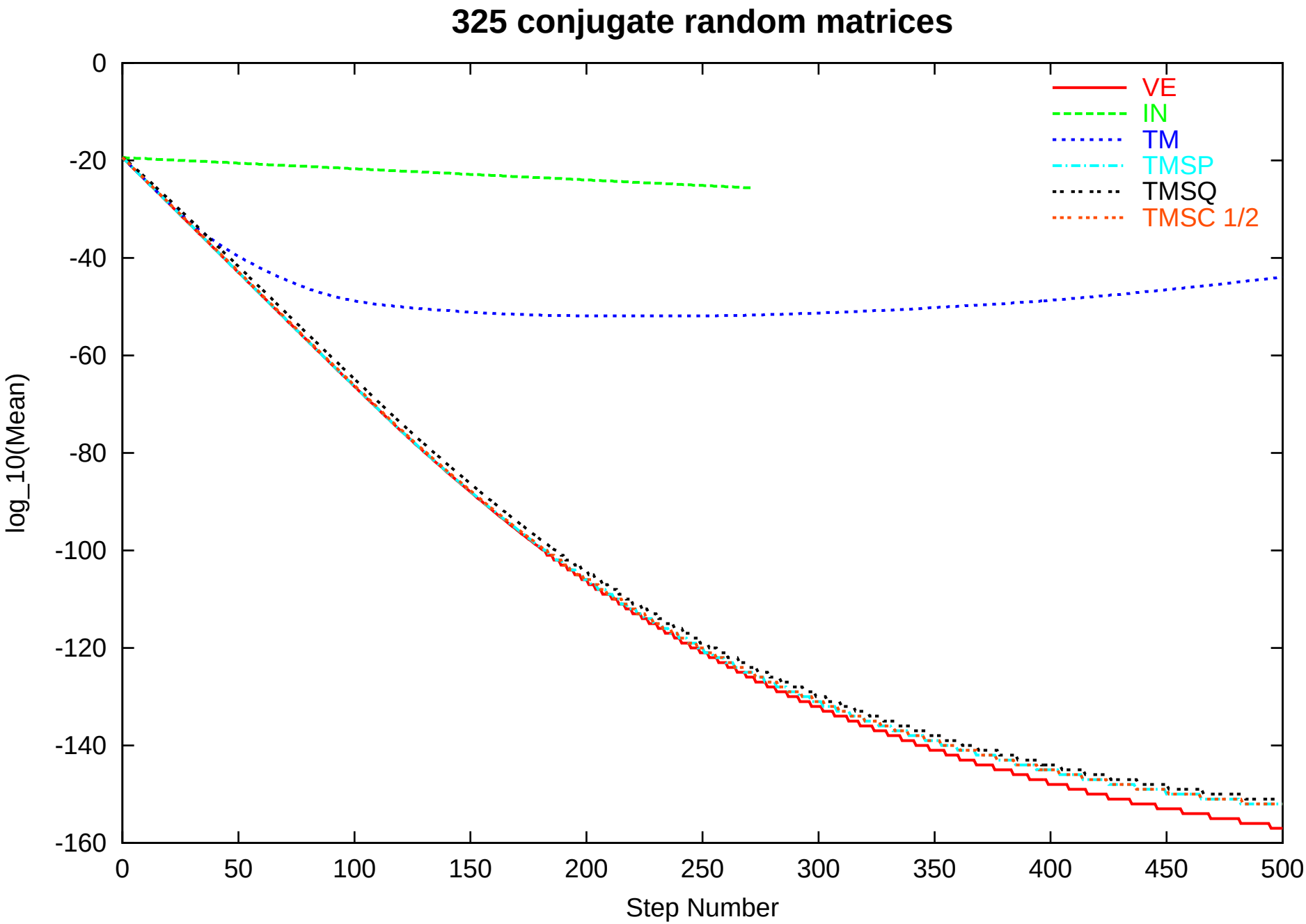
Select 1000 twodimensional random matrices with coefficients in $[-1, 1]$.Sort according to eigenvalues into seven sub-cases.

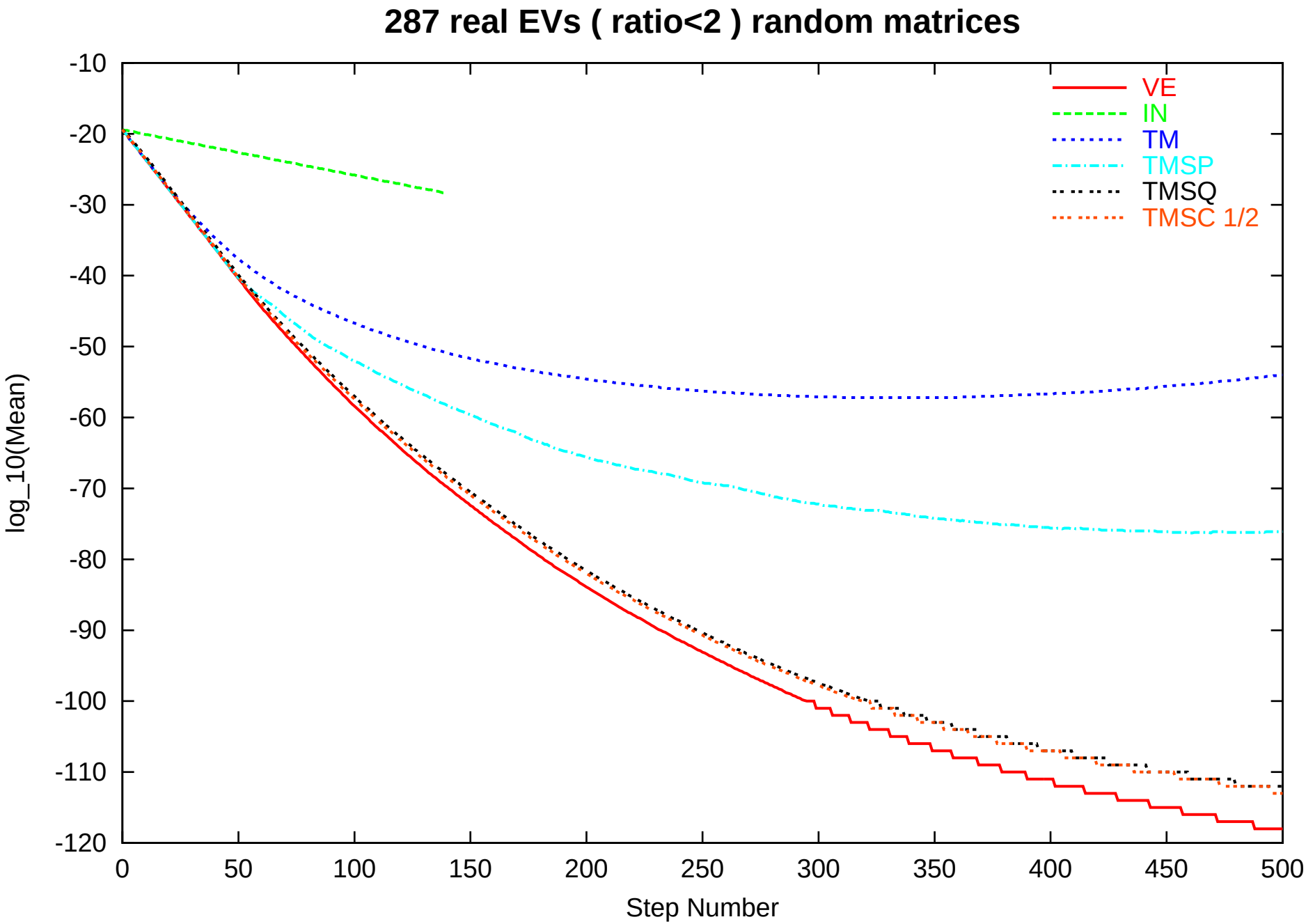
Perform iteration in the following ways:

