12-2016

(Un)Stable Manifold Computation via Iterative Forward-Backward Runge-Kutta Type Methods

Dmitriy Zhigunov
College of William and Mary

(Un)Stable Manifold Computation via Iterative Forward-Backward Runge-Kutta Type Methods

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Science in Mathematics from
The College of William and Mary

by
Dmitriy Zhigunov

Accepted for: _____

_____
Dr. Yu-Min Chung, Adviser

_____
Dr. Sarah Day

_____
Dr. Eugene R. Tracy

Williamsburg, VA
December 2016

Acknowledgements

I would like to thank Professor Yu-Min Chung for supervising me on this project, as well as for serving as the chair of my examining committee. Furthermore, I would like to thank the rest of my committee, Professors Sarah Day and Gene Tracy, for agreeing to serve on the committee.

Additionally, I would like to thank Professor Li for helping me out with the matrix analysis in my thesis.

Finally, I would like to thank the Aerobie company for creating the AeroPress, without which I would have been unable to make delicious coffee which gave me the energy to write.

Abstract

I present numerical methods for the computation of stable and unstable manifolds in autonomous dynamical systems. Through differentiation of the Lyapunov-Perron operator in [1], we find that the stable and unstable manifolds are boundary value problems on the original set of differential equation. This allows us to create a forward-backward approach for manifold computation, where we iteratively integrate one set of variables forward in time, and one set of variables backward in time. Error and stability of these methods is discussed.

Contents

1. Introduction & Background

Here I present algorithms for the computation of structures known as stable and unstable manifolds, which arise in systems of differential equations [2]. Differential equations, in the simplest sense, deal with relationships of how functions vary in time. For instance, the equations of motion for physical objects are most easily described as differential equations. The solutions to the differential equation would then give the particle trajectory and velocity as a function of time. The dynamics described by differential equations is by no means limited to particle trajectories. In general, any equation relating functions and their derivatives constitutes a system of differential equations. Functions which satisfy the system of differential equations are then called the solutions of the system differential equation.

Often of greater interest than the actual solution of a system of differential equations is its trajectory in phase space, which is the space of the solutions of the differential equation. For instance, if the functions $x(t)$ and $y(t)$ are solutions to some system of differential equations, the relevant phase space would be the xy-plane. Within a phase space, there may exist multiple objects of interest. The understanding of phase space structures is crucial for a full understanding of a given differential equation. For example, fixed points are points in phase space such that any trajectory which begins on a fixed point will remain on that fixed point (i.e. the solutions would be constant functions). More complicatedly, we have invariant sets, which are sets of phase space points such that a trajectory beginning in one of these points will remain in the invariant set.

The stable and unstable manifolds of a fixed point are special cases of invariant sets. In particular, the stable manifold is the set of all initial conditions whose trajectories asymptotically approach a given fixed point forward in time, and the unstable manifold is the set of all initial conditions whose trajectories asymptotically approach a given fixed point backwards in time. If a system is well-behaved, other phase space trajectories will asymptotically approach the stable and unstable manifolds backwards and forwards in time, allowing for them to be used to approximate the asymptotic behaviour of systems. See Figure 1 for an example.

The approach that I present here focuses on computation of stable and unstable manifolds by phrasing them as a boundary value problem on the original set of differential equations. This formulation suggests that iterative techniques can be used for computation. Thus, I will begin my thesis with a brief summary of the prior work that lead to the boundary value formulation. Afterwards, in Chapter 3, I will give descriptions of the numerical schemes for manifold computation, as well as describe their main properties. This will be the bulk of the thesis. Finally, in Chapter 4 I will take a closer look at one of the methods, in an attempt to develop a more rigorous approach to the study of the behavior of these schemes.

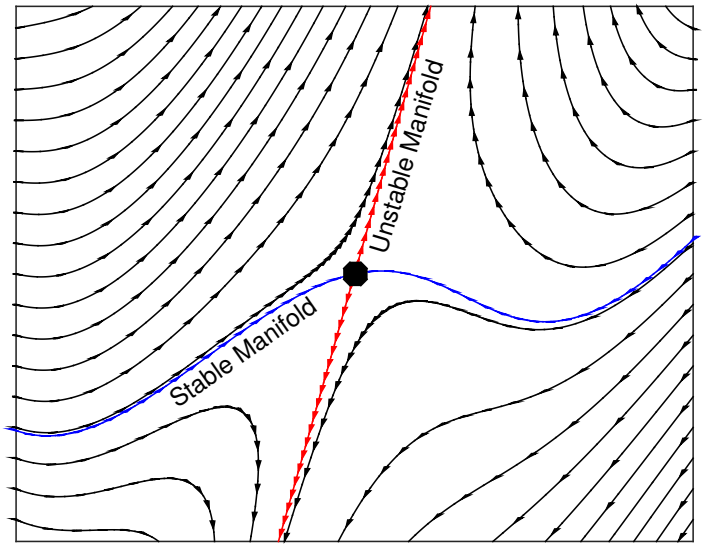

Figure 1. The stable and unstable manifold of a two dimensional, hyperbolic fixed point. This is the system analyzed in Example 15. This is a phase portrait. The curves represent the phase space trajectories of the systems, and the arrows represent the direction of motion along the trajectory forward in time.

1a. The stable and unstable manifolds

I will begin by defining the stable and unstable manifolds.

Definition 1. Consider a system of differential equations

$$\frac{d\mathbf{z}}{dt} = f(\mathbf{z}), \quad \mathbf{z} \in \mathbb{R}^n,$$

with some fixed point $\mathbf{z}^*$ – that is $\mathbf{z}(t) = \mathbf{z}^*$ is a solution. The *stable set* of $\mathbf{z}^*$ is given by

$$W_s(\mathbf{z}^*) = \left\{ \mathbf{z}_0 \,:\, \mathbf{z}(0) = \mathbf{z}_0 \implies \lim_{t\to\infty} \mathbf{z}(t) = \mathbf{z}^* \right\},$$

and the *unstable set* by

$$W_u(\mathbf{z}^*) = \left\{ \mathbf{z}_0 \,:\, \mathbf{z}(0) = \mathbf{z}_0 \implies \lim_{t\to-\infty} \mathbf{z}(t) = \mathbf{z}^* \right\}.$$

In the case that these sets are manifolds, we call them the *stable* and *unstable* manifolds.

Example 2. For the system

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -5 & 4 \\ -4 & 5 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix},$$

general solutions of the system have the form

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c_1 e^{-3t} \begin{pmatrix} 2 \\ 1 \end{pmatrix} + c_2 e^{3t} \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

where c_1, c_2 are constants determined by the initial conditions. Following from Definition 1, the stable manifold is the set of initial conditions such that $c_2 = 0$, and the unstable manifold is the set of initial conditions such that $c_1 = 0$. Some algebra leads to the stable manifold being $2x_0 - y_0 = 0$ and the unstable manifold being $x_0 - 2y_0 = 0$.

In a well-behaved system[1], the unstable manifold will attract all trajectories forward in time, and the unstable manifold will attract all trajectories backwards in time. For example, note the behaviors of the trajectories in Figure 1. Alternatively, in a hyperbolic linear system, following Definition 1, the stable manifold is the span of the eigenvectors corresponding to the negative eigenvalues, and the unstable manifold is the span of the eigenvectors corresponding to the negative eigenvalues. Clearly for large t the unstable components dominate, and for large $-t$ the stable components dominate. Hence, knowledge of the stable and unstable manifolds allows us to understand the asymptotic behavior of dynamical systems.

Furthermore, from Definition 1 it follows that the stable manifold is the unstable manifold backwards in time, and the unstable manifold is the stable manifold backwards in time. Therefore, the majority of the work I present here is written purely in the context of computation of stable manifolds. It is understood that the same algorithms can be applied to unstable manifold computation given the transformation $t' = -t$.

The crux of the numerical methods presented here rely on the ability to write the stable manifold as a boundary value problem. In most cases this requires that the system has no center component (see [2] for a discussion on center manifolds). I will thereforeassume that any given system that I am considering has no center components, which roughly correspond to pure imaginary eigenvectors of the Jacobian evaluated at the fixed point. Furthermore, I will assume that the fixed point of interest for any dynamical system is located at the origin.

1b. On the existence of stable manifolds

The optimal estimates for the existence of stable manifolds were given by Casteñeda and Rosa in [1]. They demonstrated that the stable manifold of the dynamical system[2]

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \mathcal{A}\mathbf{x} + \mathcal{F}(\mathbf{x}, \mathbf{y}), \qquad \mathbf{x} \in \mathbb{R}^m, \\ \frac{d\mathbf{y}}{dt} &= \mathcal{B}\mathbf{y} + \mathcal{G}(\mathbf{x}, \mathbf{y}), \qquad \mathbf{y} \in \mathbb{R}^n, \end{aligned} \tag{1}$$

[1] Here I mean one that has no center manifold, and the function f is sufficiently smooth. See [2] for more details.

[2] This special form isolates the linear and nonlinear components. This sort of separation was necessary to perform within their work, as it was required that the linear components somehow dominate the nonlinear components.

where $\mathcal{F}(0,0) = 0$ and $\mathcal{G}(0,0) = 0$, and $\mathcal{A}$, $\mathcal{B}$ are matrices, is given by the fixed point of the operator

$$\mathcal{T}(\mathbf{x},\mathbf{y}) = \left(e^{t\mathcal{A}}\mathbf{x}_0 + \int_0^t e^{(t-s)\mathcal{A}}\mathcal{F}(\mathbf{x},\mathbf{y})\,ds\,, - \int_t^\infty e^{(t-s)\mathcal{B}}\mathcal{G}(\mathbf{x},\mathbf{y})\,ds \right), \tag{2}$$

on the appropriate Banach space (see [1]) This is called the *Lyapunov-Perron* operator [1–5]. The stable manifold is then unique and global if $\mathcal{T}$ is a contraction mapping.

Theorem 3. [1]. *Given the dynamical system (1) If*

1. *The linear components satisfy an exponential dichotomy:*

$$\left\|e^{t\mathcal{A}}\right\| \le e^{-\alpha t}\,, \quad \left\|e^{t\mathcal{B}}\right\| \le e^{-\beta t}\,,$$

   where $\alpha, \beta \in \mathbb{R}$, and $\|\cdot\|$ is an appropriate matrix norm.

2. *The nonlinear components are globally Lipschitz:*

$$\left\| \begin{pmatrix} \mathcal{F}(\mathbf{x}_1,\mathbf{y}_1) \\ \mathcal{G}(\mathbf{x}_1,\mathbf{y}_1) \end{pmatrix} - \begin{pmatrix} \mathcal{F}(\mathbf{x}_2,\mathbf{y}_2) \\ \mathcal{G}(\mathbf{x}_2,\mathbf{y}_2) \end{pmatrix} \right\| \le \delta \left\| \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{pmatrix} - \begin{pmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{pmatrix} \right\|,$$

   where $\delta > 0$.

3. *the spectral gap condition holds:*

$$\beta + \alpha > 2\delta\,.$$

Then the operator $\mathcal{T}$ is a contraction mapping, and thus has a unique fixed point.

1c. Survey of numerical methods

Robinson in [6] provides a general justification for using numerical schemes for the computation of (un)stable manifolds. The numerical methods I will present here phrase the stable manifold of (1) as a boundary value problem on the same set of differential equations. This allows us to apply the methods used in solutions of boundary value problems to compute the stable manifold. This significantly eases the computation.

For instance, previous methods, such as those presented in [3, 5] discretize operators similar to (2) and compute the inertial manifold[3] from there. These algorithms were generalized and improved in [4]. Older methods, such as [7], applied the use of Galerkin methods for computation. Alternatively, the work in [8] demonstrates that the problem of stable manifold computation may be reduced to the solution of a system of partial differential equations.

