Computing Bifurcation Diagrams with Deflation

Casper Beentjes

St. Catherine's College

University of Oxford

A thesis submitted for the degree of

M.Sc. in Mathematical Modelling and Scientific Computing

Trinity 2015

I would like to dedicate this thesis to all my friends and family who gave me the space to grow up blissfully, but stay young in mind.

All grown-ups were once children…
but only few of them remember it.
Antoine de Saint-Exupèry (1900-1944)

Acknowledgements

This year has not been a solo-project and as a result I am indebted to many people. However, after this intense year in Oxford and meeting numerous fantastic people there is a risk of forgetting someone, know that your aid is acknowledged nonetheless!

First of all I would like to express my deepest gratitude to Dr. Patrick Farrell for his guidance in this thesis through good and bad times, his enthusiasm in supervising and above all for being a great teacher. A special thank you goes to Dr. Kathryn Gillow for her never-ending support as our course-director. Furthermore I would like to thank Jack, Stephen and my MMSC-cohort for this amazing year. You all kept me sane, in and outside the Mathematical Institute, and made me feel at home. In particular, I would like to thank Markus, Matteo and Tori for their helpful discussions over the past year and especially whilst writing this thesis.

Lastly, I am extremely grateful for the funding received from VSBfonds, Genootschap Noorthey and Stichting dr. Hendrik Muller's Vaderlandsch Fonds. Without their combined generous help this year in Oxford would not have been possible.

If our small minds, for some convenience, divide this glass of wine, this universe, into parts – physics, biology, geology, astronomy, psychology, and so on – remember that nature does not know it!
So let us put it all back together, not forgetting ultimately what it is for.
Let it give us one more final pleasure: drink it and forget it all!

Richard Feynman (1918-1988)

Abstract

One of the main deficiencies of current methods in numerical bifurcation analysis is their inability to detect solution branches without first locating a bifurcation point. As a result these methods cannot compute disconnected bifurcation branches, or branches that connect outside of the parameter domain of study. Furthermore, the detection techniques for bifurcation points are not problem-scale invariant. As a result the cost of computing bifurcation points in large-scale systems is often a prohibiting factor in the computation of bifurcation diagrams. In this thesis we propose a method to overcome these two issues.

We study the use of deflation techniques combined with Newton's method to generate multiple solution branches starting from a single initial solution. A theoretical framework for root-convergence and deflation of functions in Banach spaces is set up. In this framework we derive the first sufficient conditions for convergence towards multiple solutions if Newton's method and deflation are combined.

Deflation can be made a scalable technique and as a result could be used in tracing out bifurcation diagrams for large-scale systems. We compare arclength continuation augmented with deflation as a method to compute bifurcation diagrams with AUTO-07P, one of the most used and reliable numerical bifurcation software packages available. We find that for a range of illustrative problems AUTO-07P fails to compute complete bifurcation diagrams, whereas deflation and continuation combined yield an accurate result.

Note: this digital version contains modifications made to the paper version that has been submitted to the Examination Schools on 3 September 2015.

Contents

List of notation

$\mathbb{N}, \mathbb{R}, \mathbb{R}^+, \mathbb{C}$	Natural, real, positive real and complex numbers respectively.
$\Re(x), \Im(x)$	Real and imaginary part of x.
X, Y, Z	General Banach spaces.
$B(x, \rho), \bar{B}(x, \rho)$	Open, respectively closed, ball centred around x with radius ρ.
$L(X, Y)$	Set of bounded linear operators from X to Y.
$GL(X, Y)$	Set of invertible linear operators from X to Y.
I	Identity operator (from X to X).
$F'(x)$	Fréchet derivative of $F(x)$.
$F_x(x, y)$	Partial Fréchet derivative of $F(x, y)$ with respect to x.

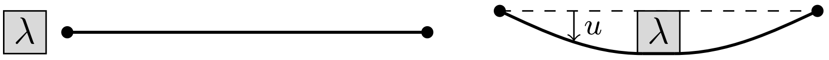
1. Introduction

As pointed out by the acclaimed physicist Eugene Wigner in his lecture *"The unreasonable effectiveness of mathematics in the natural sciences"* [47], mathematics turns up in seemingly unexpected situations and is often remarkably capable of grasping complex structures hidden beneath our observations of nature. One might therefore consider mathematics as the lingua franca among scientists as it provides a universal language able to lift physical problems into the realm of mathematical abstraction.

For many physical problems the result of this abstraction is a mathematical model which can be described by a set of equations. For this thesis we start with a very general setting, where the equations are given by

$$F(u, \lambda) = 0. \tag{1.1}$$

Without specifying the exact form of the equations F we do make a clear distinction between the roles of u and λ. We let u denote the general set of variables and λ the set of problem parameters. This difference between the two can be hard to pin down, but for most purposes we think of the variables as the measurable outcome of an experiment whereas the parameters, also called controls, govern the set-up of the experiment. One could thus say that u and λ constitute the output and input of the model respectively. Consider as a simple example the problem of the deformation of a (slender) beam under a load, see Figure 1.1. The deflection of the beam is now a variable of the problem and the external load is part of the set of control parameters.



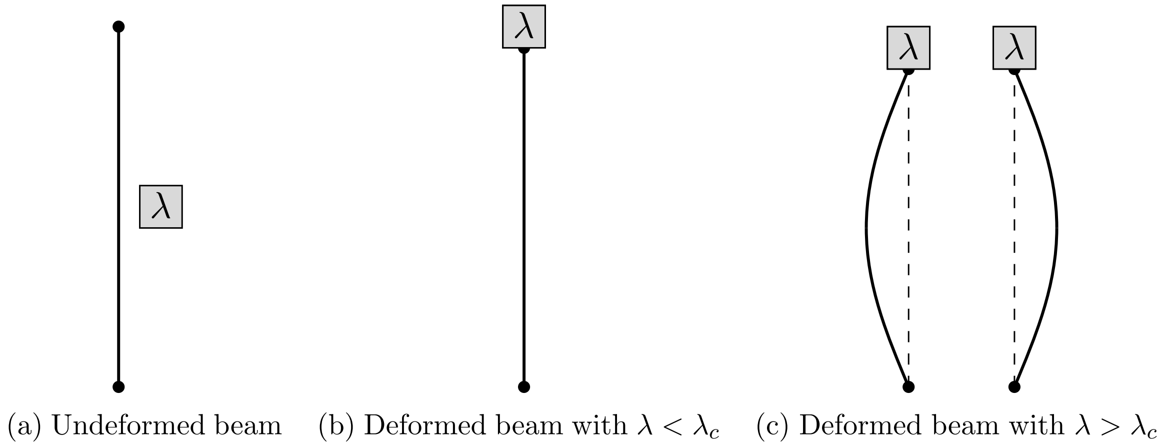
(a) Undeformed beam (b) Deformed beam

Figure 1.1: Deformation of a horizontal placed beam due to a vertical load λ.

A natural question to ask is how the output of the model, the displacement, changes if we alter one of the parameters, in this case the load λ. The reader can experimentally verify the rather uninteresting result that increasing the load increases the bending of the beam. There might be, however, a point at which the beam cannot sustain any additional small weight any longer, resulting in failure of the beam, the straw that breaks the camel's back. Here we see an example of a problem where a gradual change in the parameter can lead to a drastic change in the behaviour of its variables.

A less dramatic result is observed when we place the beam vertically and put a load on top of the beam, see Figure 1.2. Initially slowly increasing the load will not result in any bending of the beam; it will merely compress the beam in the vertical direction. There is, however, a critical value of the parameter λ_c which induces a large out of plane deformation of the beam, known as buckling.

These are examples where smoothly passing a parameter threshold value results in a qualitatively different solution, a phenomenon known as a *bifurcation*. A wide



(a) Undeformed beam (b) Deformed beam with $\lambda < \lambda_c$ (c) Deformed beam with $\lambda > \lambda_c$

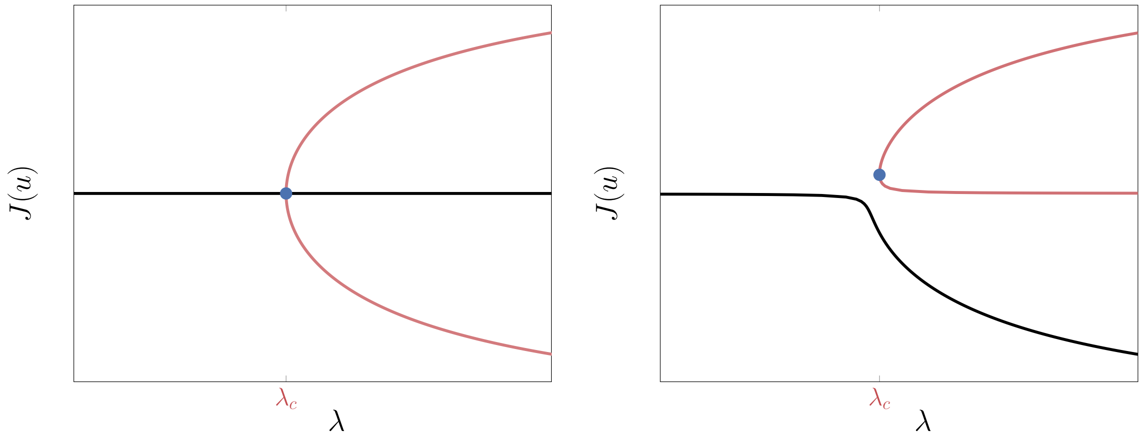
Figure 1.2: Deformation of a vertically placed beam due to a vertical load λ. The resulting deformation changes qualitatively around a critical value λ_c of the control parameter.

variety of bifurcation types exist, such as the creation or annihilation of a solution, the sudden appearance of periodic cycles or the change in stability of a solution, i.e. the response of the solution to small perturbations. For a more (mathematically) complete discussion see [3]. As the concept of a bifurcation is very general it appears in a wide range of applications ranging from continuum mechanics to ecological problems. Examples being pattern formation in chemical and biological systems, buckling of beams and structures, transitions in fluid flows such as the classical Taylor-Couette experiment, and (possibly) the Tacoma Narrows bridge collapse [43]. In engineering science bifurcations are widely studied as they potentially can induce catastrophic failure of structures.

