Numerical Analysis of Nonlinear Parameter-Dependent Systems with Continuation Methods

Max Georg Spetzler

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Anshu Narang-Siddarth, Chair

Mehran Mesbahi

Uy-Loi Ly

Program Authorized to Offer Degree:
Aeronautics and Astronautics

University of Washington

Abstract

Numerical Analysis of Nonlinear Parameter-Dependent Systems
with Continuation Methods

Max Georg Spetzler

Chair of the Supervisory Committee:
Assistant Professor Anshu Narang-Siddarth
William E. Boeing Department of Aeronautics and Astronautics

During the development of most aerospace systems, much effort is spent on deriving detailed models that describe the system dynamics. Powerful analysis tools are then required to extract a comprehensive understanding of the system's behavior throughout its operational envelope from its mathematical description. This dissertation introduces new numerical analysis methods for this purpose, with focus on studying the effect of parameters on the system dynamics. The research extends the methods of numerical bifurcation analysis to address issues specific to the aerospace sector. A framework for bifurcation analysis of multi-parameter systems in the presence of equality constraints on states and parameters is derived first, allowing analysis of particular parts of the operational envelope as specified by the constraints. The approach is then extended to bifurcation analysis of the zero dynamics for systems with input-output structure. To expose how local dynamical properties change throughout the operating envelope of the system, a method for computation of equilibrium conditions that satisfy constraints involving the eigenmodes of the linearized dynamics is developed next. A modification to the pseudo-arclength continuation algorithm underlying these methods is suggested to enable application to problems that are continuous, but only piecewise differentiable. Finally, a method to verify that the operating equilibrium of a system with parameter uncertainty does not experience bifurcation for any parameter combination is derived.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my adviser, Anshu Narang-Siddarth, for her guidance and support throughout this journey. She introduced me to the topic of bifurcation analysis and encouraged me to pursue my own ideas, while always keeping me on track.

The advice offered by the members of my doctoral supervisory committee is also greatly appreciated. The classes I have taken with Mehran Mesbahi, Kristi Morgansen, Eli Livne, and Howard Chizeck have shaped the way that I think, and the discussions outside of the classroom have helped me become a better researcher. Talking to Uy-Loi Ly when I was his teaching assistant sparked my interest in flight dynamics and control theory, which led me to join the Advanced Dynamics, Validation & Control Research Laboratory. Thank you for your support.

Many thanks also go to the other members of the lab. Without Armand, Tom, Adam, Peter, Dillon, Matt, and Colby, my time at the University of Washington would not have been the same.

Last, I would like to thank my lovely wife, Liz. She always believed in me and gave me strength when I needed it most. I would not be who I am without her.

Die Wissenschaft, sie ist und bleibt,
was einer ab vom andern schreibt –
doch trotzdem ist, ganz unbestritten,
sie immer weiter fortgeschritten.

Eugen Roth

Chapter 1
INTRODUCTION

Modern aerospace vehicles are highly complex and often require active control to fulfill their mission. In order to design a high-performance control system, the dynamics of the vehicle need to be well understood. The differential equations that are commonly used to express the vehicle dynamics are usually nonlinear and difficult to analyze, especially when they incorporate tabulated data derived from measurements. This is often the case when the model contains empirical aerodynamic models obtained through wind tunnel experiments or flight testing. Numerical analysis methods that are directly applicable to the system model can provide valuable insight into the vehicle's behavior even when analytical approaches become infeasible due to the complex structure of the model. Further analysis is required once the design of a control system has been completed, since the safety and performance of the closed-loop system must be validated before the system is cleared for flight testing [1].

The behavior of nonlinear systems operating close to an equilibrium is characterized by local properties such as stability and performance metrics including damping ratios, natural frequencies, and stability margins. However, the operating equilibrium and the corresponding nonlinear system behavior change when the system configuration, operating and environmental conditions, and/or the steady values of the control inputs are altered. These parameter changes are usually necessary to meet desired mission objectives, but may sometimes be unavoidable and occur unintentionally (during emergency situations). Even the effect of parameters that do not change may have to be considered if their value is not known precisely but only within bounds. Often, distinct parameter values that lead to significant changes in system behavior are unknown during design. It is of critical importance to identify areas

of the vehicle's operating envelope where such changes occur in order to address them appropriately during the controller design process. Conversely, control law validation requires demonstration of adequate performance of the closed-loop system throughout the operating envelope, i.e., proving the lack of undesired phenomena.

The prevalent analysis approach for investigating the effect of parameters on nonlinear systems relies on repeated time simulation. This is because few restrictions need to be imposed on the types of systems being considered. System features such as time delays, non-smooth dynamics, and switching conditions are easily accommodated in time simulations. The results, however, only reveal properties pertaining to the particular conditions of the simulation scenario, such as control inputs, parameter settings, vehicle configuration, and initial conditions. A large number simulations must be executed to obtain a comprehensive overview of the system's behavior throughout the full operation envelope. Due to the high dimensionality and complexity of aerospace systems, surveys of this type are highly demanding both in terms of computational effort and data management [2, 3].

The effect of parameters on system attractors and repellers, such as equilibrium points and limit cycles, is the subject of bifurcation theory [4–6]. Points in state-parameter space where the number and/or stability properties of these features change are called bifurcations. Numerous types of bifurcations with different characteristics have been studied. Numerical bifurcation analysis investigates a range of parameters at once and thus helps in obtaining a more global and comprehensive picture of the system dynamics. This is done by explicit computation of the location and local stability of equilibria and limit cycles for all parameter values with continuation algorithms [7]. If a bifurcation is found in this process, bifurcation analysis determines the exact parameter value at which the bifurcation occurs and examines the system behavior for values that exceed this threshold. The application of numerical bifurcation analysis tools to many aerospace problems has been discussed in [8–14] and the references therein.

The research in this dissertation builds on the principles of bifurcation analysis in multiple ways, with the intent of creating capabilities that will prove useful for the analysis of

aerospace systems. The methods of numerical bifurcation analysis are extended to include problems with multiple parameters and equality constraints, to analyze system properties beyond stability, and to consider models which do not satisfy the assumption of continuous differentiability inherent to the underlying numerical algorithm.

1.1 Contributions

This dissertation makes the following contributions to the state-of-the-art:

- Derivation of an improved method for bifurcation analysis of multi-parameter systems in the presence of equality constraints
- Extension of numerical bifurcation analysis to the study of the zero dynamics for single-input/single-output systems
- Formulation of continuation problems that enable the computation of equilibrium subsets satisfying equality constraints involving eigenvalues and eigenvectors of the linearized dynamics
- Analysis of the root cause of pseudo-arclength continuation failure when applied to continuous, but only piecewise differentiable problems as well as derivation of a mitigation strategy based on the insights gained
- A numerical method for verification of the existence and local stability of the operating equilibrium under parameter uncertainty

1.2 Organization

The remainder of this dissertation is organized as follows. Chapter 2 summarizes the theory that forms the basis for the research presented in the chapters that follow, such as continuation methods and bifurcation theory. Bifurcation analysis of multi-parameter systems in the presence of equality constraints on states and parameters is the subject of Chapter 3, which

also extends bifurcation analysis to the zero dynamics of systems with input-output structure. Chapter 4 derives continuation problems that allow the specification of constraints on the eigenmodes of the linearized dynamics, which permit the characterization of local dynamic properties throughout the operating envelope. Three analysis cases related to nonlinear aircraft dynamics are presented to demonstrate the types of results that may be obtained with this problem formulation. Application of the pseudo-arclength method to piecewise differentiable problems is the topic of Chapter 5, which analyzes the conditions that cause the algorithm to stall due to lack of differentiability and derives a strategy to alleviate the issue. Chapter 6 explores a new numerical method to verify that a system's operating equilibrium exists for all parameter combinations in a bounded set without explicitly resolving the parameter dependence. The method is based on a reliable computing technique called affine arithmetic, which is also discussed in the chapter. Some concluding remarks are provided in Chapter 7.

Chapter 2
BACKGROUND

This chapter summarizes the theory that forms the basis for continuation methods and their application in the context of the research in this dissertation. The first section describes the sets that are defined by underdetermined nonlinear systems of equations. Continuation algorithms that numerically compute these sets are discussed next, with particular focus on the pseudo-arclength method. A brief review of direct collocation methods, which enable finite-dimensional approximation of the infinite-dimensional problems that frequently arise in time-dependent problems, follows. The research in this dissertation grew out of the field of bifurcation analysis, which is summarized in the last section.

2.1 Implicitly Defined Surfaces and Implicit Functions

Continuation methods deal with the problem of solving the equation $F(z) = 0$, where F is a continuous nonlinear function mapping from $\mathcal{D} \subseteq \mathbb{R}^N$ to $\mathbb{R}^{N-M}$ (with $N > M \geq 1$). The feature distinguishing this equation from other root-finding problems lies in the fact that the problem is underdetermined, meaning that there are more scalar variables than scalar equalities. Under mild regularity assumptions, the set of all solution points (called the *kernel* of F),

$$\ker F = \{\, z \in \mathcal{D} \mid F(z) = 0 \,\}, \tag{2.1}$$

consists of M-dimensional surfaces embedded in N-dimensional space, which may be disjoint or glued together along singular lines of lower dimension. This is a consequence of the implicit function theorem, which guarantees that $\ker F$ can be locally parametrized as a function of M components of z in an open neighborhood of almost any point in $\ker F$ if the regularity assumptions are met. The goal of continuation methods is thus not merely to locate a single

solution point, but to find connected families of solution points.

The term "implicit function theorem" does not refer to one particular result, but instead comprises a whole family of theorems that are formulated to fit the settings of different problems [15]. The classical form for differentiable functions defined on Euclidean space is restated below. A second form that relaxes the differentiability requirement to Lipschitz continuity is given in chapter 5.

Theorem 2.1 (Implicit Function Theorem)**.** *Let H be a mapping of class C^k, $k \geq 1$, defined on an open set $\mathcal{U} \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and taking values in $\mathbb{R}^{n_1}$. Let $(z_1^*, z_2^*) \in \mathcal{U}$ satisfy $H(z_1^*, z_2^*) = 0$ and suppose that the Jacobian matrix of H with respect to its first argument is invertible at (z_1^*, z_2^*). Then there exists a neighborhood $\widetilde{\mathcal{U}}$ of (z_1^*, z_2^*), an open set $\mathcal{W} \subseteq \mathbb{R}^{n_2}$ containing z_2^*, and a function ϕ of class C^k defined on $\mathcal{W}$ and taking values in $\mathbb{R}^{n_1}$ such that*

$$H(\phi(z_2), z_2) = 0 \qquad \text{for all} \quad z_2 \in \mathcal{W} \tag{2.2}$$

and $\nabla\phi(z_2^*) = -\nabla_{z_1} H(z_1^*, z_2^*)^{-1} \nabla_{z_2} H(z_1^*, z_2^*)$. Furthermore, ϕ is the only function satisfying

$$\left\{ (z_1, z_2) \in \widetilde{\mathcal{U}} \;\middle|\; H(z_1, z_2) = 0 \right\} = \left\{ (z_1, z_2) \in \widetilde{\mathcal{U}} \;\middle|\; z_2 \in \mathcal{W}, z_1 = \phi(z_2) \right\}. \tag{2.3}$$

Every point z where the Jacobian matrix $\nabla F(z) \in \mathbb{R}^{(N-M)\times N}$ of $F : \mathcal{D} \subseteq \mathbb{R}^N \to \mathbb{R}^{N-M}$ (when F is differentiable) has full rank is called a *regular point* of F, and all points where the Jacobian loses rank are called *critical points* of F. Clearly, the implicit function theorem can always be applied at a regular point by splitting the variable z into two parts (z_1, z_2) such that the components of z_1 correspond to $N - M$ columns of the Jacobian matrix that form a linearly independent set.

The following

Proposition 2.2. *Let F take values in $\mathbb{R}^N$ and map to $\mathbb{R}^{N-M}$, with $N > M \geq 1$. Assume that* $\ker F$ *is not empty, that F is continuously differentiable on* $\ker F$*, and that the set $\mathcal{R}$ of all regular points of F in* $\ker F$ *is dense in* $\ker F$*. Then* $\ker F$ *is the union of one or more M-dimensional surfaces and their boundaries.*

Proof. Suppose $z_r \in \ker F$ is a regular point. Such a point always exists because $\ker F$ is not empty and $\mathcal{R}$ is dense in $\ker F$. By the implicit function theorem, there exists an open neighborhood of z_r in $\mathbb{R}^N$ such that the intersection of $\ker F$ with that neighborhood is the graph of a differentiable function of M variables. Thus $\mathcal{R}$ consists of one or more M-dimensional surfaces embedded in $\mathbb{R}^N$. Now suppose $z_c \in \ker F$ is a critical point. Because $\mathcal{R}$ is dense in $\ker F$, z_c is a limit point of $\mathcal{R}$, so z_c is in the closure of $\mathcal{R}$. As all points are either regular or critical, this concludes the proof. ■

The kernel of a function F that satisfies the conditions of Proposition 2.2 may thus be thought of a collection of surfaces that meet along lines of singular points of F. Such functions will be called *well-behaved* throughout this dissertation. The regularity conditions exclude functions whose kernel contains disjoint subsets consisting only of critical points (such as $F(z) = \|z\|^2$) as well as functions that do not sufficiently constrain the solution set (such as the trivial function $F(z) = 0$).

While many of the continuation problems derived in this dissertation are well-defined for any M, the main focus is on the case where $M = 1$. The reason for this is twofold. First, efficient and mature numerical algorithms to compute the one-dimensional solution branches for this case are available. They are discussed in the next section. Second, the visualization of higher-dimensional surfaces in two-dimensional diagrams is difficult, while a well-posed one dimensional continuation problem yields readily interpretable plots.

2.2 Continuation Algorithms

Continuation algorithms exploit continuity of F to find solutions in the neighborhood of previously calculated solution points, thereby successively expanding the set of known solutions. The process is initialized with a user-provided solution point z_0 and generally only computes those parts of the solution set that are connected to the initial solution through a continuous path contained within the solution set. The computation is limited to a finite region $\Omega \subset \mathbb{R}^N$ for obvious practical reasons. More precisely, continuation algorithms numerically

approximate the infinite set $\mathcal{S}(F, \Omega, z_0)$ defined as

$$\mathcal{S}(F, \Omega, z_0) = \{\, z \in \ker F \cap \Omega \mid \exists \text{ a continuous path in } \ker F \cap \Omega \text{ between } z \text{ and } z_0 \,\}, \quad (2.4)$$

where Ω is typically chosen as a hypercube.

While algorithms exist that can compute $\mathcal{S}(F, \Omega, z_0)$ for any value of M, the computational cost quickly becomes prohibitive for large M [16]. However, the computation for $M = 1$ is feasible even when the embedding dimension N is large. One of the most prominent methods for this case, the pseudo-arclength method, is described below.

2.2.1 Pseudo-arclength continuation

The pseudo-arclength continuation method [7, 17, 18] solves the problem defined in Eq. (2.4) with $M = 1$, so that F maps from $\mathbb{R}^N$ to $\mathbb{R}^{N-1}$. F is required to be continuously differentiable and well-behaved. With these restrictions, $\mathcal{S}(F, \Omega, z_0)$ contains continuous one-dimensional objects called *branches* which may intersect at certain *branch points*. Differentiability guarantees that a unique tangent direction to the branch exists at every regular point of F, which coincides with the null space of ∇F. Branch points occur at critical points of F where ∇F does not have full rank. Given a known solution point z_{old}, the pseudo-arclength method

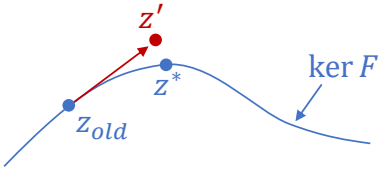

Figure 2.1: Single predictor-corrector algorithm step.

computes a neighboring solution in two steps. First, the predictor point z' is obtained by taking a finite step in the tangent direction, thereby slightly diverging from the branch (Figure 2.1). Second, Newton's method is applied to converge back onto the solution branch and find the next solution point z^*. The procedure is then repeated to obtain further points

and only stopped once the boundary of Ω is crossed. Branch points that are crossed in the process are detected and used to initialize the algorithm for continuation of the intersecting branch. They can also be used as an additional stop condition, so that Ω is dynamically adjusted during the continuation process in order to exclude branch points.

Many variations of the algorithm exist. The predictor direction is either obtained by explicit computation of an element in the null space of $\nabla F(z_{old})$ (usually called *tangent predictor*), or approximated from the difference of the two preceding solution points (*secant predictor*). Various Newton-type iterative corrector algorithms may be used, two of which are described in detail in Section 5.2.1, and the number of iterations may either be fixed or dependent on achieving a desired solution tolerance. Similarly, the step size may either be kept at a pre-defined constant value or adjusted after every step, for instance based on the number of corrector iterations that were required to achieve the desired solution tolerance. Results concerning the existence of a minimum step size for which all computed points lie within a neighborhood of the true solution branch are available for specific instances of the algorithm (see chapter 5 in [7]).

2.3 Direct Collocation Methods

Direct collocation is a discretization method for the approximate solution of differential equations and differential-algebraic equations. The infinite-dimensional problem of finding a function that satisfies the equations everywhere is relaxed to the finite-dimensional problem of determining the unique element in a family of functions parametrized by a finite number of coefficients that satisfies the equations at a corresponding number of so-called *collocation points*. Typical choices for the approximating function family are piecewise or global polynomials of finite degree.

An approximate solution for a boundary value problem involving a system of first-order

ordinary differential equations,

$$\dot{x}(t) = f(x(t)) \qquad \text{for all } \ t \in [0, T], \tag{2.5a}$$

$$0 = \phi(x(0), x(T)), \tag{2.5b}$$

with $x(t) \in \mathbb{R}^{n_x}$ and $\phi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ is obtained as follows: Let $\tilde{x}(c, \cdot)$ be the family of functions used to approximate the solution of (2.5), where $c \in \mathbb{R}^{(K+1)n_x}$ is the coefficient vector that needs to be determined. Choosing K collocation points $\{t_1, \cdots, t_K\}$ with $0 \leq t_j < t_{j+1} \leq T$ leads to the discretized problem

$$\dot{\tilde{x}}(c, t_j) = f\left(\tilde{x}(c, t_j)\right) \qquad \text{for all } \ j \in \{1, \cdots, K\}, \tag{2.6a}$$

$$0 = \phi(\tilde{x}(c, 0), \tilde{x}(c, T)), \tag{2.6b}$$

where $\dot{\tilde{x}}$ is the partial derivative of $\tilde{x}$ with respect to the second argument.

Through direct collocation, the boundary value problem has been transformed into a (square) system of $(K+1)n_x$ algebraic equations. Solving these equations for the coefficient vector c directly leads to the desired approximate solution of the original problem.

When the final time T is not known a priori and part of the solution variables, it is convenient to use the normalized time coordinate $\tau = t/T$ and solve

$$\dot{\tilde{x}}(c, \tau_j) = T f\left(\tilde{x}(c, \tau_j)\right) \qquad \text{for all } \ j \in \{1, \cdots, K\}, \tag{2.7a}$$

$$0 = \phi(\tilde{x}(c, 0), \tilde{x}(c, 1)), \tag{2.7b}$$

where $\tau_j = t_j/T$, to obtain the approximation $x(t) \approx \tilde{x}(c, t/T)$. The normalization assures that $\tilde{x}$ does not depend on T.

When the boundary condition is of simple structure (e.g., $x(0) = x_0$), it may be used to analytically eliminate n_x elements of the coefficient vector, thereby reducing the problem size. The algebraic equations for differential-algebraic problems are also enforced pointwise, either at the same locations as the differential equations or at a different set of collocation points.

The convergence properties of collocation methods for ODEs have been studied extensively in the 1970s [19–22]. It is well known that collocation of first-order ODEs with piecewise polynomials of degree d at the Gauss points leads to convergence with errors on the order of Δ^{d+1} as $\Delta \to 0$, where Δ is the length of the longest polynomial segment, and even higher accuracy (sometimes called *superconvergence*) at the mesh points that divide the segments [19]. Similar convergence results are available for differential-algebraic equations [23].

2.4 Bifurcation Analysis

Many parameter-dependent autonomous dynamical systems can be written as

$$\dot{x}(t) = f(x(t), p), \tag{2.8}$$

where $x(t) \in \mathbb{R}^{n_x}$ is the state vector and the scalar parameter $p \in \mathbb{R}$ is assumed to be static with regards to the system dynamics. The number and location of system equilibrium points where the state derivatives vanish and the associated stability properties are in general dependent on p. The properties and occurrence of other phenomena, e.g., limit cycles, also change with p. If f is well-behaved, then the set of equilibrium points in state-parameter space $\{ (x, p) \in \mathbb{R}^{n_x} \times \mathbb{R} \mid f(x, p) = 0 \}$ consists of continuous one-dimensional objects which are called *equilibrium branches* and may intersect at *branch points*.

Bifurcations are qualitative changes of the system dynamics that result from variations of parameter values, such as the creation of equilibrium points. Bifurcation theory investigates how these changes occur and classifies bifurcation types based on the observed phenomena. Necessary and sufficient conditions are derived that prove or disprove the presence of a bifurcation. Bifurcations of stationary solutions (i.e., equilibria) that are frequently found when f is continuously differentiable include the following:

- **Saddle-node bifurcations** (also called fold bifurcations) arise when a stable and an unstable equilibrium point approach and annihilate each other as the parameter is varied, leading to a fold-like shape in the equilibrium branch. They are associated

with a real eigenvalue of the Jacobian $\nabla_x f$ crossing the imaginary axis. A necessary condition for the presence of a saddle-node bifurcation at (x^*, p^*) is

$$\operatorname{rank} \nabla f(x^*, p^*) = n_x, \tag{2.9a}$$

$$\operatorname{rank} \nabla_x f(x^*, p^*) = n_x - 1, \tag{2.9b}$$

The rank condition shows that saddle-node bifurcations are regular points of f. The fold-like shape results from the fact that the p-component of the unique tangent direction at the bifurcation (which lies in the null space of $\nabla f(x^*, p^*)$) is zero, since the submatrix $\nabla_x f(x^*, p^*)$ is singular and $\nabla f(x^*, p^*)$ has full rank.

- If a complex pair of eigenvalues crosses the imaginary axis instead of a single real eigenvalue, the bifurcation point is called a **Hopf bifurcation**, which is associated with the creation and annihilation of limit cycles. No other eigenvalues are allowed to lie on the imaginary axis at the bifurcation point, which implies that Hopf bifurcations are also regular points of f.

- **Transcritical bifurcations** and **pitchfork bifurcations** occur when two equilibrium branches intersect transversally at a critical point of f. The former is present when the two intersecting branches exchange stability, while the latter implies that only one of the two changes stability. In either case, a real eigenvalue of the linearized dynamics coincides with the imaginary axis at the bifurcation point. Both transcritical and pitchfork bifurcations are common in systems that have certain symmetry properties, but rarely encountered otherwise [4]. A necessary condition for either type of bifurcation at (x^*, p^*) is the rank condition

$$\operatorname{rank} \nabla f(x^*, p^*) = n_x - 1, \tag{2.10a}$$

$$\operatorname{rank} \nabla_x f(x^*, p^*) = n_x - 1, \tag{2.10b}$$

Second order partial derivatives of f evaluated at (x^*, p^*) are required to distinguish between transcritical and pitchfork bifurcations.

The preceding list is far from complete, neglecting both bifurcations of other non-stationary attractors such as limit cycles as well as non-generic cases of stationary bifurcations. A more comprehensive discussion is available in the literature [4, 5]. Bifurcations of non-smooth systems have been investigated in [24].

Bifurcation analysis for a given system is usually carried out in multiple steps. First, the set of equilibrium points (or a subset thereof) is determined. Next, this set is searched for bifurcations by checking the necessary and sufficient conditions for the various bifurcation types. If any are found, the knowledge about the implications of these bifurcations is used to further study the system. Discovery of a Hopf bifurcation, for instance, leads to the study of the limit cycle that is known to emerge at the bifurcation point.

While simple systems with few states can often be analyzed without the help of a computer, more complex systems require the use of numerical codes such as AUTO [25]. These codes compute equilibrium branches with continuation algorithms and detect bifurcation points from changes in the Jacobian matrix of f. Limit cycle computations are performed with the same algorithms through direct collocation of the differential equations [4]. The boundary condition (2.6b) becomes $\tilde{x}(c,T) - \tilde{x}(c,0) = 0$ in this case (for a limit cycle with period T) and is usually used to analytically eliminate n_x components of the coefficient vector c, thereby restricting the approximating function family to periodic functions. It is necessary to include an additional scalar equation (called the *phase condition*) in the problem that defines the initial time along the trajectory and thus resolves the ambiguity arising from the periodic nature of limit cycles [5]. Since the period T is not known a priori, it becomes part of the solution variables. An initial solution point to start the continuation in the vicinity of a Hopf bifurcation is readily obtained from the eigenvalues and eigenvectors of the linearized system at the bifurcation. Local stability off all features is also determined from the linearized dynamics. In the case of limit cycles, this involves computing the *monodromy matrix*, which corresponds to the linearization of the Poincaré map when restricted to the subspace defined by the corresponding Poincaré section [5, 26].

Results from bifurcation analysis are summarized in bifurcation diagrams, Fig. 2.2, which

plot some projection of the equilibrium solution x over the parameter p. For limit cycles, the minimum and/or maximum value over one period is shown. Stability is indicated through coloring and line style, and bifurcation points are included to precisely identify the critical locations.

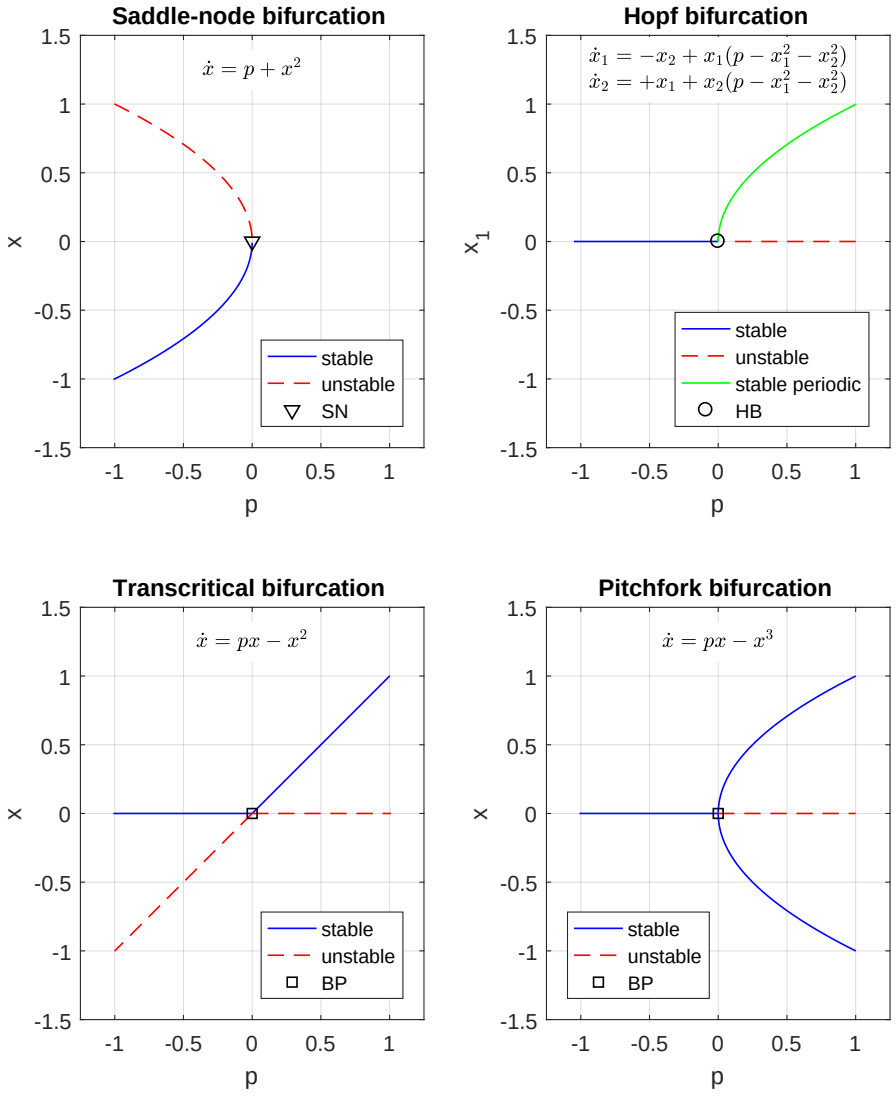

Figure 2.2: Bifurcation diagrams for simple systems that exhibit the most common bifurcation types.

Chapter 3

CONSTRAINED BIFURCATION ANALYSIS

This chapter explores ways to conduct bifurcation analysis of systems with multiple parameters in the presence of equality constraints, which may be imposed in order to address certain analysis questions. Parts of the research presented in this chapter have been published in [27] by the author.

3.1 Constraint-Satisfying Equilibrium Point Subsets

Consider a parameter-dependent autonomous dynamical system similar to (2.8),

$$\dot{x}(t) = f(x(t), p), \tag{3.1}$$

where $x(t) \in \mathbb{R}^{n_x}$ is again the state vector, but $p \in \mathbb{R}^{n_p}$ is now a parameter vector. The function f is assumed to be defined on the domain $\mathcal{D} \subseteq \mathbb{R}^{n_x} \times \mathbb{R}^{n_p}$.

The set of all equilibrium points,

$$\mathcal{M} = \{\, (x, p) \in \mathcal{D} \mid f(x, p) = 0 \,\}, \tag{3.2}$$

consists of (possibly intersecting) n_p-dimensional surfaces if f is well behaved.

For analysis purposes, it is often useful to look at subsets of $\mathcal{M}$ that contain only those equilibrium points that have particular properties as specified by equality constraints of the form $g(x, p) = 0$, where $g : \mathcal{D} \to \mathbb{R}^{n_g}$ and $n_p > n_g$. The set of equilibrium points with the desired characteristic is

$$\mathcal{M}_{(\mathrm{g})} = \{\, (x, p) \in \mathcal{D} \mid f(x, p) = 0,\ g(x, p) = 0 \,\}. \tag{3.3}$$

If the number of constraints is $n_g = n_p - 1$, then solution branches similar to those obtained through numerical bifurcation analysis of single-parameter systems can be computed through

continuation of

$$F_{(\mathrm{g})}(x,p) = \begin{bmatrix} f(x,p) \\ g(x,p) \end{bmatrix} \tag{3.4}$$

if $F_{(\mathrm{g})}$ is well-behaved. Information about the local stability of each point is again obtained from the eigenvalues of $\nabla_x f$, which is a submatrix of

$$\nabla F_{(\mathrm{g})}(x,p) = \begin{bmatrix} \nabla_x f(x,p) & \nabla_p f(x,p) \\ \nabla_x g(x,p) & \nabla_p g(x,p) \end{bmatrix}. \tag{3.5}$$

Note that the p-component of the null space of $\nabla F_{(\mathrm{g})}$ is nonzero at regular points where $\nabla_x f$ is singular, unless the rows of $\nabla_x g$ happen to be linear combinations of the rows of $\nabla_x f$ (or zero) at that point. The characteristic fold-like shape associated with saddle-node bifurcations in single-parameter systems does not occur generically in the solution branches in $\mathcal{M}_g$ when real eigenvalues cross the imaginary axis. Of course, this is only true if g indeed depends on x (if g only depends on p, then $\nabla_x g$ is always zero).

In the next section, it will be shown that singularity of $\nabla_x f$ gives rise to so-called *off-nominal* equilibrium branches that do not satisfy $g(x,p) = 0$, but exist for the parameter combinations that occur for solutions in $\mathcal{M}_{(\mathrm{g})}$. One might expect that a complex conjugate pair of eigenvalues of $\nabla_x f$ crossing the imaginary axis along a solution branch in $\mathcal{M}_{(\mathrm{g})}$ will result in the creation or destruction of a limit cycle. This is indeed the case, but the state trajectory of the cycle will in general also not obey the constraint. Such *off-nominal* limit cycles are also the subject of the next section.

3.2 Nominal and Off-Nominal Branches

All equilibrium solutions found through numerical continuation of $F_{(\mathrm{g})}(x,p) = 0$ in (3.4) satisfy the equality constraints $g(x,p) = 0$ and are called *nominal* solution branches. The off-nominal equilibrium branches and limit cycles that exist for the parameter combinations occurring in the nominal branches (but violate the constraints) may also be identified with continuation methods.

3.2.1 Equilibrium solutions

The set of all nominal *and* off-nominal equilibrium branches is

$$\mathcal{M}_{(\mathrm{g}+)} = \{\, (x,p) \in \mathcal{D} \mid f(x,p) = 0,\ f(x_n,p) = 0,\ g(x_n,p) = 0 \text{ for some } x_n \in \mathbb{R}^{n_x} \,\}. \quad (3.6)$$

Note that $\mathcal{M}_{(\mathrm{g})} \subseteq \mathcal{M}_{(\mathrm{g}+)}$. The above set may be approximated numerically through continuation of the function

$$F_{(\mathrm{g}+)}(x, x_n, p) = \begin{bmatrix} f(x,p) \\ f(x_n,p) \\ g(x_n,p) \end{bmatrix}, \quad (3.7)$$

with

$$\nabla F_{(\mathrm{g}+)}(x, x_n, p) = \begin{bmatrix} \nabla_x f(x,p) & 0 & \nabla_p f(x,p) \\ 0 & \nabla_x f(x_n,p) & \nabla_p f(x_n,p) \\ 0 & \nabla_x g(x_n,p) & \nabla_p g(x_n,p) \end{bmatrix} \quad (3.8)$$

Note that $(x,p) \in \mathcal{M}_{(\mathrm{g})}$ implies $(x,x,p) \in \ker F_{(\mathrm{g}+)}$. The set $\mathcal{M}_{(\mathrm{g}+)}$ is the projection of $\ker F_{(\mathrm{g}+)}$ into two of its components,

$$\mathcal{M}_{(\mathrm{g}+)} = \{\, (x,p) \mid (x, x_n, p) \in \ker F_{(\mathrm{g}+)} \text{ for some } x_n \in \mathbb{R}^{n_x} \,\}. \quad (3.9)$$

Nominal branches intersect with off-nominal branches when real eigenvalues of $\nabla_x f$ cross the imaginary axis, since $\nabla F_{(\mathrm{g}+)}$ drops rank at these points:

Proposition 3.1. *Let (x^*, p^*) be a regular point of $F_{(g)}$. Then (x^*, x^*, p^*) is a critical point of $F_{(g+)}$ if and only if $\nabla_x f(x^*, p^*)$ is singular.*

Proof. The rows of the $(n_x + n_g) \times (n_x + n_p)$-matrix

$$\nabla F_{(\mathrm{g})}(x^*, p^*) = \begin{bmatrix} \nabla_x f(x^*,p^*) & \nabla_p f(x^*,p^*) \\ \nabla_x g(x^*,p^*) & \nabla_p g(x^*,p^*) \end{bmatrix}$$

form a linearly independent set, because (x^*, p^*) is a regular point of $F_{(\mathrm{g})}$ and $n_p > n_g$. Due to the structure of $\nabla F_{(\mathrm{g}+)}$, this means that the rows of $\nabla F_{(\mathrm{g}+)}(x^*, x^*, p^*)$ form a linearly

independent set whenever $\nabla_x f(x^*, p^*)$ is invertible. Conversely, $n_p > n_g$ implies that there exist matrices $V_x \in \mathbb{R}^{n_x \times (n_p - n_g)}$ and $V_p \in \mathbb{R}^{n_p \times (n_p - n_g)}$ such that

$$\operatorname{rank} \begin{bmatrix} V_x \\ V_p \end{bmatrix} = n_p - n_g \qquad \text{and} \qquad \begin{bmatrix} \nabla_x f(x^*, p^*) & \nabla_p f(x^*, p^*) \\ \nabla_x g(x^*, p^*) & \nabla_p g(x^*, p^*) \end{bmatrix} \begin{bmatrix} V_x \\ V_p \end{bmatrix} = 0,$$

If $\nabla_x f(x^*, p^*)$ is singular, then there also exists a nonzero vector $v_x \in \mathbb{R}^{n_x}$ with $\nabla_x f(x^*, p^*) v_x = 0$ so that

$$\begin{bmatrix} \nabla_x f(x^*, p^*) & 0 & \nabla_p f(x^*, p^*) \\ 0 & \nabla_x f(x^*, p^*) & \nabla_p f(x^*, p^*) \\ 0 & \nabla_x g(x^*, p^*) & \nabla_p g(x^*, p^*) \end{bmatrix} \begin{bmatrix} v_x & V_x \\ 0 & V_x \\ 0 & V_p \end{bmatrix} = 0.$$

The null space of $\nabla F_{(\mathrm{g}+)}(x^*, x^*, p^*)$ is at least of dimension $n_p - n_g + 1$, which means that $\nabla F_{(\mathrm{g}+)}$ does not have full rank at (x^*, x^*, p^*). ■

3.2.2 Off-nominal limit cycles

The crossing of the imaginary axis by a complex conjugate pair of eigenvalues along a nominal branch is associated with the creation or destruction of an off-nominal limit cycle:

Proposition 3.2. *Suppose f and g have continuous second derivatives and $n_p = n_g + 1 > 1$. Let λ_k, $k \in \{1, \cdots, n_x\}$ be continuous functions of x and p such that $\{\lambda_k(x, p)\}$ are the eigenvalues of $\nabla_x f(x, p)$ and let $(x^*, p^*) \in \ker F_{(g)}$ be a regular point of $F_{(g)}$ with $\lambda_1(x^*, p^*) = i\omega$, $\lambda_2(x^*, p^*) = -i\omega$, $\omega > 0$, and $\operatorname{Re} \lambda_k(x^*, p^*) \neq 0$ for $k > 2$. Furthermore, assume that $[\nabla_x \operatorname{Re} \lambda_1(x^*, p^*) \quad \nabla_p \operatorname{Re} \lambda_1(x^*, p^*)] v \neq 0$ for any v in the null space of $\nabla F_{(g)}(x^*, p^*)$. Then an off-nominal limit cycle emerges from the nominal equilibrium branch at (x^*, p^*).*

Proof. Let $(\bar{x}, \bar{p}) : (-1, 1) \to \mathbb{R}^{n_x} \times \mathbb{R}^{n_p}$ be a parametrization of the graph of the implicit function defined by $F_{(\mathrm{g})} = 0$ in a neighborhood of (x^*, p^*) such that $\bar{x}(0) = x^*$ and $\bar{p}(0) = p^*$, i.e., $F_{(\mathrm{g})}(\bar{x}(s), \bar{p}(s)) = 0$ for all $s \in (-1, 1)$. Now define the function $\bar{f} : \mathbb{R}^{n_x} \times (-1, 1) \to \mathbb{R}^{n_x}$ by $\bar{f}(x, s) = f(x, \bar{p}(s))$. Note that $(\bar{x}, \bar{p})$ is twice differentiable by the implicit function theorem, and hence $\bar{f}$ is also twice differentiable. Furthermore, $\bar{f}(x^*, 0) = 0$ holds and $\{\lambda_k(x, \bar{p}(s))\}$

are the eigenvalues of $\nabla_x \bar{f}(x,s)$, so that $\nabla_x \bar{f}(x^*,0)$ has a complex conjugate pair of eigenvalues on the imaginary axis while all other eigenvalues have nonzero real part. Note that $\nabla_s \lambda_1(\bar{x},\bar{p})(0) = \nabla_x \mathrm{Re}\,\lambda_1(x^*,p^*)\nabla_s \bar{x}(0) + \nabla_p \mathrm{Re}\,\lambda_1(x^*,p^*)\nabla_s \bar{p}(0)$ since, by the implicit function theorem, $\begin{bmatrix}\nabla_s \bar{x}(0)^T & \nabla_s \bar{x}(0)^T\end{bmatrix}^T$ is in the null space of $\nabla F_{(\mathrm{g})}(x^*,p^*)$. Thus $\bar{f}$ satisfies the conditions of Theorem 2.11 in [5], which states that the system $\dot{x}(t) = \bar{f}(x(t),s)$ undergoes a Hopf bifurcation at $(x^*,0)$ as the parameter s changes sign. The associated limit cycle exists in the original system $\dot{x}(t) = f(x(t),p)$ for the parameter combinations that occur in the nominal branch which includes (x^*,p^*), but the cycle's state trajectory does not in general satisfy the constraint. ■

To compute these off-nominal limit cycles, the standard problem for the continuation of limit cycles in numerical bifurcation analysis is extended to include the nominal equilibrium x_n and the constraints similar to the off-nominal equilibrium branch case. This leads to

$$F_{(\mathrm{g+lc})}(c,x_n,p,T) = \begin{bmatrix} \dot{\tilde{x}}(c,\tau_1) - Tf\left(\tilde{x}(c,\tau_1),p\right) \\ \vdots \\ \dot{\tilde{x}}(c,\tau_K) - Tf\left(\tilde{x}(c,\tau_K),p\right) \\ \theta(c,p) \\ f(x_n,p) \\ g(x_n,p) \end{bmatrix}, \tag{3.10}$$

where $c \in \mathbb{R}^{Kn_x}$ is the coefficient vector of the parametrized limit cycle and $\theta : \mathbb{R}^{Kn_x} \times \mathbb{R}^{n_p} \to \mathbb{R}$ defines the phase condition that determines the initial time along the periodic trajectory.