Our methods are significantly simpler to implement.[4] Since they implement a boundary value problem on the original system of differential equations, there is minimal additional

[3] In hyperbolic systems which satisfy the conditions in Theorem 3, the inertial manifold is the unstable manifold [4].

[4] See Appendix C for a sample implementation of the algorithms.

work to be done in implementation. The methods themselves are similar to waveform relaxation methods [9], and are likewise highly parallelizable, suggesting that they may ultimately be superior in speed as well.

1d. A note on the global Lipschitz condition

A lot of the work I present here requires functions to be globally Lipschitz. Such a strict condition may have the reader thinking that such work would be very difficult to apply to real problems. However, we can note that in real world applications, only a finite region of space is of interest. For the purposes of such an application, we can then design functions to be Lipschitz by introducing a smooth fall off to zero away from the area of interest. In this sense, the algorithms presented here would be self-validating: if given that the stable manifold stays within the locally Lipschitz region, then for that given application a smooth fall off can be used to force the function to be globally Lipschitz. Such methods are used in [7].

2. The Numerical Methods & Their Properties

Once again, we consider differential equations in the form

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \mathcal{A}\mathbf{x} + \mathcal{F}(\mathbf{x},\mathbf{y}) \equiv f(\mathbf{x},\mathbf{y})\,, \\ \frac{d\mathbf{y}}{dt} &= \mathcal{B}\mathbf{y} + \mathcal{G}(\mathbf{x},\mathbf{y}) \equiv g(\mathbf{x},\mathbf{y})\,, \end{aligned} \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$. Further, we assume that the stable manifold of (3) is a fixed point of the Lyapunov-Perron operator (2), and that all conditions of Theorem 3 are satisfied. Then,

$$\begin{aligned} \mathbf{x}(t) &= e^{t\mathcal{A}}\mathbf{x}_0 + \int_0^t e^{(t-s)\mathcal{A}}\mathcal{F}(\mathbf{x},\mathbf{y})\,ds\,, \\ \mathbf{y}(t) &= -\int_t^\infty e^{(t-s)\mathcal{B}}\mathcal{G}(\mathbf{x},\mathbf{y})\,ds\,. \end{aligned}$$

Differentiation with respect to t then recovers the initial set of differential equation, but subject to the boundary conditions.

$$\mathbf{x}(0) = \mathbf{x}_0\,, \quad \mathbf{y}(\infty) = 0\,. \tag{4}$$

Hence the stable manifold is a boundary value problem on an infinite interval. Note: the unstable manifold would be instead subject to the boundary conditions

$$\mathbf{x}(-\infty) = 0\,, \quad \mathbf{y}(0) = \mathbf{y}_0\,.$$

The boundary value formulation (4) reduces the question of finding the stable manifold to finding the corresponding $\mathbf{y}_0$ for a given $\mathbf{x}_0$, as we can trace the remainder of the trajectory using standard methods for solving differential equations. In particular we are interested in writing

$$\mathbf{y}_0 = \Phi(\mathbf{x}_0)\,.$$

Going back to Example 2, we could derive the same result by assuming some fixed $\mathbf{x}_0$ and then finding the $\mathbf{y}_0$ which satisfies the condition $\mathbf{y}(\infty) = 0$. One may also note that the stable manifold satisfies the boundary conditions $\mathbf{x}(\infty) = 0$ and $\mathbf{y}(0) = \mathbf{y}_0$. Though this is true, the exponential dichotomy condition prefers $\mathbf{x}$ as the stable direction, meaning it is possible to construct an example in which there are multiple trajectories which satisfy these boundary conditions, with only one being the stable manifold. However, if the conditions of Theorem 3 hold, there will only be one trajectory which satisfies (4).

The first key step to developing a numerical algorithm is to truncate the interval of the boundary value problem to a finite one. That is to say, we consider solutions of (3) which satisfy

$$\mathbf{x}(0) = \mathbf{x}_0\,, \quad \mathbf{y}(T) = 0 \tag{5}$$

where T is some sufficiently large number. Given the form of this boundary value problem, it stands to reason that we should iteratively integrate $\mathbf{x}$ forward in t, and $\mathbf{y}$ backward in t. We call this the *forward-backward* method.[1] Then, given some initial grid[2] $\mathbf{x}^0$, $\mathbf{y}^0$, we can write down the Euler method with a Jacobi update scheme:

$$\begin{aligned} \mathbf{x}^{i+1}_{j+1} &= \mathbf{x}^i_j + hf(\mathbf{x}^i_j, \mathbf{y}^i_j)\,, \quad \mathbf{x}^i_0 = \mathbf{x}_0\,, \\ \mathbf{y}^{i+1}_{j-1} &= \mathbf{y}^i_j - hg(\mathbf{x}^i_j, \mathbf{y}^i_j)\,, \quad \mathbf{y}^i_N = 0\,. \end{aligned} \tag{6}$$

As mentioned before, the superscripts refer to the iteration of the method, and the subscripts refer to the elements of $\mathbf{x}$ and $\mathbf{y}$. I have also introduced the total number of grid points, N.

I have referred to (6) as a **Jacobi** scheme, as it relies exclusively on the previous iteration of the method in its update. It may instead seem like a good idea to include information of $\mathbf{x}^{i+1}_j$ and $\mathbf{y}^{i+1}_j$ into the generation of $\mathbf{x}^{i+1}_{j+1}$ and $\mathbf{y}^{i+1}_{j-1}$, respectively, as that information is available to us. This idea leads to the **Gauss-Seidel** update scheme:

$$\begin{aligned} \mathbf{x}^{i+1}_{j+1} &= \mathbf{x}^{i+1}_j + hf(\mathbf{x}^{i+1}_j, \mathbf{y}^i_j)\,, \quad \mathbf{x}^i_0 = \mathbf{x}_0\,, \\ \mathbf{y}^{i+1}_{j-1} &= \mathbf{y}^{i+1}_j - hg(\mathbf{x}^i_j, \mathbf{y}^{i+1}_j)\,, \quad \mathbf{y}^i_N = 0\,. \end{aligned} \tag{7}$$

It is also possible to create more exotic update schemes; however I will limit the discussion here to the Jacobi (6) and Gauss-Seidel (7) schemes.

Hitherto, I have only mentioned variations of the Euler method. However, there is no reason for us to restrict ourselves to the Euler method. For instance a **Runge-Kutta order 4** (RK4) with the Jacobi update scheme can be written as

$$\begin{aligned} \xi_1 &= hf\left(x^i_j, y^i_j\right)\,, & \upsilon_1 &= hg\left(x^i_j, y^i_j\right)\,, \\ \xi_2 &= hf\left(x^i_j + \frac{\xi_1}{2}, y^i_j + \frac{\upsilon_1}{2}\right)\,, & \upsilon_2 &= hg\left(x^i_j + \frac{\xi_1}{2}, y^i_j + \frac{\upsilon_1}{2}\right)\,, \\ \xi_3 &= hf\left(x^i_j + \frac{\xi_2}{2}, y^i_j + \frac{\upsilon_2}{2}\right)\,, & \upsilon_3 &= hg\left(x^i_j + \frac{\xi_2}{2}, y^i_j + \frac{\upsilon_2}{2}\right)\,, \\ \xi_4 &= hf\left(x^i_j + \xi_3, y^i_j + \upsilon_3\right)\,, \end{aligned}$$

$$x^{i+1}_{j+1} = x^i_j + \frac{1}{6}\left(\xi_1 + 2\xi_2 + 2\xi_3 + \xi_4\right)\,.$$

[1] Similar forward-backward methods were used in [4], though they relied on direct discretization of the Lyapunov-Perron operator.

[2] It typically makes sense to initialize the $\mathbf{x}$ grid as an array of $\mathbf{x}_0$ values, and the $\mathbf{y}$ grid as an array of zeros.

$$\begin{aligned}
\xi_1 &= hf\left(x_j^i, y_j^i\right), & \upsilon_1 &= hg\left(x_j^i, y_j^i\right), \\
\xi_2 &= hf\left(x_j^i - \frac{\xi_1}{2}, y_j^i - \frac{\upsilon_1}{2}\right), & \upsilon_2 &= hg\left(x_j^i - \frac{\xi_1}{2}, y_j^i - \frac{\upsilon_1}{2}\right), \\
\xi_3 &= hf\left(x_j^i - \frac{\xi_2}{2}, y_j^i - \frac{\upsilon_2}{2}\right), & \upsilon_3 &= hg\left(x_j^i - \frac{\xi_2}{2}, y_j^i - \frac{\upsilon_2}{2}\right), \\
& & \upsilon_4 &= hg\left(x_j^i - \xi_3, y_j^i - \upsilon_3\right),
\end{aligned}$$

$$y_{j-1}^{i+1} = y_j^i - \frac{1}{6}\left(\upsilon_1 + 2\upsilon_2 + 2\upsilon_3 + \upsilon_4\right).$$

Other order finite difference methods, such as linear multistep methods, may be developed in a similar manner.

2a. Errors due to truncation

In the implementation of the method, we must pick some cutoff time T that serves as the endpoint of the time grid. Since the analytical boundary condition occurs at $t = \infty$, creating a cutoff time inherently creates a error in our calculation of the manifold. This error can be described explicitly for a two-dimensional linear system.

Theorem 4. *Let $y(0) = y_0$ be the initial condition for the two-dimensional boundary value problem*

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{A}\begin{pmatrix} x \\ y \end{pmatrix}, \quad x(0) = x_0, , \quad y(\infty) = 0\,, \tag{8}$$

where $\mathbf{A}$ has spectrum $\{\lambda_-, \lambda_+\}$ where $\lambda_- < 0 < \lambda_+$. Let $b(0) = b_0$ be the initial condition for

$$\frac{d}{dt}\begin{pmatrix} a \\ b \end{pmatrix} = \mathbf{A}\begin{pmatrix} a \\ b \end{pmatrix}, \quad a(0) = x_0, , \quad b(T) = 0\,. \tag{9}$$

Then the error between y_0 and b_0 is given by

$$|y_0 - b_0| = \left|\frac{\xi e^{\lambda_- T}}{\eta e^{\lambda_+ T} - \gamma e^{\lambda_- T}}\right|, \tag{10}$$

where ξ, η, and γ are constants independent of T.

Proof. Recall that the general solution of the linear system

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{A}\begin{pmatrix} x \\ y \end{pmatrix},$$

is given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = c_+\mathbf{v}_+ e^{\lambda_+ t} + c_-\mathbf{v}_- e^{\lambda_- t},$$

where $\mathbf{v}_\pm$ denotes eigenvector of $\mathbf{A}$ associated with $\lambda_\pm$, and c_1, c_2 are constants denoted by

the boundary conditions. A little bit of algebra yields the solutions to (8) and (9):

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{x_0}{v_{-,x}} \mathbf{v}_- e^{\lambda_- t} ,$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \left(\frac{x_0 v_{+,y} e^{\lambda_+ T}}{v_{-,x} v_{+,y} e^{\lambda_+ T} - v_{-,y} v_{+,x} e^{\lambda_- T}} \right) \mathbf{v}_- e^{\lambda_- t} - \left(\frac{x_0 v_{-,y} e^{\lambda_- T}}{v_{-,x} v_{+,y} e^{\lambda_+ T} - v_{-,y} v_{+,x} e^{\lambda_- T}} \right) \mathbf{v}_+ e^{\lambda_+ t} .$$

Some algebra then yields:

$$y_0 = \frac{x_0 v_{-,y}}{v_{-,x}} .$$

$$b_0 = \frac{x_0 v_{-,y} v_{+,y} \left(e^{\lambda_+ T} - e^{\lambda_- T} \right)}{v_{-,x} v_{+,y} e^{\lambda_+ T} - v_{-,y} v_{+,x} e^{\lambda_- T}} .$$

As expected, $b_0 \to y_0$ when $T \to \infty$. It then follows that

$$|y_0 - b_0| = \left| \frac{x_0 v_{-,y} \left(v_{+,y} - \frac{v_{-,y} v_{+,x}}{v_{-,x}} \right) e^{\lambda_- T}}{v_{-,x} v_{+,y} e^{\lambda_+ T} - v_{-,y} v_{+,x} e^{\lambda_- T}} \right| .$$

□

The difficulty in extending this proof to higher dimensions lies in the fact that it is no longer trivial to write the explicit forms of y_0 and b_0, as the components of the eigenvectors are no longer scalars. Certainly, a closed form must be possible, but would likely be considerably more difficult to write down. We will see in Section 2d that the error due to truncation appears to follow the pattern in (10) in higher dimensional and nonlinear systems.