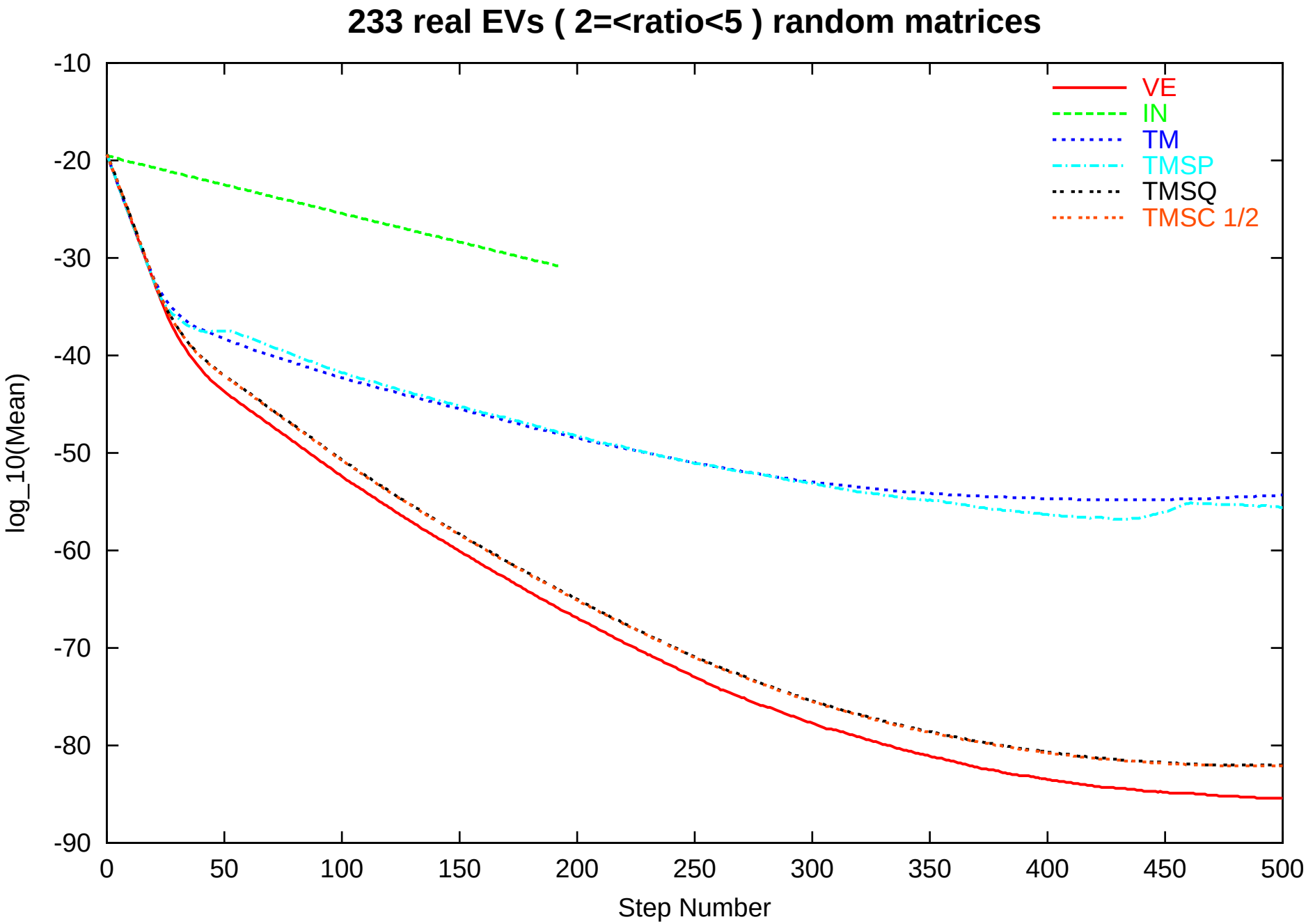
- Naive Interval
- Naive Taylormodel
- Parallelepiped-preconditioned Taylormodel
- QR-preconditioned Taylormodel
- Blunted QR-preconditioned TM, various blunting factors
- Set of four floating point corner points for volume estimation

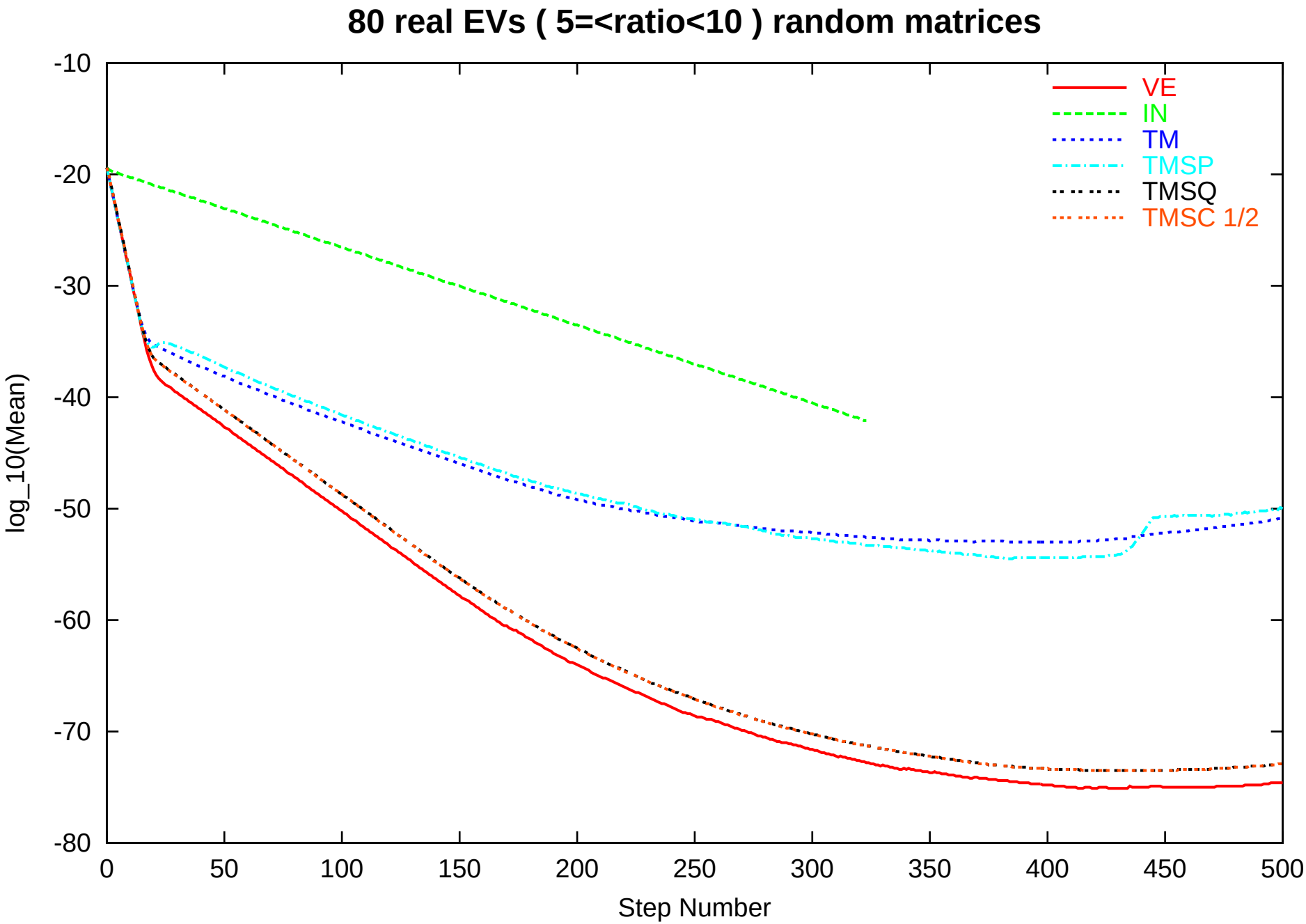
Perform the following tasks:

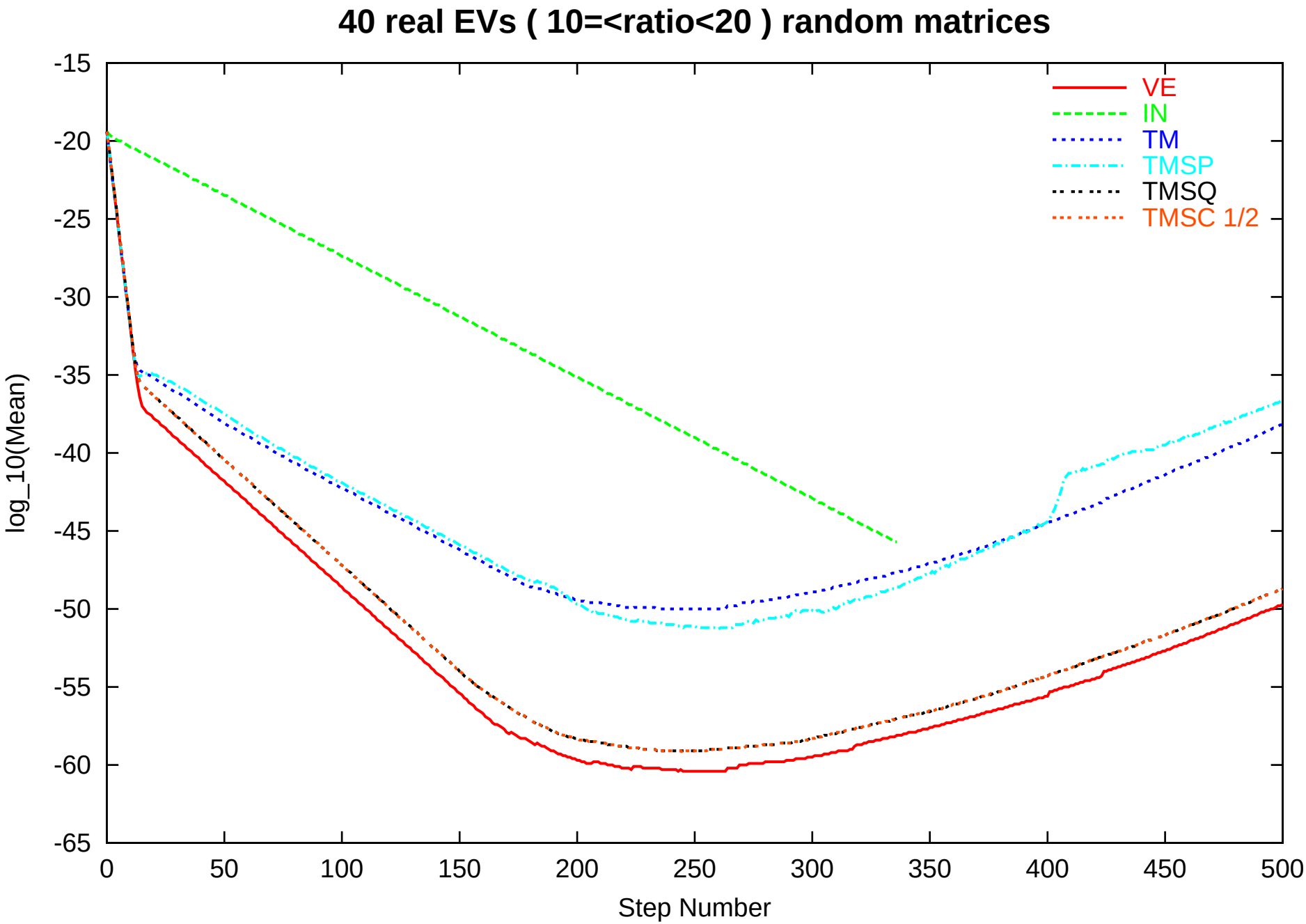
- 500 iterations through matrix
- 25 sets of iterations through matrix and its inverse

