Abstract

DICKSON, KELLY IRENE. Condition Estimates for Numerical Continuation with Applications to Atomic and Molecular Fluids. (Under the direction of C.T. Kelley.)

Numerical continuation is the process of solving nonlinear equations of the form $G(u, \lambda) = 0$ for various parameter values, λ. We discuss established numerical continuation techniques for solution paths of $G(u, \lambda) = 0$ containing regular points and simple folds. Pseudo-arclength continuation is a widely used technique that no longer solves $G(u, \lambda) = 0$ directly, but solves a newly parameterized version $F(u(s), \lambda(s)) = 0$ where s is arclength. We present a new characterization for certain classes of points that leads to an upper bound on $\|(F')^{-1}\|$. This bound is needed to meaningfully quantify the convergence of Newton's method in the context of pseudo-arclength continuation.

In particle fluids, it is important to accurately predict conditions under which the structure and thermodynamics of a fluid change. Of particular interest are phase transitions, for instance, when water turns from liquid to vapor as a function of temperature. Such information can often be obtained through rigorous and accurate continuation studies, yet is historically burdensome to produce computationally. To this end, we present new integral equation theory that model atomic and molecular fluids developed

by the Institute of Molecular Design at the University of Houston. We show cases for which this new theory results in more accurate fluid structure and thermodynamics than previously reported in the literature. We discuss the numerical continuation problems of interest for both atomic and molecular fluids and how to implement them using this new theory and software developed at Sandia National Laboratories. Employing these theoretical and computational tools, we provide evidence for the potential to uncover new chemical and physical properties of fluids in an accurate and efficient way.

Condition Estimates for Numerical Continuation with Applications to Atomic and Molecular Fluids

by
Kelly Irene Dickson

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Applied Mathematics

Raleigh, North Carolina

2008

APPROVED BY:

C. T. Kelley Chair of Advisory Committee	I. C. F. Ipsen
Z. Li	D. V. Zenkov

Dedication

This dissertation is dedicated to my new salary, without which I could not afford the many upcoming years of therapy and rehabilitation needed to heal from the effects of graduate school.

Biography

The majority of Kelly Irene Dickson's life was spent in Bow, New Hampshire, a town with no traffic lights unless you count the blinking yellow light in front of the middle school. At 17, Kelly decided she was going to go to college to become a writer and expert in English literature. By 18, she was pre-med. In her first semester of college she failed Calculus I despite the fact she worked homework problems four hours a day in the math tutoring center. A C- in chemistry and a C in zoology filled out her stellar performance that semester.

Kelly went on to get a Bachelor and Master of Science in Applied Mathematics and, with the acceptance of this dissertation, a Doctor of Philosophy. Remember kids, math dreams really can come true.

Acknowledgements

- For Tim Kelley and the characters he resembles:

  Like Santa Claus, he brings me books, computers, and all-expense paid trips to the desert and ski resorts. Better than Santa Claus, he slips into my bank account every other Friday and leaves a generous sum of cash.

  Like Mr. Rogers, he wears many hats (cardigan sweaters, shoes, whatever) including a financial supporter, a career counselor, a mathematical mentor, and the guy who brings me back to reality when I've hopped that trolley to the Neighborhood of Make-Believe.

  Like Buddha, he is truly an "Enlightened One", also translated "Awakened One". But mostly he is "Awakened". As in, awake. Around the clock. With Blackberry in one hand and cup of tea in the other, ready to work.

- For the support and guidance of my all-star committee, Tim Kelley, Ilse Ipsen,

Dmitry Zenkov, Zhilin Li, and Richard Bernhard.

- For Tim Kelley, Monte Pettitt, Marcelo Marucho, Ilse Ipsen, Andy Salinger, Eric Sills, and Gary Howell

  without whom, the subsequent pages of this thesis would be blank.

- For Rik and Jeryl

  whose sacrifices extend beyond the scope of acknowledgement.

- For Charles Martin,

  the One Most.

- For Tanya and Kathy

  and their never wilted support and intrigue.

- For Beth, Erica, Kerry, Kimmy, Melissa, and Steph,

  my home.

- For Bob and Marilyn

  and their 27 years of encouragement and pride.

- For Jon

  and the 2008 World Champion Boston Celtics!

- For Chris Baker and Anne Costolanski,

  my Prozac.

- For the research group past and present

Karen Dillard, Dan Finkel, Chris Kees, Matthew Lasater, Jill Reese, Corey Winton, Scott Pope, Anne Costolanski, and Dave Mokrauer.

- For the Sandia National Lab community and the mathematical, computer, career, and personal support of Andy Salinger, Danny Dunlavy, Brian Adams, Chris Baker, Roger Pawlowski, Eric Phipps, Tammy Kolda, Heidi Thornquist, Genetha Gray, Josh Griffin, Bill Goldman, and Roger Retallack.

- For the career support of

  Paul Hovland, Mihai Anitescu, Rao Kotamarthi, Sven Leyffer, Melissa Choi, Aryeh Feder, Paul Hezel, Dan Finkel, Carol Woodward, Lori Diachin, Owen Eslinger, Andrei Draganescu, and Jason Howell.

- For The Angels,

  Karen and Jill.

- For the Graduate Student Numerical Analysis Seminar members, especially Tim Kelley, Ilse Ipsen, Pierre Gremaud, and Steve Campbell.

- For those who inspired me mathematically at an early age

  Shashi Ramakrishna, John Klippert, Paul Warne, Debra Warne, Dave Pruett, James Liu, Daniel Robinson, Mary Lee, and Steve Penny.

- For the assistance and guidance of

  Steve Campbell, Ernie Stitzinger, Jeff Scroggs, Denise Seabrooks, Carolyn Gunton, Seyma Bennett Shabbir, Adam Attarian, Di Bucklad, Janice Gaddy, Mary Byrd, Brad Priser, Charlene Wallace, Miriam Ansley, and Erica Cutchins.

- Finally, for my (previously unmentioned) peers, without whom, I would have taken the Bachelor's and called it a day

  Teresa Selee, Ryan Siskind, Rebecca Kalhorn, Laurie Zack, Monique Taylor, Laura Ellwein, April Alston, Jimena Davis, Johnny Samuels, Jon Brown, Iti Janovitz Freireich, Kristen DeVault, Louis Levy, Drew Pasteur, Rebecca Wills, Nicole Kroeger, Christina Oxendine, Francesca Reale-Levis, Rizwana Rehman, Anna Hart, Kelly Sweetingham, Julie Beier, and Ryan Therkelsen.

Table of Contents

List of Figures

Chapter 1

Introduction

Numerical continuation is the process of solving systems of nonlinear equations

$$G(u, \lambda) = 0$$

for various values of a real parameter λ. Here, $u \in \Re^N$, $\lambda \in \Re$, and $G : \Re^{N+1} \to \Re^N$.

An obvious approach for implementing numerical continuation, called *natural parameterization* [16, 13, 20, 35], traces out a solution path by repeatedly incrementing λ until the desired value of λ is reached. In each such iteration, the current solution u is used as an initial iterate for obtaining the next solution corresponding to the next value of λ. Although natural parameterization is simple and intuitive, it fails at points (u, λ) where the Jacobian G_u is singular. While this thesis does not address classic singularity (or catastrophe) theory directly, we provide an introductory summary of its major concepts in the Appendix. The type of singularity we consider in this thesis is called a *simple fold*.

For simple folds, the standard way to remedy the failure of natural parameterization is to reparameterize the problem by introducing the arclength parameter, s, so that both u and λ depend on s. This idea, known as *pseudo-arclength continuation* (PAC) [16, 13, 20, 35], implements natural parameterization on $F(u(s), \lambda(s)) = 0$ with s as the parameter and $x = (u^T, \lambda)^T$ as the unknown. Thus PAC requires that the Jacobian F_x of F be nonsingular. It is known that F_x is nonsingular at simple folds and points where G_u is nonsingular [16].

There are two main results presented in this thesis. The first is a new quantification of a simple fold in terms of a singular value decomposition (SVD) of G_u. We use this quantification to present a new upper bound on $\|(F_x)^{-1}\|_2$ which is needed to guarantee the convergence of Newton's method. As a byproduct, we obtain a refinement of Weyl's monotonicity theorem [29] for the smallest eigenvalue of a symmetric positive semi-definite matrix plus a rank-one update.

The second main contribution of this thesis is a new way to implement continuation for studying structural and thermodynamic properties of atomic and molecular fluids. For example, suppose one is interested in conditions under which water transitions from liquid to vapor. This transition depends on the water's temperature, for example. Thus using equations that model water, one can define temperature as a parameter and perform a continuation study to observe how critical properties of water change with increasing temperature. Using new integral equation theory developed by the Institute of Molecular Design at the University of Houston and continuation software developed by Sandia National Laboratories, we present significant headway for discovering new physical properties of fluids.

Chapter 2

Preliminaries

In this chapter, we review in detail the two main concepts that motivate the findings in Chapter 4. Newton's method, discussed in §2.1, highlights major convergence theorems and considerations for solving systems of nonlinear equations. Newton's method is the primary numerical solver involved in all numerical continuation techniques as described in Chapter 3. For now, we focus on Newton's method for general nonlinear systems without the context of continuation. In §2.2, we outline the well-known *implicit function theorem*, the theoretical core in describing solution path behavior from a continuation viewpoint.

2.1 Newton's Method

Newton's method [17, 18] solves systems of nonlinear equations of the form

$$G(x) = 0 \tag{2.1}$$

where $G : \Re^N \to \Re^N$. The method is given by

$$x_{n+1} = x_n - G_x(x_n)^{-1} G(x_n) \tag{2.2}$$

where G_x denotes the Jacobian matrix of G. This iterative formulation arises from the first degree Taylor expansion of the function G about the current iteration. That is,

$$T(x) = G(x_n) + G_x(x_n)(x - x_n).$$

The root of $T(x)$ is given by x_{n+1} in (2.2). Throughout this thesis we denote $\mathcal{B}_\epsilon(y)$ the ball of radius ϵ about y. That is,

$$\mathcal{B}_\epsilon(y) \equiv \{z \in \Omega \mid \|z - y\| < \epsilon\}$$

where Ω is an open subset of $\Re^N$. Although the results in this chapter hold for any vector norm (and induced matrix norm), it is most convenient for future chapters to consider the discrete Euclidean norm defined by

$$\|y\|_2 = \left(\sum_{i=1}^{N} |y_i|^2 \right)^{1/2}$$

where $y = (y_1\ y_2\ y_3 \dots y_N)^T \in \Re^N$. We will also make use of

Definition 2.1. *Let $\Omega \subset \Re^N$ be an open set and let $G : \Omega \to \Re^N$. G is* **Lipschitz continuous** *[28] on Ω with Lipschitz constant γ if*

$$\|G(x) - G(y)\| \leq \gamma \|x - y\| \quad \forall \quad x, y \in \Omega.$$

The following lemma, found in [17], provides a useful tool for proving the local convergence associated with Newton's method which we present momentarily. In essence, the lemma gives an upper bound on the inverse Jacobian matrix $G_x(x)^{-1}$ for all x in a neighborhood surrounding the exact solution x^*.

Lemma 2.1. *Let $G(x) = 0$ have a solution x^*. Let $\Omega \subset \Re^N$ be an open set containing x^* and $G_x : \Omega \to \Re^{N\times N}$ be Lipschitz continuous with Lipschitz constant γ. Let $G_x(x^*)$ be nonsingular. Then there is a $\delta > 0$ such that for all $x \in \mathcal{B}_\delta(x^*)$,*

$$\|G_x(x)^{-1}\| \leq 2\|G_x(x^*)^{-1}\|.$$

Proof. Let δ be small enough so that $B_\delta(x^*) \subset \Omega$. Choosing

$$\delta < \frac{\|G_x(x^*)^{-1}\|^{-1}}{2\gamma}$$

yields

$$\|I - G_x(x^*)^{-1}G_x(x)\| = \|G_x(x^*)^{-1}(G_x(x^*) - G_x(x))\|$$

$$\leq \gamma\|G_x(x^*)^{-1}\|\|x - x^*\| \leq \gamma\delta\|G_x(x^*)^{-1}\| < 1/2.$$

Then the Banach lemma [23] implies that $G_x(x^*)^{-1}G_x(x)$ is nonsingular with

$$\|G_x(x)^{-1}\|\|G_x(x^*)^{-1}\|^{-1} \leq \|G_x(x^*)^{-1}G_x(x)\| \leq \frac{1}{1-\frac{1}{2}} = 2$$

so that

$$\|G_x(x)^{-1}\| \leq 2\|G_x(x^*)^{-1}\|.$$

□

The well-known *fundamental theorem of Calculus* [27] for many variables will also come to our aid in the subsequent section. It can be stated as

Theorem 2.1. *Let G be differentiable in an open set $\Omega \subset \Re^N$ with $x^* \in \Omega$. Then for all $x \in \Omega$ sufficiently near x^*,*

$$G(x) - G(x^*) = \int_0^1 G_x(x^* + t(x - x^*))(x - x^*)dt.$$

Under appropriate conditions, Newton's method exhibits *q-quadratic* convergence:

Definition 2.2. *Let $\{x_n\} \subset \Re^N$ and $x^* \in \Re^N$. Then $x_n \to x^*$* **q-quadratically** *[18] if $x_n \to x^*$ and there is $K > 0$ such that*

$$\|x_{n+1} - x^*\| \leq K\|x_n - x^*\|^2.$$

2.1.1 Local Convergence

With the use of Lemma 2.1, we prove the q-quadratic convergence of Newton's method in Theorem 2.3. First, however, we have

Theorem 2.2. *[17] Let $G(x) = 0$ have a solution x^*. Let $G_x : \Omega \to \Re^{N \times N}$ be Lipschitz continuous with Lipschitz constant γ where $\Omega \subset \Re^N$ is an open set containing x^*. Let $G_x(x^*)$ be nonsingular. Then there are $K > 0$ and $\delta > 0$ such that if $x_n \in \mathcal{B}_\delta(x^*)$, the*

Newton iterate $x_{n+1} = x_n - G_x(x_n)^{-1}G(x_n)$ *satisfies*

$$\|x_{n+1} - x^*\| \ \leq \ K\|x_n - x^*\|^2.$$

Proof. Let δ be small enough so that Lemma 2.1 holds. Taking the definition of the Newton iteration and subtracting x^* from both sides yields

$$x_{n+1} - x^* = x_n - x^* - G_x(x_n)^{-1}G(x_n) = x_n - x^* - G_x(x_n)^{-1}\left[G(x_n) - G(x^*)\right] \quad (2.3)$$

since $G(x^*) = 0$. Using the fundamental theorem (Theorem 2.1) on the expression $G(x_n) - G(x^*)$ from the last equality above and the fact that $x_n - x^* = \int_0^1 (x_n - x^*)dt$, we see that

$$x_n - x^* - G_x(x_n)^{-1}\left[G(x_n) - G(x^*)\right] = \int_0^1 \left[I - G_x(x_n)^{-1}G_x(x^* + t(x_n - x^*))\right](x_n - x^*)dt.$$

Factoring out a $G_x(x_n)^{-1}$ from the last equality above and equating back to (2.3) gives

$$x_{n+1} - x^* = G_x(x_n)^{-1}\int_0^1 \left[G_x(x_n) - G_x(x^* + t(x_n - x^*))\right](x_n - x^*)dt.$$

Taking norms in the above equation and using norm properties yields

$$\begin{aligned}
\|x_{n+1} - x^*\| &\leq \|G_x(x_n)^{-1}\| \int_0^1 \|G_x(x_n) - G_x(x^* + t(x_n - x^*))\| \ \|x_n - x^*\| dt \\
&\leq 2\|G_x(x^*)^{-1}\| \|x_n - x^*\| \int_0^1 \|G_x(x_n) - G_x(x^* + t(x_n - x^*))\| dt \\
&\leq 2\gamma \|G_x(x^*)^{-1}\| \ \|x_n - x^*\| \int_0^1 \|x_n - x^* - t(x_n - x^*))\| dt \\
&\leq 2\gamma \|G_x(x^*)^{-1}\| \ \|x_n - x^*\|^2 \int_0^1 (1 - t) dt \\
&= \gamma \|G_x(x^*)^{-1}\| \ \|x_n - x^*\|^2.
\end{aligned}$$

In the second inequality above, we use Lemma 2.1 and factor out $\|x_n - x^*\|$ from the integral since it does not depend on t. In the third inequality, we use Lipschitz continuity of G_x (Definition 2.1) inside the integral. In the fourth inequality, we factor out another $\|x_n - x^*\|$ and since the remaining integral in t is equal to one half, we arrive at the final equality. Choosing $K = \gamma \|G_x(x^*)^{-1}\|$ concludes the proof. □

Using Theorem 2.2, we finally present the convergence result for the complete Newton iteration.

Theorem 2.3. *[17] Let $G(x) = 0$ have a solution x^*. Let $G_x : \Omega \to \Re^{N \times N}$ be Lipschitz continuous with Lipschitz constant γ where $\Omega \subset \Re^N$ is an open set. Let $G_x(x^*)$ be nonsingular. Then there is a $\delta > 0$ such that if $x_0 \in \mathcal{B}_\delta(x^*)$, the Newton iteration given by $x_{n+1} = x_n - G_x(x_n)^{-1} G(x_n)$ converges q-quadratically to x^*.*

Proof. Reduce the δ from Theorem 2.2 so that $K\delta \equiv \eta < 1$. Let $n \geq 0$ and $x_n \in \mathcal{B}_\delta(x^*)$.

Then Theorem 2.2 gives

$$\|x_{n+1} - x^*\| \leq K\|x_n - x^*\|^2 \leq K\delta\|x_n - x^*\| = \eta\|x_n - x^*\| < \|x_n - x^*\| \tag{2.4}$$

so that $x_{n+1} \in \mathcal{B}_{\eta\delta}(x^*) \subset \mathcal{B}_\delta(x^*)$. Since $x_0 \in \mathcal{B}_\delta(x^*)$ by assumption, so is the entire sequence x_n, and (2.4) implies $x_n \to x^*$ q-quadratically. □

One iteration of Newton's method requires evaluating $G(x_n)$, solving for the *Newton step* s

$$G_x(x_n)s = -G(x_n), \tag{2.5}$$

computing the next iterate $x_{n+1} = x_n + \nu s$ where ν is chosen to decrease $\|G\|$, and finally, testing for convergence subject to some tolerance. We now elaborate on these points and discuss computational issues for Newton's method.

2.1.2 The Jacobian Matrix and the Newton Step

The most difficult task in implementing Newton's method is the calculation of the Newton step in (2.5). One must consider the computational effort required to make this linear solve involving the Jacobian matrix G_x and the function G. Depending on the size of the problem, availability of an analytic Jacobian with or without sparsity patterns, and other issues addressed in [17], a direct or iterative method may be preferred. The example problems solved in §4.1 employ direct methods only, in particular, PLU *factorization* in which an analytic Jacobian is discretized and factored as $G_x = PLU$. The matrix P is a permutation matrix while L and U are lower and upper triangu-

lar factors, respectively, with the diagonal of L containing all ones. Using the PLU factorization of the Jacobian to solve for the Newton step is equivalent to the well-known Gaussian elimination scheme with partial pivoting [25]. Thus when discussing the linear solve within Newton's method, we often refer to it as a factorization.