A common way to visualise bifurcations is by means of a *bifurcation diagram*, in which we plot $J(u)$, a scalar measure of the outcome u, as a function of the parameters λ. This way we can depict the structure of the solution curves in parameter space, which are known as solution branches, and their interactions at bifurcation points. See Figure 1.3 for sketches of bifurcation diagrams.

The analysis of the solution behaviour of (1.1) as a function of the parameter set λ and the construction of an accompanying bifurcation diagram is, however, often difficult as the systems of equations for which bifurcations occur are necessarily of non-linear nature. One of the ways to nonetheless investigate the behaviour of these systems is by use of numerical methods, i.e. by numerically approximating their solutions.

In order to find a numerical solution to non-linear equations we typically use iterative methods. If we provide an initial guess of the root of (1.1) then these methods try on every iteration to find a better approximation to the root. One of the most popular methods in the class of iterative methods is Newton's method, due to its fast convergence properties close to actual roots of the equation. Note though that whereas starting close to a solution with Newton's method (or any other iterative method) can result in rapid convergence, starting just slightly too far away



(a) Connected branches

(b) Disconnected branches

Figure 1.3: Schematic bifurcation diagrams showing a scalar measure $J(u)$ of the output u as a function of the parameter λ. Bifurcation points are denoted by ● and are located at λ_c and $\tilde{\lambda}_c$ respectively. Although the diagrams bear a resemblance to each other the solution branch structure is fundamentally different in terms of connectedness.

could result in extremely fast divergence, especially in the case of highly non-linear problems.

Iterative methods provide us with the most basic framework for the computation of bifurcation diagrams, *numerical continuation*. If we are given an initial point on our solution branch we can use this point as an initial guess to compute a solution on the same solution branch for a different parameter value using iterative methods and thus continue our solution along a branch. If we take small enough steps so that the iterative methods will converge on each step we can proceed iteratively along branches and thus trace out full solution branches.

To illustrate the working of numerical bifurcation algorithms let us look at the bifurcation diagrams in Figure 1.3. Suppose we start off in the left part of the bifurcation diagrams and trace out the diagrams in the positive λ-direction. Most numerical bifurcation algorithms are able to correctly trace out the black solution branch in Figure 1.3 by making use of continuation techniques. Having found this one solution branch the next question becomes: how can we detect and trace out the red branches?

In the case of Figure 1.3a we encounter a bifurcation point on this curve and we might detect this using numerical bifurcation techniques, see for example [44]. Having found this special point we ideally want to be able to switch to the red branches in order to fully trace out the diagram. In order to do so we recall that iterative methods need to be started with sufficiently good initial guesses in order to yield convergence. Various techniques, based on results in bifurcation theory, exist to construct initial guesses for the red branches close to the bifurcation point. These are implemented in contemporary numerical bifurcation software packages and allow users to switch branches if the location of the bifurcation point is known, see for instance [28]. Current bifurcation algorithms are therefore able to successfully trace out Figure 1.3a.

The case of disconnected branches as in Figure 1.3b on the other hand often poses a more severe problem. First of all if one does not know beforehand that a disconnected branch of solutions exists it is likely that it will remain hidden in the course of the construction of the bifurcation diagram as we do not find a bifurcation point on the branch, which was our indicator in Figure 1.3a for multiple solutions. Even if we do know that extra branches of solutions for fixed λ exist the question still remains; how can we construct an initial solution on this other branch which would allow us to follow this branch using continuation? Standard branch switching techniques cannot be applied now as we have not located a bifurcation point which is needed for the standard methods to work. This leaves bifurcation diagrams like Figure 1.3b uncomputable by current numerical bifurcation algorithms.

We are thus left in the situation where multiple solutions to (1.1) for fixed λ exist but we are not able to locate a bifurcation point which prevents us from calculating the full bifurcation diagram. This gives rise to the question: can we devise a numerical technique that is able not only to find *a solution*, but also, if they exist, to find *multiple solutions*? One such technique has been known for some decades to work, but only for a special class of functions, namely scalar polynomials. Wilkinson considered a technique, now known as deflation, to find multiple roots of polynomials starting from a single initial guess [48]. Based on work by Brown and Gearhart [8] this deflation technique was recently generalised by Farrell, Birkisson and Funke [18], to apply to a more general class of functions, which provides a possible solution to the problem sketched in Figure 1.3b. In this thesis we will investigate the use of deflation in robustly tracing out bifurcation diagrams, especially those with disconnected branches. In particular, we combine deflation for fixed parameters to generate multiple solutions on distinct solution branches with standard continuation techniques to trace out branches, which allows us to find more complete bifurcation diagrams.

1.1 Outline of work

The work in this thesis can be roughly divided in two; a theoretical section and a more practical, numerical section, which are linked together by the main idea of the thesis, the deflation technique.

First in the the theoretical part we will try to answer the question of whether we can derive sufficient conditions for the deflation method in combination with Newton's method to converge to multiple solutions if they exist. In order to achieve this we will start with background material on Newton's method and deflation techniques. To incorporate a wide range of applications we look at functions $F : X \to Y$, where X and Y are Banach spaces. This allows us to discuss a very general class of problems that are of interest to the scientific community, namely that of partial differential equations (PDEs), ordinary differential equations (ODEs), integral equations and algebraic problems. We will for sake of notation consider $X = U \times \Lambda$, a product of the space of variables U and that of parameters Λ.

We first prove a new result on the convergence of Wilkinson deflation on scalar polynomials with real roots, which shows that indeed sufficient conditions can be found in this special case. In order to derive more general sufficient conditions for convergence of deflation to multiple roots we review the (local) convergence theory available for Newton's method in Banach space setting. We bring together some well and lesser known convergence theorems and their proofs and look at their similarities and differences. No such review has been published before.

Next we prove the novel result that under certain restrictions on the type of deflation, some local Newton convergence theorems are still applicable after our function has been deflated. These results then allow us to start the derivation of sufficient conditions for convergence towards multiple solutions of the deflation technique in Banach spaces.

The final chapter looks at a practical numerical bifurcation algorithm built on the deflation technique. In this chapter we will show that certain types of problems which were previously inaccessible by traditional implementations of numerical bifurcation techniques, such as disconnected branches, can be tackled with the addition of deflation. We will benchmark our implementation for some illustrative problems with one of the most robust and reliable numerical bifurcation software packages available, AUTO-07P [15]. We find that our algorithm succeeds in cases where the methods in AUTO-07P fail, demonstrating the power of the approach.

2. Background on Newton's method and deflation

For this thesis we will look at

$$F(u, \lambda) = 0, \tag{2.1}$$

in the general setting of Banach spaces so that its theory can be applied to a wide variety of problems, including ODEs and PDEs. In order to simplify notation, let the Banach spaces of the variables and parameters be denoted by U and Λ respectively and let $X = U \times \Lambda$ be their product space.

In order to study (2.1) we can therefore consider X and Y general Banach spaces and $F : X \to Y$. Solving (2.1) now translates to finding the roots of F, i.e. $x \in X$ that satisfy $F(x) = 0$, which can now be amongst other things scalar solutions to algebraic problems or functions solving a differential equation.

First we will introduce a common technique, Newton's method, to approximate a solution to $F(x) = 0$. This method is only concerned with finding a single root to this equation. There are, however, functions F for which multiple roots do exist. Particularly of interest to us is the problem of detection of multiple branches in the construction of bifurcation diagrams. A natural question thus is whether there exist methods to find more than one root. To this end we will introduce one computational technique, deflation, which can make, under certain conditions, Newton's method find multiple roots starting from just one initial guess.

2.1 Newton's method

Exact solutions to the root-finding problem are often hard to acquire. One can, however, try to approximate the roots by means of various schemes. A popular class of approximation schemes uses an iterative method of the form,

$$x_{k+1} = x_k + \Delta x_k, \tag{2.2}$$

where the update Δx_k depends on the type of iterative method. Starting from an initial point x_0 one hopes to produce a sequence by the iterative method which converges towards a root of F.

A specific and widely studied iterative method is Newton's method, or the Newton-Raphson method. The classical version of Newton's method defines the updates by the equation

$$F'(x_k)\Delta x_k = -F(x_k), \tag{2.3}$$

where $F'(x)$ denotes the Fréchet derivative of $F(x)$, which we have to assume to exist in order for Newton's method to be well-defined. One motivation for this specific form of the update can be found by looking at the local Taylor expansion of F around the current iterate x_k.

Affine transformation invariance

An important observation about Newton's method is its behaviour under affine transformations of the domain or codomain of the function F. If $A \in GL(Y)$, then one can see from (2.3) that the functions $F(x)$ and $AF(x)$ yield exactly the same Newton sequence, i.e. Newton's method is affine covariant. Furthermore we see that if $B \in GL(X)$, then the functions $F(x)$ and $G(y) = F(By)$ yield Newton sequences $\{x_k\}$ and $\{y_k\}$ that are simply related by a scaling $x_k = By_k$. Newton's method is thus said to be affine contravariant as well. These properties have important implications, as we will see later on.

2.2 Deflation techniques

The aforementioned Newton's method can compute approximate roots starting from an initial guess x_0. If we are to find multiple solutions starting from x_0 we will have to extend our method and one possible candidate to do so is deflation. To introduce the general technique of deflation we first consider a specific type of deflation which can be applied to find the roots of scalar polynomials, now known as Wilkinson deflation, which was the first deflation technique developed [48].

2.2.1 Wilkinson deflation of polynomials

Suppose we are given a scalar polynomial p with distinct roots $r_1, \cdots, r_N \in \mathbb{C}$ so that we can write $p(x) = \prod_{i=1}^{N}(x - r_i)$. If by some method we have acquired a set of roots with indices $i \in \mathcal{S}$ then we can attempt to find the unknown roots by our original method if we can hide the known roots from our root-finder. In the case of scalar polynomials this can be achieved by considering a deflated polynomial

$$q(x) = \frac{p(x)}{\prod_{i \in \mathcal{S}}(x - r_i)}, \tag{2.4}$$

where we have now filtered the known roots from the original polynomial by polynomial division. Applying our root-finding technique of choice to this deflated polynomial will, if it converges, converge to a root which has not yet been found.