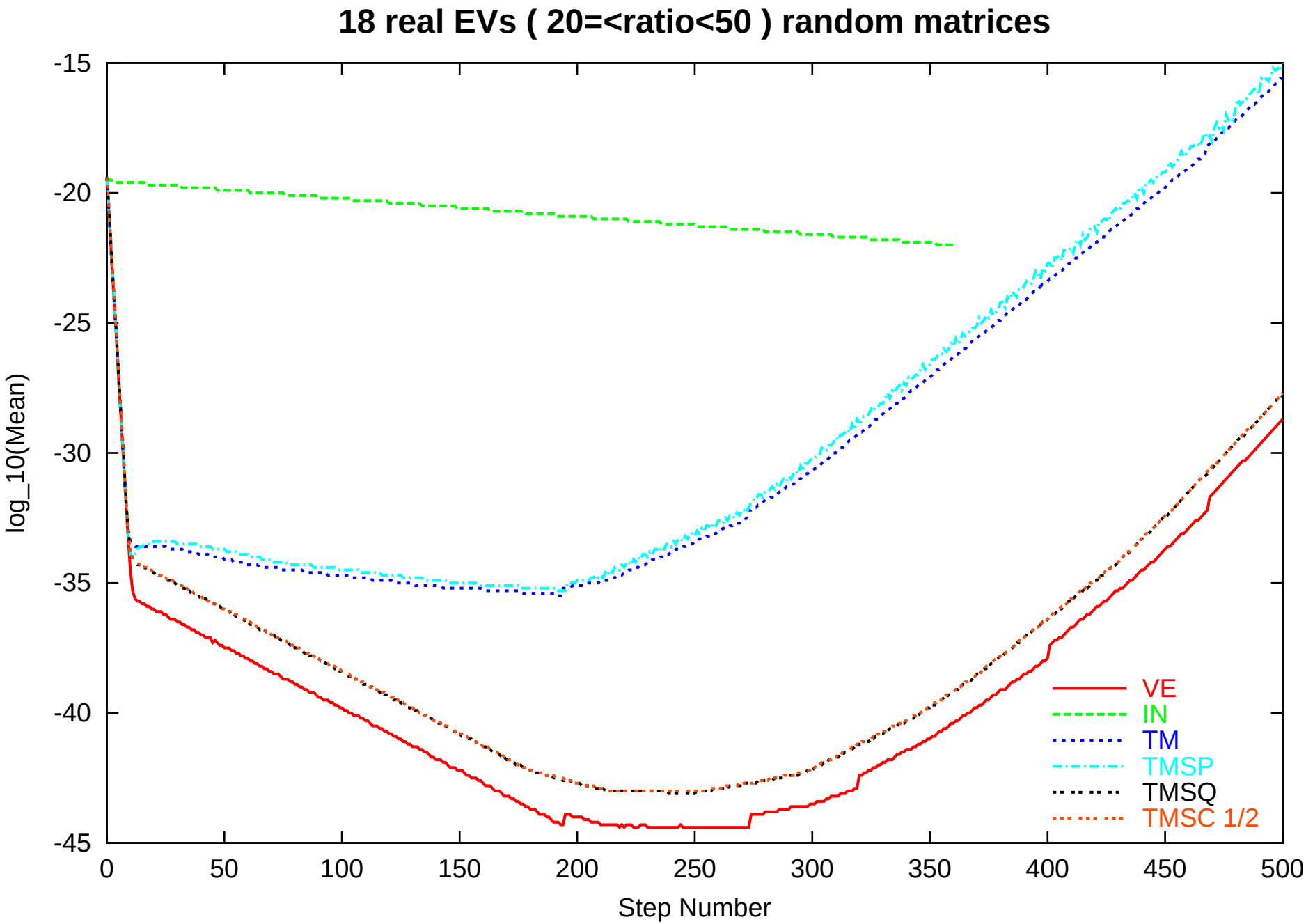

325 conjugate random matrices
log_10(Mean)
Step Number
0
-20
-40
-60
-80
-100
-120
-140
-160
0
50
100
150
200
250
300
350
400
450
500
VE
IN
TM
TMSP
TMSQ
TMSC 1/2

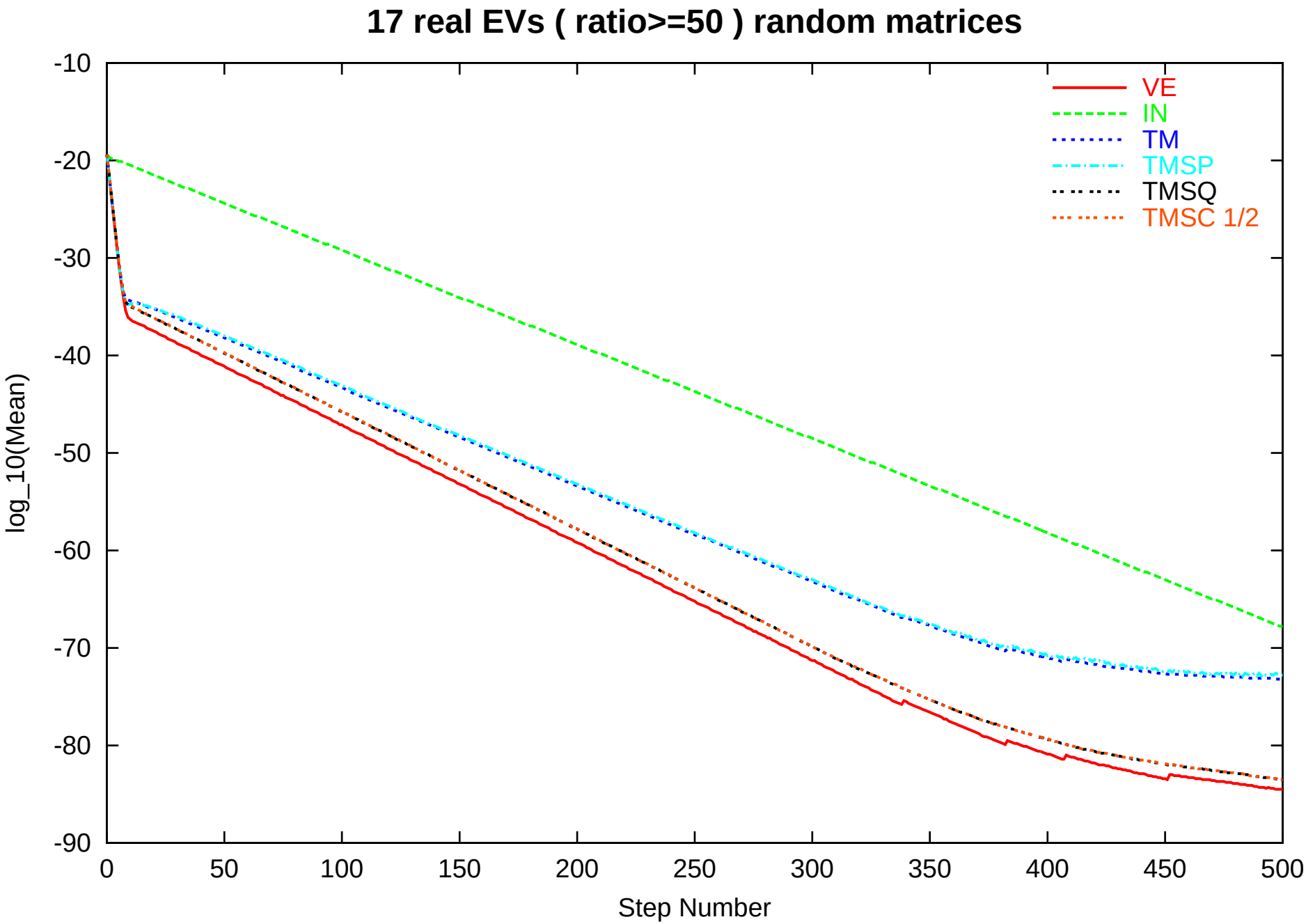

287 real EVs (ratio<2) random matrices
log_10(Mean)
Step Number
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
0
50
100
150
200
250
300
350
400
450
500
VE
IN
TM
TMSP
TMSQ
TMSC 1/2