The orbital stability [26] of the off-nominal limit cycle is determined from the eigenvalues of the monodromy matrix, which is computed exactly as in the case of unconstrained bifurcation analysis [5].

3.2.3 Plotting of results

To facilitate the interpretation of results, it is recommended to choose an injective projection $y = h(x_n,p)$ of the nominal state and parameter vector for the horizontal axis of any plot

containing nominal and off-nominal branches (i.e., h is "one-to-one"). The plot then shows the system features that exist in the dynamics as a function of y, each value of which is associated with a unique value of x_n and p.

3.2.4 Comparison with two-step procedure from literature

A two-step procedure (named extended bifurcation analysis procedure) has been proposed in the literature [28–30] to compute off-nominal solution branches with legacy continuation codes for numerical bifurcation analysis such as AUTO [25]. The two steps consist of a first continuation run to calculate the nominal, i.e. constraint-satisfying roots, followed by a second run where $n_p - 1$ of the parameters are scheduled as a function of the remaining parameter, based upon the results from the first run. The solution set to this second, unconstrained problem, contains the nominal roots of the system that satisfy $g(x, p) = 0$ as well as the off-nominal features that exist for the parameter combinations resulting from the equality constraints.

A major disadvantage of this approach is that the scheduling of $m - 1$ parameters as a function of the free parameter in the second step has to be done by interpolating between the solution points found in the first step. The error introduced by the interpolation destroys the symmetry properties that lead to the occurrence of branch points [4] where nominal and off-nominal branches intersect, as branch points are known to vanish under perturbations [7]. For this reason, application of the two-step procedure is frequently plagued by numerical difficulties which are not encountered when the off-nominal branches are computed with the method derived earlier in this section.

3.3 Zero-Dynamics Bifurcation Analysis

If one of the parameters is a system input and the constraint is defined by a (scalar) system output that is forced to zero at all times by manipulation of the input, then the set of all points in state space that satisfy the constraint is rendered invariant. This means that the state trajectory remains in the set for any constraint-satisfying initial condition. This section

deals with bifurcation analysis of the dynamics restricted to this invariant set, commonly referred to as the *zero dynamics*.

The system is assumed to have the following single input, single output structure throughout this section:

$$\dot{x}(t) = f(x(t), u(t), p) \tag{3.11a}$$

$$y(t) = h(x(t), p) \tag{3.11b}$$

In the above $p \in \mathbb{R}$ is a scalar parameter, $u(t) \in \mathbb{R}$ is the control input, and $y(t) \in \mathbb{R}$ is the output.

The system is said to have *relative degree* r at (x^*, u^*) if f and h satisfy $\nabla_u L_f^k h(x, u) = 0$ for all (x, u) in a neighborhood of (x^*, u^*) and all $0 \leq k < r$ and $\nabla_u L_f^r h(x^*, u^*) \neq 0$, where $L_f^k h$ is the k-th Lie derivative of h with respect to f, which is defined recursively as $L_f^k h = \nabla_x L_f^{k-1} h\, f$ with $L_f^0 h = h$, see [31].

If the system has relative degree $r < n_x$ at (x^*, u^*), then there exists a locally defined coordinate transformation such that the system has the form

$$\dot{\xi}_k(t) = \xi_{k+1}(t) \qquad \text{for all } k \in \{1, \cdots, r-1\} \tag{3.12a}$$

$$\dot{\xi}_r(t) = a(\xi(t), \eta(t), u(t), p) \tag{3.12b}$$

$$\dot{\eta}(t) = q(\xi(t), \eta(t), p) \tag{3.12c}$$

$$y(t) = \xi_1(t) \tag{3.12d}$$

in the new coordinates, with $\xi(t) = [\xi_1(t) \cdots \xi_r(t)]^T \in \mathbb{R}^r$ and $\eta(t) \in \mathbb{R}^{n_x - r}$ [31]. Requiring the output to be zero at all times implies $\xi(t) = 0$, so that the zero dynamics are governed by

$$\dot{\eta}(t) = q(0, \eta(t), p) \tag{3.13}$$

in a neighborhood of (x^*, u^*). This directly leads to the following observation:

Proposition 3.3. *Let the system defined in* (3.11) *have relative degree* $r < n_x$ *at* (x^*, u^*). *Then its zero dynamics may exhibit any bifurcation phenomenon encountered in parameter-dependent autonomous nonlinear systems with a state space of dimension* $n_x - r$.

Proof. The zero dynamics state equation (3.13) defines a parameter-dependent autonomous nonlinear system of dimension $n_x - r$ similar to (2.8). To see that there are in general no restrictions on the form of q in (3.12c), consider that the system may already have the structure of (3.12) in the original coordinates. ■

Determining the nonlinear coordinate transformation that takes the system to the normal form in (3.12) is nontrivial in general. The remainder of this section is concerned with numerical bifurcation analysis of the zero dynamics in the original coordinates, thus enabling direct application of the method to the original system model. Such analysis may for instance support the identification of suitable input-output pairs for the design of nonlinear control laws based on feedback linearization [32].

3.3.1 Equilibrium points

Analogous to $F_{(\mathrm{g})}$ in (3.4), equilibrium branches of the zero dynamics are computed through continuation of

$$F_{(\mathrm{zd})}(x,u,p) = \begin{bmatrix} f(x,u,p) \\ h(x,p) \end{bmatrix} \tag{3.14}$$

with

$$\nabla F_{(\mathrm{zd})}(x,u,p) = \begin{bmatrix} \nabla_x f(x,u,p) & \nabla_u f(x,u,p) & \nabla_p f(x,u,p) \\ \nabla_x h(x,p) & 0 & \nabla_p h(x,p) \end{bmatrix}. \tag{3.15}$$

However, local stability can no longer be determined from the eigenvalues of $\nabla_x f$ evaluated at the equilibrium, since the input is assumed to respond to any (allowable) disturbance in such a way that the output remains at zero. Instead, local stability of equilibrium points in the zero dynamics is associated with the transmission zeros of the input-output transfer function of the linearized system. These transmission zeros are known to coincide with the eigenvalues of the linearized zero dynamics [32] and are readily computed [33] through numerical solution of the generalized eigenvalue problem

$$\begin{bmatrix} \nabla_x f(x,u,p) & \nabla_u f(x,u,p) \\ \nabla_x h(x,p) & 0 \end{bmatrix} \begin{bmatrix} v_x \\ v_u \end{bmatrix} = \lambda \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_x \\ v_u \end{bmatrix} \tag{3.16}$$

Note that saddle-node bifurcations in the zero dynamics result in the same fold-like shape in the equilibrium branch as in the familiar case of unconstrained dynamics shown in Figure 2.2: The p-component of the tangent vector at that point is zero, because $[v_x^T\ v_u^T\ 0]^T$ is in the null space of $\nabla F_{(\mathrm{zd})}$ when λ is zero in (3.16) as a real transmission zero of the linearized system crosses the imaginary axis.

When a complex conjugate pair of transmission zeros crosses the imaginary axis along an equilibrium branch, then the zero dynamics experience a Hopf bifurcation (since the transmission zeros of the linearized system are the eigenvalues of the linearized zero dynamics). A limit cycle in the zero dynamics is therefore created or destroyed at such points.

3.3.2 Limit cycles in the zero dynamics

As the input is assumed to adjust its value in such a way that the output remains at zero for all times, it will in general not be constant along a system trajectory that forms a limit cycle in the zero dynamics. Numerical computation of such limit cycles therefore requires not only finite-dimensional approximation of the state trajectory, but also of the control trajectory. Let $\tilde{x}(c_x, \cdot)$ and $\tilde{u}(c_u, \cdot)$ be suitable approximating function families parametrized by their arguments c_x and c_u that satisfy the periodicity conditions $\tilde{x}(c_x, 1) = \tilde{x}(c_x, 0)$ and $\tilde{u}(c_u, 1) = \tilde{u}(c_u, 0)$. Direct collocation of the differential-algebraic equations leads to the function

$$F_{(\mathrm{zd,lc})}(c_x, c_u, p, T) = \begin{bmatrix} \dot{\tilde{x}}(c_x, \tau_1) - T f\left(\tilde{x}(c_x, \tau_1), \tilde{u}(c_u, \tau_1), p\right) \\ h\left(\tilde{x}(c_x, \tau_1), p\right) \\ \vdots \\ \dot{\tilde{x}}(c_x, \tau_K) - T f\left(\tilde{x}(c_x, \tau_K), \tilde{u}(c_u, \tau_K), p\right) \\ h\left(\tilde{x}(c_x, \tau_K), p\right) \\ \theta(c_x, p) \end{bmatrix} \tag{3.17}$$

where θ defines the phase condition and τ_j are the collocation points, with $j \in \{1, \cdots, K\}$. The c_x and T components of the elements in $\ker F_{(\mathrm{zd,lc})}$ define the desired approximations of the system's limit cycles as $x(t) \approx \tilde{x}(c_x, t/T)$. The associated control that forces the output

to zero along the cycle is approximated as $u(t) \approx \tilde{u}(c_u, t/T)$.

Continuation of $F_{(\text{zd,lc})}$ may be initialized at a Hopf bifurcation $(x^*, u^*, p^*) \in \ker F_{(\text{zd})}$ based on the generalized eigenvector $[v_x^T \; v_u^T]^T$ pertaining to the generalized eigenvalue $\lambda = i\omega$ associated with the bifurcation in (3.16). The state offset and input offset trajectories

$$\Delta x(t) = x(t) - x^* = \operatorname{Re} v_x \cos \omega t - \operatorname{Im} v_x \sin \omega t \tag{3.18a}$$

$$\Delta u(t) = u(t) - u^* = \operatorname{Re} v_u \cos \omega t - \operatorname{Im} v_u \sin \omega t \tag{3.18b}$$

are known to satisfy the linearized dynamics while keeping the output of the linearized system at zero [34]. Since the linearized system approximates the nonlinear system to first order, the periodic functions defined by $x^* + \epsilon\, \Delta x(t)$ and $u^* + \epsilon\, \Delta u(t)$ for $0 < \epsilon \ll 1$ are "close" to the limit cycle that emerges from the Hopf bifurcation. Let π_{c_x} and π_{c_u} be projections from the space of periodic functions into coefficient space of the approximating function families satisfying $F_{(\text{zd,lc})}(\pi_{c_x}(\breve{x}), \pi_{c_u}(\breve{u}), p, T) = 0$ when $(\breve{x}, \breve{u})$ is a periodic system trajectory with period T for parameter value p. Then an initial point to start continuation of $F_{(\text{zd,lc})}$ is obtained as $\big(\pi_{c_x}(x^* + \epsilon\, \Delta x), \pi_{c_u}(u^* + \epsilon\, \Delta u), p^*, 2\pi/\omega\big)$.

The orbital stability of limit cycles in the zero dynamics may be determined from the eigenvalues of the *zero dynamics monodromy matrix*, which is introduced in the following theorem. For conciseness, the functional dependence of f and h on the parameter p is dropped throughout the theorem and proof, as it is assumed to be fixed.

Theorem 3.4. *Let the periodic trajectory defined by* $\breve{x} : [0, T] \to \mathbb{R}^{n_x}$ *and* $\breve{u} : [0, T] \to \mathbb{R}^{n_u}$ *with* $\breve{x}(T) = \breve{x}(0)$ *and* $\breve{u}(T) = \breve{u}(0)$ *satisfy*

$$\dot{\breve{x}}(t) = f(\breve{x}(t), \breve{u}(t)) \tag{3.19a}$$

$$0 = h(\breve{x}(t)) \tag{3.19b}$$

with $\dot{\breve{x}}(0) \neq 0$ *and suppose the system has relative degree* r *at all points on the trajectory. Furthermore, let the columns of* $Z \in \mathbb{R}^{n_x \times n_x - r}$ *form an orthonormal basis for the null space of the matrix whose rows are formed by* $\nabla_x L_f^k h(\breve{x}(0), \breve{u}(0))$ *for* $0 \leq k < r$.

Then there exist $\hat{\Phi} : [0,T] \to \mathbb{R}^{n_x \times (n_x - r)}$, $\hat{\Gamma} : [0,T] \to \mathbb{R}^{n_u \times (n_x - r)}$ *satisfying*

$$\dot{\hat{\Phi}}(t) = \nabla_x f(\breve{x}(t), \breve{u}(t))\hat{\Phi}(t) + \nabla_u f(\breve{x}(t), \breve{u}(t))\hat{\Gamma}(t) \tag{3.20a}$$

$$0 = \nabla_x h(\breve{x}(t), \breve{u}(t))\hat{\Phi}(t) \tag{3.20b}$$

$$\hat{\Phi}(0) = Z \tag{3.20c}$$

and the zero-dynamics monodromy matrix $M_z = Z^T \hat{\Phi}(T)$ *has an eigenvalue at 1. Furthermore, the periodic trajectory is orbitally stable if this eigenvalue has algebraic multiplicity one and all other eigenvalue lie within the open unit circle. It is unstable if at least one eigenvalue has magnitude greater than one.*

Proof. The system has relative degree r on the trajectory, which means that $\nabla_u L_f^r h(\breve{x}(t), \breve{u}(t))$ is nonzero for all $t \in [0,T]$. By the implicit function theorem, this implies that the equation $L_f^r h(x,u) = 0$ is locally solvable for u in a neighborhood of the trajectory. Therefore, there exists an open neighborhood $\mathcal{X}_0$ of $\breve{x}(0)$ and two functions $\hat{x} : [0, T+\epsilon) \times \mathcal{X}_0 \to \mathbb{R}^{n_x}$ and $\hat{u} : [0, T+\epsilon) \times \mathcal{X}_0 \to \mathbb{R}^{n_u}$ for any $\epsilon > 0$ that satisfy

$$\dot{\hat{x}}(t, x_0) = f(\hat{x}(t, x_0), \hat{u}(t, x_0)) \tag{3.21a}$$

$$0 = L_f^r h(\hat{x}(t, x_0), \hat{u}(t, x_0)) \tag{3.21b}$$

$$\hat{x}(0, x_0) = x_0 \tag{3.21c}$$

where the abbreviation $\dot{\hat{x}} = \nabla_t \hat{x}$ is used. Integrating (3.21b) r times with respect to t and considering the initial condition (3.21c) leads to the equivalent equation

$$\sum_{k=0}^{r-1} L_f^k h(x_0, \hat{u}(0, x_0)) t^k = h(\hat{x}(t, x_0)) \tag{3.22}$$

Differentiating both sides of (3.21a), (3.22), and (3.21c) with respect to x_0 gives

$$\begin{aligned} \nabla_{x_0} \dot{\hat{x}}(t, x_0) = {} & \nabla_x f(\hat{x}(t, x_0), \hat{u}(t, x_0)) \nabla_{x_0} \hat{x}(t, x_0) \\ & + \nabla_u f(\hat{x}(t, x_0), \hat{u}(t, x_0)) \nabla_{x_0} \hat{u}(t, x_0) \end{aligned} \tag{3.23a}$$

$$\sum_{k=0}^{r-1} \nabla_x L_f^k h(x_0, \hat{u}(0, x_0)) t^k = \nabla_x h(\hat{x}(t, x_0)) \nabla_{x_0} \hat{x}(t, x_0) \tag{3.23b}$$

$$\nabla_{x_0} \hat{x}(0, x_0) = I \tag{3.23c}$$

Now exchange the order of differentiation on the left side of (3.23a), multiply (3.23a)–(3.23c) with Z from the right, and define $\hat{\Phi}(t) = \nabla_{x_0}\hat{x}(t, \breve{x}(0))Z$ as well as $\hat{\Gamma}(t) = \nabla_{x_0}\hat{u}(t, \breve{x}(0))Z$ to obtain (3.20). Note that by definition, Z is orthogonal to all terms on the left side of (3.23b).

Periodicity of $\breve{x}$ and (3.21c) lead to $\hat{x}(T, \breve{x}(t)) = \hat{x}(0, \breve{x}(t)) = \breve{x}(t)$, which yields

$$\nabla_{x_0}\hat{x}(T, \breve{x}(t))\,\dot{\breve{x}}(t) = \dot{\breve{x}}(t) \tag{3.24}$$

when differentiated with respect to t. Let $Y \in \mathbb{R}^{n_x \times r}$ be such that $[Z\ Y]$ is an orthonormal matrix and note that $I = ZZ^T + YY^T$. From the definition of Z and (3.19b), it follows that $Y^T\dot{\breve{x}}(0) = 0$ holds true. Multiplying both sides of (3.24) with Z^T from the left and inserting the identity matrix between the two terms on the left hand side demonstrates that M_Z has an eigenvalue of 1 with eigenvector $Z^T\dot{\breve{x}}(0)$,

$$Z^T\nabla_{x_0}\hat{x}(T, \breve{x}(t))(ZZ^T + YY^T)\dot{\breve{x}}(t) = Z^T\dot{\breve{x}}(t) \tag{3.25}$$

$$\implies \qquad \underbrace{Z^T\nabla_{x_0}\hat{x}(T, \breve{x}(0))Z}_{=M_z}\, Z^T\dot{\breve{x}}(0) = Z^T\dot{\breve{x}}(0) \tag{3.26}$$

Let $\widetilde{\mathcal{X}}_0$ be the subset of $\mathcal{X}_0$ whose members satisfy $L_f^k h(x,u) = 0$ for $k \in \{0, \cdots, r\}$. By the implicit function theorem, applied at $(\breve{x}(0), \breve{u}(0))$, the set $\widetilde{\mathcal{X}}_0$ may be parametrized as a function of $n_x - r$ coordinates. Let $\mathcal{S}$ be an open neighborhood of the origin in $\mathbb{R}^{n_x - r}$ such that $\chi : \mathcal{S} \to \widetilde{\mathcal{X}}_0$ is such a parametrization satisfying $\chi(0) = \breve{x}(0)$ and $\nabla\chi(0) = Z$. Note that by construction, the inverse map $\chi^{-1} : \widetilde{\mathcal{X}}_0 \to \mathcal{S}$ exists. Differentiating both sides of the equality $\chi^{-1}(\chi(s)) = s$ leads to $\nabla\chi^{-1}(\chi(s))\nabla\chi(s) = I$, so that $\nabla\chi^{-1}(\breve{x}(0))Z = I$ holds. Because the columns of Z are orthonormal, it follows that $\nabla\chi^{-1}(\breve{x}(0)) = Z^T$. Also note that $\hat{x}(t, \chi(s)) \in \widetilde{\mathcal{X}}_0$ for all s in a subset $\mathcal{S}_r$ of $\mathcal{S}$ containing the origin and all t in a neighborhood $\mathcal{T}_r$ of T. Now define the map $f_s : \mathcal{T}_r \times \mathcal{S}_r \to \mathcal{S}$ as $f_s(t,s) = \chi^{-1}\big(\hat{x}\big(t, \chi(s)\big)\big)$ and see that

$$\nabla_s f_s(T,0) = \nabla\chi^{-1}\big(\underbrace{\hat{x}\big(T, \breve{x}(0)\big)}_{=\breve{x}(0)}\big)\ \underbrace{\nabla_{x_0}\hat{x}\big(T, \breve{x}(0)\big)\,\nabla\chi(0)}_{=\hat{\Phi}(T)} = Z^T\hat{\Phi}(T) = M_z \tag{3.27}$$

So the zero-dynamics monodromy matrix is the linearization of the map that associates all points in $\widetilde{\mathcal{X}}_0$ with their position after being propagated for one period of the cycle. This

map is locally stable when all eigenvalues of M_z not associated with the periodic motion have magnitude less than one, and unstable if at least one eigenvalue has magnitude greater than one. ■

The collocation equations for approximation of limit cycles in the zero dynamics with piecewise polynomials are derived in Appendix A, as are the equations for computation of the associated zero-dynamics monodromy matrix.

3.4 Examples

The last section of this chapter presents two examples which are analyzed with the methods presented in the preceding sections.

3.4.1 Intersection of nominal and off-nominal equilibrium branches

The first example illustrates how off-nominal equilibrium branches intersect the constraint-satisfying branches when a real eigenvalue of $\nabla_x f$ crosses the imaginary axis at a regular point of $F_{(\mathrm{g})}$. The set of equilibrium points of the system

$$\dot{x} = f(x,p) = \left(x - \frac{p_2}{8}\right)^2 - p_1 \tag{3.28}$$

with one state and two parameters is a two-dimensional surface embedded in state-parameter space, plotted in Figure 3.1 as $\ker f$. For any fixed value of p_2, the system has a saddle-node bifurcation at $p_1 = 0$ and $x = p_2/8$. These bifurcation points form a line that subdivides $\ker f$ into a lower part of stable equilibrium points and an upper part of unstable equilibrium points.

Now suppose the constraint

$$g(x,p) = x - p_2 = 0 \tag{3.29}$$

is imposed to select a subset of $\ker f$ for analysis purposes. The set of equilibrium points satisfying these constraints lie in the intersection of $\ker f$ and $\ker g$, shown in the Figure 3.1,

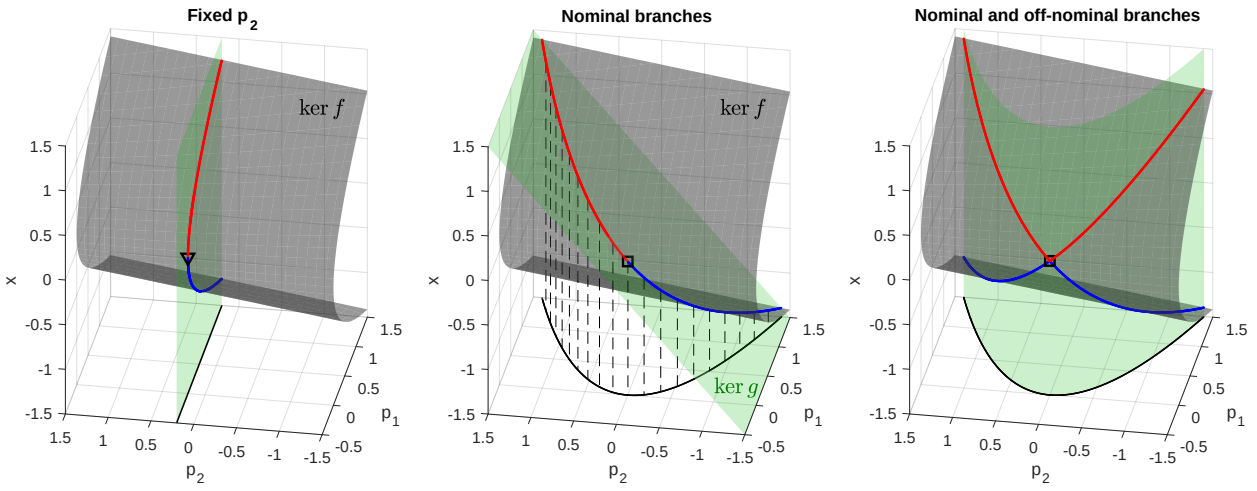

Figure 3.1: Equilibrium branches form subsets of $\ker f$. Stable and unstable branches are plotted in blue and red, respectively.

and comprises a branch of stable equilibrium points (plotted in blue) and a branch of unstable equilibrium points (plotted in red).

The combinations of p_1 and p_2 that occur in these nominal branches are shown in Figure 3.1 as the projection of the branches into the parameter plane. Apart from the nominal equilibrium branches, further equilibrium points exist for these parameter combinations that do not satisfy the constraint. These off-nominal branches are plotted on the right in Figure 3.1 as the intersection of $\ker f$ and the surface obtained by extending the parameter-plane projection of the nominal branches into state-parameter space. The two surfaces have a parabolic shape whose extremal lines intersect at the branch point where the nominal and off-nominal branches meet. Note that a small random perturbation in either of these surfaces (as are introduced by the interpolation in the two-step procedure discussed in 3.2.4) would move these lines such that they no longer intersect and cause the branch point to vanish.

3.4.2 Zero-dynamics bifurcation analysis of modified van-der-Pol oscillator

Consider the single input, single output system defined in (3.11) with

$$f(x,u,p) = \begin{bmatrix} x_2 + \frac{u}{4} \\ -x_1 + (x_3 - x_1^2 - 1)x_2 \\ x_2 - x_3 + u \end{bmatrix} \tag{3.30a}$$

$$h(x,p) = x_3 - p \tag{3.30b}$$

The system has relative degree one everywhere, since $\nabla_u L_f^1 h = 1$, and will be used to demonstrate bifurcation analysis of the zero dynamics. The parameter p may be understood as a setpoint for the third state, so that the output is the offset from that setpoint.

With a feedback control law of the form

$$u = -x_2 + x_3 - k(x_3 - p), \tag{3.31}$$

the closed-loop dynamics for the output become $\dot{y} = \dot{x}_3 = -k(x_3 - p) = -ky$, thus driving the output to zero when $k > 0$. Furthermore, when the output is at zero, then it remains zero at all time. The behavior of the zero dynamics, however, is strongly dependent on the particular value of p. Figure 3.2 shows time simulation results of the closed-loop system for three different values of p, where the inital state was chosen as the origin. Depending on p, the first two states either diverge, converge towards an equilibrium condition, enter a stable limit cycle, or diverge after completing what resembles a single cycle of the periodic motion observed for other parameter values.

The qualitative changes that the zero dynamics undergo as p changes are readily interpreted by inspection of the zero dynamics bifurcation diagram shown on the left in Figure 3.3, which was generated with the methods derived in Section 3.3. For $-1.1 < p < 1$, the zero dynamics have a stable equilibrium point. System trajectories that resemble the second simulation case in Figure 3.2 are thus expected in this parameter regime. At $p = -1.1$, the zero dynamics experience a saddle-node bifurcation at which the stable equilibrium vanishes, so that the trajectory from the simulation with $p = -2$ diverges. A stable limit cycle is created

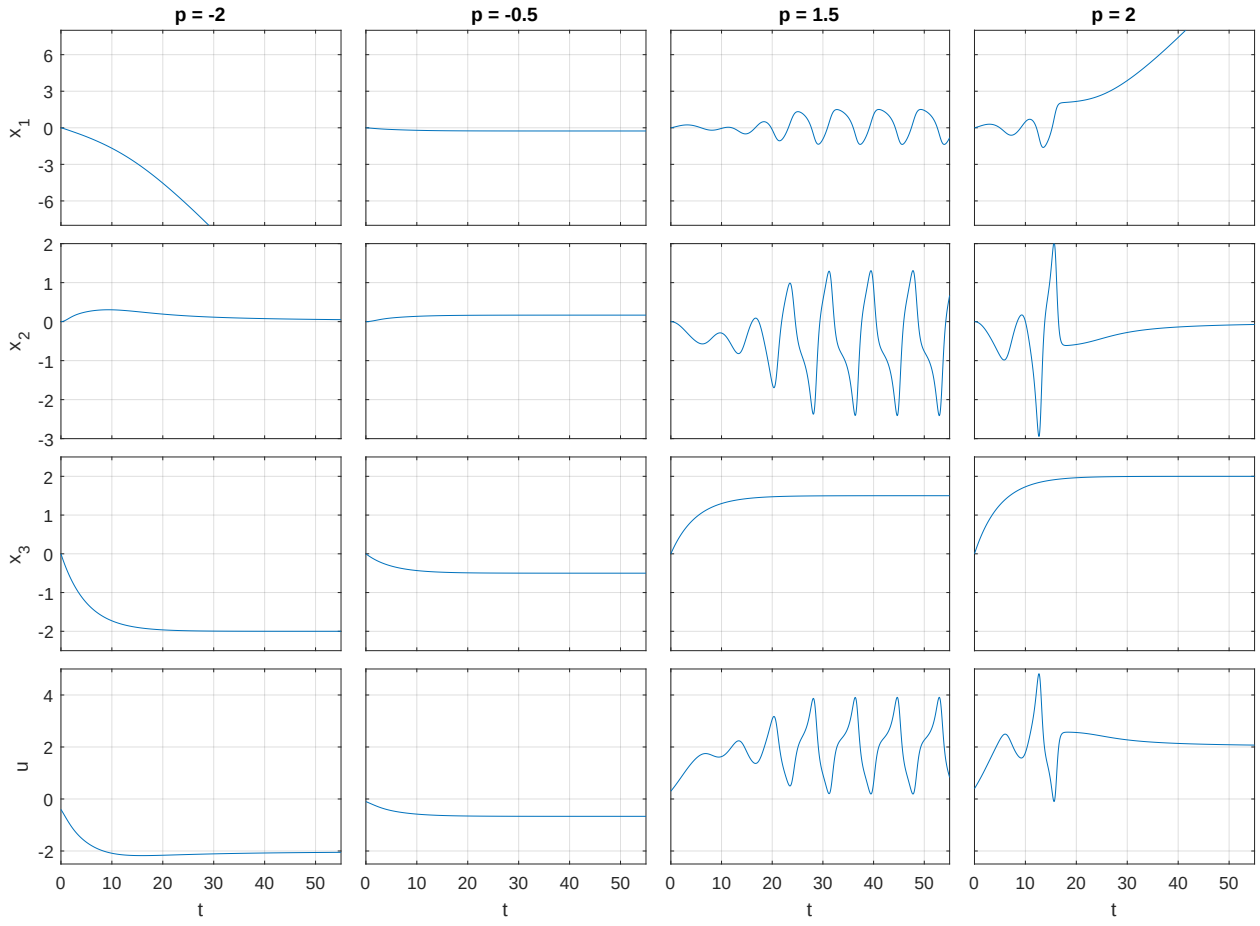

Figure 3.2: Time simulation of the closed-loop system with different setpoints for x_3.

in the zero dynamics at $p = 1$, which undergo a Hopf bifurcation at this parameter value. The cycle exists until p reaches 1.85, at which point it collides with an unstable equilibrium point in a homoclinic bifurcation. The green curve in Figure 3.3 corresponds to the maximum value x_1 achieves along the cycle. The shape of the limit cycle for various parameter values, computed through continuation of $F_{(\mathrm{zd,lc})}$, is plotted on the right in Figure 3.3. Note how the limit cycle develops a sharp corner as it gets closer to the unstable equilibrium point with increasing p. At the homoclinic bifurcation, the corner point and the equilibrium merge as the homoclinic orbit (a closed trajectory that starts and ends at the same equilibrium point) is created. The trajectory of the third simulation case converges to the limit cycle

with $p = 1.5$ on the right in Figure 3.3. The last simulation case illustrates how the state trajectory first approaches the homoclinic orbit that exist for $p = 2$, thus resembling a single cycle of a periodic motion, only to diverge along the unstable manifold of the equilibrium point on the orbit.

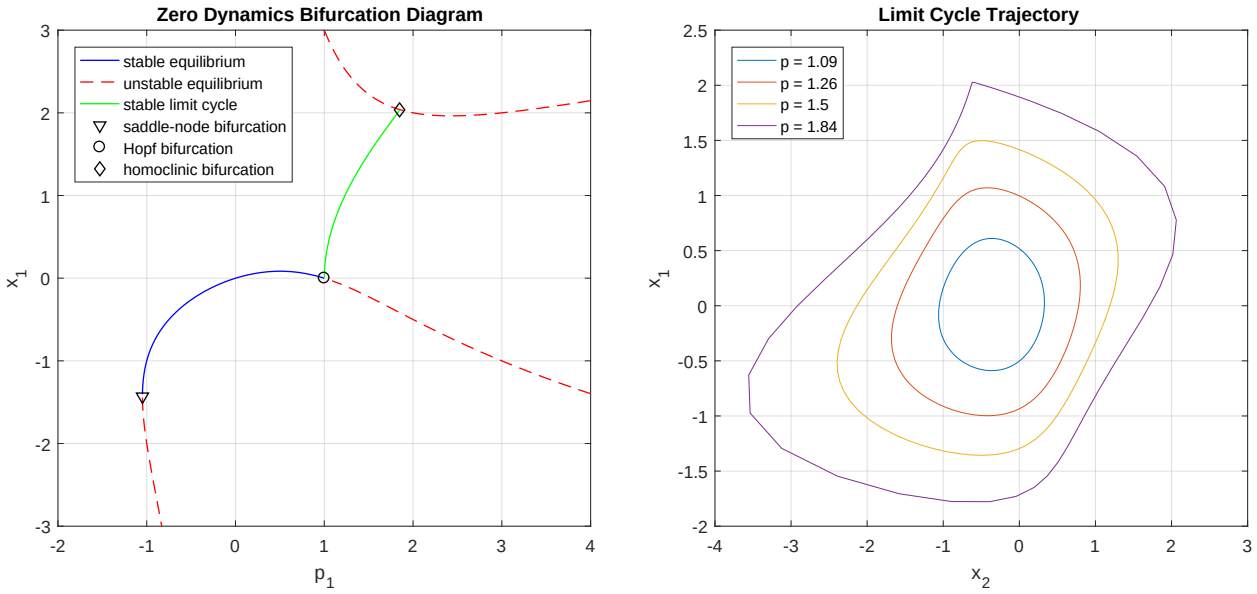

Figure 3.3: Zero dynamics bifurcation diagram and limit cycle shapes for the second example.

Chapter 4

GLOBAL ANALYSIS OF LOCAL DYNAMICAL PROPERTIES

Many local properties that characterize the system dynamics in the neighborhood of an equilibrium point depend on the eigenvalues and eigenvectors of the linearized system. Similar to Section 3.1, this chapter derives methods to compute equilibrium point subsets of particular interest with continuation methods. However, the equality constraints that specify the properties of these subsets are now allowed to include conditions on the eigenmodes of the linearized dynamics. Such constraints enable targeted analysis not only of the system's steady-state behavior, but also of dynamic characteristics beyond local stability. Results of such analysis reveal how these properties depend on the particular operating conditions of the system. Much of the research presented in this chapter have been published in [27,35,36] by the author.

4.1 Continuation of Equilibrium Solutions Subject to Eigenmode Constraints

The systems of consideration throughout this chapter are assumed to have the form

$$\dot{x}(t) = f(x(t), u(t), p) \tag{4.1a}$$

$$y(t) = h(x(t), u(t), p) \tag{4.1b}$$

where $x(t) \in \mathbb{R}^{n_x}$ is the state, $u(t) \in \mathbb{R}^{n_u}$ the input, $y(t) \in \mathbb{R}^{n_y}$ the output, and $p \in \mathbb{R}^{n_p}$ a vector of static system parameters. The functions f and h are assumed to have continuous second derivatives on the domain $\mathcal{D} \subseteq \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_p}$.

The set of equilibrium points of the system that satisfy equality constraints involving an

eigenvalue λ of $\nabla_x f$ with associated right eigenvector v or left eigenvector w are defined as

$$\mathcal{M}_{(\mathrm{v})} = \big\{(x,u,p) \in \mathcal{D} \;\big|\; f(x,u,p) = 0,\ \bar{g}(x,u,p,\lambda,v) = 0,\ \nabla_x f(x,u,p)v = \lambda v \\ \text{for some } \lambda \in \mathbb{C},\ v \in \mathbb{C}^{n_x} \backslash \{0\}\big\} \quad (4.2)$$

$$\mathcal{M}_{(\mathrm{w})} = \big\{(x,u,p) \in \mathcal{D} \;\big|\; f(x,u,p) = 0,\ \bar{g}(x,u,p,\lambda,w) = 0,\ \nabla_x f^T(x,u,p)w = \lambda w \\ \text{for some } \lambda \in \mathbb{C},\ v \in \mathbb{C}^{n_x} \backslash \{0\}\big\} \quad (4.3)$$

Note that the constraint defined by $\bar{g} : \mathcal{D} \times \mathbb{C} \times \mathbb{C}^{n_x} \to \mathbb{R}^{n_{\bar{g}}}$, unlike g used in (3.3), also depends on λ and v or w. It can thus be chosen such that $\mathcal{M}_{(\mathrm{v})}$ or $\mathcal{M}_{(\mathrm{w})}$ contains only those equilibrium points that possess certain eigenmode characteristics of the linearized system.