The practical implementation of this technique has to overcome several issues due to the ill-conditioning of the root-finding problem of polynomials and finite arithmetic, see for example [36, 48], but we will not explore this matter further here. Instead we will now look at the combination of Wilkinson deflation with Newton's method and show that this combination can provably yield multiple roots starting from a single initial guess, a result which has not been published before.

Convexity & global convergence

If we have a polynomial with distinct real roots, then we can prove that there exist infinitely many points which will converge to all the roots of the polynomial by making use of local convexity of the polynomial. The rate of convergence is, however, not specified. It is thus not necessarily quadratic, the fast convergence rate which can be proven by the local convergence theorems in the next chapter. In order to prove this statement we need the following lemma, which uses the geometric interpretation of Newton's method and the convexity of a function.

Lemma 1. *Let $f : \mathcal{I} \subseteq \mathbb{R} \to \mathbb{R}$ be a convex, differentiable function where $\mathcal{I}$ is an open interval such that there exists precisely one $x^* \in \mathcal{I}$ with $f(x^*) = 0$. Starting from $x_0 \in \mathcal{I}$ with $f'(x_0) \neq 0$, Newton's method converges to x^* if $x_1 \in \mathcal{I}$.*

A similar lemma can of course be formulated for concave functions. For a proof, see Appendix A.1. With this lemma we are now able to prove the following corollary.

Corollary 1 (Wilkinson deflation for purely real roots)**.** *Let $p(x)$ be a polynomial with N distinct real roots*

$$p(x) = \sigma \prod_{i=1}^{N} (x - r_i), \tag{2.5}$$

such that $r_1 < r_2 < \cdots < r_N$ and $\sigma \in \mathbb{R}$. Starting from $x_0 \geq r_N$ or $x_0 \leq r_1$ Newton's method with Wilkinson deflation converges to all roots of $p(x)$.

Note that this theorem could be extended to complex polynomials with collinear roots in the complex plane, but we will not consider this further here.

Proof. W.l.o.g. assume that $\sigma = 1$ as Newton's method is affine invariant and consider the case $x_0 \geq r_N$. The case $x_0 \leq r_1$ automatically follows from taking the transformation $x \mapsto -x$ and applying Newton's method to the transformed polynomial.

As polynomials are twice differentiable we can use the second derivative to look at the convexity of the polynomial

$$p''(x) = 2 \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \prod_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{N} (x - r_k). \tag{2.6}$$

Assuming $N > 1$ it follows that $p''(x) > 0$ if $x \geq r_N$ and thus that p is convex on $[r_N, \infty)$. From the continuity of p and p'' it follows that there exists an open interval $\mathcal{I} \subset \mathbb{R}$ such that $r_N, x_0 \in \mathcal{I}$, $r_{N-1} \notin \mathcal{I}$ and p is convex on $\mathcal{I}$. In order to apply lemma 1 it remains to show that $x_1 \in \mathcal{I}$. If $x_0 = r_N$ we are done as the Newton iterates will repeatedly yield $r_N \in \mathcal{I}$. So let us assume $x_0 > r_N$. As $p'(x_0) > 0$ and $p(x_0) > 0$ we know that $x_1 < x_0$. The tangent at x_0 is given by $l(x) = p(x_0) + p'(x_0)(x - x_0)$ and from convexity we know that for all $x \in \mathcal{I}$ we have $p(x) \geq l(x)$. Suppose $x_1 < r_N$. By definition of x_1 we have $l(x_1) = 0$ and $l(x_0) = p(x_0) > 0$. As a result $l(r_N) > 0$, but as $r_N \in \mathcal{I}$ this implies $0 = p(r_N) \geq l(r_N) > 0$ which is a contradiction. Therefore we conclude that $r_N \leq x_1 \leq x_0$ and thus $x_1 \in \mathcal{I}$.

By applying the above reasoning to a polynomial which deflates r_N we can continue deflating roots and using Newton's method. The result will be convergence (in decreasing order) to $r_{N-1}, \cdots, r_2$. By deflating these roots we arrive at $q(x) = x - r_1$ which is a linear polynomial. Therefore Newton's method will yield $x = r_1$ after one iteration and as a result x_0 will have converged to all the roots of $p(x)$. □

Note that this theorem implies that there are indeed functions which allow one to find multiple roots starting from a fixed x_0, but the class of functions is unfortunately rather limited and the theorem does not have a natural extension into the more general Banach space setting. Therefore we will now look at a generalisation of Wilkinson deflation which is applicable in Banach space setting.

2.2.2 General deflation in Banach space setting

Instead of considering a scalar polynomial we return to the problem of a general (non-linear) function $F : X \to Y$, where X and Y are Banach spaces. Brown and Gearhart [8] generalised the technique of Wilkinson deflation for the case where X and Y are finite dimensional real or complex vector spaces, say $\mathbb{R}^n$ or $\mathbb{C}^n$, by considering deflation matrices of the form

$$M(x; r) = \frac{A}{\|x - r\|}, \tag{2.7}$$

where A is a non-singular matrix. This concept of deflation matrices has a natural extension to our previous framework of general Banach spaces, as was set out by Farrell, Birkisson and Funke [18].

Definition 1 (Deflation operator on a Banach space [6])**.** *Let X, Y and Z be Banach spaces, and $D \subseteq X$ be an open subset. Let $F : D \subset X \to Y$ be a Fréchet differentiable operator with derivative F'. For each $r \in D$, let $\mathcal{M}(\cdot; r) : D\backslash\{r\} \to GL(Y, Z)$. We say that $\mathcal{M}$ is a* deflation operator *if for any F such that $F(r) = 0$ and $F'(r)$ is nonsingular, we have*

$$\liminf_{i \to \infty} ||\mathcal{M}(x_i; r)F(x_i)|| > 0 \tag{2.8}$$

for any sequence $\{x_i\}$ converging to r, $x_i \in D \setminus \{r\}$.

The most common and practical case is where the deflation operator is a linear operator in $GL(Y)$. For this thesis we will use a generalisation of the deflation matrices by Brown and Gearhart, namely the class of shifted deflation operators.

Definition 2 (Shifted deflation [18])**.** *Shifted deflation specifies*

$$\mathcal{M}_{p,\alpha}(x; r) = \left(\frac{1}{||x - r||^p} + \alpha\right) I, \tag{2.9}$$

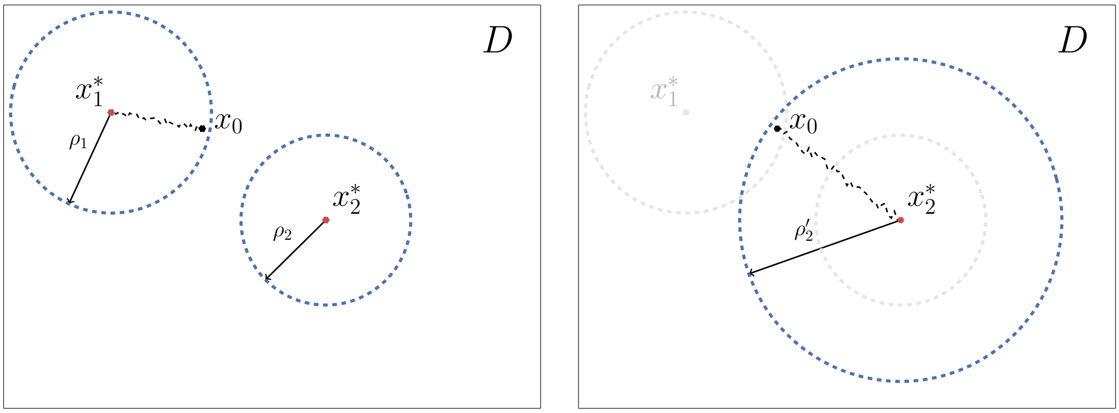
where $\alpha \geq 0$ is the shift, $p \geq 1$ and I is the identity operator on Y.

The shift and exponent parameter of the deflation operator can be chosen so as to improve the convergence of the root-finding technique applied to the deflated function, see [18]. Note that any other $A \in GL(Y)$ can be chosen instead of $I \in GL(Y)$, but as Newton's method is affine covariant this would not change the Newton sequences.

Convergence to multiple roots

The objective of our theoretical endeavour is to derive conditions which can guarantee us to find multiple roots, starting from a single initial guess x_0. Given an arbitrary initial guess and function F then we are, however, not guaranteed to even find a single approximate root of F using Newton's method. It is therefore of interest to see under what conditions we can establish guaranteed convergence of Newton's method. To this end we present in the next chapter a review of various convergence theorems for the classical version of Newton's method as defined by (2.2,2.3).

These local convergence theorems will yield open neighbourhoods around roots of the function which form regions of guaranteed convergence. Given a region $D \subset X$ with two roots x_1^* and x_2^* we can then sketch the general idea for a convergence result towards multiple roots. As deflation is changing the objective function we expect these convergence regions to change and we therefore look for conditions in which the regions before and after deflation have a non-empty overlap for different roots, as in Figure 2.1, so that any point in this overlap will converge to at least two solutions.



(a) Convergence before deflation

(b) Convergence after deflation

Figure 2.1: Illustration of local convergence regions around roots. The approach we are taking in deriving sufficient conditions for convergence of x_0 towards multiple roots is sketched here. The initial guess x_0 lies within a convergence region for x_1^* initially. After deflation x_0 will no longer converge to x_1^*, but instead lies within a new region of convergence around x_2^*, which has increased relative to the convergence region for the undeflated function.

3. A review of local convergence theorems for Newton's method

Recall Newton's method for finding approximate solutions to $F(x) = 0$

$$x_{k+1} = x_k + \Delta x_k, \tag{3.1a}$$

$$F'(x_k)\Delta x_k = -F(x_k), \tag{3.1b}$$

which we need to initialise by an initial guess x_0. We hope to get a Newton sequence which converges to a root of F, but as mentioned before Newton's method is not guaranteed to succeed in doing so. Several sufficient conditions do exist which can guarantee convergence of a root if we start in a neighbourhood close to a root and we present here a non-exhaustive review of various local convergence theorems for the classical version of Newton's method as defined by (3.1) in finding solutions to $F(x) = 0$.