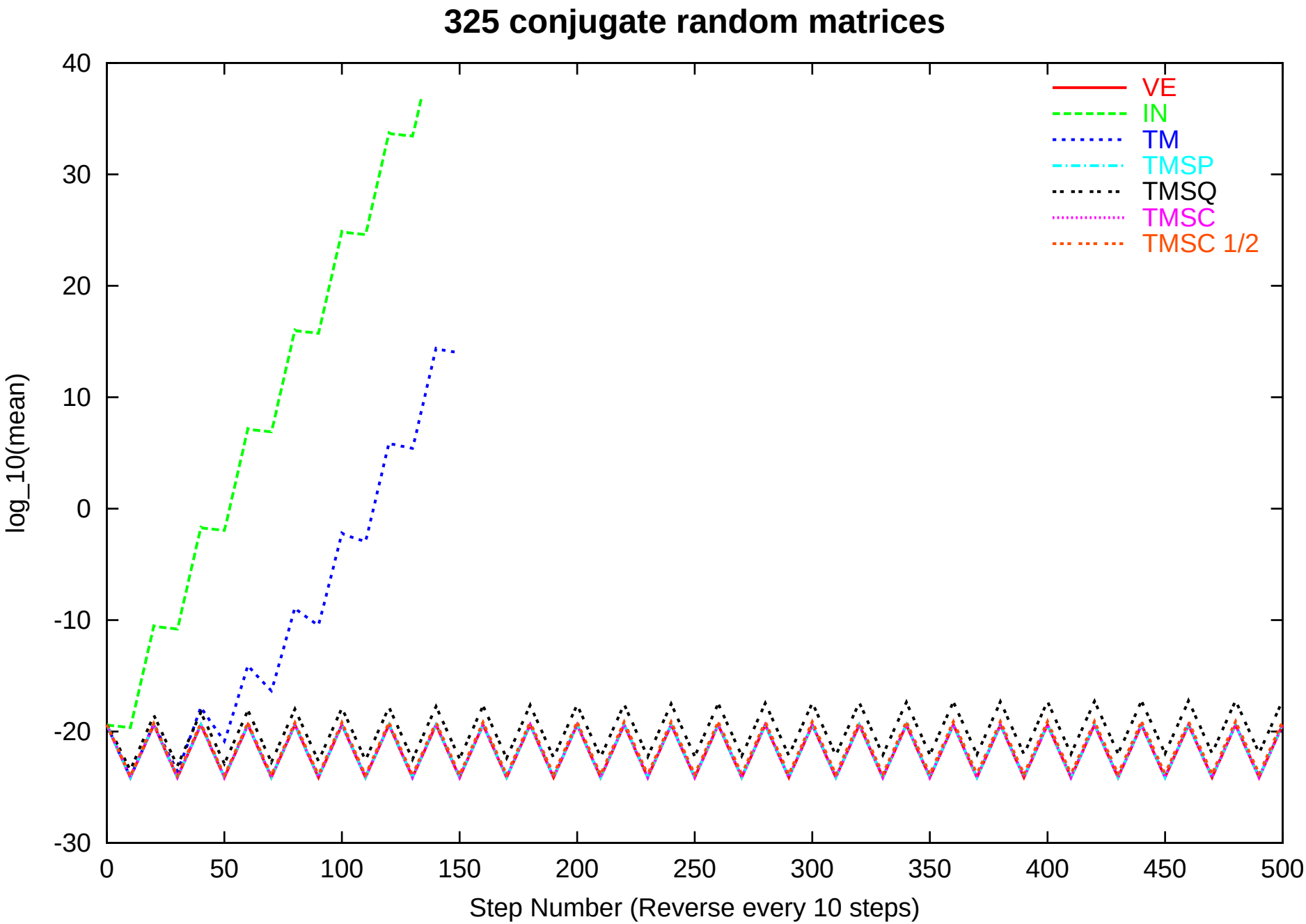

233 real EVs (2=<ratio<5) random matrices
log_10(Mean)
Step Number
-10
-20
-30
-40
-50
-60
-70
-80
-90
0
50
100
150
200
250
300
350
400
450
500
VE
IN
TM
TMSP
TMSQ
TMSC 1/2

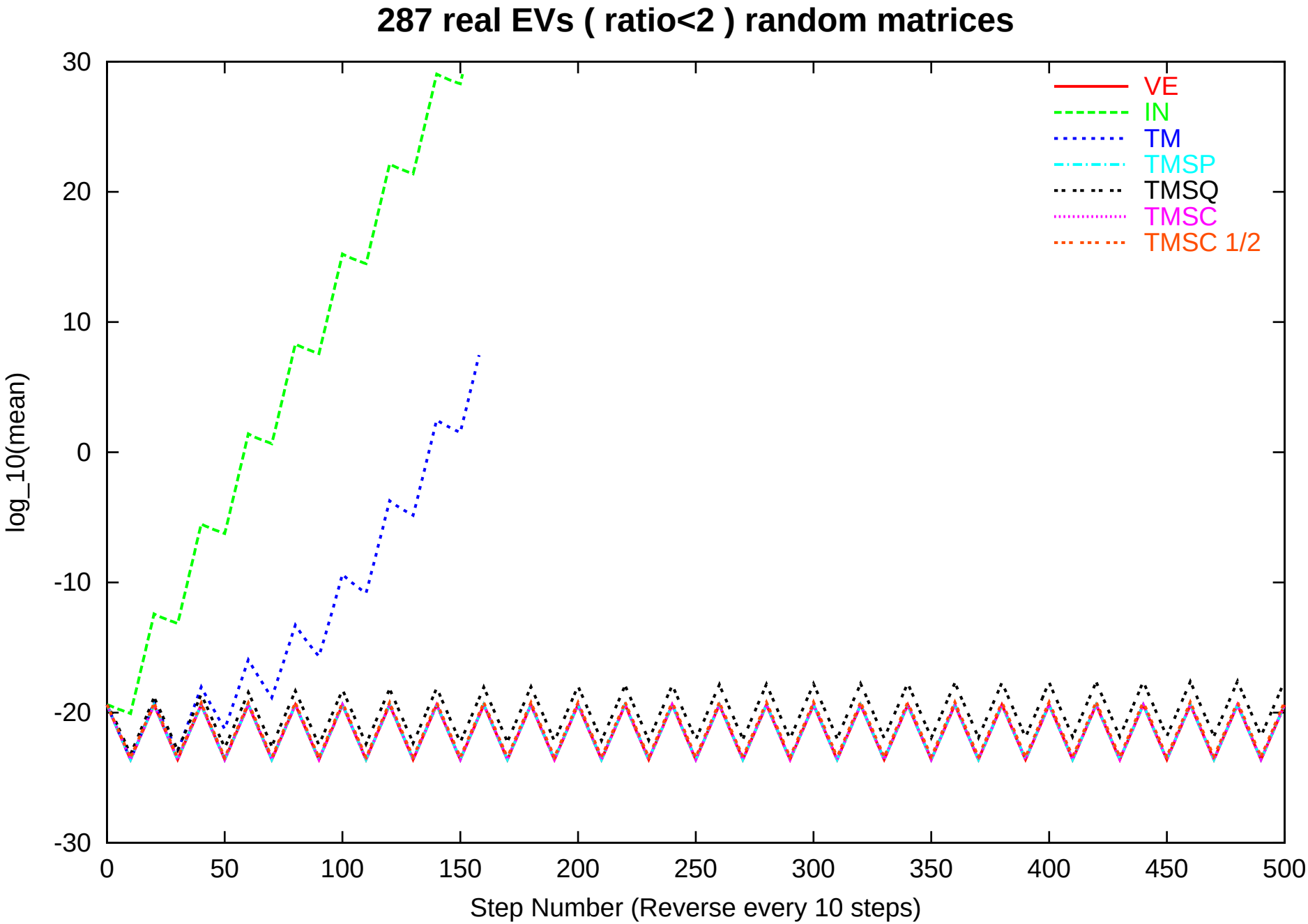

80 real EVs (5=<ratio<10) random matrices
log_10(Mean)
Step Number
-10
-20
-30
-40
-50
-60
-70
-80
0
50
100
150
200
250
300
350
400
450
500
VE
IN
TM
TMSP
TMSQ
TMSC 1/2