Instead of $\mathcal{M}_{(\mathrm{v})}$ and $\mathcal{M}_{(\mathrm{w})}$, it will be convenient to work with the sets

$$\widetilde{\mathcal{M}}_{(\mathrm{v})} = \big\{(x,u,p,\sigma,\eta,v_r,v_i) \in \widetilde{\mathcal{D}} \;\big|\; f(x,u,p) = 0,\ \bar{g}(x,u,p,\sigma+i\eta,v_r+iv_i) = 0, \\ \nabla_x f(x,u,p)v_r = \sigma v_r - \eta v_i,\ \nabla_x f(x,u,p)v_i = \sigma v_i + \eta v_r, \\ v_r^T v_r + v_i^T v_i = 1,\ v_r^T v_i = 0\big\} \quad (4.4)$$

$$\widetilde{\mathcal{M}}_{(\mathrm{w})} = \big\{(x,u,p,\sigma,\eta,w_r,w_i) \in \widetilde{\mathcal{D}} \;\big|\; f(x,u,p) = 0,\ \bar{g}(x,u,p,\sigma+i\eta,w_r+iw_i) = 0, \\ \nabla_x f^T(x,u,p)w_r = \sigma w_r - \eta w_i,\ \nabla_x f^T(x,u,p)w_i = \sigma w_i + \eta w_r, \\ w_r^T w_r + w_i^T w_i = 1,\ w_r^T w_i = 0\big\}, \quad (4.5)$$

where $\widetilde{\mathcal{D}} = \mathcal{D} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$.

In the definitions of $\widetilde{\mathcal{M}}_{(\mathrm{v})}$ and $\widetilde{\mathcal{M}}_{(\mathrm{w})}$ above, the real and imaginary parts of the eigenvalue $\lambda = \sigma + i\eta$ are included as components of the elements in the sets, as are the real and imaginary parts of the eigenvectors $v = v_r + iv_i$ and $w = w_r + iw_i$. The complex eigenmode equation is rewritten in (4.4) in terms of two real equations obtained by equating the real and imaginary parts on the left and right hand side,

$$\nabla_x f(x,u,p)(v_r + iv_i) = (\sigma + i\eta)(v_r + iv_i) \qquad \Leftrightarrow \qquad \begin{cases} \nabla_x f(x,u,p)v_r = \sigma v_r - \eta v_i \\ \nabla_x f(x,u,p)v_i = \sigma v_i + \eta v_r \end{cases} \quad (4.6)$$

Additionally, two equations to normalize the complex magnitude of the eigenvector are added in order to resolve the ambiguity of eigenvector scaling. The first equation normalizes the

magnitude, while the second enforces orthogonality between the real and imaginary parts of the eigenvector:

$$v_r^T v_r + v_i^T v_i = 1, \tag{4.7a}$$

$$v_r^T v_i = 0. \tag{4.7b}$$

The additional equations in the definition of $\widetilde{\mathcal{M}}_{(\mathrm{w})}$ are derived analogously.

All conditions on the elements in $\widetilde{\mathcal{M}}_{(\mathrm{v})}$ and $\widetilde{\mathcal{M}}_{(\mathrm{w})}$ are expressed as nonlinear equations involving the various components of the elements, so that the "for some" qualifier in (4.2) and (4.3) is no longer needed. This allows computation with a continuation method as discussed in Section 2.2. The sets $\mathcal{M}_{(\mathrm{v})}$ and $\mathcal{M}_{(\mathrm{w})}$ are obtained from $\widetilde{\mathcal{M}}_{(\mathrm{v})}$ and $\widetilde{\mathcal{M}}_{(\mathrm{w})}$, respectively, by projection into $\mathcal{D}$,

$$\mathcal{M}_{(\mathrm{v})} = \pi_{\mathcal{D}}\big(\widetilde{\mathcal{M}}_{(\mathrm{v})}\big) \qquad \text{and} \qquad \mathcal{M}_{(\mathrm{w})} = \pi_{\mathcal{D}}\big(\widetilde{\mathcal{M}}_{(\mathrm{w})}\big). \tag{4.8}$$

This amounts to simply removing (or ignoring) the eigenvalue and eigenvector components of the solution points, so that the projection $\pi_{\mathcal{D}}$ is defined as

$$\pi_{\mathcal{D}}(x, u, p, \sigma, \eta, v_r, v_i) = (x, u, p). \tag{4.9}$$

Note that the normalization defined in (4.7) is not unique: if (v_r, v_i) satisfies (4.6) and (4.7), so do $(-v_r, -v_i)$, $(v_i, -v_r)$ and $(-v_i, v_r)$. This ambiguity means that each element in $\mathcal{M}_{(\mathrm{v})}$ corresponds to four elements in $\widetilde{\mathcal{M}}_{(\mathrm{v})}$. When computing connected subsets of $\widetilde{\mathcal{M}}_{(\mathrm{v})}$ with continuation algorithms, only one of those elements will be obtained for each point in $\mathcal{M}_{(\mathrm{v})}$ (which one depends on how the algorithm is initialized). If λ is real, (4.7b) ensures that v is real[1] when v_r is non-zero, i.e., $v_i = 0$. Due to the non-uniqueness of the normalization, however, the equation also admits the case that v is purely imaginary with $v_r = 0$.

Algorithms that numerically calculate eigenvalues and eigenvectors often normalize them differently. However, if an eigenvector $\hat{v} \in \mathbb{C}^{n_x}$ has been computed, it can always be re-normalized to satisfy (4.7):

[1]This is not automatically true: if $Av = \lambda v$ for $\lambda \in \mathbb{R}$ and $v \in \mathbb{R}^n$, then $A\bar{v} = \lambda\bar{v}$ with $\bar{v} = e^{i\theta}v \in \mathbb{C}^n$ for any $\theta \in [0, 2\pi]$.

Proposition 4.1. *Suppose $\hat{v} \in \mathbb{C}^{n_x}$ is an eigenvector of the matrix $A \in \mathbb{R}^{n_x \times n_x}$ with eigenvalue λ. Then $v = v_r + iv_i \in \mathbb{C}^{n_x}$ defined by*

$$v = \frac{e^{i\theta}}{\|\hat{v}\|_2}\hat{v} \qquad \textit{with} \qquad \theta = -\frac{1}{2}\operatorname{atan2}\left(\operatorname{Im}\hat{v}^T\hat{v},\ \operatorname{Re}\hat{v}^T\hat{v}\right) \tag{4.10}$$

is an eigenvector of A satisfying (4.7) for the same eigenvalue, where atan2 is the four-quadrant inverse tangent function.

Proof. Since $A\hat{v} = \lambda\hat{v}$ holds, so does $Aa\hat{v} = \lambda a\hat{v}$ for any $a \in \mathbb{C}\backslash\{0\}$. Choose $a = e^{i\theta}/\|\hat{v}\|_2$ to see that v is an eigenvector for the same eigenvalue. Equation (4.7a) is satisfied, since

$$v_r^T v_r + v_i^T v_i = \|v\|_2^2 = \left(\frac{|e^{i\theta}|}{\|\hat{v}\|_2}\|\hat{v}\|_2\right)^2 = 1. \tag{4.11}$$

Define $\hat{v}_r = \operatorname{Re}\hat{v}$ and $\hat{v}_i = \operatorname{Im}\hat{v}$ as well as $r = 1/\|\hat{v}\|_2$ so that

$$v = \underbrace{r(\hat{v}_r\cos\theta - \hat{v}_i\sin\theta)}_{=v_r} + i\,\underbrace{r(\hat{v}_r\sin\theta + \hat{v}_i\cos\theta)}_{=v_i}. \tag{4.12}$$

With some algebra and trigonometric identities, it is found that

$$\begin{aligned} v_r^T v_i &= r^2\left(\hat{v}_r\cos\theta - \hat{v}_i\sin\theta\right)^T\left(\hat{v}_r\sin\theta + \hat{v}_i\cos\theta\right) \\ &= \frac{1}{2}r^2\left[(\hat{v}_r^T\hat{v}_r - \hat{v}_i^T\hat{v}_i)\sin(2\theta) + 2\hat{v}_r^T\hat{v}_i\cos(2\theta)\right] \\ &= \frac{1}{2}r^2 C\sin(2\theta + \phi), \end{aligned} \tag{4.13}$$

where $C > 0$ and $\phi = \operatorname{atan2}\left(2\hat{v}_r^T\hat{v}_i,\ \hat{v}_r^T\hat{v}_r - \hat{v}_i^T\hat{v}_i\right)$. With $\hat{v}^T\hat{v} = \hat{v}_r^T\hat{v}_r - \hat{v}_i^T\hat{v}_i + i2\hat{v}_r^T\hat{v}_i$ and the definition of θ in (4.10), it can be seen that θ equals $-\frac{1}{2}\phi$, so that the sine term vanishes. The vector v therefore satisfies (4.7b). ■

It will be useful to separate the constraints into two groups, namely those that only depend on x, u and p and those that also depend on the eigenvalues and/or eigenvectors. The constraint function $\bar{g}$ in (4.2–4.5) is therefore defined to consist of two components,

$$\bar{g}(x, u, p, \lambda, v) = \begin{bmatrix} g(x, u, p) \\ \tilde{g}(x, u, p, \sigma, \eta, v_r, v_i) \end{bmatrix}, \tag{4.14}$$

where $\sigma = \operatorname{Re}\lambda$, $\eta = \operatorname{Im}\lambda$, $v_r = \operatorname{Re} v$ and $v_i = \operatorname{Im} v$. The function $\tilde{g} : \widetilde{\mathcal{D}} \to \mathbb{R}^{n_{\tilde{g}}}$ is one of the eigenmode constraints defined in Section 4.2 (or any other function of the eigenmodes) and $g : \mathcal{D} \to \mathbb{R}^{n_g}$ allows for definition of further properties on the set of equilibrium points.

The equations that define the sets $\widetilde{\mathcal{M}}_{(\mathrm{v})}$ and $\widetilde{\mathcal{M}}_{(\mathrm{w})}$ are now grouped to form expressions of the type $F(z) = 0$, so that continuation algorithms can be applied to solve for the variables in z when the functions are well-behaved. To meet the differentiability requirement of the pseudo-arclength method, f must have continuous second derivatives, while g and $\tilde{g}$ are required to have continuous first derivatives. Conditions under which the Jacobian matrices of these functions have full rank or drop rank are given later in this section.

4.1.1 Extended system with right eigenvectors

Collecting the conditions on the elements in $\widetilde{\mathcal{M}}_{(\mathrm{v})}$ from (4.4) gives the function

$$F_{(\mathrm{v})}(x,u,p,\sigma,\eta,v_r,v_i) = \begin{bmatrix} f(x,u,p) \\ g(x,u,p) \\ \nabla_x f(x,u,p)v_r - \sigma v_r + \eta v_i \\ \nabla_x f(x,u,p)v_i - \sigma v_i - \eta v_r \\ v_r^T v_r + v_i^T v_i - 1 \\ v_r^T v_i \\ \tilde{g}(x,u,p,\sigma,\eta,v_r,v_i) \end{bmatrix} \tag{4.15}$$

where the Jacobian matrix

$$\nabla F_{(\mathrm{v})} = \begin{bmatrix} \nabla_x f & \nabla_u f & \nabla_p f & 0 & 0 & 0 & 0 \\ \nabla_x g & \nabla_u g & \nabla_p g & 0 & 0 & 0 & 0 \\ [\nabla^2_{xx} f v_r] & [\nabla^2_{xu} f v_r] & [\nabla^2_{xp} f v_r] & -v_r & v_i & \nabla_x f - \sigma I & \eta I \\ [\nabla^2_{xx} f v_i] & [\nabla^2_{xu} f v_i] & [\nabla^2_{xp} f v_i] & -v_i & -v_r & -\eta I & \nabla_x f - \sigma I \\ 0 & 0 & 0 & 0 & 0 & 2v_r^T & 2v_i^T \\ 0 & 0 & 0 & 0 & 0 & v_i^T & v_r^T \\ \nabla_x \tilde{g} & \nabla_u \tilde{g} & \nabla_p \tilde{g} & \nabla_\sigma \tilde{g} & \nabla_\eta \tilde{g} & \nabla_{v_r} \tilde{g} & \nabla_{v_i} \tilde{g} \end{bmatrix} \tag{4.16}$$

has dimension $(3n_x+n_g+2+n_{\tilde{g}})\times(3n_x+n_u+n_p+2)$ and the function arguments have been omitted for brevity. The terms in brackets involve second partial derivatives of f and are defined as

$$[\nabla^2_{xx} f v_*] = \left[\frac{\partial(\nabla_x f)}{\partial x_1} v_* \quad \cdots \quad \frac{\partial(\nabla_x f)}{\partial x_{n_x}} v_*\right], \tag{4.17a}$$

$$[\nabla^2_{xu} f v_*] = \left[\frac{\partial(\nabla_x f)}{\partial u_1} v_* \quad \cdots \quad \frac{\partial(\nabla_x f)}{\partial u_{n_u}} v_*\right], \tag{4.17b}$$

$$[\nabla^2_{xp} f v_*] = \left[\frac{\partial(\nabla_x f)}{\partial p_1} v_* \quad \cdots \quad \frac{\partial(\nabla_x f)}{\partial p_{n_p}} v_*\right]. \tag{4.17c}$$

When only real eigenmodes are of interest, Eq. (4.15) simplifies by setting $\eta = 0$ and $v_i = 0$. The resulting system of equations then becomes

$$F_{(\mathrm{v,r})}(x,u,p,\sigma,v_r) = \begin{bmatrix} f(x,u,p) \\ g(x,u,p) \\ \nabla_x f(x,u,p) v_r - \sigma v_r \\ v_r^T v_r - 1 \\ \tilde{g}(x,u,p,\sigma,0,v_r,0) \end{bmatrix} \tag{4.18}$$

with

$$\nabla F_{(\mathrm{v,r})} = \begin{bmatrix} \nabla_x f & \nabla_u f & \nabla_p f & 0 & 0 \\ \nabla_x g & \nabla_u g & \nabla_p g & 0 & 0 \\ [\nabla^2_{xx} f v_r] & [\nabla^2_{xu} f v_r] & [\nabla^2_{xp} f v_r] & -v_r & \nabla_x f - \sigma I \\ 0 & 0 & 0 & 0 & 2v_r^T \\ \nabla_x \tilde{g} & \nabla_u \tilde{g} & \nabla_p \tilde{g} & \nabla_\sigma \tilde{g} & \nabla_{v_r} \tilde{g} \end{bmatrix}. \tag{4.19}$$

4.1.2 Extended system with left eigenvectors

Collecting the conditions on the elements in $\widetilde{\mathcal{M}}_w$ from Eq. (4.5) leads to

$$F_{(\mathrm{w})}(x,u,p,\sigma,\eta,w_r,w_i) = \begin{bmatrix} f(x,u,p) \\ g(x,u,p) \\ \nabla_x f^T(x,u,p)w_r - \sigma w_r + \eta w_i \\ \nabla_x f^T(x,u,p)w_i - \sigma w_i - \eta w_r \\ w_r^T w_r + w_i^T w_i - 1 \\ w_r^T w_i \\ \tilde{g}(x,u,p,\sigma,\eta,w_r,w_i) \end{bmatrix} \tag{4.20}$$

with Jacobian matrix

$$\nabla F_{(\mathrm{w})} = \begin{bmatrix} \nabla_x f & \nabla_u f & \nabla_p f & 0 & 0 & 0 & 0 \\ \nabla_x g & \nabla_u g & \nabla_p g & 0 & 0 & 0 & 0 \\ [\nabla^2_{xx} f^T w_r] & [\nabla^2_{xu} f^T w_r] & [\nabla^2_{xp} f^T w_r] & -w_r & w_i & \nabla_x f^T - \sigma I & \eta I \\ [\nabla^2_{xx} f^T w_i] & [\nabla^2_{xu} f^T w_i] & [\nabla^2_{xp} f^T w_i] & -w_i & -w_r & -\eta I & \nabla_x f^T - \sigma I \\ 0 & 0 & 0 & 0 & 0 & 2w_r^T & 2w_i^T \\ 0 & 0 & 0 & 0 & 0 & w_i^T & w_r^T \\ \nabla_x \tilde{g} & \nabla_u \tilde{g} & \nabla_p \tilde{g} & \nabla_\sigma \tilde{g} & \nabla_\eta \tilde{g} & \nabla_{w_r} \tilde{g} & \nabla_{w_i} \tilde{g} \end{bmatrix}. \tag{4.21}$$

The entries that contain second partial derivatives of f are understood as follows:

$$[\nabla^2_{xx} f^T w_*] = \begin{bmatrix} \frac{\partial(\nabla_x f^T)}{\partial x_1} w_* & \cdots & \frac{\partial(\nabla_x f^T)}{\partial x_{n_x}} w_* \end{bmatrix} \tag{4.22a}$$

$$[\nabla^2_{xu} f^T w_*] = \begin{bmatrix} \frac{\partial(\nabla_x f^T)}{\partial u_1} w_* & \cdots & \frac{\partial(\nabla_x f^T)}{\partial u_{n_u}} w_* \end{bmatrix} \tag{4.22b}$$

$$[\nabla^2_{xp} f^T w_*] = \begin{bmatrix} \frac{\partial(\nabla_x f^T)}{\partial p_1} w_* & \cdots & \frac{\partial(\nabla_x f^T)}{\partial p_{n_p}} w_* \end{bmatrix} \tag{4.22c}$$

A reduced system of equations for continuation of equilibrium points with constraints on

a real eigenmode is again obtained by letting $\eta = 0$ and $w_i = 0$:

$$F_{(\mathrm{w,r})}(x,u,p,\sigma,w_r) = \begin{bmatrix} f(x,u,p) \\ g(x,u,p) \\ \nabla_x f^T(x,u,p) w_r - \sigma w_r \\ w_r^T w_r - 1 \\ \tilde{g}(x,u,p,\sigma,0,w_r,0) \end{bmatrix} \tag{4.23}$$

The corresponding Jacobian matrix is

$$\nabla F_{(\mathrm{w,r})} = \begin{bmatrix} \nabla_x f & \nabla_u f & \nabla_p f & 0 & 0 \\ \nabla_x g & \nabla_u g & \nabla_p g & 0 & 0 \\ [\nabla^2_{xx} f^T w_r] & [\nabla^2_{xu} f^T w_r] & [\nabla^2_{xp} f^T w_r] & -w_r & \nabla_x f^T - \sigma I \\ 0 & 0 & 0 & 0 & 2w_r^T \\ \nabla_x \tilde{g} & \nabla_u \tilde{g} & \nabla_p \tilde{g} & \nabla_\sigma \tilde{g} & \nabla_{w_r} \tilde{g} \end{bmatrix}. \tag{4.24}$$

The functions $F_{(\mathrm{v})}$, $F_{(\mathrm{w})}$, $F_{(\mathrm{v,r})}$, and $F_{(\mathrm{w,r})}$ will be called *extended system functions* in the following sections.

4.1.3 Regular and critical points of the extended system functions

Before defining some useful eigenmode constraints in the next section, the conditions under which the Jacobian matrices of the functions defined above have full rank or drop rank are investigated. It is always assumed that $n_u + n_p > n_g + n_{\tilde{g}}$, so that the various Jacobian matrices are "fat" and have full rank if and only if the sets formed by their rows are linearly independent.

A necessary condition that any regular point of the extended system functions must satisfy is regularity of its (x,u,p)-component with respect to the functions f and g that specify the set of equilibrium points from which a subset that satisfies the eigenmode constraint is selected.

Proposition 4.2. *Suppose* $(x^*, u^*, p^*) \in \mathcal{D}$ *is a critical point of the function*

$$F_{(g)}(x, u, p) = \begin{bmatrix} f(x, u, p) \\ g(x, u, p) \end{bmatrix}. \tag{4.25}$$

Then $(x^*, u^*, p^*, \sigma, \eta, v_r, v_i) \in \widetilde{\mathcal{D}}$ *is a critical point of* $F_{(v)}$ *for any* $\sigma, \eta \in \mathbb{R}$ *and* $v_r, v_i \in \mathbb{R}^{n_x}$.

Proof. By definition of a critical point, the rows of $\nabla F_{(\mathrm{g})}(x^*, u^*, p^*)$ form a linearly dependent set, and hence the same is true for the first $n_x + n_g$ rows of $\nabla F_{(\mathrm{v})}(x^*, u^*, p^*, \sigma, \eta, v_r, v_i)$. ■

Corollary 4.3. *Similar statements hold for* $F_{(v,r)}$, $F_{(w)}$, *and* $F_{(w,r)}$.

The conditions leading to rank deficiency in the Jacobian matrices of the extended system functions at regular points of $F_{(\mathrm{g})}$ are considered next.

Proposition 4.4. *Suppose that the rank condition*

$$\operatorname{rank} \begin{bmatrix} [\nabla^2_{xx} f v_r] & [\nabla^2_{xu} f v_r] & [\nabla^2_{xp} f v_r] \\ [\nabla^2_{xx} f v_i] & [\nabla^2_{xu} f v_i] & [\nabla^2_{xp} f v_i] \\ \nabla_x \tilde{g} & \nabla_u \tilde{g} & \nabla_p \tilde{g} \end{bmatrix} \geq n_{\tilde{g}} \tag{4.26}$$

does not hold at $z = (x, u, p, \sigma, \eta, v_r, v_i)$. *Then* z *is a critical point of* $F_{(v)}$.

Proof. The submatrix

$$\begin{bmatrix} -v_r & v_i & A - \sigma I & \eta I \\ -v_i & -v_r & -\eta I & A - \sigma I \\ 0 & 0 & 2v_r^T & 2v_i^T \\ 0 & 0 & v_i^T & v_r^T \\ \nabla_\sigma \tilde{g} & \nabla_\eta \tilde{g} & \nabla_{v_r} \tilde{g} & \nabla_{v_i} \tilde{g} \end{bmatrix} \tag{4.27}$$

of $\nabla F_{(\mathrm{v})}$ in (4.16) has dimension $(2n_x + 2 + n_{\tilde{g}}) \times (2n_x + 2)$ and hence at most rank $2n_x + 2$. If (4.26) does not hold, then the rows of the matrix in (4.26) do not sufficiently augment the rank of the (fat) submatrix formed by the bottom $2n_x + 2 + n_{\tilde{g}}$ rows of $\nabla F_{(\mathrm{v})}$ to eliminate its left null space. ■

Corollary 4.5. *Similar statements hold for $F_{(v,r)}$, $F_{(w)}$, and $F_{(w,r)}$.*

The rank condition (4.26) ensures that $\nabla_x f$ depends sufficiently on x, u and p in order to admit satisfaction of the eigenmode constraints. Consider the case with a scalar eigenmode constraint ($n_{\tilde{g}} = 1$) when $\tilde{g}$ does not depend on x, u, p. The rank condition will be violated when the second derivatives of f are all zero, so that $\nabla_x f$ is constant. This implies that $\ker F_{(\mathrm{v})}$ is either empty (no point satisfies the constraint), or the eigenmode constraint is trivially satisfied everywhere and therefore does not restrict the solution set.

Critical points of the extended system functions also occur at points that correspond to repeated eigenvalues of the Jacobian matrix $\nabla_x f(x, u, p)$ under some conditions, which are examined below. Some intermediate results are derived first in the following Lemma and Corollary.

Lemma 4.6. *Let $A \in \mathbb{R}^{n_x \times n_x}$, $\sigma, \eta \in \mathbb{R}$, and $v_r, v_i \in \mathbb{R}^{n_x}$ be such that $\sigma + i\eta$ is an eigenvalue of A with (right) eigenvector $v_r + iv_i$ and assume v_r and v_i satisfy the normalization defined in* (4.7)*. Then*

$$V = \begin{bmatrix} -v_r & v_i & A - \sigma I & \eta I \\ -v_i & -v_r & -\eta I & A - \sigma I \\ 0 & 0 & 2v_r^T & 2v_i^T \\ 0 & 0 & v_i^T & v_r^T \end{bmatrix} \tag{4.28}$$

is singular if and only if $\sigma + i\eta$ is a repeated eigenvalue of A.

Proof. If V is singular, then there exist $r_r, r_i \in \mathbb{R}$ and $q_r, q_i \in \mathbb{R}^{n_x}$ not all zero such that

$$\begin{bmatrix} -v_r & v_i & A - \sigma I & \eta I \\ -v_i & -v_r & -\eta I & A - \sigma I \\ 0 & 0 & 2v_r^T & 2v_i^T \\ 0 & 0 & v_i^T & v_r^T \end{bmatrix} \begin{bmatrix} r_r \\ r_i \\ q_r \\ q_i \end{bmatrix} = 0. \tag{4.29}$$

Note that the normalization as defined in 4.7 implies that

$$v_r^T v_r - v_i^T v_i = \mathrm{Re}\big[(v_r^T + iv_i^T)(v_r + iv_i)\big] \neq 0, \tag{4.30}$$

because (4.10) has to hold with $\theta = 0$ when evaluated on $\hat{v} = v_r + iv_i$ and the four-quadrant inverse tangent function is never zero when the second argument equals zero.

It is first shown by contradiction that $v_r + iv_i$ and $q_r + iq_i$ are linearly independent. Suppose that they are linearly dependent, so that there exist $a_r, a_i \in \mathbb{R}$ not both zero such that

$$q_r + iq_i = (a_r + ia_i)(v_r + iv_i) \qquad \Longleftrightarrow \qquad q_r = a_r v_r - a_i v_i \quad \text{and} \quad q_i = a_r v_r + a_i v_r \tag{4.31}$$

Inserting this into the third and fourth rows of (4.29) gives

$$v_r^T q_r + v_i^T q_i = a_r v_r^T v_r - a_i v_r^T v_i + a_r v_i^T v_i + a_i v_i^T v_r = a_r \underbrace{(v_r^T v_r + v_i^T v_i)}_{=1} = 0 \tag{4.32}$$

$$v_i^T q_r + v_r^T q_i = a_r \underbrace{v_i^T v_r}_{=0} - a_i v_i^T v_i + a_r \underbrace{v_r^T v_i}_{=0} + a_i v_r^T v_r = a_i \underbrace{(v_r^T v_r - v_i^T v_i)}_{\neq 0} = 0 \tag{4.33}$$

This is a contradiction, since it was assumed that a_r and a_i are not both zero. The two vectors are therefore linearly independent.

The first and second row of (4.29) lead to

$$Aq_r - \sigma q_r + \eta q_i - r_r v_r + r_i v_i = 0 \quad \text{and} \quad Aq_i - \sigma q_i - \eta q_r - r_r v_i - r_i v_r = 0 \tag{4.34}$$

$$\Longleftrightarrow \qquad Aq_r - \sigma q_r + \eta q_i - r_r v_r + r_i v_i + i(Aq_i - \sigma q_i - \eta q_r - r_r v_i - r_i v_r) = 0 \tag{4.35}$$

$$\Longleftrightarrow \qquad (A - (\sigma + i\eta)I)\,(q_r + iq_i) = (r_r + ir_i)(v_r + iv_i) \tag{4.36}$$

If r_r and r_i are zero, then q_r and q_i cannot both be zero. It follows that $q_r + iq_i$ is an eigenvector of A corresponding to the eigenvalue $\sigma + i\eta$. If r_r and r_i are not both zero, then q_r and q_i cannot both be zero because $v_r + iv_i$ is nonzero. It follows from

$$(A - (\sigma + i\eta)I)^2\,(q_r + iq_i) = (r_r + ir_i)\,(A - (\sigma + i\eta)I)\,(v_r + iv_i) = 0 \tag{4.37}$$

that $q_r + iq_i$ is a generalized eigenvector of A corresponding to the eigenvalue $\sigma + i\eta$. In either case, the algebraic multiplicity of $\sigma + i\eta$ is greater than one, i.e., it is a repeated eigenvalue. Since the argument is based on equivalence, this proves both necessity and sufficiency. ■

Corollary 4.7. *Suppose that $A \in \mathbb{R}^{n_x \times n_x}$ has a real eigenvalue $\sigma \in \mathbb{R}$ with eigenvector $v_r \in \mathbb{R}^{n_x}$ which satisfies $v_r^T v_r = 1$. Then*

$$V_r = \begin{bmatrix} -v_r & A - \sigma I \\ 0 & 2v_r^T \end{bmatrix} \tag{4.38}$$

is singular if and only if σ is a repeated eigenvalue of A.

Proof. Since A, σ, and v_r satisfy the conditions of Lemma 4.6 with $\eta = 0$ and $v_i = 0$, there exist $r_r, r_i \in \mathbb{R}$ and $q_r, q_i \in \mathbb{R}^{n_x}$ not all zero such that

$$\begin{bmatrix} -v_r & 0 & A - \sigma I & 0 \\ 0 & -v_r & 0 & A - \sigma I \\ 0 & 0 & 2v_r^T & 0 \\ 0 & 0 & 0 & v_r^T \end{bmatrix} \begin{bmatrix} r_r \\ r_i \\ q_r \\ q_i \end{bmatrix} = 0. \tag{4.39}$$

if and only if σ is a repeated eigenvalue. So if σ is a repeated eigenvalue, then $[r_r \quad q_r^T]^T$ is in the null space of V_r. Conversely, if there exists a nonzero vector $[r \quad q^T]^T$ in the null space of V_r, where $r \in \mathbb{R}$ and $q \in \mathbb{R}^{n_x}$, then (4.39) holds with $r_r = r$, $r_i = 0$, $q_r = q$, and $q_i = 0$. ■

Corollary 4.8. *Suppose that σ is a repeated real eigenvalue of $\nabla_x f(x, u, p)$ with eigenvector v_r satisfying $v_r^T v_r = 1$. Then $(x, u, p, \sigma, 0, v_r, 0)$ is a critical point of $F_{(v)}$.*

Proof. By Corollary 4.7, the matrix V_r in (4.38) is singular and hence has a nontrivial left null space, so that there exist $\hat{q} \in \mathbb{R}^{n_x}$ and $\hat{r} \in \mathbb{R}$ satisfying the equalities $\hat{q}^T v_r = 0$ and $\hat{q}^T(\nabla_x f(x, u, p) - \sigma I) + 2\hat{r} v_r^T = 0$. It follows that

$$\begin{bmatrix} 0 & 0 & 0 & \hat{q}^T & 0 & \hat{r}/2 & 0 \end{bmatrix} \nabla F_{(\mathrm{v})}(x, u, p, \sigma, 0, v_r, 0) = 0 \tag{4.40}$$

where the dimensions of the zero entries are implied from the block structure of $\nabla F_{(\mathrm{v})}$ in (4.16). This means that the rows of $F_{(\mathrm{v})}$ form a linearly dependent set. ■

Corollary 4.9. *A similar statement holds for $F_{(w)}$.*

4.2 Some Eigenmode Constraints

This section derives some constraints that may be used with the extended system functions defined in the preceding section.

4.2.1 Stability boundary

Choosing the eigenvalue constraint function

$$\tilde{g}_{(\mathrm{s})}(x, u, p, \sigma, \eta, v_r, v_i) = \sigma \tag{4.41}$$

forces the eigenvalue to lie on the imaginary axis. All equilibrium points that satisfy this constraint are therefore non-hyperbolic and constitute bifurcation points of the nonlinear system. In fact, using $\tilde{g}_{(\mathrm{s})} = 0$ to eliminate σ from the unknowns gives an extended system similar to what is found in the literature [4] for the continuation of Hopf bifurcations (characterized by a complex pair of eigenvalues on the imaginary axis) in two parameters when no equality constraints $g(x, u, p) = 0$ are enforced:

$$F_{(\mathrm{s})}(x, u, p, \eta, v_r, v_i) = \begin{bmatrix} f(x, u, p) \\ \nabla_x f(x, u, p) v_r + \eta v_i \\ \nabla_x f(x, u, p) v_i - \eta v_r \\ v_r^T v_r + v_i^T v_i - 1 \\ v_r^T v_i \end{bmatrix} \tag{4.42}$$

Note that solutions of (4.42) comprise not only Hopf bifurcations but also bifurcations that occur when a real eigenvalue equals zero. The problem can be further reduced by setting $\eta = 0$ and $v_i = 0$ when only real eigenmodes are of interest, leading to the standard augmented system for continuation of saddle-node bifurcations in two parameters,

$$F_{(\mathrm{s,r})}(x, u, p, v_r) = \begin{bmatrix} f(x, u, p) \\ \nabla_x f(x, u, p) v_r \\ v_r^T v_r - 1 \end{bmatrix} \tag{4.43}$$

4.2.2 Performance measures

Typical measures that characterize the system performance around equilibrium points include properties of the eigenmodes of the linearized dynamics such as natural frequency, damping ratio, and time-to-double. The definitions of these properties directly lead to eigenvalue constraints that enable the specification of equilibrium points having these characteristics. Since the constraints do not involve the eigenvector, they can be used with either $F_{(\mathrm{v})}$ or $F_{(\mathrm{w})}$.

Natural frequency constraint

To find equilibrium points that have a mode with natural frequency ω_n^*, the constraint function

$$\tilde{g}_{(\omega_n)}(\sigma,\eta) = \sqrt{\sigma^2+\eta^2} - \omega_n^* \tag{4.44}$$

with partial derivatives

$$\nabla_\sigma \tilde{g}_{(\omega_n)}(\sigma,\eta) = \frac{\sigma}{\sqrt{\sigma^2+\eta^2}} \qquad \text{and} \qquad \nabla_\eta \tilde{g}_{(\omega_n)}(\sigma,\eta) = \frac{\eta}{\sqrt{\sigma^2+\eta^2}} \tag{4.45}$$

can be used.

Damping coefficient constraint

Equilibrium branches that have a mode with damping coefficient ζ^* can be found with the following eigenvalue constraint function:

$$\tilde{g}_{(\zeta)}(\sigma,\eta) = -\frac{\sigma}{\sqrt{\sigma^2+\eta^2}} - \zeta^* \tag{4.46}$$

The corresponding partial derivatives are

$$\nabla_\sigma \tilde{g}_{(\zeta)}(\sigma,\eta) = -\frac{\eta^2}{(\sigma^2+\eta^2)^{\frac{3}{2}}} \qquad \text{and} \qquad \nabla_\eta \tilde{g}_{(\zeta)}(\sigma,\eta) = \frac{\sigma\eta}{(\sigma^2+\eta^2)^{\frac{3}{2}}}. \tag{4.47}$$

Time-to-halve / time-to-double

For equilibrium branches containing a mode with time to double or halve $T_2^* > 0$, the constraint function is

$$\tilde{g}_{(T_2)}(\sigma) = \frac{\ln 2}{|\sigma|} - T_2^* \tag{4.48}$$

with

$$\nabla_\sigma \tilde{g}_{(T_2)}(\sigma) = -\frac{\ln 2}{|\sigma|\sigma}. \tag{4.49}$$

4.2.3 Local linear controllability

Establishing controllability of the nonlinear system (4.1) is in general a difficult problem, and often requires tools of differential geometry [37]. However, it has been shown [38, 39] that linear controllability of the linearized system at an equilibrium point (x^*, u^*, p^*) guarantees local controllability of the nonlinear system at (x^*, u^*, p^*). This means that for fixed p^* there exists an open neighborhood $\mathcal{X}$ around x^* and control $u(t)$ such that for any initial state $x(t_0) \in \mathcal{X}$, the state can be driven to the equilibrium point in finite time T without leaving the neighborhood, i.e., $x(t_0+T) = x^*$ and $x(t) \in \mathcal{X}$ for all $t \in [t_0, t_0+T]$. The converse is not true: A system may still be locally controllable at the equilibrium point even if the linearized system is uncontrollable. However, a nonlinear controller is then required to stabilize the system [40]. Three eigenmode constraints for the analysis of local controllability with linear methods are derived below.

Uncontrollable complex eigenmode constraint

A constraint function $\tilde{g}$ to be used with $F_{(\mathrm{w})}$ defined in (4.20) that allows identification of those equilibrium points that are linearly uncontrollable can be derived from the Popov-Belevitch-Hautus (PBH) controllability test [41], which states that the linear system $\dot{x} = Ax + Bu$ is uncontrollable if and only if

$$\operatorname{rank}\begin{bmatrix} A - \lambda I & B \end{bmatrix} \neq n_x \qquad \Longleftrightarrow \qquad w^H B = (w_r^T - i w_i^T) B = 0 \tag{4.50}$$

for some eigenvalue λ of A with left eigenvector $w = w_r + iw_i$, where w^H is the complex conjugate transpose of w. A constraint function on the linearized dynamics of the nonlinear system (4.1) to be used with (4.20) is obtained by requiring that both the real and imaginary parts of $w^H \nabla_u f(x, u, p)$ vanish,

$$\tilde{g}_{(\mathrm{uc})}(x, u, p, w_r, w_i) = \begin{bmatrix} \nabla_u f^T(x, u, p) w_r \\ \nabla_u f^T(x, u, p) w_i \end{bmatrix} \tag{4.51}$$

Assuming that $\nabla_u f$ has full rank, the constraint consists of $n_{\tilde{g}} = 2n_u$ real scalar equalities, since $\nabla_u f$ has dimension $n_x \times n_u$. This is consistent with the fact that the (left) invariant subspace associated with a complex eigenmode has dimension two, and that the condition requires this subspace to be orthogonal to the n_u columns of $\nabla_u f$. The partial derivatives of $\tilde{g}_{(\mathrm{uc})}$ for the Jacobian matrix are

$$\nabla_x \tilde{g}_{(\mathrm{uc})} = \begin{bmatrix} [\nabla^2_{ux} f^T w_r] \\ [\nabla^2_{ux} f^T w_i] \end{bmatrix}, \quad \nabla_u \tilde{g}_{(\mathrm{uc})} = \begin{bmatrix} [\nabla^2_{uu} f^T w_r] \\ [\nabla^2_{uu} f^T w_i] \end{bmatrix}, \quad \nabla_p \tilde{g}_{(\mathrm{uc})} = \begin{bmatrix} [\nabla^2_{up} f^T w_r] \\ [\nabla^2_{up} f^T w_i] \end{bmatrix}, \tag{4.52a}$$

$$\nabla_{w_r} \tilde{g}_{(\mathrm{uc})} = \begin{bmatrix} \nabla_u f^T \\ 0 \end{bmatrix}, \qquad \nabla_{w_i} \tilde{g}_{(\mathrm{uc})} = \begin{bmatrix} 0 \\ \nabla_u f^T \end{bmatrix}, \tag{4.52b}$$

where

$$[\nabla^2_{ux} f^T w_*] = \begin{bmatrix} \frac{\partial(\nabla_u f^T)}{\partial x_1} w_* & \cdots & \frac{\partial(\nabla_u f^T)}{\partial x_{n_x}} w_* \end{bmatrix} \tag{4.53a}$$

$$[\nabla^2_{uu} f^T w_*] = \begin{bmatrix} \frac{\partial(\nabla_u f^T)}{\partial u_1} w_* & \cdots & \frac{\partial(\nabla_u f^T)}{\partial u_{n_u}} w_* \end{bmatrix} \tag{4.53b}$$

$$[\nabla^2_{up} f^T w_*] = \begin{bmatrix} \frac{\partial(\nabla_u f^T)}{\partial p_1} w_* & \cdots & \frac{\partial(\nabla_u f^T)}{\partial p_{n_p}} w_* \end{bmatrix} \tag{4.53c}$$

Uncontrollable real eigenmode constraint

One may expect that $F_{(\mathrm{w})} = 0$ with $\tilde{g}_{(\mathrm{uc})}$ as defined in (4.51) for the complex eigenmode case can also be used to find equilibria with real eigenmodes that are linearly uncontrollable. Such equilibrium points do indeed satisfy this equation. However, the equation becomes unsuitable for pseudo-arclength continuation because the Jacobian matrix $\nabla F_{(\mathrm{w})}$ loses rank

when evaluated at $\eta = 0$ and $w_i = 0$ with this constraint function, meaning that the kernel consists only of critical points. The loss of rank results from the fact that the subspace associated with a real eigenmode is only one-dimensional, so that the condition of orthogonality between this subspace and the columns of $\nabla_u f$ only comprises n_u scalar conditions instead of the $2n_u$ conditions in the complex eigenmode case. Accordingly, the term $\nabla_u f^T w_i$ in $\tilde{g}_{(\mathrm{uc})}$ is always zero when evaluated with $w_i = 0$.