Note that our general objective is finding sufficient conditions for convergence to multiple solutions, an idea sketched in Figure 2.1, which we will keep in mind while looking at the different approaches of the theorems.

3.1 Classical theorems

As can be seen from (3.1) Newton's method relies upon the solvability of a linear system involving the Fréchet derivative of F at the current iterate. In order to guarantee a solution to the Newton update equation (3.1b) one often requires invertibility of this Fréchet derivative. The following standard result from linear functional analysis gives sufficient conditions in order for a linear operator to be invertible.

Lemma 2 (Banach perturbation lemma)**.** *Let $T : X \to Y$ be a linear operator such that $\|T\| < 1$. Then $I + T$ is invertible and*

$$\frac{1}{1 + \|T\|} \leq \|(I + T)^{-1}\| \leq \frac{1}{1 - \|T\|}. \tag{3.2}$$

For a proof, see, for example, [42, Theorem 4.40]. This will be a key result in the following convergence theorems as it can be used to show that (3.1) has a well-defined solution. Another important result is the extension of the mean-value theorem to functions on Banach spaces.

Lemma 3 (Mean-value theorem)**.** *Let $F : X \to Y$ be a continuously Fréchet differentiable operator on the open subset $D \subseteq X$. For any $x, y \in D$ we have*

$$F(x) - F(y) = \int_0^1 F'(tx + (1 - t)y)(x - y)\,\mathrm{d}t. \tag{3.3}$$

For a proof, see, for example, [26, Theorem 13.3].

The first theorem is the classical convergence theorem by Kantorovich [27], who first brought convergence theorems into the framework of functions on Banach spaces.

Theorem 1 (Newton-Kantorovich [27]). *Let $F : D \to Y$ be a continuously Fréchet differentiable function on the open convex subset $D \subseteq X$. Starting at $x_0 \in D$, assume that*

i) $F'(x_0)^{-1}$ *exists and set* $\|F'(x_0)^{-1}\| = \beta$, $\|F'(x_0)^{-1}F(x_0)\| = \alpha$,

ii) $\|F'(x) - F'(y)\| \leq \gamma\|x - y\|$ *for all* $x, y \in D$ *(Lipschitz continuity of* $F'(x)$*),*

iii) $h_0 = \alpha\beta\gamma \leq \frac{1}{2}$,

iv) $\mathcal{B} = \bar{B}(x_0, \rho_0) \subset D$ *for* $\rho_0 = \frac{1-\sqrt{1-2h_0}}{\gamma\beta} = \frac{2\alpha}{1+\sqrt{1-2h_0}}$.

Then the Newton sequence defined by (3.1) is well-defined and remains within $\mathcal{B}$*. The Newton sequence converges to a* $x^* \in \mathcal{B}$ *with* $F(x^*) = 0$*. Furthermore, if we define* $\rho^+ = \frac{1+\sqrt{1-2h_0}}{\gamma\beta}$*, then* x^* *is unique within* $D \cap B(x_0, \rho^+)$*.*

The above theorem turns out to be versatile as it has found use in convergence proofs of numerical algorithms as well as existence and uniqueness results of roots in (non-linear) functional analysis. One of the main advantages of the theorem is that its assumptions are mostly checked at the initial guess x_0 and an open neighbourhood around it and thus yields the option of an a-priori check of convergence. One does not need to assume the existence of a root beforehand either.

The last part of the Newton-Kantorovich theorem, regarding uniqueness, is often viewed as a benefit if interested in existence and uniqueness of roots. However, in our case, we are interested in a number of roots and not just a single root. The Newton-Kantorovich uniqueness result then forms a potential downside as it can limit the size of the convergence balls $\mathcal{B}$ in the case of multiple roots, which by the theorem are not allowed to overlap.

We present a modification by Ortega of one of the original proofs by Kantorovich using majorant functions in full detail as this illustrates some common proof steps for other convergence theorems as well. First we need to introduce two lemmas.

Lemma 4 (Majorant sequences). *Let* $\{x_k\}$ *be a sequence in* X *and* $\{t_k\}$ *a sequence in* $\mathbb{R}^+$ *with the property*

$$\|x_{k+1} - x_k\| \leq t_{k+1} - t_k, \tag{3.4}$$

where $t_k \to t^*$ *with* $t^* < \infty$*. Then there exists a* $x^* \in X$ *such that* $x_k \to x^*$*.*

The proof of this lemma shows that $\{x_k\}$ is a Cauchy sequence in a Banach space and thus must have a limit in X. The sequence $\{x_k\}$ is said to be majorised by $\{t_k\}$.

Lemma 5 (Invertibility of the Fréchet derivative). *Let* $F : D \to Y$ *be a continuously Fréchet differentiable function on the open convex subset* $D \subseteq X$*. Let* $\tilde{x} \in D$ *and assume that* *i)* $F'(\tilde{x})^{-1}$ *exists and set* $\|F'(\tilde{x})^{-1}\| = \beta$,

ii) $\|F'(x) - F'(y)\| \leq \gamma \|x - y\|$ *for all* $x, y \in D$ *(Lipschitz continuity of* $F'(x)$*).*

Then for all $x \in B(\tilde{x}, 1/\beta\gamma)$ *the Fréchet derivative* $F'(x)$ *is invertible and its norm is bounded by*

$$\|F'(x)^{-1}\| \leq \frac{\beta}{1 - \beta\gamma\|x - \tilde{x}\|}. \tag{3.5}$$

Proof. If $x \in B(\tilde{x}, 1/\beta\gamma)$ we find that $G(x) = I - F'(\tilde{x})^{-1}F'(x)$ satisfies $\|G\| < 1$ by use of the Lipschitz continuity of the Fréchet derivative on D. By lemma 2 we then know that $(I + G)^{-1}$ exists, i.e. $(F'(\tilde{x})^{-1}F'(x))^{-1}$ is well-defined, which proves that $F'(x)^{-1}$ exists. The bound on the norm follows from (3.2). □

With these lemmas we can now give the proof by Ortega which constructs an explicit majorant sequence to the Newton sequence to prove convergence.

Proof of Theorem 1 [34]. Observation 1 (Difference in Newton iterates)
Given x_k for $k = 1, \ldots, N$ and using lemma 5 we find that the difference between the next iterates satisfies

$$\begin{aligned} \|x_{N+1} - x_N\| &= \|F'(x_N)^{-1}F(x_N)\| && (3.6)\\ &\leq \|F'(x_N)^{-1}\| \; \|F(x_N)\| && (3.7)\\ &\leq \frac{\beta}{1 - \beta\gamma\|x_N - x_0\|}\|F(x_N)\|. && (3.8) \end{aligned}$$

To find an upper bound for the last term note that we can apply the mean-value theorem and then use convexity of D to apply the Lipschitz continuity

$$\begin{aligned} \|F(x_N)\| &= \|F(x_N) - F(x_{N-1}) - F'(x_{N-1})(x_N - x_{N-1})\| && (3.9)\\ &= \|\int_0^1 [F'(sx_N + (1-s)x_{N-1}) - F'(x_{N-1})] \, \mathrm{d}s \; (x_N - x_{N-1})\| && (3.10)\\ &\leq \int_0^1 \|F'(sx_N + (1-s)x_{N-1}) - F'(x_{N-1})\| \, \mathrm{d}s \; \|x_N - x_{N-1}\| && (3.11)\\ &\leq \frac{\gamma}{2}\|x_N - x_{N-1}\|^2. && (3.12) \end{aligned}$$

Note that we have used (3.1) at step $N - 1$ in the first line.

Observation 2 (Construction of an explicit majorant sequence)
Let $\phi(t) = \frac{\beta\gamma}{2}t^2 - t + \alpha$. Note that this is a convex quadratic. By $h_0 \leq 1/2$ its roots, given by ρ_0 and ρ^+, are real. Then applying Newton to this function will result in a monotonically converging sequence to the roots. If we start with $t_0 = 0$ we will thus create a sequence $\{t_k\}$ such that $t_k \uparrow \rho_0$ as $\rho_0 \leq \rho^+$. This sequence is given by

$$t_{k+1} = t_k - \frac{\frac{\beta\gamma}{2}t_k^2 - t_k + \alpha}{\beta\gamma t_k - 1}. \tag{3.13}$$

Claim: $\{t_k\}$ is a majorising sequence for the Newton sequence started from x_0. Note that $t_1 = \alpha$ and $\|x_1 - x_0\| = \|F'(x_0)^{-1}F(x_0)\| = \alpha$. Now suppose that for all $k = 0, 1, \dots N$ the Newton sequence is majorised, i.e. $\|x_k - x_{k-1}\| \leq t_k - t_{k-1}$.

Then we find by observation 1 that the difference in Newton iterates is bounded by

$$\|x_{N+1} - x_N\| \leq \frac{\beta\gamma}{2 - 2\beta\gamma\|x_N - x_0\|}\|x_N - x_{N-1}\|^2 \leq \frac{\beta\gamma/2}{1 - \beta\gamma t_N}(t_N - t_{N-1})^2. \quad (3.14)$$

Using the definition of the Newton sequence from $\phi(t)$ and

$$t_{k+1} = \frac{t_k}{2} + \frac{\frac{t_k}{2} - \alpha}{\beta\gamma t_k - 1}, \quad (3.15)$$

we can deduce that the last term in (3.14) is equal to $t_{N+1} - t_N$, which proves the claim. Therefore there exists a $x^* \in D$ such that $x_k \to x^*$. Additionally we have that $\|x_N - x_0\| \leq \sum_0^{N-1} \|x_{i+1} - x_i\| \leq t_N - t_0 = t_N \leq \rho_0$ so that the Newton sequence remains in $\mathcal{B}$. Thus $x^* \in \mathcal{B}$, and from convergence of Newton's method it follows that $F(x^*) = 0$.
If $h_0 < 1/2$ we find that $\rho_0 < (\beta\gamma)^{-1}$ and thus from (3.14) that the convergence is in fact quadratic by using that $\|x_N - x_0\| \leq \rho_0$.