For the atomic and molecular fluids application presented in Chapter 5, the problem is far too large with no analytic Jacobian matrix available. Instead, an iterative Krylov method (namely GMRES) is used as the linear solver for the Newton step with each nonlinear iteration. In Chapter 5, the Jacobian matrix is approximated with a finite difference which is classically used to compute directional derivatives for the matrix-vector product routine in GMRES. Details on numerical linear algebra extend beyond the scope of this thesis, but good references are [17, 18, 25, 36]. The important distinction to keep in mind throughout this thesis is the difference between the linear and nonlinear iteration. Again, the linear iteration refers to the iterative linear solver that solves for the Newton step within each Newton (nonlinear) iteration. For more discussion to this end, see the next section under *Iterative Variations of Newton's Method*.

2.1.3 Variations of Newton's Method

To be complete in our discussion of Newton's method, we must mention a few widely used variations of the classical Newton iteration described at the beginning of this section. When using direct methods (think PLU factorization) for solving for the Newton step, it is often impractical or unnecessary to refactor the Jacobian matrix with every Newton iteration. We first present two options that reduce the number of Jacobian fac-

tors throughout the nonlinear iteration, but at a penalty of slower convergence. Next, we discuss *inexact Newton methods* which are used when an iterative linear solver is used for computing the Newton step (think GMRES). Inexact Newton methods accept an adequate Newton step by testing to see if the relative linear residual is sufficiently small. We now elaborate on these variations which are discussed further in [17, 18].

Direct Variations of Newton's Method. The chord and Shamanskii methods [18] are close relatives to Newton's method, but do not require refactoring the Jacobian matrix with each nonlinear iteration.

The chord method uses the Jacobian from the initial iterate in all subsequent nonlinear iterations. Formally, the chord method is given by [17]

$$x_{n+1} = x_n - G_x(x_0)^{-1} G(x_n). \tag{2.6}$$

The chord method converges *q-linearly* [17]:

Definition 2.3. *Let* $\{x_n\} \subset \Re^N$ *and* $x^* \in \Re^N$. *Then* $x_n \to x^*$ **q-linearly with q-factor** $\sigma \in (0,1)$ *if*

$$\|x_{n+1} - x^*\| \leq \sigma \|x_n - x^*\|$$

for sufficiently large n.

The convergence estimate for the chord method is as follows. The proof can be found in [17] and follows a very similar argument as in the proof of Theorems 2.2 and 2.3.

Theorem 2.4. *Let the assumptions of Theorem* 2.2 *hold. Then there are* $K > 0$ *and* $\delta > 0$ *such that if* $x_0 \in \mathcal{B}_\delta(x^*)$*, the chord iterates* (2.6) *converge q-linearly to* x^* *and*

$$\|x_{n+1} - x^*\| \leq K\|x_0 - x^*\|\|x_n - x^*\|.$$

The Shamanskii method represents a compromise between Newton's method and the chord method. A new Jacobian is refactored with every m nonlinear iterations. The Shamanskii method converges *q-superlinearly*:

Definition 2.4. *[17] Let* $\{x_n\} \subset \Re^N$ *and* $x^* \in \Re^N$*. Then* $x_n \to x^*$ **q-superlinearly with q-order** $\alpha > 1$ *if* $x_n \to x^*$ *and there is* $K > 0$ *such that*

$$\|x_{n+1} - x^*\| \leq K\|x_n - x^*\|^\alpha.$$

The convergence result associated with the Shamanskii method is stated here and proved in [17] where, once again, the proof mirrors that of Theorems 2.2 and 2.3.

Theorem 2.5. *Let the assumptions of Theorem* 2.2 *hold and let an integer* $m \geq 1$ *be given. Then there are* $K > 0$ *and* $\delta > 0$ *such that if* $x_0 \in \mathcal{B}_\delta(x^*)$*, the Shamanskii iterates converge q-superlinearly to* x^* *with q-order* $m + 1$ *and*

$$\|x_{n+1} - x^*\| \leq K\|x_n - x^*\|^{m+1}.$$

Iterative Variations of Newton's Method. *Inexact Newton methods* [17] do not solve for the Newton step exactly, but only require the relative linear residual to be small enough, that is, s must satisfy

$$\|G_x(x_n)s + G(x_n)\| \leq \eta_n \|G(x_n)\|. \tag{2.7}$$

The term η_n is called the *forcing term*. Inexact methods are employed when the linear solve for s is done iteratively.

The following result [17] describes convergence of inexact Newton methods based on the behavior of the sequence $\{\eta_n\}$. It essentially says that when using an inexact Newton method as described above, convergence is q-linear (Definition 2.3), and when the forcing term sequence sufficiently decays, is q-superlinear (Definition 2.4).

Theorem 2.6. *Let the assumptions of Theorem 2.2 hold. Then there are $\delta > 0$ and $\bar{\eta} > 0$ such that if $x_0 \in \mathcal{B}_\delta(x^*)$ and $0 \leq \eta_n \leq \bar{\eta}$, then the inexact Newton iterates $x_{n+1} = x_n + s_n$ where*

$$\|G_x(x_n)s_n + G(x_n)\| \leq \eta_n \|G(x_n)\|,$$

converge q-linearly to x^*. Moreover,

- *if $\eta_n \to 0$ the convergence is q-superlinear, and*
- *if $\eta_n \leq K\|G(x_n)\|^p$ for some $K > 0$, the convergence is q-superlinear with q-order $1 + p$.*

2.1.4 Termination Criteria

When writing a Newton method code, exact or inexact, one must consider when to terminate the iteration. The Newton code implemented in this thesis terminates when

$$\|G(x_n)\| \leq tol_{rel}\|G(x_0)\| + tol_{abs} \tag{2.8}$$

where tol_{rel} and tol_{abs} are relative and absolute tolerances, respectively. In practice, this is generally a reliable condition provided the assumptions from Theorem 2.2 hold and the initial iterate x_0 is sufficiently near the root [18]. If these conditions hold, it can be shown [18] that in general,

$$\frac{\|x_n - x^*\|}{4\kappa\|x_0 - x^*\|} \leq \frac{\|G(x_n)\|}{\|G(x_0)\|} \leq \frac{4\kappa\|x_n - x^*\|}{\|x_0 - x^*\|}$$

where $\kappa \equiv \|G_x(x^*)\|\|G_x(x^*)^{-1}\|$ is the condition number (with respect to inversion) of $G_x(x^*)$. This inequality says if the sequence of iterates stays within the ball of convergence (which is less than one in radius by assumption from Theorem 2.2), then the relative nonlinear residuals stay bounded provided the Jacobian matrix of $G(x^*)$ is well-conditioned. Note that the tol_{abs} term in (2.8) is needed in the event that the initial iterate is the exact solution up to round-off error. That is, if $G(x_0)$ is nearly zero, then you will need the absolute tolerance to terminate on.

Another termination option measures the size of the Newton step, s_n. That is, one terminates the iteration when $\|s_n\|$ is sufficiently small since $\|x_n - x^*\| = \|s_n\| + O(\|x_n - x^*\|^2)$ (from the definition of the Newton iteration). In other words, near x^* the magnitude of the Newton step is nearly that of the error.

Regardless of which Newton variation and termination criteria is used, declaring a maximum number of iterations on the Newton solver is smart: if one fails to reach the requested termination criterion, doing this will guarantee that the iteration terminates

eventually (avoid infinite loops!). If the Newton solver reaches the maximum number of iterations, it means the termination criterion was not met, in which case, one must consider why. This might occur if there is a bug in the code or if one or more of the standard assumptions from Theorem 2.2 is violated.

2.1.5 Line search methods

Another concern is the magnitude of the Newton step. After computing the Newton step, one might find that it points in the direction of the root, but is so large in magnitude, it overshoots the root. This can happen when local convergence results do not apply because the initial iterate is far from the root. Safeguards against such behavior are called *line search methods* in which one seeks a decrease in $\|G\|$ along the line segment $[x_n, x_n + d_n]$ where d_n is the *Newton direction*

$$d_n = -G_x(x_n)^{-1}G(x_n).$$

Here, we make the distinction between the *step* and the *direction* to indicate that the actual step s_n is not chosen until the described decrease is found along $[x_n, x_n + d_n]$. The Newton codes from [18] and [32] use the *Armijo rule*, a type of line search method, which finds the smallest integer $k \geq 0$ such that

$$\|G(x_n + 2^{-k}d_n)\| < (1 - \alpha 2^{-k})\|G(x_n)\|.$$

This is called the *sufficient decrease* of $\|G\|$ and the 2^{-k} term keeps the direction from going too far. Choosing α to be small helps to make the sufficient decrease easy to

satisfy. Once an appropriate k is settled upon, set $s_n = 2^{-k}d_n$ so that $x_{n+1} = x_n + 2^{-k}d_n$. In the code from [18], $\alpha = 10^{-4}$.

If G is Lipschitz continuously differentiable, there are only three possibilities for global Newton-Armijo convergence [18]:

- $\{x_n\} \to x^*$,
- $\|x_n\| \to \infty$, or
- $G_x(x_n)$ becomes singular for some x_n.

Additionally,

Theorem 2.7. *[18] Let $x_0 \in \Re^N$ and $\alpha \in (0,1)$ be given. Suppose that $\{x_n\}$ are the Newton-Armijo iterates and that G is Lipschitz continuously differentiable in $\Re^N$. Let $0 < \eta_n \leq \bar{\eta} < 1 - \alpha$ and $\{x_n\}, \{\|G_x(x_n)^{-1}\|\}$ be bounded. Then $\{x_n\}$ converges to a root x^* of G at which the assumptions from Theorem 2.2 hold, full steps ($k = 0$) are taken for n sufficiently large, and the convergence behavior in the final phase of the iteration is that given by Theorem 2.6.*

2.2 Implicit Function Theorem

We now switch gears from solving the problem $G(x) = 0$ to solving a system of *parameter dependent* nonlinear equations of the form

$$G(u, \lambda) = 0 \tag{2.9}$$

where $u \in \Re^N$, $\lambda \in \Re$, and $G : \Re^{N+1} \to \Re^N$. Now, u is the unknown and λ is a real number parameter. The solution to (2.9) depends on the value of λ. The remainder of this thesis is dedicated to solution procedures (utilizing Newton's method) and numerical analysis involved in solving (2.9). Here, we present the *implicit function theorem* [23, 16, 28], an important ingredient in the analysis throughout Chapters 3 and 4.

The implicit function theorem gives conditions under which a unique solution to (2.9) exists. The theme of this thesis is to investigate instances where the implicit function theorem is violated and how solution behavior and numerical methods are affected in this case. This study is called *bifurcation analysis*. In other words, certain parameter values λ may violate the assumptions of the implicit function theorem. Thus there may be more than one solution to (2.9) for that particular parameter value. Finding all such solutions (and being aware that they exist!) is a very important and difficult task. For example, if the function G represents a physical model (as we will see in Chapter 5) and there are many solutions to (2.9) for a given λ, it is important to distinguish the physically meaningfully solutions from the physically irrelevant ones.

In the statement of the theorem below, $\|\cdot\|$ denotes the Euclidean norm, and $C^k(\Omega)$ denotes the space of k times continuously differentiable functions from an open subset $\Omega \subset R^{N+1}$ to R^N. The Jacobian matrix is denoted G_u and G_λ is the derivative of G with respect to λ.

Theorem 2.8. *Let Ω be an open subset of R^{N+1} and let $G \in C^k(\Omega)$ for some integer $k > 0$. Let G_u and G_λ be Lipschitz continuous in $\bar{\Omega}$, the closure of Ω. If*

- $(u_0, \lambda_0) \in \Omega$,

- $G(u_0, \lambda_0) = 0$, *and*
- $G_u(u_0, \lambda_0)$ *is nonsingular,*

then there are $\rho > 0$ *and* $\epsilon > 0$ *such that there is a unique solution*

$$v \in \mathcal{B}_\rho(u_0) = \{u \,|\, \|u - u_0\| < \rho\}$$

of $G(u, \lambda) = 0$ *for all* $\lambda \in (\lambda_0 - \epsilon, \lambda_0 + \epsilon)$ *and* $u \in \mathcal{B}_\rho(u_0)$. *Furthermore,* v *is a* k *times continuously differentiable function of* λ.

Chapter 3

Numerical Continuation Basics

This chapter describes procedures for solving

$$G(u, \lambda) = 0 \tag{3.1}$$

where $G : \Re^{N+1} \to \Re^N$.

As mentioned in Chapter 1, the process of finding solutions u to (3.1) for various parameter values λ is called numerical continuation. In §3.1, we discuss *natural parameterization*, the intuitive approach for solving (3.1) for various parameter values. We will see where and how this approach fails and turn to a well-known alternative strategy in §3.2 called *pseudo-arclength continuation*. Pseudo-arclength continuation is the focus of our new analysis in Chapter 4. The remaining sections in this chapter outline the key ingredients needed for a continuation algorithm, be it natural parameterization or pseudo-arclength. The last section is a summary of existing continuation literature and the formulations they implement in continuation algorithms.

3.1 Natural Parameterization

Natural parameterization (or *parameter continuation*) [16, 13, 20, 35] is the simplest and most intuitive method for solving $G(u, \lambda) = 0$. The idea is to start at a point $\lambda = \lambda_{init}$ and solve $G(u, \lambda)$ for u using Newton's method as described in §2.1. Then use the solution $u_0 \equiv u$ as the initial iterate to solve the next problem $G(u, \lambda + d\lambda) = 0$ where $d\lambda$ is the parameter increment size $|\lambda_{current} - \lambda_{previous}|$. Algorithm `npalg` below is a simple implementation of natural parameterization from an initial parameter value λ_{init} to a desired stopping value λ_{end}.

`npalg`$(u, G, \lambda_{init}, \lambda_{end}, d\lambda)$
  Set $\lambda = \lambda_{init}$, $u_0 = u$
  while $\lambda \leq \lambda_{end}$ **do**
    Solve $G(u, \lambda) = 0$ via Newton's method with u_0 as the initial iterate to obtain u_1
    $u_0 = u_1$
    $\lambda = \lambda + d\lambda$
  end while

Parameter continuation is a reasonable method for solving $G(u, \lambda) = 0$ as long as the assumptions of the implicit function theorem (§2.2) and Theorem 2.3 remain satisfied with each increment in the parameter. This is because Theorem 2.8 says a unique solution to (3.1) exists for each parameter, and as we will show below, Newton's method will happily converge.

Corollary 3.1. *[8] Let G be Lipschitz continuously differentiable, $G(u_0, \lambda_0) = 0$, and $G_u(u_0, \lambda_0)$ be nonsingular. Then there is a $\delta > 0$, which depends only on $\|G_u^{-1}(u_0, \lambda_0)\|$ and the Lipschitz constants of G_u and u, such that if $|\lambda - \lambda_0| < \delta$ then Newton's method*

with initial iterate u_0 converges q-quadratically to the solution $u(\lambda)$ of $G(u,\lambda)=0$, i.e.,

$$\|u_{n+1}-u(\lambda)\| = O(\|u_n-u(\lambda)\|^2),$$

where, for $n \geq 0$,

$$u_{n+1} = u_n - G_u(u_n,\lambda)^{-1}G(u_n,\lambda).$$

Proof. Define the Lipschitz constant

$$\|G_u(u,\lambda)-G_u(v,\mu)\| \leq \gamma_G(\|u-v\|+|\lambda-\mu|).$$

Differentiating $G(u,\lambda)=0$ with respect to λ gives

$$du/d\lambda = -G_u^{-1}G_\lambda.$$

The implicit function theorem implies that there is a δ_1 such that if

$$|\lambda-\lambda_0| \leq \delta_1$$

then there is a solution arc $u(\lambda)$ defined for $|\lambda-\lambda_0| \leq \delta_1$. Since $G_u^{-1}G_\lambda$ is Lipschitz continuous, there is a γ_u, which depends only on $\|G_u^{-1}(u_0,\lambda_0)\|$ and the Lipschitz constants of G_u and G_λ, such that

$$\|du/d\lambda\| = \|G_u^{-1}G_\lambda\| \leq \gamma_u.$$

A lower bound for the radius of the ball of convergence for the Newton iteration (see Theorem 2.2) is

$$\frac{1}{2\gamma_G \|G_u^{-1}(u_0, \lambda_0)\|},$$

so choosing

$$\delta = \min\left(\delta_1, \frac{1}{2\gamma_u \gamma_G \|G_u^{-1}(u_0, \lambda_0)\|}\right)$$

completes the proof. □

This corollary also highlights the importance of the size increment in the parameter, $d\lambda$. This issue, along with other computational details for `npalg`, are addressed in detail in §3.2.

While Corollary 3.1 gives conditions under which natural parameterization will succeed, this algorithm will fail at *singular points* where the assumptions to Theorems 2.8 and 2.3 are violated.

Definition 3.1. *A* **singular point** *is a solution* (u_0, λ_0) *to* $G(u, \lambda) = 0$ *for which* $G_u(u_0, \lambda_0)$ *is singular.*

So far in this thesis, we have emphasized the nonsingularity of the Jacobian matrix of G. The next section and Chapter 4 emphasize solution paths to $G(u, \lambda) = 0$ (3.1) that involve singular points. Singular points are an indication that more than one solution path to (3.1) exists since, at such points, the implicit function theorem no longer guarantees a unique solution to (3.1). Moreover, Corollary 3.1 says Newton's method will not converge q-quadratically during the natural parameterization scheme. The concept of singular points is extremely important when G represents a physical

model. In physics, singular points and multiple solution paths represent major transitions in physical properties of the system being modeled. This is exactly the focus of Chapter 5. Without robust algorithms that trace solution paths to $G(u, \lambda) = 0$ accurately (particularly in the presence of singular points), the science of the problem can easily be misinterpreted.

Once again, Algorithm `npalg` is insufficient when the Jacobian matrix G_u becomes singular for some (u_0, λ_0). Thus we seek an alternative continuation method.

3.2 Pseudo-Arclength Continuation (PAC)

Pseudo-arclength continuation (PAC) [16, 13, 20, 35] avoids the problems of Algorithm `npalg` at certain types of singular points through an arclength parameterization. PAC is the standard continuation technique for solution paths containing certain kinds of singularities (we will discuss one such singularity in the next chapter).

The idea of PAC is to choose arclength s as the parameter instead of λ where now $x(s) = (u(s)^T, \lambda(s))^T$ is the unknown. Then compute the solution curve to $G(u(s), \lambda(s)) = G(x(s)) = 0$ with natural parameterization (Algorithm `npalg`, §3.1) using this reparameterization. The definition of arclength with respect to the Euclidean norm from calculus is

$$\|\dot{x}\|^2 = \|\dot{u}\|^2 + |\dot{\lambda}|^2 = 1 \tag{3.2}$$

where the "dot" notation represents differentiation with respect to s. Since we introduced the new parameter s, we must add an equation to $G(u, \lambda) = 0$ so that the number of equations equals the number of unknowns. Hence, we work with the ex-

tended equations

$$F(x,s) = \begin{pmatrix} G(x) \\ N(x,s) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{3.3}$$

The normalization equation $N = 0$ is an approximation of (3.2) where

$$N(x,s) = \dot{x}_0^T(x - x_0) - (s - s_0) = 0. \tag{3.4}$$

This equation says that the new point on the path lies on the hyper-plane perpendicular to the tangent vector at a distance $|s - s_0|$ through the current point x_0.

Given a point (x_0, s_0), PAC increments arclength by ds, and solves (3.3) with the normalization (3.4) by Newton's method with initial iterate x_0. Algorithm `paalg` is a simple implementation of pseudo-arclength continuation. Note that `paalg` is nothing but `npalg` for the new problem (3.3).