2b. Matrix representations and stability

As usual in stability analysis [10], I will consider the stability of these methods on linear systems. This sort of analysis is referred to as A-stability analysis and gives the stability of schemes on exponentially decaying "stiff" systems. This is particularly relevant to us, as the **x**-direction is stiff forward in time, and the **y**-direction is stiff backwards in time.

The analysis here is similar to the stability analysis of waveform relaxation methods in [9]. However, the forward-backward nature of the scheme introduces additional subtleties in the stability making the results of [9] inapplicable.

Definition 5 (Stability)**.** A method is called *stable* for some linear system $\dot{\mathbf{z}} = \mathbf{M}\mathbf{z}$ if when the the numerical scheme is written as $\mathbf{z}^{i+1} = \mathbf{A}\mathbf{z}^i$, the spectral radius of $\mathbf{A}$ is less than or equal to 1, implying that

$$\lim_{i \to \infty} \mathbf{z}^{\mathbf{i}}$$

converges.

Proposition 6 (Jacobi)**.** *Given a linear system of equations*

$$\frac{d}{dt}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \mathcal{M}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix},$$

where $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$, *and* $\mathcal{M}$ *is a* $m+n$ *square matrix, then the iterations of the Jacobi update scheme for any one step method can be written as*

$$\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^{i+1} = \left(\begin{array}{ccc|ccc} I_{m\times m} & & & B_{m\times n} & & \\ A_{m\times m} & \ddots & & & \ddots & \\ & \ddots & & & \ddots & \\ \hline & \ddots & & & \ddots & \\ & & C_{n\times m} & & & D_{n\times n} \\ & & & & & I_{n\times n} \end{array}\right) \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^{i}, \tag{11}$$

Subscripts on matrices are used here to denote their size. In each quadrant, the dimensions of the blocks are the same.

Proof. On a linear system, the Jabobi update for a one step method may be written as

$$\begin{aligned}\mathbf{x}_{j+1}^{i+1} &= A_{m\times m}\mathbf{x}_j^i + B_{m\times n}\mathbf{y}_j^{,},\\ \mathbf{y}_{j-1}^{i+1} &= C_{n\times m}\mathbf{x}_j^i + D_{n\times n}\mathbf{y}_j^i\,,\end{aligned}$$

where A, B, C, and D are determined by $\mathcal{M}$, the step size, and numerical method. Noting that the initial set of points in $\mathbf{x}$ and the ending points in $\mathbf{y}$ are fixed as per the condition in (5), the result of (11) follows. □

Theorem 7. [11]. *Let* $a, b, c, d > 0$ *and*

$$M = \begin{pmatrix} aJ_n & bI_n \\ cI_n & dJ_n^T \end{pmatrix}, \tag{12}$$

where J_n *is an* n*-dimensional Jordan block.*[3] *Then the spectral radius of* M *satisfies the relation*

$$\rho(M) \leq \sqrt{ad} + \sqrt{bc}\,. \tag{13}$$

The requirements on the positivity of a, b, c, d seem strict, but we shall see that these come out naturally in linear systems which satisfy both exponential dichotomy, and the spectral gap condition. At first, it may appear that that matrix in (11) does not have the right form.

[3]See [12] for a discussion.

However, we can note that the matrix in (11) is block lower triangular:

$$\left(\begin{array}{c|ccccccc} I_{m\times m} & & & & & & & \\ \hline A_{m\times m} & & & & B_{m\times n} & & & \\ & A_{m\times m} & & & & \ddots & & \\ & & \ddots & & & & \ddots & \\ & \ddots & & & & \ddots & & \\ & & \ddots & & & & \ddots & \\ & & & C_{m\times m} & & & & D_{n\times n} \\ & & & & & & & I_{n\times n} \end{array}\right),$$

meaning that n of its eigenvalues are equal to one, and the rest are equal to the eigenvalues of the lower block. But, this block is upper block triangular:

$$\left(\begin{array}{ccccccc|c} & & & & B_{m\times n} & & & \\ A_{m\times m} & & & & & \ddots & & \\ & \ddots & & & & & \ddots & \\ \ddots & & & & & \ddots & & \\ & \ddots & & & & & \ddots & \\ & & C_{n\times m} & & & & & D_{n\times n} \\ \hline & & & & & & & I_{n\times n} \end{array}\right),$$

meaning that m of the eigenvalues are also equal to one. Therefore, the question of stability of the Jacobi method reduces to the question of finding the spectral radius of the matrix

$$\left(\begin{array}{ccc|ccc} & & & B_{m\times n} & & \\ A_{m\times m} & & & & \ddots & \\ & \ddots & & & & \ddots \\ \hline \ddots & & & & \ddots & \\ & \ddots & & & & D_{n\times n} \\ & & C_{n\times m} & & & \end{array}\right).$$

This matrix now reassembles the form of the one in theorem 7.

Example 8 (The Jacobi Euler scheme)**.** The Jacobi Euler scheme, given in (6), for the linear system

$$\frac{d}{dt}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}, \tag{14}$$

may be written as

$$\begin{aligned} \mathbf{x}_{j+1}^{i+1} &= (I + hM_{11})\mathbf{x}_j^i + hM_{12}\mathbf{y}_j^i\,, \\ \mathbf{y}_{j+1}^{i+1} &= -hM_{21}\mathbf{x}_{\mathbf{j}}^{\mathbf{i}} + (I - hM_{22})\mathbf{y}_j^i\,. \end{aligned}$$

The full matrix form of this scheme may be written by following the proof of Proposition 6. However, in order to fit the format of Theorem 13, let us consider the case where $\dim \mathbf{x} = \dim \mathbf{y} = 1$; and $M_{11} = -A$, $M_{22} = B$, and $M_{12} = -M_{21} = \delta$, where $A, B, \delta > 0$, and $A + B > 2\delta$. The matrix form of the scheme can then be written as

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^{i+1} = \left(\begin{array}{cccc|cccc} 1 & & & & & & & \\ 1-hA & \ddots & & & h\delta & \ddots & & \\ \ddots & \ddots & & & \ddots & \ddots & & \\ & \ddots & & & & \ddots & & \\ \hline & & \ddots & & & & \ddots & \\ & & \ddots & & & & \ddots & \\ & & & h\delta & & & & 1-hB \\ & & & & & & & 1 \end{array}\right) \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^{i}$$

The relevant spectral radius is that of the inner block, i.e. excluding the boundary points. Following the Theorem 7, that spectral radius is bounded by $\rho \leq \sqrt{(1-hA)(1-hB)} + h\delta$. Our ultimate goal is to find a value of h such that $\rho < 1$. Some algebra yields the condition

$$h(AB - \delta^2) - A - B + 2\delta < 0$$

We then need to consider a few cases:

1. If the gap condition holds and $AB \leq \delta^2$, then the scheme will converge for all $h > 0$.
2. If the gap condition holds and $AB > \delta^2$, then we need
$$h < \frac{A + B - 2\delta}{AB - \delta^2}\,. \tag{15}$$
3. If the gap condition does not hold then the scheme will never converge.

Example 9 (Jacobi Runge-Kutta order 2)**.** Recall that for the differential equation $\dot{x} = f(x)$ the scheme

$$x_{j+1} = x_j + hf\left(x_j + \frac{h}{2}f(x_j)\right)\,,$$

has order 2 accuracy. Then for (14), the Jacobi scheme is given by

$$\mathbf{x}_{j+1}^{i+1} = \mathbf{x}_j^i + h\left(M_{11}\left(\mathbf{x}_j^i + \frac{h}{2}M_{11}\mathbf{x}_j^i + \frac{h}{2}M_{12}\mathbf{y}_j^i\right) + M_{12}\left(\mathbf{y}_j^i + \frac{h}{2}M_{22}\mathbf{y}_j^i + \frac{h}{2}M_{21}\mathbf{x}_j^i\right)\right)\,,$$
$$\mathbf{y}_{j-1}^{i+1} = \mathbf{y}_\mathbf{j}^\mathbf{i} - h\left(M_{21}\left(\mathbf{x}_j^i - \frac{h}{2}M_{11}\mathbf{x}_j^i - \frac{h}{2}M_{12}\mathbf{y}_j^i\right) + M_{22}\left(\mathbf{y}_j^i - \frac{h}{2}M_{22}\mathbf{y}_j^i - \frac{h}{2}M_{21}\mathbf{x}_j^i\right)\right)\,.$$

Reorganising like terms, we have

$$\mathbf{x}_{j+1}^{i+1} = \left(I + hM_{11} + \frac{h^2}{2}M_{11}^2 + \frac{h^2}{2}M_{12}M_{21}\right)\mathbf{x}_j^i + \left(hM_{12} + \frac{h^2}{2}M_{12}M_{22} + \frac{h^2}{2}M_{11}M_{12}\right)\mathbf{y}_j^i\,,$$
$$\mathbf{y}_{j-1}^{i+1} = \left(-hM_{21} + \frac{h^2}{2}M_{21}M_{11} + \frac{h^2}{2}M_{22}M_{21}\right)\mathbf{x}_j^i + \left(I - hM_{22} + \frac{h^2}{2}M_{22}^2 + \frac{h^2}{2}M_{21}M_{12}\right)\mathbf{y}_j^i\,.$$

Once again, we will consider the case where $\dim \mathbf{x} = \dim \mathbf{y} = 1$; and $M_{11} = -A$, $M_{22} = B$, and $M_{12} = -M_{21} = \delta$, where $A, B, \delta > 0$, and $A + B > 2\delta$. The schemes then become

$$x_{j+1}^{i+1} = \left(1 - hA + \frac{h^2}{2}A^2 - \frac{h^2}{2}\delta^2\right)x_j^i + \left(h\delta + \frac{h^2}{2}B\delta - \frac{h^2}{2}A\delta\right)y_j^i\,,$$
$$y_{j-1}^{i+1} = \left(h\delta + \frac{h^2}{2}A\delta - \frac{h^2}{2}B\delta\right)x_j^i + \left(1 - hB + \frac{h^2}{2}B^2 - \frac{h^2}{2}\delta^2\right)y_j^i\,.$$

We can use Theorem 7 to derive the stability condition:

$$\sqrt{\left(1 - hA + \frac{h^2}{2}A^2 - \frac{h^2}{2}\delta^2\right)\left(1 - hB + \frac{h^2}{2}B^2 - \frac{h^2}{2}\delta^2\right)}$$
$$+\sqrt{\left(h\delta + \frac{h^2}{2}B\delta - \frac{h^2}{2}A\delta\right)\left(h\delta + \frac{h^2}{2}A\delta - \frac{h^2}{2}B\delta\right)} \leq 1\,.$$

Unfortunately, these do not simplify nicely. However, we can consider the special case when $A = B$. Then we have

$$1 - hA + \frac{h^2}{2}\left(A^2 - \delta^2\right) + h\delta \leq 1\,,$$

and some algebra recovers the conditions of the Euler method:

$$A^2 - \delta^2 > 0\,, \quad h < 2\frac{A - \delta}{A^2 - \delta^2}\,.$$

Proposition 10 (Gauss-Seidel)**.** *Given a linear system of equations*

$$\frac{d}{dt}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \mathcal{M}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix},$$

where $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$, and $\mathcal{M}$ is a $m + n$ square matrix, then the iterations of the Gauss-

Seidel update scheme for any one step method can be written as

$$\left(\begin{array}{cc|cc} I_{m\times m} & & & \\ -A_{m\times m} & I_{m\times m} & & \\ \ddots & \ddots & & \\ & \ddots & \ddots & \\ \hline & & \ddots & \ddots \\ & & \ddots & \ddots \\ & & I_{n\times n} & -D_{n\times n} \\ & & & I_{n\times n} \end{array}\right) \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^{i+1} = \left(\begin{array}{cc|cc} & & 0_{m\times n} & \\ & & B_{m\times n} & \\ & & & \ddots \\ & & & \ddots \\ \hline \ddots & & & \\ & \ddots & & \\ & & C_{n\times m} & \\ & & 0_{n\times m} & \end{array}\right) \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^{i}. \tag{16}$$

Subscripts on matrices are used here to denote their size. In each quadrant, the dimensions of the blocks are the same.