For pseudo-arclength continuation of equilibrium points with an uncontrollable real eigenmode, $F_{(\mathrm{w,r})}$ defined in (4.23) is therefore used with the constraint function $\tilde{g}_{(\mathrm{uc,r})}$, defined as

$$\tilde{g}_{(\mathrm{uc,r})}(x, u, p, w_r) = \nabla_u f^T(x, u, p) w_r, \tag{4.54}$$

with partial derivatives

$$\nabla_x \tilde{g}_{(\mathrm{uc,r})} = [\nabla^2_{ux} f^T w_r], \quad \nabla_u \tilde{g}_{(\mathrm{uc,r})} = [\nabla^2_{uu} f^T w_r], \quad \nabla_p \tilde{g}_{(\mathrm{uc,r})} = [\nabla^2_{up} f^T w_r], \quad \nabla_{w_r} \tilde{g}_{(\mathrm{uc,r})} = \nabla_u f^T \tag{4.55a}$$

Measure of modal controllability constraint

A linear system may pass the PBH controllability test and thus be technically controllable, but be very close to being uncontrollable. Continuous measures of controllability have therefore been developed in the literature [42–44]. The following measure is taken from [43]. It defines the measure of controllability of the k-th eigenmode from all inputs of the linear system $\dot{x} = Ax + Bu$ as

$$m_{c,k} = \left\| w^H_{(k)} B \right\|_2, \tag{4.56}$$

where $w_{(k)}$ is a normalized left eigenvector associated with the k-th eigenmode. When applying this controllability measure, the inputs need to be scaled appropriately because the magnitude of B directly affects the measure. The scaling should be such that unit magnitude of each component of the input u corresponds to a similar amount of control effort or actuator power. Arguably, (4.56) provides a suitable measure of the difficulty to regulate the nonlinear system with linear methods when operating close to an equilibrium point. An

eigenmode constraint function for use with $F_{(\mathrm{w})}$ that identifies equilibrium solutions with a particular modal controllability measure value m_c of the linearized system is readily obtained from (4.56),

$$\tilde{g}_{(\mathrm{cm})}(x,u,p,w_r,w_i) = \left\|(w_r^T - iw_i^T)\nabla_u f(x,u,p)\right\| - m_c. \tag{4.57}$$

where

$$\left\|(w_r^T - iw_i^T)\nabla_u f(x,u,p)\right\| = \sqrt{w_r^T\nabla_u f(\cdot)\nabla_u f^T(\cdot)w_r + w_i^T\nabla_u f(\cdot)\nabla_u f^T(\cdot)w_i} \tag{4.58}$$

The derivatives of the eigenmode constraint are

$$\nabla_x \tilde{g}_{(\mathrm{cm})} = \frac{w_r^T\nabla_u f[\nabla^2_{ux} f^T w_r] + w_i^T\nabla_u f[\nabla^2_{ux} f^T w_i]}{\|(w_r^T - iw_i^T)\nabla_u f\|} \tag{4.59a}$$

$$\nabla_u \tilde{g}_{(\mathrm{cm})} = \frac{w_r^T\nabla_u f[\nabla^2_{uu} f^T w_r] + w_i^T\nabla_u f[\nabla^2_{uu} f^T w_i]}{\|(w_r^T - iw_i^T)\nabla_u f\|} \tag{4.59b}$$

$$\nabla_p \tilde{g}_{(\mathrm{cm})} = \frac{w_r^T\nabla_u f[\nabla^2_{up} f^T w_r] + w_i^T\nabla_u f[\nabla^2_{up} f^T w_i]}{\|(w_r^T - iw_i^T)\nabla_u f\|} \tag{4.59c}$$

$$\nabla_{w_r} \tilde{g}_{(\mathrm{cm})} = \frac{w_r^T\nabla_u f\nabla_u f^T}{\|(w_r^T - iw_i^T)\nabla_u f\|} \tag{4.59d}$$

$$\nabla_{w_i} \tilde{g}_{(\mathrm{cm})} = \frac{w_i^T\nabla_u f\nabla_u f^T}{\|(w_r^T - iw_i^T)\nabla_u f\|} \tag{4.59e}$$

and the terms containing second derivatives of f are defined in (4.53). Equation (4.57) is scalar, so that $n_{\tilde{g}} = 1$.

It is also possible to use $\left\|w^H\nabla_u f\right\|^2 - m_c^2 = 0$ instead of $\tilde{g}_{(\mathrm{cm})} = 0$ as eigenmode constraint. While this simplifies the partial derivatives (thereby eliminating the undifferentiable point where the denominator is zero), it is observed that the Newton corrector step of the pseudo-arclength algorithm required significantly more iterations to converge. One may also think that this modification would allow continuation of uncontrollable equilibrium points by choosing $m_c = 0$. However, the partial derivatives of this modified constraint are all zero when evaluated at a point that satisfies $\left\|w^H\nabla_u f\right\|^2 = 0$. This means that $\nabla F_{(\mathrm{w})}$ does not have full rank anywhere on $\ker F_{(\mathrm{w})}$ and the pseudo-arclength method cannot be used.

4.2.4 Local linear observability

Similar to the controllability case, observability of the linearized system is sufficient but not necessary for local observability of the nonlinear system [45]. Constraints for the analysis of the system's local linear observability are provided below.

Unobservable complex eigenmode constraint

The PBH observability test [41] states that the linear system defined by $\dot{x} = Ax + Bu$ and $y = Cx + Du$ has an unobservable eigenmode if and only if

$$\operatorname{rank} \begin{bmatrix} A - \lambda I \\ C \end{bmatrix} \neq n_x \qquad \Longleftrightarrow \qquad Cv = 0 \tag{4.60}$$

for some eigenvalue λ of A with right eigenvector v. This condition leads to $n_{\tilde{g}} = 2n_y$ real scalar equalities $\tilde{g}_{(\mathrm{uo})} = 0$ that allow identification of equilibria where the linearized system has an unobservable complex eigenmode, where

$$\tilde{g}_{(\mathrm{uo})}(x, u, p, v_r, v_i) = \begin{bmatrix} \nabla_x h(x, u, p) v_r \\ \nabla_x h(x, u, p) v_i \end{bmatrix}. \tag{4.61}$$

In the above equation, the Jacobian matrix $\nabla_x h$ of the output function h from (4.1) with regards to the state vector is assumed to have full rank. Partial derivatives of the eigenmode constraint (4.61) are

$$\nabla_x \tilde{g}_{(\mathrm{uo})} = \begin{bmatrix} [\nabla^2_{xx} h v_r] \\ [\nabla^2_{xx} h v_i] \end{bmatrix}, \quad \nabla_u \tilde{g}_{(\mathrm{uo})} = \begin{bmatrix} [\nabla^2_{xu} h v_r] \\ [\nabla^2_{xu} h v_i] \end{bmatrix}, \quad \nabla_p \tilde{g}_{(\mathrm{uo})} = \begin{bmatrix} [\nabla^2_{xp} h v_r] \\ [\nabla^2_{xp} h v_i] \end{bmatrix}, \tag{4.62a}$$

$$\nabla_{v_r} \tilde{g}_{(\mathrm{uo})} = \begin{bmatrix} \nabla_x h \\ 0 \end{bmatrix}, \qquad \nabla_{v_i} \tilde{g}_{(\mathrm{uo})} = \begin{bmatrix} 0 \\ \nabla_x h \end{bmatrix}, \tag{4.62b}$$

where the terms with second derivatives of h are

$$[\nabla^2_{xx} h v_*] = \begin{bmatrix} \frac{\partial(\nabla_x h)}{\partial x_1} v_* & \cdots & \frac{\partial(\nabla_x h)}{\partial x_{n_x}} v_* \end{bmatrix} \tag{4.63a}$$

$$[\nabla^2_{xu} h v_*] = \begin{bmatrix} \frac{\partial(\nabla_x h)}{\partial u_1} v_* & \cdots & \frac{\partial(\nabla_x h)}{\partial u_{n_u}} v_* \end{bmatrix} \tag{4.63b}$$

$$[\nabla^2_{xp} h v_*] = \begin{bmatrix} \frac{\partial(\nabla_x h)}{\partial p_1} v_* & \cdots & \frac{\partial(\nabla_x h)}{\partial p_{n_p}} v_* \end{bmatrix} \tag{4.63c}$$

Unobservable real eigenmode constraint

As in the controllability case, components of $F_{(\mathrm{v})}$ from (4.15) with constraint $\tilde{g}_{(\mathrm{uo})}$ become functionally dependent on $\eta = 0$ and $v_i = 0$ since the last n_y components of the equation, $\nabla_x h v_i = 0$, are always satisfied for real eigenmodes. Instead, the constraint

$$\tilde{g}_{(\mathrm{uo,r})}(x, u, p, v_r) = \nabla_x h(x, u, p) v_r = 0 \tag{4.64}$$

with partial derivatives

$$\nabla_x \tilde{g}_{(\mathrm{uo,r})} = [\nabla^2_{xx} h v_r], \quad \nabla_u \tilde{g}_{(\mathrm{uo,r})} = [\nabla^2_{xu} h v_r], \quad \nabla_p \tilde{g}_{(\mathrm{uo,r})} = [\nabla^2_{xp} h v_r], \quad \nabla_{v_r} \tilde{g}_{(\mathrm{uo,r})} = \nabla_x h. \tag{4.65}$$

can be used with $F_{(\mathrm{v,r})}$ when analyzing an unobservable real eigenmode with the pseudo-arclength method. The number of scalar conditions is $n_{\tilde{g}} = n_y$ in this case.

Measure of modal observability constraint

Analogous to the controllability measure, the following continuous measure of modal observability taken from [43] is useful to determine the degree of observability of the k-th eigenmode of the linear system $\dot{x} = Ax + Bu$ and $y = Cx + Du$ from all outputs:

$$m_{o,k} = \left\| C v_{(k)} \right\|_2 \tag{4.66}$$

In the above, $v_{(k)}$ is a normalized right eigenvector associated with the k-th eigenmode. Equilibrium points with a particular modal observability measure value m_o of the linearized system dynamics are identified by rewriting (4.66) as an eigenmode constraint to be used with (4.15),

$$\tilde{g}_{(\mathrm{om})}(x, u, p, v_r, v_i) = \|\nabla_x h(x, u, p)(v_r + i v_i)\| - m_o = 0 \tag{4.67}$$

where

$$\|\nabla_x h(x, u, p)(v_r + i v_i)\| = \sqrt{v_r^T \nabla_x h^T(\cdot) \nabla_x h(\cdot) v_r + v_i^T \nabla_x h^T(\cdot) \nabla_x h(\cdot) v_i} \tag{4.68}$$

Partial derivatives of the eigenmode constraint are

$$\nabla_x \tilde{g}_{(\mathrm{om})} = \frac{v_r^T \nabla_x h^T [\nabla_{xx}^2 h v_r] + v_i^T \nabla_x h^T [\nabla_{xx}^2 h v_i]}{\|\nabla_x h (v_r + i v_i)\|} \tag{4.69a}$$

$$\nabla_u \tilde{g}_{(\mathrm{om})} = \frac{v_r^T \nabla_x h^T [\nabla_{xu}^2 h v_r] + v_i^T \nabla_x h^T [\nabla_{xu}^2 h v_i]}{\|\nabla_x h (v_r + i v_i)\|} \tag{4.69b}$$

$$\nabla_p \tilde{g}_{(\mathrm{om})} = \frac{v_r^T \nabla_x h^T [\nabla_{xp}^2 h v_r] + v_i^T \nabla_x h^T [\nabla_{xp}^2 h v_i]}{\|\nabla_x h (v_r + i v_i)\|} \tag{4.69c}$$

$$\nabla_{v_r} \tilde{g}_{(\mathrm{om})} = \frac{v_r^T \nabla_x h^T \nabla_x h}{\|\nabla_x h (v_r + i v_i)\|} \tag{4.69d}$$

$$\nabla_{v_i} \tilde{g}_{(\mathrm{om})} = \frac{v_i^T \nabla_x h^T \nabla_x h}{\|\nabla_x h (v_r + i v_i)\|} \tag{4.69e}$$

where the terms with second derivatives of h are defined in Eq. (4.63).

4.3 Identifying and Analyzing Particular Eigenmodes

Many aerospace systems display certain typical modal characteristics when operated at a typical equilibrium point (x^*, u^*, p^*). Conventional airplanes, for instance, usually expose modes that are labeled Short-Period, Phugoid, Dutch Roll, etc. The labels are assigned based on the eigenvalues and components of the eigenvectors of the linearized dynamics that indicate how the various state variables change relative to each other when the corresponding mode is excited. One may reasonably expect that a slight perturbation to the steady-state input values or parameters (and thus to the equilibrium state) will lead to a small change in the eigenmodes, but that their general characteristics will be preserved so that the same labels can still be assigned. This assignment may however not be obvious anymore for system equilibrium points that are not close to (x^*, u^*, p^*).

This leads to two interesting questions: If both (x^*, u^*, p^*) and (x^a, u^a, p^a) are equilibrium points of the system (4.1), is it possible to uniquely assign labels to the eigenmodes at (x^a, u^a, p^a) based on the labels of the eigenmodes at (x^*, u^*, p^*) through a continuity argument? If so, is it always the same eigenmode that satisfies $\tilde{g} = 0$ when computing the sets $\mathcal{S}(F^{(\mathrm{v})}, \Omega, z_0)$ or $\mathcal{S}(F^{(\mathrm{w})}, \Omega, z_0)$ defined in Section 2.2 by application of a continuation

method to the extended system functions? The following discussion addresses both of these questions.

Suppose $\Xi \subset \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_p}$ and g are chosen such that

$$\mathcal{M}_g^* = \{(x, u, p) \in \mathcal{D} \cap \Xi \mid f(x, u, p) = 0,\ g(x, u, p) = 0\} \tag{4.70}$$

is simply connected (i.e., there are no holes in the set) and such that $\nabla_x f$ does not have repeated eigenvalues anywhere on $\mathcal{M}_g^*$. Furthermore, assume that

$$(x^*, u^*, p^*) \in \mathcal{M}_g^* \qquad \text{and} \qquad (x^a, u^a, p^a) \in \mathcal{M}_g^*.$$

Because $\mathcal{M}_g^*$ is connected, there exists a continuous path Γ from (x^*, u^*, p^*) to (x^a, u^a, p^a) within $\mathcal{M}_g^*$. Since $\nabla_x f$ is continuous and does not have repeated eigenvalues on any such path, and because the eigenvalues of a matrix depend continuously on its entries, the set

$$\Gamma^{(\lambda)} = \{(x, u, p, \lambda) \in \Gamma \times \mathbb{C} \ \mid \ \nabla_x f(x, u, p) v = \lambda v \text{ for some } v \in \mathbb{C}^{n_x}\} \tag{4.71}$$

is the union of n_x disjoint paths $\Gamma_k^{(\lambda)}$ in $\Gamma \times \mathbb{C}$. Each element on Γ corresponds to exactly n_x elements in $\Gamma^{(\lambda)}$ (one for each eigenvalue of $\nabla_x f$) and to one element in each $\Gamma_k^{(\lambda)}$.

This allows assignment of labels to the system's eigenmodes at (x^a, u^a, p^a) based on the labels at (x^*, u^*, p^*): if $(x^*, u^*, p^*, \lambda_k^*)$ and $(x^a, u^a, p^a, \lambda_k^a)$ are both in $\Gamma_k^{(\lambda)}$, then λ_k^* and λ_k^a correspond to the same eigenmode. The assignment is unique: since $\mathcal{M}_g^*$ is simply connected, any two paths in $\mathcal{M}_g^*$ from the first point to the other can be continuously transformed into each other without leaving the set. The labeling is therefore independent of the actual path taken. Note that if $\nabla_x f$ was allowed to have repeated eigenvalues on $\mathcal{M}_g^*$, the sets $\Gamma_k^{(\lambda)}$ would not necessarily be disjoint, in which case it would not be possible to uniquely associate λ_k^a with λ_k^* for all $k \in \{1, \ldots, n_x\}$. Because the two points can be chosen arbitrarily, a one-to-one correspondence between the eigenmodes at any two equilibrium points in $\mathcal{M}_g^*$ exists. Furthermore, the modes at any point in $\mathcal{M}_g^*$ can be labeled based on the eigenmodes at one particular point in $\mathcal{M}_g^*$. This answers the first question: labels may be assigned to the eigenmodes at (x^a, u^a, p^a) based on the labels at (x^*, u^*, p^*) under the conditions that both

points lie on a simply connected surface of equilibrium points on which the Jacobian matrix $\nabla_x f$ does not have repeated eigenvalues.

The no-repeated-eigenvalues condition may be relaxed if the correspondence is to be established for only one particular eigenmode with index k^*. In this case, the path $\Gamma_{k^*}^{(\lambda)}$ must be disjoint from all other paths, but the other paths may intersect. This means that repeated eigenvalues are admissible for eigenmodes that do not correspond to $\lambda_{k^*}^*$ and $\lambda_{k^*}^a$.

Now suppose a continuation method is used to compute

$$\mathcal{S}\big(F_{(\mathrm{v})}, \Omega, z^*\big), \qquad \Omega = \Xi \times \Psi, \qquad \Psi \subset \mathbb{R}^{2+2n_x}, \qquad z^* = (x^*, u^*, p^*, \sigma^*, \eta^*, v_r^*, v_i^*)\,, \quad (4.72)$$

where $F_{(\mathrm{v})}$ was defined in (4.15), Ξ and the g part of $F_{(\mathrm{v})}$ is the same as in the definition of $\mathcal{M}_g^*$, and $(\sigma^*, \eta^*, v_r^*, v_i^*)$ correspond to a particular eigenmode of the system linearized at (x^*, u^*, p^*), so that $\nabla_x f(x^*, u^*, p^*)(v_r^* + i v_i^*) = (\sigma^* + i\eta^*)(v_r^* + i v_i^*)$. One can think of $\pi_{\mathcal{D}}\big(\mathcal{S}\big(F_{(\mathrm{v})}, \Omega, z^*\big)\big)$ as a subset of $\mathcal{M}_g^*$ with points that satisfy the eigenmode constraint $\tilde{g}(\cdot) = 0$, where $\pi_{\mathcal{D}}$ is the projection into $\mathcal{D}$ defined in (4.9). Because continuation methods compute connected sets, the elements in $\mathcal{S}\big(F_{(\mathrm{v})}, \Omega, z^*\big)$ are associated with the same eigenmode. This answers the second question: if Ξ is chosen such that $\pi_{\mathcal{D}}\big(\mathcal{S}\big(F_{(\mathrm{v})}, \Omega, z^*\big)\big)$ satisfies the two conditions of simple connectedness and absence of repeated eigenvalues, then it is always the same eigenmode that satisfies the constraint. However, it is difficult to define Ξ (and thus Ω) in advance in such a way that these requirements are met. Instead, the conditions can be checked after the computation is finished, or the continuation algorithm can be configured to dynamically restrict Ω during execution of the algorithm upon encountering certain stop conditions.

Simple connectedness of $\mathcal{M}_g^*$ may be enforced by stopping the continuation whenever the distance between the projection of the latest solution point into $\mathcal{D}$ and the projection of a previously encountered solution is less than some predefined tolerance.

To satisfy the second condition, the continuation must be stopped whenever a point is encountered at which $\sigma + i\eta$ becomes a repeated eigenvalue. Such points can be identified by monitoring the (square) submatrix V defined in (4.28) of the Jacobian $\nabla F_{(\mathrm{v})}$, which is

singular if and only if $\sigma + i\eta$ is a repeated eigenvalue of $\nabla_x f(x, u, p)$ (see Lemma 4.6).

A similar discussion holds for $F_{(\mathrm{w})}$ and the corresponding submatrix of $\nabla F_{(\mathrm{w})}$.

4.4 Example: Nonlinear Airplane Dynamics

Three analysis cases for a highly nonlinear 6-DOF aircraft model that make use of the analysis method are presented in this section. They serve to demonstrate the type of results that can be generated with the method and show how the constraint functions g and $\tilde{g}$ can be chosen to select particular trim conditions from the set of all possible equilibrium conditions.

4.4.1 Aircraft model

The aircraft is modeled as a rigid body with translational and rotational degrees of freedom subject to gravitational and aerodynamic forces [46]. The state vector consists of the translational velocity vector (U, V, W) and the angular velocity vector (P, Q, R), both expressed in the body frame, as well as the aircraft's attitude in terms of bank angle ϕ and pitch attitude θ. The third Euler angle (heading angle) and the position vector are not included in this formulation. The aircraft is controlled via the four inputs throttle δ_T, elevator angle δ_E, aileron deflection δ_A and rudder angle δ_R. The airplane's inertial properties are defined by its mass M_{ac} and the components J_x, J_y, J_z and J_{xz} of its moment of inertia tensor. The remaining inertia tensor components J_{xy} and J_{yz} are assumed to be zero due to symmetry of the airframe. The aerodynamic forces (D_s, Y, L_s) and moments (l, m, n) are functions of the states and inputs which are defined later. T_{max} denotes the maximum combined thrust

of the two engines. With these definitions, the state equations of the aircraft are

$$\begin{bmatrix} \dot{U} \\ \dot{V} \\ \dot{W} \\ \dot{P} \\ \dot{Q} \\ \dot{R} \\ \dot{\phi} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} VR - WQ - g\sin\theta + (T_{max}\delta_T + L_s\sin\alpha - D_s\cos\alpha)/M_{ac} \\ WP - UR + g\sin\phi\cos\theta + Y/M_{ac} \\ UQ - VP - g\cos\phi\cos\theta - (L_s\cos\alpha + D_s\sin\alpha)/M_{ac} \\ [J_{xz}(J_x - J_y + J_z)PQ - (J_z^2 - J_yJ_z + J_{xz}^2)QR + J_zl + J_{xz}n]/[J_xJ_z - J_{xz}^2] \\ [(J_z - J_x)PR - J_{xz}(P^2 - R^2) + m]/J_y \\ [(J_x^2 - J_xJ_y + J_{xz}^2)PQ - J_{xz}(J_x - J_y + J_z)QR + J_{xz}l + J_xn]/[J_xJ_z - J_{xz}^2] \\ P + (R\cos\phi + Q\sin\phi)\tan\theta \\ Q\cos\phi - R\sin\phi \end{bmatrix}. \tag{4.73}$$

Assuming flight in still air (no wind), the translational velocity is related to the aircraft's airspeed V_a, angle of attack α, and sideslip angle β through the following equations,

$$V_a = \sqrt{U^2 + V^2 + W^2}, \tag{4.74a}$$

$$\alpha = \text{atan2}\left(W, U\right), \tag{4.74b}$$

$$\beta = \text{atan2}\left(V, \sqrt{U^2 + W^2}\right). \tag{4.74c}$$

The aerodynamic model is taken from [47], which defines the aerodynamic forces

$$D_s = \bar{q}SC_{D_s}, \qquad Y = \bar{q}SC_Y, \qquad L_s = \bar{q}SC_{L_s}, \tag{4.75a}$$

in the stability frame and moments

$$l = \bar{q}bSC_l, \qquad m = \bar{q}\bar{c}SC_m, \qquad n = \bar{q}bSC_n \tag{4.75b}$$

in the body frame. In the above, S, b, and $\bar{c}$ are the aircraft's planform area, wing span, and mean aerodynamic chord, while $\bar{q} = \frac{1}{2}\rho V_a^2$ is dynamic pressure. The aerodynamic coefficients

are defined as multivariate polynomials

$$C_{D_s} = \theta_1 + \theta_2\alpha + \theta_3\alpha\tilde{q} + \theta_4\alpha\delta_E + \theta_5\alpha^2 + \theta_6\alpha^2\tilde{q} + \theta_7\alpha^2\delta_E + \theta_8\alpha^3 + \theta_9\alpha^3\tilde{q} + \theta_{10}\alpha^4 \quad (4.76a)$$

$$C_Y = \theta_{11}\beta + \theta_{12}\tilde{p} + \theta_{13}\tilde{r} + \theta_{14}\delta_A + \theta_{15}\delta_R \quad (4.76b)$$

$$C_{L_s} = \theta_{16} + \theta_{17}\alpha + \theta_{18}\tilde{q} + \theta_{19}\delta_E + \theta_{20}\alpha\tilde{q} + \theta_{21}\alpha^2 + \theta_{22}\alpha^3 + \theta_{23}\alpha^4 \quad (4.76c)$$

$$C_l = \theta_{24}\beta + \theta_{25}\tilde{p} + \theta_{26}\tilde{r} + \theta_{27}\delta_A + \theta_{28}\delta_R \quad (4.76d)$$

$$C_m = \theta_{29} + \theta_{30}\alpha + \theta_{31}\tilde{q} + \theta_{32}\delta_E + \theta_{33}\alpha\tilde{q} + \theta_{34}\alpha^2\tilde{q} + \theta_{35}\alpha^2\delta_E + \theta_{36}\alpha^3\tilde{q} + \theta_{37}\alpha^3\delta_E + \theta_{38}\alpha^4 \quad (4.76e)$$

$$C_n = \theta_{39}\beta + \theta_{40}\tilde{p} + \theta_{41}\tilde{r} + \theta_{42}\delta_A + \theta_{43}\delta_R + \theta_{44}\beta^2 + \theta_{45}\beta^3 \quad (4.76f)$$

where $\tilde{p}$, $\tilde{q}$, and $\tilde{r}$ are the non-dimensionalized angular rates,

$$\tilde{p} = \frac{b}{2V_a}P, \qquad \tilde{q} = \frac{\bar{c}}{2V_a}Q, \qquad \tilde{r} = \frac{b}{2V_a}R. \quad (4.77)$$

The values of the coefficients θ_j are published in [47] for a number of different aircraft. In this example, the coefficients for NASA's Generic Transport Model are used. This model aircraft is dynamically scaled to resemble a transport category aircraft. It has maximum total thrust $T_{max} = 32$ lbf, nominal weight $M_{ac} = 49.6$ lbm $= 1.542$ slug, planform area $S = 5.902$ ft^2, wing span $b = 6.849$ ft, and mean aerodynamic chord $\bar{c} = 0.915$ ft. The moments of inertia are $J_x = 1.327$ slug ft^2, $J_y = 4.254$ slug ft^2, $J_z = 5.454$ slug ft^2, and $J_{xz} = 0.120$ slug ft^2.

The aircraft's flight path angle γ and turning rate (heading angle rate) $\dot{\psi}$ are defined as

$$\gamma = \sin^{-1}\left[(U\sin\theta - V\sin\phi\cos\theta - W\cos\phi\cos\theta)/V_a\right] \quad (4.78)$$

$$\dot{\psi} = (Q\sin\phi + R\cos\phi)/\cos\theta \quad (4.79)$$

In all the following analysis cases, the computational domain Ω is chosen such that continuation stops when any of the inputs reach their upper or lower bounds. Further bounds are defined on all state variables (the bank angle ϕ, for instance, is limited to $\pm 90°$).

4.4.2 Analysis case 1: Pitch damper feedback gain calculation

As a first example, a pitch damper control law for the aircraft is synthesized. The design objective is to achieve a Short-Period mode damping coefficient of $\zeta_{sp}^{*} = 0.7$ at all flight speeds for three different types of flight conditions. For this, a feedback gain schedule is to be determined.

Continuation analysis enables direct computation of the feedback gain schedule required to obtain the desired damping while simultaneously solving the trim problem. The resulting gain schedule can either be used directly or serve as a starting point for the design of a simplified control law. Obtaining these results with conventional methods would require multiple steps: After trimming the aircraft to a large number of flight conditions and deriving linearized models, the gain value that results in the desired Short-Period mode damping would have to be determined at each trim point (for instance with root-locus techniques). The resulting gain-scheduled controller would then have to be inserted into the original model in order to analyze the closed-loop dynamics. The continuation-based approach consolidates these steps into a single computation, and information about all other eigenmodes is also obtained. Specification of the damping ratio constraint is only made possible by the extended system of equations introduced in this work. An idea related to this analysis case has been presented in [48], where the gain schedule to place the poles of a simple two-state system is computed with a continuation algorithm. The method presented in this chapter is more general, since pole placement can be achieved through a particular choice of eigenmode constraints.

For simplicity, the control law for the pitch damper is chosen as proportional pitch-rate to elevator feedback,

$$\delta_E = \delta_{E,\mathrm{cmd}} + k_Q Q, \tag{4.80}$$

where $\delta_{E,\mathrm{cmd}}$ is the pilot's elevator command input. In this analysis case, the closed-loop system is analyzed. The function f in (4.1) that defines the closed-loop system dynamics is obtained by substituting (4.80) into the equations that define the open-loop system, i.e.,

(4.73–4.77). The external inputs of the closed-loop system are $u = [\delta_T \quad \delta_{E,\text{cmd}} \quad \delta_A \quad \delta_R]^T$. The feedback gain k_Q is considered a system parameter, i.e., $p = k_Q$. It is positive due to the sign conventions for Q and δ_E.

The goal of this analysis is to determine the value of k_Q that is required to obtain the desired Short-Period mode damping for all airspeeds. Three types of conditions are considered, each of which is specified through three equality constraints:

- Straight flight at constant altitude: $\dot{\psi} = 0$, $\gamma = 0$ and $\beta = 0$
- Straight climbing flight: $\dot{\psi} = 0$, $\gamma - \gamma^* = 0$ and $\beta = 0$ (with $\gamma^* = 15°$)
- Coordinated banked turn at constant altitude: $\phi - \phi^* = 0$, $Y = 0$, and $\gamma = 0$ (with $\phi^* = 30°$)

These three constraints define the function g for each type of flight condition. Results for all airspeeds will be computed with the continuation algorithm.

The eigenvalue constraint $\tilde{g}_{(\zeta)} = 0$ as defined in (4.46) is imposed to fix the damping coefficient of the Short-Period mode at the desired value of ζ_{sp}^*. Note that the turning flight conditions may not have well-defined Short-Period and Phugoid modes due to the coupling of the longitudinal and lateral-directional dynamics. However, based on the continuity argument in Section 4.3, these names are still used in the following diagrams and discussions.

With eight states, four inputs, and one parameter, the number of variables is 13. A total of 12 equations are specified (eight equilibrium equations, three flight condition constraints, and one eigenvalue constraint). The number of degrees of freedom in the continuation problem is therefore $M = 1$.

For each of the three cases (and the corresponding constraints), the extended system of equations $F_{(\text{v})} = 0$ defined in (4.15) is solved with the pseudo-arclength method. The initial solution point used to initialize the algorithm is chosen such that the components of the point that represent the eigenvalue and eigenvector belong to the Short-Period mode. The eigenvector is normalized with (4.10) to ensure that the initial point satisfies (4.7).

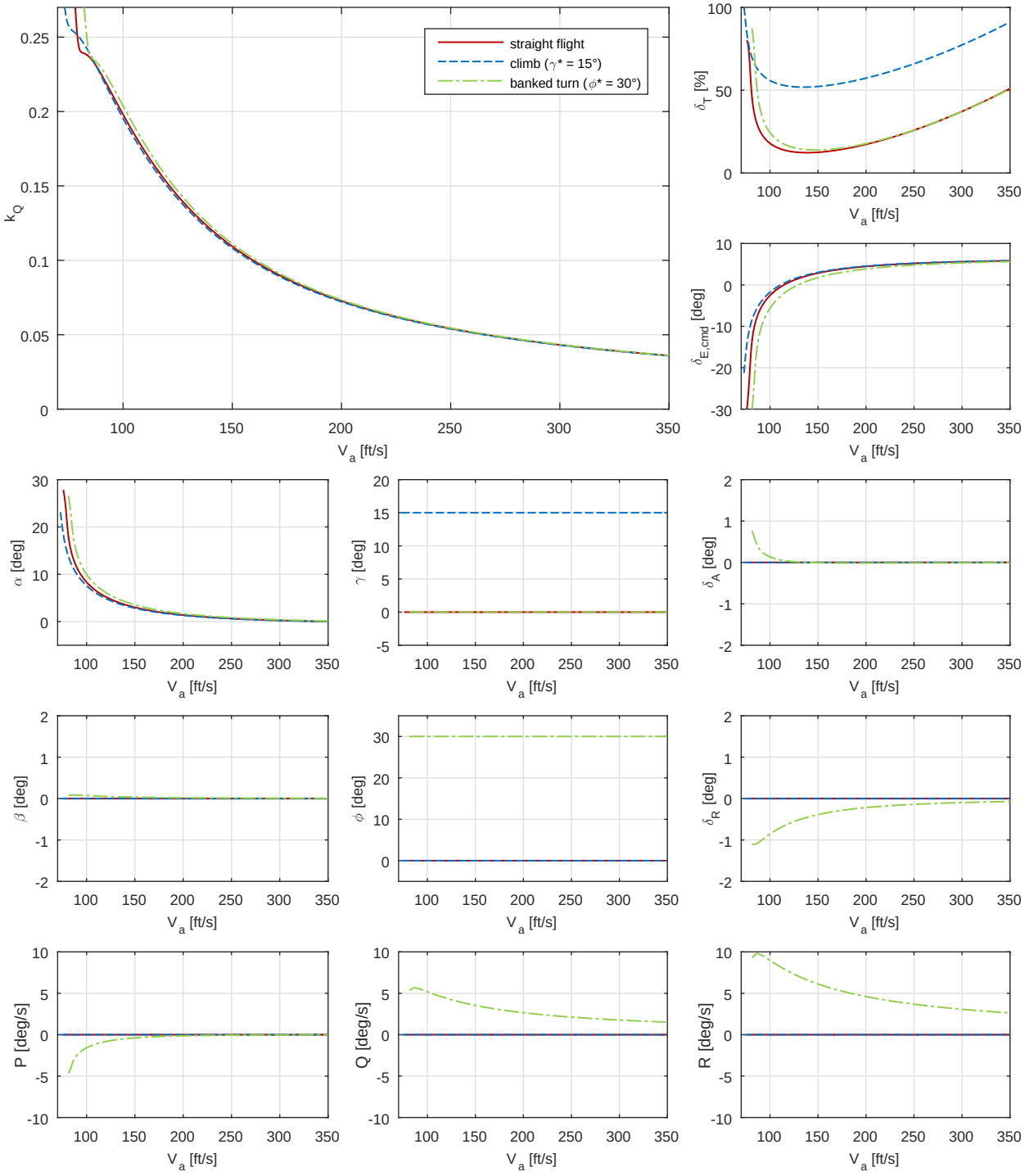

Figure 4.1: Pitch damper feedback gain schedules and trim values of states and inputs.

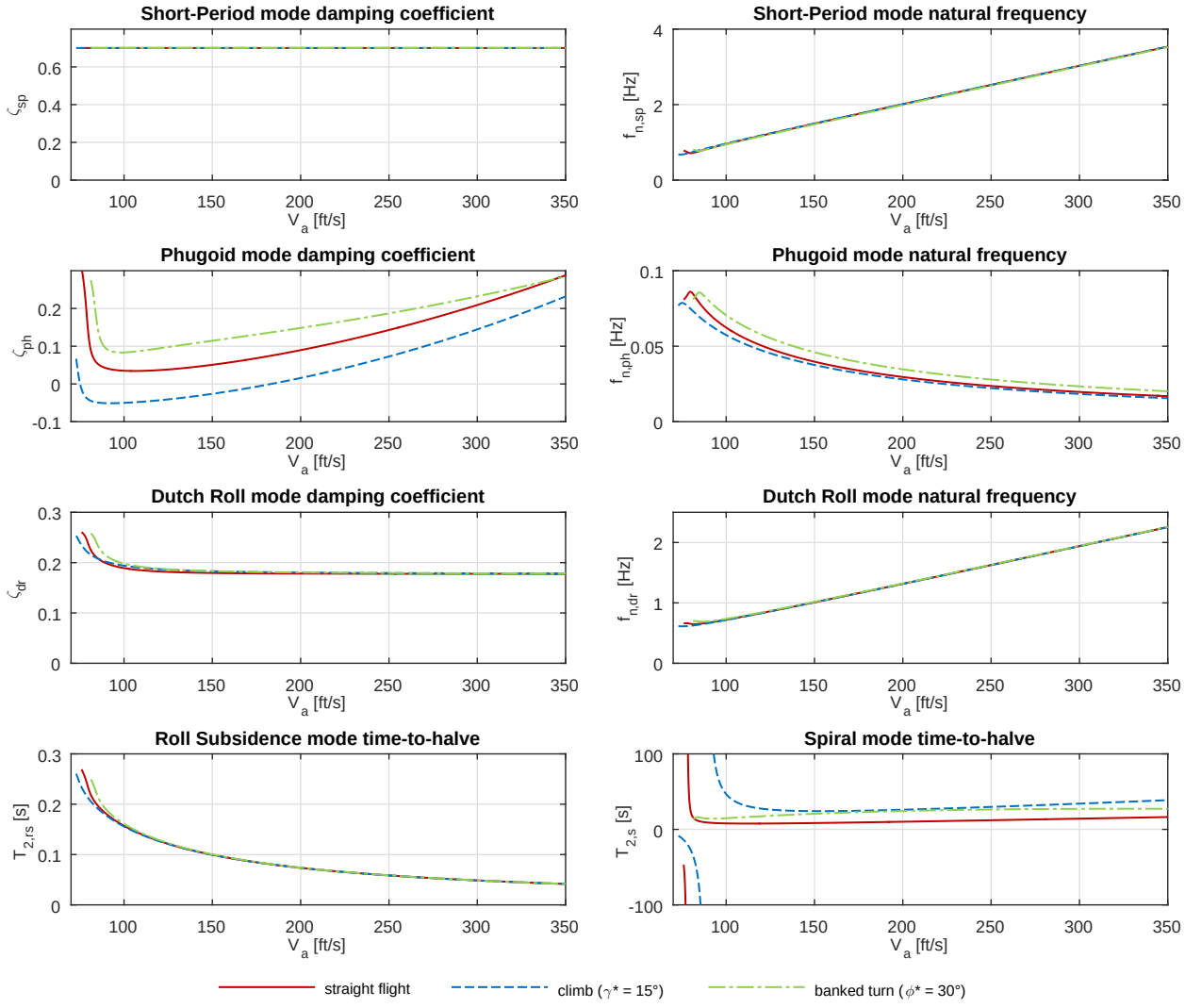

Figure 4.2: Eigenmode characteristics for different trim conditions with pitch damper and computed gain schedules.