Observation 3 (Uniqueness of solution)
We define $\mathcal{U} = D \cap B(x_0, \rho^+)$ and $H(x) = F'(x_0)^{-1}F(x)$. Then $H'(x) = F'(x_0)^{-1}F'(x)$ and as a result $H'(x_0) = I$. Note that H is Lipschitz continuous on D as well with Lipschitz constant $\beta\gamma$.

Suppose there exist $a, b \in B(x_0, \frac{1}{\beta\gamma})$ such that $F(a) = F(b) = 0$, but $a \neq b$. Then we find

$$\|a - b\| = \|H(a) - H(b) - I(a - b)\| = \|H(a) - H(b) - H'(x_0)(a - b)\|. \quad (3.16)$$

By use of the mean value theorem we then find

$$\begin{aligned}
\|a - b\| &\leq \int_0^1 \|H'(sa + (1-s)b) - H'(x_0)\| \,\mathrm{d}s \|a - b\| && (3.17)\\
&\leq \sup_{s\in[0,1]} \|H'(sa + (1-s)b) - H'(x_0)\| \; \|a - b\| && (3.18)\\
&\leq \beta\gamma\|a - b\| \sup_{s\in[0,1]} \|sa + (1-s)b - x_0\| && (3.19)\\
&= \beta\gamma\|a - b\| \sup_{s\in[0,1]} \|s(a - x_0) + (1-s)(b - x_0)\| && (3.20)\\
&\leq \beta\gamma\|a - b\| \max\left(\|a - x_0\|, \|b - x_0\|\right) && (3.21)\\
&< \frac{\beta\gamma}{\beta\gamma}\|a - b\| = \|a - b\|. && (3.22)
\end{aligned}$$

Which is a contradiction, so if there exists such a, b they must be unique (this in fact already proves uniqueness in $\mathcal{B}$).

To prove uniqueness in $\mathcal{U}$ we show that in $\mathcal{U} \setminus \mathcal{B}$ there exists no zero in the case $h_0 < 1/2$ (for the case $h_0 = 1/2$, see [9]). Note that now

$$\|H(x) - H(x_0) + H'(x_0)(x - x_0)\| \leq \frac{\beta\gamma}{2}\|x - x_0\|^2, \tag{3.23}$$

by use of Lipschitz continuity and the mean value theorem. From the reverse triangle inequality we then deduce that

$$\begin{aligned} \|H(x)\| &\geq -\frac{\beta\gamma}{2}\|x - x_0\|^2 - \|H(x_0)\| + \|x - x_0\| && (3.24) \\ &= -\frac{\beta\gamma}{2}\|x - x_0\|^2 - \alpha + \|x - x_0\| && (3.25) \\ &= -\phi(\|x - x_0\|) > 0, && (3.26) \end{aligned}$$

since $\rho_0 < \|x - x_0\| < \rho^+$ and thus ϕ is negative. As a result we cannot have $F(x) = 0$ and thus there is no zero of F in $\mathcal{U} \setminus \mathcal{B}$, which proves uniqueness of x^* in the ball $B(x_0, \rho^+)$. □

Under slightly different assumptions, strengthening condition *i)* and weakening condition *iii)*, we arrive at another classical theorem by Mysovskikh, a contemporary of Kantorovich. A part of the original paper has been often omitted in literature [11, 12, 21, 49], which states conditions for uniqueness, but we will include this here for completeness.

Theorem 2 (Newton-Mysovskikh [32]). *Let $F : D \to Y$ be a continuously Fréchet differentiable function on the convex subset $D \subseteq X$ with invertible Fréchet derivative $F'(x)$ for all $x \in D$. Starting at $x_0 \in D$ let $\alpha = \|F'(x_0)^{-1}F(x_0)\|$ and assume that*

i) $\|F'(x)^{-1}\| \leq \beta$ for all $x \in D$,

ii) $\|F'(x) - F'(y)\| \leq \gamma\|x - y\|$ for all $x, y \in D$ (Lipschitz continuity of $F'(x)$),

iii) $h_0 = \alpha\beta\gamma < 2$,

iv) $\mathcal{B} = \bar{B}(x_0, \rho_0) \subset D$ for $\rho_0 = \alpha H$, where $H = \sum_{k=0}^{\infty} \left(\frac{h_0}{2}\right)^{2^k - 1} \leq \frac{1}{1 - h_0/2}$.

Starting at $x_0 \in \mathcal{B}$, the Newton sequence defined by (3.1) is well-defined and remains within $\mathcal{B}$. The Newton sequence converges to a $x^* \in \mathcal{B}$ for which $F(x^*) = 0$. Furthermore, if $h_0 < \tilde{h}$, where $\tilde{h} \approx 0.71$ is the solution to $H = 1/h_0$, then the solution is unique within $\mathcal{B}$.

The proof is based on the original proof by Mysovskikh (in Russian) and supplemented by the often overlooked uniqueness result found in the original article.

Proof [32]. We use (3.7) and (3.12) from the previous proof with the boundedness of the Fréchet derivative to derive

$$\|x_{N+1} - x_N\| \leq \frac{\beta\gamma}{2}\|x_N - x_{N-1}\|^2. \tag{3.27}$$

Because of the assumption for Newton-Mysovskikh we can now find an explicit upper-bound for $\|x_{N+1} - x_N\|$

$$\|x_{N+1} - x_N\| \leq \left(\frac{\beta\gamma}{2}\right)^{2^N-1} \|x_1 - x_0\|^{2^N} = \left(\frac{\beta\gamma}{2}\right)^{2^N-1} \alpha^{2^N} = \alpha \left(\frac{h_0}{2}\right)^{2^N-1}. \quad (3.28)$$

As the base case $N = 0$ is true a proof by induction can show that this must be true for all the elements in the Newton sequence. This in turn can be used to deduce that

$$\|x_{N+1} - x_0\| \leq \sum_{i=0}^{N} \|x_{i+1} - x_i\| \leq \alpha \sum_{i=0}^{N} \left(\frac{h_0}{2}\right)^{2^i-1} \leq \alpha H = \rho_0, \quad (3.29)$$

and thus $x_k \in \mathcal{B}$ for all k. A slight modification of the above argument can show that the Newton sequence is a Cauchy sequence and therefore must converge towards an $x^* \in \mathcal{B}$. By the definition of the Newton iterates the point x^* must then satisfy $F(x^*) = 0$.

Finally, let us assume the conditions for uniqueness, i.e. $h_0 < \tilde{h}$, then we see that $\rho_0 = \alpha H < \frac{\alpha}{h_0} = \frac{1}{\beta\gamma}$. Now the uniqueness of x^* in $\mathcal{B} = \bar{B}(x_0, \rho_0) \subset B\left(x_0, \frac{1}{\beta\gamma}\right)$ follows immediately from the result in the Newton-Kantorovich theorem. □

The uniqueness result is often forgotten or omitted in literature which could give one the impression that there is a fundamental difference with Newton-Kantorovich. We see, however, that Newton-Mysovskikh proves existence of a root and shares many other characteristics with Newton-Kantorovich, such as α, γ. So naturally the question arises how they differ. Apart from stricter restriction on the Fréchet derivative in the case of Newton-Mysovskikh the main difference is the constraint on h_0 and the resulting size of the convergence ball. Note that under the condition $h_0 \leq 1/2$, i.e. the case where condition *iii)* holds for both theorems, the size of the convergence balls is larger for Newton-Kantorovich. Newton-Mysovskikh can, however, be used to prove uniqueness for a wider range of h_0.

The previous two classical theorems derive convergence balls centred at the initial guess x_0 without assuming the existence of a root x^*. In the case that one is given that roots of the function F exist, a different approach can be taken, which uses evaluations at the roots. This approach might be more natural for our purposes as we are considering the case where we assume that multiple solutions do exist and are thus not concerned with existence proofs of roots.

Theorem 3 (Rall-Rheinboldt [39, 40])**.** *Let $F : D \to Y$ be a continuously Fréchet differentiable function on the open convex subset $D \subseteq X$. Suppose that there exists an $x^* \in D$ such that $F(x^*) = 0$. Assume that*

i) $F'(x^*)^{-1}$ exists and set $\beta = \|F'(x^*)^{-1}\|$,

ii) $\|F'(x) - F'(y)\| \leq \gamma\|x - y\|$ for all $x, y \in D$ (Lipschitz continuity of $F'(x)$).

Then any $\rho^* \leq 2/(3\gamma\beta)$ such that $\mathcal{B} = B(x^*, \rho^*) \subset D$ has the property that starting at $x_0 \in \mathcal{B}$, the Newton sequence defined by (3.1) is well-defined and remains within $\mathcal{B}$. The Newton sequence converges to $x^* \in \mathcal{B}$. Furthermore, if we define $\rho^+ = \frac{1}{\gamma\beta}$, then x^* is unique within $D \cap B(x^*, \rho^+)$.

The proof is very similar to the proof of the Newton-Kantorovich theorem but uses x^* instead of x_0.

Proof [40]. If $x, y \in \mathcal{B}$ we find by lemma 5 a bounded Fréchet derivative

$$\|F'(x)^{-1}\| \leq \frac{\beta}{1 - \beta\gamma\|x - x^*\|} \leq 3\beta. \tag{3.30}$$

Suppose that we have a Newton sequence $\{x_k\}$ for $k = 1, \ldots, N$ such that $x_k \in \mathcal{B}$. Similar to the Newton-Kantorovich proof we find that the next iterate satisfies

$$\begin{align} \|x_{N+1} - x^*\| &= \|F'(x_N)^{-1} \left[F(x^*) - F(x_N) + F'(x_N)(x_N - x^*)\right] \| \tag{3.31} \\ &\leq 3\beta\|F(x^*) - F(x_N) + F'(x_N)(x_N - x^*)\| \tag{3.32} \\ &\leq \frac{3\beta\gamma}{2}\|x_N - x^*\|^2 \tag{3.33} \\ &\leq \|x_N - x^*\|^2/\rho^*. \tag{3.34} \end{align}$$

As $x_N \in \mathcal{B}$ we find $\|x_N - x^*\|/\rho^* < 1$ and thus $\|x_{N+1} - x^*\| < \|x_N - x^*\|$, which proves that the sequence remains in $\mathcal{B}$ as the base case $k = 0$ is satisfied trivially. Additionally we observe that $x_k \to x^*$, at at least a quadratic rate.