Proof. On a linear system, the Gauss-Seidel scheme can be written as

$$\begin{aligned} \mathbf{x}_{j+1}^{i+1} &= A_{m\times m}\mathbf{x}_j^{i+1} + B_{m\times n}\mathbf{y}_j^i\,, \\ \mathbf{y}_{j-1}^{i+1} &= C_{n\times m}\mathbf{x}_j^i + D_{n\times n}\mathbf{y}_j^{i+1}\,. \end{aligned} \tag{17}$$

Moving like super-indices to each side, we can write (17) in the matrix form of (16)

□

Theorem 11. [11]. *Suppose* $L = \frac{1}{d}(I_n - aJ_n)$ *and* $U = \frac{1}{c}(I_n - bJ_n^T)$ *where* $a, b, c, d > 0$ *and* $ab \in (0, 1)$. *Then the spectral radius*

$$\rho(LU) > \frac{1 - ba}{cd}\,.$$

Example 12 (Gauss-Seidel Euler scheme)**.** Using the same example as in Example 8, we can write down the Gauss-Seidel Euler scheme as

$$\begin{aligned} x_{j+1}^{i+1} &= (1 - hA)\, x_j^{i+1} + h\delta y_j^i\,, \\ y_{j-1}^{i+1} &= h\delta x_j^i + (1 - hB)\, y_j^{i+1}\,. \end{aligned}$$

Though we can write this in the form of (16), but in this case it is possible to do better.

This scheme can be rewritten as

$$x_j^i = \frac{1}{h\delta} y_{j-1}^{i+1} + \frac{hB-1}{h\delta} y_j^{i+1}\,,$$
$$y_j^i = \frac{hA-1}{h\delta} x_j^{i+1} + \frac{1}{h\delta} x_{j+1}^{i+1}\,.$$

Therefore, the iteration can be written as

$$\frac{1}{h\delta}\left(\begin{array}{cccc|cccc} h\delta & & & & 1 & hB-1 & & \\ & & & & & \ddots & \ddots & \\ & & & & & & \ddots & \ddots \\ \hline \ddots & \ddots & & & & & & \\ & \ddots & \ddots & & & & & \\ & & hA-1 & 1 & & & & \\ & & & & & & & h\delta \end{array}\right)\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^{i+1} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^{i}.$$

In this case, we are interested in the spectral radius of the *inverse* of the matrix we see on the lefthand side. Hence, we want the spectral radius of that matrix to be less than or equal to one. Following the arguments we used for the Euler method, we can reduce the matrix of interest to

$$\frac{1}{h\delta}\left(\begin{array}{cccc|cccc} & & & & 1 & hB-1 & & \\ & & & & & \ddots & \ddots & \\ & & & & & & \ddots & \\ \hline \ddots & & & & & & & \\ \ddots & \ddots & & & & & & \\ & hA-1 & 1 & & & & & \end{array}\right).$$

Notice that this matrix is of the form

$$\frac{1}{h\delta}\left(\begin{array}{c|c} & U \\ \hline L & \end{array}\right).$$

Squaring it, we obtain a block-diagonal matrix with each block in the form of Theorem 11. Hence, Since the iteration is really the inverse of this matrix, we want the spectral radius to be greater than one. Thus we want

$$\frac{1-(1-hB)(1-hA)}{h^2\delta^2} > 1\,.$$

Some algebra then yields that we need

$$h < \frac{A+B}{AB^2+\delta^2}\,.$$

Note that this condition is more strict than the Jacobi method. In general, it appears that the Gauss-Seidel scheme offers between convergence, but with the tradeoff of worse stability.

2c. Errors in the Euler method, first approach

In this section, I present a method to derive the error for the Euler method, given stringent conditions on the nature of our dynamical system. This approach follows the methodology in [10], which presents the traditional approach for calculating the error term for the Euler method. A more general method for finding error will be given in the next chapter.

It is first important to note that for the Euler method, all update schemes converge to the fixed point $(\mathbf{a}, \mathbf{b})^T$, satisfying

$$\begin{aligned} &1. \quad \mathbf{a}_0 = \mathbf{x}_0\,, \\ &2. \quad \mathbf{b}_N = 0\,, \\ &3. \quad \mathbf{a}_{j+1} = \mathbf{a}_j + hf(\mathbf{a}_j, \mathbf{b}_j)\,, \quad j \in \mathbb{N} \cap [0, N-1]\,, \\ &4. \quad \mathbf{b}_{j-1} = \mathbf{b}_j - hg(\mathbf{a}_j, \mathbf{b}_j)\,, \quad j \in \mathbb{N} \cap [1, N]\,. \end{aligned} \tag{18}$$

Theorem 13. *Let $\mathbf{x}, \mathbf{y}$ be Taylor expandable functions which satisfy the boundary value problem*

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \mathcal{A}\mathbf{x} + \mathcal{F}(\mathbf{x}, \mathbf{y})\,, \qquad \mathbf{x}(0) = \mathbf{x}_0\,, \\ \frac{d\mathbf{y}}{dt} &= \mathcal{B}\mathbf{y} + \mathcal{G}(\mathbf{x}, \mathbf{y})\,, \qquad \mathbf{y}(T) = 0\,, \end{aligned}$$

with

1. *For all $h < h_0$, under some induced matrix norm $|||\cdot|||$, $|||I + h\mathcal{A}||| < 1 - h\alpha$ and $|||I - h\mathcal{B}||| < 1 - h\beta$ for some suitable positive constants α, β.*

2. *There exist constants L_F and L_G such that*

$$\begin{aligned} \|\mathcal{F}(\mathbf{x}_1, \mathbf{y}_1) - \mathcal{F}(\mathbf{x}_2, \mathbf{y}_2)\| &\leq L_{\mathcal{F}} \left(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|\right), \\ \|\mathcal{G}(\mathbf{x}_1, \mathbf{y}_1) - \mathcal{G}(\mathbf{x}_2, \mathbf{y}_2)\| &\leq L_{\mathcal{G}} \left(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|\right), \end{aligned} \tag{19}$$

   for all choices of $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2$.

3. *There exists some γ such that $\alpha - L_{\mathcal{F}} \geq \gamma > 0$ and $\beta - L_{\mathcal{G}} \geq \gamma > 0$.*

4. *$L_{\mathcal{F}} L_{\mathcal{G}} \leq \gamma^2$.*

5. *We are in a regime where $h\gamma < 1$.*

Let a, b satisfy the conditions of (18). Then

$$\begin{aligned} \|\mathbf{x} - \mathbf{a}\|_\infty &\leq h \frac{L_{\mathcal{F}}\eta + \gamma\xi}{2(\gamma^2 - L_{\mathcal{F}} L_{\mathcal{G}})}\,, \\ \|\mathbf{y} - \mathbf{b}\|_\infty &\leq h \frac{L_{\mathcal{G}}\xi + \gamma\eta}{2(\gamma^2 - L_{\mathcal{F}} L_{\mathcal{G}})}\,, \end{aligned}$$

where η and ξ are positive constants.

Proof. Since $\mathbf{x}$ and $\mathbf{y}$ are Taylor expandable we can write

$$\begin{aligned} \mathbf{x}_{j+1} &= \mathbf{x}_j + h\mathcal{A}\mathbf{x}_j + h\mathcal{F}(\mathbf{x}_j, \mathbf{y}_j) + \frac{h^2\xi_j}{2}\,, \\ \mathbf{y}_{j-1} &= \mathbf{y}_j - h\mathcal{B}\mathbf{y}_j - h\mathcal{G}(\mathbf{x}_j, \mathbf{y}_j) + \frac{h^2\eta_j}{2}\,, \end{aligned} \tag{20}$$

where the subscript j denotes at time jh. For convenience, we define the error terms, ϵ and δ, by

$$\begin{aligned} \epsilon_j &= \|\mathbf{x}_j - \mathbf{a}_j\|\,, \\ \delta_j &= \|\mathbf{y}_j - \mathbf{b}_j\|\,. \end{aligned}$$

A substitution of (20) yields

$$\begin{aligned} \epsilon_{j+1} &\leq \epsilon_j + h\,\|\mathcal{A}(\mathbf{x}_j - \mathbf{a}_j) + \mathcal{F}(\mathbf{x}_j, \mathbf{y}_j) - \mathcal{F}(\mathbf{a}_j, \mathbf{b}_j)\| + \frac{h^2\xi}{2}\,, \\ \delta_{j-1} &\leq \delta_j + h\,\|\mathcal{B}(\mathbf{y}_j - \mathbf{b}_j) + \mathcal{G}(\mathbf{x}_j, \mathbf{y}_j) - \mathcal{G}(\mathbf{a}_j, \mathbf{b}_j)\| + \frac{h^2\eta}{2}\,, \end{aligned}$$

where we have defined $\xi = \max_j(\|\xi_j\|)$ and $\eta = \max_j(\|\eta_j\|)$. Applying the triangle inequality, as well as the first three conditions, it follows that

$$\begin{aligned} \epsilon_{j+1} &\leq (1 - h\gamma)\epsilon_j + hL_{\mathcal{F}}\delta_j + \frac{h^2\xi}{2} \\ &\leq (1 - h\gamma)\epsilon_j + hL_{\mathcal{F}}\delta_{\max} + \frac{h^2\xi}{2}\,. \end{aligned}$$

Likewise,

$$\delta_{j-1} \leq (1 - h\gamma)\delta_j + hL_{\mathcal{G}}\epsilon_{\max} + \frac{h^2\eta}{2}\,.$$

We can note that the expression for ϵ is equivalent to

$$\epsilon_{j+1} \leq (1 - h\gamma)^{j+1}\epsilon_0 + \left(hL_{\mathcal{F}}\delta_{\max} + \frac{h^2\xi}{2}\right)\sum_{n=0}^{j}(1 - h\gamma)^n\,.$$

The latter term is a geometric series, but we will sum to infinity as the extra terms are not useful in constructing the bounds. Furthermore, we note that by construction $\epsilon_0 = 0$. Hence

$$\epsilon_{j+1} \leq \frac{L_{\mathcal{F}}\delta_{\max}}{\gamma} + \frac{h\xi}{2\gamma}\,.$$

Since this is true for all ϵ_j, it is true for $\epsilon_{\max}$:

$$\epsilon_{\max} \leq \frac{L_{\mathcal{F}}\delta_{\max}}{\gamma} + \frac{h\xi}{2\gamma}\,.$$

Likewise, for δ, we find that

$$\delta_{\max} \leq \frac{L_{\mathcal{G}}\epsilon_{\max}}{\gamma} + \frac{h\eta}{2\gamma}\,.$$

Given the fourth condition, we find that

$$\epsilon_{\max} \leq h\frac{L_{\mathcal{F}}\eta + \gamma\xi}{2(\gamma^2 - L_{\mathcal{F}}L_{\mathcal{G}})}\,,$$
$$\delta_{\max} \leq h\frac{L_{\mathcal{G}}\xi + \gamma\eta}{2(\gamma^2 - L_{\mathcal{F}}L_{\mathcal{G}})}\,.$$

□

Interestingly, the fourth condition for this error, $L_{\mathcal{F}}L_{\mathcal{G}} < \gamma^2$, is analogous to the first part of the stability condition we found for the Jacobi scheme in (15).

