Using Numerical Continuation Methods and Galerkin's Method for Finding and Tracing Periodic Solutions in Nonlinear Dynamic Systems

Fabian Späh Mats Bosshammer

SS 2016

Contents

1 Introduction

2 Modeling Generic Periodic Solutions

3 An Optimality Criterion for Periodic Solutions, Galerkin's Method

Given a system of $n \in \mathbb{N}$ real valued, possibly non-linear, stationary, ordinary, coupled, first-order differential equations $\mathbf{x}$

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}), \tag{1}$$

we are interested in numerically computed, periodic solutions. That is, solutions $\mathbf{y} : \mathbb{R} \to \mathbb{R}^n$ which obey $\mathbf{y}(t) = \mathbf{y}(t + T)$ for all $t \in \mathbb{R}$ and some period $T \in \mathbb{R}$. This is the general case, as differential equations of any degree can be converted to a system of degree-1 differential equations.

Model Solution candidates need to be modeled in a certain way. The periodicity constraint suggests using a multidimensional trigonometric polynomial of degree $m \in \mathbb{N}$

$$\mathbf{y} \coloneqq \sum_{k=-m}^{m} \mathbf{y}_k \exp\left(i\omega k t\right),$$

where $\mathbf{y}_k \in \mathbb{C}$, $\mathbf{y}_{-k} = \Re(\mathbf{y}_k) - i\Im(\mathbf{y}_k)$ for $k \in \mathbb{N}$, $-m \leq k \leq m$, $\omega = \frac{2\pi}{T}$. Solution candidates of this form satisfy $\mathbf{y}(t) = \mathbf{y}(t + T)$ by definition. A function of this form is defined solely by its $m + 1$ unique coefficients $\mathbf{y}_k$ for $0 \leq k \leq m$.

Optimality Criterion Finding good solutions, that is, functions $\mathbf{y}$ which at least approximate $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$, requires a measure of fit of the solution. In this case, Galerkin's method takes this role. A useful property of the trigonometric polynomial is, that it can be trivially differentiated

$$\mathbf{y}' = \sum_{k=-m}^{m} i\omega k \mathbf{y}_k \exp\left(i\omega k t\right).$$

Employing this property in the definition of the differential equation system yields

$$\begin{aligned} & \mathbf{y}' = \mathbf{f}(t, \mathbf{y}) \\ \Leftrightarrow\ & \mathbf{f}(t, \mathbf{y}) - \mathbf{y}' = 0 \\ \Leftrightarrow\ & \mathbf{f}(t, \mathbf{y}) - \sum_{k=-m}^{m} i\omega k \mathbf{y}_k \exp\left(i\omega k t\right) = 0. \end{aligned}$$

The difference between these two functions is called the *residual* $\mathbf{r}(t) \coloneqq \mathbf{f}(t, \mathbf{y}) - \mathbf{y}'$. A candidate $\mathbf{y}$ is a solution iff $\mathbf{r} = 0$. In general the solution of the system of differential equations will not be representable by a trigonometric polynomial. Galerkin's method relaxes the equality requirement such that only projections on a set of so called *trial vectors*, needs to be zero. This is equivalent to requiring a projection of $\mathbf{r}(t)$ onto the subspace spanned by the trial vectors to be zero. Choosing the complex oscillations as a basis for the subspace as well is a

solid choice. Basically the residual is approximated by a trigonometric polynomial and a solution is required to only minimize this representation. There are other factors supporting this choice: The residual is periodic as well, because of orthogonality many terms can cancel each other out, it allows us to employ the FFT for many operations, and the resulting system of equations is almost balanced.