`paalg`(u, F, s_{end}, ds)
  Set $s = 0$, $x_0 = (u_0^T, \lambda_0)^T$
  while $s \leq s_{end}$ **do**
    Approximate $\dot{x}$
    Solve $F(x,s) = 0$ via Newton's method with x_0 as the initial iterate to obtain x_1
    $x_0 = x_1$
    $s = s + ds$
  end while

We now take a moment to describe some important details for implementing `paalg`. Note that the following considerations also apply to Algorithm `npalg` from the previous section where u replaces x and λ replaces s.

The examples in §4.1 and the work from Chapter 5 approximate $\dot{x}$ using the *secant*

formulation [35, 13, 20, 16]. That is, if $x_0 = (u_0^T, \lambda_0)^T$ and $x_1 = (u_1^T, \lambda_1)^T$ are the two previous known solutions to $F(x, s) = 0$, form

$$\dot{x} = \frac{x_1 - x_0}{s_1 - s_0}.$$

Algorithm `paalg` (and Algorithm `npalg`) both use the previous known solution x_0 as an initial iterate for the next step in arclength. This is known as a *zeroth order predictor* or a *trivial predictor* [35, 20]. A *first order predictor*, however, is more commonly used [35, 16, 20] and is the initial iterate formulation used for results presented in this thesis. A first order predictor is simply given by $2x_1 - x_0$ where x_1 and x_0 are two previous known solutions to $F(x, s) = 0$.

Choosing a reasonable parameter step size is also important, not just for theoretical purposes (as we will demonstrate below), but for computational purposes as well: if one is overly conservative (very small step length), the continuation strategy will spend a lot of time on parts of a solution path that may not be of much interest. In contrast, choosing too large of a step size might cause the continuation scheme to "step over" important solution path behavior, especially in the presence of singular points. Most algorithms [33, 35, 31, 30, 20, 13] measure the performance of the Newton iteration to guide the step size. More specifically, if Newton's method is converging quickly with each step in parameter, the step length is relaxed. If it's converging slowly (often this indicates an approaching singular point), then one shrinks the step size by some amount (usually a half), and starts the Newton iteration over. This process is known as *step size control* and is implemented in [31, 30, 33, 9]. The step size reduction factor and the number of Newton iterations that differentiates "bad" from "good" convergence is

usually left to the user's specification.

We have motivated PAC as an improvement on natural parameterization for solution paths containing certain types of singular points. The problem with `npalg` arises when the Jacobian matrix G_u becomes singular for some pair (u_0, λ_0). In order for `paalg` to truly be an improvement under this condition, the Jacobian matrix associated with the reparameterized problem (3.3) (F_x) must be nonsingular. If it is, then we can apply Corollary 3.1 to the problem (3.3) with x taking the place of u and s taking the place of λ. At this time, we have made no statement that assures F_x stays nonsingular even in the presence of a singular point. This will be addressed in the next chapter when we study a particular singular point of interest. For now, assuming F_x is indeed nonsingular for each (x, s), we have

Corollary 3.2. *Let F be Lipschitz continuously differentiable, $F(x_0, s_0) = 0$, and $F_x(x_0, s_0)$ be nonsingular. Then there is a $\delta > 0$, which depends only on $\|F_x^{-1}(x_0, s_0)\|$ and the Lipschitz constants of F_x and x, such that if $|s - s_0| < \delta$ then Newton's method with initial iterate s_0 converges q-quadratically to the solution $x(s)$ of $F(x, s) = 0$, i.e.,*

$$\|x_{n+1} - x(s)\| = O(\|x_n - x(s)\|^2),$$

where, for $n \geq 0$,

$$x_{n+1} = x_n - F_x(x_n, s)^{-1} F(x_n, s).$$

We noted in §3.1 that Corollary 3.1 implies the parameter step size $d\lambda$ in Algorithm `npalg`. In particular, the corollary states that $d\lambda$ depends explicitly on the Lipschitz constants of G_u and u and $\|G_u^{-1}(u_0, \lambda_0)\|$. Similarly, Corollary 3.2 implies the step

length ds for Algorithm `paalg`. Now, however, a proper step size depends on the Lipschitz constants of F_x and x and $\|F_x^{-1}(x_0, s_0)\|$. We can see from the dependence on $\|F_x^{-1}(x_0, s_0)\|$ that an upper bound on this quantity is crucial in order for `paalg` to succeed. This is exactly the motivation for the new analysis presented in the next chapter. In Chapter 4, we present a singular point of interest and describe solution paths to $G(u, \lambda) = 0$ that contain such singular points. Our analysis will give an upper bound on $\|F_x^{-1}(x_0, s_0)\|$ which is important for choosing a proper step size ds in pseudo-arclength continuation as stated by Corollary 3.2.

Chapter 4

Numerical Continuation for Simple Folds

In Chapter 3 we suggested that, for certain types of solution paths, pseudo-arclength continuation (PAC, §3.2) is preferable to natural parameterization (§3.1). In this chapter, we consider solution paths to $G(u, \lambda) = 0$ that contain *regular points* and *simple folds* which we define shortly. First in §4.1, we present two examples of simple folds. Next in §4.2, we present the main result of this thesis which is a new characterization of simple folds in terms of the singular value decomposition (SVD) of G_u. From the SVD, we derive a new bound for $\|F_x^{-1}\|$ as motivated by Corollary 3.2 (and corroborate the nonsingularity of F_x for paths containing regular points and simple folds). As a byproduct, we obtain a refinement of Weyl's monotonicity theorem [29] for the smallest eigenvalue of a symmetric positive semi-definite matrix (§4.2.1).

In this chapter we are interested in solution paths to $G(u, \lambda)$ that contain only

regular points [16, 13] and simple fold singularities [16].

Definition 4.1. *A solution* (u_0, λ_0) *of* $G(u, \lambda) = 0$ *is a* **regular point** *if* $G(u_0, \lambda_0)$ *is nonsingular.*

Definition 4.2. *A solution* (u_0, λ_0) *of* $G(u, \lambda) = 0$ *is a* **simple fold** *if*

- $dim(\mathcal{N}(G_u(u_0, \lambda_0))) = 1$ *and*
- $G_\lambda(u_0, \lambda_0) \notin \mathcal{R}(G_u(u_0, \lambda_0))$

where $\mathcal{N}$ is the nullspace of a matrix defined by $\mathcal{N}(A) = \{x \in \Re^N \mid Ax = 0, A \in \Re^N \times \Re^N\}$ and $\mathcal{R}$ is the range of a matrix defined by $\mathcal{R}(A) = \{Ax \mid A \in \Re^N \times \Re^N, x \in \Re^N\}$. The dimension of a space is the number of vectors in any basis for that space and is denoted here by *dim*. Finally, recall that G_λ denotes differentiation with respect to λ.

Before we justify the use of PAC for paths containing the above points, we first present some examples of simple folds.

4.1 Simple Fold Examples

We now employ the pseudo-arclength continuation algorithm from §3.2 to follow solution paths containing regular points and simple fold singularities.

4.1.1 A Boundary Value Problem Simple Fold

Consider the boundary value problem proposed in [16]:

$$G(u, \lambda) \equiv u''(t) + \frac{u'(t)}{t} + u(t)^3 + \lambda = 0 \quad \text{in} \quad 0 < t < 1$$

subject to

$$u'(0) = 0, \quad u(1) = 0.$$

We discretize $t \in [0,1]$ by $t_i = i/h$, $i = 0, 1, \ldots, N-1$ where $h = 1/(N-1)$. We will denote $u(t_i) \equiv u_i$. We approximate the problem with finite difference schemes for the derivative terms to obtain

$$G_i(u_i, \lambda) = \frac{u_{i+2} - 2u_i + u_{i-2}}{4h^2} + \frac{u_{i+1} - u_{i-1}}{2h} \cdot \frac{1}{t_i} + u_i^3 + \lambda.$$

The Jacobian matrix is the $N \times N$ tridiagonal matrix whose superdiagonal, diagonal, and subdiagonal are made up of the coefficients of u_{i-1}, u_i, and u_{i+1}, respectively in the above formulation. Using $N = 20$, we employ algorithm `paalg` from §3.2 starting with the initial iterate $x = (u^T, \lambda)^T = 0_{(N+1)\times 1}$. We use a PLU factorization (§2.1.2) of the Jacobian matrix G_u to solve for the Newton step in each Newton iteration. Figure 4.1 shows the solution path traced out by PAC as a function of the parameter λ. The second coordinate axis is the scalar measure $f(u) = \sqrt{|u'(1)u(0)|}sgn(-u'(1))$ where $sgn(x)$ is -1 if $x < 0$, 0 if $x = 0$, and 1 if $x > 0$. There are two simple folds at approximately $\lambda = .5$ and $\lambda = -11.5$. Notice the "folding" behavior of the path at such points which gives this singularity its name.

4.1.2 Chandrasekhar H-Equation

The equation of interest for this example is called the Chandrasekhar H-equation [7, 17, 26] from radiative transfer theory:

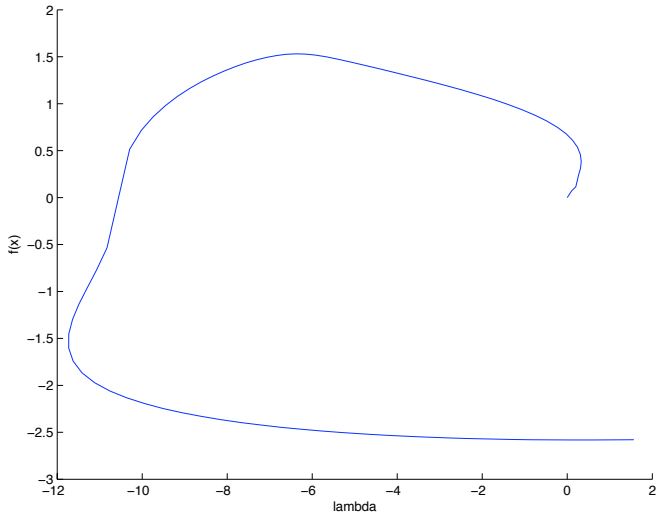


Figure 4.1: Two Simple Folds

$$G(H,c)(\mu) \equiv H(\mu) - \left(1 - \frac{c}{2}\int_0^1 H(\nu)\frac{\mu}{\mu+\nu}d\nu\right)^{-1} = 0. \tag{4.1}$$

We compute the l^1 norm of the solution to (4.1) for various natural parameter values c. That is, we compute

$$\|H\|_1 = \int_0^1 H(\nu)\,d\nu$$

as a function of c where now H takes on the role of u and c takes on the role of λ. It can be shown [18] that a simple fold occurs at $c = 1$. We demonstrate this in Figure 4.3. We discretize the integral with the composite midpoint rule and 200 nodes. Once again, we employ PAC as the continuation strategy with a PLU factorization of the Jacobian matrix G_H used in Newton's method. We will return to the H-equation in more detail in §4.3.

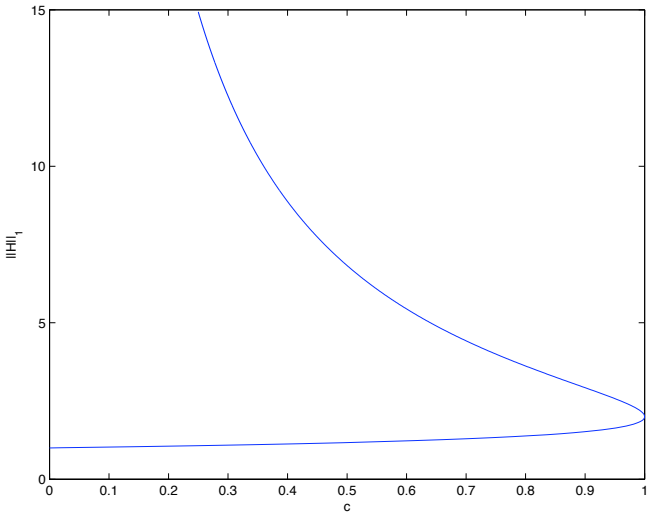


Figure 4.2: $\|H\|_1$ as a function of c

At simple folds, G_u is singular and natural parameterization fails for reasons discussed in Chapter 3. We dedicate the next section to understanding why pseudo-arclength continuation is indeed a remedy to natural parameterization for paths containing regular points and simple folds.

4.2 Upper Bound on $\|F_x^{-1}\|$

For a solution $x_0 = (u_0, \lambda_0)$ to $G(u, \lambda) = 0$, we present an upper bound on $\|F_x^{-1}(x_0, s_0)\|$ in the case that

- $G_u(u_0, \lambda_0)$ is nonsingular or
- (u_0, λ_0) is a simple fold of $G(u, \lambda) = 0$.

In order to derive the bound, we first introduce a new characterization of a simple fold which is based on the singular value decomposition of G_u. After doing this, we

present the main theorem, Theorem 4.1. Before proving the main theorem, however, we take a detour to §4.2.1 which presents a refinement of Weyl's monotonicity theorem [29] for the smallest eigenvalue of a rank-one update to a symmetric positive semi-definite matrix. We then return to the proof of our main theorem in §4.2.2 using this refinement as a tool. These results can be found in [8]. Throughout, we assume that G is a Lipschitz continuously differentiable function in some open subset Ω of $\Re^N$ and all norms are Euclidean.

Let

$$G_u(u, \lambda) = U \Sigma V^T$$

be a singular value decomposition (SVD) of $G_u(u, \lambda)$ where U and V are orthogonal matrices,

$$\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_N), \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N \quad , \qquad u_N \equiv U e_N,$$

and e_N is the last column of the $N \times N$ identity matrix. The expression $diag(\sigma_1, \sigma_2, \ldots, \sigma_N)$ is the $N \times N$ matrix with diagonal elements $\sigma_1, \sigma_2, \ldots, \sigma_N$ (starting in the top left corner) and zeros in the remaining entries. The trailing column u_N of U is a left singular vector associated with the smallest singular value σ_N. When necessary, we will make the dependence on λ or u, by writing, for example, $\sigma_N(u, \lambda)$ or $u_N(u, \lambda)$.

Since the singular values are continuous functions of the elements in $G_u(u, \lambda)$, they are also continuous in λ. If

$$\sigma_{N-1} \geq \bar{\sigma} > 0$$

for all $(u, \lambda) \in \Omega$, then the nullity of $G_u(u, \lambda)$ is at most one. If in addition $\sigma_N = 0$ then u_N spans the left nullspace of $G_u(u, \lambda)$. From the direct sum

$$\mathcal{N}(G_u^T(u_0, \lambda_0)) \oplus \mathcal{R}(G_u(u_0, \lambda_0)) = R^N$$

we see that $G_\lambda(u_0, \lambda_0)$ is not in $\mathcal{N}(G_u(u_0, \lambda_0))$ if and only if $G_\lambda(u_0, \lambda_0)^T u_N \neq 0$. Hence we have a new, equivalent definition of a simple fold.

Definition 4.3 (Simple Fold via SVD)**.** *Let (u_0, λ_0) be a solution of $G(u, \lambda) = 0$, and let $u_N(u_0, \lambda_0)$ be a left singular vector of $G_u(u_0, \lambda_0)$ associated with σ_N.*

Then (u_0, λ_0) is a **simple fold** *if*

- $\sigma_{N-1}(u_0, \lambda_0) > 0$,
- $dim(\mathcal{N}(G_u(u_0, \lambda_0))) = 1$, *and*
- $u_N(u_0, \lambda_0)^T G_\lambda(u_0, \lambda_0) \neq 0$.

We will use Definition 4.3 to motivate the assumptions in Theorem 4.1. Suppose (u_0, λ_0) is a regular point (G_u nonsingular) or a simple fold. Since G is Lipschitz continuously differentiable, we can, by requiring $u_N^T(\lambda_0) u_N(\lambda) > 0$ for example, define u_N as a continuous function of u and λ. Hence $G_\lambda(u, \lambda)^T u_N(u, \lambda)$ is a continuous function of (u, λ). So there is $\alpha > 0$ such that for all (u, λ) sufficiently near (u_0, λ_0),

$$\max\left(\sigma_N(u, \lambda)^2, |u_N(u, \lambda)^T G_\lambda(u, \lambda)|^2 \frac{\text{gap}}{\text{gap} + \xi^2}\right) \geq \alpha > 0, \tag{4.2}$$

where

$$\text{gap} \equiv \sigma_{N-1}(u, \lambda)^2 - \sigma_N(u, \lambda)^2 \tag{4.3}$$

and

$$\xi \equiv |u_N((u,\lambda)^T G_\lambda(u,\lambda)| + \|(I - u_N(u,\lambda)u_N(u,\lambda)^T)G_\lambda(u,\lambda)\|. \tag{4.4}$$

Equation (4.2) is a way to quantify the statement that all points on a solution arc are either regular points or simple folds by saying that either $\sigma_N > 0$ (regular point) or the conditions in Definition 4.3 hold.

The main result of this thesis is

Theorem 4.1. *Let $\bar{\Omega}$ be the closure of an open subset $\Omega \in R^{N+1}$, and let G be continuously differentiable in $\bar{\Omega}$. Let $x_0 = (u_0, \lambda_0)$ in $\bar{\Omega}$ be a solution to $G(u_0, \lambda_0) = 0$, and $\mathcal{N}(x_0, s_0) = 0$ with $\|\dot{x}_0\| = 1$. Let $\tau \geq 0$ be such that $\|G_u(u_0, \lambda_0)\dot{u}_0 + G_\lambda(u_0, \lambda_0)\dot{\lambda}_0\| \leq \tau$.*

Assume that for all (u, λ) in $\bar{\Omega}$ there exists $\alpha > 0$ such that

$$\sigma_{N-1}(u,\lambda) > 0 \text{ and } \max\left(\sigma_N(u,\lambda)^2, |u_N(u,\lambda)^T G_\lambda(u,\lambda)|^2 \frac{\text{gap}}{\text{gap} + \xi^2}\right) \geq \alpha,$$

where gap *and ξ are defined by* (4.3) *and* (4.4).

If $\tau < \min\{\alpha, 1\}$, then for all $x = (u, \lambda)$ in $\bar{\Omega}$, the smallest singular value $\sigma_{\min}(F_x)$ of the Jacobian F_x of $F(x, s)$ is bounded from below with

$$\sigma_{\min}(F_x) \geq \sqrt{1 - \tau \max\left\{\frac{1}{\alpha}, 1\right\}}.$$

We postpone the proof of Theorem 4.1 until §4.2.2 in order to derive an auxiliary result first.

4.2.1 Lower Bound for the Smallest Eigenvalue

We derive a lower bound for the smallest eigenvalue of the rank-one update $A + yy^T$, where A is a real symmetric positive semi-definite matrix of order N, and y is a real $N \times 1$ vector.

Let $\beta_1 \geq \ldots \geq \beta_N \geq 0$ be the eigenvalues of A. Weyl's monotonicity theorem [29, Theorem (10.3.1)] implies bounds for the smallest eigenvalue of $A + yy^T$:

$$\beta_N \leq \lambda_{\min}(A + yy^T) \leq \beta_{N-1}.$$

Intuitively one would expect that $\lambda_{\min}(A + yy^T)$ is larger if y is close to an eigenvector of β_N. We confirm this by deriving lower bounds for $\lambda_{\min}(A + yy^T)$ that incorporate the contribution of y in the eigenspace of β_N. Recent related work on this topic can be found in [15].