2d. Test cases & numerical results

Though the theory is helpful, it is likewise useful to demonstrate the capabilities of these numerical schemes on a few test problems. In particular, I will consider two test problems: a linear problem and a nonlinear problem.

Example 14 (Linear)**.** In a linear system

$$\frac{d}{dt}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \mathcal{M}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\,,$$

for which the stable manifold satisfies the boundary value problem

$$\mathbf{x}(0) = \mathbf{x}_0\,, \quad \mathbf{y}(\infty) = 0\,,$$

the stable manifold in phase space is the span of the eigenvectors of $\mathcal{M}$ which have negative eigenvalues. In particular, I will consider the system

$$\frac{dx}{dt} = -x + \frac{1}{10}y\,, \quad \frac{dy}{dt} = -\frac{1}{10}x + y\,, \tag{21}$$

which has a stable manifold given by

$$y = \frac{\sqrt{11}}{33 + 10\sqrt{11}}x\,.$$

As mentioned before, the two relevant errors are with respect to the step size h, and the cutoff time T. Since the system is determined entirely by the initial condition $y(0)$, the natural error term to use would be $|y^*(0) - y(0)|$, where y^* denotes the analytical form of the stable manifold. Holding the step size fixed, the error with respect to the cutoff time is given in Figure 2 and the error with respect to the step size, holding the cutoff time fixed, is given in Figure 3. The asymptotic region in Figure 2 is where the error from the step

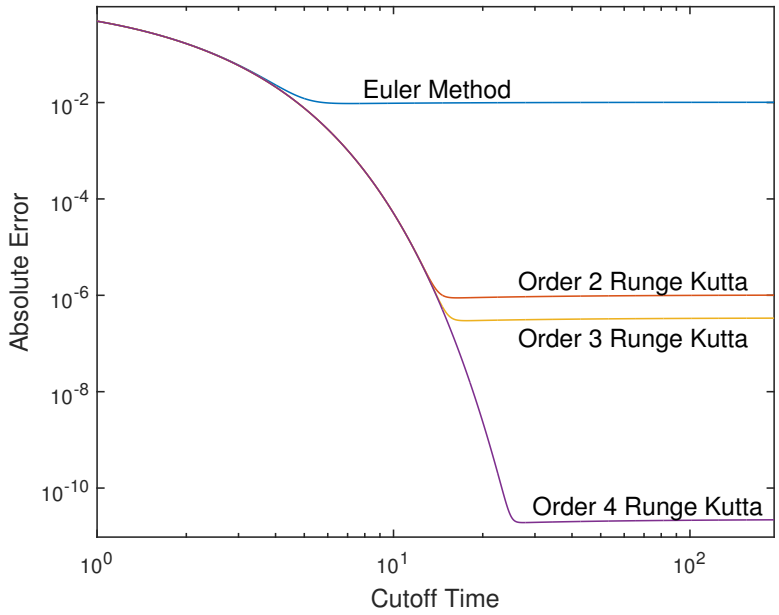

Figure 2. Error of various methods in computing the stable manifold for the linear test system (21). The step size was fixed at $h = 0.02$, the tolerance was 10^{-10}, and the initial condition was $x_0 = 4$.

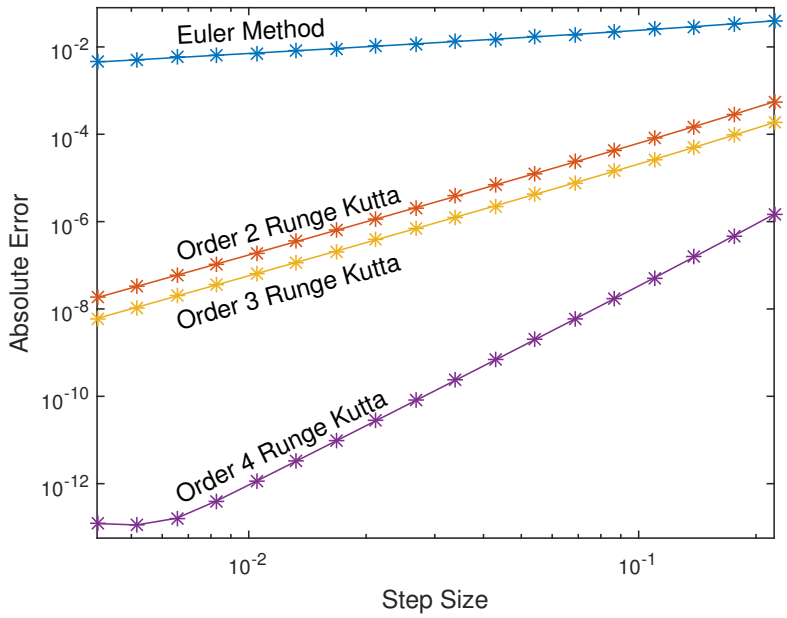

Figure 3. Error of various methods in computing the stable manifold for the linear test system (21). The cutoff time was fixed at $T = 80$, the tolerance was 10^{-12}, and the initial condition was $x_0 = 4$.

size dominates. We can note that the error due to step size for the order 2 Runge-Kutta scheme is actually order 3 in accuracy. This appears to be a property that the method has on *linear systems*. We will see in Example 15 that this is not necessarily the case for a nonlinear system. Though these calculations were carried out using the Jacobi method, the error curves would be identical for the Gauss-Seidel method, within the order of the tolerance.

Example 15 (Nonlinear)**.** The system

$$\begin{aligned} \frac{dx}{dt} &= -\frac{1+c_1c_2}{1-c_1c_2}x - \frac{-2c_1}{1-c_1c_2}y - \frac{c_1}{1-c_1c_2}\left((x+c_1y)\sin(c+c_1y) - k(x+c_1y)^3\right) , \\ \frac{dy}{dt} &= \frac{2c_2}{1-c_1c_2}x + \frac{1+c_1c_2}{1-c_1c_2}y + \frac{1}{1-c_1c_2}\left((x+c_1y)\sin(c+c_1y) - k(x+c_1y)^3\right) . \end{aligned} \tag{22}$$

has the implicitly defined stable manifold

$$\frac{k}{4}(x+c_1y)^3 + \cos(x+c_1y) - \frac{\sin(x+c_1y)}{x+c_1y} - c_2x - y = 0\,.$$

This sort of system can be computed by solving explicitly for a nonlinear system in which one of the variables is decoupled from the other, and then performing some linear transformation. See Appendix A for a full derivation. In this case c_1, c_2 are the parameters for that linear transformation, and k is a constant to keep the nonlinear terms small. This is the system depicted in Figure 1. The corresponding behaviour of the error can be seen in Figures 4 and 5. We can note that the error with respect to T behaves similarly to the linear case, with an extra sudden increase in accuracy near $T = 5$. This sort of behavior often appears in nonlinear systems, and is most likely due to the nature of the phase portrait of the system. For the error with respect to h, all methods have the expected order of accuracy.

Note: Though these algorithms were tested on two dimensional systems, they are valid for any hyperbolic system of dimension greater or equal to two. The purpose of the linear example was to draw parallels between Examples 8, 9, and 12; and it is difficult to create a solvable nonlinear example problem with a global stable manifold.

Example 16 (Convergence of methods)**.** We can consider the convergence of methods by plotting the tolerance $\|\mathbf{x}^i - \mathbf{x}^{i-1}\| + \|\mathbf{y}^i - \mathbf{y}^{i-1}\|$ versus the number of steps, i. These results can be seen in Figures 6 and 7. Clearly, the Gauss-Seidel schemes converge much faster. For reference, in the linear example it took the Jacobi scheme 2133 steps to reach the same tolerance that the Gauss-Seidel scheme reached in 20 steps, and in the nonlinear example it took the Jacobi scheme 2916 steps to reach the same tolerance that the Gauss-Seidel scheme reached in 60 steps. However, it would be incorrect to discount the Euler method, as it has much more predictable stability behaviour, as seen when comparing the results of Examples 8 and 12, as well as the general forms of Proposition 6 and 10.

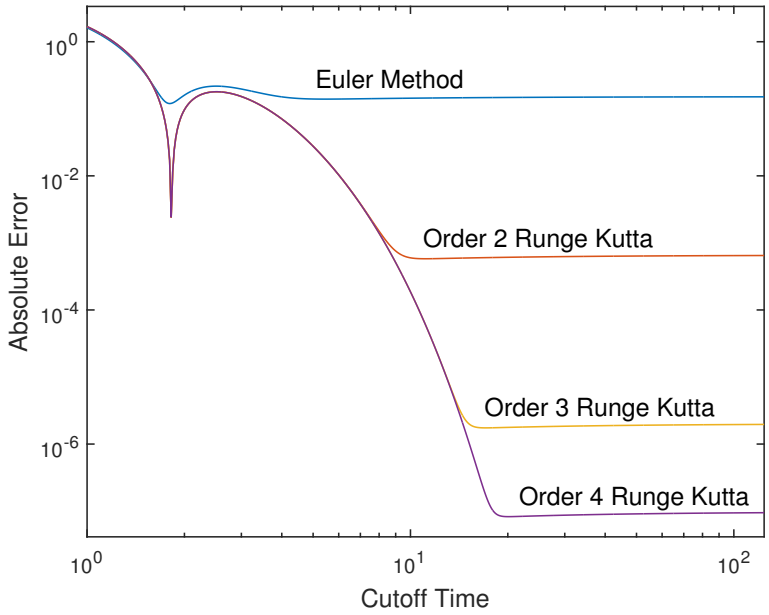

Figure 4. Error of various methods in computing the stable manifold for the nonlinear test system (22) with $k = 0.001$, $c_1 = -0.3$, $c_2 = -0.2$. The step size was fixed at $h = 0.02$, the tolerance was 10^{-10}, and the initial condition was $x_0 = 4$.

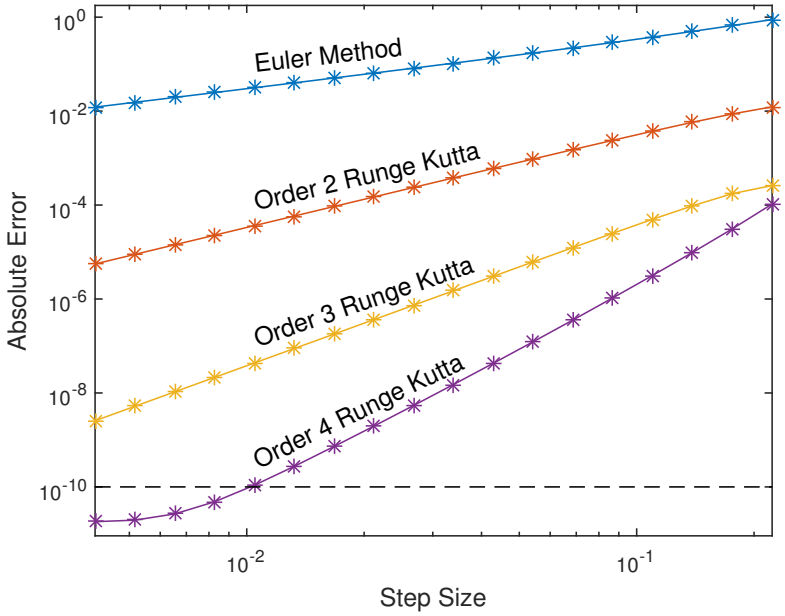

Figure 5. Error of various methods in computing the stable manifold for the nonlinear test system (22) with $k = 0.001$, $c_1 = -0.3$, $c_2 = -0.2$. The cutoff time was fixed at $T = 80$, the tolerance was 10^{-10}, and the initial condition was $x_0 = 4$. The dotted line is the tolerance of the method.