System of Equations Because we are only interested in real systems, this yields $m+1$ equations, one for each trial vector $\mathbf{v}_k = \exp(i\omega k t)$ for $0 \leq k \leq m$

$$\begin{aligned} & \langle \mathbf{r}, \mathbf{v}_k \rangle \\ = & \langle \mathbf{f}(t, \mathbf{y}), \mathbf{v}_k \rangle - \langle \mathbf{y}', \mathbf{v}_k \rangle \\ = & \langle \mathbf{f}(t, \mathbf{y}), \mathbf{v}_k \rangle - i\omega k \mathbf{y}_k. \end{aligned}$$

The last step exploits that there is always exactly one component of the candidate function which is not orthogonal to the trial vectors. However, there are $m+2$ variables: $m+1$ unique coefficients and ω. This represents the situation, that at this point there is still one degree of freedom: Each phase shifted version of a solution is still a solution. We thus introduce another generic equation called the *anchor* equation, which basically chooses one of these solutions. In this case $\langle \mathbf{y}(0), (\delta_{1i})_{i \in \mathbb{N}_n} \rangle = 0$ is used: For $t = 0$, the solution needs to intersect the hyperplane defined by being zero in the first component. This can be formulated by requiring the corresponding coefficients to sum up to zero. The anchor equation needs to be adapted to the system considered: If there are no intersections with this plane, another equation needs to be chosen.

When transferring these theoretical constructs to a practical setting, the main change is that solutions $\mathbf{y}$ are not considered to be functions of continuous time, but vectors representing discrete time. Let $N = 2m + 1$, then

$$\mathbf{y} = \left(\sum_{k=0}^{N-1} \mathbf{y}_k \exp\left(i \frac{2\pi}{N} jk \right) \right)_{0 \leq j \leq N-1}.$$

This is no real limitation at that point, as solution candidates are band limited by construction. Define the *DFT matrix* $\mathbf{F} \in \mathbb{R}^{N \times N}$ and the coefficient vector $\mathbf{Y} \in (\mathbb{C}^n)^N$

$$\begin{aligned} \mathbf{F} &\coloneqq N^{-1} \left(\exp\left(-i \frac{2\pi}{N} jk \right) \right)_{0 \leq j,k \leq N-1} \\ \mathbf{Y} &\coloneqq (\mathbf{y}_k)_{0 \leq k \leq N-1}. \end{aligned}$$

$\mathbf{Y}$ is a vector of vectors, while somewhat unusual, allows for more concise notation than notation as a matrix. Using these constructs allows us to reformulate

$$\mathbf{y} = \mathbf{F}^{-1}\mathbf{Y}.$$

Considering only the equations from projection onto a trial vector and considering ω to be known, and let $\mathbf{K} \in \mathbb{R}^{N \times N}$ be the diagonal matrix with entries going from 0 to m to $-m$, the system of equations becomes

$$\mathbf{F}\tilde{\mathbf{f}}(\mathbf{F}^{-1}\mathbf{Y}) - i\omega\mathbf{K}\mathbf{Y} = 0,$$

where $\tilde{\mathbf{f}} : (\mathbb{R}^n)^N \mapsto (\mathbb{R}^n)^N$ simply row wise applies $\mathbf{f}$, the function defining the system of differential equations, see Equation 1.

Introducing these new constructs might at first glance appear unnecessarily complicated, because the system of equations was already known, and could be used in the state it was in. However, an implementation in the basic form would have required different treatment for each dynamic system, while this reworked form is only trivially dependent of the system and the degree of the trigonometric polynomial. That is, the auxiliary vectors and matrices are trivially constructed, as is $\tilde{\mathbf{f}}$ from $\mathbf{f}$. While knowing the system of equations is important, the core task of finding a good solution requires the Jacobian of the system. The by far largest advantage is that from this form, the system can be derived in a general way.