Theorem 4.2. *Let A be an $N \times N$ real symmetric positive semi-definite matrix, u_N an eigenvector of A associated with β_N, $\|u_N\| = 1$, and $y \neq 0$ a real $N \times 1$ vector. Set $y_N \equiv u_N^T y$. Then*

$$\lambda_{\min}(A + yy^T) \geq \max\left\{\beta_N,\, y_N^2\, \frac{\text{gap}}{\text{gap} + \xi^2}\right\} \tag{4.5}$$

where $\text{gap} \equiv \beta_{N-1} - \beta_N$ *and* $\xi \equiv |y_N| + \sqrt{\|y\|^2 - y_N^2}$.

Proof. We first show that

$$\lambda_{\min}(A + yy^T) \geq \min\{\beta_N + y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2},\, \beta_{N-1} \frac{y_N^2}{\xi^2}\} \tag{4.6}$$

is a lower bound for $\lambda_{\min}(A + yy^T) = \min_{\|x\|=1} x^T(A + yy^T)x$.

Let

$$A = U \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_N \end{pmatrix} U^T$$

be an eigendecomposition of A, and x be any real vector with $\|x\| = 1$. Partition

$$U^T x = \begin{pmatrix} \bar{x} \\ x_N \end{pmatrix}, \qquad U^T y = \begin{pmatrix} \bar{y} \\ y_N \end{pmatrix}$$

so that $\xi = |y_N| + \|\bar{y}\|$. Then

$$x^T(A + yy^T)x \geq \beta_{N-1}\|\bar{x}\|^2 + \beta_N x_N^2 + (y^T x)^2.$$

If $\|\bar{x}\| \geq |y_N|/\xi$ then

$$x^T(A + yy^T)x \geq (\beta_{N-1} y_N^2)/\xi^2,$$

which proves the second part of the bound in (4.6).

If $\|\bar{x}\| < |y_N|/\xi$ then $|y_N| - \|\bar{x}\|\xi > 0$, and it makes sense to use $|x_N| \geq 1 - \|\bar{x}\|$ in

$$|y^T x| = |y_N x_N + \bar{y}^T \bar{x}| \geq |y_N x_N| - \|\bar{x}\|\|\bar{y}\| \geq |y_N| - \|\bar{x}\|\xi.$$

Hence

$$x^T(A + yy^T)x \geq \beta_{N-1}\|\bar{x}\|^2 + \beta_N x_N^2 + (y^T x)^2 \geq \beta_N + y_N^2 + (\text{gap} + \xi^2)\|\bar{x}\|^2 - 2\xi\|\bar{x}\||y_N|.$$

This is a function of $\|\bar{x}\|$ which has a minimum at $\|\bar{x}\| = |y_N|\xi/(\text{gap} + \xi^2)$. Hence

$$x^T(A + yy^T)x \geq \beta_N + y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2},$$

which proves the first part of the bound in (4.6). With the help of (4.6) we now show the desired bound (4.5). Weyl's theorem [29, Theorem (10.3.1)] implies $\lambda_{\min}(A+yy^T) \geq \beta_N$, which proves the first part of the bound in (4.5). For the second part of the bound in (4.5), we use the fact that the eigenvalues of A are non-negative, hence $\beta_{N-1} \geq \text{gap}$ and

$$\frac{\beta_{N-1}}{\xi^2} \geq \frac{\text{gap}}{\text{gap} + \xi^2}.$$

Substituting this into (4.6) gives the second part of the bound in (4.5)

$$\begin{aligned} \min(A + yy^T) &\geq \min\{\beta_N + y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2},\ y_N^2 \frac{\beta_{N-1}}{\xi^2}\} \\ &\geq \min\{\beta_N + y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2},\ y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2}\} = y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2}. \end{aligned}$$

□

The quantity gap in Theorem 4.2 is the absolute gap between the smallest and next smallest eigenvalues, and ξ is an approximation for $\|y\|$ since $\|y\| \leq \xi \leq \sqrt{2}\|y\|$. The theorem shows that $\lambda_{\min}(A + yy^T)$ is likely to be larger if y has a substantial contribution in the eigenspace of β_N.

Now we are in a position to complete the proof of Theorem 4.1.

4.2.2 Proof of Main Theorem 4.1

Define the residual

$$r \equiv G_u(u_0, \lambda_0)\dot{u}_0 + G_\lambda(u_0, \lambda_0)\dot{\lambda}_0.$$

Letting $G_u = G_u(u, \lambda)$, $G_\lambda = G_\lambda(u, \lambda)$, and $F_x = F_x(x, s)$, we have

$$F_x F_x^T = \begin{pmatrix} G_u & G_\lambda \\ \dot{u}_0^T & \dot{\lambda}_0 \end{pmatrix} \begin{pmatrix} G_u^T & \dot{u}_0 \\ G_\lambda^T & \dot{\lambda}_0 \end{pmatrix} = \begin{pmatrix} G_u G_u^T + G_\lambda G_\lambda^T & r \\ r^T & 1 \end{pmatrix}.$$

Note that the 1 entry in the above matrix comes from the arclength property (3.2). The eigenvalues of $F_x F_x^T$ are the squares of the singular values of F_x. Applying Theorem 4.2 to $G_u G_u^T + G_\lambda G_\lambda^T$ with $A = G_u G_u^T$, $y = G_\lambda$, $\beta_N = \sigma_N(u, \lambda)^2$, $\beta_{N-1} = \sigma_{N-1}^2(u, \lambda)$ and gap $= \sigma_{N-1}(u, \lambda)^2 - \sigma_N(u, \lambda)^2$ shows $\lambda_{\min}(G_u G_u^T + G_\lambda G_\lambda^T) \geq \alpha$. Hence we can write

$$\begin{pmatrix} G_u G_u^T + G_\lambda G_\lambda^T & 0 \\ 0 & 1 \end{pmatrix}^{-1} F_x F_x^T = I + E,$$

where $\|E\| \leq \tau \max\left\{\frac{1}{\alpha}, 1\right\}$. If $\tau < \min\{\alpha, 1\}$ then $\|E\| < 1$, $I + E$ is nonsingular, and

$$\frac{1}{\|(F_x F_x^T)^{-1}\|} \geq 1 - \tau \max\left\{\frac{1}{\alpha}, 1\right\}.$$

4.3 Return to the Chandrasekhar H-Equation

In §4.1, we introduced the Chandrasekhar H-equation [7, 17, 26] as a problem with a solution path containing regular points and a simple fold. We now revisit this ex-

ample with a slightly different motivation. While we will restate the problem, the emphasis here is that the smallest singular value of the Jacobian matrix associated with pseudo-arclength continuation ($F_{(H,c)}$) stays bounded away from zero from below. We show that in contrast, the smallest singular value of the Jacobian matrix associated with natural parameterization (G_H) is near zero at the singularity. Further, using GMRES for the linear solve in Newton's method, we compare convergence results for pseudo-arclength continuation versus natural parameterization. These results combined support the conclusions from our main theorem, Theorem 4.1.

The H-equation is given by

$$H(\mu) - \left(1 - \frac{c}{2}\int_0^1 H(\nu)\frac{\mu}{\mu+\nu}d\nu\right)^{-1} = 0. \tag{4.7}$$

The goal is to compute $H(\mu)$ for $\mu \in [0,1]$ as a function of c. There is a simple fold at $c = 1$ [26] which is also the case for any discretization of the equation which uses a quadrature rule that integrates constant functions exactly.

In this section we use a Newton-GMRES version of pseudo-arclength continuation [11], fixing the step in arclength to $ds = .02$, using a secant predictor [16], and beginning the continuation at $c = 0$, where $H = 1$ is the solution. The vector with components all equal to one is the solution of the discrete problem as well. We discretize the integral with the composite midpoint rule using 200 nodes. A consequence of this discretization is that all scalar products of discretized functions in the GMRES solves were scaled by $1/200$.

Figure 4.3 is a plot of $\|H\|_1$ against c. A fixed value of ds, as we use here, causes problems as the L^1 norm of H increases. The reason for this is that the solution

develops very large derivatives, and the predictor becomes very poor. We stopped the continuation at $c = .3$ on the upper branch for that reason.

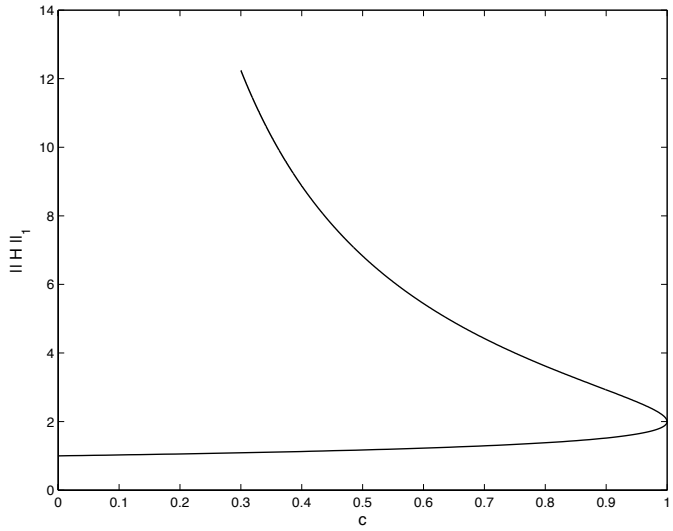


Figure 4.3: $\|H\|_1$ as a function of c

For this example we can also compute the L^1 norm as a function of c analytically, and verify the results in Figure 4.3. We can rewrite (4.7) as

$$H(\mu) = 1 + \frac{c}{2}\int_0^1 H(\mu)H(\nu)\frac{\mu}{\mu+\nu}d\nu. \tag{4.8}$$

Integrating (4.8) with respect to μ yields

$$\|H\|_1 = 1 + \frac{c}{2}\int_0^1\int_0^1 H(\mu)H(\nu)\frac{\mu}{\mu+\nu}d\mu d\nu = 1 + \frac{c}{4}\|H\|_1^2,$$

and so

$$\|H\|_1 = \frac{1 \pm \sqrt{1-c}}{c/2}. \tag{4.9}$$

As a demonstration of the result in Theorem 4.1, we calculate the smallest singular value of the Jacobian matrix associated with the augmented system for the H-equation with each continuation iteration. We used the MATLAB `svds` command for this. In the language of § 4.2.1, we find $\sigma_{\min}(F_{(H,c)})$ for various c where $F_{(H,c)}$ denotes the Jacobian of $\begin{pmatrix} G(H,c) \\ \mathcal{N}(H,c,s) \end{pmatrix}$. Figure 4.4 shows that the smallest singular value of $F_{(H,c)}$ for each c stays away from zero keeping $F_{(H,c)}$ nonsingular, even at the simple fold ($c = 1$).

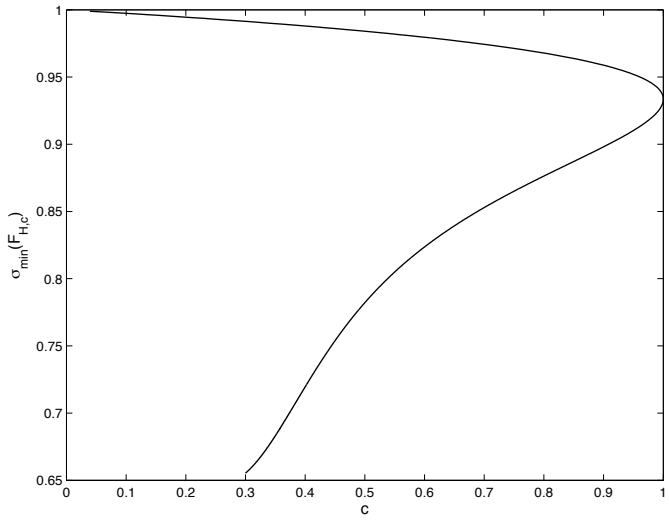


Figure 4.4: $\sigma_{\min}(F_{(H,c)})$ as a function of c

It is interesting to compare Figure 4.4 with a plot of the smallest singular value of G_H, which we can also compute on the path. In Figure 4.5, one can see the singularity at $c = 1$.

In Figure 4.6 we plot the average number of GMRES iterations per Newton iteration as a function of c. The lower curve corresponds to the continuation from $c = 0$ to $c = 1$, and the upper from $c = 1$ to $c = .3$. The computation in this figure was done for pseudo-

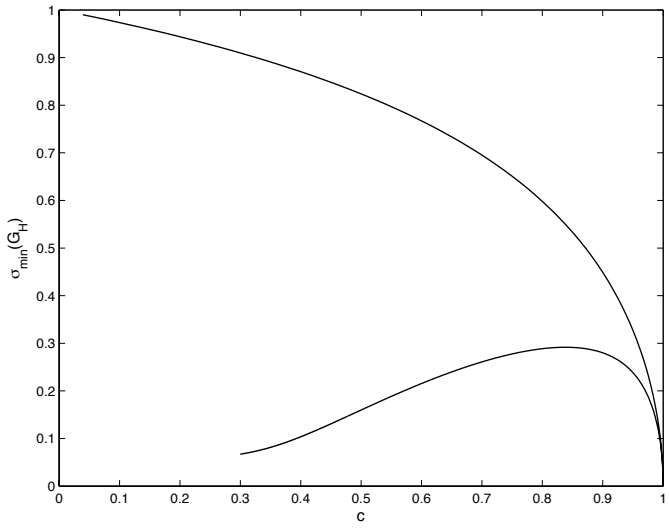


Figure 4.5: $\sigma_{\min}(G_H)$ as a function of c

arclength continuation, which we compare with parameter continuation in Figure 4.7. As one moves further on the path, the predictor becomes less effective, and the number of Newton iterations increases. The predictor is also different for the first two points on the path, because we do not have the data we need to build the secant predictor before we have computed two points. The initial point for $c = 0$ is the vector with 1 in each component, which is the solution, so the plots begin with the first non-zero value of c.

Figure 4.7 is the result of a simple parameter continuation for each of the upper and lower branches. The lower curve is for values of $c \in [0, .9]$, where the problem is quite easy. The linear solver takes fewer GMRES iterations per Newton iteration on this branch, and we observe that the difference in linear iterations from the lower branch in Figure 4.6 is at most 1, consistent with the theory. On the upper branch, where $c \in [.9, .3]$, the performance of parameter continuation is significantly worse than that

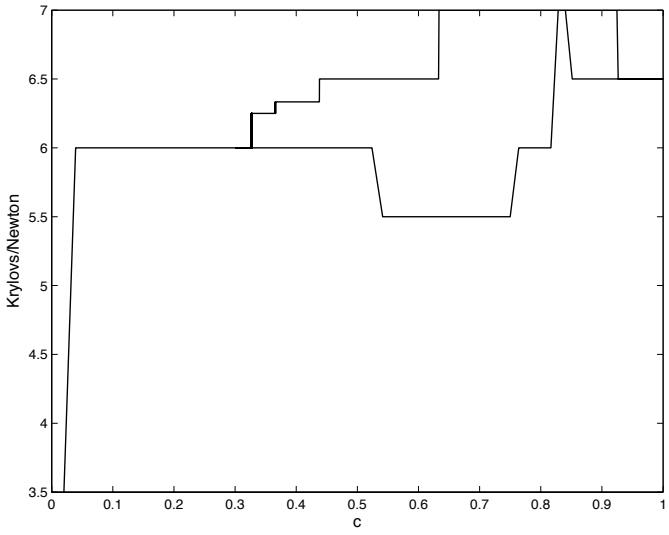


Figure 4.6: Krylov's per Newton: Pseudo-Arclength Continuation

of pseudo-arclength continuation, and the linear solver performs significantly less well in the parameter continuation solver. This is consistent with the singular value results in Figure 4.5.

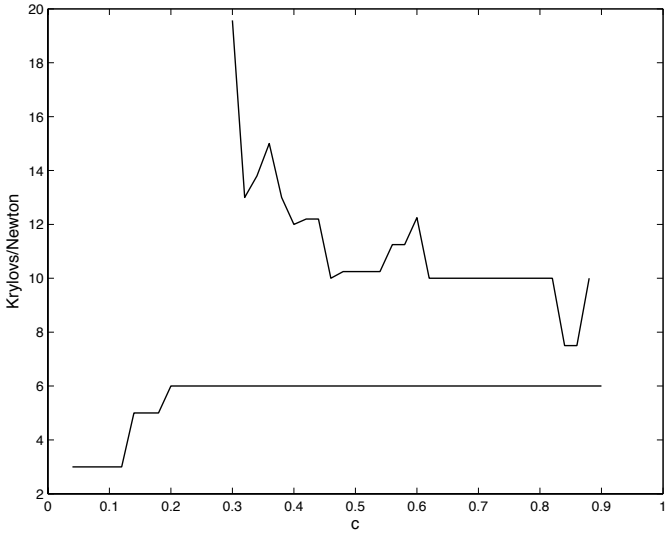


Figure 4.7: Krylov's per Newton, Parameter Continuation in c

Chapter 5

Continuation Applied to Atomic and Molecular Fluids

5.1 Molecular Dynamics and Integral Equation Theory

Many physical systems that surround us are best described on an atomic or molecular level. In fact, American physicist Richard Feynman said in 1963, "Everything that living things do can be understood in terms of the jiggling and wiggling of atoms." Two ways of analyzing molecular properties and interactions are known as molecular dynamics and integral equation theory.

Molecular dynamics is an atomistic simulation of molecules and their interactions in time. The input into a molecular dynamics code is simply Newton's laws of motion, any forces involved, and some properties of the atoms being studied, such as atomic mass,

charge, etc. Then Newton's laws, in the form of ordinary differential equations, are solved using some integrator. The results of the simulation are positions and velocities of the atoms as well as some thermodynamic properties such as heat capacity, pressure, compressibility, or chemical potential. Simulation is a very useful and accessible means of analyzing molecular systems. The down side to simulation is that it can be very time-consuming and costly depending on the system of consideration.

However, often one is interested in molecular structure and thermodynamic properties of a system at thermodynamic equilibrium (time independent). Integral equation theorists seek to achieve accurate structural and thermodynamic information with less computation than simulation. This thesis focuses on improving integral equation theory to recover fluid structure and thermodynamic properties through analyzing *closure equations*. In particular, we will focus on a new closure equation [24] recently developed by Marcelo Marucho and B.M. Pettitt of the University of Houston's Department of Chemistry.

In the next section, we will explain the terms "structure" and "thermodynamic properties" by considering *atomic* and *molecular* fluids. We will introduce integral equation theory and closure equations by first considering atomic fluids and then later extending the theory to molecular fluids. Along the way, we will point out some crucial deficiencies in existing equations found in the literature. This will motivate the need for improvements in integral equation theory, and in particular, the one we present in this thesis based on the works of Marucho and Pettitt [24]. We then present the main numerical continuation problems of interest related to both the new and existing fluid theory. We discuss in detail new implementations for these continuation studies

which employ software developed at Sandia National Laboratories. The purpose of these studies is to identify structural and thermodynamic properties of a fluid more accurately and computationally efficient than has been previously done. In the case of molecular fluids, the theory and technology presented in this thesis may reveal physics that is currently unknown. This is the ultimate goal of this chapter and its advancements.

This thesis presents new results regarding continuation for atomic and molecular fluids. Specifically, in §5.3 we present

- a new closure equation developed by Marucho and Pettitt [24] for atomic fluids,
- a new continuation problem formulation for this closure developed by Andrew Salinger of Sandia National Labs, and
- new continuation results (Dickson) using the new closure equation and problem formulation for atomic fluids.