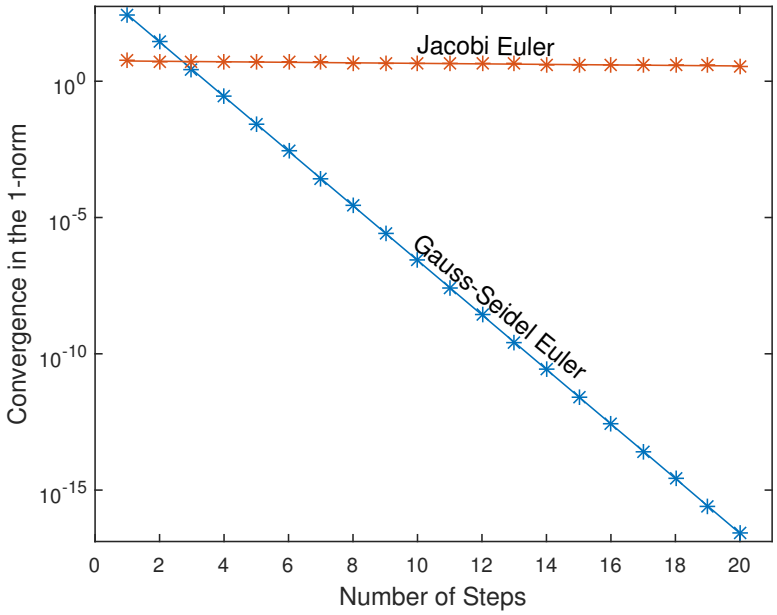

Figure 6. The rates of convergence of the Jacobi and Gauss-Seidel Euler schemes in computing the stable manifold for the linear test system (21). The step size was fixed at $h = 0.02$, the cutoff time at $T = 80$, and the initial condition was $x_0 = 4$.

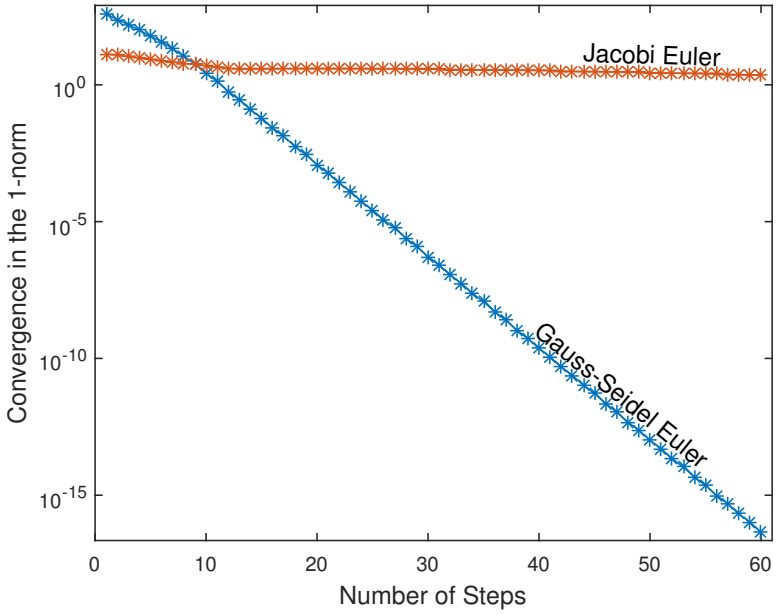

Figure 7. The rates of convergence of the Jacobi and Gauss-Seidel Euler schemes in computing the stable manifold for the nonlinear test system (22) with $k = 0.001$, $c_1 = -0.3$, $c_2 = -0.2$. The step size was fixed at $h = 0.02$, the cutoff time at $T = 80$, and the initial condition was $x_0 = 4$.

3. Discrete Operator Framework for the Jacobi Scheme

In this chapter we discuss the theoretical properties of the Jacobi method. We will demonstrate that under similar conditions as those imposed by Theorem 3, the Jacobi Euler method is also a contraction mapping. Furthermore, we will demonstrate that under these conditions the method has an upper bound on the error with respect to the truncated boundary value problem (23) that is proportional to h.

3a. Framework

Given some $\mathbf{x}_0$, we begin with the autonomous system of integral equations

$$\begin{aligned} \mathbf{x} &= \mathbf{x}_0 + \int_0^t f\left(\mathbf{x}(s), \mathbf{y}(s)\right)\, ds\,, \\ \mathbf{y} &= -\int_t^T g\left(\mathbf{x}(s), \mathbf{y}(s)\right)\, ds\,, \end{aligned} \tag{23}$$

where $(\mathbf{x}, \mathbf{y})^T \in (X, Y)^T$, where $X = L^1([0,T], \mathbb{R}^m)$, $Y = L^1([0,T], \mathbb{R}^n)$, and f, g are sufficiently smooth. We then proceed to define the operator $\mathcal{J}_{N,T} : (X,Y)^T \to (X,Y)^T$ by the rules

1. When $t = jh$, where $h = (T+1)/N$ and $j \in \mathbb{N} \cap [1, N-1]$

$$\mathcal{J}_{N,T} \left(\begin{array}{c} \mathbf{a} \\ \mathbf{b} \end{array} \right) \Bigg|_{t=jh} = \left(\begin{array}{c} \mathbf{a}(t-h) + hf\left(\mathbf{a}(t-h), \mathbf{b}(t-h)\right) \\ \mathbf{b}(t+h) - hg\left(\mathbf{a}(t+h), \mathbf{b}(t+h)\right) \end{array} \right).$$

2. When $t = 0$

$$\mathcal{J}_{N,T} \left(\begin{array}{c} \mathbf{a} \\ \mathbf{b} \end{array} \right) \Bigg|_{t=0} = \left(\begin{array}{c} \mathbf{x}_0 \\ \mathbf{b}(t+h) - hg\left(\mathbf{a}(t+h), \mathbf{b}(t+h)\right) \end{array} \right).$$

3. When $t = T$

$$\mathcal{J}_{N,T} \left(\begin{array}{c} \mathbf{a} \\ \mathbf{b} \end{array} \right) \Bigg|_{t=T} = \left(\begin{array}{c} \mathbf{a}(t-h) + hf\left(\mathbf{a}(t-h), \mathbf{b}(t-h)\right) \\ 0 \end{array} \right).$$

4. Otherwise, define ϵ as the closest distance to the lowest nearby jh, and δ as the distance

to the higher nearby jh Then

$$\mathcal{J}_{N,T}\left(\begin{array}{c}\mathbf{a}\\ \mathbf{b}\end{array}\right)\Big|_t = \left(\begin{array}{c}\mathbf{a}(t-h-\epsilon)+(h+\epsilon)f\left(\mathbf{a}(t-h-\epsilon),\mathbf{b}(t-h-\epsilon)\right)\\ \mathbf{b}(t+h+\delta)-(h+\delta)g\left(\mathbf{a}(t+h+\delta),\mathbf{b}(t+h+\delta)\right)\end{array}\right). \tag{24}$$

Note that this operator is simply the Jacobi Euler method with an interpolation (24). Iterations of $T_{N,\tau}$ correspond to the iterations of the numerical method. Since the map is piecewise integrable, $\mathcal{J}_{N,T} : (X,Y)^T \to (X,Y)^T$.

3b. Contraction conditions

Here I briefly examine under what conditions $\mathcal{J}_{N,T}$ is a contraction mapping. First, consider the operator $\tilde{\mathcal{J}}_{N,T} : \mathbb{R}^{(m+n)\times N} \to \mathbb{R}^{(m+n)\times N}$ defined by

1. When $0 < j \in \mathbb{N} < N$, then

$$\tilde{\mathcal{J}}_{N,T}\left(\begin{array}{c}\tilde{\mathbf{a}}\\ \tilde{\mathbf{b}}\end{array}\right)_j = \left(\begin{array}{c}\tilde{\mathbf{a}}_{j-1}+hf\left(\tilde{\mathbf{a}}_{j-1},\tilde{\mathbf{b}}_{j-1}\right)\\ \tilde{\mathbf{b}}_{j+1}+hg\left(\tilde{\mathbf{a}}_{j+1},\tilde{\mathbf{b}}_{j+1}\right)\end{array}\right).$$

2. When $j = 0$

$$\tilde{\mathcal{J}}_{N,T}\left(\begin{array}{c}\tilde{\mathbf{a}}\\ \tilde{\mathbf{b}}\end{array}\right)_j = \left(\begin{array}{c}\mathbf{x}_0\\ \tilde{\mathbf{b}}_{j+1}+hg\left(\tilde{\mathbf{a}}_{j+1},\tilde{\mathbf{b}}_{j+1}\right)\end{array}\right).$$

3. When $j = N$

$$\tilde{\mathcal{J}}_{N,T}\left(\begin{array}{c}\tilde{\mathbf{a}}\\ \tilde{\mathbf{b}}\end{array}\right)_j = \left(\begin{array}{c}\tilde{\mathbf{a}}_{j-1}+hf\left(\tilde{\mathbf{a}}_{j-1},\tilde{\mathbf{b}}_{j-1}\right)\\ 0\end{array}\right).$$

This operator corresponds to the Jacobi scheme on a grid, and as would be expected, has related behaviour to $\mathcal{J}_{N,T}$.

Lemma 17. *The operators $\mathcal{J}_{N,T}$ and $\tilde{\mathcal{J}}_{N,T}$ have the same number of fixed points.*

Proof. Consider some fixed point of $\tilde{\mathcal{J}}_{N,T}$, $\left(\begin{array}{cc}\tilde{\mathbf{a}} & \tilde{\mathbf{b}}\end{array}\right)^T$. Define the functions

$$\left(\begin{array}{c}\mathbf{a}'(t)\\ \mathbf{b}'(t)\end{array}\right) = \left(\begin{array}{c}\tilde{\mathbf{a}}_{\lfloor t/h\rfloor}\\ \tilde{\mathbf{b}}_{\lfloor t/h\rfloor}\end{array}\right),$$

where $\lfloor\,\rfloor$ denotes the floor of a number. Then, by construction of $\tilde{\mathcal{J}}_{N,T}$,

$$\mathcal{J}_{N,T}\left(\begin{array}{c}\mathbf{a}'\\ \mathbf{b}'\end{array}\right)\Big|_{t=jh} = \left(\begin{array}{c}\mathbf{a}'\\ \mathbf{b}'\end{array}\right)\Big|_{t=jh},$$

for all $j \in \mathbb{N} \cap [0,N]$. For $t \neq jh$, the values of the operator are given by

$$\mathcal{J}_{N,T}\left(\begin{array}{c}\mathbf{a}'\\ \mathbf{b}'\end{array}\right)\Big|_{t\neq jh} = \left(\begin{array}{c}\mathbf{a}'(t-h-\epsilon)+(h+\epsilon)f\left(\mathbf{a}'(t-h-\epsilon),\mathbf{b}'(t-h-\epsilon)\right)\\ \mathbf{b}'(t+h+\delta)-(h+\delta)g\left(\mathbf{a}'(t+h+\delta),\mathbf{b}'(t+h+\delta)\right)\end{array}\right),$$

where ϵ and δ have the same meaning as they do in (24). As the mapping of points at $t \neq jh$ depends exclusively on points at $t = jh$, it follows

$$\mathcal{J}_{N,T}\mathcal{J}_{N,T}\begin{pmatrix} \mathbf{a}' \\ \mathbf{b}' \end{pmatrix} = \mathcal{J}_{N,T}\begin{pmatrix} \mathbf{a}' \\ \mathbf{b}' \end{pmatrix}.$$

Hence, the existence of a fixed point of $\tilde{\mathcal{J}}_{N,T}$ implies the existence of a fixed point of $\mathcal{J}_{N,T}$.