Let D denote a differential operator. The only variable in the system is $\mathbf{Y}$, thus

$$\begin{aligned} & D\left(\mathbf{F}\tilde{\mathbf{f}}(\mathbf{F}^{-1}\mathbf{Y}) - i\omega\mathbf{K}\mathbf{Y}\right) \\ = \ & \left(\mathbf{F}\mathbf{J}_{\tilde{\mathbf{f}}}\left(\mathbf{F}^{-1}\mathbf{Y}\right)\mathbf{F}^{-1} - i\omega\mathbf{K}\right) \cdot D\mathbf{Y}, \end{aligned} \tag{2}$$

where $\mathbf{J}_{\tilde{\mathbf{f}}} \coloneqq \frac{d\tilde{\mathbf{f}}(\mathbf{Y})}{d\mathbf{Y}}$ is the Jacobian of $\tilde{\mathbf{f}}$, a vector-by-vector derivative, which yields a matrix. However, because both vector's elements are vectors themselves, the entries of the matrix are again vector-by-vector derivatives. This results in an $N \times N$ matrix of matrices with

$$\left(\mathbf{J}_{\tilde{\mathbf{f}}}\right)_i^j = \frac{d}{d\mathbf{x}_j}\mathbf{f}(\mathbf{x}_i) = \delta_{ij}\mathbf{J}_{\mathbf{f}}(\mathbf{x}_i),$$

where $\mathbf{J}_{\mathbf{f}} \coloneqq \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}}$ is the Jacobian of $\mathbf{f}$. The matrix $\mathbf{J}_{\tilde{\mathbf{f}}}$ is diagonal, thus

$$\mathbf{J}_{\tilde{\mathbf{f}}}(\mathbf{Y}) = \operatorname{diag}\left(\left(\mathbf{J}_{\mathbf{f}}\left(\mathbf{Y}_i\right)\right)_{i\in\mathbb{N}_N}\right)$$

Because different cells of $\mathbf{J}_{\mathbf{f}}$ are independent, the expression for the Jacobian of the whole system can be considered separately for each partial derivative in $\mathbf{J}_{\mathbf{f}}$. For $k, l \in \mathbb{N}_n$ define the $N \times N$ diagonal matrix which extracts a single partial derivative

$$\mathbf{A}_{kl}(\mathbf{Y}) = \operatorname{diag}\left(\left(\frac{\partial f_k}{\partial x_l}\left(\mathbf{Y}_i\right)\right)_{i\in\mathbb{N}_N}\right).$$

For each k and l Equation 2 then becomes

$$\left(\mathbf{F}\mathbf{A}_{kl}(\mathbf{F}^{-1}\mathbf{Y})\mathbf{F}^{-1} - i\omega\mathbf{K}\right) \cdot D\mathbf{Y} \tag{3}$$

This expression is again free of nested vectors.

Because $\mathbf{A}_{kl}(\mathbf{F}^{-1}\mathbf{Y})$ is diagonal, $\mathbf{A}_{kl}(\mathbf{F}^{-1}\mathbf{Y})\mathbf{F}^{-1}$ is the rows of $\mathbf{F}^{-1}$ each multiplied by a constant. This can also be seen as an element wise multiplication of the columns of $\mathbf{F}^{-1}$ by a column vector defined by the diagonal of $\mathbf{A}_{kl}(\mathbf{F}^{-1}\mathbf{Y})$:

$$\begin{aligned} \mathbf{c}_k &\coloneqq \left(\exp\left(i\frac{2\pi}{N}jk\right)\right)_{0\le j\le N-1} \\ \mathbf{d} &\coloneqq \frac{\partial f_k}{\partial x_l}\{\mathbf{Y}\}. \end{aligned}$$

A column of $\mathbf{F}\mathbf{J}_{\tilde{\mathbf{f}}}\left(\mathbf{F}^{-1}\mathbf{Y}\right)\mathbf{F}^{-1}$ is then the transform of the product of two signals $\mathcal{F}(\mathbf{d}\cdot\mathbf{c}_k)$. Because of the structure of $\mathbf{c}_k$, this corresponds to a simple periodic shift of the Fourier coefficients of the discrete signal $\mathbf{d}$ by k steps. Using this for every column of the complete term yields

$$\mathbf{F}\mathbf{A}_{kl}(\mathbf{F}^{-1}\mathbf{Y})\mathbf{F}^{-1} = \left(\mathcal{F}(\mathbf{d})_{((i-j))_N}\right)_{0\le i,j\le N},$$

that is, a simple circulant matrix.