The analysis results are plotted in Figure 4.1. Each of the diagrams in the figure displays a different projection of the three solution branches that correspond to the three types of flight conditions defined earlier. Every point on these branches corresponds to one particular trim point of the aircraft and includes the feedback gain value required to obtain the desired Short-Period mode damping at this trim condition.

For non-turning flight, the lateral-directional variables such as bank angle, aileron input, and rudder input are zero when the aircraft is in trim. As one would expect from basic energy considerations, the climbing flight conditions require much higher trim throttle settings than the constant-altitude cases. The coordinated flight constraint (zero side force) leads to non-zero sideslip angles for the banked turns. In all cases, trim airspeeds below 100 ft/s require excessively high angles of attack and the aerodynamic model loses validity.

While the feedback gain required to obtain the desired Short-Period mode damping depends strongly on the trim airspeed, it is virtually the same for all three types of flight conditions at any given airspeed. Scheduling the gain as a function of airspeed according to the results for the straight, constant-altitude flight case may thus be a good option for the pitch damper design.

The Jacobian matrix $\nabla F_{(\mathrm{v})}$ is evaluated at every solution point by the pseudo-arclength continuation method. The submatrix $\nabla_x f$ that corresponds to the linearized (closed-loop) system is therefore available as part of the continuation results so that the eigenvalues are readily calculated. To ensure consistency of the results, the eigenmode characteristics of the aircraft (obtained from the eigenvalues) are displayed in Figure 4.2. The figure demonstrates that the Short-Period mode damping is indeed equal to 0.7 in all cases for the computed gain schedules. The Phugoid mode characteristics differ considerably between the different flight conditions. For climbing flight, Phugoid mode damping actually becomes negative for low to medium airspeeds, indicating that the mode is unstable under these conditions. The second analysis case will investigate the Phugoid mode characteristics in more detail. The spiral mode is stable for all investigated flight conditions except at very low airspeeds, where the aerodynamic model breaks down.

Figure 4.3 shows the results of a time simulation of the closed-loop system for the particular flight condition of straight constant-altitude flight at 150 ft/s. The control input trim settings, initial state values, and feedback gain $k_Q = 0.11s$ are taken from the continuation results (Figure 4.1). After two seconds, an elevator doublet is commanded to excite the longitudinal modes. The nicely damped Short-Period mode response is clearly visible in the

angle of attack and pitch rate time histories, indicating that the pitch damper is working as expected.

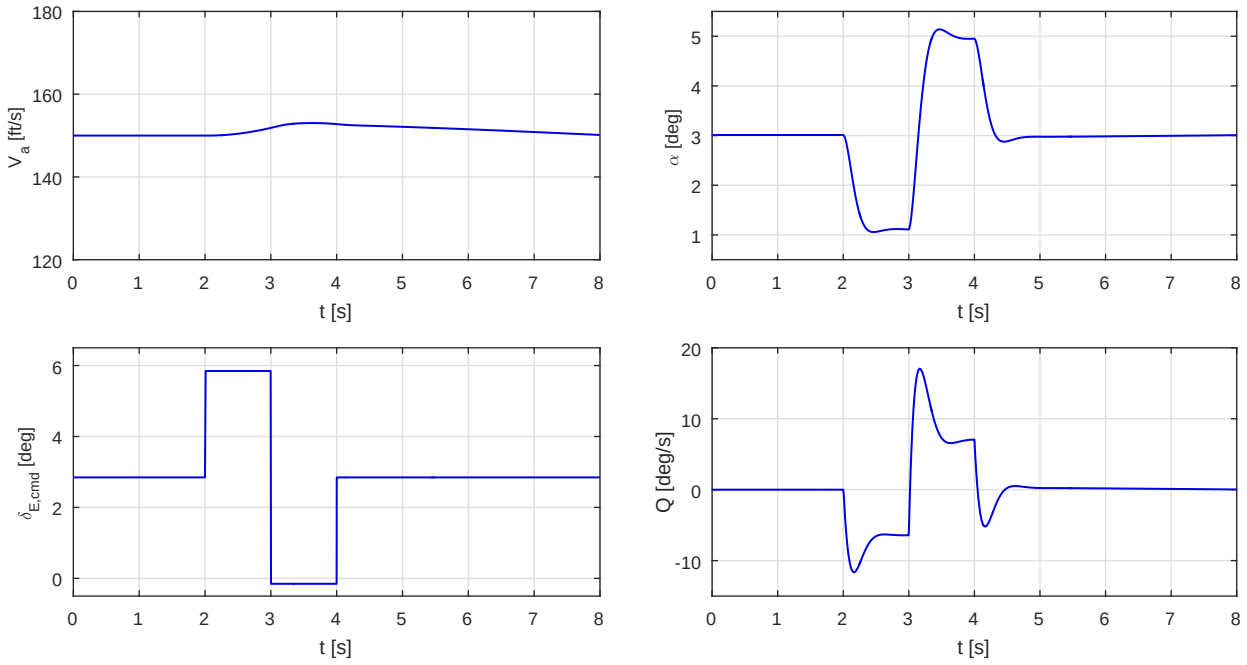

Figure 4.3: Time response to elevator doublet.

In practice, one may be interested in choosing a particular constant value of k_Q that keeps the Short-Period mode damping within certain bounds for all flight conditions instead of implementing a tabulated gain schedule. Repeating the above analysis for different values of ζ_{sp}^* (or just for the bounds) provides insight into how the constant gain value should be chosen.

4.4.3 Analysis case 2: Phugoid mode characteristics in non-turning flight

The purpose of the second analysis case is to gain insight into the Phugoid mode characteristics of the open-loop system. In particular, it is investigated how the Phugoid mode behaves at different airspeeds and flight path angles. Such information is for instance useful for handling quality assessments and to identify areas of the flight envelope where the

Phugoid mode is insufficiently damped.

Continuation analysis is first used to determine the limits of the flight envelope within which the aircraft can be trimmed without exceeding the control input limits. The Phugoid mode characteristics for the possible trim conditions within this envelope are then visualized through contour plots of damping coefficient (where the eigenmode is stable) and time-to-double (where it is unstable). Each contour line is obtained as the solution of a continuation problem with an eigenmode constraint and cannot be computed directly without the technique derived in this work. The resulting figure clearly exposes how the Phugoid mode changes for different trim airspeeds and flight path angles.

Equations (4.73–4.77) define the system function f in (4.1) for the open-loop system. The input vector of the open-loop system is $u = [\delta_T \quad \delta_E \quad \delta_A \quad \delta_R]^T$. Multiple sets of equilibrium points that satisfy different equality constraints are computed with the pseudo-arclength method. In all cases, the two constraints $\dot{\psi} = 0$ and $\beta = 0$ are imposed so that the equilibria correspond to non-turning flight with zero sideslip. Each case furthermore enforces one of the following constraints:

- Control input limits: $\delta_{T,\min} = 0$ (zero thrust), $\delta_{T,\max} - 1 = 0$ (full thrust), or $\delta_{E,\min} - (-30^\circ) = 0$ (minimum elevator angle)
- Angle of attack contours: $\alpha - \alpha^* = 0$, with $\alpha^* \in \{0.5^\circ, 1^\circ, 1.5^\circ, 2^\circ, 5^\circ, 10^\circ, 15^\circ, 20^\circ\}$
- Stability boundary: $\tilde{g}_{(\mathrm{s})} = 0$
- Damping coefficient contours: $\tilde{g}_{(\zeta)} = 0$, with $\zeta^* \in \{0.04, 0.08, 0.12, 0.16, 0.20, 0.24, 0.28\}$
- Time-to-double contours: $\tilde{g}_{(T_2)} = 0$, with $T_2^* \in \{15s, 25s, 35s, 45s, 55s\}$

Figure 4.4 shows the results for the Phugoid mode characteristics in straight flight. Every line in the diagram is the result of continuation analysis with one particular choice of constraints as discussed above. As in the first analysis case, this implies that a complete trim

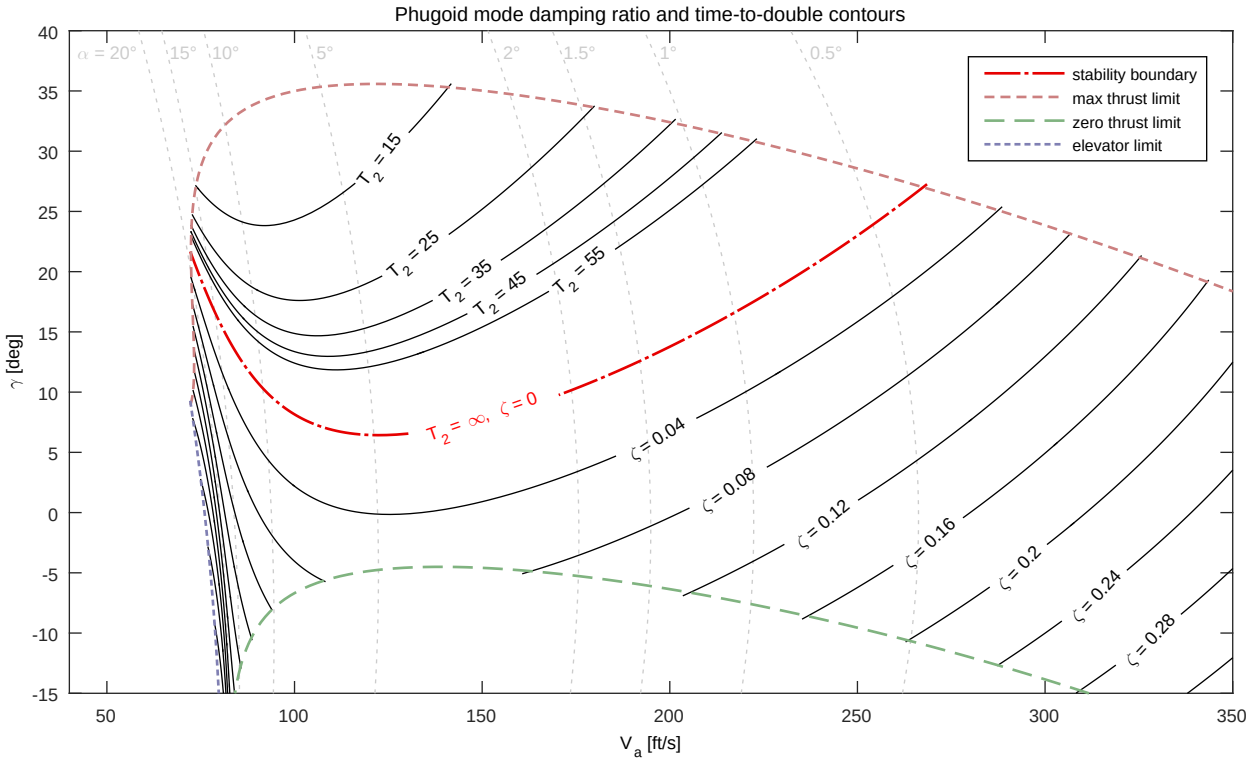

Figure 4.4: Phugoid mode characteristics for different trim airspeeds and trim flight path angles, non-turning flight.

solution is obtained as part of the results for every point on these lines. The projection of the data in Figure 4.4 is chosen such that every trim condition for non-turning flight with zero sideslip (characterized by particular values of airspeed and flight path angle) corresponds to one particular point in the diagram plane. The different contour lines thus indicate how various system properties associated with these trim conditions depend on airspeed and flight path angle. The lines of constant angle of attack, for instance, visualize the trim angle of attack for these conditions.

Flight envelope limits exist due to three actuator bounds. First, the maximum flight path angle at any airspeed is limited by engine thrust. The maximum of 35.6° marks the speed of steepest climb at 123 ft/s. When climb rate $\dot{h} = V_a \sin\gamma$ is projected onto the vertical axis instead of flight path angle, the maximum occurs at a higher airspeed (283 ft/s), known

as the speed of fastest climb. Second, for any given airspeed the maximum (negative) glide slope depends on how fast potential energy can be dissipated by drag at zero thrust. The aircraft's maximum glide ratio of $(\tan 4.5°)^{-1} = 12.7$ occurs at the maximum of the zero-thrust curve at 139 ft/s. To descend at a steeper angle, the aircraft must either increase its airspeed (and thus dynamic pressure and drag), decrease its airspeed (and therefore increase angle of attack, lift coefficient, and eventually drag), or use additional control surfaces such as spoilers and flaps. Third, at very low airspeeds, the elevator deflection required to trim the aircraft may reach the limit of the actuator. This bound is however to be considered with care, as the trim angle of attack in this region is very large and exceeds the limits of validity of the aerodynamic model.

Figure 4.4 indicates that the Phugoid mode is either stable or unstable, depending on airspeed and flight path angle. Contour lines indicate the damping ratio in the stable region. The stability boundary, at which the complex pair of eigenvalues crosses the imaginary axis, is plotted as a red line. In the unstable region, the contours show the time it takes for the mode to double its amplitude. Note that time-to-double tends towards infinity when approaching the stability boundary from the unstable side, so that no contour lines for $T_2 > 55s$ were calculated.

The U.S. military has published specifications in MIL-F-8785C, Flying Qualities of Piloted Airplanes, that classify an aircraft's flying qualities with regards to the Phugoid mode as follows: flying quality level 1 requires Phugoid mode damping $\zeta \geq 0.04$; level 2 requires $\zeta \geq 0.0$; and level 3 requires Phugoid mode time-to-double $T_2 \geq 55$. Figure 4.4 clearly visualizes how the aircraft ranks for different non-turning trim conditions.

4.4.4 Analysis case 3: Phugoid mode controllability in straight constant-altitude flight

The control effectiveness of the aircraft's control surfaces may be reduced when flying at atypical trim conditions that involve significant sideslip or other factors diverging from the usual coordinated flight trims. While the aircraft is rarely ever flown under such conditions, it is useful to identify areas within the envelope of possible trims where the effectiveness is

reduced, since emergency situations may require operation of the aircraft at atypical trim points. The specific case of interest for this example is the controllability of the Phugoid mode when flying straight with non-zero bank angle. Similar to the previous analysis case, two contour plots are generated by solving continuation problems subject to eigenvalue constraints. The contour lines in the two plots visualize Phugoid mode controllability from different subsets of the aircraft's control inputs and clearly illustrate how the controllability depends on trim bank angle and airspeed.

Typically, the aircraft's wings are aligned with the horizontal direction when it is not turning. It is however possible to trim the aircraft for straight flight with non-zero bank and sideslip angles. For sake of the example, it will be investigated how the choice of bank angle affects Phugoid mode controllability at different airspeeds if the mode is controlled from the elevator input or from the three other inputs combined. This type of information, for instance, may be of interest to determine emergency procedures for the case that elevator control is suddenly lost in flight. Based on the continuity argument from Section 4.3, the eigenmode is called Phugoid mode even when the trim bank angle becomes large and the eigenvector no longer displays the characteristics that allows identification of the Phugoid mode at straight-and-level trims.

To establish the conditions of straight flight and constant altitude, the constraint $g = [\dot{\psi} \quad \gamma]^T = 0$ is defined. This constraint is used with $F_{(\mathrm{w})}$ and $\tilde{g}_{(\mathrm{cm})}$ defined in (4.20) and (4.57) to obtain contours of modal controllability by repeating the continuation with different values of m_c. The analysis is carried out for two scenarios. First, only the elevator is considered as input, so that $\nabla_u f$ has one column. Throttle, aileron and rudder are thus deemed parameters. Next, the roles are reversed, so that $\nabla_u f$ has three columns that correspond to throttle, aileron and rudder. Controllability measure contours are obtained for both cases, so that Phugoid mode controllability can be compared. No further scaling is applied to the inputs. For the aircraft model, elevator deflection ranges from $-30°$ to $20°$, and aileron and rudder deflections are limited to $\pm 25°$. In all cases, this gives an input value range of approximately 0.87 radians, which is close to the throttle input range of 0 to 1.

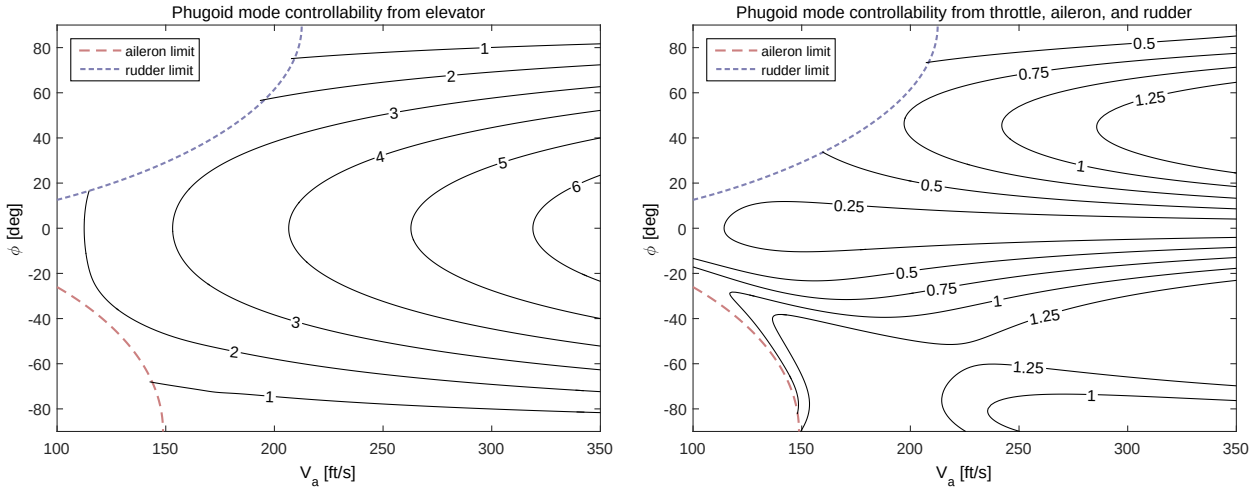

Figure 4.5: Phugoid mode controllability measure contours for controllability from elevator (left) and all other inputs (right) for different trim airspeeds and trim bank angles.

Results for the two analysis scenarios described above are plotted in Figure 4.5, where the solution branches are projected into the two coordinates airspeed and bank angle. The results are not symmetrical with regards to bank angle due to the β^2 term in the definition of the aerodynamic coefficient C_n in (4.76f).

Phugoid mode controllability from the elevator is largest for wings-level flight ($\phi = 0$) at all trim airspeeds and decreases with increasing bank angle. This is expected, since the control torque generated by the elevator becomes increasingly misaligned with the vertical plane in which the Phugoid motion's exchange of potential and kinetic energy takes place. The elevator effectiveness increases with airspeed, since the aerodynamic forces and moments depend on dynamic pressure.

The situation changes considerably when combined controllability from throttle, aileron, and rudder is considered. The controllability measure is lowest for wings-level flight, because the two aerodynamic surfaces only directly affect the lateral-directional modes in this type of trim condition. Only the throttle can be used to control the Phugoid mode. If the aircraft is banked so that the wings do not line up with the horizon, controllability is increased

significantly as the control surfaces start generating forces that affect the Phugoid mode.

At low airspeeds, the aircraft cannot be trimmed for straight flight with high bank angles because actuator limits are reached. The two curves in the figure that mark the boundary due to aileron and rudder saturation are obtained through continuation without the eigenmode constraint but with the additional condition that δ_A or δ_R are constant and equal to their respective limiting values.

4.5 Comparison with Gridded Analysis Approach

Instead of using continuation methods, analysis results similar to the figures in Section 4.4 may be obtained through a brute-force approach that involves computing a grid of trim points and analyzing the linearized model at all grid points. The grid point values are then interpolated by a contour plotting routine to produce diagrams resembling the continuation analysis plots.

To demonstrate the advantages of continuation analysis over the gridded approach, the Phugoid mode damping contour lines from the second analysis case in Section 4.4 are generated with both methods, Figure 4.6. The computations are carried out multiple times with different grid spacing and with different continuation step sizes and the resulting solution accuracies and computation times are recorded.

The gridded approach is described as follows: First, a regularly spaced grid of trim points is generated by repeatedly solving the equilibrium equation $\dot{x} = 0$ together with the constraint that airspeed and flight path angle have the particular values corresponding to each grid point. All numerical solutions are obtained with a standard nonlinear solver (MATLAB's "fsolve" function). For every grid point, the solver is initialized with a previously computed solution from an adjacent grid point. Next, the nonlinear model is linearized about each trim point using a central difference approximation to obtain linear models for eigenmode analysis. The complex conjugate pairs of eigenvalues and eigenvectors pertaining to the Phugoid mode are identified based on the eigenvector characteristics (largest components of airspeed and flight path angle), so that the Phugoid mode damping ratio can be calculated for

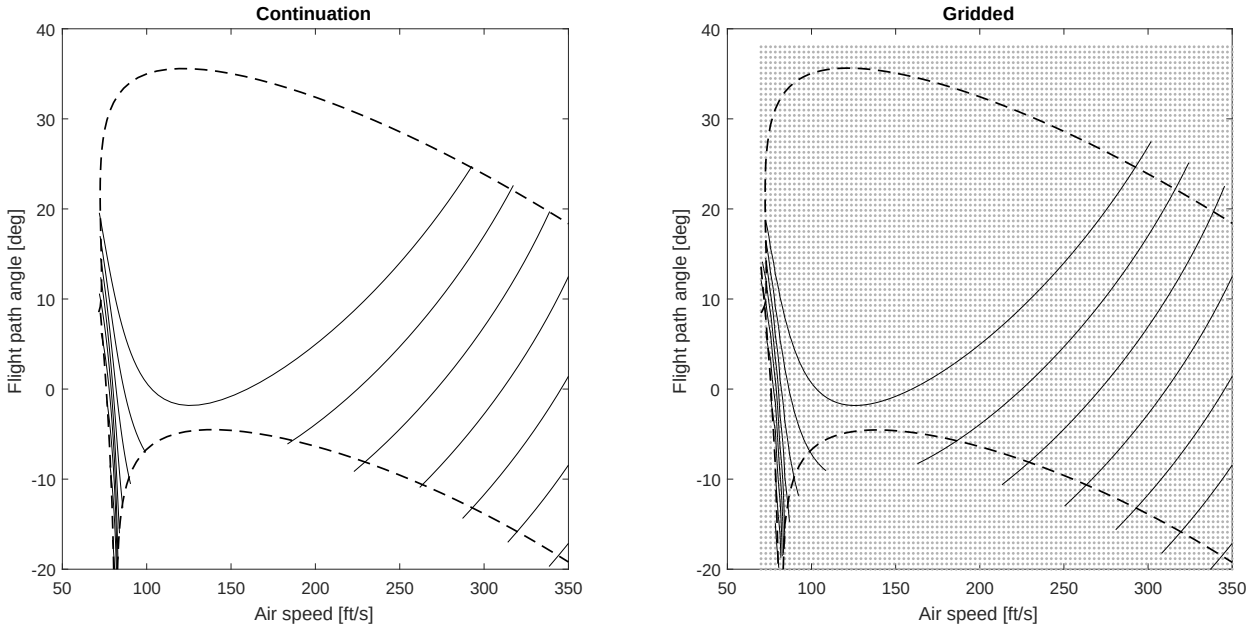

Figure 4.6: Damping ratio contour lines generated with continuation (left) and the gridded approach (right).

each grid point. Last, a standard contour plotting routine (MATLAB's "countour" function) interpolates the gridded data and generates the desired contour lines.

The contour lines obtained with either method are lists of points that specify air speeds and flight path angles at which the Phugoid damping is approximately equal to the corresponding contour value. The accuracy of a contour line depends both on the individual points and on the spacing in between the points. For this comparison, the accuracy is quantified according to the following procedure. First, all midpoints between the contour line points are linearly interpolated from the data and added to the original points. Second, complete trim solutions for these airspeeds and flight path angles are computed with the same solver used for the grid points. Next, the nonlinear model is linearized at all these trim conditions, the Phugoid mode damping is computed, and the difference between this value and the nominal value of the corresponding contour line is taken as the error for each point.

The mean and maximum error over all points and midpoints of all contour lines are

plotted over the associated computation time in Figure 4.7. The gridded approach requires significant more computation time to achieve the same accuracy as the continuation-based approach. All computations were carried out on a workstation with an Intel i7-4770 CPU and 8 GB of memory, using only a single core of the CPU.

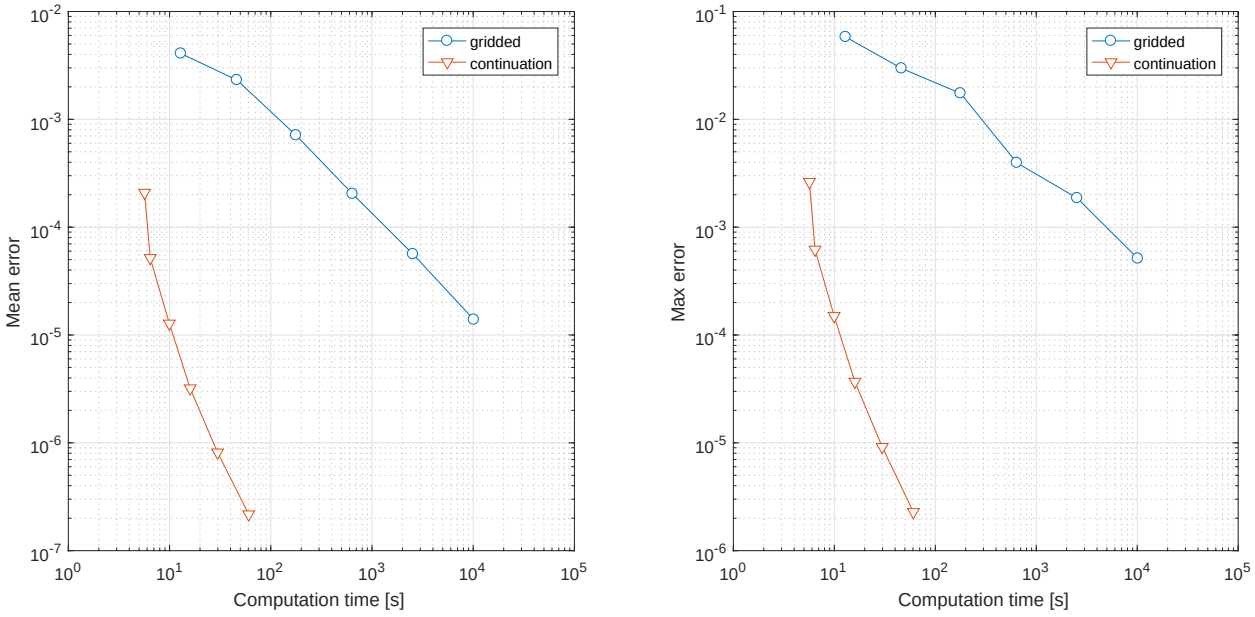

Figure 4.7: Solution error versus computation time.

Chapter 5

PSEUDO-ARCLENGTH CONTINUATION FOR PIECEWISE DIFFERENTIABLE PROBLEMS

The differentiability requirement imposed by the pseudo-arclength method is often violated by practical engineering problems, for instance due to linearly interpolated data tables contained in the models that are analyzed. In cases where the equations are continuous, but only piecewise differentiable, the continuation process may or may not fail upon crossing of a non-differentiable hypersurface.

This section investigates conditions under which the algorithm succeeds or fails for such problems. A mitigation strategy is proposed based on the insights gained. Two examples illustrate the effectiveness of this strategy. Much of the research in this chapter has been published in [49] by the author.

5.1 Implicit Lipschitz Functions

Differentiability of the function F is not essential for the existence of a locally defined implicit function whose graph coincides with the kernel of F. The mathematical area of nonsmooth analysis [50,51] defines the *generalized Jacobian* matrix $\widehat{\nabla}h(x)$ of a locally Lipschitz function $h : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$ as

$$\widehat{\nabla}h(x) = \mathrm{co}\Big\{ \lim_{i\to\infty} \nabla h(x_i) \;\Big|\; \nabla h(x_i) \text{ exists}, x_i \to x \Big\}, \tag{5.1}$$

where $\mathrm{co}\{\cdot\}$ denotes the convex hull. The generalized Jacobian thus assigns to every x a set that consists of the convex hull of the values that the Jacobian matrix ∇h approaches as it is evaluated on any sequence that tends towards x on which ∇h is defined. Such sequences always exist, since Rademacher's theorem states that Lipschitz functions are differentiable almost everywhere. The generalized Jacobian is a *set-valued function* (also called *multifunction*

or *correspondence* in the literature) that is upper-semicontinuous and produces non-empty convex compact subsets of $\mathbb{R}^{n_2 \times n_1}$ when evaluated at any point in its domain [50]. Note that if h is differentiable at x, then $\widehat{\nabla} h(x) = \{\nabla h(x)\}$. The generalized Jacobian is said to be of *maximal rank* at x if every matrix in $\widehat{\nabla} h(x)$ is of maximal rank.

An implicit function theorem for Lipschitz functions may be formulated by replacing the invertability condition on the Jacobian matrix in Theorem 2.1 with one on the generalized Jacobian. The following statement of the theorem is adapted from [50, 52]:

Theorem 5.1 (Implicit Lipschitz Function Theorem)**.** *Let H be a Lipschitz mapping defined on an open set $\mathcal{U} \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and taking values in $\mathbb{R}^{n_1}$. Let $(z_1^*, z_2^*) \in \mathcal{U}$ satisfy $H(z_1^*, z_2^*) = 0$ and suppose that all matrices in the set*

$$\left\{ J_1 \in \mathbb{R}^{n_1 \times n_1} \;\middle|\; \begin{bmatrix} J_1 & J_2 \end{bmatrix} \in \widehat{\nabla} H(z_1^*, z_2^*) \quad \text{for some } \; J_2 \in \mathbb{R}^{n_1 \times n_2} \right\} \tag{5.2}$$

are invertible. Then there exists a neighborhood $\widetilde{\mathcal{U}}$ of (z_1^*, z_2^*), an open set $\mathcal{W} \subseteq \mathbb{R}^{n_2}$ containing z_2^*, and a Lipschitz function ϕ defined on $\mathcal{W}$ and taking values in $\mathbb{R}^{n_1}$ such that

$$H(\phi(z_2), z_2) = 0 \qquad \text{for all} \quad z_2 \in \mathcal{W} \tag{5.3}$$

and ρ defined by

$$2\rho^{-1} = \min_{\begin{bmatrix} J_1 & J_2 \end{bmatrix} \in \widehat{\nabla} H(z_1^*, z_2^*)} \underline{\sigma} \left(\begin{bmatrix} J_1 & J_2 \\ 0 & I \end{bmatrix} \right) \tag{5.4}$$

is a Lipschitz constant of ϕ. Furthermore, ϕ is the only function satisfying

$$\left\{ (z_1, z_2) \in \widetilde{\mathcal{U}} \;\middle|\; H(z_1, z_2) = 0 \right\} = \left\{ (z_1, z_2) \in \widetilde{\mathcal{U}} \;\middle|\; z_2 \in \mathcal{W}, z_1 = \phi(z_2) \right\}. \tag{5.5}$$

The notation $\underline{\sigma}(\cdot)$ refers to the minimum singular value of the argument. While the original theorem does not state the expression for the Lipschitz constant, it follows directly from the proofs in [52] and [53].

The applicability of the above theorem depends on the coordinate system used, since the invertibility condition on the set (5.2) is coordinate dependent [54]. The following example illustrates this: Plotted on the left in Figure 5.1 is the kernel of the function $F(z_1, z_2) =$

$z_1 + z_2 - 2|z_1 - z_2|$. It is not possible to locally parametrize this set in terms of either coordinate at the origin. However, when the coordinate transformation $\hat{z}_1 = z_1 - z_2$ and $\hat{z}_2 = z_1 + z_2$ is applied (shown on the right in Figure 5.1), $\ker F$ may be expressed as the graph of a function of $\hat{z}_1$.

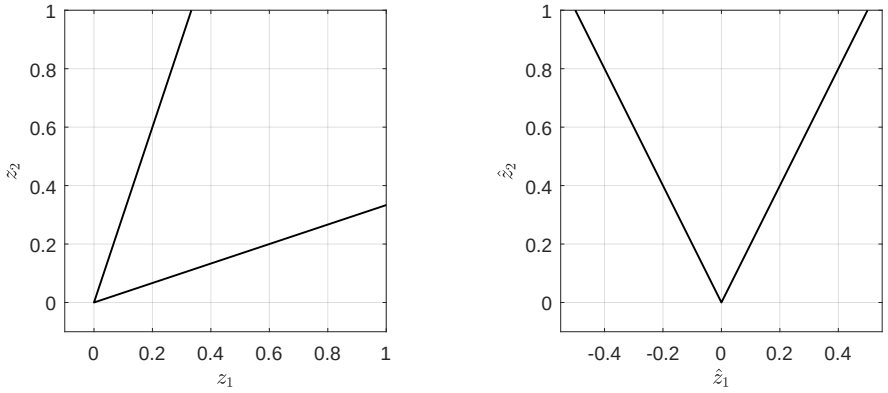

Figure 5.1: Kernel of the example function in the original and new coordinates.

The following theorem introduces conditions under which coordinate transformations of this type exists for more general continuous, piecewise differentiable functions:

Theorem 5.2. *Let $\mathcal{Z}$ be a neighborhood of $z^* \in \mathbb{R}^N$ and consider the continuous, piecewise differentiable function $F : \mathcal{Z} \to \mathbb{R}^{N-M}$ defined by the three continuously differentiable functions $F_b : \mathcal{Z} \to \mathbb{R}^{N-M}$, $F_c : \mathcal{Z} \to \mathbb{R}^{N-M}$, and $F_{sw} : \mathcal{Z} \to \mathbb{R}$ as*

$$F(z) = \begin{cases} F_b(z) & \text{if } F_{sw}(z) \geq 0 \\ F_c(z) & \text{otherwise} \end{cases} \tag{5.6}$$

satisfying $F(z^*) = 0$ and $F_{sw}(z^*) = 0$. Suppose the Jacobian matrices of F_b, F_c, F_{sw} have full rank at z^* and the null space of $\nabla F_b(z^*)$ is not a subspace of the null space of $\nabla F_{sw}(z^*)$.

Then there exist a neighborhood $\mathcal{Y}$ of the origin in $\mathbb{R}^{N-M} \times \mathbb{R}^M$ as well as matrices $E_1 \in \mathbb{R}^{N \times (N-M)}$ and $E_2 \in \mathbb{R}^{N \times M}$ such that $[E_1 \quad E_2]$ is orthogonal and the function $H : \mathcal{Y} \to \mathbb{R}^{N-M}$ defined by $H(y_1, y_2) = F(z^* + E_1 y_1 + E_2 y_2)$ satisfies the conditions of Theorem 5.1 at the origin.

Proof. Let $\bar{z}_{sw} : \mathcal{S} \subset \mathbb{R}^{N-1} \to \mathbb{R}^N$ be a parametrization of the graph of the implicit function defined by $F_{sw}(z) = 0$ in a neighborhood of z^* satisfying $\bar{z}_{sw}(0) = z^*$. Due to continuity of F, the equality $F_b(\bar{z}_{sw}(s)) = F_c(\bar{z}_{sw}(s))$ holds for all $s \in \mathcal{S}$. Differentiating both sides with respect to s yields $\nabla F_b(\bar{z}_{sw}(s))\, \nabla \bar{z}_{sw}(s) = \nabla F_c(\bar{z}_{sw}(s))\, \nabla \bar{z}_{sw}(s)$ and in particular, with $s = 0$,

$$\nabla F_b(z^*)\, \nabla \bar{x}_{sw}(0) = \nabla F_c(z^*)\, \nabla \bar{x}_{sw}(0) \tag{5.7}$$

Note that the columns of $\nabla \bar{z}_{sw}(0)$ form a minimal spanning set for the null space of $\nabla F_{sw}(z^*)$ and that the intersection of the null space of $\nabla F_{sw}(z^*)$ and $\nabla F_b(z^*)$ is $(M-1)$-dimensional. It is therefore always possible to find a matrix $\hat{S} \in \mathbb{R}^{(N-1)\times(N-1)}$ such that the columns of $\nabla \bar{x}_{sw}(0)\hat{S} = [E_1 \quad \hat{E}_2]$ form an orthonormal set and such that $\nabla F_b(z^*)E_1$ is invertible and $\nabla F_b(z^*)\hat{E}_2 = 0$ holds, where $E_1 \in \mathbb{R}^{N\times(N-M)}$, $\hat{E}_2 \in \mathbb{R}^{N\times(M-1)}$. Multiply (5.7) with $\hat{S}$ from the right to see that the equalities $\nabla F_c(z^*)E_1 = \nabla F_b(z^*)E_1$ and $\nabla F_c(z^*)\hat{E}_2 = \nabla F_b(z^*)\hat{E}_2$ hold. Now let $\hat{e}$ be such that $[E_1 \quad \hat{E}_2 \quad \hat{e}]$ is an orthogonal matrix and define $E_2 = [\hat{E}_2 \quad \hat{e}]$ as well as $A = \nabla F_b(z^*)E_1 = \nabla F_c(z^*)E_1$. With $H(y_1, y_2) = F(z^* + E_1 y_1 + E_2 y_2)$, the definition of the generalized Jacobian in (5.1) leads to

$$\begin{aligned}
\widehat{\nabla} H(0,0) &= \mathrm{co}\{[\nabla F_b(z^*)E_1 \quad \nabla F_b(z^*)E_2],\ [\nabla F_c(z^*)E_1 \quad \nabla F_c(z^*)E_2]\} && (5.8)\\
&= \mathrm{co}\{[A \quad \nabla F_b(z^*)E_2],\ [A \quad \nabla F_c(z^*)E_2]\} && (5.9)\\
&= \left\{ [A \quad B] \;\middle|\; B = \alpha \nabla F_b(z^*)E_2 + (1-\alpha)\nabla F_c(z^*)E_2 \text{ for some } \alpha \in [0,1] \right\} && (5.10)
\end{aligned}$$

Since A is invertible, H satisfies the conditions of Theorem 5.1 at the origin. ■

Corollary 5.3. *Suppose F satisfies the conditions of Theorem 5.2 at z^*. Then* $\ker F$ *is an M-dimensional surface in a neighborhood of z^*.*

Proof. By Theorem 5.2, $\ker F$ is the graph of a Lipschitz function of M variables near z^*. ■

The theorem essentially states that $\ker F$ is an M-dimensional surface near any point on the switching surface whenever the implicit surfaces defined by $F_b(z) = 0$ and $F_c(z) = 0$ intersect the switching surface transversally. The remainder of this chapter is concerned with pseudo-arclength continuation of problems where $M = 1$, so that $\ker F$ consists of solution branches (i.e, one-dimensional surfaces).