Now consider some fixed point of $\mathcal{J}_{N,T}$, $\begin{pmatrix} \mathbf{a} & \mathbf{b} \end{pmatrix}^T$. By construction of $\tilde{\mathcal{J}}_{N,T}$, the vector defined by

$$\begin{pmatrix} \tilde{\mathbf{a}}_j \\ \tilde{\mathbf{b}}_j \end{pmatrix} = \begin{pmatrix} \mathbf{a}(jh) \\ \mathbf{b}(jh) \end{pmatrix}$$

is a fixed point of $\tilde{\mathcal{J}}_{N,T}$. □

Lemma 18. *$\tilde{\mathcal{J}}_{N,T}$ is a contraction mapping if given two arbitrary vectors, $(\mathbf{x}_1, \mathbf{y}_1)^T$ and $(\mathbf{x}_2, \mathbf{y}_2)^T$, then*

$$\sum_{j=1}^{N} \|\mathbf{x}_{1,j} + hf(\mathbf{x}_{1,j}, \mathbf{y}_{1,j}) - \mathbf{x}_{2,j} - hf(\mathbf{x}_{2,j}, \mathbf{y}_{2,j})\|$$
$$+ \|\mathbf{y}_{1,j} + hg(\mathbf{x}_{1,j}, \mathbf{y}_{1,j}) - \mathbf{y}_{2,j} - hg(\mathbf{x}_{2,j}, \mathbf{y}_{2,j})\|) < \sum_{j=1}^{N} \|\mathbf{x}_{1,j} - \mathbf{x}_{2,j}\| + \|\mathbf{y}_{1,j} - \mathbf{y}_{2,j}\|).$$

Proof. Let us use the vector norm defined by

$$\left\|(\mathbf{x}, \mathbf{y})^T\right\| = \sum_{j=1}^{N} (\|\mathbf{x}_j\| + \|\mathbf{y}_j\|).$$

For $\tilde{\mathcal{J}}_{N,T}$ to be a contraction mapping, it must satisfy the condition

$$\left\|\tilde{\mathcal{J}}_{N,T}\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{pmatrix} - \tilde{\mathcal{J}}_{N,T}\begin{pmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{pmatrix}\right\| < \left\|\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{pmatrix} - \begin{pmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{pmatrix}\right\|.$$

From the definition of $\tilde{\mathcal{J}}_{N,T}$, we can note that

$$\left\|\tilde{\mathcal{J}}_{N,T}\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{pmatrix} - \tilde{\mathcal{J}}_{N,T}\begin{pmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{pmatrix}\right\| \leq \sum_{j=1}^{N} \|\mathbf{x}_{1,j} + hf(\mathbf{x}_{1,j}, \mathbf{y}_{1,j}) - \mathbf{x}_{2,j} - hf(\mathbf{x}_{2,j}, \mathbf{y}_{2,j})\|$$
$$+ \|\mathbf{y}_{1,j} + hg(\mathbf{x}_{1,j}, \mathbf{y}_{1,j}) - \mathbf{y}_{2,j} - hg(\mathbf{x}_{2,j}, \mathbf{y}_{2,j})\|),$$

and the result follows. □

Theorem 19. *Suppose that we have a system of differential equations*

$$\frac{d\mathbf{x}}{dt} = \mathcal{A}\mathbf{x} + \mathcal{F}(\mathbf{x}, \mathbf{y}),$$
$$\frac{d\mathbf{y}}{dt} = \mathcal{B}\mathbf{y} + \mathcal{G}(\mathbf{x}, \mathbf{y}).$$

as long as we have that

1. *For sufficiently small h, under some induced matrix norm $|||\cdot|||$, $|||I + h\mathcal{A}||| < 1 - h\alpha$ and $|||I - h\mathcal{B}||| < 1 - h\beta$ for some suitable positive constants α, β.*
2. *There exist constants $L_{\mathcal{F}}$ and $L_{\mathcal{G}}$ such that*

$$\|\mathcal{F}(\mathbf{x}_1, \mathbf{y}_1) - \mathcal{F}(\mathbf{x}_2, \mathbf{y}_2)\| \leq L_{\mathcal{F}} \left(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|\right),$$
$$\|\mathcal{G}(\mathbf{x}_1, \mathbf{y}_1) - \mathcal{G}(\mathbf{x}_2, \mathbf{y}_2)\| \leq L_{\mathcal{G}} \left(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|\right),$$

   for all choices of $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2$.

3. *$\alpha, \beta > L_{\mathcal{F}} + L_{\mathcal{G}}$*

then $\tilde{\mathcal{J}}_{N,T}$ will be a contraction mapping.

Proof. From Lemma 18 we need to prove that

$$\sum_{j=1}^{N} \left(\|(I + h\mathcal{A})(\mathbf{x}_{1,j} - \mathbf{x}_{2,j}) + h\left(\mathcal{F}(\mathbf{x}_{1,j}, \mathbf{y}_{1,j}) - \mathcal{F}(\mathbf{x}_{2,j}, \mathbf{y}_{2,j})\right)\| \right.$$
$$\left. + \|(I - h\mathcal{B})(\mathbf{y}_{1,j} - \mathbf{y}_{2,j}) - h\left(\mathcal{G}(\mathbf{x}_{1,j}, \mathbf{y}_{1,j}) - \mathcal{G}(\mathbf{x}_{2,j}, \mathbf{y}_{2,j})\right)\|\right)$$
$$< \sum_{j=1}^{N} \|\mathbf{x}_{1,j} - \mathbf{x}_{2,j}\| + \|\mathbf{y}_{1,j} - \mathbf{y}_{2,j}\|).$$

It is sufficient to prove that this inequality holds for all j, so we will drop the sum. Applying the triangle inequality, we have

$$\begin{aligned} L.H.S. &\leq \|(I + h\mathcal{A})(\mathbf{x}_{1,j} - \mathbf{x}_{2,j})\| + h\|\mathcal{F}(\mathbf{x}_{1,j}, \mathbf{y}_{1,j}) - \mathcal{F}(\mathbf{x}_{2,j}, \mathbf{y}_{2,j})\| \\ &\quad + \|(I - h\mathcal{B}))(\mathbf{y}_{1,j} - \mathbf{y}_{2,j})\| + h\|\mathcal{G}(\mathbf{x}_{1,j}, \mathbf{y}_{1,j}) - \mathcal{G}(\mathbf{x}_{2,j}, \mathbf{y}_{2,j})\| \\ &\leq |||I + h\mathcal{A}|||\,\|\mathbf{x}_{1,j} - \mathbf{x}_{2,j}\| + h\|\mathcal{F}(\mathbf{x}_{1,j}, \mathbf{y}_{1,j}) - \mathcal{F}(\mathbf{x}_{2,j}, \mathbf{y}_{2,j})\| \\ &\quad + |||I - h\mathcal{B}|||\,\|\mathbf{y}_{1,j} - \mathbf{y}_{2,j}\| + h\|\mathcal{G}(\mathbf{x}_{1,j}, \mathbf{y}_{1,j}) - \mathcal{G}(\mathbf{x}_{2,j}, \mathbf{y}_{2,j})\|, \end{aligned}$$

where $|||\cdot|||$ is the matrix norm induced by $\|\cdot\|$. Applying conditions 1 and 2, we get

$$L.H.S < (1 - h\alpha + hL_{\mathcal{F}} + hL_{\mathcal{G}})\|\mathbf{x}_{1,j} - \mathbf{x}_{2,j}\| + (1 - h\beta + hL_{\mathcal{F}} + hL_{\mathcal{G}})\|\mathbf{y}_{1,j} - \mathbf{y}_{2,j}\|.$$

Applying condition 3, the result follows. □

We must note that the third condition imposed is much stronger than the gap condition used by [1]. It is unclear whether this condition can be reduced by a more careful analysis of the norm inequalities in Theorems 18 and 19, or if it simply a result of the discretization.

3c. Step size error

We now have a much more general framework for describing the Euler method. As promised, We can now lay out a more general condition for the error. In order to aid this discussion, I will define

$$\mathcal{J}^X_{N,T}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix} \equiv \mathrm{proj}_X\left(\mathcal{J}_{N,T}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\right), \quad \mathcal{J}^Y_{N,T}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix} \equiv \mathrm{proj}_Y\left(\mathcal{J}_{N,T}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\right).$$

Theorem 20. *If the operator $\mathcal{J}_{N,T} : (X,Y)^T \to (X,Y)^T$ is a contraction mapping with rate constant L and fixed point $(a,b)^T$, and $(x,y)^T$ satisfies (23) and has a Taylor sequence with radius of convergence of at least $2h$ around all $t_0 \in [0,T]$. Then*

$$\left\|\begin{pmatrix}\mathbf{x}(t)\\ \mathbf{y}(t)\end{pmatrix} - \begin{pmatrix}\mathbf{a}(t)\\ \mathbf{b}(t)\end{pmatrix}\right\| \leq \frac{2Kh^2}{1-L}, \tag{25}$$

where K is some finite, positive constant. Furthermore, when $t = jh$, for some $j \in \mathbb{N}$,

$$\left\|\begin{pmatrix}\mathbf{x}(jh)\\ \mathbf{y}(jh)\end{pmatrix} - \begin{pmatrix}\mathbf{a}(jh)\\ \mathbf{b}(jh)\end{pmatrix}\right\| \leq \frac{Kh^2}{2(1-L)}. \tag{26}$$

Proof. Recall that for any given function of one variable, $q(t)$, that is Taylor expandable within some ball of radius ϵ around some point t_0, for some $\delta < \epsilon$

$$q(t_0 \pm \delta) = q(t_0) \pm \ell q'(t_0) + \frac{\ell^2}{2} q''(\xi)'.$$

where $\xi \in [t_0, t_0 \pm \ell]$. Therefore, for some $\mu_1 \leq 2h$ we can write

$$\mathbf{x}(t+\mu_1) = \mathbf{x}(t) + \mu_1 f\left(\mathbf{x}(t), \mathbf{y}(t)\right) + \frac{\mu_1^2}{2}\mathbf{x}''(\xi).$$

However, if we define μ_1 as $h + \epsilon$ from (24), it follows that

$$\mathbf{x}(t+\mu_1) = \mathcal{J}^X_{N,T}\begin{pmatrix}\mathbf{x}(t+\mu_1)\\ \mathbf{y}(t+\mu_1)\end{pmatrix} + \frac{\mu_1^2}{2}x''(\xi). \tag{27}$$

Likewise, we can define $\mu_2 = h + \delta$

$$\mathbf{y}(t-\mu_2) = \mathcal{J}^Y_{N,T}\begin{pmatrix}\mathbf{x}(t-\mu_2)\\ \mathbf{y}(t-\mu_2)\end{pmatrix} + \frac{\mu_2^2}{2}\mathbf{y}''(\xi)). \tag{28}$$

Defining

$$K = \sup_{t\in[0,T]}\left\|\begin{pmatrix}\mathbf{x}''(\xi)\\ \mathbf{y}''(\xi)\end{pmatrix}\right\|$$

and noting that $\mu \leq 2h$, we find that

$$\left\| \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{pmatrix} - \begin{pmatrix} \mathbf{a}(t) \\ \mathbf{b}(t) \end{pmatrix} \right\| \leq \left\| \mathcal{J}_{N,T} \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{pmatrix} - \mathcal{J}_{N,T} \begin{pmatrix} \mathbf{a}(t) \\ \mathbf{b}(t) \end{pmatrix} \right\| + 2hK \,.$$