Uniqueness is proved by defining $H(x) = F'(x^*)^{-1}F(x)$ and follows the same proof as before. □

The Rall-Rheinboldt significantly differs from the previous two theorems and appears not to be as widespread in literature, perhaps due to its lack of existence results. This might, however, work as an advantage in our case as we mentioned earlier. The framework of the theorem, which is now centred around x^*, naturally yields so called basins of attraction for the roots x^*, regions in X which are guaranteed to converge to x^*. There is still a uniqueness result, but as this is now centred around a root instead of an initial guess this does not form a disadvantage for our purposes as was sketched in Figure 2.1.

There is, however, a major problem with these classic results, as pointed out by Deuflhard [12]. The Newton sequences are invariant under affine transformations as we saw in Section 2.1. As a result we have to use $\beta(A)$ and $\gamma(A)$, because the constants depend on the linear transformation chosen. One can easily observe that $\beta(A) \leq \beta(I)\|A^{-1}\|$ and $\gamma(A) \leq \gamma(I)\|A\|$ which therefore yields that $\beta(A)\gamma(A) \leq \beta(I)\gamma(I)\text{cond}(A)$ where $\text{cond}(A)$ is the condition number of A. As the Newton balls all depend in one way or another on $1/\beta\gamma$ we can thus make these radii of the theorems shrink to zero by a suitable choice of A, even though the Newton sequences still converge. The constants appearing in the classical theorems can therefore not be fundamental constants and we need to set the theorems in a different framework.

3.2 Affine covariant theorems

When monitoring the error $\|x_N - x^*\|$ as a measure for convergence, the natural framework to cope with convergence results is that of affine covariant theorems. To do so we note that by combining assumptions in the preceding Newton theorems we can cast these theorems in affine covariant form, so that multiplying by a $A \in GL(Y)$ does not affect the statement of the theorem.

Theorem 4 (Affine covariant Newton-Kantorovich)**.** *Let $F : D \to Y$ be a continuously Fréchet differentiable function on the open convex subset $D \subseteq X$. Starting at $x_0 \in D$, assume that*

i) $F'(x_0)^{-1}$ *exists and set* $\alpha = \|F'(x_0)^{-1}F(x_0)\|$,

ii) $\|F'(x_0)^{-1}(F'(x) - F'(y))\| \leq \omega_0\|x - y\|$ *for all* $x, y \in D$
(affine covariant Lipschitz continuity of $F'(x)$*),*

iii) $h_0 = \alpha\omega_0 \leq \frac{1}{2}$,

iv) $\mathcal{B} = \bar{B}(x_0, \rho_0) \subset D$ *for* $\rho_0 = \frac{1-\sqrt{1-2h_0}}{\omega_0}$.

Then the Newton sequence defined by (3.1) is well-defined and remains within $\mathcal{B}$*. The Newton sequence converges to an* $x^* \in \mathcal{B}$ *with* $F(x^*) = 0$*. Furthermore, if we define* $\rho^+ = \frac{1+\sqrt{1-2h_0}}{\omega_0}$*, then* x^* *is unique within* $D \cap B(x_0, \rho^+)$*.*

The proofs of the affine covariant theorems follow the exact same lines as that of the classical theorems and we will therefore only highlight some small differences if necessary. In the above theorem the proof just follows from the classical version if one lumps β, γ together in ω_0.

Theorem 5 (Affine covariant Newton-Mysovskikh)**.** *Let $F : D \to Y$ be a continuously Fréchet differentiable function on the convex subset $D \subseteq X$ with invertible Fréchet derivative $F'(x)$ for all $x \in D$. Starting at $x_0 \in D$ let $\alpha = \|F'(x_0)^{-1}F(x_0)\|$ and assume that*

i) $\|F'(z)^{-1}(F'(x) - F'(y))\| \leq \omega\|x - y\|$ *for all collinear* $x, y, z \in D$
(affine covariant Lipschitz continuity of $F'(x)$*),*

ii) $h_0 = \alpha\omega < 2$,

iii) $\mathcal{B} = \bar{B}(x_0, \rho_0) \subset D$ *for* $\rho_0 = \alpha H$*, where* $H = \sum_{k=0}^{\infty}\left(\frac{h_0}{2}\right)^{2^k-1} \leq \frac{1}{1-h_0/2}$.

Starting at $x_0 \in \mathcal{B}$*, the Newton sequence defined by (3.1) is well-defined and remains within* $\mathcal{B}$*. The Newton sequence converges to an* $x^* \in \mathcal{B}$ *for which* $F(x^*) = 0$*.*
Furthermore, if $h_0 < \tilde{h}$*, where* $\tilde{h} \approx 0.71$ *is the solution to* $H = 1/h_0$*, then the solution is unique within* $\mathcal{B}$*.*

Strictly speaking we could take the first assumption to be

$$\|F'(x)^{-1}(F'(y + s(y - x)) - F'(y))\| \leq s\omega\|y - x\|, \qquad (3.35)$$

and derive the same proof, which only differs from the classical version by replacing all occurrences of $\beta\gamma$ by ω.

Theorem 6 (Affine covariant Rall-Rheinboldt). *Let $F : D \to Y$ be a continuously Fréchet differentiable function on the open convex subset $D \subseteq X$. Suppose that there exists an $x^* \in D$ such that $F(x^*) = 0$. Assume that*

i) $F'(x^*)^{-1}$ exists,

ii) $\|F'(x^*)^{-1} (F'(x) - F'(y)) \| \leq \omega^* \|x - y\|$ for all $x, y \in D$ (affine covariant Lipschitz continuity of $F'(x)$).

Then any $\rho^* \leq 2/(3\omega^*)$ such that $\mathcal{B} = B(x^*, \rho^*) \subset D$ has the property that starting at $x_0 \in \mathcal{B}$, the Newton sequence defined by (3.1) is well-defined and remains within $\mathcal{B}$. The Newton sequence converges to $x^* \in \mathcal{B}$. Furthermore, if we define $\rho^+ = \frac{1}{\omega^*}$, then x^* is unique within $D \cap B(x^*, \rho^+)$.

The proof again follows from its classical counterpart by replacing $\beta\gamma$ by ω^*. A combination of elements from Rall-Rheinboldt and the Newton-Mysovskikh theorem yields the theorem called refined Newton-Mysovskikh which centres around roots x^* but uses stronger assumptions on the invertibility of the Fréchet derivative.

Theorem 7 (Refined Newton-Mysovskikh [13]). *Let $F : D \to Y$ be a continuously Fréchet differentiable function on the convex subset $D \subseteq X$ with invertible Fréchet derivative $F'(x)$ for all $x \in D$. Suppose that there exists an $x^* \in D$ such that $F(x^*) = 0$. Assume that starting at $x^0 \in D$*

i) $\|F'(x)^{-1} (F'(x) - F'(y)) \| \leq \bar{\omega} \|x - y\|$ for all $x, y \in D$ (affine covariant Lipschitz continuity of $F'(x)$),

ii) $\mathcal{B} = \bar{B}(x^*, \bar{\rho}) \subset D$ for $\bar{\rho} = \|x_0 - x^*\|$,

iii) $\bar{\omega}\bar{\rho} < 2$.

Starting at $x_0 \in \mathcal{B}$, the Newton sequence defined by (3.1) is well-defined and remains within $\mathcal{B}$. The Newton sequence converges to $x^* \in \mathcal{B}$. Furthermore, if we define $\rho^+ = \frac{2}{\bar{\omega}}$, then x^* is unique within $D \cap B(x^*, \rho^+)$.

A proof based on [13, Theorem 1.1] uses the same techniques as in the proof of the Rall-Rheinboldt theorem.

Proof [13]. The start of the proof establishes again an estimate for the difference of the Newton solution with x^* by use of the Lipschitz condition and the mean-value theorem

$$\|x_{N+1} - x^*\| \leq \frac{\bar{\omega}}{2} \|x_N - x^*\|^2 < \|x_N - x^*\|^2 / \bar{\rho} \tag{3.36}$$

As before this proves that the Newton sequence remains in $\mathcal{B}$ and that the sequence converges to x^* at a rate at least equal to quadratic.

To prove uniqueness assume that there exists a $y^* \in D \cap B(x^*, \rho^+)$ with $x^* \neq y^*$. Then if we choose $x_0 = y^*$ we know that $x_k = y^*$ for all k. Therefore we see that

$$\|y^* - x^*\| = \|y_{N+1} - x^*\| \leq \frac{\bar{\omega}}{2} \|y_N - x^*\|^2 = \frac{\bar{\omega}}{2} \|y^* - x^*\|^2 < \|y^* - x^*\|, \tag{3.37}$$

which yields a contradiction and thus we must have $x^* = y^*$. □

All the above theorems are stated in a general setting, with the only severe restriction being a local (affine covariant) Lipschitz continuity of the Fréchet derivative of the function. Of course there are situations in which the function is in fact smoother than demanded by the preceding theorems and this can lead to new convergence theorems, which we will consider in the next section.

3.3 Convergence theorems for analytic functions

We step from continuously differentiable functions to the class of functions that are infinitely differentiable, a concept known in finite dimensions as analyticity. In doing so we skip the classes of three, four, etc. times differentiable functions for which some literature does exist, see for example [23]. The style and approach to those functions in literature is, however, completely in line with the previous section and authors merely devise more complex majorant sequences.

Analytic functions are considered to be functions which can (at least locally) be expanded in a convergent Taylor series, which is likely to be familiar for functions on $\mathbb{R}$ or $\mathbb{C}$. To extend the ideas of analytic functions to general Banach spaces we first need to redefine a polynomial, which is a crucial concept in the definition of a power series.

Definition 3. *A continuous m-homogeneous polynomial P is a function from X to Y such that there exists a multi-linear map A of degree m with the property that for every $x \in X$*

$$P(x) = A(x, \ldots, x) = A(x^m) := Ax^m. \tag{3.38}$$

A remark on notation, $x^m = (x, \ldots, x) \in X^m$ is a m-tuple on which A acts. Norms on multi-linear maps $G : X^n \to Y$ can be defined as induced norms by

$$\|G\| = \sup_{x_1 \neq 0, \ldots x_n \neq 0} \frac{\|G(x_1, \ldots, x_n)\|}{\|x_1\| \ldots \|x_n\|}. \tag{3.39}$$

Note that $\|P\| \leq \|A\|$ with equality if A is a symmetric multi-linear map. This now allows us to define the concept of a power series.