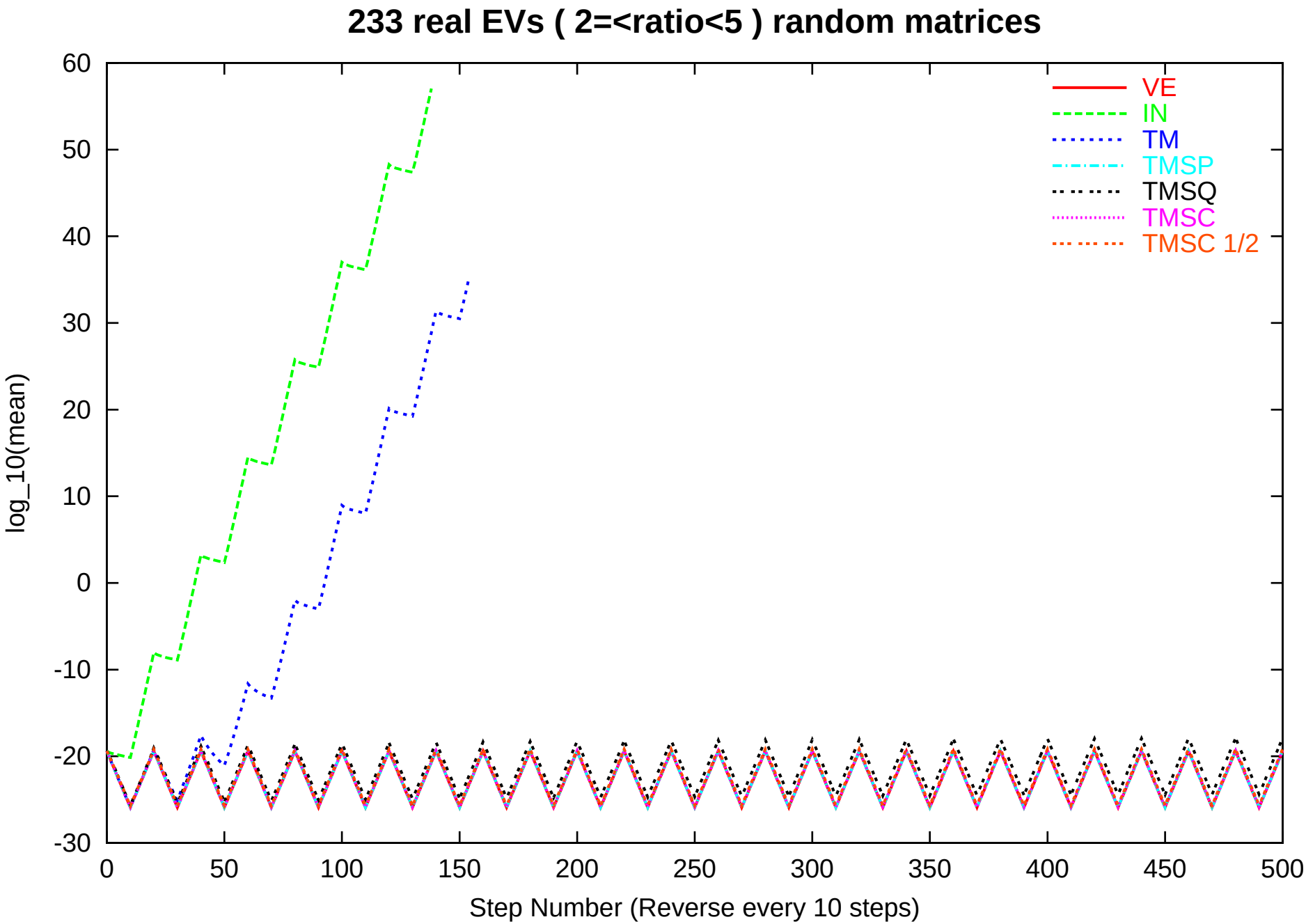

40 real EVs (10=<ratio<20) random matrices
log_10(Mean)
Step Number
-15
-20
-25
-30
-35
-40
-45
-50
-55
-60
-65
0
50
100
150
200
250
300
350
400
450
500
VE
IN
TM
TMSP
TMSQ
TMSC 1/2

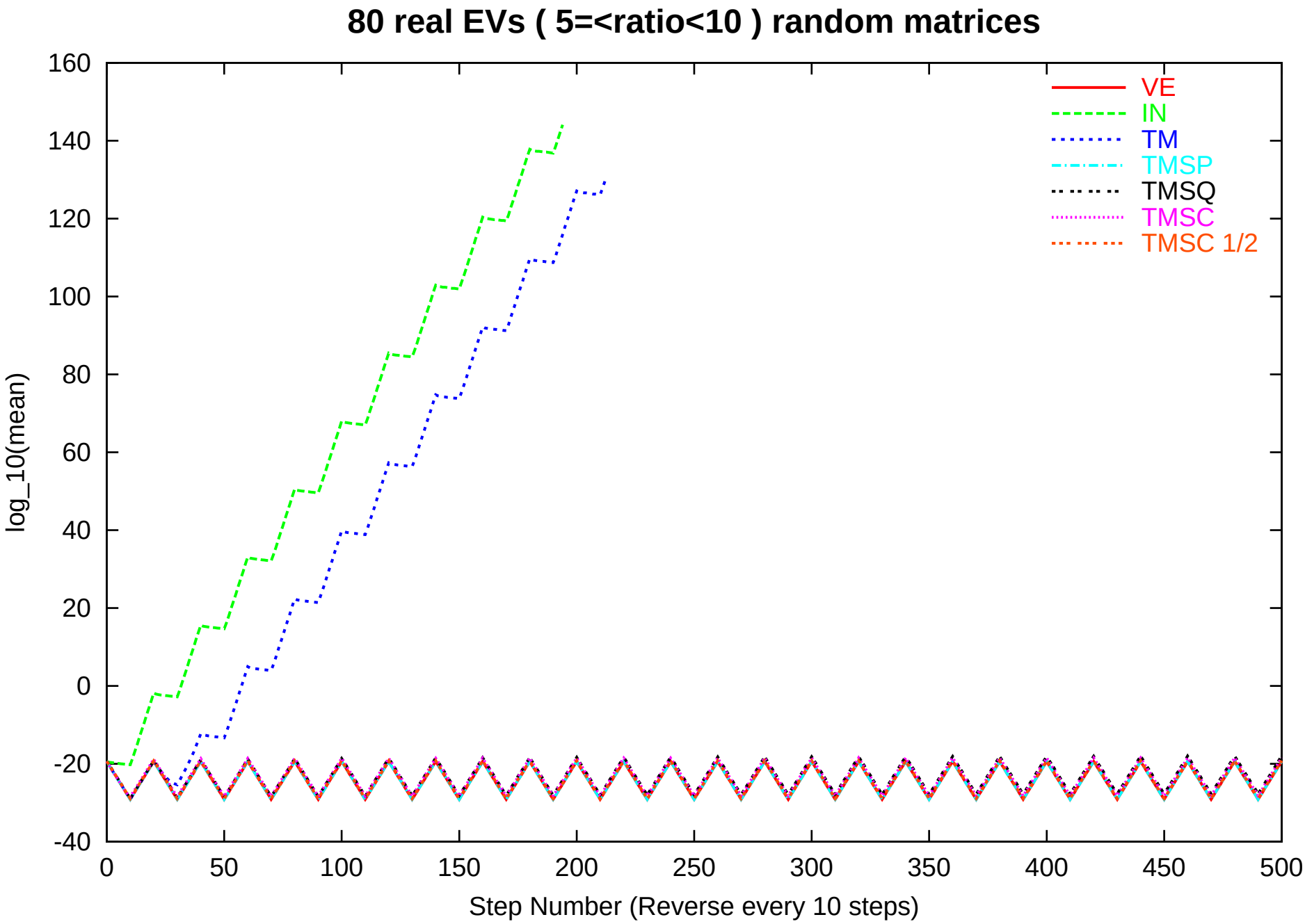

18 real EVs (20=<ratio<50) random matrices
log_10(Mean)
Step Number
-15
-20
-25
-30
-35
-40
-45
0
50
100
150
200
250
300
350
400
450
500
VE
IN
TM
TMSP
TMSQ
TMSC 1/2