Since $\mathcal{J}_{N,T}$ is a contraction mapping, meaning that

$$\left\| \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{pmatrix} - \begin{pmatrix} \mathbf{a}(t) \\ \mathbf{b}(t) \end{pmatrix} \right\| \leq L \left\| \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{pmatrix} - \begin{pmatrix} \mathbf{a}(t) \\ \mathbf{b}(t) \end{pmatrix} \right\| + 2hK \,,$$

or (since $L < 1$)

$$\left\| \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{pmatrix} - \begin{pmatrix} \mathbf{a}(t) \\ \mathbf{b}(t) \end{pmatrix} \right\| \leq \frac{2Kh^2}{1-L} \,.$$

Furthermore, we can note that when $t = jh$, $\mu_1 = \mu_2 = h$, and equations (27) and (28) become

$$\mathbf{x}(jh + \mu_1) = \mathcal{J}^X_{N,T} \begin{pmatrix} \mathbf{x}(jh + \mu_1) \\ \mathbf{y}(jh + \mu_1) \end{pmatrix} + \frac{h^2}{2} \mathbf{x}''(\xi) \,,$$

$$\mathbf{y}(jh - \mu_2) = \mathcal{J}^Y_{N,T} \begin{pmatrix} \mathbf{x}(jh - \mu_2) \\ \mathbf{y}(jh - \mu_2) \end{pmatrix} + \frac{h^2}{2} \mathbf{y}''(\xi)) \,.$$

Then, it follows that

$$\left\| \begin{pmatrix} \mathbf{x}(jh) \\ \mathbf{y}(jh) \end{pmatrix} - \begin{pmatrix} \mathbf{a}(jh) \\ \mathbf{b}(jh) \end{pmatrix} \right\| \leq \frac{Kh^2}{2(1-L)} \,.$$

□

Though this gives an error term for the numerical method, without an idea of what L is, we cannot say if taking the limit $h \to 0$ converges. Under the assumptions of Theorem 19, we can note that $L = 1 - \gamma h$ for some $\gamma > 0$, meaning ther error become

$$\frac{2Kh}{\gamma} \,,$$

which is first order in h. This is what we would expect for the Euler method.

Corollary 21. *If $\mathcal{J}_{N,T}$ is a contraction mapping with rate L such that $L < 1 - h^2$, and fixed point $(\mathbf{a}, \mathbf{b})^T$, then*

$$\lim_{h \to 0} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} ,$$

is the unique Taylor expandable solution of (23).

Proof. From (25) it follows that under the condition that the operator $\mathcal{J}_{N,T}$ is a contraction mapping, its fixed point $(\mathbf{a}, \mathbf{b})^T$ limits to a solution of (23). There can only be one Taylor expandable solution of (23) since (25) is true with respect to all Taylor expandable solutions. □

3d. Conclusions

The methodology presented in this chapter can be understood in two ways. First of all, it provides a rigorous method through which we can determine whether our schemes are able to converge to a unique fixed point, and what is the global error of these schemes with respect to the truncated boundary value problem. Viewing this formalism through such a sense makes no reference to the fact that the presented schemes are meant to calculate stable and unstable manifolds, and the similarity between Theorem 19 to the conditions of Castaneda and Rosa is a coincidence.

Alternatively, the formalism may be seen as the first step to a solution to the integral equations

$$\mathbf{x} = \mathbf{x}_0 + \int_0^t f\left(\mathbf{x}(s), \mathbf{y}(s)\right)\, ds\,,$$
$$\mathbf{y} = -\int_t^\infty g\left(\mathbf{x}(s), \mathbf{y}(s)\right)\, ds$$

for we have demonstrated that the system for any finite T has a unique solution. In a sense, we are one limit away from a novel proof of the existence of a unique stable manifold in a dynamical system. If this is achieved, this methodology would be much more robust than that of Castañeda and Rosa [1], since the vector norm equations of Theorem 18 puts no requirement on the existence of linear and nonlinear components.

A. Derivation of Nonlinear Test System

Consider the differential equation of the form

$$\frac{du}{dt} = -u\,, \quad \frac{dv}{dt} = v + u\sin u - ku^3\,.$$

This system has a solution of the form

$$u(t) = u_0 e^{-t}\,, \quad v(t) = \alpha e^t - \frac{ku_0^3}{4}e^{-3t} + \cos\left(u_0 e^{-t}\right) - \frac{1}{u_0}\sin\left(u_0 e^{-t}\right)e^t\,,$$

where α is a constant determined by the initial condition on v. From Definition 1, the stable manifold will be given by some set of conditions on u_0, v_0 such that $\alpha = 0$. This is satsifed by

$$v_0 = \frac{k}{4}u_0^3 + u_0\cos u_0 + \frac{1}{u_0}\sin u_0\,.$$

Now consider the linear transformation to the function x, y:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 1 & c_1 \\ c_2 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}, \quad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{1-c_1c_2} & -\frac{c_1}{1-c_1c_2} \\ -\frac{c_2}{1-c_1c_2} & \frac{1}{1-c_1c_2} \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

Then, the differential equations governing x and y are given by

$$\begin{aligned}
\frac{dx}{dt} &= \frac{-1}{1-c_1c_2}(x + c_1y) - \frac{c_1}{1-c_1c_2}\left(c_2x + y + (x+c_1y)\sin(c+c_1y) - k(x+c_1y)^3\right)\,, \\
\frac{dy}{dt} &= \frac{c_2}{1-c_1c_2}(x + c_1y) + \frac{1}{1-c_1c_2}\left(c_2x + y + (x+c_1y)\sin(c+c_1y) - k(x+c_1y)^3\right)\,.
\end{aligned}$$

Simplifying, we get

$$\begin{aligned}
\frac{dx}{dt} &= -\frac{1+c_1c_2}{1-c_1c_2}x - \frac{-2c_1}{1-c_1c_2}y - \frac{c_1}{1-c_1c_2}\left(+(x+c_1y)\sin(c+c_1y) - k(x+c_1y)^3\right)\,, \\
\frac{dy}{dt} &= \frac{2c_2}{1-c_1c_2}x + \frac{1+c_1c_2}{1-c_1c_2}y + \frac{1}{1-c_1c_2}\left((x+c_1y)\sin(c+c_1y) - k(x+c_1y)^3\right)\,.
\end{aligned}$$

The stable manifold also transforms linearly, and becomes

$$\frac{k}{4}(x+c_1y)^3 + \cos(x+c_1y) - \frac{\sin(x+c_1y)}{x+c_1y} - c_2x - y = 0\,.$$

B. Local Boundary Conditions

We saw in Chapter 1 that the boundary value formulation of the stable manifold requires the conditions of Theorem 3. Nevertheless, we can note that our numerical methods accurately converge to the stable manifold for systems which do not satisfy these requirements, as seen in Example 15. Though all of the theorems on the properties of the numerical methods which I present do require some variation of these conditions, it is important to note that the boundary value formulation itself, at least locally, does not require them.

Hand-wavy Theorem 1. *Given the dynamical system*

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, \mathbf{y}), \quad \frac{d\mathbf{y}}{dt} = g(\mathbf{x}, \mathbf{y}),$$

where $\dim \mathbf{x} = m$ *and* $\dim \mathbf{y} = n$ *has a hyperbolic fixed point at the origin, a diagonalizable Jacobian matrix at the origin, and an* m*-dimensional stable manifold, then there exists some set of coordinates* $(\mathbf{u}, \mathbf{v})$*, where* $\dim \mathbf{u} = m$ *and* $\dim \mathbf{v} = n$*, which are related linearly to* $(\mathbf{x}, \mathbf{y})$*:*

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \mathbf{M} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix},$$

such that near the origin, the stable manifold of the system satisfies the boundary values $\mathbf{u}(0) = \mathbf{u_0}$*,* $\mathbf{v}(\infty) = 0$*.*

Proof. Since the system has an m dimensional stable manifold, we know that the Jacobian of the system at the origin $\mathbf{J}(f, g; \mathbf{x} = \mathbf{0}, \mathbf{y} = 0)$ has m negative eigenvalues and n positive eigenvalues. We can diagonalize the Jacobian by similarity through some matrix M such that $\mathbf{M}^{-1}\mathbf{J}\mathbf{M}$ is a diagonal matrix. The linear system corresponding to this matrix has a stable manifold corresponding to the boundary values $\mathbf{u}(0) = \mathbf{u_0}$, $\mathbf{v}(\infty) = 0$. This system is locally a topological conjugacy (see [13]) to the original dynamical system in coordinates given

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \mathbf{M} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix},$$

meaning that locally, the boundary conditions will apply to the stable manifold. □

Though the methods that I present here are not guaranteed to converge for all dynamical systems with stable manifolds, it is clear that they will *locally* converge for a lot of them. Therefore, it is not too unexpected to see the numerical schemes find stable manifolds for systems which are not covered by the theory.

C. Sample MATLAB Code

```
function [x,y] = jb_euler(f,g,x0,yT,h,N,tol,maxiter)
% jacobi euler method

% initialize xy
x = x0*ones(1,N);
y = yT*ones(1,N);

e = 2*tol;
i = 0;
while e > tol && i < maxiter
    i = i + 1;
    xl = x;
    yl = y;
    x(:,2:end) = x(:,1:end-1) + h*f(x(:,1:end-1),y(:,1:end-1));
    y(1:end-1) = y(:,2:end) - h*g(x(:,2:end),y(:,2:end));
    e = norm(x-xl)+norm(y-yl);
end
```

Jacobi Euler

```
function [x,y] = gs_euler(f,g,x0,yT,h,N,tol,maxiter)
% gauss-seidel euler method

% initialize xy
x = x0*ones(1,N);
y = yT*ones(1,N);

e = 2*tol;
i = 0;
while e > tol && i < maxiter
    i = i + 1;
    xl = x;
    yl = y;
    for j = 1:N-1
        x(:,j+1) = x(:,j) + h*f(x(:,j),yl(:,j));
        y(:,N-j) = y(:,N-j+1) - h*g(xl(:,N-j+1),y(:,N-j+1));
    end
    e = norm(x-xl)+norm(y-yl);
end
```

Gauss Seidel Euler

Bibliography

[1] N. Castañeda and R. Rosa, "Optimal estimates for the uncoupling of differential equations," *Journal of Dynamics and Differential Equations*, vol. 8, no. 1, pp. 103–139, 1996.

[2] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. Springer, 2 ed., 2003.

[3] R. Rosa, "Approximate inertial manifolds of exponential order," *Discrete and Continuous Dynamical Systems*, vol. 1, no. 3, pp. 421–448, 1995.

[4] Y. M. Chung and M. S. Jolly, "A unified approach to compute foliations, inertial manifolds, and tracking solutions," *Mathematics of Computation*, vol. 81, no. 294, pp. 1729–1751, 2015.

[5] M. S. Jolly, R. Rosa, and R. Temam, "Accurate computations on inertial manifolds," *SIAM Journal on Scientific Computing*, vol. 22, no. 6, pp. 2216–2238, 2001.

[6] J. C. Robinson, "Convergent families of inertial manifolds for convergent approximations," *Numerical Algorithms*, vol. 14, no. 1, pp. 179–188, 1997.

[7] C. Foias, M. S. Jolly, I. G. Kevrekidis, G. R. Sell, and E. S. Titi, "On the computation of inertial manifolds," *Physics Letters A*, vol. 131, no. 7, pp. 433 – 436, 1988.

[8] J. Guckenheimer and A. Vladimirsky, "A fast method for approximating invariant manifolds," *SIAM Journal on Applied Dynamical Systems*, vol. 3, no. 3, pp. 232–260, 2004.

[9] K. Burrage, *Parallel and Sequential Methods for Ordinary Differential Equations*. Oxford University Press, 1 ed., 1995.

[10] R. L. Burden and J. D. Faires, *Numerical Analysis*. Cengage Learning, 9 ed., 2010.

[11] Y. M. Chung, C. K. Li, and Y. Liu, "Spectral radius estimate." unpublished.

[12] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2 ed., 2013.

[13] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview Press, 2 ed., 2014.