The new results for atomic fluids presented in §5.3.4 that employ the above closure equation and problem formulation include

- continuation studies for particular isothermal parameters,
- a verification of correctness for these continuation studies, and
- evidence that our continuation codes can recover physically relevant solution paths (not always easy [4, 37]) while bypassing irrelevant "no solution zones".

In §5.4, we present

- a new closure equation developed by Marucho and Pettitt [24] for diatomic molecular fluids,
- a new continuation problem formulation for this closure developed by Salinger, and
- a new preliminary continuation study (Dickson) for Chlorine-like molecular fluids employing the closure equation and problem formulation just mentioned.

Currently, we are working to improve the molecular code from above to potentially observe new physics for some diatomic molecular fluids. In the future, we hope to extend this code to more complicated fluids, such as water, to observe new, more accurate physical properties than have been previously discovered.

5.2 Introduction to Atomic and Molecular Fluids

This section defines some terminology associated with fluids [14] that we will use throughout this chapter. We will elaborate on these ideas in subsequent sections.

As mentioned earlier, integral equation theory is a route to uncovering structural and thermodynamic properties of fluids. The two general categorizations of fluids we will discuss are *atomic fluids* and *molecular fluids*. Atomic fluids are ones that are comprised of only single atoms. For example, a fluid that is made up of only chlorine atoms (Cl on the periodic table of elements) is an atomic fluid. Molecular fluids on the other hand are made up of molecules. Molecules are just groups of atoms bonded together. So for instance hydrogen chloride (HCl) is a molecule with one hydrogen

atom (H) bonded to one chloride atom (Cl). In particular, this is called a diatomic molecule because it is made up of only two atoms bonded together. Water, H_2O, is a triatomic molecule because it is made up of three atoms (two hydrogens and an oxygen).

We will consider homogeneous atomic and molecular fluids in this thesis, that is, fluids that only contain one kind of atom or one kind of molecule. For molecular fluids, we will only consider diatomic fluids in detail. The general theory, however, can be extended to n-atomic molecules. When convenient, we will use the term "particle" to refer to either an atom or molecule.

We have mentioned that integral equation theory is useful for finding structural and thermodynamic properties of particle fluids. In particular, it is useful for finding *pair correlation functions* [3] which we will denote $g(r)$. Here, r represents the distance between two particles in a fluid. The pair correlation function, $g(r)$, is the relative probability of finding a particle at a distance r from the reference particle ($r = 0$) compared to that same probability for an uncorrelated fluid. Figure 5.1 is an example of what a pair correlation function may look like. A general characterization of a correlation function is that the closer two particles are (small r), the higher their correlation. Likewise, the farther the particles (large r), the smaller their correlation. Understanding how particles in a fluid are correlated is known as the "structure" of a fluid.

Additionally, it is often of great importance to know thermodynamic properties associated with a fluid. We can relate properties of atoms and molecules on a microscopic level to properties of materials that we observe in everyday life on a macroscopic level

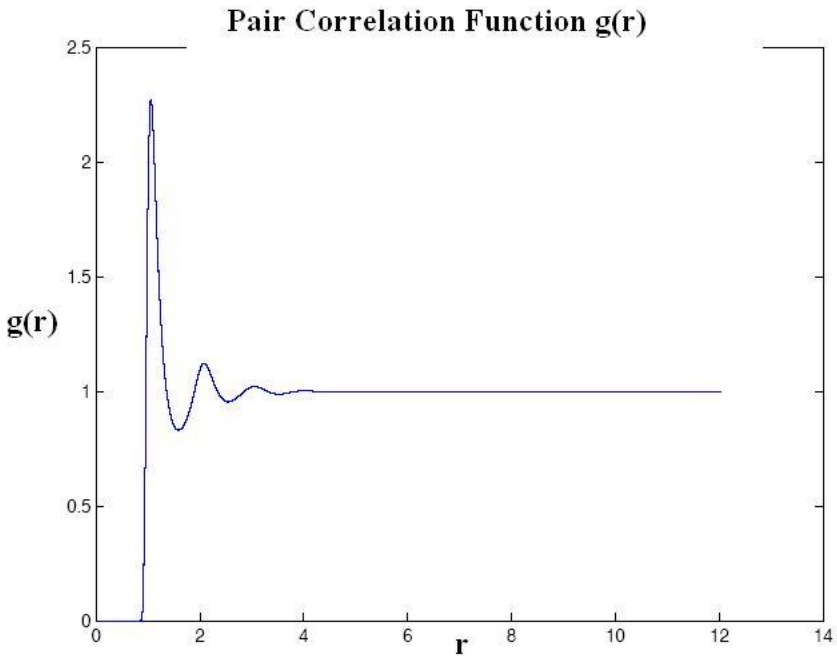


Figure 5.1: Pair correlation function $g(r)$ for a particular density and temperature

through thermodynamics. Thermodynamics are the effects of changes in temperature, pressure, chemical potential, compressiblity, etc. on a physical system. Many thermodynamic quantities for fluids are functions of the pair correlation function, $g(r)$. That is, $g(r)$ gives rise to many thermodynamic properties of interest.

To understand the equations that describe particle fluids, we need to discuss a few more types of correlation functions. The first is called the *direct correlation function* which we will denote $c(r)$. The function $c(r)$ represents the direct correlation between two particles in a fluid. This is opposed to $t(r)$, called the *indirect correlation function*, which also describes the correlation between two particles in a fluid, but takes into consideration that the two particles are surrounded by other particles of the same type. That is, $t(r)$ measures correlations due to the presence of other particles in the

fluid. Then the *total correlation function*, denoted $h(r)$, is the direct correlation plus indirect correlation. In other words, $h(r) = c(r) + t(r)$. Finally, the pair correlation function is given by $g(r) = h(r) + 1$.

With a little terminology under our belts, we now are ready to introduce the underlying integral equation for atomic fluids.

5.3 Atomic Fluids

One way to recover structural, and therefore thermodynamic, properties of atomic fluids is through the Ornstein-Zernike (OZ) equation [14, 3].

5.3.1 Integral Equation Theory for Atomic Fluids

The OZ equation is an integral equation with two unknowns, namely, the total correlation function $h(r)$ and the direct correlation function $c(r)$ discussed in §5.2. The OZ equation is given by

$$h(r) = c(r) + \rho(h * c)(r) \tag{5.1}$$

where

$$(h * c)(r) = \int_{\Re^3} c(\|\mathbf{r} - \mathbf{r}'\|)h(\|\mathbf{r}'\|)d\mathbf{r}'.$$

Here,

- $\mathbf{r} \in \Re^3$,
- $r = \|\mathbf{r}\|$ is the Euclidean distance of $\mathbf{r}$ from the origin,

- ρ is density, and
- $h(r), c(r) \in C[0, \infty)$ are the unknowns.

Let's try to gain a bit of intuition regarding OZ's formulation. Recall that the total correlation $h(r)$ is given by the equation $h(r) = c(r) + t(r)$ where $c(r)$ is the direct correlation between two atoms in a fluid and $t(r)$ is the indirect correlation that takes into account the presence of nearby particles. As one can see from (5.1), the indirect correlation $t(r)$ is given by $t(r) = \rho(h * c)(r)$. This is because the indirect correlation can be thought of as a sum (integral!) over all possible paths between one atom and another. In this case "path" means the contribution to the correlation from neighboring atoms in the fluid. You can think of it as a road map. There are lots of ways to get from point A to point B. You can get there directly, that is, a straight line from A to B (direct correlation $c(r)$). But you can still get there, even if you have to make an intermediate stop at point C. That is, you can get from A to B by taking the A to C to B route. It's indirect, but it is still a way to get from A to B. Sticking to our analogy, this route would contribute to the indirect correlation function, $t(r)$. But you could also make more than one stop between points A and B. Perhaps you want to stop at points D and E and F too before reaching point B. All of these paths contribute to the indirect correlation. By now it is clear that the number of paths to consider is dependent on the number of stops one can make on the road map. If you're in a remote area, there may be only a few possible stops to make. If you're in a crowded city, maybe you have lots of possible routes to choose from. This is why the density of the fluid (like the number of possible stops given the area) is a factor in the indirect correlation term $t(r)$.

In order to resolve (5.1) and recover the pair correlation function $g(r)$, we need to solve for the two unknowns $h(r)$ and $c(r)$. In order to do this, we need one more equation that relates the two unknowns. This equation is called a *closure equation* or better yet, a *closure approximation*. Most closure equations involve an infinite sum of multidimensional integrals that are difficult to evaluate, aside from the fact there are infinitely many. Thus closure equations must be approximated, often by truncating this infinite sum in some way. While there are some cases in which common closure approximations yield an accurate pair correlation function, they can also result in thermodynamic inconsistencies. This occurs when two different formulations for a thermodynamic quantity produce two different answers when they should be the same. This is exactly the kind of inconsistency we wish to avoid since, as earlier discussed, accurate thermodynamic properties are important for understanding a fluid.

5.3.2 Closure Equations for Atomic Fluids

Here we define two common closure equations used to find the solutions $h(r)$ and $c(r)$ to the OZ equation (5.1). Then we present a new closure approximation developed in [24] for atomic fluids.

The hypernetted chain equation [14, 3] (HNC) is one popular closure approximation used to resolve (5.1). This equation, however, can produce inaccurate pair correlation functions and/or thermodynamic inconsistencies under certain conditions. The HNC approximation is given by

$$\exp(-u(r)/(TK_B) + h(r) - c(r)) - h(r) - 1 = 0, \quad 0 \leq r \leq \infty. \tag{5.2}$$

In this equation, $u(r)$ is called the pair potential between particles. Throughout this paper, we will consider only the Lennard-Jones potential which describes both an attractive and repulsive interaction between particles. It is given by

$$u(r) = 4\epsilon \left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right). \tag{5.3}$$

Additionally,

- T is absolute temperature (Kelvin),
- K_B is Boltzmann's constant (kcal/Kelvin),
- ϵ is the depth of the attractive interaction (kcal/mol), and
- σ is the width of the repulsive core (Angstroms).

Similar to the HNC closure is the Perkus-Yevick [14, 3] (PY) closure given by

$$\exp(-\beta u(r) + ln(h(r) - c(r) + 1)) - h(r) - 1 = 0. \tag{5.4}$$

At this time, we skip the computational details involved in solving OZ combined with HNC or PY. They can be found at the end of this chapter in §5.6.

As we briefly mentioned before, the HNC and PY closure approximations when paired with the OZ equation result in thermodynamic and structural inconsistencies under some conditions. Integral equation theorists test their theories by trying to match results ($g(r)$, some thermodynamic quantity, etc.) against that of molecular dynamics simulation. In Figure 5.2, we see four examples of different fluids and their

corresponding pair correlation function $g(r)$. In each of the four quadrants, two HNC-like theories like we described in (5.2) are used to approximate $g(r)$ and are represented by the solid and dash-doted lines. The dotted line represents the pair correlation obtained by molecular dynamics simulation. Ideally, the integral equation theories would match very closely to the MD result, but it is clear for these fluids that the HNC-like theories do a poor job of predicting the pair correlation function. Note that we say "HNC-*like*" theories here because we do not yet have the background needed to fully describe them. We will discuss these theories in more detail in §5.4, but for now, they serve as motivation for new, more accurate closure equations.

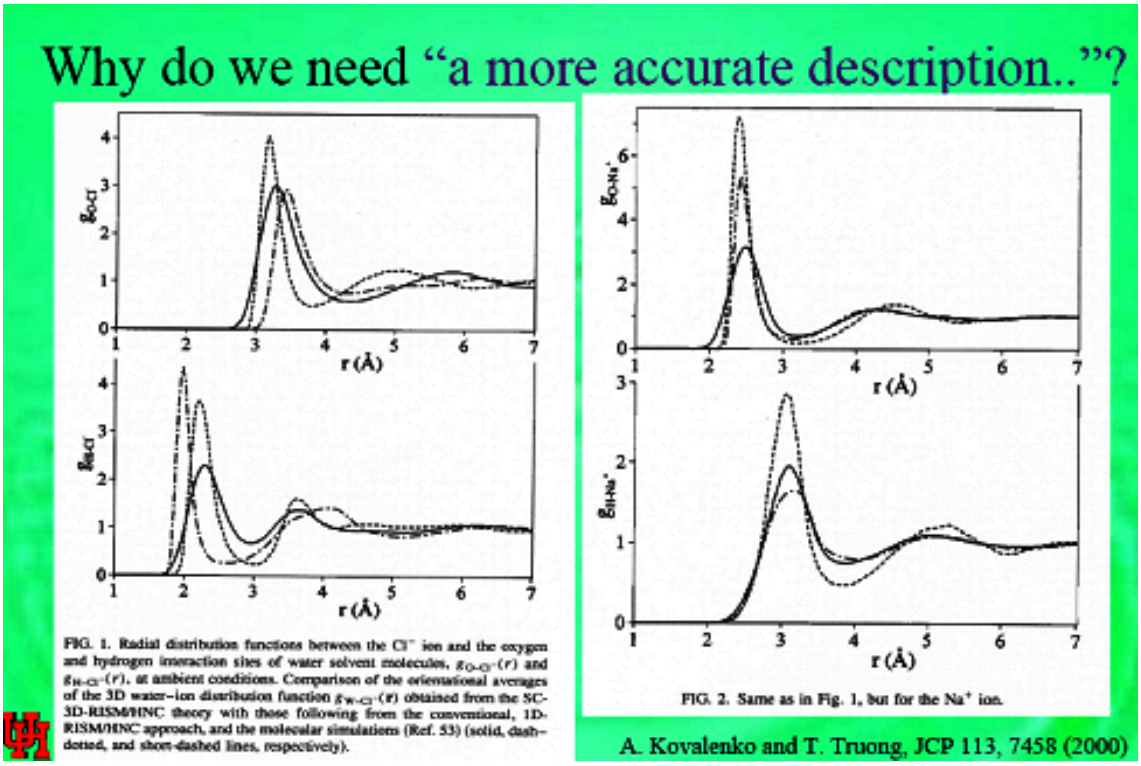


Figure 5.2: HNC-like theories poorly predicting pair correlations from MD simulations, A. Kovalenko and T. Truong, JPC 113, 7458 (2000)

Due to the structural and thermodynamic inconsistencies resulting from the HNC and PY closure equations, Marcelo Marucho and B.M. Pettitt of the University of Houston have developed a new closure equation that "interpolates" between the HNC and PY closures [24]. The new interpolating closure for atomic fluids, which we will denote IC, is given by

$$\exp(-\beta u(r))\left(-a+(a+1)\exp\left(\frac{h(r)-c(r)}{a+1}\right)\right)-h(r)-1=0 \tag{5.5}$$

where a is chosen to minimize excess chemical potential (free energy), $\mu(h(r), c(r), a)$ which we will define momentarily. Note that when $a = 0$, IC reduces to HNC. Additionally, when $a \to \infty$, the IC equation becomes the PY equation. Thus we can think of IC as "interpolating" between HNC and PY in a way that chooses a to minimize free energy. The idea of minimizing free energy comes from first principles in physics.

The excess chemical potential μ can be approximated with the formula [24]

$$\beta\mu(h(r), c(r), a) \approx -\rho \int_{\Re^3} \{h(r) - (h(r) - c(r))\mathcal{I}(h(r), c(r), a)\}\mathbf{dr} \tag{5.6}$$

where

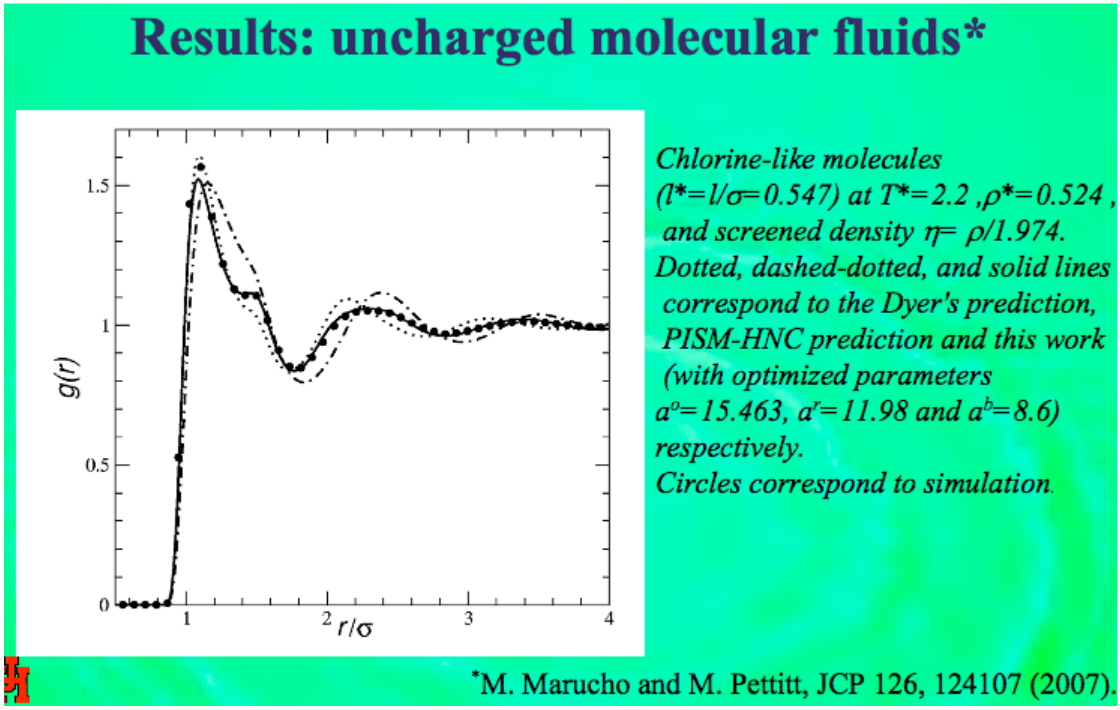
$$\mathcal{I}(h(r), c(r), a) = \{[h(r)+1]ln[y(r)/a] + ln[a] +$$
$$[h(r)(a+1)/(h(r)-c(r))]Re[li_2(y(r)/a+1) - li_2((a+1)/a)]\}/(h(r)-c(r)).$$

Here, li_2 denotes the dilogarithm function [1] and $y(r) = -a + (a+1)\exp[(h(r) - c(r))/(a+1)]$.

Marucho and Pettitt implement IC together with OZ for atomic fluids in [24] where

they further justify this closure from a scientific and mathematical perspective. One point to emphasize from [24] is that the idea of "interpolating" closures is not new. The way the interpolation parameter a is found through enforcing the physical property of minimizing free energy, is new however. The findings from [24] display more accurate structural and thermodynamic properties than the HNC-like closures for some fluids. Figure 5.3 shows pair correlation function approximations $g(r)$ from the interpolating closure IC (5.5) and HNC-like closures. The dots represent the molecular dynamics simulation and the solid line represents the prediction using the IC closure. It is apparent that for this particular fluid, the IC closure produces a better, more accurate approximation to $g(r)$ than the other existing theories.

Recall the OZ equation (5.1) depends on the fluid density ρ and the closure equations depend on temperature, T. When we use the phrase "under certain conditions", we are referring to the values of density and temperature. So far in Figures 5.2 and 5.3 we have been considering particular values of density and temperature. In this case, one just plugs the density and temperature values of interest into the OZ equation and closure equation and solves for the pair correlation function and some thermodynamic properties. However, often one is interested in how the structure or thermodynamics of the fluid changes as the density and temperature change. At this point, we see the makings of a numerical continuation problem introduced in Chapter 3. We now discuss numerical continuation and its application to the OZ equation.