5.2 Continuation Algorithm Failure at the Switching Surface

The pseudo-arclength continuation method may stall at the switching surface, resulting in incomplete computation of the desired solution branch. This is the case even when the conditions of Theorem 5.1 are satisfied, so that the branch is well-defined on both sides of as well as on the switching surface. Analysis in this section reveals that convergence of the algorithm's corrector step is closely related to the angle between the two solution branch segments that meet at the switching surface.

5.2.1 The corrector step of the pseudo-arclength method

As discussed in Section 2.2 and illustrated in Figure 2.1, the pseudo-arclength algorithm successively computes points in $\ker F$ by first taking a predictive step along the tangential direction of the solution branch, followed by an iterative Newton-type method to correct for the curvature in the solution branch.

If the corrector step does not converge due to the nonlinearity of F, the step size is reduced and the predictor and corrector steps are repeated from the last valid solution point. The predictor point is now closer to the solution branch, so that the corrector step is more likely to succeed. Differentiability of F guarantees that it is always possible to proceed along the branch by making the step size small enough, because sufficient reduction in step size will always reduce the local nonlinearity in the region around the predictor point to the extent that the iterative corrector succeeds.

Two commonly used variations of Newton's method for the corrector step are considered in the following analysis.

The standard Newton-Raphson method can be applied by appending a step size constraint g to F, such that $g(z^*) = 0$ when $\|z^* - z_{old}\| = \Delta$, where Δ is the step size. With

$$H(z) = \begin{bmatrix} F(z) \\ g(z) \end{bmatrix}, \tag{5.11}$$

and $z_0 = z'$ the Newton iteration rule is

$$z_{k+1} = z_k - \left(\nabla_z H(z_k)\right)^{-1} H(z_k). \tag{5.12}$$

The step size constraint may be defined in several ways. The pseudo-arclength algorithm implementation in the popular bifurcation analysis code AUTO [25] defines

$$g(z) = \frac{\|z - z_{old}\|^2}{\Delta} - \Delta, \tag{5.13}$$

which will also be the constraint used throughout this chapter.

A modified Newton-Raphson method that works without the specification of a step size constraint uses the Moore-Penrose pseudo-inverse (indicated by the superscript †) in the iteration rule,

$$z_{k+1} = z_k - \left(\nabla F(z_k)\right)^{\dagger} F(z_k). \tag{5.14}$$

where

$$\nabla F(z_k)^{\dagger} = \nabla F(z_k)^T \left(\nabla F(z_k) \nabla F(z_k)^T\right)^{-1}. \tag{5.15}$$

With this corrector, the distance between z_{old} and the converged corrector point z^* will only approximately equal the step size Δ that was taken during the predictor step.

The number of iterations is either fixed or limited to a maximum value, and convergence is determined by comparison of $F(z_k)$ to a numerical solution tolerance $0 < \epsilon \ll 1$. As will become apparent in the following section, the difficulty of crossing a nondifferentiable surface results from the fact that step size reductions do not make the problem locally "more linear" in the neighborhood of the surface. In practice, the corrector step may not converge for any step size, causing the algorithm to reduce the step size until some pre-defined minimum value is reached, at which point the continuation process is aborted due to non-convergence.

5.2.2 A simple example

Consider the following nonlinear function of the two scalar variables $x, p \in \mathbb{R}$ with constant parameters $b, c > 0$:

$$F(x,p) = \begin{cases} x - bp & \text{if } p \geq 0 \\ x + cp & \text{otherwise} \end{cases} \tag{5.16}$$

The function F is continuous and piecewise smooth, but not differentiable along the switching surface defined by the line $\{\,(x,0) \mid x \in \mathbb{R}\,\}$. The solution branch that satisfies $F(x,p) = 0$ consists of two rays that start at the origin, indicated with green and blue lines in Figure 5.2. When applied to this problem, the pseudo-arclength algorithm will successfully proceed along the solution branch as long as it does not cross the switching surface where $p = 0$ during the predictor step. The remainder of this section investigates what exactly happens once the algorithm proceeds past this critical value.

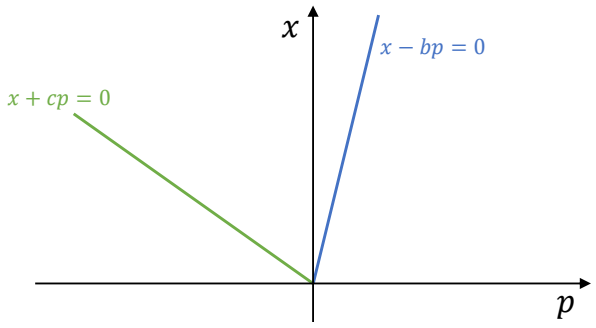

Figure 5.2: Kernel of the piecewise differentiable function defined in 5.16.

The last solution point before the switching surface is reached is called (x_{old}, p_{old}) in the following. Without loss of generality, it is assumed that $x_{old} > 0$ and $p_{old} < 0$. Because the solution is assumed to have been obtained numerically during the previous corrector step, it only solves the problem in an approximate sense, i.e., $|x_{old} + cp_{old}| < \epsilon \ll 1$, so that it is close to but not necessarily on the true solution branch. The following analysis assumes that $x_{old}^2 + p_{old}^2 < \Delta^2$. This inequality is always satisfied if the predictor step crosses the switching

surface from a true solution point along the exact tangential direction, but may not hold true under approximate conditions when $x_{old}^2 + p_{old}^2 \approx \Delta^2$.

Newton-Raphson corrector with step size constraint

Substitution of (5.16) and (5.13) into (5.11) leads to the following corrector iteration rule according to (5.12) after some algebraic manipulations:

$$\begin{bmatrix} x_{k+1} \\ p_{k+1} \end{bmatrix} = \begin{cases} \dfrac{x_k^2 + p_k^2 - x_{old}^2 - p_{old}^2 + \Delta^2}{2(bx_k - bx_{old} + p_k - p_{old})} \begin{bmatrix} b \\ 1 \end{bmatrix} & \text{if } p_k > 0 \\ \dfrac{x_k^2 + p_k^2 - x_{old}^2 - p_{old}^2 + \Delta^2}{2(cx_k - cx_{old} - p_k + p_{old})} \begin{bmatrix} c \\ -1 \end{bmatrix} & \text{if } p_k < 0 \end{cases} \tag{5.17}$$

Note that

$$x_{k+1} = \begin{cases} bp_{k+1} & \text{if } p_k > 0 \\ -cp_{k+1} & \text{if } p_k < 0 \end{cases} \tag{5.18}$$

Now consider the situation depicted in Figure 5.3. The last approximate solution point (x_{old}, p_{old}) lies so close to the switching surface that the desired new solution point (at a distance of one step size Δ), designated P_1 in the figure, is located on the other side. However, based on the Jacobian of H evaluated at (x_{old}, p_{old}), the predictor step will estimate that the new solution point lies close to P_2. As will be shown, successful convergence towards the desired solution point P_1 depends on wether or not the predictor point z' lies within region A indicated in Figure 5.3.

First, it is shown that for any starting point (of the Newton iterations) in region A or B in Figure 5.3, the sign of p_k switches with every iteration if

$$bc > 1 \qquad \text{and} \qquad bx_{old} + p_{old} > 0. \tag{5.19}$$

For any (x_k, p_k) that lies in region A, defined by $p_k > 0$ and $bx_k - bx_{old} + p_k - p_{old} < 0$, the iteration rule (5.17) then leads to

$$p_{k+1} = \frac{x_k^2 + p_k^2 - x_{old}^2 - p_{old}^2 + \Delta^2}{2(bx_k - bx_{old} + p_k - p_{old})} < 0, \tag{5.20}$$

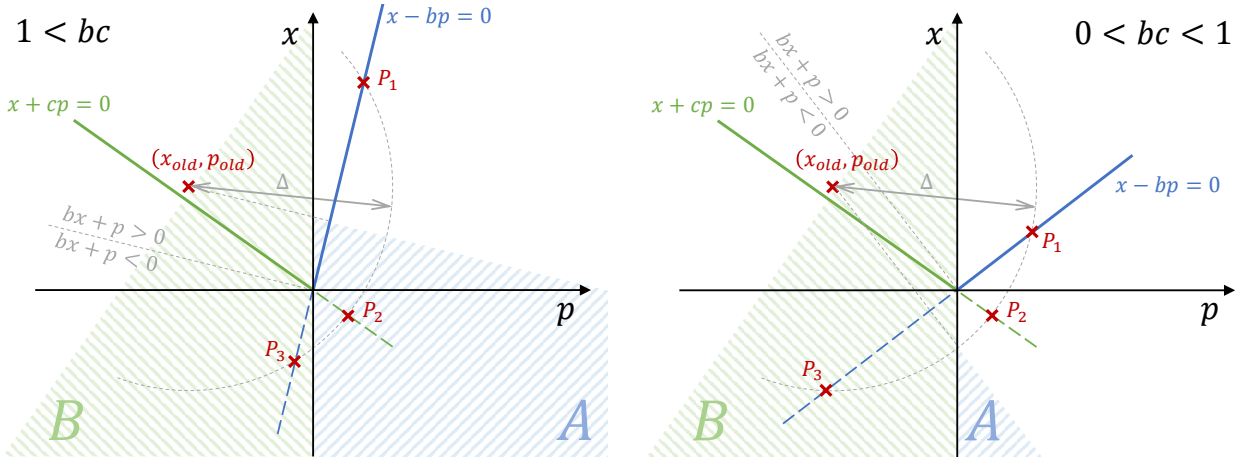

Figure 5.3: Geometry associated with the Newton-Raphson corrector.

since $x_{old}^2 + p_{old}^2 < \Delta^2$. Furthermore, considering (5.18), the inequality

$$cx_{k+1} - cx_{old} - p_{k+1} + p_{old} = \underbrace{(bc-1)}_{>0}\underbrace{p_{k+1}}_{<0} - \underbrace{cx_{old}}_{>0} + \underbrace{p_{old}}_{<0} < 0 \tag{5.21}$$

holds. Similarly, for any (x_k, p_k) that lies in region B, defined by $p_k < 0$ and $cx_k - cx_{old} - p_k + p_{old} < 0$, the following inequalities hold if (5.19) is satisfied:

$$p_{k+1} = -\frac{x_k^2 + p_k^2 - x_{old}^2 - p_{old}^2 + \Delta^2}{2(cx_k - cx_{old} - p_k + p_{old})} > 0, \tag{5.22}$$

$$bx_{k+1} - bx_{old} + p_{k+1} - p_{old} = \underbrace{(1-bc)}_{<0}\underbrace{p_{k+1}}_{>0} - \underbrace{(bx_{old} + p_{old})}_{>0} < 0 \tag{5.23}$$

A single iteration from any point in region A thus always results in a point in region B and vice versa, so that the sign of p_k changes with every iteration. The relations between x_{k+1} and p_{k+1} mean that the points always lie on the dashed blue and green lines in Figure 5.3 after the first iteration.

Figure 5.4 shows several iterations (indicated by red arrows) starting from a point one step size away from the known solution on the extension of the green line for three values of bc. The value of b is changed to obtain $bc \in \{0.4, 1.4, 2.0\}$. For $bc > 1$, the discrete system defined by the iteration rule either is unstable or enters a stable period-2 cycle.

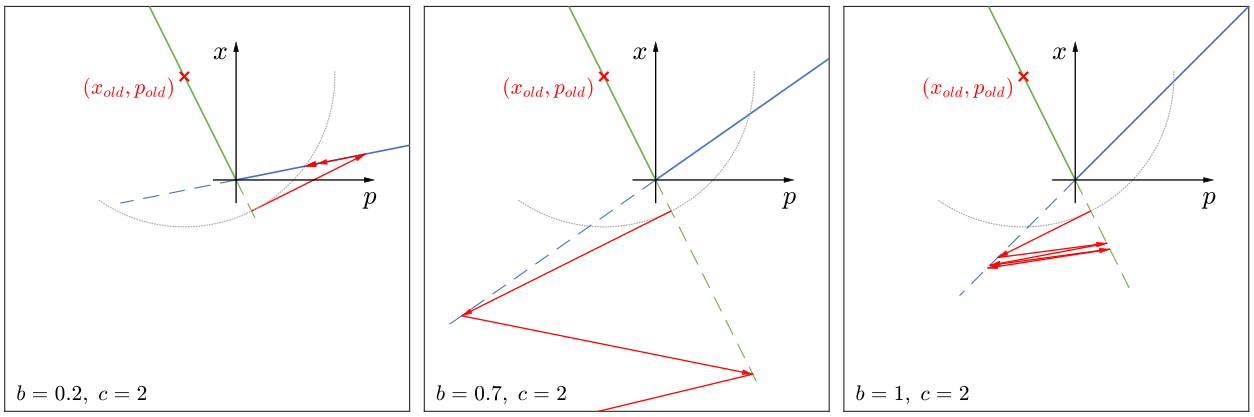

Figure 5.4: Several iterations of the Newton-Raphson corrector starting from the predictor point for different values of bc.

At this point, the two conditions defined in (5.19) deserve some discussion. Geometrically, the inequality $bc > 1$ means that the angle between the solid blue and green lines in Figure 5.3 is less than 90°. The second relation $bx_{old} + p_{old} > 0$ ensures that the dashed green line is completely contained inside region A when $bc > 1$. Figure 5.3 shows the relevant geometry for both $bc > 1$ and $1 > bc > 0$. Applying the iteration rule (5.17) to any point in region A or B still results in a sign change of p_{k+1}, since (5.20) and (5.22) still hold. In particular, any point in B is mapped onto the dashed green line in Figure 5.3 due to (5.18). However, the dashed green line is no longer contained in region A. The sign of p_{k+1} remains positive when applying (5.17) to any point with $p_k > 0$ that is not in A (so that $bx_k - bx_{old} + p_k - p_{old} > 0$), since the denominator in (5.20) is now positive. Furthermore, with $0 < bc < 1$ and $bx_{old} + p_{old} < 0$, the relation

$$bx_{k+1} - bx_{old} + p_{k+1} - p_{old} = \underbrace{(b^2+1)}_{>0}\underbrace{p_{k+1}}_{>0} - \underbrace{(bx_{old} + p_{old})}_{<0} > 0 \tag{5.24}$$

shows that the new point remains outside of region A. Therefore, the iterations will not oscillate between positive and negative values of p in this case, so that the solution will eventually converge. Furthermore, when starting in region B or the area with $p > 0$ not contained in A, the solution always converges to the desired point P_1. This is also true

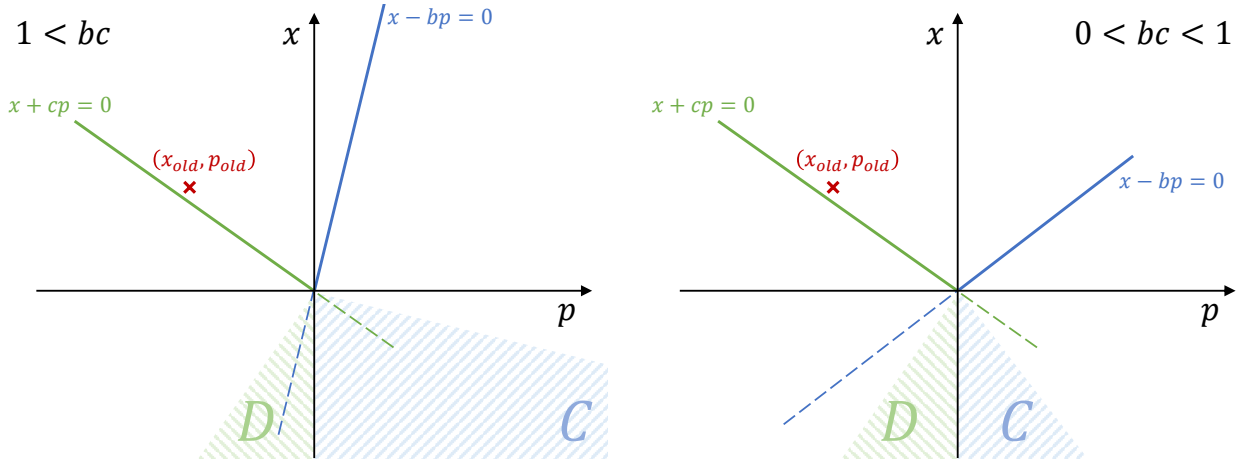

Figure 5.5: Geometry associated with pseudo-inverse corrector.

for those points in region A that are mapped to region B in the first iteration. Note that $bc > 1$ implies $bx_{old} + p_{old} > 0$ whenever (x_{old}, p_{old}) solves the equation precisely such that $x_{old} - cp_{old} = 0$. Similarly, $bc < 1$ implies $bx_{old} + p_{old} < 0$ when this is the case. Since the predictor point is chosen based on the Jacobian matrix evaluated at (x_{old}, p_{old}), it will always lie close to the dashed green line in Figure 5.3. For any reasonable solution tolerance, the value of bc is thus the factor that determines wether or not the corrector iterations converge.

Pseudo-inverse corrector without step size constraint

Inserting (5.16) into the iteration rule with the modified Newton method (5.14) yields

$$\begin{bmatrix} x_{k+1} \\ p_{k+1} \end{bmatrix} = \begin{cases} \dfrac{p_k + bx_k}{1+b^2} \begin{bmatrix} b \\ 1 \end{bmatrix} & \text{if } p_k > 0 \\ \dfrac{p_k - cx_k}{1+c^2} \begin{bmatrix} -c \\ 1 \end{bmatrix} & \text{if } p_k < 0 \end{cases} \tag{5.25}$$

Again, note that (5.18) holds.

For any (x_k, p_k) that lies in region C in Figure 5.5, defined by $p_k > 0$ and $x_k < -p_k/b$, the inequality

$$p_{k+1} = \frac{p_k + bx_k}{1+b^2} < \frac{p_k + b(-p_k/b)}{1+b^2} = 0 \tag{5.26}$$

holds. Furthermore, assuming $bc > 1$ gives

$$x_{k+1} = bp_{k+1} < \frac{p_{k+1}}{c}. \tag{5.27}$$

Similarly, for any (x_k, p_k) in region D in Figure 5.5, defined by $p_k < 0$ and $x_k < p_k/c$, the relation

$$p_{k+1} = \frac{p_k - cx_k}{1 + c^2} > \frac{p_k - c(p_k/c)}{1 + c^2} = 0 \tag{5.28}$$

holds and assuming $bc > 1$ gives

$$x_{k+1} = -cp_{k+1} < -\frac{p_{k+1}}{b}. \tag{5.29}$$

It follows that the iterates alternate between the two regions for any initial point in region C or D when $bc > 1$.

Application of (5.25) to any (x_k, p_k) outside of regions C and D in Figure 5.5 immediately results in a point that solves $F(x_{k+1}, p_{k+1}) = 0$ with $\operatorname{sgn} p_{k+1} = \operatorname{sgn} p_k$.

If $bc < 1$, then the directions of the inequalities in (5.27) and (5.29) reverse so that any point in C is mapped to the area where $p_{k+1} < 0$ that is not in D, and any point in D is mapped to the area where $p_{k+1} > 0$ that is not in C. This means that the iterations will converge from any starting point. Furthermore, the dashed green line in Figure 5.5 lies outside of region C in this case. If the iterations are initialized with a predictor point on or near the dashed green line (but outside of region C), then the iterates will converge towards a point with $p > 0$, so that the switching surface is successfully passed.

Figure 5.6 shows several iterations of (5.25) for different values of bc. When $bc > 1$, the iterations converge towards the origin (it is shown below that this is always the case), and the convergence rate depends on the values of b and c. This is a problem for two reasons: First, if convergence is slow than the algorithm will terminate after the maximum number of iterations is reached without achieving the required solution tolerance. Second, if convergence is fast, the iterations will terminate at a point close to the origin. The next predictor step is then uncertain, since it is based on the Jacobian matrix computed numerically close to

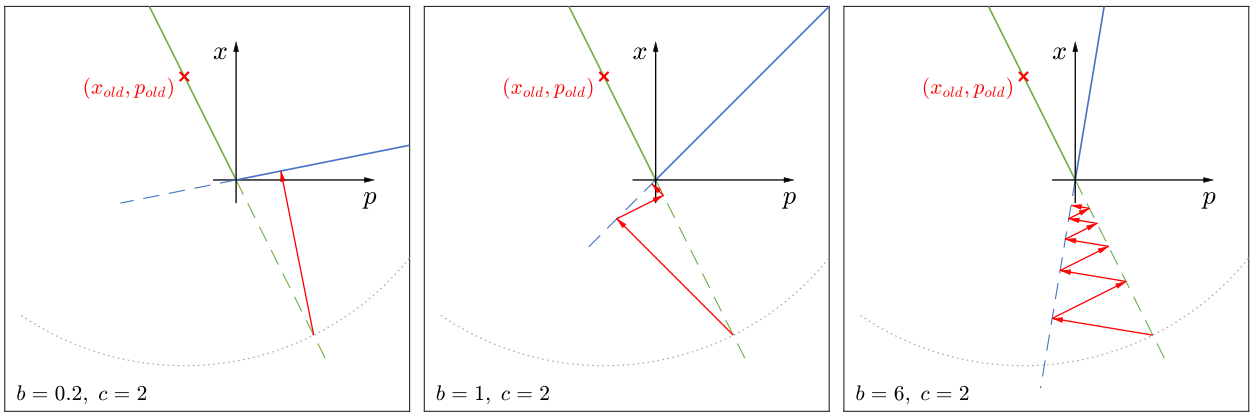

Figure 5.6: Several iterations of the pseudo-inverse corrector starting from the predictor point for different values of bc.

the (undifferentiable) origin. The corrector immediately converges to the desired solution for $bc < 1$ in Figure 5.6.

Now consider the discrete dynamics of every two iterations of the pseudo-inverse corrector with $bc > 1$. For $p_k > 0$ and $k \geq 1$, inserting (5.25) into itself leads to

$$\begin{bmatrix} x_{k+2} \\ p_{k+2} \end{bmatrix} = \underbrace{\frac{(1-bc)^2}{(1+b^2)(1+c^2)}}_{>0} \begin{bmatrix} x_k \\ p_k \end{bmatrix} = \underbrace{\frac{1+b^2c^2-2bc}{1+b^2c^2+b^2+c^2}}_{<1} \begin{bmatrix} x_k \\ p_k \end{bmatrix}, \tag{5.30}$$

implying convergence to the origin.

5.2.3 The general case

Now consider a general continuous, piecewise differentiable function $F : \mathcal{D} \subseteq \mathbb{R}^N \to \mathbb{R}^{N-1}$ and suppose $z^* \in \ker F$ lies on a single switching surface of F, so that the restriction of F to some neighborhood $\mathcal{Z}$ of z^* matches the form of the function from Theorem 5.2 in (5.6). Then F is locally defined by

$$F(z) = \begin{cases} F_b(z) & \text{if } F_{sw}(z) \geq 0 \\ F_c(z) & \text{otherwise} \end{cases} \tag{5.31}$$

for all $z \in \mathcal{Z}$ and $F(z) = F_b(z) = F_c(z)$ holds for all $z \in \ker F_{sw}$ due to continuity.

The following discussion illustrates that the conditions at z^* locally resemble the simple example discussed above if F satisfies the conditions of Theorem 5.2 at z^*, which is assumed from here on.

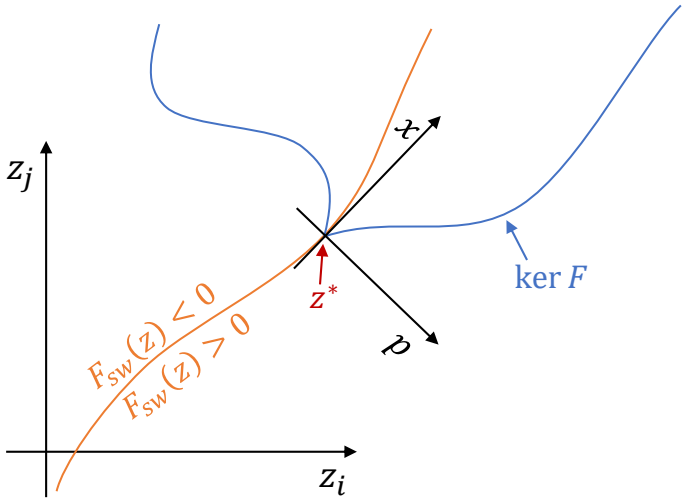

Figure 5.7: Coordinate transformation based on the direction normal to the switching surface.

By the assumptions of the theorem, the Jacobian matrices of F_b and F_c have full rank at z^*, so that these functions are locally well approximated by their linearizations about z^*. With $F_b(z^*) = F_c(z^*) = 0$, Taylor expansion of F_b and F_c leads to

$$F_b(z) = \nabla F_b(z^*)(z - z^*) + O(\|z - z^*\|^2) \tag{5.32}$$

$$F_c(z) = \nabla F_c(z^*)(z - z^*) + O(\|z - z^*\|^2) \tag{5.33}$$

With E_1 and $E_2 = \hat{e}$ (since $M = 1$) from the theorem, define $A \in \mathbb{R}^{(N-1)\times(N-1)}$ and $b, c \in \mathbb{R}^{N-1}$ as

$$A = \nabla F_b(z^*)E_1 = \nabla F_c(z^*)E_1 \tag{5.34}$$

$$b = \nabla F_b(z^*)\hat{e} \tag{5.35}$$

$$c = \nabla F_c(z^*)\hat{e} \tag{5.36}$$

as well as the coordinate transformation (see Figure 5.7)

$$x = E_1^T(z - z^*), \qquad p = \hat{e}^T(z - z^*) \tag{5.37}$$

to obtain the local first-order approximation of (5.31) defined by

$$F(z) \approx \hat{F}(E_1^T(z - z^*), \hat{e}^T(z - z^*)) \tag{5.38}$$

with

$$\hat{F}(x, p) = \begin{cases} Ax + bp & \text{if } p \geq 0 \\ Ax + cp & \text{otherwise} \end{cases} \tag{5.39}$$

For small step sizes, as would be used after non-convergence of the corrector for larger step sizes, all quadratic terms are small, so that this first-order approximation is accurate. Note that, by construction in the proof of Theorem 5.2, A is invertible.

Only the modified Newton-Raphson corrector without the step size constraint is considered in the following.

Pseudo-inverse corrector without step size constraint

Substitution of the approximation (5.38–5.39) into (5.14) leads to

$$\begin{bmatrix} x_{k+1} \\ p_{k+1} \end{bmatrix} = \begin{cases} \dfrac{p_k - b^T(A^T)^{-1}x_k}{1 + b^T(A^T)^{-1}A^{-1}b} \begin{bmatrix} -A^{-1}b \\ 1 \end{bmatrix} & \text{if } p_k > 0 \\ \dfrac{p_k - c^T(A^T)^{-1}x_k}{1 + c^T(A^T)^{-1}A^{-1}c} \begin{bmatrix} -A^{-1}c \\ 1 \end{bmatrix} & \text{if } p_k < 0 \end{cases} \tag{5.40}$$

after some algebraic manipulations. Note that

$$x_{k+1} = \begin{cases} -A^{-1}bp_{k+1} & \text{if } p_k > 0 \\ -A^{-1}cp_{k+1} & \text{if } p_k < 0 \end{cases} \tag{5.41}$$

Since both denominators in (5.40) are positive, the sign of p_{k+1} depends on the relative magnitudes of the two terms in the numerators. If $p_k > 0$ and $b^T(A^T)^{-1}x_k > p_k$ (which defines

a region similar to region C in Figure 5.5), then $p_{k+1} < 0$. Furthermore, if $-b^T(A^T)^{-1}A^{-1}c > 1$, then

$$c^T(A^T)^{-1}x_{k+1} = \underbrace{-c^T(A^T)^{-1}A^{-1}b}_{>1}\underbrace{p_{k+1}}_{<0} < p_{k+1} \tag{5.42}$$

Similarly, if $p_k < 0$ and $c^T(A^T)^{-1}x_k < p_k$ (which defines a region similar to region D in Figure 5.5), then $p_{k+1} > 0$, and if also $-b^T(A^T)^{-1}A^{-1}c > 1$ then

$$b^T(A^T)^{-1}x_{k+1} = \underbrace{-b^T(A^T)^{-1}A^{-1}c}_{>1}\underbrace{p_{k+1}}_{>0} > p_{k+1}. \tag{5.43}$$

The same sign switching behavior of p_k is thus present in the general case if the iteration is started in one of these regions. Analogous to the simple example, the iteration immediately converges from any point outside of these regions. The direction of the inequalities in (5.42) and (5.43) reverse when $-b^T(A^T)^{-1}A^{-1}c < 1$, leading to eventual convergence even from within the two regions.

Again, the success or failure of the corrector step depends on the angle between the two solution branch segments at the switching surface: Note that

$$\begin{bmatrix} -A^{-1}b \\ 1 \end{bmatrix} \in \ker([A \quad b]) \quad \text{and} \quad \begin{bmatrix} A^{-1}c \\ -1 \end{bmatrix} \in \ker([A \quad c]), \tag{5.44}$$

so that any positive scalar multiple of these vectors is a root of $\hat{F}$. The angle of interest is therefore related to the inner product of these vectors, which is $-b^T(A^T)^{-1}A^{-1}c - 1$. The angle is acute if $-b^T(A^T)^{-1}A^{-1}c > 1$ and obtuse if $-b^T(A^T)^{-1}A^{-1}c < 1$.

5.3 Mitigation Strategy

The previous section shows that non-convergence of the pseudo-arclength continuation method upon crossing a switching surface is closely related to the local angle between the incoming and outgoing branches. For acute angles, the corrector iterations may get stuck in a condition of jumping between the two sides of the switching surface, thereby preventing the continuation of the solution branch beyond the surface.

To overcome this problem, it is suggested to change the notion of geometry through use of a weighted inner product of the form

$$\langle u, v \rangle_W = u^T W v, \qquad W = W^T > 0. \tag{5.45}$$

This approach allows to stretch space in such a way that the local angle between the incoming and outgoing branches becomes large enough to prevent the undesirable behavior. The norm induced by the weighted inner product is

$$\|u\|_W = \sqrt{u^T W u}. \tag{5.46}$$

To account for the weighted inner product in the case of the classic Newton-Raphson corrector (5.12), the step size constraint (5.13) must be expressed in terms of the weighted norm instead of the 2-norm. Under the weighted inner product, the minimum-norm solution based on the induced norm leads to the weighted pseudo-inverse

$$\begin{aligned} M_W^\dagger &= W^{-1} M^T (M W^{-1} M^T)^{-1} \\ &= W^{-1/2} (M W^{-1/2})^\dagger \end{aligned} \tag{5.47}$$

which may be used in the pseudo-inverse corrector iteration rule (5.14) in lieu of the regular pseudo-inverse.

The weight matrix W can either be chosen a priori based on knowledge about the particular problem to be solved with the pseudo-arclength method, or adapted automatically by the continuation algorithm. Suppose it is defined as

$$W = I + (w - 1) u u^T \tag{5.48}$$

$$\iff \qquad W^{-1/2} = I + (1/\sqrt{w} - 1) u u^T \tag{5.49}$$

in order to stretch space by a factor of $w > 0$ in the direction $u \in \mathbb{R}^{N+1}$, where $\|u\|_2 = 1$ and I is the identity matrix. Under the weighted inner product, the relevant angle (with respect to the discussion in Section 5.2.3) is obtuse if

$$\begin{bmatrix} -b^T A^{-T} & 1 \end{bmatrix} \left(I + (w - 1) u u^T \right) \begin{bmatrix} A^{-1} c \\ -1 \end{bmatrix} < 0 \tag{5.50}$$

holds. This inequality is guaranteed to be satisfied if $u = [0 \quad \cdots \quad 0 \quad \pm 1]^T$ and $w > -b^T A^{-T} A^{-1} c$ are chosen. In the original coordinates, defined by (5.37), this corresponds to selecting $u = \pm\hat{e}$, which coincides with the gradient of the function F_{sw} that defines the switching surface.

Figure 5.8 shows several iterations of both corrector algorithms with the weighted inner product for the simple example defined in (5.16). The matrix W was chosen based on (5.48) with u chosen orthogonal to the x-axis (which marks the switching surface) and $w \in \{1, 3, 6\}$, where $w = 1$ corresponds to the unweighted case. The gray ellipses in the plots indicate points at a distance of one step size under the weighted norm.

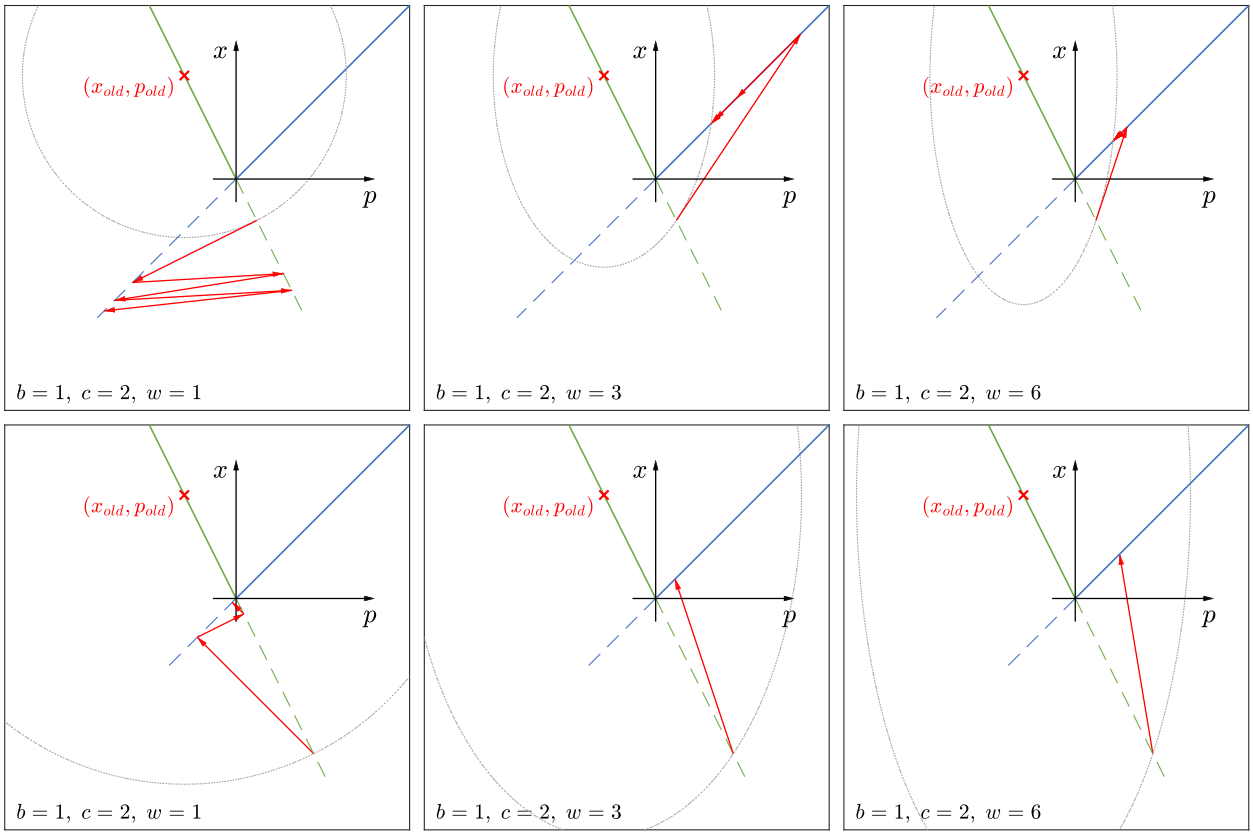

Figure 5.8: Several iterations of the Newton-Raphson corrector (top) and pseudo-inverse corrector (bottom) under the weighted inner product for different values of w.

A fixed weighting matrix is useful when the switching function is affine (or approximately affine), so that chosing a constant stretching direction u is reasonable. The pseudo-arclength algorithm may be modified as follows to enable continuation of piecewise differentiable functions if this is not the case: After computing the predictor point z' based on the last solution z_{old}, check if a sign change has occurred in the switching function, i.e., if $F_{sw}(z_{old})F_{sw}(z') < 0$. If so, choose W based on (5.48) with $u = \nabla F_{sw}(z')^T / \left\|\nabla F_{sw}(z')\right\|_2$, $w \gg 1$, and execute the corrector step with the weighted step size constraint or weighted pseudo-inverse defined in (5.47). Otherwise, use the usual corrector iteration rule.

Since multiple functions may be monitored for sign changes, the method is also applicable for problems with several switching surfaces, as long as no more than one of them is crossed during a single continuation step. It is noted that simply scaling the variables may resolve convergence issues for some problems, since this is equivalent to using a weighted inner product with a diagonal matrix W.

5.4 Example

Figure 5.9 shows continuation results for computation of the kernel of

$$F(z_1, z_2) = \left|z_1^2 + z_2^2 - 1\right| - \frac{z_1}{2} - \frac{z_2}{3} - \frac{1}{2}, \tag{5.51}$$

obtained with the strategy of using a weighting matrix based on the gradient of F_{sw} upon crossing of the switching surface as described in the previous section. The switching surface is defined by the equation $F_{sw}(z_1, z_2) = z_1^2 + z_2^2 - 1 = 0$ and plotted as a dashed line in the figure. The black points indicate the sequence of solutions computed by the pseudo-arclength method with pseudo-inverse corrector, where the point at the bottom right was provided to start the continuation process. The algorithm successfully crossed the switching surface twice despite the sharp corners in the solution branch at these points.