4 Finding Periodic Solutions

A separate problem from evaluating, optimizing and continuing periodic solutions is finding an initial candidate solution. Because continuation is a crucial part of this project, having just one single periodic solution might enable to find many others, through tracing and switching solution branches. Because the systems considered in this work have stable periodic solutions, we focus on this case. When this is not the case, as mentioned in the section about continuation methods (section 5), a homotopy between a trivial system and the target system, combined with continuation methods, might be a promising approach.

Starting from a point in the periodic solution's basin of attraction, one can simply forward integrate such a system. There are several possible problems involved:

- Forward integration can accumulate errors.
- Even if the starting point lies exactly on the periodic trajectory, the sampling interval would probably not be an integer fraction of the period of the periodic trajectory. Periodicity in the sequence of points are not directly related to periodicity in the continuous system.
- Given the nature of the project, it is very likely that period doubling bifurcations are encountered. These provoke situations where two periodic solutions exist which are difficult to distinguish.

Forward integration yields a sequence of points in phase space $(\mathbf{s}_i)_{i\in\mathbb{N}}$. To obtain more manageable data only intersections of the trajectory with a hyperplane $p(\mathbf{x}) = \langle \mathbf{n}, \mathbf{x} - \mathbf{x}_0 \rangle = 0$ in a single direction are considered (so called Poincaré sections). To find these, $\mathbf{s}_i, \mathbf{s}_{i+1} = \mathbf{s}_i + h_0 \cdot \mathbf{f}(\mathbf{s}_i)$ with $p(s_i) \leq 0 < p(s_{i+1})$ are searched. Because of continuity, there needs to exists $h \in [0, h_0]$ such that $p(\mathbf{s}_i + h \cdot \mathbf{f}(\mathbf{s}_i)) = 0$, which is found via bisection. Let $(\mathbf{u}_i)_{i\in\mathbb{N}}$ be the sequence of intersections.

To find the number of intersections per period, the intersections are partitioned into $k \in \mathbb{N}$ disjoint clusters $V_{k,i} = \{\mathbf{u}_m \mid m \in k\mathbb{N} + i\}$ for $i \in \mathbb{N}$, $i \leq k$. The relative quality of the i-th cluster can then be assessed using the within-cluster variance $\sum_{v\in V_{k,i}} ||v - \mathrm{E}(V_{k,i})||^2$. The correct number $k_{\min} \in \mathbb{N}$ of intersections in one period is then taken to be the one minimizing the sum of within-cluster variances:

$$k_{\min} \coloneqq \arg\min_{k\in\mathbb{N}} \sum_{i=1}^{k} \sum_{v\in V_{k,i}} ||v - \mathrm{E}(V_{k,i})||^2.$$

For further information about these measures see for example [Halkidi et al., 2001].

In this form the criterion might at best work if the intersection sequence is infinite. When dealing with finite sequences increasing the number of clusters inevitably leads to lower total within-cluster variances. It is thus necessary to constrain the available values for k and discourage the method of overestimating the number of clusters. Trivially an upper limit for k needs to be introduced.

Furthermore, the minimality criterion needs to be relaxed: Suppose there are $k_{\min}$ intersections per period, then all multiples of $k_{\min}$ yield equal or lower ratings. One thus wants to choose the minimum k which is in some sense still almost optimal.

5 Tracing Periodic Solutions, Predictor-Corrector Continuation

A central part of this project are numerical continuation methods. Being a large topic, the details are out of the scope of this work, we thus concentrate on conveying a general idea of the concepts. This is especially true for the wealth of methods from other fields, numerical continuation draws upon. Among the things required in the following are basic knowledge of Newton's method in multiple dimensions, forward integrating differential equations (Runge-Kutta methods) and basic vector calculus (Jacobian matrix). For a more thorough introduction to the topic, see [Allgower and Georg, 1990], on which the continuation part of project and this section of this document is based. The following is a reduced introduction of the concepts used for this project.