Dots --> Simulation, Solid line --> Interpolating closure

Figure 5.3: Interpolating closure (solid line) predicts MD simulation (dots) better than other closures, M. Marucho and M. Pettitt, JCP 126, 124107 (2007)

5.3.3 Numerical Continuation for the OZ Equation

As previously mentioned, in order to solve the OZ equation together with a closure relation, one must specify parameter values for density and temperature. Here, we focus on the density parameter ρ. We can get a sense of how solution behavior (correlation functions and thermodynamic properties) changes as density values vary using the

numerical continuation techniques described in Chapter 3.

Consider once again our main equation

$$G(u, \lambda) = 0 \tag{5.7}$$

where $u \in \Re^N$ is the unknown and $\lambda \in \Re$ is a real number parameter. We want to find solutions u corresponding to various values of λ. In the context of the OZ equation, we can think of correlation functions $h(r)$ and $c(r)$ as the unknown u and the parameter ρ as λ. G is then like the OZ equation paired with the closure approximation. We would like to understand what happens to the solutions $h(r)$ and $c(r)$ as the parameter ρ varies. In particular, the goal is to do this using the interpolating closure (5.5) and compare it to solution paths obtained from the HNC or PY closures. Before we discuss implementing the continuation problem, we make one final point regarding continuation with the HNC and PY closures.

In addition to structural and thermodynamic inconsistencies, the continuation problem we have just presented is ill-posed with regard to the HNC and PY equations. In 2006, Beardmore and Peplow [4] proved that when truncating and discretizing the integral in OZ paired with HNC (or PY), there are infinitely many solution branches and bifurcation points for very low temperatures with repulsive, homogeneous potential.

Figure 5.4 [4] displays two plots. Both are continuation plots with density as the continuation parameter. Note that $\gamma = h(r) - c(r)$ here and R is the upper limit on the truncated integral from the OZ equation. Here they are plotting density versus the L_1 norm of γ. The figure on the left is a schematic. It shows three types of expected solution branches for different ranges of density and temperature. The branch labeled

C_1 has a familiar simple fold (or fold bifurcation) at ρ_{FB}. The branch labeled C_2 displays a spinoidal bifurcation at ρ_{spi}. A spinoidal bifurcation can be thought of like a vertical asymptote so that the solution branch tends toward infinity as ρ approaches ρ_{spi}. Finally, the branch labeled C_3 has no bifurcation at all.

On the right hand side of Figure 5.4, we see some spurious solution branches to the discrete problem as computed by Beardmore and Peplow in [4]. Along the branches computed here, there exist h and c (for particular values of ρ and T) that are solutions to the discrete OZ equation, but not the continuous one!

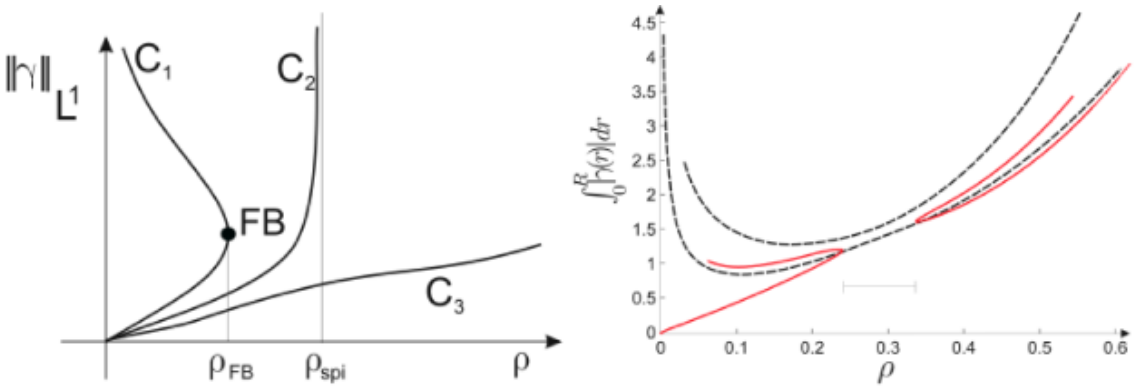


Figure 5.4: Beardmore and Peplow 2006 ($\gamma = h - c$, R =upper limit on integral for truncated problem)

Currently there is no mathematical evidence that says the interpolating closure IC (5.5) will avoid spurious, non-physical solution branches like Beardmore and Peplow discovered with the HNC and PY closures. However using IC in conjunction with robust continuation software may improve on the accuracy and efficiency of existing continuation studies that use HNC or PY closures. The next section presents new results to this end.

5.3.4 Continuation for the OZ and IC Equations for Atomic Fluids

The primary focus of this section is to present new continuation results that employ the OZ and IC equations for atomic fluids. These new results include

- continuation studies for particular isothermal parameters,
- a verification of correctness for these continuation studies, and
- evidence that our continuation codes can recover physically relevant solution paths while bypassing irrelevant "no solution zones".

We want to solve the OZ and IC equations but for various density values. This means for each density value ρ, we must solve the optimization problem of minimizing the excess chemical potential $\mu(h(r), c(r), a)$ from (5.6). With each continuation step, in order to minimize (5.6) with respect to a, one could simply run a suitable optimization scheme such as a descent or simplex method. Then once the minimizer a is found, it can be plugged into the closure equation to solve for $h(r)$ and $c(r)$ as described in §5.6. While this is the most intuitive approach, there may be a sufficiently accurate way to minimize the free energy by simply tweaking the continuation problem a bit.

The idea, proposed by Andrew Salinger of Sandia National Laboratories, is to add a few more equations to our main equation (5.7) so that when the new system is solved, we will approximate the value of a that minimizes $\mu(h(r), c(r), a)$ *and* simultaneously solve the OZ equation paired with IC. Then the continuation problem will be applied to this new system of equations.

We will formulate this new system in a way that requires a first order difference approximation to the derivative of $\mu(h(r), c(r), a)$ with respect to a to be small. That is, we wish to find a that makes $d\mu/da = 0$ which signifies that we are at a critical point. Thus if $\mu(h(r), c(r), a)$ has a concave up nature (and it does for atomic fluids), this will imply we are at a minimum.

To see how to formulate the new system, suppose we arrange (5.1) and (5.5) into the form $G = 0$ as done in §5.6. Call $u = (h, c)$ and consider a to be another unknown. Let $y = (u, v, a)$ be the new unknown (where v is a vector of the same length as u). Then instead of solving $G(u) = 0$, solve

$$P(y) = \begin{pmatrix} G(u, a) \\ G(u + \epsilon v, a + \epsilon) \\ [\mu(u + \epsilon v, a + \epsilon) - \mu(u, a)]/\epsilon \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The first two equations give two points (u, a) and $(u + \epsilon v, a + \epsilon)$ on the curve $G = 0$. The third equation says that the first order approximation to the derivative of μ at the first point is small, meaning a is a minimizer of μ. Finding the y that satisfies $P(y) = 0$ means solving the original problem $G(u) = 0$ and the minimization problem simultaneously. Then apply continuation in ρ on

$$P(y, \rho) = 0. \tag{5.8}$$

We have just described a way to implement continuation for the OZ and IC equations for atomic fluids. As we've previously mentioned, this problem is important for understanding structural and thermodynamic changes to a fluid as a function of

density. In the case of the HNC and PY closures, not only is it easy to detect spurious solutions when running a continuation scheme [4], but there are also regions in which there is no solution [37]. Additionally, finding a continuation scheme that is computationally efficient and can find physically relevant solution paths is a challenge. This section is an initial investigation into these issues for atomic fluids. The results presented here begin to validate the OZ equation paired with IC and sophisticated continuation software (§5.5) as a potential improvement upon existing continuation studies, in particular, the ones from [37].

Before we describe the results from [37] and the new ones presented here, we must explain a fact about the IC formulation that we have (intentionally) neglected up to this point. For atomic fluids, the value of a that minimizes the excess chemical potential (5.6) is infinity. Recall from earlier in this chapter that as $a \to \infty$, the interpolating closure IC goes to the PY closure (5.4). In other words, in theory we could have just used the PY closure to begin with instead of IC which requires the extra optimization problem. Yet still the use of the interpolating closure IC with robust numerical continuation software may improve upon some of the issues discussed in the last paragraph. Further, these results will validate the new methodology, code, and software that we use here and wish to extend to molecular fluids (§5.4).

Now we compare results from [37] to new results obtained using the OZ equation paired with IC (and our own PY code). We use the software from §5.5 and methodology discussed at the beginning of this section to solve (5.8).

The first task is to establish that, for a particular density and temperature value, the following yield the same correlation functions $h(r)$ and $c(r)$:

- the result from [37],
- the result from our code that uses the OZ equation paired with the PY closure, and
- the result from our code that uses the OZ equation paired with the IC closure.

Doing this will verify that

- our code for the PY closure is accurate, and
- our code for the IC closure is accurate.

Note that in the chemistry literature, it is conventional to plot quantities in terms of their *reduced dimension*. In the next several plots we will use the dimensionless parameters defined by

- $\rho^* = \rho\sigma^3$ where σ is from the Lennard-Jones potential (5.3) (ρ is original density), and
- $T^* = TK_B/\epsilon$ where ϵ is also from the Lennard-Jones formula (T is the original temperature and K_B is Boltzmann's constant).

Figure 5.5 compares the total correlation function $h(r)$ for $T^* = 1.2$, $\rho^* = .6$, and $\sigma = 5.851$ for OZ paired with the PY and IC closure approximations. As expected, the results from PY and IC are indistinguishable, and further, they agree with the physically relevant result from [37] on p. 53, Figure 2. The result from OZ paired with HNC is plotted as well for the sake of comparison. Note that $L = 3.5\sigma$ is the upper limit on the integral in the OZ equation. Additionally, we plot the direct correlation

$c(r)$ in Figure 5.6 for just OZ paired with IC (with the same parameters from Figure 5.5). The result obtained from the PY closure was identical (so we do not show it here) and matches the one from [37].

Thus Figures 5.5 and 5.6 validate our PY and IC codes in the sense that they produce the same correlation functions and match the result from [37]. Additionally, we reiterate that our codes produced the physically relevant (non-spurious) solutions that were also found in [37].

The next task is to reproduce Figure 5 from p. 54 of [37], a continuation study in density versus the thermodynamic quantity *virial pressure* P given by the formulation

$$\frac{P}{\rho T K_B} = 1 - \frac{2\pi\rho}{3}\int_0^\infty r^3 u'(r) g(r) dr \tag{5.9}$$

where once again, T is temperature, K_B is Boltzmann's constant, and $u(r)$ is the Lennard-Jones pair potential (5.3). In Figure 5.7 we plot (in reduced variables) $1/\rho^*$ versus P/TK_B. We run this continuation study for four reduced temperature values T^* using the OZ equation paired with both PY (lines) and IC (dots). Again as expected, the results from our PY and IC codes match. An interesting characteristic of Figure 5.7, however, is that our continuation codes find solutions in a range of reduced density values where PY theoretically has no solutions. While these solutions are physically irrelevant, our continuation scheme is able to recover from this "no solution zone" and continue to compute correct solutions along the remaining path where PY does have solutions. For example, the curve in Figure 5.7 for $T^* = 1.2$ matches the result from [37]. There is, however, a theoretical no solution zone in the approximate range $3 \leq 1/\rho^* \leq 6.5$. In this region, our continuation code finds solutions, but is able to

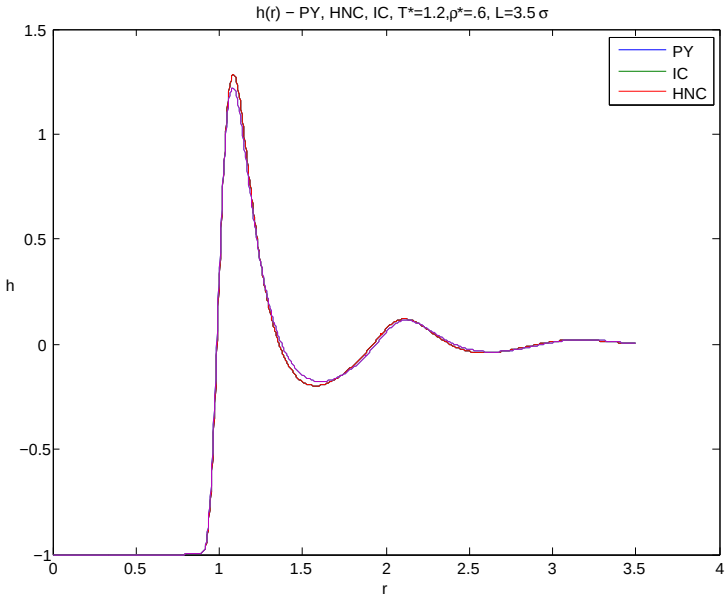


Figure 5.5: Total correlation function $h(r)$ for OZ paired with HNC, PY, and IC closures with $T^* = 1.2$, $\rho^* = .6$, and $\sigma = 5.851$

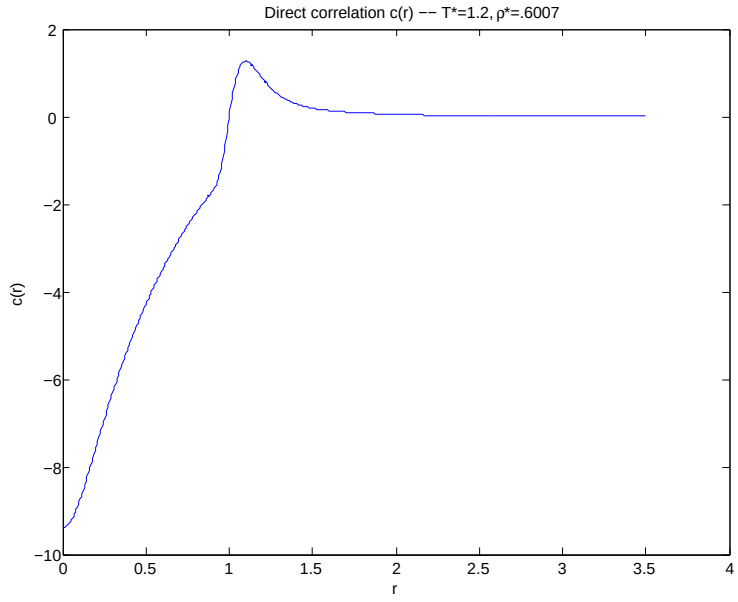


Figure 5.6: Direct correlation function $c(r)$ for OZ paired with IC closures with $T^* = 1.2$, $\rho^* = .6$, and $\sigma = 5.851$

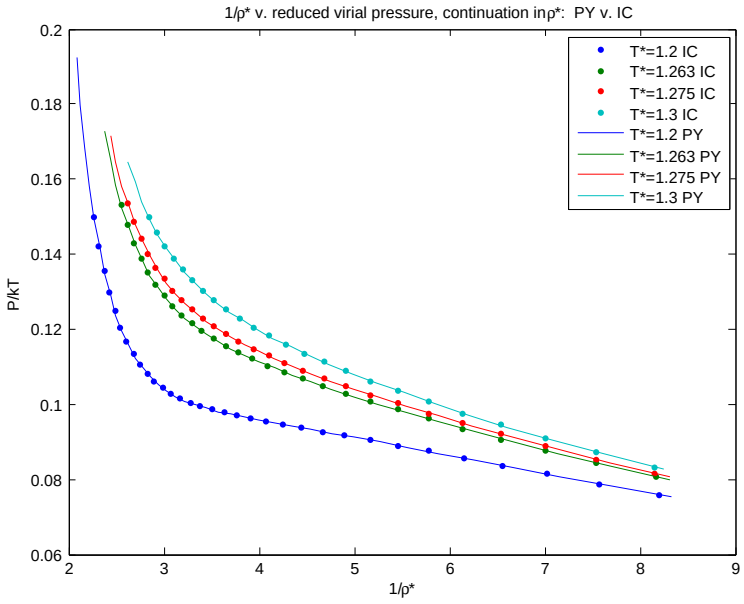


Figure 5.7: HNC and IC closures: pressure as a function of reduced density

resume on the correct path ($1/\rho^* > 6.5$) which matches the result from [37]. It is not totally understood why our codes produce solutions in the no solution zones. However, it does bring up an important distinction between the plots from [37] and our PY and IC plots: in [37] (and many other sources from the literature), the continuation plots are generated *by hand*. In other words, one obtains the continuation studies by manually plugging in various (reduced) density values (in this case) into the OZ equation, solving for the quantity of interest (like reduced virial pressure), and connecting the dots. That is, there are no automated continuation techniques or software used to produce the plots. In contrast, our plots implement the software from §5.5 and algorithms from Chapter 3 to efficiently generate the continuation study in Figure 5.7 (as well as the others in this chapter). This is a substantial improvement and advantage over techniques used in [37] and other "hand made" continuation studies.

Despite the fact that IC is equivalent to PY in the case of atomic fluids, we still have recovered some interesting information by combining the continuation formulation (5.8) with powerful continuation software [33] described in §5.5. Further, we have now provided a complete picture to the extension of the interpolating closure theory for *molecular fluids*, to which we now turn.

5.4 Extension to Molecular Fluids

In this section, we present

- a new closure equation developed by Marucho and Pettitt [24] for diatomic molecular fluids,

- a new continuation problem formulation for this closure developed by Salinger, and
- a new preliminary continuation study (Dickson) for Chlorine-like molecular fluids employing the closure equation and problem formulation just mentioned.

So far we have described a way to recover accurate structural and thermodynamic properties of atomic fluids with varying densities using the interpolating closure. We now extend this idea to more complicated molecular fluids. First, we present the "analog" of the OZ and IC equations for molecular fluids.

5.4.1 Integral Equation Theory for Molecular Fluids

Integral equation theory for molecular fluids is more complex than the OZ equation for atomic fluids. As mentioned in §5.2, we will discuss the theory for diatomic molecular fluids specifically. An integral equation theory for diatomic molecular fluids, and the one considered here, was developed by Dyer, Perkyns, and Pettitt [10] and was found to be an improvement upon existing theories such as PISM (proper interaction site model)[21]. The equations are in matrix-product form and given in Fourier space. The "hat" notation denotes the Fourier transform, and k is the independent variable in Fourier space. In particular, we use a fast sine transform, the definition of which can be found in §5.6. The molecular fluid equations are given by

$$\hat{\mathbf{H}}(k) = \hat{\mathbf{C}}(k) + [\hat{\mathbf{C}}(k) + \hat{\mathbf{S}}(k)]\bar{\rho}_m[\hat{\mathbf{H}}(k) + \hat{\mathbf{S}}(k)]. \tag{5.10}$$

Here,

$$\rho_m = \begin{pmatrix} \rho & \eta \\ \eta & 0 \end{pmatrix}, \bar{\rho}_m = \begin{pmatrix} \rho_m & \mathbf{0} \\ \mathbf{0} & \rho_m \end{pmatrix}, \hat{\mathbf{S}}(k) = \begin{pmatrix} \mathbf{0} & s(k) \\ s(k) & \mathbf{0} \end{pmatrix}, s(k) = \eta^{-1} \begin{pmatrix} 0 & 0 \\ 0 & \frac{sin(kL)}{kL} \end{pmatrix},$$

where η is called the screened density, and L is the bond length between atoms in the molecule. The notations ρ_m and $\bar{\rho}_m$ differentiate the matrix forms of ρ and the regular scalar ρ. The unknowns here are the matrix entries of

$$\mathbf{Q}(k) = \begin{pmatrix} \mathbf{Q}_{11}(k) & \mathbf{Q}_{12}(k) \\ \mathbf{Q}_{21}(k) & \mathbf{Q}_{22}(k) \end{pmatrix}, \mathbf{Q}_{\alpha\gamma}(k) = \begin{pmatrix} q^o_{\alpha\gamma}(k) & q^r_{\alpha\gamma}(k) \\ q^l_{\alpha\gamma}(k) & q^b_{\alpha\gamma}(k) \end{pmatrix},$$

where $\mathbf{Q}$ represents either $\mathbf{H}$ or $\mathbf{C}$. The subscript $\alpha = 1, 2$ represents one type of atom in the diatomic molecule and $\gamma = 1, 2$ the other so that $\mathbf{H}_{\alpha\gamma}$ or $\mathbf{C}_{\alpha\gamma}$ represents a correlation between atom α in one molecule and atom γ in another. Finally, the superscripts o, r, l, and b represent further classifications of each atom-atom correlation (they can be thought of as four "orientations" of the molecule). Just as in the case for atomic fluids, we wish to recover accurate correlation functions for molecular fluids. For diatomic molecules in particular, we now see the growth of correlations we must keep track of (namely, the $\alpha\alpha$, $\alpha\gamma = \gamma\alpha$, and $\gamma\gamma$ correlations). In general, the above equations can be extended to n-atomic molecules where the matrices and number of unknown correlations grow considerably in size as n gets large.