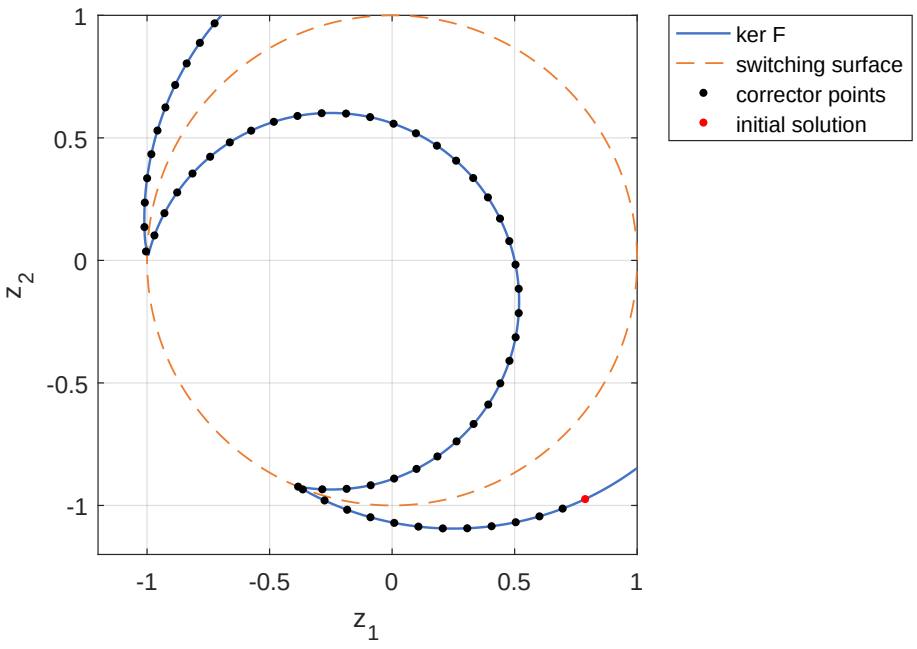

Figure 5.9: Pseudo-arclength continuation results for the example defined in (5.51) with adaptive weighting.

Chapter 6

NUMERICAL VERIFICATION OF EQUILIBRIUM LOCATION BOUNDS AND LOCAL STABILITY

This chapter presents a method to numerically verify bounds on the equilibrium location of a nonlinear system subject to bounded parameter uncertainty and assure local stability. In contrast to the material presented in the previous chapters, the method does not rely on explicit computation of the equilibrium surface with continuation algorithms. Instead, sufficient conditions which guarantee the existence of an equilibrium within a bounded subset of state space for all parameter combinations and which prove that the location of this equilibrium is a continuous function of the parameters are derived.

The conditions are verified numerically with reliable computing techniques that enable the numerical evaluation of functions and their derivatives on sets, yielding conservative results. The particular reliable computing method used in this chapter is called *affine arithmetic*. A polytopic linear system that contains the linearized system dynamics at the operating equilibrium is obtained as part of the computation, and its stability is established with standard methods. Part of the research in this chapter has been published in [55] by the author.

6.1 Sufficient Conditions for Existence of the Operating Equilibrium and Absence of Bifurcations

The goal of the numerical analysis method developed in this chapter is to verify that the operating equilibrium of the parameter-dependent system

$$\dot{x}(t) = f(x(t), p) \tag{6.1}$$

depends continuously on the parameter p, does not experience bifurcation, and is locally stable for all p in a bounded parameter set $\mathcal{P} \subset \mathbb{R}^{n_p}$. This section develops sufficient conditions on which the method is based.

6.1.1 Implicit functions on compact domains

The two versions of the Implicit Function Theorem presented in Sections 2.1 and 5.1 are local results, and the existence of a globally defined implicit function cannot be implied when the conditions hold everywhere. A well-known example demonstrating this fact is the function

$$F(x,p) = \begin{bmatrix} e^{x_1}\cos x_2 - p_1 \\ e^{x_1}\sin x_2 - p_2 \end{bmatrix} \tag{6.2}$$

which satisfies $\det\left(\nabla_x F\right) = e^{2x_1} > 0$, so that $\nabla_x F$ is invertible everywhere and $\ker F$ is locally always the graph of a function of p. However, no global function of this type exists, as the projection of $\ker F$ on the right in Figure 6.1 illustrates.

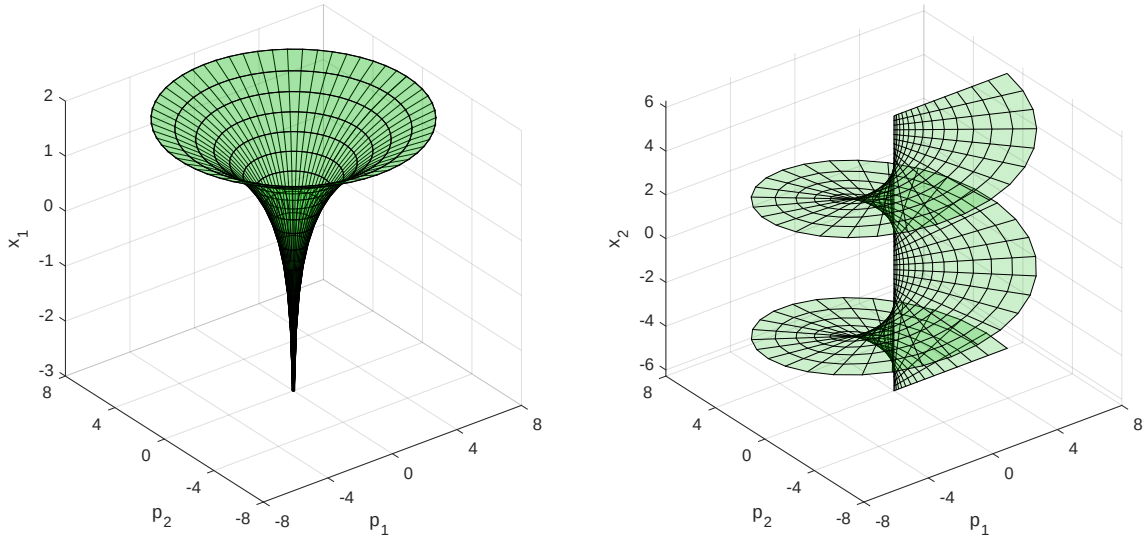

Figure 6.1: Two projections of the kernel of the function defined in (6.2).

Additional conditions must be imposed to guarantee the existence of an implicit function defined on a domain of known size. The following theorem introduces such conditions. A

set that is simply connected and has the property that there exists a Lipschitz continuous path contained within the set between any two elements of the set is called *simply Lipschitz connected* in the following. The difference to a simply connected set is that pathological cases (such as the graph of the Weierstrass function) are excluded.

Theorem 6.1. *Let $\mathcal{U}$ be a compact subset of $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and suppose its projection into $\mathbb{R}^{n_2}$, i.e., $\mathcal{W} = \pi_2(\mathcal{U})$ with $\pi_2(z_1, z_2) = z_2$, is simply Lipschitz connected. Furthermore, let $H : \mathcal{D} \subseteq \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_1}$ be locally Lipschitz on an open set containing $\mathcal{U}$ and define $\mathcal{S} = \{(z_1, z_2) \mid (z_1, z_2) \in \partial\mathcal{U},\ z_2 \notin \partial\mathcal{W}\}$. Suppose that*

1. *there exists $(z_1^*, z_2^*) \in \mathcal{U}$ such that $H(z_1^*, z_2^*) = 0$,*
2. *the image of H from $\mathcal{S}$ does not contain the origin, i.e., $0 \notin H(\mathcal{S})$, and*
3. *all matrices in the set*

$$\left\{ J_1 \in \mathbb{R}^{n_1 \times n_1} \;\middle|\; [J_1 \quad J_2] \in \widehat{\nabla} H(z_1, z_2) \quad \text{for some } J_2 \in \mathbb{R}^{n_1 \times n_2} \right\} \tag{6.3}$$

are invertible for all $(z_1, z_2) \in \mathcal{U}$

Then there exists a Lipschitz continuous function ϕ defined on $\mathcal{W}$ and taking values in $\mathbb{R}^{n_1}$ such that

$$H(\phi(z_2), z_2) = 0 \qquad \text{for all} \quad z_2 \in \mathcal{W}. \tag{6.4}$$

Furthermore, there exists a neighborhood $\widehat{\mathcal{U}}$ of the graph of ϕ such that ϕ is the only function satisfying

$$\left\{ (z_1, z_2) \in \widehat{\mathcal{U}} \;\middle|\; H(z_1, z_2) = 0 \right\} = \left\{ (z_1, z_2) \;\middle|\; z_2 \in \mathcal{W}, z_1 = \phi(z_2) \right\}. \tag{6.5}$$

Proof. First, note that since $\widehat{\nabla} H$ is upper semi-continuous and the set obtained by evaluating it at any point in its domain is compact (and thus closed), the graph of its restriction to the closed set $\mathcal{U}$ (the set $\{(z_1, z_2, J) \mid J \in \widehat{\nabla} H(z_1, z_2),\ (z_1, z_2) \in \mathcal{U}\}$) is closed (see Proposition 1.4.8 in [56]). It follows, again considering that $\widehat{\nabla} H(z_1, z_2)$ is compact (and thus bounded), that the projection of this graph into the codomain of $\widehat{\nabla} H$ (the set $\bigcup_{(z_1,z_2)\in\mathcal{U}} \widehat{\nabla} H(z_1, z_2)$) is

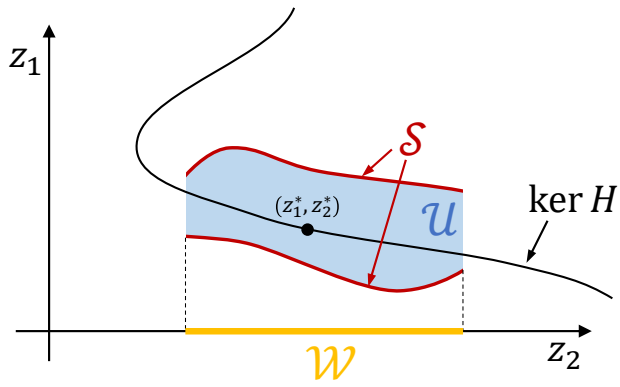

Figure 6.2: Visualization of the sets in Theorem 6.1.

compact. Therefore, with condition 3 of the theorem, there exists a positive number $\hat{\rho}$ such that

$$2\hat{\rho}^{-1} \leq \min_{[J_1\ J_2] \in \widehat{\nabla} H(z_1, z_2)} \underline{\sigma} \left(\begin{bmatrix} J_1 & J_2 \\ 0 & I \end{bmatrix} \right) \tag{6.6}$$

for all $(z_1, z_2) \in \mathcal{U}$. It follows that whenever Theorem 5.1 is applied at a point in $\mathcal{U}$, $\hat{\rho}$ is a Lipschitz constant of the implicit function in a neighborhood of that point.

Next, consider the case with $n_2 = 1$, $\mathcal{W} = [0, 1]$, and $z_2^* = 0$. The following argument is adapted from the section on homotopy in [15]. From Theorem 5.1, it follows that there exists a Lipschitz function ψ defined on an open neighborhood of 0 with $\psi(0) = z_1^*$ such that $H(\psi(z_2), z_2) = 0$ holds on that neighborhood. Therefore, there exists $Z_2 > 0$ such that the following holds:

a) ψ is defined on $[0, Z_2)$

b) $\psi(0) = z_1^*$

c) ψ is Lipschitz on $[0, Z_2)$ with Lipschitz constant $\hat{\rho}$

d) $H(\psi(z_2), z_2) = 0$ holds for all $z_2 \in [0, Z_2)$

Let $\hat{Z}_2$ be the supremum of all such numbers, i.e.,

$$\hat{Z}_2 = \sup \{\, Z_2 \mid \text{there exists } \psi \text{ satisfying the above conditions} \,\}$$

The theorem claims that $1 \in [0, \hat{Z}_2)$, so that $\hat{Z}_2 > 1$, which is proven by contradiction. Suppose that $\hat{Z}_2 \leq 1$. This implies that the graph of ψ is contained in $\mathcal{U}$ due to condition 2 of the theorem. Because ψ is Lipschitz (and thus uniformly continuous), it can be continuously extended to the closure of its domain (see Theorem 10.9.1 in [57]), i.e., the limit $\hat{Z}_1 = \lim_{z_2 \uparrow \hat{Z}_2} \psi(z_2)$ exists. Note that $(\hat{Z}_1, \hat{Z}_2) \in \mathcal{U}$ and that $H(\hat{Z}_1, \hat{Z}_2) = 0$ due to continuity of H. Now Theorem 5.1 can be applied again at $(\hat{Z}_1, \hat{Z}_2)$, thus extending ψ to a larger interval. The extended function also has Lipschitz constant $\hat{\rho}$ and therefore satisfies conditions a) through d) above. This contradicts the definition of $\hat{Z}_2$.

Last, consider the general case with $n_2 \geq 1$. Pick any point $\tilde{z}_2 \in \mathcal{W}$ and let the path $\mu : [0, 1] \to \mathcal{W}$ with $\mu(0) = z_2^*$ and $\mu(1) = \tilde{z}_2$ be Lipschitz continuous. The function g defined by $g(z_1, q) = H(z_1, \mu(q))$ satisfies the conditions of the theorem for the one-dimensional case proven above, so that a function $\psi : [0, 1] \to \mathbb{R}^{n_1}$ with $H(\psi(q), \mu(q)) = 0$ exists. Now apply Theorem 5.1 to H at every point in the set $\{(\psi(q), \mu(q)) \mid q \in [0, 1]\}$ and let $\mathcal{M}$ be the union of the resulting neighborhoods $\widetilde{\mathcal{U}}$. For any sufficiently small perturbation of the path that preserves Lipschitz continuity and the endpoints, the relation $\{(\psi_\delta(q), \mu_\delta(q)) \mid q \in [0, 1]\} \subset \mathcal{M}$ holds, where μ_δ is the perturbed path and $\psi_\delta(q)$ the corresponding perturbed function with $H(\psi_\delta(q), \mu_\delta(q)) = 0$. This implies that $\psi_\delta(1) = \psi(1)$, which means that no continuous transformation between any two Lipschitz paths from z_2^* to $\tilde{z}_2$ can change the value of the implicit function ψ at $q = 1$. Because $\mathcal{W}$ is simply Lipschitz connected, any Lipschitz path from z_2^* to $\tilde{z}_2$ can be continuously transformed into any other such path and the value of $\psi(1)$ is independent of the particular path taken. Define $\phi(\tilde{z}_2) = \psi(1)$ and note that $\tilde{z}_2$ can be picked freely to see that a function ϕ with $H(\phi(z_2), z_2) = 0$ exists. Clearly, this function ϕ is locally Lipschitz by pointwise application of Theorem 5.1 to its graph, and $\widehat{\mathcal{U}}$ is the union of the local sets $\widetilde{\mathcal{U}}$ from the theorem. Since ϕ is defined on a compact domain, it is also Lipschitz (though not necessarily with constant $\hat{\rho}$). ■

Notice in Figure 6.1 that the example function defined in (6.2) never satisfies condition 2 of the theorem with $z_1 = x$ and $z_2 = p$ when $\mathcal{W}$ includes the origin.

6.1.2 Application to parameter-dependent systems

The theorem directly leads to the desired sufficient conditions for the system defined in (6.1).

Corollary 6.2. *Let $f : \mathcal{D} \subseteq \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_x}$ and $\mathcal{U} \subset \mathcal{D}$ satisfy the conditions of Theorem 6.1. Then there exists a Lipschitz continuous function $\bar{x}$ such that $(\bar{x}(p), p)$ is an equilibrium point of the dynamical system $\dot{x}(t) = f(x(t), p)$ with $(\bar{x}(p), p) \in \mathcal{U}$ for all parameters $p \in \mathcal{P} = \pi_2(\mathcal{U}) \subset \mathbb{R}^{n_p}$. Furthermore, no other equilibrium points exist in a neighborhood of the graph of $\bar{x}$.*

Corollary 6.3. *Suppose furthermore that f is continuously differentiable and the Jacobian matrix $\nabla_x f(x, p)$ is Hurwitz for all $(x, p) \in \mathcal{U}$. Then $(\bar{x}(p), p)$ is a locally exponentially stable equilibrium for all $p \in \mathcal{P}$ and does not experience bifurcation.*

The next sections derive a way to check the assumptions of the theorem and its corollaries with affine arithmetic.

6.2 Affine Arithmetic

Reliable computing methods were originally developed to endow the results of numerical computations with known lower and upper bounds that quantify the otherwise unknown error resulting from the finite precision of digital number representation [58]. The bounded results are obtained through implementation of a data type that represents sets of numbers and redefinition of all mathematical operations for this data type such that the results are always conservative. In essence, reliable computing methods permit numerical evaluation of functions on sets such that the resulting set includes the image of the function from the input set: Given a set of inputs X, these methods compute a superset Y of the image of F from X, Figure 6.3.

The most well-known reliable computing method is called *interval analysis* [59] and represents sets as hypercubes parametrized by lower and upper bounds on each variable. While interval analysis has been successfully used in a number of system analysis applications [60–63], it is known to be highly conservative unless the hypercubes are rather small.

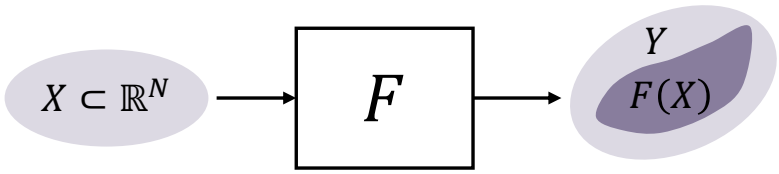

Figure 6.3: Conservative evaluation of functions with reliable computing methods.

Affine arithmetic reduces this conservatism, at the cost of a more complex representation of the sets involved in the computations. The definitions and notation used here deviate somewhat from the affine arithmetic literature [64–66], which is largely concerned with aspects of implementation. This is done to facilitate the development of the results presented in the next section. The notation is loosely based on [66].

Affine arithmetic defines objects called *affine quantities* that represent sets of elements from a vector space. For a vector space $\mathbb{V}$ over the real numbers, the set of affine quantities on $\mathbb{V}$ is defined as

$$\mathcal{A}(\mathbb{V}) = \{\, \langle \hat{\gamma}, (\gamma_1, \gamma_2, \dots) \rangle \mid \hat{\gamma}, \gamma_i \in \mathbb{V},\ i = 1, \dots, \infty \,\} \tag{6.7}$$

where $\hat{\gamma}$ is called the *midpoint* (or *central value*) and γ_i are the *error terms* (or *partial deviations*). The number of non-zero error terms is assumed to be finite. The following notation will be used from here on: if $c \in \mathcal{A}(\mathbb{V})$, then $\hat{c}$ is its midpoint and c_i is its i-th error term.

The subset of $\mathbb{V}$ represented by $c \in \mathcal{A}(\mathbb{V})$ is called the *range* of c, which is defined to be

$$\text{range}(c) = \{\, z \mid z = \hat{c} + \sum\nolimits_i c_i \epsilon_i,\ |\epsilon_i| \leq 1 \,\}, \tag{6.8}$$

where the summation index is implied to go over the indices associated with all nonzero error terms (or, equivalently, from one to infinity). The sets that can be represented with affine quantities are thus centrally symmetric convex polytopes called *zonotopes*. They are the image of the infinity-norm unit ball (of dimension equal to the number of nonzero error terms) under the affine projections defined by the midpoint and nonzero error terms of the affine quantities. An example is shown in Figure 6.4.

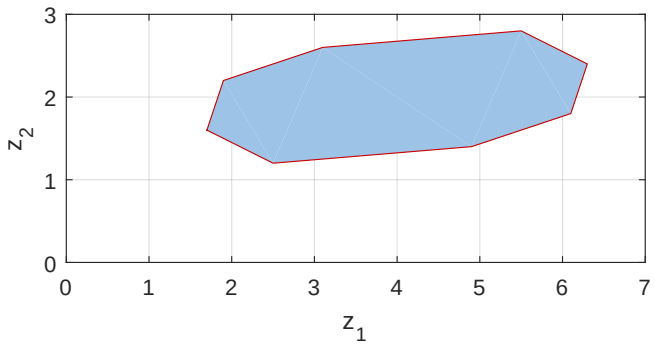

Figure 6.4: The range of the affine quantity $\langle \left[\begin{smallmatrix}4\\2\end{smallmatrix}\right], (\left[\begin{smallmatrix}1.2\\0.1\end{smallmatrix}\right], \left[\begin{smallmatrix}0.1\\0.3\end{smallmatrix}\right], \left[\begin{smallmatrix}-0.4\\0.2\end{smallmatrix}\right], \left[\begin{smallmatrix}0.6\\0.2\end{smallmatrix}\right], 0, \cdots)\rangle \in \mathcal{A}(\mathbb{R}^2)$.

Two affine quantities $c \in \mathcal{A}(\mathbb{V})$ and $d \in \mathcal{A}(\mathbb{W})$ form the *joint affine quantity*

$$(c, d) = \langle (\hat{c}, \hat{d}), ((c_1, d_1), (c_2, d_2), \dots) \rangle \in \mathcal{A}(\mathbb{V} \times \mathbb{W}) \tag{6.9}$$

with *joint range*

$$\text{range}(c, d) = \{\, (y, z) \mid (y, z) = (\hat{c}, \hat{d}) + \sum\nolimits_i (c_i, d_i)\epsilon_i,\ |\epsilon_i| \leq 1 \,\}. \tag{6.10}$$

Note that the joint range of two affine quantities is a subset of the Cartesian product of their individual ranges,

$$\text{range}(c, d) \subseteq \text{range}(c) \times \text{range}(d) \subset \mathbb{V} \times \mathbb{W}. \tag{6.11}$$

A point $z^* \in \mathbb{V}$ is in the range of $c \in \mathcal{A}(\mathbb{V})$ if the problem

$$z^* = \hat{c} + \sum\nolimits_i c_i \epsilon_i \quad \text{subject to} \quad |\epsilon_i| \leq 1 \tag{6.12}$$

has a solution. This is a linear feasibility problem for which efficient numerical solution methods exist.

A function $H : \mathcal{A}(\mathbb{V}) \to \mathcal{A}(\mathbb{W})$ is called an *affine approximation* of $h : \mathbb{V} \to \mathbb{W}$ if the following inclusion property holds for all $X \in \mathcal{A}(\mathbb{V})$ whenever h is defined on range(X):

$$\{\, (x, y) \mid y = h(x),\ x \in \text{range}(X) \,\} \subseteq \text{range}(X, H(X)) \tag{6.13}$$

The number of non-zero error terms of $H(X)$ will typically be larger than that of X, because the image of h from a zonotope is in general not a zonotope (unless h is an affine operation)

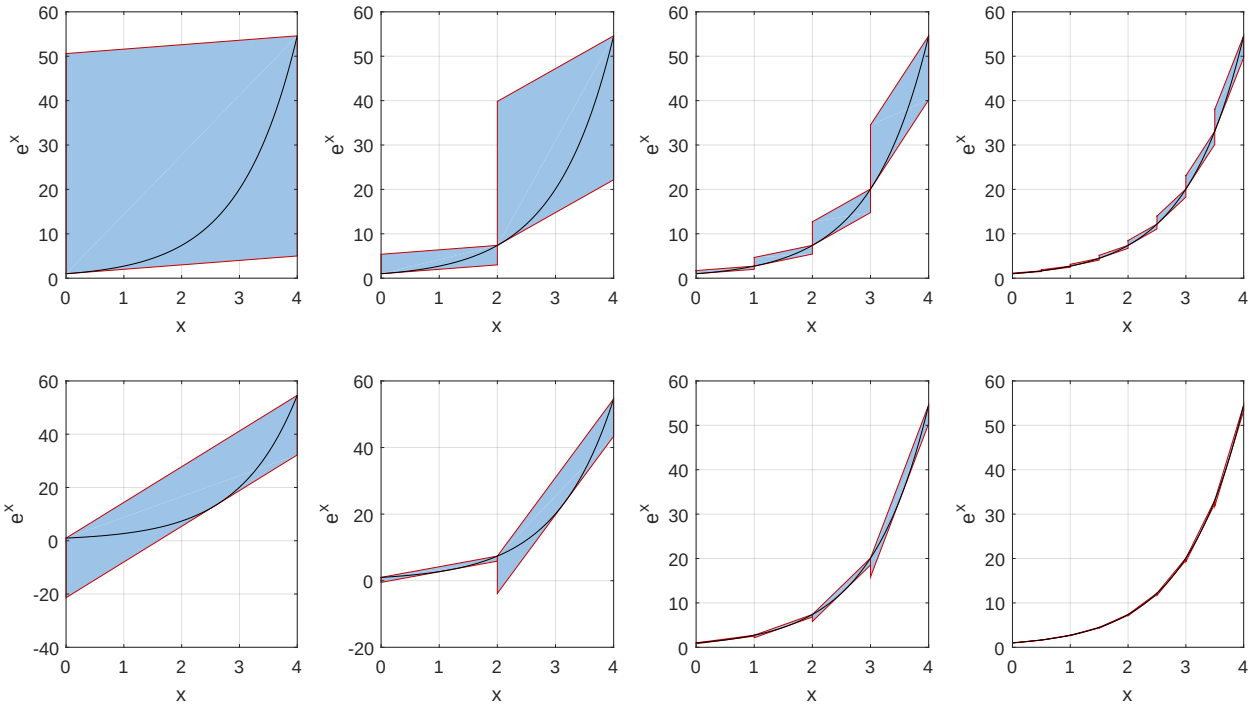

Figure 6.5: Two different affine approximations of the exponential function (top and bottom) evaluated on affine quantities representing input sets of decreasing size (left to right).

and therefore cannot be represented precisely by an affine quantity. This is the source of conservatism of the method. Figure 6.5 shows two affine approximations of the exponential function evaluated on affine quantities representing various subsets of the interval $[0, 4]$. Note that the graph of the exponential function is contained in the blue zonotopes that correspond to range$(X, H(X))$ in (6.13).

If h is the composition of many primitive operations, such as addition, multiplication, and basic nonlinear functions with well-known properties (sin, cos, exp, log, sqrt, ...), an affine approximation of h is obtained by defining affine approximations for each of the primitives. The composition then automatically[1] satisfies (6.13).

Since evaluating an affine approximation always yields an affine quantity representing a superset of the graph of the approximated function, an extension to allow for the set-valued

[1]Some care concerning the management of error term indices must be taken for this to be true.

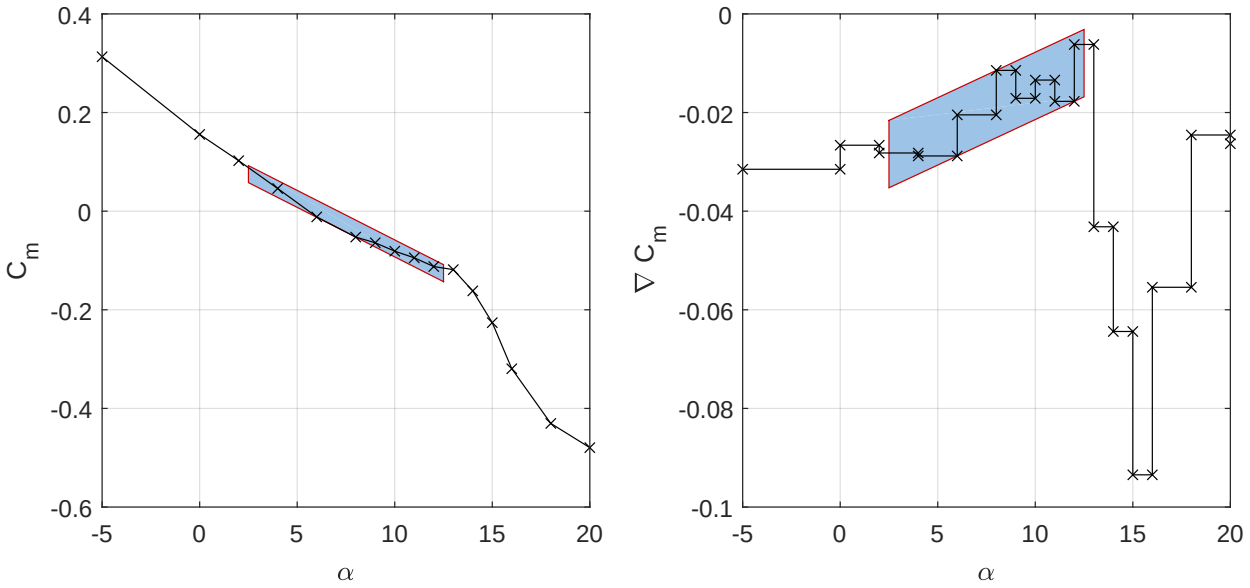

Figure 6.6: Affine approximation of linearly interpolated table data and its generalized Jacobian.

generalized Jacobian from (5.1) may be defined as follows. A function $H : \mathcal{A}(\mathbb{V}) \to \mathcal{A}(\mathbb{W})$ is an affine approximation of the set-valued function h mapping from $\mathbb{V}$ to the space of subsets of $\mathbb{W}$ if the following inclusion property holds for all $X \in \mathcal{A}(\mathbb{V})$:

$$\{\, (x, y) \mid y \in h(x),\ x \in \text{range}(X) \,\} \subseteq \text{range}(X, H(X)) \tag{6.14}$$

An affine approximation of a function defined by linearly interpolated data points and of its generalized Jacobian is shown in Figure 6.6.

No discussion of the details of deriving affine approximations for all common primitive operations is offered here. Suitable methods are the subject of most of the literature concerned with affine arithmetic [64–66]. It is noted, however, that affine arithmetic codes are typically implemented through operator overloading in object oriented programming languages. If h is defined as a subroutine in the programming environment, an affine approximation of h is evaluated by passing an object that represents the affine quantity to this subroutine. The technique can be combined with forward-mode automatic differentiation [58,67], so that

affine approximations of function derivatives can be evaluated without specifying an analytic expression.

Similar to other reliable computing methods, affine arithmetic may be implemented such that results have proof-like quality in the sense that all numerical errors incurred throughout the computation are properly accounted for. This requires appropriate control of the rounding mode used by the CPU on the workstation which executes the code.

6.3 Verifying the Sufficient Conditions with Affine Arithmetic

To numerically check the conditions of Theorem 6.1 with affine arithmetic, all sets that occur in the theorem (or supersets thereof) must be represented by affine quantities. The following two lemmas establish two results needed for this purpose.

Lemma 6.4. *Suppose that the set of nonzero error terms of $c \in \mathcal{A}(\mathbb{V})$ is linearly independent and spans $\mathbb{V}$. Let I be the index set of all nonzero error terms of c (i.e., $c_i \neq 0$ if $i \in I$ and $c_i = 0$ if $i \notin I$). Then the interior of range(c) is not empty, and the boundary and interior of range(c) are*

$$\partial \mathit{range}(c) = \Big\{ z \mid z = \hat{c} + \sum\nolimits_i c_i \epsilon_i, |\epsilon_i| \leq 1,\ |\epsilon_k| = 1 \text{ for some } k \in I \Big\}, \tag{6.15}$$

$$\mathit{range}(c) \setminus \partial \mathit{range}(c) = \{\, z \mid z = \hat{c} + \sum\nolimits_i c_i \epsilon_i,\ |\epsilon_i| < 1 \,\}. \tag{6.16}$$

Proof. The interior of range(c) is not empty because the nonzero error terms span $\mathbb{V}$. The affine operation that maps the infinity-norm unit ball to range(c) defined in (6.8) is invertible since the nonzero error terms of c also form a linearly independent set. Every point on the boundary (in the interior) of the infinity-norm unit ball is thus mapped to the boundary (to the interior) of range(c). ■

The conditions of Lemma 6.4 ensure that range(c) is not embedded in a (translated) lower-dimensional subspace of $\mathbb{V}$ (which would be the case if the nonzero error terms did not span $\mathbb{V}$) and that all boundary points of the infinity-norm unit ball are mapped to the

boundary of range(c) (which would not be the case if the set of nonzero error terms was not linearly independent).

Lemma 6.5. *Let $c \in \mathcal{A}(\mathbb{V})$ and $d \in \mathcal{A}(\mathbb{W})$ be such that both d and the joint affine quantity (c, d) satisfy the conditions of Lemma 6.4. Let I_c and I_d be the index sets of all nonzero error terms of c and d, respectively, and define $I = I_c \setminus I_d$. Furthermore, define*

$$\mathcal{S} = \{(y, z) \mid (y, z) \in \partial range(c, d)\,,\ z \notin \partial range(d)\} \tag{6.17}$$

and $c^{(k+)}, c^{(k-)} \in \mathcal{A}(\mathbb{V})$ with midpoints $\hat{c}^{(k+)} = \hat{c} + c_k$, $\hat{c}^{(k-)} = \hat{c} - c_k$, and error terms $c_i^{(k+)} = c_i^{(k-)} = c_i$ for $i \neq k$ and $c_k^{(k+)} = c_k^{(k-)} = 0$. Then the following relation holds:

$$\mathcal{S} \subset \bigcup_{k \in I} \Big(range\big(c^{(k+)}, d\big) \cup range\big(c^{(k-)}, d\big) \Big). \tag{6.18}$$

Proof. The set $\mathcal{S}$ contains all points on the boundary of range(c, d) whose second component is in the interior of range(d). Considering Lemma 6.4, $\mathcal{S}$ can therefore be expressed as follows:

$$\begin{aligned}
\mathcal{S} &= \Big\{(y, z) \mid (y, z) = (\hat{c}, \hat{d}) + \textstyle\sum_i (c_i, d_i)\epsilon_i,\ \ |\epsilon_i| \leq 1, \\
&\qquad\qquad |\epsilon_k| = 1 \text{ for some } k \in I_c \cup I_d, z = \hat{d} + \textstyle\sum_j d_j \gamma_j,\ |\gamma_j| < 1\Big\} \\
&= \Big\{(y, z) \mid (y, z) = (\hat{c}, \hat{d}) + \textstyle\sum_i (c_i, d_i)\epsilon_i,\ \ |\epsilon_i| \leq 1, \\
&\qquad\qquad |\epsilon_j| < 1 \text{ for all } j \in I_d,\ |\epsilon_k| = 1 \text{ for some } k \in I\Big\} \\
&\subset \Big\{(y, z) \mid (y, z) = (\hat{c}, \hat{d}) + \textstyle\sum_i (c_i, d_i)\epsilon_i,\ \ |\epsilon_i| \leq 1,\ |\epsilon_k| = 1 \text{ for some } k \in I\Big\} \\
&= \bigcup_{k \in I} \Big(\Big\{(y, z) \mid (y, z) = (\hat{c} + c_k, \hat{d}) + \textstyle\sum_{i \neq k} (c_i, d_i)\epsilon_i,\ \ |\epsilon_i| \leq 1\Big\} \\
&\qquad\qquad \cup \Big\{(y, z) \mid (y, z) = (\hat{c} - c_k, \hat{d}) + \textstyle\sum_{i \neq k} (c_i, d_i)\epsilon_i,\ \ |\epsilon_i| \leq 1\Big\}\Big) \\
&= \bigcup_{k \in I} \Big(\text{range}\big(c^{(k+)}, d\big) \cup \text{range}\big(c^{(k-)}, d\big) \Big)
\end{aligned}$$

■

Based on the two lemmas above, the conditions of Theorem 6.1 are now reformulated in terms of affine quantities and affine function approximations.

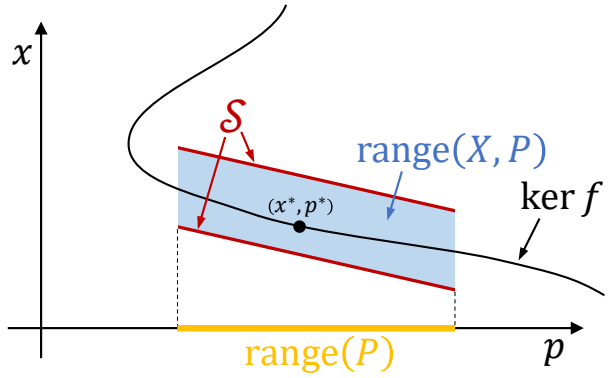

Figure 6.7: Visualization of the sets in Theorem 6.6.

Theorem 6.6. *Let $X \in \mathcal{A}(\mathbb{R}^{n_x})$ and $P \in \mathcal{A}(\mathbb{R}^{n_p})$ be such that the nonzero error terms of P form a basis for $\mathbb{R}^{n_p}$ and the nonzero error terms of the joint affine quantity (X, P) form a basis for $\mathbb{R}^{n_x} \times \mathbb{R}^{n_p}$. Let I_X and I_P be the index sets of all nonzero error terms of X and P, respectively, and define $I = I_X \setminus I_P$. Furthermore, define $X^{(k+)}, X^{(k-)} \in \mathcal{A}(\mathbb{R}^{n_x})$ with $\hat{X}^{(k+)} = \hat{X} + X_k$, $\hat{X}^{(k-)} = \hat{X} - X_k$, and $X_i^{(k+)} = X_i^{(k-)} = X_i$ for $i \neq k$ and $X_k^{(k+)} = X_k^{(k-)} = 0$.*

Assume f is locally Lipschitz. Let $Y^{(k+)}$ and $Y^{(k-)}$ be the results of evaluating an affine function approximation of f on $(X^{(k+)}, P)$ and $(X^{(k-)}, P)$, respectively, and let J be the result of evaluating an affine function approximation of the generalized Jacobian $\widehat{\nabla} f$ on (X, P). Suppose that

1. *there exists (x^*, p^*) in the interior of range(X, P) such that $f(x^*, p^*) = 0$,*
2. *the origin is not a member of range$\left(Y^{(k+)}\right)$ or range$\left(Y^{(k-)}\right)$ for any $k \in I$, and*
3. *all matrices in the set*

$$\left\{ J_x \in \mathbb{R}^{n_x \times n_x} \;\middle|\; \begin{bmatrix} J_x & J_p \end{bmatrix} \in \text{range}(J) \quad \text{for some } J_p \in \mathbb{R}^{n_x \times n_p} \right\} \tag{6.19}$$

are invertible.

Then f and range(X, P) satisfy the conditions of Theorem 6.1.