Definition 4 ([33])**.** *A power series from X to Y about $\tilde{x} \in X$ is a series in $x \in X$ of the form*

$$\sum_{k=0}^{\infty} P_k(x - \tilde{x}), \tag{3.40}$$

where P_k is a continuous k-homogeneous polynomial from X to Y.

A power series is said to be convergent around $\tilde{x}$ if there exists some $r > 0$ such that for all $x \in B(\tilde{x}, r)$ the power series converges uniformly in norm. Analytic functions are then defined by use of power series.

Definition 5. *A function $F : D \to Y$, with $D \subseteq X$ an open subset is called analytic if for all $x \in D$ there exists a $r > 0$ such that the function can be expanded in a convergent power series, $F(z) = \displaystyle\sum_{k=0}^{\infty} \frac{1}{k!} F^{(k)}(x)(z-x)^k$, for all $z \in B(x, r) \subset D$.*

Here $F^{(n)}(x)$ is the n-th Fréchet derivative at the point x and this is a symmetric multi-linear map of degree n from $X^n = X \times \cdots \times X$ to Y. The notation in Definition 5 is thus equivalent to $(F^{(n)}(x))(z-x, \ldots, z-x)$. The series expansion of F around x is called the Taylor series. One can now see that the power series in Definition 5 is in fact a direct generalisation of the power series for functions of one variable. In fact, the Taylor series is the unique power series expansion of F at x just as is the case for functions of one variable.

Before we proceed to the convergence theorems we need another auxiliary functional analysis result, the Cauchy-Hadamard theorem [33, Proposition 4.1], which extends the idea of a radius of convergence for power series to a Banach space setting.

Theorem 8 (Cauchy-Hadamard [33])**.** *Given a power series $\sum_{k=0}^{\infty} P_k(x - \tilde{x})$ from X to Y about $\tilde{x} \in X$, The largest R such that the power series is uniformly convergent on every $B(\tilde{x}, \rho)$ for $0 \leq \rho < R$ is given by*

$$\frac{1}{R} = \limsup_{k \to \infty} \|P_k\|^{1/k}, \tag{3.41}$$

and is called the radius of convergence of the power series.

Note that although F might be an entire function, i.e. analytic on X, this does not imply that its Taylor series around some $\tilde{x} \in X$ converges on the whole of X as this is only true for the finite dimensional case. See, for instance, [33, Remark 7.1] for a counter example, signifying that there are in fact differences between the finite dimensional case which we normally encounter and the formalism in (infinite dimensional) Banach spaces.

A related concept in complex analysis to analytic functions is that of a holomorphic function.

Definition 6. *Let X and Y be complex Banach spaces. A function $F : D \to Y$, with $D \subseteq X$ an open subset, is called* holomorphic *if F is Fréchet differentiable for all $x \in D$.*

We thus see that all functions that we considered in the preceding section are in fact holomorphic if X and Y are complex Banach spaces. Just as in finite dimensional complex analysis the notion of holomorphic and analytic functions are equivalent if we have Banach spaces over $\mathbb{C}$ as pointed out by the following theorem.

Theorem 9 (Goursat [33])**.** *Let X and Y be complex Banach spaces and $F : D \to Y$, with $D \subseteq X$ an open subset, then the following statements are equivalent:*

i) F is holomorphic,

ii) F is analytic.

Consequently, the results in this section will only consider a truly different class of functions F if we use real Banach spaces. In the case of complex Banach spaces the assumptions in the previous section are actually more restrictive than we will see in this section, because we previously considered Lipschitz continuous Fréchet derivatives, whereas for the analytic functions in this section we just have continuous Fréchet derivatives.

Now consider an $F : X \to Y$ analytic and the application of Newton's method to find its roots. A different class of convergence theorems can be derived in this case, based on initial work by Smale [45].

Smale convergence theorems

First we define an auxiliary quantity introduced by Smale

$$\gamma(x) = \sup_{\substack{k\in\mathbb{N}\\ k\geq 2}} \left\| \frac{F'(x)^{-1}F^{(k)}(x)}{k!} \right\|^{1/(k-1)} . \tag{3.42}$$

This is a well-defined expression for analytic functions at points where the Fréchet derivative is invertible. This follows from application of theorem 8 to the Taylor series of an analytic function F. Note furthermore that $\gamma(x)$ is an affine covariant quantity.

We start with a convergence theorem similar in spirit to the Rall-Rheinboldt theorem which, instead of using an affine covariant Lipschitz condition on F', makes use of evaluation of $\gamma(x)$ at the root of F.

Theorem 10 (Smale's γ-theorem [7])**.** *Let $F : D \to Y$ with $D \subseteq X$ an open subset be an analytic function. Suppose that there exists an $x^* \in D$ for which $F(x^*) = 0$ and $F'(x^*)^{-1}$ exists. Let $\gamma^* = \gamma(x^*)$, then any $\rho^* \leq \frac{5-\sqrt{17}}{4\gamma^*}$ such that $\mathcal{B} = B(x^*, \rho^*) \subseteq D$ has the property that starting at $x_0 \in \mathcal{B}$, the Newton sequence defined by (3.1) is well-defined, remains within $\mathcal{B}$. The Newton sequence converges to $x^* \in \mathcal{B}$, which is unique in $\mathcal{B}$.*

We note that this theorem does not seem to appear in literature as a result on its own and is often merely used to proof another Smale theorem which we will present hereafter. Nonetheless we state the theorem here as it is of interest for our purposes, because of the x^*-centred framework in which it is formulated. The theorem and proof are taken from [7, Theorem 8.1], but with some minor modifications to illustrate some subtleties.

Proof [7]. Observation 1 (Invertibility of the Jacobian)
Using the fact that F is analytic we expand its Fréchet derivative in a Taylor series.

We find for some $r > 0$ and $x \in B(x^*, r)$

$$F'(x^*)^{-1}F'(x) = F'(x^*)^{-1}\left(F'(x^*) + \sum_{m=1}^{\infty} \frac{F^{(m+1)}(x^*)}{m!}(x - x^*)^m\right) \tag{3.43}$$

$$= I + \sum_{m=1}^{\infty}\left(\frac{1}{m!}F'(x^*)^{-1}F^{(m+1)}(x^*)\right)(x - x^*)^m \tag{3.44}$$

$$= I + G, \tag{3.45}$$

which is now of the form $I + G$ with $G \in GL(Y)$.

Our objective is to prove that the Fréchet derivative of F exists on $\mathcal{B}$. In order to do so we have to use that the Taylor series is defined on the whole of $\mathcal{B}$, or simply said that the radius of convergence R of the Taylor series satisfies $R \geq \rho^*$. This is a subtle point and has apparently been overlooked by many authors. The radius of convergence can be found by applying theorem 8 which yields

$$\frac{1}{R} = \limsup_{\substack{n\to\infty \\ n\geq 1}} \|P_n\|^{1/n} \tag{3.46}$$

$$\leq \limsup_{\substack{n\to\infty \\ n\geq 1}} \left\|\frac{1}{n!}F'(x^*)^{-1}F^{(n+1)}(x^*)\right\|^{1/n} \tag{3.47}$$

$$\leq \limsup_{\substack{n\to\infty \\ n\geq 1}} (n+1)^{1/n} \cdot \limsup_{\substack{n\to\infty \\ n\geq 1}} \left\|\frac{1}{(n+1)!}F'(x^*)^{-1}F^{(n+1)}(x^*)\right\|^{1/n} \tag{3.48}$$

$$= 1 \cdot \limsup_{\substack{m\to\infty \\ m\geq 2}} \left\|\frac{1}{m!}F'(x^*)^{-1}F^{(m)}(x^*)\right\|^{1/(m-1)} \leq \gamma^*. \tag{3.49}$$

From this we can conclude that the radius of convergence satisfies $R \geq 1/\gamma^* > \rho^*$ and thus indeed we can use the Taylor series on $\mathcal{B}$.

Let $u = \|x - x^*\|\gamma^*$. Under the assumption that $u < 1 - \sqrt{2}/2$, which is guaranteed by $x \in \mathcal{B}$, we find that using the power series

$$\|G\| = \left\|\sum_{m=1}^{\infty} P_m(x - x^*)\right\| \tag{3.50}$$

$$\leq \sum_{m=1}^{\infty} \|P_m(x - x^*)\| \tag{3.51}$$

$$\leq \sum_{m=1}^{\infty} \|P_m\| \, \|x - x^*\|^m. \tag{3.52}$$

Next we use the explicit form of the terms in the series expansion of G to find

$$\|G\| \leq \sum_{m=1}^{\infty} \left\| \frac{1}{m!} F'(x^*)^{-1} F^{(m+1)}(x^*) \right\| \|x - x^*\|^m \tag{3.53}$$

$$\leq \sum_{n=2}^{\infty} n \left\| \frac{1}{n!} F'(x^*)^{-1} F^{(n)}(x^*) \right\| \|x - x^*\|^{n-1} \tag{3.54}$$

$$\leq \sum_{n=2}^{\infty} n \left(\gamma^* \|x - x^*\| \right)^{n-1} \tag{3.55}$$

$$= \sum_{n=2}^{\infty} n u^{n-1} \leq \frac{1}{(1-u)^2} - 1 < 1. \tag{3.56}$$

By using lemma 2 we then find that

$$\|F'(x)^{-1} F'(x^*)\| \leq \frac{1}{1 - \|G\|} \leq \frac{1}{1 - \left(\frac{1}{(1-u)^2} - 1 \right)} = \frac{(1-u)^2}{\psi(u)}, \tag{3.57}$$

where $\psi(u) = 1 - 4u + 2u^2$.

Observation 2 (Evolution of Newton iterates)
We will use an induction argument to show that the Newton sequence remains bounded within $\mathcal{B}$ and forms a converging sequence.