The base of numerical continuation is formed by the fact that in the vicinity of a solution of an underdetermined continuous system, there are almost always other, similar solutions. Iterative application of this, in the context of systems having one constraint less than unknowns, yields that solutions of the system form a curve (which might be closed). Continuation methods provide means to trace these curves, and thus to derive infinitely many other solutions from a single given solution. A very central use-case of such a method is being able to numerically solve a system without requiring an otherwise needed good starting value. This works by continuously blending two systems with equal numbers of unknowns and equations using a homotopy, which adds one degree of freedom. This way, a known solution from a system can be traced to one of the harder system. However, in the context of this work, where the underdetermination occurs naturally, we are not interested in single solutions, but rather in obtaining the whole continua of solutions.

This project uses a single continuation method: The predictor-corrector method. As the name suggests, it continues the solution curves, by predicting the next point on the curve through linearization at the current point, and afterwards correcting it, to compensate for the non-linear influences.

For a matrix $A \in \mathbb{R}^{N\times(N+1)}$ with full rank, let $\operatorname{tang}(A) \in \mathbb{R}^{N+1}$ denote its *tangent*, that is the unique vector with

- $A \cdot \operatorname{tang}(A) = 0$
- $||\operatorname{tang}(A)|| = 1$
- $\det \begin{pmatrix} A \\ \operatorname{tang}(A)^T \end{pmatrix} > 0$

For a given function $H : \mathbb{R}^{N+1} \to \mathbb{R}^N$ for $N \in \mathbb{N}$, let $J : \mathbb{R}^{N+1} \to \mathbb{R}^{N\times(N+1)}$ denote its Jacobian. The tangent can be used to define an initial value problem specific to H, the Davidenko equation ([Davidenko, 1953]):

- $\dot{c} = \operatorname{tang}(J(c))$
- $c(0) = c_0 \in \mathbb{R}^{N+1}$

Solutions $c : \mathbb{R} \to \mathbb{R}^{N+1}$ to this problem satisfy $H(c(t)) = \text{const}$ for all $t \in \mathbb{R}$, as can be seen on

$$\frac{d}{dt} H(c(t)) = J(c(t)) \cdot \dot{c}(t) = J(c(t)) \cdot t(J(c(t))) = 0.$$

From a theoretical perspective, the Davidenko equation is the sole thing needed for continuation. Tracing solutions becomes a matter of finding a solution for the differential equation. In a practical setting, this also includes the same numerical obstacles, and ignores that the underlying functions H and J are known. Numerically integrating the Davidenko equation is the prediction part as indicated earlier. It introduces errors, such that the constancy is no longer guaranteed. However, since H and J are known, the point can be *corrected* towards the constant solution curve.

The correction part of the method uses a variant of Newton's method, working in multiple dimensions and with non-square Jacobians. Finding zeros using Newton's method in multiple dimensions works similar to the single dimensional case: A function is linearized, and the zero of the linear problem is considered a good approximation of a zeros of the more complex problem. In non-degenerate cases, iterating this procedure quickly converges to a root, the fixed points of this iteration.

First we consider function $H : \mathbb{R}^N \to \mathbb{R}^N$ with square Jacobian $J : \mathbb{R}^N \to \mathbb{R}^{N \times N}$. In point $x_i \in \mathbb{R}^N$ the next point $x_{i+1} \in \mathbb{R}^N$ is defined as the root of the linearization in x_i, $x \mapsto H(x_i) + J(x_i) \cdot (x - x_i) \stackrel{!}{=} 0$, which becomes

$$x_{i+1} \coloneqq x_i - H(x_i) J^{-1}(x_i).$$

Generalizing the method to non-square Jacobians involves finding a substitute for inverting the Jacobian, which is not defined for non-square matrices. This can be done using *Moore-Penrose* pseudoinverse.