Although in matrix form, however, we still have two primary unknowns as in the atomic case, $\mathbf{H}$ and $\mathbf{C}$. Thus we must close the system with a closure equation. Like atomic fluids, the popular HNC and PY closures for molecular fluids result in structural

and thermodynamic inconsistencies. Thus we seek a new closure that will remedy these effects, namely, an "interpolating closure" for molecular fluids.

5.4.2 Interpolating Closure for Molecular Fluids

In [24], Marucho and Pettitt introduce the molecular "analog" of the interpolating closure presented in §5.3.2. The interpolating closure equations for molecular fluids are given by

$$
\begin{aligned}
c^o_{\alpha\gamma}(r) &= -a^o e^{-\beta u(r)} + (1+a^o) e^{[-\beta u(r) + (h^o_{\alpha\gamma}(r) - c^o_{\alpha\gamma}(r))/(1+a^o)]} - 1 - h^o_{\alpha\gamma}(r) + c^o_{\alpha\gamma}(r), \\
c^r_{\alpha\gamma}(r) &= \frac{(1+a^o)}{(1+a^r)} (h^r_{\alpha\gamma}(r) - c^r_{\alpha\gamma}(r)) e^{[-\beta u(r) + (h^o_{\alpha\gamma}(r) - c^o_{\alpha\gamma}(r))/(1+a^o)]} - h^r_{\alpha\gamma}(r) + c^r_{\alpha\gamma}(r), \\
c^l_{\alpha\gamma}(r) &= \frac{(1+a^o)}{(a+a^r)} (h^l_{\alpha\gamma}(r) - c^l_{\alpha\gamma}(r)) e^{[-\beta u(r) + (h^o_{\alpha\gamma}(r) - c^o_{\alpha\gamma}(r))/(1+a^o)]} - h^l_{\alpha\gamma}(r) + c^l_{\alpha\gamma}(r), \\
c^b_{\alpha\gamma}(r) &= (1+a^o) \left[\frac{(h^b_{\alpha\gamma}(r) - c^b_{\alpha\gamma}(r))}{(1+a^b)} + \frac{(h^r_{\alpha\gamma}(r) - c^r_{\alpha\gamma}(r))(h^l_{\alpha\gamma}(r) - c^l_{\alpha\gamma}(r))}{(1+a^r)^2} \right] \times \\
&\quad e^{[-\beta u(r) + (h^o_{\alpha\gamma}(r) - c^o_{\alpha\gamma}(r))/(1+a^o)]} - h^b_{\alpha\gamma}(r) + c^b_{\alpha\gamma}(r).
\end{aligned}
\tag{5.11}
$$

Notice there are now three interpolation parameters for the molecular case, a^o, a^r, and a^b. Once again, $u(r)$ is the Lennard-Jones pair potential between particles (5.3), and β is the inverse of temperature times Boltzmann's constant. Similar to §5.3.2, we now must choose the values of a^o, a^r, and a^b to minimize the approximate excess *molecular* chemical potential given by

$$\beta\mu(h_{\alpha\gamma}, c_{\alpha\gamma}, a^o, a^r, a^b) \approx -\rho \sum_{\alpha\gamma} \int \{h_{\alpha\gamma} - \mathcal{I}\left[h^o_{\alpha\gamma}(r), c^o_{\alpha\gamma}(r), a^o\right] \times \{h^o_{\alpha\gamma}(r) - c^o_{\alpha\gamma}(r) + \frac{(1+a^o)\left[h^r_{\alpha\gamma}(r) - c^r_{\alpha\gamma}(r) + h^l_{\alpha\gamma}(r) - c^l_{\alpha\gamma}(r)\right]}{(1+a^r)} + \frac{(1+a^o)\left[h^b_{\alpha\gamma}(r) - c^b_{\alpha\gamma}(r)\right]}{2(1+a^b)}\} - \frac{h^r_{\alpha\gamma}\left[h^l_{\alpha\gamma}(r) - c^l_{\alpha\gamma}(r)\right]}{(1+a^r)} - \frac{\left[h^r_{\alpha\gamma}(r) + h^l_{\alpha\gamma}(r) + h^b_{\alpha\gamma}(r)\right]\left[h^o_{\alpha\gamma}(r) - c^o_{\alpha\gamma}(r)\right]}{2(1+a^o)}\} d\mathbf{r}, \tag{5.12}$$

where $\mathcal{I}(h^o_{\alpha\gamma}(r), c^o_{\alpha\gamma}(r), a^o)$ is obtained by replacing $h(r), c(r)$, and a by $h^o_{\alpha\gamma}(r), c^o_{\alpha\gamma}(r)$, and a^o, respectively in the formulation for $\mathcal{I}$ in §5.3.2. Now the goal is to solve the matrix equations and molecular interpolating closure just defined for $\mathbf{H}$ and $\mathbf{C}$. This is done in [24] for a few values of density. However, we wish to do a continuation study for molecular fluids as was done in §5.3.4. Thus the next step is to formulate a problem that will allow a continuation study in ρ while solving the new molecular equations, keeping in mind that we must now minimize the molecular free energy with each continuation step.

5.4.3 Continuation for Molecular Fluids

As done for atomic fluids in §5.3.4, we can formulate a new continuation problem for molecular fluids that will resolve the optimization problem from §5.4.2. We can do this by requiring a first order difference approximation to $d\mu/da^o = d\mu/da^r = d\mu/da^b = 0$. Of course now that we have three interpolation parameters to resolve, the continuation problem in (5.8) grows in size.

To see how to formulate the new system, suppose we arrange (5.10) and (5.11) into

the form $G = 0$. We will write the elements of the matrix unknowns $\mathbf{H}$ and $\mathbf{C}$ as vectors of length $16N$ where N is the size of the discretization (like the one from §5.6). This is done by stacking the rows of $\mathbf{H}$ and $\mathbf{C}$ so that

$$\mathbf{Q} = \begin{pmatrix} q^o_{\alpha\alpha} \\ q^r_{\alpha\alpha} \\ q^l_{\alpha\alpha} \\ q^b_{\alpha\alpha} \\ q^o_{\alpha\gamma} \\ q^r_{\alpha\gamma} \\ q^l_{\alpha\gamma} \\ q^b_{\alpha\gamma} \\ q^o_{\gamma\alpha} \\ q^r_{\gamma\alpha} \\ q^l_{\gamma\alpha} \\ q^b_{\gamma\alpha} \\ q^o_{\gamma\gamma} \\ q^r_{\gamma\gamma} \\ q^l_{\gamma\gamma} \\ q^b_{\gamma\gamma} \end{pmatrix} \tag{5.13}$$

where $\mathbf{Q}$ is either $\mathbf{H}$ or $\mathbf{C}$ and $\alpha, \gamma = 1, 2$ where the two indices run over the possible atoms that make up the diatomic molecule.

Call $u = (\mathbf{H}, \mathbf{C})$ in the vector format (5.13), and consider a^o, a^r, and a^b to be

unknowns. Let $y = (u, v_1, v_2, v_3, a^o, a^r, a^b)$ be the new unknown (where v_1, v_2, and v_3 are vectors of the same length as u). Then instead of solving $G(u) = 0$, solve

$$P(y) = \begin{pmatrix} G(u, a^o, a^r, a^b) \\ G(u + \epsilon v_1, a^o + \epsilon, a^r, a^b) \\ G(u + \epsilon v_2, a^o, a^r + \epsilon, a^b) \\ G(u + \epsilon v_3, a^o, a^r, a^b + \epsilon) \\ [\mu(u + \epsilon v_1, a^o + \epsilon, a^r, a^b) - \mu(u, a^o, a^r, a^b)]/\epsilon \\ [\mu(u + \epsilon v_2, a^o, a^r + \epsilon, a^b) - \mu(u, a^o, a^r, a^b)]/\epsilon \\ [\mu(u + \epsilon v_3, a^o, a^r, a^b + \epsilon) - \mu(u, a^o, a^r, a^b)]/\epsilon \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

In the same fashion as in §5.3.4, finding the y that satisfies $P(y) = 0$ means solving the original problem $G(u) = 0$ combined with the minimization problem. Then one applies continuation in ρ on

$$P(y, \rho) = 0. \tag{5.14}$$

The formulation of this continuation problem for diatomic molecules is simply the obvious extension of what was done in §5.3 for atoms: we minimize the excess potential for molecular fluids using a first order difference approximation of μ in each of the three interpolation parameters a^o, a^r, and a^b. For atomic fluids, the interpolating closure IC was equivalent to the PY closure. For molecules, however, the interpolating closure paired with the integral equation theory from [10] is unique. In [24], there is already strong evidence that these combined theories result in more accurate correlation functions than existing HNC-PISM approaches for certain values of density and

temperature. In fact, Figure 5.3 shows how the molecular interpolating closure (5.11) predicts a more accurate pair correlation for Chlorine-like diatomic molecules. (We simply didn't have the vocabulary to say so at the time.)

Implementing continuation for simple atomic fluids is not easy and even more difficult for diatomic molecular fluids (and so on...). Specifically, *phase transitions* where a fluid changes state (say, from liquid to vapor) as a function of temperature and density boils down to singular points mathematically. That is, to track difficult and important phase behavior for molecular fluids, one must employ the methods of Chapter 3 in an efficient and accurate way. It is the hope that using the above continuation formulation and the software from §5.5, new chemical and physical territory may be revealed with little extra computational effort than the current literature reports. We have already taken strides toward this goal. Figure 5.8 is a preliminary continuation study in (reduced) density with $T^* = 2.2$ and constant screened density η for Chlorine-like molecules. These results were produced using the new technology described in this section and verified (for constant η) using results from [24]. These results combined with those from §5.3 indicate the potential for uncovering new physics for complicated fluids (such as water) in a computationally accurate and efficient way.

5.5 Trilinos

We now take some time to discuss the software responsible for producing the results in this chapter. Trilinos is a collection of free software packages developed by Sandia National Laboratories. It is written in the C++ programming language, and while

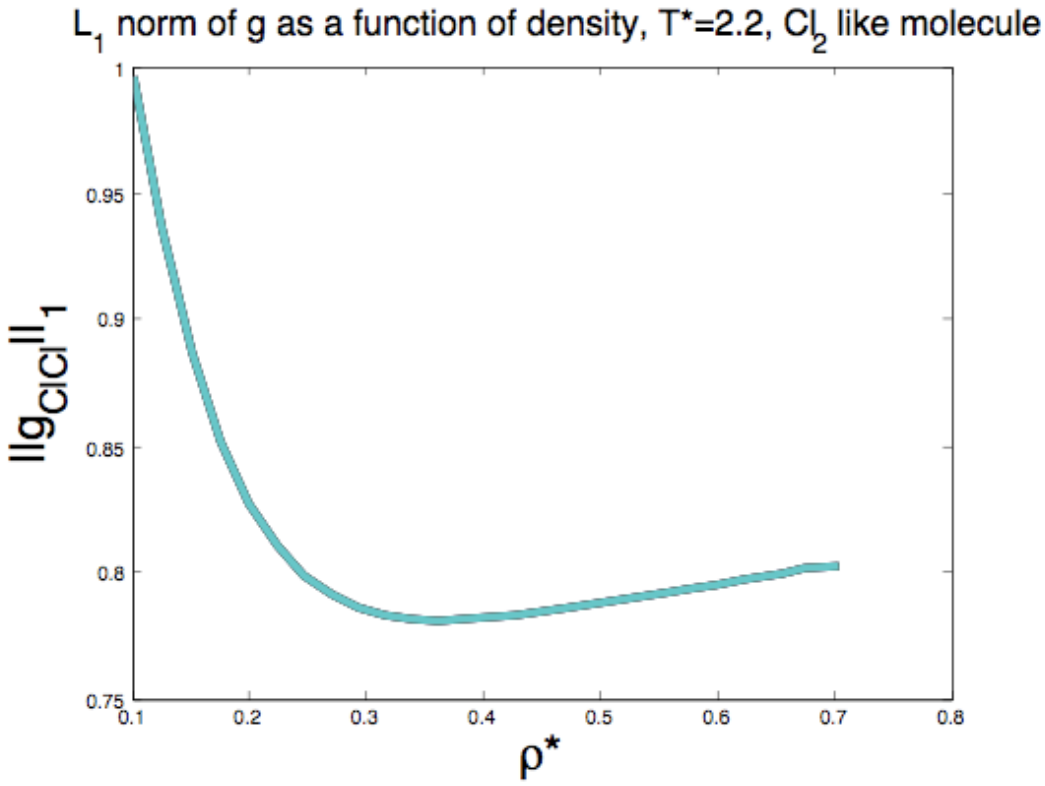


Figure 5.8: Continuation in density (with surrounding pair correlation snapshots) for Chlorine-like molecules with $T^* = 2.2$ and constant η

most of the packages within Trilinos have the ability to stand alone, they are intended to integrate with one another. For example, the nonlinear solver package containing Newton's method has the ability to call the linear solver package to solve for the Newton step. To find out more about Trilinos, visit

`http://trilinos.sandia.gov`.

In particular, we are interested in three packages within the Trilinos suite. The first is the nonlinear solver package NOX (Nonlinear Object-Oriented Solutions). This is the package that controls the nonlinear solver needed to solve the nonlinear problem with each continuation step. Second is the linear solver package AztecOO, which contains both direct and iterative method linear solvers (needed to solve for the Newton step within NOX). Finally, and most important for us, is LOCA (Library of Continuation Algorithms). LOCA is responsible for using the technology discussed in Chapter 3 (and more!) to track solution branches as a function of a continuation parameter.

The main goal in the context of this chapter is to use LOCA to track solution (pair correlation functions) branches for the integral equations paired with the IC closure (in both the atomic and molecular settings). We describe some details in implementing LOCA for this problem. In order to solve the problem (5.8) for atomic fluids and (5.14) for molecular fluids in LOCA, we are required to write two major components of C++ code.

The first of the two is an interface between LOCA and the problem itself. The primary purpose for this is to define options for the continuation (LOCA), the nonlinear solver (NOX), and the linear solver (AztecOO). Specific solver methods, convergence tolerances, and maximum iterations are only a few examples of solver options that must be specified in this part of the code. For the results presented in §5.3.4, we used pseudo-arclength as the continuation strategy, Newton's method with line search as the nonlinear solver, and matrix-free GMRES to solve for the Newton step.

The second necessary piece of code is the one that defines the problem. In general,

this is where the residual (5.7) is coded and queried by the solvers. For the problem of interest, it is the place that implements (5.8) for atomic fluids and (5.14) for molecular fluids. Other methods can also be defined here such as a method to print out the solution with each continuation step or a method that defines a Jacobian matrix if the user wishes to supply one.

5.6 Solving the OZ and HNC Equations

Here we give a detailed strategy for solving the OZ equation. We do this in the context of the HNC closure, although it is very similar for other closures discussed in this chapter.

In order to solve the OZ and HNC equations, we first express the OZ equation as a compact fixed point problem in the form of (5.7). We then discretize the resulting problem with the trapezoid rule. This gives us the format for the continuation problems to be solved using LOCA in §5.3.4, 5.4.3.

5.6.1 The Fixed Point Problem

Recall that the OZ equation is given by

$$h(r) = c(r) + \rho(h * c)(r) \tag{5.15}$$

where

$$(h * c)(r) = \int_{\Re^3} c(\|\mathbf{r} - \mathbf{r}'\|)h(\|\mathbf{r}'\|)d\mathbf{r}'. \tag{5.16}$$

The convolution $h * c$ can be computed with only one-dimensional integrals using the spherical-Bessel transform. Assuming h decays sufficiently rapidly, we define

$$\hat{h}(k) = \mathcal{H}(h)(k) = 4\pi \int_0^\infty \frac{sin(kr)}{kr} h(r) r^2 dr$$

and

$$h(r) = \mathcal{H}^{-1}(\hat{h})(r) = \frac{1}{2\pi^2} \int_0^\infty \frac{sin(kr)}{kr} \hat{h}(k) k^2 dk.$$

We compute $h * c$ by discretizing the formula (see §5.6.2)

$$h * c = \mathcal{H}^{-1}(\hat{h}\hat{c}) \tag{5.17}$$

where $\hat{h}\hat{c}$ denotes the pointwise product of functions. Transforming (5.15) gives

$$\hat{h} = \hat{c} - \rho \hat{h} \hat{c}$$

so that, given c, we can compute h as

$$h = H(c) = \mathcal{H}^{-1}\left(\frac{\hat{c}}{1 - \rho\hat{c}}\right).$$

Now we use the HNC closure

$$\exp(-u(r)/(TK_B) + h(r) - c(r)) - h(r) - 1 = 0, \qquad 0 \le r \le \infty. \tag{5.18}$$

to recover the fixed point map. Here u is the Lennard-Jones potential defined in (5.3). Having computed $h = H(c)$, we define

$$y = h - c$$

so $h = y + c$, and the HNC closure becomes

$$\exp(-u(r)/(TK_B) + y(r)) - y(r) - 1 = c(r).$$

Substituting $H(c) - c$ in for y gives

$$\mathcal{K}(c) = \exp(-u(r)/(TK_B) + (H(c) - c)(r)) - (H(c) - c)(r) - 1$$

where $\mathcal{K}$ is a compact fixed point map in c [19] with $c = \mathcal{K}(c)$ at a solution. The final residual is

$$G(c) = c - \mathcal{K}(c).$$

Since h can be recovered once c is known, here we can solve

$$G(h, c, \rho = 0)$$

as ρ varies by first solving for c and then simply evaluating h with each continuation step. This also is the case for implementing the interpolating closure (5.5) for atomic fluids. For solving (5.14), the unknowns $\mathbf{H}$ and $\mathbf{C}$ must be solved for simultaneously with each continuation step.

5.6.2 The Discretization

As alluded to in §5.6.1, we discretize the convolution in (5.17) by discretizing the transform and then defining the discrete convolution as the inverse transform of the product of the transforms. We discretize the transform variable in a way that allows a fast Fourier transform (FFT) to evaluate the transform and its inverse.

First, we truncate the radial variable at L so that $r \in [0, L]$. The nodes in r are defined as

$$\{r_i^\delta\}_{i=1}^N$$

where $r_i^\delta = (i-1)\delta$ and $\delta = L/(N-1)$ is the mesh width.