Proof. Let $\mathcal{P}$ and $\mathcal{U}$ be the interiors of range(P) and range(X, P), respectively. Both sets are compact and convex, since they are the image of the closed infinity-norm unit ball under an affine transformation. This implies simple Lipschitz connectedness of $\mathcal{P}$. Note from the definition of the joint range in (6.10) that range$(P) = \pi_p(\text{range}(X, P))$, so that $\mathcal{P} = \pi_p(\mathcal{U})$ holds (where $\pi_p(x, p) = p$). Condition 1 asserts the existence of a point in $\mathcal{U}$ where f is zero. From Lemma 6.5 and condition 2, it follows that $0 \notin f(\{(x, p) \mid (x, p) \in \partial\mathcal{U},\ p \notin \partial\mathcal{P}\})$. By the definition of affine approximations for set-valued functions in (6.14), range(J) is a superset of $\bigcup_{(x,p)\in\text{range}(X,P)} \widehat{\nabla} f(x, p)$, so that condition 3 implies the last condition of Theorem 6.1. ■

Proposition 6.7. *Suppose f is continuously differentiable and let A be the result of evaluating an affine function approximation of the Jacobian $\nabla_x f$ on (X, P). Then the set defined in* (6.19) *is equal to range(A).*

Proof. When f is continuously differentiable, then $\widehat{\nabla} f(x, p) = \left\{ \begin{bmatrix} \nabla_x f(x, p) & \nabla_p f(x, p) \end{bmatrix} \right\}$ holds everywhere. ■

Note that (6.19) defines a polytope of matrices. Standard methods based on linear matrix inequalities [68, 69] may therefore be used to numerically establish that all matrices in this polytope are Hurwitz, which is sufficient for condition 3 of Theorem 6.6. Condition 2 can be checked by solving $2n_x$ linear feasibility problems, for instance by using a standard solver for linear programs. The first condition is most easily satisfied by construction of X and P based on a known equilibrium point of the system (see below). If all conditions are verified with these methods, then the conclusion of Corollary 6.2 and Corollary 6.3 (if f is continuously differentiable) immediately follows.

6.3.1 Specifying the affine quantities

The question remains on how the affine quantities X and P in Theorem 6.6 should be chosen. Suppose that independent lower and upper bounds are specified for all parameters as $\underline{p}, \bar{p} \in \mathbb{R}^{n_p}$, and that the nominal parameter vector is $p^* = \frac{1}{2}(\underline{p} + \bar{p})$. Choosing $P \in \mathcal{A}(\mathbb{R}^{n_p})$

with $\hat{P} = p^*$ and P_i nonzero for $n_x < i \leq n_x + n_p$ and zero otherwise such that the nonzero error terms form the columns of a diagonal matrix, i.e., $[P_{n_x+1} \ \cdots \ P_{n_x+n_p}] = \text{diag}\big(\frac{1}{2}(\bar{p} - \underline{p})\big)$, results in an affine quantity that satisfies the conditions of Lemma 6.4 and whose range represents the hypercube of possible parameter combinations.

Similarly, $X \in \mathcal{A}(\mathbb{R}^{n_x})$ should be defined such that $\hat{X} = x^*$ (the nominal equilibrium location) and with X_i nonzero for $1 \leq i \leq n_x + n_p$ and zero otherwise. If the first n_x error terms are linearly independent, then (X, P) satisfies the conditions of Lemma 6.4. A suitable choice is the following: Let the nonzero error terms form the columns of the matrix $[X_1 \ \cdots \ X_{n_x+n_p}] = [\text{diag}(\Delta x) \qquad \nabla_p \bar{x}(p^*)\text{diag}\big(\frac{1}{2}(\bar{p} - \underline{p})\big)]$, where $\Delta x \in \mathbb{R}^{n_x}$ and $\nabla_p \bar{x}(p^*) = -(\nabla_x f(x^*, p^*))^{-1} \ \nabla_p f(x^*, p^*)$.

With X and P specified this way, range(X, P) can be understood as the Minkowski sum of the hyperplane tangential to the graph of $\bar{x}$ at (x^*, p^*), cropped at the parameter bounds, and a n_x-dimensional hypercube whose side lengths are specified by the components of Δx. To obtain the least conservative results, the components of Δx should be chosen as small as possible but large enough for condition 2 in Theorem 6.6 to hold.

6.4 Example

The well-understood glycolytic oscillator dynamics [70] are analyzed in this section to illustrate the method. The state equation is

$$\dot{x} = \begin{bmatrix} -x_1 + p_1 x_2 + x_1^2 x_2 \\ p_2 - p_1 x_2 - x_1^2 x_2 \end{bmatrix} \tag{6.20}$$

For this system, an analytical solution of the (unique) equilibrium in terms of the parameters is readily obtained and will be used for comparison:

$$\bar{x}(p) = \begin{bmatrix} p_2 \\ p_2(p_1 + p_2^2)^{-1} \end{bmatrix} \tag{6.21}$$

Depending on the parameter values, the equilibrium is either locally stable or unstable. The stability boundary is known to be defined by $p_2^2 = \frac{1}{2}\left(1 - 2p_1 \pm \sqrt{1 - 8p_1}\right)$. It is plotted in Figure 6.8.

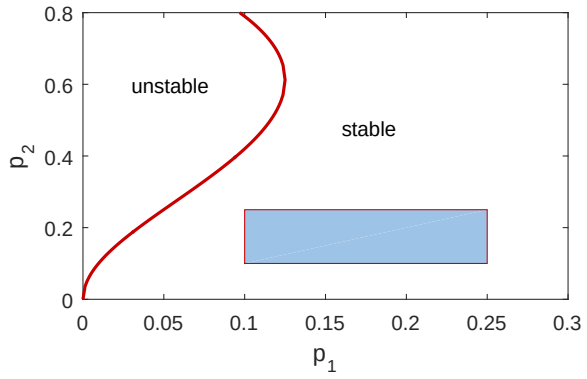

Figure 6.8: Stable and unstable parameter regions and range(P) (rectangle).

The analysis method derived in this chapter is applied to the system in order to show that the nominal equilibrium which exists for $p_1 = p_2 = 0.175$ at $x_1 = 0.175$, $x_2 = 0.851$ depends continuously on the parameters, does not bifurcate, and is stable for all $p_1 \in [0.1, 0.25]$, $p_2 \in [0.1, 0.25]$. For this, the affine quantities P and X are defined as discussed in the previous section:

$$P = \langle \begin{bmatrix} 0.175 \\ 0.175 \end{bmatrix}, \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.075 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0.075 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \dots \right) \rangle \tag{6.22}$$

$$X = \langle \begin{bmatrix} 0.175 \\ 0.851 \end{bmatrix}, \left(\begin{bmatrix} 0.02 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0.6 \end{bmatrix}, \begin{bmatrix} 0 \\ -0.31 \end{bmatrix}, \begin{bmatrix} 0.075 \\ 0.256 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \dots \right) \rangle \tag{6.23}$$

Figure 6.9 shows two projections of range(X, P) and the graph of the analytical expression for the equilibrium location given in (6.21).

All affine function approximations are evaluated with an affine arithmetic and automatic differentiation code implemented by the author. Condition 2 of the theorem is shown to be satisfied by demonstrating that (6.12) does not have a solution (with $z^* = 0$) for $c \in \{Y^{(k+)}, Y^{(k-)}\}$ and $k \in I$ using a linear programming solver. Indeed, the components of Δx that define X were chosen by starting with rather loose bounds and reducing the values until further reduction prevented verification of this condition. All matrices in the set defined by (6.19) (obtained by evaluating an affine approximation of $\nabla_x f$ on (X, P) to obtain $A \in \mathcal{A}(\mathbb{R}^{2\times 2})$, see Proposition 6.7) are shown to be Hurwitz with Matlab's Robust

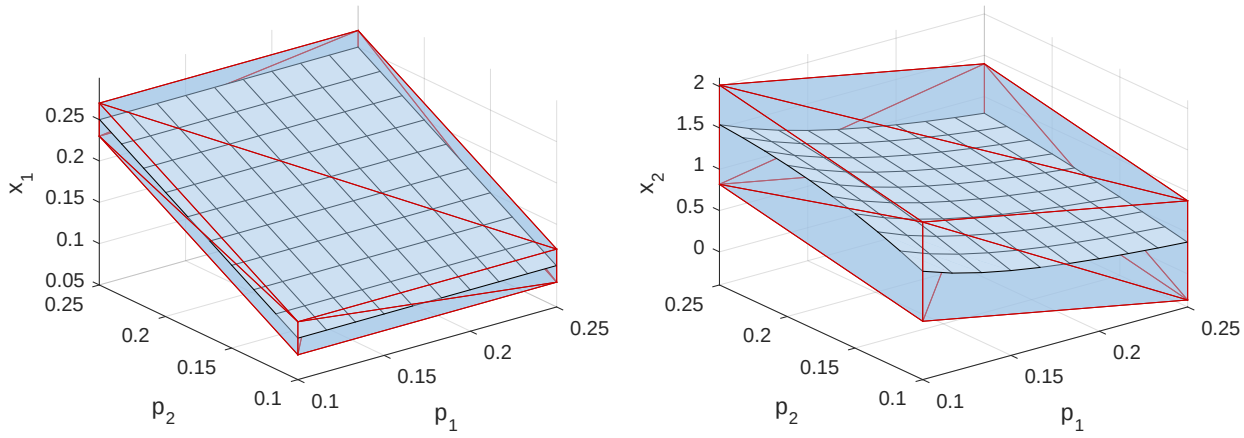

Figure 6.9: Two projections of the graph of $\bar{x}(p)$ (black wireframe mesh) and range(X, P) (surrounding blue box).

Control Toolbox. The conditions of Theorem 6.6 are therefore satisfied and the conclusion of Corollary 6.3 follows.

To validate the results, the eigenvalues of elements in range(A) are compared to eigenvalues of an analytic expression for the linearized dynamics at the equilibrium. Figure 6.10 plots the eigenvalues of 5064 matrices in range(A) obtained by applying the affine projection defined in (6.8) to the 64 vertices and to 5000 uniform samples of the infinity-norm unit ball (blue) and eigenvalues of the Jacobian matrix evaluated at 100 grid points of the analytic solution (black). The conservatism of the method is clearly visible.

If a different parameter region is chosen such that it overlaps with the unstable region in Figure 6.8 or is very close to it, the LMI-based stability test cannot establish that all matrices in range(A) are Hurwitz. In this case, the analysis results are inconclusive since they neither prove nor disprove stability and existence of the operating equilibrium.

It is emphasized that the analytical solution of the equilibrium location given in (6.21) is not required to apply the method. An example for which an analytical solution is available was chosen so that the bounding set range(X, P) and the actual equilibrium location can be plotted in the same figure for comparison, see Figure 6.9. If the nominal equilibrium (x^*, p^*)

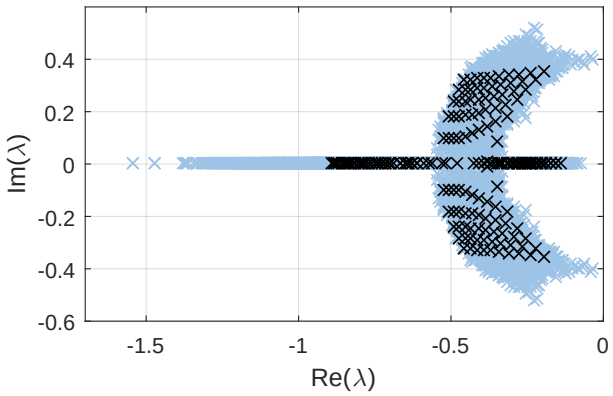

Figure 6.10: Eigenvalue comparison.

is obtained with numerical methods (so that $\|f(x^*, p^*)\| \leq \mu$ for some small solution tolerance $\mu > 0$), Theorem 6.6 guarantees the existence of a function $\tilde{x}(p)$ such that $f\,\|(\tilde{x}(p), p)\| \leq \mu$. To see this, apply the theorem to the function $g(x, p) = f(x, p) - f(x^*, p^*)$. The derivatives of f at (x^*, p^*) that are used to define X can be approximated with finite differences, since accuracy is not required.

Chapter 7
CONCLUSION

The research in this dissertation has introduced several new tools for the analysis of parameter-dependent nonlinear systems. First, the methods of bifurcation analysis have been extended to include equality constraints on states and parameters, so that multi-parameter systems can be analyzed by enforcing the parameter combinations that satisfy the constraints. Continuation problems were also derived for the computation of features that exist for these parameter combinations, but violate the constraints. The approach was then extended to bifurcation analysis of the zero dynamics, which comprise the remaining degrees of freedom in the dynamics when a system output is constrained to zero by manipulation of the input. Next, a framework for the computation of equilibrium point subsets satisfying constraints on the eigenvalues and eigenvectors of the linearized dynamics was derived. It enables the systematic study of local dynamical properties over the operating envelope, since the dynamics of nonlinear systems operated around an equilibrium point are locally characterized by properties of the linearized system. Continuation with the pseudo-arclength algorithm requires continuous differentiability of the problem equations. This assumption is frequently violated by models of aerospace systems. For piecewise differentiable problems, the convergence issues that occur have been thoroughly analyzed, and a modification to the algorithm was suggested to mitigate the problem. Last, a novel method to numerically verify the existence of an operating equilibrium and establish the absence of bifurcations for systems with parameter uncertainty was introduced. Instead of explicitly resolving the equilibrium location with continuation algorithms, it is based on the verification of sufficient conditions with reliable computing techniques.

BIBLIOGRAPHY

[1] Christopher Fielding, Andras Varga, Samir Bennani, and Michiel Selier. *Advanced techniques for clearance of flight control laws*, volume 283. Springer, 2002.

[2] Adam Steltzner, Devin Kipp, Allen Chen, Dan Burkhart, Carl Guernsey, Gavin Mendeck, Robert Mitcheltree, Richard Powell, Tommaso Rivellini, Miguel San Martin, et al. Mars science laboratory entry, descent, and landing system. In *IEEE Aerospace Conference*, pages 15 pp.–. IEEE, 2006.

[3] Ethan Baumann, Catherine Bahm, Brian Strovers, Roger Beck, and Michael Richard. The X-43A six degree of freedom monte carlo analysis. In *46th AIAA Aerospace Sciences Meeting and Exhibit*, 2008. AIAA Paper 2008-203.

[4] Yuri Aleksandrovich Kuznetsov. *Elements of applied bifurcation theory*, volume 112. Springer, 3 edition, 2004.

[5] Rüdiger Seydel. *Practical bifurcation and stability analysis*. Springer, 2010.

[6] John Guckenheimer and Philip J. Holmes. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, volume 42. Springer, 2013.

[7] Eugene L. Allgower and Kurt Georg. *Introduction to Numerical Continuation Methods*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2003.

[8] MG Goman, GI Zagainov, and AV Khramtsovsky. Application of bifurcation methods to nonlinear flight dynamics problems. *Progress in Aerospace Sciences*, 33(9):539–586, 1997.

[9] Brad S. Liebst and Robert C. Nolan. Method for the prediction of the onset of wing rock. *Journal of Aircraft*, 31(6):1419–1421, 1994.

[10] Mark H. Lowenberg. Bifurcation analysis of multiple-attractor flight dynamics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, pages 2297–2320, 1998.

[11] F.B.J. Macmillen and J.M.T. Thompson. Bifurcation analysis in the flight dynamics design process? a view from the aircraft industry. *Philosophical Transactions of the Royal Society of London*, 356(1745):2321–2333, 1998.

[12] Harry G. Kwatny, Jean-Etienne T. Dongmo, Bor-Chin Chang, Gaurav Bajpai, Murat Yasar, and Christine Belcastro. Nonlinear analysis of aircraft loss of control. *Journal of Guidance, Control, and Dynamics*, 36(1):149–162, 2012.

[13] Sanjiv Sharma, Etienne B. Coetzee, Mark H. Lowenberg, Simon A. Neild, and Bernd Krauskopf. Numerical continuation and bifurcation analysis in aircraft design: an industrial perspective. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 373(2051), 2015.

[14] Stephen J. Gill, Mark H. Lowenberg, Simon A. Neild, Luis G. Crespo, Bernd Krauskopf, and Guilhem Puyou. Nonlinear dynamics of aircraft controller characteristics outside the standard flight envelope. *Journal of Guidance, Control, and Dynamics*, 38(12):2301–2308, 2015.

[15] Steven G. Krantz and Harold R. Parks. *The implicit function theorem: history, theory, and applications*. Springer, 2012.

[16] Michael E. Henderson. Multiple parameter continuation: Computing implicitly defined k-manifolds. *International Journal of Bifurcation and Chaos*, 12(03):451–476, 2002.

[17] Herbert B. Keller. Numerical solution of bifurcation and nonlinear eigenvalue problems. *Applications of Bifurcation Theory*, pages 359–384, 1977.

[18] E. Riks. An incremental approach to the solution of snapping and buckling problems. *International Journal of Solids and Structures*, 15(7):529–551, 1979.

[19] Carl De Boor and Blair Swartz. Collocation at gaussian points. *SIAM Journal on Numerical Analysis*, 10(4):582–606, 1973.

[20] Thomas R. Lucas and George W. Reddien, Jr. Some collocation methods for nonlinear boundary value problems. *SIAM Journal on Numerical Analysis*, 9(2):341–356, 1972.

[21] R. D. Russell and Lawerence F. Shampine. A collocation method for boundary value problems. *Numerische Mathematik*, 19(1):1–28, 1972.

[22] K. A. Wittenbrink. High order projection methods of moment-and collocation-type for nonlinear boundary value problems. *Computing*, 11(3):255–274, 1973.

[23] Peter Kunkel and Volker Mehrmann. *Differential-algebraic equations: Analysis and numerical solution*. European Mathematical Society, 2006.

[24] Remco Leine and Henk Nijmeijer. *Dynamics and bifurcations of non-smooth mechanical systems*, volume 18. Springer, 2013.

[25] Eusebius J. Doedel and Bart E. Oldeman. AUTO-07P: Continuation and bifurcation software for ordinary differential equations, 2012. http://cmvl.cs.concordia.ca/auto.

[26] Miklós Farkas. *Periodic motions*, volume 104. Springer, 2013.

[27] Max G. Spetzler and Anshu Narang-Siddarth. Continuation analysis of nonlinear systems with equality constraints on states, parameters, and eigenvalues. In *AIAA Guidance, Navigation and Control Conference*, 2015. AIAA Paper 2015-1320.

[28] Narayan Ananthkrishnan and Nandan Sinha. Level flight trim and stability analysis using extended bifurcation and continuation procedure. *Journal of Guidance, Control, and Dynamics*, 24(6):1225–1228, 2001.

[29] Nandan Sinha and Narayan Ananthkrishnan. Use of the extended bifurcation analysis method for flight control law design. In *40th AIAA Aerospace Sciences Meeting & Exhibit*, page 249, 2002.

[30] Aditya Paranjape, Nandan Sinha, and Narayan Ananthkrishnan. Use of bifurcation and continuation methods for aircraft trim and stability analysis – a state-of-the-art. In *45th AIAA Aerospace Sciences Meeting & Exhibit*, page 1051, 2007.

[31] Michael A. Henson and Dale E. Seborg. *Nonlinear process control*. Prentice Hall, 1997.

[32] Alberto Isidori. *Nonlinear control systems*. Springer, 1995.

[33] Alan J. Laub and B.C. Moore. Calculation of transmission zeros using QZ techniques. *Automatica*, 14(6):557–566, 1978.

[34] A.G.J. MacFarlane and N. Karcanias. Poles and zeros of linear multivariable systems: a survey of the algebraic, geometric and complex-variable theory. *International Journal of Control*, 24(1):33–74, 1976.

[35] Max G. Spetzler and Anshu Narang-Siddarth. Local linear controllability and observability analysis of nonlinear systems with continuation methods. In *AIAA Guidance, Navigation and Control Conference*, 2016. AIAA Paper 2016-0080.

[36] Max G. Spetzler and Anshu Narang-Siddarth. Increased functionality of continuation-based nonlinear system analysis. *Journal of Guidance, Control, and Dynamics*, 39(6):1206–1222, 2016.

[37] Robert Hermann and Arthur J. Krener. Nonlinear controllability and observability. *Automatic Control, IEEE Transactions on*, 22(5):728–740, 1977.

[38] Lawrence Markus and Ernest Bruce Lee. On the existence of optimal controls. *Journal of Basic Engineering*, 84(1):13–20, 1962.

[39] Rudolf E. Kalman. Discussion to [38]. Ibid, pp. 21-22.

[40] Eduardo D. Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer, 1998.

[41] Thomas Kailath. *Linear systems*, volume 1. Prentice-Hall, 1980.

[42] Rikus Eising. Between controllable and uncontrollable. *Systems & Control Letters*, 4(5):263–264, 1984.

[43] A.M.A. Hamdan and A.H. Nayfeh. Measures of modal controllability and observability for first-and second-order linear systems. *Journal of Guidance, Control, and Dynamics*, 12(3):421–428, 1989.

[44] M. Tarokh. Measures for controllability, observability and fixed modes. *Automatic Control, IEEE Transactions on*, 37(8):1268–1273, 1992.

[45] Ernest Bruce Lee and Lawrence Markus. *Foundations of optimal control theory*. Wiley, 1967.

[46] Brian L. Stevens and Frank L. Lewis. *Aircraft control and simulation*, volume 2. Wiley New York, 2003.

[47] Jared A. Grauer and Eugene A. Morelli. A generic nonlinear aerodynamic model for aircraft. In *AIAA Atmospheric Flight Mechanics Conference*, 2014. AIAA Paper 2014-0542.

[48] Thomas Richardson, Mark Lowenberg, Mario DiBernardo, and Guy Charles. Design of a gain-scheduled flight control system using bifurcation analysis. *Journal of Guidance, Control, and Dynamics*, 29(2):444–453, 2006.

[49] Max G. Spetzler and Anshu Narang-Siddarth. Pseudo-arclength method convergence for continuous, piecewise-differentiable systems. In *American Control Conference (ACC)*. IEEE, 2017.

[50] Francis H. Clarke. *Optimization and Nonsmooth Analysis*, chapter 7, pages 252–283. Society for Industrial and Applied Mathematics, 1990.

[51] Francis H. Clarke, Yuri S. Ledyaev, Ronald J. Stern, and Peter R. Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer, 2008.

[52] J. B. Hiriart-Urruty. Tangent cones, generalized gradients and mathematical programming in banach spaces. *Mathematics of Operations Research*, 4(1):79–97, 1979.

[53] Francis H. Clarke. On the inverse function theorem. *Pacific Journal of Mathematics*, 64(1):97–102, 1976.

[54] H. Th. Jongen, Jan-J. Rückmann, and Vladimir Shikhman. On stability of the feasible set of a mathematical problem with complementarity problems. *SIAM Journal on Optimization*, 20(3):1171–1184, 2009.

[55] Max G. Spetzler and Anshu Narang-Siddarth. Numerical verification of equilibrium location bounds and local stability for nonlinear systems with parameter uncertainty. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 4072–4077. IEEE, 2016.

[56] Jean-Pierre Aubin and Hélène Frankowska. *Set-valued analysis*. Springer, 2009.

[57] Mícheál Ó Searcóid. *Metric spaces*. Springer, 2006.

[58] Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010.

[59] Ramon E. Moore. *Interval analysis*, volume 4. Prentice-Hall, 1966.

[60] Andrew P. Smith, Luis G. Crespo, César A. Munoz, and Mark H. Lowenberg. Bifurcation analysis using rigorous branch and bound methods. In *Control Applications (CCA), 2014 IEEE Conference on*, pages 2095–2100. IEEE, 2014.

[61] Luc Jaulin and Fabrice Le Bars. An interval approach for stability analysis: Application to sailboat robotics. *Robotics, IEEE Transactions on*, 29(1):282–287, 2013.

[62] Sascha Warthenpfuhl, Bernd Tibken, and Sascha Mayer. An interval arithmetic approach for the estimation of the domain of attraction. In *Computer-Aided Control System Design (CACSD), 2010 IEEE International Symposium on*, pages 1999–2004. IEEE, 2010.

[63] Robert Swiatlak, Bernd Tibken, Thomas Paradowski, and Robert Dehnert. An interval arithmetic approach for the estimation of the robust domain of attraction for nonlinear autonomous systems with nonlinear uncertainties. In *American Control Conference (ACC), 2015*, pages 2679–2684. IEEE, 2015.

[64] Luiz Henrique De Figueiredo and Jorge Stolfi. Affine arithmetic: concepts and applications. *Numerical Algorithms*, 37(1-4):147–158, 2004.

[65] Jorge Stolfi and Luiz Henrique De Figueiredo. Self-validated numerical methods and applications. In *Monograph for the 21st Brazilian Mathematics Colloquium*, 1997.

[66] Siegfried M. Rump and Masahide Kashiwagi. Implementation and improvements of affine arithmetic. *Nonlinear Theory and Its Applications, IEICE*, 6(3):341–359, 2015.

[67] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Society for Industrial and Applied Mathematics, 2008.

[68] Pascal Gahinet, Pierre Apkarian, and Mahmoud Chilali. Affine parameter-dependent Lyapunov functions and real parametric uncertainty. *Automatic Control, IEEE Transactions on*, 41(3):436–442, 1996.

[69] Pierre-Alexandre Bliman. Nonconservative LMI approach to robust stability for systems with uncertain scalar parameters. In *Decision and Control, 2002, Proceedings of the 41st IEEE Conference on*, volume 1, pages 305–310. IEEE, 2002.

[70] Steven H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology and chemistry*. Perseus Publishing, 2001.

Appendix A

DIRECT COLLOCATION WITH PIECEWISE POLYNOMIALS

As discussed in Section 2.3, direct collocation is a method that transcribes a set of differential equations or differential-algebraic equations into a system of nonlinear algebraic equations with finitely many unknowns, the solution of which yields an approximate solution of the original problem.

This section derives the collocation equations for the differential-algebraic problem

$$\dot{x}(t) = f(x(t), u(t), p) \tag{A.1a}$$

$$0 = h(x(t), p) \tag{A.1b}$$

$$x(T) = x(0) \tag{A.1c}$$

with piecewise polynomials. These equations may be used in (3.17) for computation of limit cycles in the zero dynamics as discussed in Section 3.3.2.

Parametrization of Periodic Piecewise Polynomials

Let $\{\tau_0, \cdots, \tau_N\}$ define a partition of the interval $[0, 1]$ with $0 = \tau_0 < \tau_1 < \cdots < \tau_N = 1$ and let $\tilde{x}_j(\cdot)$, where $j \in \{1, \cdots, N\}$, be polynomials of degree m taking values in $\mathbb{R}^{n_x}$ defined on the interval $[\tau_{j-1}, \tau_j]$ by

$$\tilde{x}_j(\tau) = \sum_{k=0}^{m} x_{j,k} L_{j,k}(\tau). \tag{A.2}$$

In the above, $x_{j,k} \in \mathbb{R}^{n_x}$ are support points which parametrize $\tilde{x}_j$ through the Lagrange polynomials

$$L_{j,k}(\tau) = \prod_{s=0, s\neq k}^{m} \frac{\tau - \tau_{j,s}}{\tau_{j,k} - \tau_{j,s}}, \tag{A.3}$$

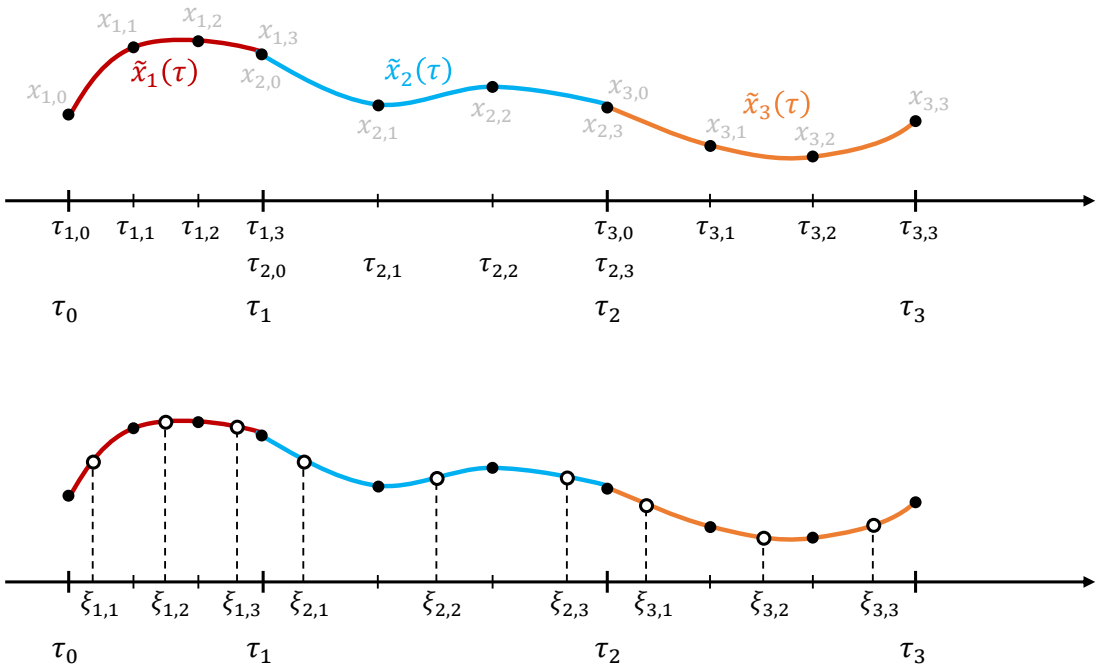

Figure A.1: Piecewise polynomial with support points $x_{j,k}$ and collocation points $\xi_{j,l}$ for $N = m = 3$.

where $\{\tau_{j,0}, \cdots \tau_{j,m}\}$ defines a sub-partition of $[\tau_{j-1}, \tau_j]$ such that $\tau_{j-1} = \tau_{j,0} < \cdots < \tau_{j,m} = \tau_j$. Note that $\tilde{x}_j(\tau_{j,k}) = x_{j,k}$ holds and that the derivative of $\tilde{x}_j$ with respect to τ is

$$\dot{\tilde{x}}_j(\tau) = \sum_{k=0}^{m} x_{j,k} \dot{L}_{j,k}(\tau) \tag{A.4}$$

where

$$\dot{L}_{j,k}(\tau) = \sum_{s=0, s\neq k}^{m} \frac{L_{j,k}(\tau)}{\tau - \tau_{j,s}}. \tag{A.5}$$

The periodic piecewise polynomial $\tilde{x} : \mathbb{R} \to \mathbb{R}^{n_x}$ is now defined as

$$\tilde{x}(\tau) = \tilde{x}_j(\tau - n) \qquad \text{if } \tau_{j-1} \leq \tau - n < \tau_j \text{ for some integer } n \tag{A.6}$$

and continuity is enforced by requiring that

$$x_{j,m} = x_{j+1,0} \text{ for } j \in \{1, \cdots, N-1\} \quad \text{and} \quad x_{N,m} = x_{1,0}. \tag{A.7}$$

The piecewise polynomial is thus parametrized by Nm support points in $\mathbb{R}^{n_x}$. Similarly, let $\tilde{u} : \mathbb{R} \to \mathbb{R}^{n_u}$ be a periodic piecewise polynomial parametrized by the Nm support points $u_{j,k} \in \mathbb{R}^{n_u}$.

Collocation at the Gauss Points

Next, define the m collocation points $\{\xi_{j,1}, \cdots, \xi_{j,m}\}$ on each interval $[\tau_{j-1}, \tau_j]$ as

$$\xi_{j,l} = \tau_{j-1} + \frac{1}{2}(\gamma_l + 1)(\tau_j - \tau_{j-1}) \tag{A.8}$$

where $\{\gamma_1, \cdots, \gamma_m\}$ are the Gauss points on the normalized quadrature interval $[-1, 1]$. With some ambiguity, both the normalized times $\xi_{j,l}$ and the polynomial pieces evaluated at these instants, $\tilde{x}_j(\xi_{j,l})$ and $\tilde{u}_j(\xi_{j,l})$, are referred to as collocation points.

The collocation equations are obtained by enforcing the differential-algebraic equations at the collocation points,

$$\dot{\tilde{x}}_j(\xi_{j,l}) = T f\left(\tilde{x}_j\left(\xi_{j,l}\right), \tilde{u}_j\left(\xi_{j,l}\right), p\right) \tag{A.9a}$$

$$0 = h\left(\tilde{x}_j\left(\xi_{j,l}\right), p\right) \tag{A.9b}$$

This leads to $F^{j,l}_{(\text{coll,pp})}(T, x_{j,0}, \cdots, x_{j,m}, u_{j,0}, \cdots, u_{j,m}, p) = 0$ for all $j \in \{1, \cdots, N\}$ and $l \in \{1, \cdots, m\}$, where

$$F^{j,l}_{(\text{coll,pp})}(\cdot) = \begin{bmatrix} T f\big(\sum_{k=0}^{m} x_{j,k}\Lambda_{j,k,l},\, \sum_{k=0}^{m} u_{j,k}\Lambda_{j,k,l},\, p\big) - \sum_{k=0}^{m} x_{j,k}\dot{\Lambda}_{j,k,l} \\ h\big(\sum_{k=0}^{m} x_{j,k}\Lambda_{j,k,l},\, p\big) \end{bmatrix} \tag{A.10}$$

and

$$\Lambda_{j,k,l} = L_{j,k}(\xi_{j,l}) \qquad \text{and} \qquad \dot{\Lambda}_{j,k,l} = \dot{L}_{j,k}(\xi_{j,l}) \tag{A.11}$$

are constants when the partition and sub-partitions are fixed.

The partial derivatives with respect to the support point $x_{j,k}$ are

$$\nabla_{x_{j,k}} F^{j,l}_{(\text{coll,pp})}(\cdot) = \begin{bmatrix} T\, A_{j,l}\Lambda_{j,k,l} - I\dot{\Lambda}_{j,k,l} & T\, B_{j,l}\Lambda_{j,k,l} \\ C_{j,l}\Lambda_{j,k,l} & 0 \end{bmatrix} \tag{A.12}$$

where I is the identity matrix of size $(n_x \times n_x)$,

$$A_{j,l} = \nabla_x f\Big(\sum_{k=0}^{m} x_{j,k}\Lambda_{j,k,l},\, \sum_{k=0}^{m} u_{j,k}\Lambda_{j,k,l},\, p\Big) \tag{A.13a}$$

$$B_{j,l} = \nabla_u f\Big(\sum_{k=0}^{m} x_{j,k}\Lambda_{j,k,l},\, \sum_{k=0}^{m} u_{j,k}\Lambda_{j,k,l},\, p\Big) \tag{A.13b}$$

$$C_{j,l} = \nabla_x h\Big(\sum_{k=0}^{m} x_{j,k}\Lambda_{j,k,l},\, \sum_{k=0}^{m} u_{j,k}\Lambda_{j,k,l},\, p\Big) \tag{A.13c}$$

and

$$\nabla_{x_{j,k}} F^{j_2,l}_{(\text{coll,pp})}(\cdot) = 0 \qquad \text{for all } j \neq j_2 \tag{A.14}$$

unless $x_{j,k}$ happens to be one of the endpoints that "overlaps" with the neighboring polynomial piece as defined in (A.7).

Zero-Dynamics Monodromy Matrix

In Section 3.3.2, the zero-dynamics monodromy matrix is defined as $M_z = Z^T\hat{\Phi}(1)$, where $\hat{\Phi}$ satisfies

$$\dot{\hat{\Phi}}(\tau) = TA(\tau)\hat{\Phi}(\tau) + TB(\tau)\hat{\Gamma}(\tau) \tag{A.15a}$$

$$0 = C(\tau)\hat{\Phi}(\tau) \tag{A.15b}$$

$$\hat{\Phi}(0) = Z \tag{A.15c}$$

in the normalized time $\tau = t/T$ and A, B, C are obtained by evaluating $\nabla_x f$, $\nabla_u f$, $\nabla_x h$, respectively, along the limit cycle trajectory.

This linear time-varying ODE is again solved with collocation. Using the same partition and sub-partitions as for the state and control trajectories, the matrix-valued functions $\hat{\Phi}$ and $\hat{\Gamma}$ are approximated with (non-periodic) piecewise polynomials

$$\tilde{\Phi}(\tau) = \tilde{\Phi}_j(\tau) \quad \text{and} \quad \tilde{\Gamma}(\tau) = \tilde{\Gamma}_j(\tau) \qquad \text{if } \tau_{j-1} \leq \tau < \tau_j \tag{A.16}$$

parametrized by the support points $\hat{\Phi}_{j,k}$ and $\hat{\Gamma}_{j,k}$ through the definition of the polynomials

$$\tilde{\Phi}_j(\tau) = \sum_{k=0}^{m} \hat{\Phi}_{j,k} L_{j,k}(\tau) \quad \text{and} \quad \tilde{\Gamma}_j(\tau) = \sum_{k=0}^{m} \hat{\Gamma}_{j,k} L_{j,k}(\tau) \tag{A.17}$$

Continuity is again enforced by requiring

$$\hat{\Phi}_{j,m} = \hat{\Phi}_{j+1,0} \quad \text{and} \quad \hat{\Gamma}_{j,m} = \hat{\Gamma}_{j+1,0} \qquad \text{for } \; j \in \{1, \cdots, N-1\} \tag{A.18}$$

Using the collocation points defined in (A.8), the collocation equations become

$$\sum_{k=0}^{m} \begin{bmatrix} T\, A_{j,l}\Lambda_{j,k,l} - I\dot{\Lambda}_{j,k,l} & T\, B_{j,l}\Lambda_{j,k,l} \\ C_{j,l}\Lambda_{j,k,l} & 0 \end{bmatrix} \begin{bmatrix} \hat{\Phi}_{j,k} \\ \hat{\Gamma}_{j,k} \end{bmatrix} = 0 \tag{A.19}$$

for $l \in \{1, \cdots, m\}$ and $j \in \{1, \cdots, N\}$. The initial conditions in (A.15c) may be used to eliminate $\hat{\Phi}_{1,0}$ or simply appended to the equations defined by (A.19). Note that the equations are linear in the unknowns $\hat{\Phi}_{j,k}$ and $\hat{\Gamma}_{j,k}$, and that the matrices multiplying these unknowns are the same as in (A.12). Because the problem does not provide a boundary condition for $\hat{\Gamma}$, the number of unknowns (the components of $\hat{\Phi}_{j,k}$ and $\hat{\Gamma}_{j,k}$) is greater than the number of linear equations in (A.19). A simple solution to obtain a square system of equations is to further constrain the approximating function family for $\hat{\Gamma}$ by requiring that the highest derivative of the last polynomial piece $\tilde{\Gamma}_N$ be zero at the final time $\tau = 1$.

The zero-dynamics monodromy matrix $M_z \approx Z^T \hat{\Phi}_{N,m}$ is thus approximated by solving a sparse linear system of equations based on the Jacobian matrix of the collocation equations associated with the problem defined in (A.1).