6 Case Study: Lorenz System

7 Case Study: Rössler System

The well known Rössler system is given by the equations

$$\begin{aligned} \dot{x} &= -y - z \\ \dot{y} &= x + a \cdot y \\ \dot{z} &= b + z \cdot (x - c) \end{aligned}$$

with parameters a, b, and c. For this example, a and b are fixed to 0.1, while we are varying c.

The system has several periodic solutions for each value of c with different periodicities, though only one is stable at a time. We are interested in the origin and branching of those solutions and, thus, drawing a bifurcation diagram using the map

$$f \mapsto \{\|f(t)\|_2 \mid t \in [0, 2\pi),\ f(t)_1 = 0\} \ .$$

Here, f denotes a single periodic solution. Note, that we use only the approximation given by Galerkin's method.

Now, the exploration of the bifurcation with the given tools follows rougly this steps:

1. As a starting point, search for a value of c such that the system has a stable, periodic solution and use the method described in section 4 to find it. 4000 iterations in the transient part with step size of 0.01, 120 generated intersections with the $(x = 0)$-plane and tests for at most 30 periods were sufficient for all initial solutions of the Rössler system. We started at $c = 4$ with 64 samples.

2. Trace out the branch just by following the newly found solution in both directions using the predictor-corrector continuation method (section 5). The parameters $\kappa = 0.4$, $\delta = 3.0$, and $\alpha = 10.0$ are a decent choice for the whole diagram as a good tradeoff between performance and accuracy.

3. It is possible to detect a pitchfork bifurcation by doubling up the periodicity artifically. Using the described predictor-corrector method, the value of c will eventually converge to the bifurcation point, i.e., a point with a singular tangent. Since the method is now stuck, reaching the doubly-periodic solution can be done by solving a slightly perturbed problem for a short time. A short perturbation of strength 0.2 was enough to reach the doubly-periodic solution. Also, it might be necessary to increase the sample size due to increased complexity in the descending branch.