The nodes in the transform variable are $k_j = (j-1)\delta_k$ where $\delta_k = \pi/L$ (so $\delta_k\delta = \pi/(N-1)$). We define, for $2 \le j \le N-1$ and $v \in \Re^N$,

$$\begin{aligned}
\hat{v}_j &= \mathcal{H}(v)_j \\
&= \tfrac{4\pi\delta^2}{(j-1)\delta_k} \sum_{i=2}^{N-1} (i-1) v_i \sin((i-1)(j-1)\delta_k\delta) \\
&= \tfrac{4\pi\delta^3(N-1)}{j-1} \sum_{i=2}^{N-1} (i-1) v_i \sin((i-1)(j-1)\pi/(N-1)).
\end{aligned}$$

Then, for $2 \le i \le N-1$,

$$\mathcal{H}^{-1}(\hat{v})_i = \frac{1}{2(i-1)\pi\delta^3} \sum_{j=2}^{N-1} k\hat{v}_j \sin((i-1)(j-1)\pi/(N-1)).$$

Finally, define for $2 \leq i \leq N-1$,

$$(u * v)_i = \mathcal{H}^{-1}(\hat{u}\hat{v})$$

where $\hat{u}\hat{v}$ denotes the component-wise product. We set $(u*v)_N = 0$ and define $(u*v)_1$ by linear interpolation

$$(u * v)_1 = 2(u * v)_2 - (u * v)_3.$$

We use a fast sine transform to compute the transform and inverse transform.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1972.

[2] V. I. Arnold. *The Theory of Singularities and Its Applications*. Press Syndicate of the University of Cambridge, Cambridge, 1991.

[3] Phil Attard. *Thermodynamics and Statistical Mechanics: Equilibrium by Entropy Maximisation*. Academic Press, Inc., London, 2002.

[4] R. E. Beardmore and A. Peplow. A bifurcation analysis of the Ornstein-Zernike equation with hypernetted chain closure, 2006. preprint.

[5] T. H. Bröcker. *Differentiable Germs and Catastrophes*. Cambridge University Press, Cambridge, 1975.

[6] Domenico P. L. Castrigiano and Saundra A. Hayes. *Catastrophe Theory*. Westview Press, Boulder, second edition, 2004.

[7] S. Chandrasekhar. *Radiative Transfer*. Dover, New York, 1960.

[8] K. I. Dickson, C. T. Kelley, I. C. F. Ipsen, and I. G. Kevrekidis. Condition estimates for pseudo-arclength continuation. *SIAM J. Numer. Anal.*, 45:263–276, 2006.

[9] Eusebius J. Doedel, Alan R. Champneys, Thomas F. Fairgrieve, Yuri A. Kuznetsov, Bjorn Sandstede, and Xianjun Wang. Auto 97: Continuation and bifurcation software for ordinary differential equations (with HomCont). Manual, March 1998.

[10] K. M. Dyer, J. S. Perkyns, and B. M. Pettitt. Effective density terms in proper integral equations. *J. Chem. Phys.*, 123:204512, 2005.

[11] W. R. Ferng and C. T. Kelley. Mesh independence of matrix-free methods for path following. *SIAM J. Sci. Comput.*, 21:1835–1850, 2000.

[12] Robert Gilmore. *Catastrophe Theory for Scientists and Engineers*. John Wiley and Sons, New York, 1981.

[13] W. J. F. Govaerts. *Numerical Methods for Bifurcations of Dynamic Equilibria*. SIAM, Philadelphia, 2000.

[14] Jean Pierre Hansen and Ian R. McDonald. *Theory of Simple Liquids*. Academic Press, Inc., London, second edition, 1986.

[15] I. C. F. Ipsen and B. Nadler. Refined perturbation bounds for eigenvalues of Hermitian and non-Hermitian matrices. *SIAM J. Matrix Anal. Appl.* to appear.

[16] H. B. Keller. *Lectures on Numerical Methods in Bifurcation Theory*. Tata Institute of Fundamental Research, Lectures on Mathematics and Physics. Springer-Verlag, New York, 1987.

[17] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. Number 16 in Frontiers in Applied Mathematics. SIAM, Philadelphia, 1995.

[18] C. T. Kelley. *Solving Nonlinear Equations with Newton's Method*. Society for Industrial and Applied Mathematics, Philadelphia, 2003.

[19] C. T. Kelley and B. Montgomery Pettitt. A fast solver for the Ornstein-Zernike equations. *J. of Comp. Phys.*, 197:491–501, 2004.

[20] Yuri A. Kuznetsov. *Elements of Applied Bifurcation Theory*. Springer, New York, 1998.

[21] L. Luie and D. Blankschtein. *J. Chem. Phys.*, 102:5427, 1995.

[22] Antal Majthay. *Foundations of Catastrophe Theory*. Pitman Advanced Publishing Program, Marshfield, Massachusetts, 1985.

[23] Jerrold E. Marsden and Michael J. Hoffman. *Elementary Classical Analysis*. W.H. Freeman and Company, New York, second edition, 1993.

[24] M. Marucho and B. Montgomery Pettitt. An optimized theory for simple and molecular fluids. *J. Chem. Phys.*, 126:124107–124107–9, 2007.

[25] C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, 2000.

[26] T. W. Mullikin. Some probability distributions for neutron transport in a half space. *J. Appl. Prob.*, 5:357–374, 1968.

[27] I. Newton. *The Mathematical Papers of Isaac Newton (7 volumes)*. Cambridge University Press, Cambridge, 1967–1976. D. T. Whiteside, Editor.

[28] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, 1970.

[29] B. N. Parlett. *The Symmetric Eigenvalue Problem*. Prentice Hall, Englewood Cliffs, 1980.

[30] Werner C. Rheinboldt and John V. Burkardt. PITCON. Technical report, University of Pittsburg, 1983.

[31] Werner C. Rheinboldt and John V. Burkardt. PITCON. Technical report, University of Pittsburg, Pittsburg, June 2000.

[32] Marzio Sala, Michael A. Heroux, and David M. Day. Trilinos Tutorial. Technical Report SAND 2004-2189, Sandia National Laboratories, August 2007.

[33] Andrew G. Salinger, Nawaf M. Bou-Rabee, Roger P. Pawlowski, Edward D. Wilkes, Elizabeth A. Burroughs, Richard B. Lehoucq, and Louis A. Romero. LOCA 1.1: Library of continuation algorithms: Theory and implementation manual. Manual, Sandia National Laboratories, Albuquerque, October 2002.

[34] P.T. Saunders. *An Introduction to Catastrophe Theory*. Cambridge University Press, Cambridge, 1980.

[35] Rudiger Seydel. *Practical Bifurcation and Stability Analysis: From Equilibrium to Chaos*. Springer-Verlag, New York, second edition, 1994.

[36] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, Philadelphia, 1996.

[37] R. O. Watts. Percus-Yevick equation applied to a Lennard-Jones fluid. *J. Chem. Phys.*, 48:50–55, 1968.

Appendix

An Introduction to Singularity Theory

Singularity theory and *catastrophe theory* [2, 5, 12, 6, 34, 22] is an extremely important study in the physical sciences primarily, but also has applications in biological, financial, and social sciences. Examples include the breaking of a wave, division of a cell, collapse of a bridge, turbulence, phase transition in chemical substances, and well-known chaotic systems such as the Lorenz attractor [34, 12].

Consider a physical system with *state variables* $(u_1, u_2, \ldots, u_N)$ for some (often very large, perhaps a million) positive integer value of N. In general, the state variables are what completely describe the system at any given time (or some other given parameters of the system). Now consider M independent *control variables* $(\lambda_1, \lambda_2, \ldots, \lambda_M)$ which influence the state variables in a possibly non-unique way, as we will see. While M is also some positive integer, it is usually small. For the purposes of this chapter (and thesis), we will assume $M \leq 4$. In the context of this section only, we will assume the system is a single function of the state and control variables and will call it g later on. Singularity theory is the underlying interest of this thesis and is used to analyze

a system that has mostly smooth solution behavior (solution meaning state variables), but undergoes discontinuities in some places (meaning for some control variables). It is at such discontinuities that critical physical behavior occurs like the examples from the above paragraph. In general, the assumption that M is small is not restrictive in the sense that most discontinuities of a system are only influenced by one (or a few at most) control variables at a time. When more than a few are involved, most any analysis on the discontinuity becomes extremely difficult or impossible [34]. Catastrophe theory tells us that the qualitative configurations of discontinuities depends on the number of control variables, not on the number of state variables.

Considering the above discussion, we have *Thom's theorem* [34, 2, 12, 5, 22] named after René Thom, a pioneer in the field:

Theorem 5.1. *Let M be a postitive integer such that $M \leq 4$. That is, let there be at most four control variables for a given system. Let the number of state variables, N, be a positive integer that is arbitrary, but finite. Then there are only seven generic distinct characterizations of any discontinuity of the system. Moreover, none of these discontinuities involve contributions from more than two state variables.*

The last sentence of the theorem requires some elaboration. Under the assumptions of Theorem 5.1, a discontinuity (which we also refer to as singularity or catastrophe throughout this section) may involve more than two state variables. However it is always possible to find a change of variables (into *eigenvariables*) so that the resulting system (equivalent to the original via transformation) only involves two state variables. The proof of this theorem is difficult and rigorously considered in [5]. An extension of this theorem proves that if there are five control variables ($M = 5$), then the number

of distinct generic catastrophes grows from seven to eleven. Beyond that, however, things become much more complicated [34], the biggest reason being that the defining characteristics of the catastrophes for $M \leq 5$ no longer exist for $M \geq 6$.

We have just seen that for $M \leq 4$, only two state variables at most are critically involved in a discontinuity. It is due to this fact that when analyzing a discontinuity, we need only consider a Taylor expansion of the system in question in two variables, at most. So far we have been referring to the seven "characterizations" of discontinuities when there are less than five control variables. These characteristics are distinguished in the following way: we say that two catastrophes are equivalent if one can be transformed to the other by (1) a diffeomorphism (a one-to-one, differentiable mapping that is invertible with differentiable inverse) of the control variables and (2) at each point in the control space, a diffeomorphism of the state variables. The resulting family of state-variable diffeomorphisms must be continuously differentiable when considered as a function of the control variables [34]. A slightly inaccurate, yet meaningful way to visualize this equivalency is to think of preserving the topology of the singularity (we will discuss these geometries shortly): if you can deform the geometry without tearing or cutting and pasting in addition to no creasing or flattening of creases, the topologies are said to be equivalent. The rest of this section describes the analysis for each of the seven catastrophies in their "canonical" form. That is, we have taken the simplest form of the system to evaluate, and it is this system that represents the catastrophe characterization for all equivalent systems as just defined.

Analyzing the seven catastrophes involves the following steps. First, given a system

$$g(u_1, u_2, \ldots, u_N, \lambda_1, \lambda_2, \ldots, \lambda_M),$$

we find the *equilibrium surface*

$$\nabla g(u_1, u_2, \ldots, u_N, \lambda_1, \lambda_2, \ldots, \lambda_M) = 0$$

where ∇ denotes the vector of partial derivatives of g with respect to the state variables only. The set of solutions to this equation is called the *equilibrium set* and is simply the critical points of g as we think of them in calculus. Next, we find the *singularity set* which is the subset of the equilibrium set with the added condition that the determinant of the Hessian matrix of g is zero ($\det H = 0$). The Hessian matrix is defined by the matrix entries $H_{ij} \equiv \frac{\partial^2 g}{\partial u_i \partial u_j}$, $i, j = 1, 2, \ldots, N$. The next step is to project the singularity set down to the control space to obtain the *bifurcation set*. The bifurcation set is the set of all points in the control space at which changes in the "form" of g occur (see below). Finally, we determine the form of g at every point in the control space. We will get a better feel for what these last two items mean in the upcoming analysis. Note that even for the "simple" catastrophes we are about to describe, there is quite a bit of analysis required. For most cases, we only give a brief description of the analysis leaving the details to [2, 5, 12, 6, 34, 22]. In summary, the goal of the analysis for each catastrophe is to separate the control space into regions for which we understand the "form" of the system. The "form" is described by the number and type of critical points of the system in a given control space region. At

this time, recall that we are only considering canonical polynomial systems in up to two state variables and four control variables. Thus instead of working with a system of the form $g(u_1, u_2, \ldots, u_N, \lambda_1, \lambda_2, \ldots, \lambda_M)$, we will simply look at $g(u, v, \lambda, \mu, \gamma, \xi)$ in the most complicated setting. Here, u and v will denote state variables and $\lambda, \mu, \gamma, \xi$ control variables. Recall from calculus that the critical points of g in this setting (for $\det H \neq 0$) will be (1) a maximum if $\det H > 0$ and $\partial^2 g/\partial u^2 < 0$, (2) a minimum if $\det H > 0$ and $\partial^2 g/\partial u^2 > 0$, or (3) a saddle point if $\det H < 0$. We are ready to describe the seven catastrophes of interest.

The fold. The simplest singularity is called a *fold*. The canonical form of a system involving a fold looks like

$$g(u, \lambda) = u^3 + \lambda u$$

where λ is the single control variable and u is the state. The equilibrium surface is

$$3u^2 + \lambda = 0 \tag{5.19}$$

and the singularity set satisfies the equilibrium surface and $6u = 0$. Thus the singularity set simply contains the point $(u = 0, \lambda = 0)$. The bifurcation set is just the projection of this point onto the control space (the line $u = 0$) which is just $\lambda = 0$. Now, when $\lambda > 0$, there are no real roots to (5.19) and therefore no critical points. When $\lambda < 0$, there is a maximum and a minimum and when $\lambda = 0$, we have a single point of inflection. Thus we have completely analyzed all regions of control space (just the real line here) and the corresponding critical points (form) of g.

The cusp. The cusp singularity has canonical form

$$g(u, \lambda, \mu) = u^4 + \lambda u^2 + \mu u$$

where λ and μ are the control variables and u is still the state. The equilibrium surface here is

$$4u^3 + 2\lambda u + \mu = 0 \tag{5.20}$$

and the singularity set satisfies (5.20) and $12u^2 + 2\lambda = 0$. Solving for λ in this last equation and substituting it into (5.20) gives $8\lambda^3 + 27\mu^2 = 0$, solutions to which make up the bifurcation set. The bifurcation set in the $\lambda\mu$ plane is a typical cusp described by two smooth curves intersecting at a sharp point which occurs at the origin here. While we do not provide all of the details here, inside the cusp there are two minima and one maximum for g and outside the cusp, a single minimum.

The swallowtail. The swallowtail has canonical form

$$g(u, \lambda, \mu, \gamma) = u^5 + \lambda u^3 + \mu u^2 + \gamma u$$

where λ, μ, and γ are the control variables and u the state. The equilibrium surface is

$$5u^4 + 3\lambda u^2 + 2\mu u + \gamma = 0 \tag{5.21}$$

and the singularity set satisfies (5.21) and $20u^3 + 6\lambda u + 2\mu = 0$. After some work,

it turns out when $\lambda > 0$, the bifurcation set in the $\mu\gamma$ plane is just a parabola with minimum at the origin. When $\lambda < 0$, the bifurcation set is more complicated in the $\mu\gamma$ plane in a way that mimics a swallowtail. There are three primary restrictions on γ that define the form of g. The first has no critical points, the second has two minima and two maxima, and the last has one maximum and one minimum.

The elliptic umbilic. Here the canonical form is

$$g(u, v, \lambda, \mu, \gamma) = \frac{1}{3}u^3 - uv^2 + \gamma(u^2 + v^2) - \lambda u + \mu v.$$

The equilibrium surface is found by taking the gradient of g and setting it equal to zero. The singularity set is the subset of the equilibrium set that has the determinant of the 2×2 Hessian matrix of g equal to zero. When all is said and done, the bifurcation set in the $\lambda\mu$ plane is three cusp points; it looks like three points connected like a triangle with sides caved in. Inside the cusped triangle (or cones in $\lambda\mu\gamma$ space), there are three saddle points and one maximum or minimum depending on whether $\gamma < 0$ and $\gamma > 0$, respectively. Outside this region there are two saddle points.

The hyperbolic umbilic. The canonical form here is

$$g(u, v, \lambda, \mu, \gamma) = u^3 + v^3 + \gamma uv - \lambda u - \mu v$$

where the control space is three dimensional and the state space is two dimensional. Again the equilibrium surface is found by setting the gradient of g equal to zero and

the singularity set is the subset of the equilibrium set that also has determinant of the 2×2 Hessian matrix of g equal to zero. The bifurcation diagram is difficult to visualize in this case. Imagine two separate sheets of paper, each folded like a tunnel or concave down arc. Imagine on each of the two tunnels, one half of the sheet is pinched right at the top of the arc. So now each sheet has a half a tunnel and half a rooftop. Now, take both sheets at the point of transition between the arc and rooftop and intersect the two sheets at the origin of the $\lambda\mu\gamma$ plane to define the bifurcation surface for the hyperbolic umbilic. There are four regions in which to consider the form of the system: above both sheets, below both sheets, and in between the two sheets on either side of the intersection point. Below both sheets there are four critical points: a minimum, a maximum, and two saddles. Above the two sheets is a maximum and a saddle. Between the two sheets for $\gamma > 0$ there are no critical points and between the sheets for $\gamma < 0$ there is one minimum and one saddle.

The butterfly. The canonical form in this case is

$$g(u, \lambda, \mu, \gamma, \xi) = u^6 + \xi u^4 + \lambda u^3 + \mu u^2 + \gamma u$$

where, this time, the control space is four dimensional and we have just one state variable. Thus the bifurcation set is impossible to visualize. A useful technique for getting a sense of the bifurcation set, however, is to take a sampling of points in the $\xi\lambda$ plane and for those values of ξ and λ, draw a bifurcation set in the $\mu\gamma$ plane. Such a sampling is well done in [34]. Since the breakdown of the bifurcation set is complicated, we don't describe the system forms here. However it is interesting that the butterfly

characteristic can be seen in a $\gamma\mu$ cross section of the bifurcation set for the case that $\lambda = 0$ and $\xi < 0$ [34].

The parabolic umbilic. The canonical form here is

$$g(u, v, \lambda, \mu, \gamma, \xi) = v^4 + u^2 v + \gamma u^2 + \xi v^2 - \lambda u - \mu v$$

so that we have a two dimensional state space and four dimensional control space. Once again the bifurcation set is impossible to visualize, but on the same order of difficulty as the butterfly. Texts by Thom (1972), Poston and Stewart (1978), and Godwin (1971) discuss this catastrophe in more detail.

Before we conclude this brief introduction to singularity theory, we must mention the issue of stability. Since stability is not considered a central theme in this thesis, we do not explain it in detail here. However, stability is an important aspect of catastrophe theory and bifurcation analysis, and we must at least mention it for the sake of completeness. A system is said to be structurally stable or unstable depending on how close the form of a neighboring system is to that of the original system g. If we consider a canonical polynomial form of g and take some arbitrarily small perturbation of its coefficients, the question of stability boils down to analyzing the resulting critical points. For example, $g(u) = u^4$ is not structurally stable because there are polynomials close to it which give birth to new and/or different types of critical points. One such polynomial is $f(u) = u^4 + \alpha u^3$ where $|\alpha|$ is small, but not zero. Since $u = 0$ is a minimum for g, but f has a point of inflection at the origin and a minimum at

$u = -\frac{3}{4}\alpha$, g is not structurally stable. Further discussion on stability can be found in [2, 5, 12, 6, 34, 22].