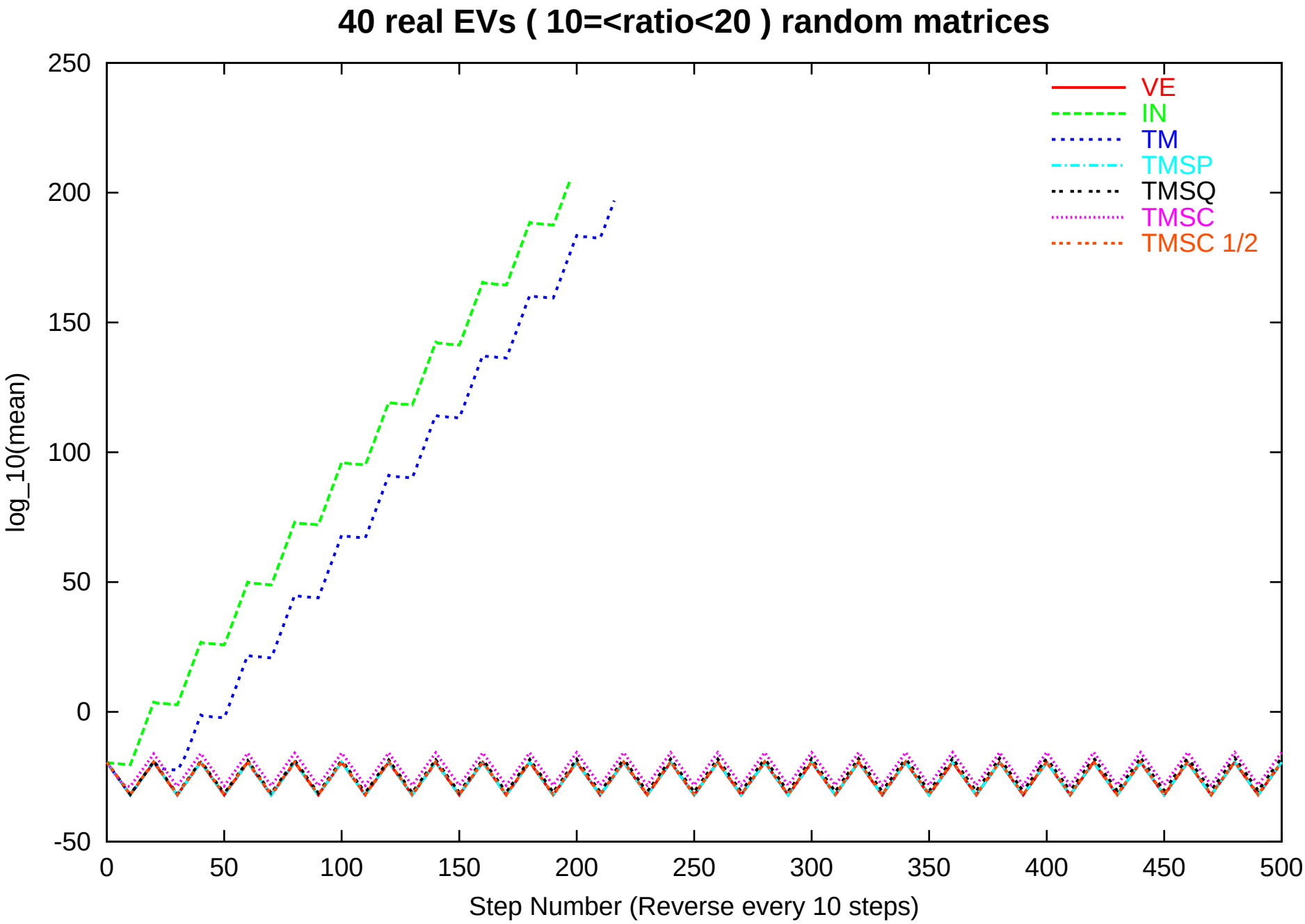

17 real EVs (ratio>=50) random matrices
log_10(Mean)
Step Number
-10
-20
-30
-40
-50
-60
-70
-80
-90
0
50
100
150
200
250
300
350
400
450
500
VE
IN
TM
TMSP
TMSQ
TMSC 1/2