4. Remove the perturbation and continue tracing out the bifurcation, i.e. start over in 1 or 2.

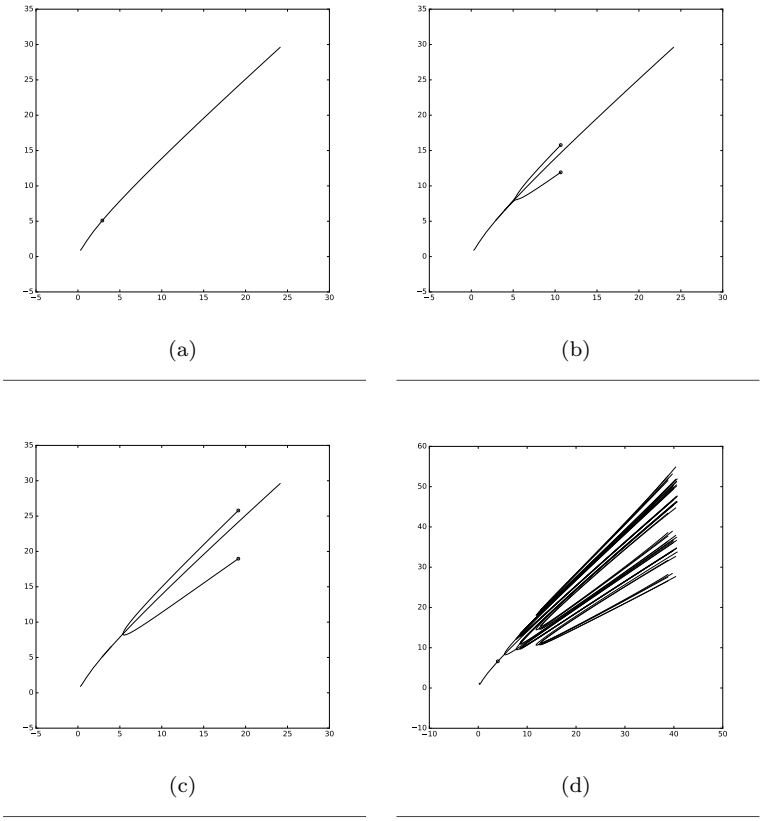

Figure 1: Tracing out the branch of solutions beginning with the solution for $c = 4$ (a). The bifurcation point in $c \approx 5.376$ is avoided by using a permutation (b), and the two-periodic solutions can be traced out after removing the inaccurate values from before (c). Eventually, we reach (d), a widely traced out bifurcation diagram. Note the solutions with periodicity three and descendants, that do not originate in the peridocity-one branch.

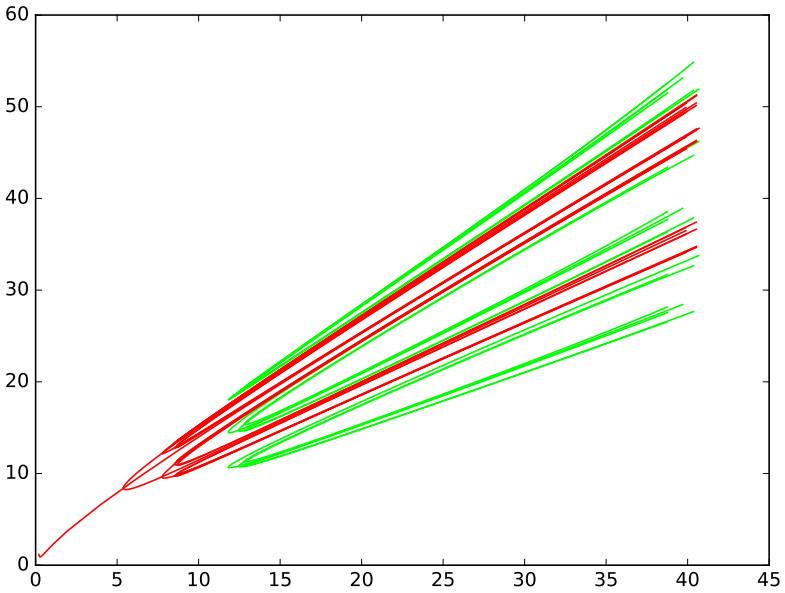

Figure 2: The bifurcation diagram of the Rössler system for varying c. The even-periodic solutions are colored red, the odd-periodic ones green.

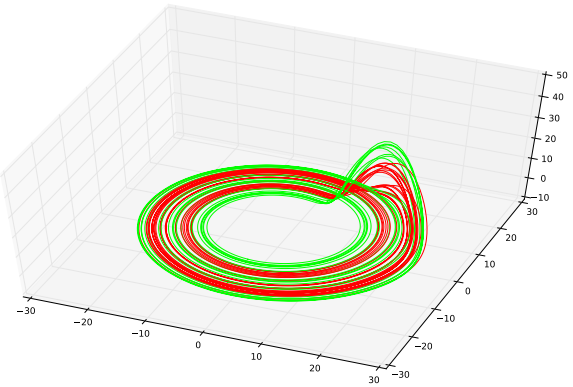

Figure 3: All found solutions for $c = \ldots$ plotted into one diagram. Observe how the odd-periodic and even-periodic solutions dodge since they do not originate from each other.

---

8 Outlook

References

[Allgower and Georg, 1990] Allgower, E. L. and Georg, K. (1990). *Numerical Continuation Methods: An Introduction.* Springer Verlag Berlin Heidelberg.

[Davidenko, 1953] Davidenko, D. (1953). On a new method of numerical solution of systems of nonlinear equations. In *Dokl. Akad. Nauk SSSR*, volume 88, pages 601–602.

[Halkidi et al., 2001] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems*, 17(2-3):107–145.