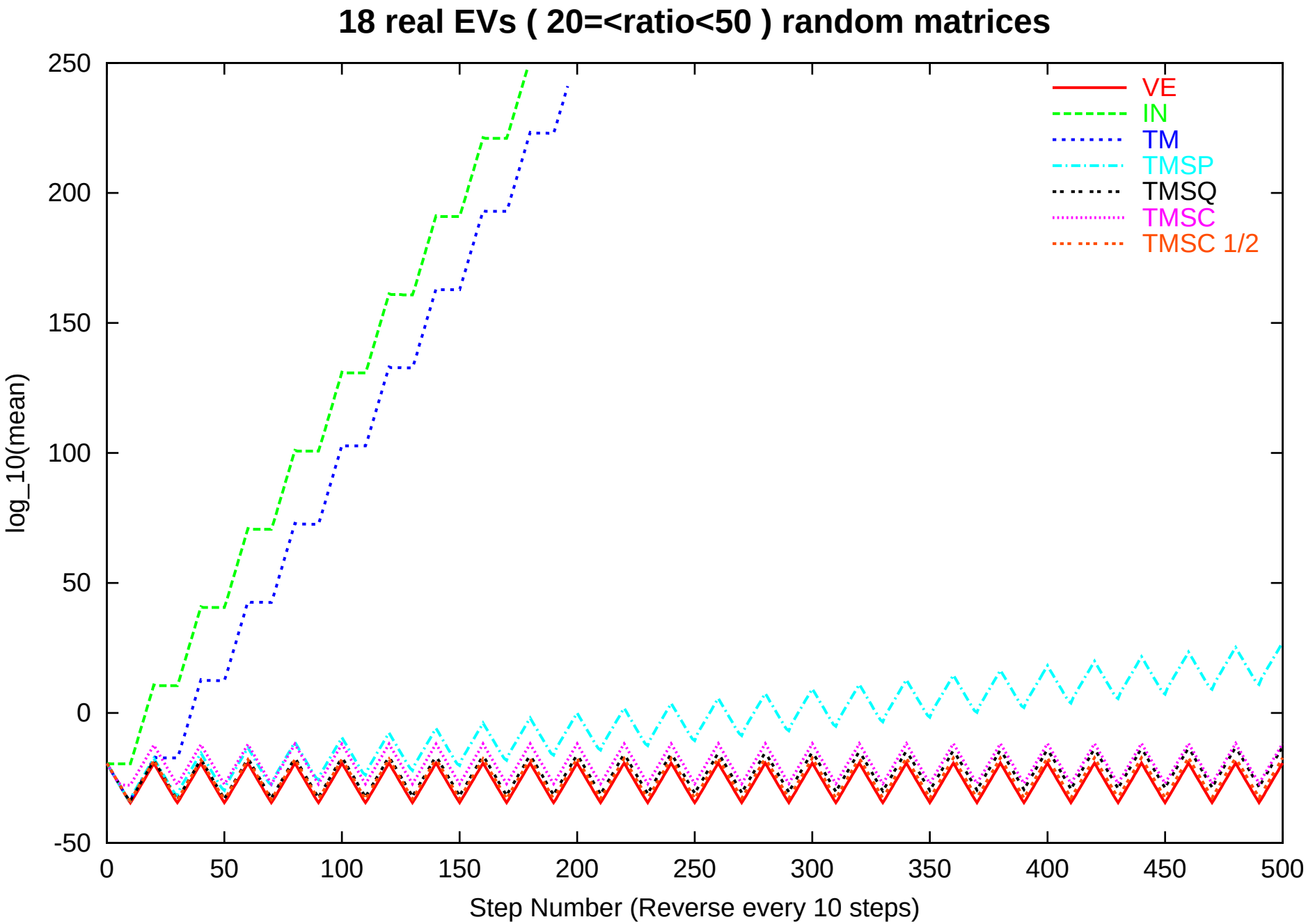

325 conjugate random matrices
log_10(mean)
Step Number (Reverse every 10 steps)
VE
IN
TM
TMSP
TMSQ
TMSC
TMSC 1/2
40
30
20
10
0
-10
-20
-30
0
50
100
150
200
250
300
350
400
450
500

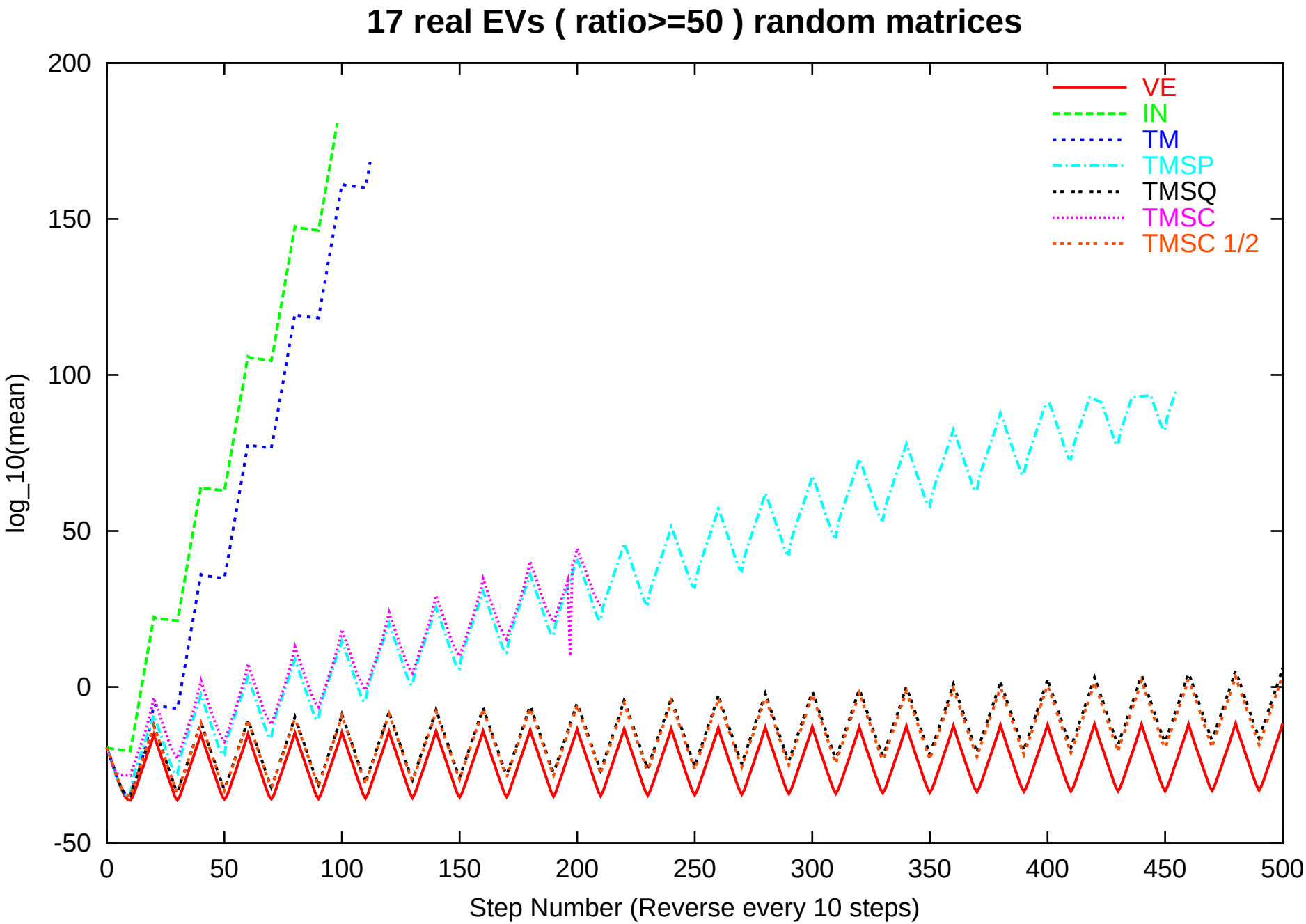

287 real EVs (ratio<2) random matrices
log_10(mean)
Step Number (Reverse every 10 steps)
30
20
10
0
-10
-20
-30
0
50
100
150
200
250
300
350
400
450
500
VE
IN
TM
TMSP
TMSQ
TMSC
TMSC 1/2


233 real EVs (2=<ratio<5) random matrices
log_10(mean)
Step Number (Reverse every 10 steps)
VE
IN
TM
TMSP
TMSQ
TMSC
TMSC 1/2
60
50
40
30
20
10
0
-10
-20
-30
0
50
100
150
200
250
300
350
400
450
500


80 real EVs (5=<ratio<10) random matrices
log_10(mean)
Step Number (Reverse every 10 steps)
VE
IN
TM
TMSP
TMSQ
TMSC
TMSC 1/2
160
140
120
100
80
60
40
20
0
-20
-40
0
50
100
150
200
250
300
350
400
450
500


40 real EVs (10=<ratio<20) random matrices
log_10(mean)
Step Number (Reverse every 10 steps)
250
200
150
100
50
0
-50
0
50
100
150
200
250
300
350
400
450
500
VE
IN
TM
TMSP
TMSQ
TMSC
TMSC 1/2


18 real EVs (20=<ratio<50) random matrices
log_10(mean)
Step Number (Reverse every 10 steps)
250
200
150
100
50
0
-50
0
50
100
150
200
250
300
350
400
450
500
VE
IN
TM
TMSP
TMSQ
TMSC
TMSC 1/2


17 real EVs (ratio>=50) random matrices
log_10(mean)
Step Number (Reverse every 10 steps)
VE
IN
TM
TMSP
TMSQ
TMSC
TMSC 1/2
200
150
100
50
0
-50
0
50
100
150
200
250
300
350
400
450
500