Suppose that we have a Newton sequence $\{x_k\}$ for $k = 1, \ldots, N$ such that $x_k \in \mathcal{B}$. If we use the Taylor series for $F(x_N)$ and $F'(x_N)$ around x^* we find that the next iterate satisfies

$$\|x_{N+1} - x^*\| = \left\| F'(x_N)^{-1} F'(x^*) \left[-F'(x^*)^{-1} F(x_N) + F'(x^*)^{-1} F'(x_N)(x_N - x^*) \right] \right\| \tag{3.58}$$

$$= \left\| F'(x_N)^{-1} F'(x^*) \left[\sum_{m=1}^{\infty} (m-1) \frac{F'(x^*)^{-1} F^{(m)}(x^*)}{m!} (x_N - x^*)^m \right] \right\| \tag{3.59}$$

Using the triangle inequality and the γ^*-condition we can derive

$$\|x_{N+1} - x^*\| \leq \frac{(1-u)^2}{\psi(u)} \sum_{m=1}^{\infty} \left\| (m-1) \frac{F'(x^*)^{-1} F^{(m)}(x^*)}{m!} (x_N - x^*)^m \right\| \tag{3.60}$$

$$\leq \frac{(1-u)^2}{\psi(u)} \sum_{m=1}^{\infty} (m-1) \left\| \frac{F'(x^*)^{-1} F^{(m)}(x^*)}{m!} \right\| \|x_N - x^*\|^m \tag{3.61}$$

$$\leq \frac{(1-u)^2}{\psi(u)} \sum_{m=1}^{\infty} (m-1) (\gamma^*)^{m-1} \|x_N - x^*\|^m \tag{3.62}$$

$$= \|x_N - x^*\| \frac{(1-u)^2}{\psi(u)} \sum_{m=1}^{\infty} (m-1) u^{m-1} \tag{3.63}$$

$$= \|x_N - x^*\| \frac{(1-u)^2}{\psi(u)} \frac{u}{(1-u)^2} = \|x_N - x^*\| \frac{u}{\psi(u)}. \tag{3.64}$$

As $x_N \in \mathcal{B}$ we find $u < (5-\sqrt{17})/4$ which implies $u/\psi(u) < 1$. Therefore we conclude that $\|x_{N+1} - x^*\| < \|x_N - x^*\|$. This proves that the sequence remains in $\mathcal{B}$ as the base case $k = 0$ is satisfied trivially.

Using induction we can now show that $\|x_N - x^*\| < (u/\psi(u))^N \|x_0 - x^*\|$ and thus, we observe that $x_k \to x^*$.

The uniqueness proof follows that of the refined Newton-Mysovskikh theorem. □

With this result Smale then derived his point estimate theorem which establishes local convergence of a point purely based on function evaluations at the starting point and is more similar to Newton-Kantorovich and Newton-Mysovskikh in spirit.

Theorem 11 (Smale's α-theorem [45])**.** *Let $F : D \to Y$ with $D \subseteq X$ an open subset be an analytic function. Assume that, starting at $x_0 \in X$,*

i) $F'(x_0)^{-1}$ exists and let $\beta_0 = \|F'(x_0)^{-1}F(x_0)\|$,

ii) $\alpha(x_0) = \beta_0\gamma(x_0) < \alpha_0$, where α_0 is a universal constant ($\alpha_0 \approx 0.1307$),

iii) $x_0 \in \mathcal{B} = B(x_0, \rho_0) \subseteq D$, where $\rho_0 = \frac{1-\sqrt{2}/2}{\gamma(x_0)}$.

Then the Newton sequence started at x_0 converges to a unique $x^* \in \bar{B}(x_0, 2\beta_0)$ for which $F(x^*) = 0$ holds.

The universal constant α_0 appears to be subject to debate as the optimal value has not been established as of yet and the best version we found in the literature is $\alpha_0 = 3 - 2\sqrt{2}$ [37].

3.4 Summary of local convergence theorems

In this chapter we have reviewed a collection of local convergence theorems for the classical Newton's method. Following an important observation by Deuflhard [12] we state all the final versions of the theorems in an affine covariant form. The theorems share many general characteristics and mainly differ on just two points. A classification based on these differences is given in Table 3.1.

The first one is the reference frame of the theorem. With that we mean whether the theorems are centred around an initial guess x_0 or around a root x^*. This will prove to be important in considering the deflation technique in the next chapter, where we will show that to incorporate deflation there is a clear preference for root-centred theorems.

The other main difference between theorems is their restriction on the smoothness of the function. Here we make a distinction between real and complex Banach spaces. In the case of real Banach spaces, the minimal requirement seems to be a Lipschitz continuous Fréchet derivative on some open set $D \subseteq X$, which is equivalent to a bounded second Fréchet derivative if a function is in fact twice differentiable. By requiring more smoothness in the form of analytic functions we can use a different approach and use Smale's point estimation theories. In the case of a complex Banach space the requirements for the theorems coincide, all theorems now require the function to be holomorphic. The Lipschitz continuity of the Fréchet derivative is now actually a more severe constraint than analyticity, which merely implies a continuous Fréchet derivative.

	Lipschitz continuous Fréchet derivative	Analytic function
x_0	Newton-Kantorovich theorem 4 *Newton-Mysovskikh theorem 5	Smale's α-theorem 11
x^*	Rall-Rheinboldt theorem 6 *Refined Newton-Mysovskikh theorem 7	Smale's γ-theorem 10

Table 3.1: Summary of local convergence theorems for Newton's method. All theorems are understood to be in affine covariant form. The theorems are ordered by their center of convergence balls and requirements on the objective functions smoothness. *-Theorems assume invertibility of the Fréchet derivative throughout the domain, whereas the others just assume invertibility at their respective centres of the convergence balls.

4. Local convergence of Newton's method and deflation

The local convergence theorems in the preceding chapter have been concerned with the problem of finding a single solution to $F(x) = 0$. For our purposes, however we want to be able to find multiple solutions. In this chapter we will therefore start deriving sufficient conditions that can guarantee the converge of Newton's method combined with deflation to multiple solutions starting from a single initial guess x_0.

4.1 Examples of local convergence to multiple solutions

The theorems in chapter 3 can be used to prove quadratic convergence of an initial guess x_0 to a root. We will first show by simple examples that it is possible to prove local quadratic convergence of a single initial point to multiple roots. These examples also serve as an illustration for the differences in the convergence theorems for Newton's method.

Example 1 (Quadratic polynomials)**.** Let $X = Y = \mathbb{C}$ and $f(x) = (x - 1)(x + 1)$. Note that due to the affine-covariant and affine-contravariant properties of Newton's method any quadratic polynomial with distinct roots can be brought into this form, so that the analysis of quadratic polynomials can be done purely on this specific example.

It is sufficient to prove convergence to one of the roots of the quadratic polynomial since after Wilkinson deflation the polynomial is linear which will result in Newton's method converging in one step. As the function is invariant under the transformation $x \mapsto -x$ we only consider the case $\Re(x_0) > 0$.

Under the assumption $x_0 \neq 0$, the affine covariant Lipschitz constant for Newton-Kantorovich is given by $\omega_0 = 1/|x_0|$ and the initial Newton step by $\alpha = |x_0^2 - 1|/2|x_0|$. To apply theorem 4 we need $|1 - 1/x_0^2| \leq 1$. As a result we find quadratic convergence towards $x = 1$ for $U_{\text{NK}} = \{x \in \mathbb{C} : |1 - 1/x^2| \leq 1\}$.

The affine covariant Lipschitz constant in the Rall-Rheinboldt (Theorem 6) is given by $\omega^* = 1$ and thus for all $x \in U_{\text{RR}} = \{x \in \mathbb{C} : |1 - x| < 2/3\}$ we find convergence towards $x = 1$.

Note that for the refined Newton-Mysovskikh theorem we can choose the domain D for which we assume that $f'(x)^{-1}$ exists, and we take it to be equal to $\bar{B}(1, |x_0 - 1|)$. Looking at the affine covariant Lipschitz constant we find $\bar{\omega} = 1/(1 - |x_0 - 1|)$ and we find the exact same convergence region as for the Rall-Rheinboldt theorem.

In this simple case we can calculate Smale's γ explicitly, which gives $\gamma(\pm 1) = 1/2$ and therefore $\rho^* = (5 - \sqrt{17})/2 \approx 0.438$. Convergence to $x = 1$ is thus guaranteed for $U_{\text{S}\gamma} = \{x \in \mathbb{C} : |1 - x| < (5 - \sqrt{17})/2\} \subset U_{\text{RR}}$.

Evaluation of Smale's $\alpha(x_0)$ can be done explicitly as well and yields $\alpha(x_0) = |1-1/x_0^2|/4$. Convergence to $x=1$ is thus guaranteed for $U_{\mathrm{S}\alpha} = \{x \in \mathbb{C} : |1-1/x^2| \leq 4\alpha_0\} \subset U_{\mathrm{NK}}$.

The resulting convergence regions are sketched in Figure 4.1. A clear distinction in the shape of the local convergence regions for the theorems which center convergence balls around x_0 versus those centred around x^* can be seen.

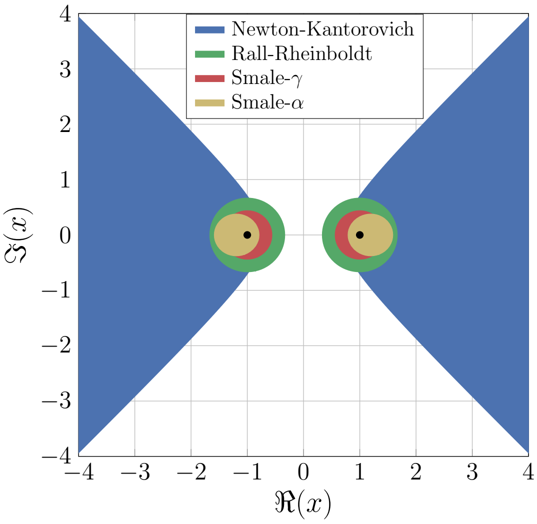


Figure 4.1: Local convergence regions given by theorems in chapter 3 for the quadratic polynomial $f(x) = (x-1)(x+1)$. The roots are indicated by •.

Example 2 (Cubic polynomials)**.** Just as for the quadratic polynomials we use the invariant properties of Newton's method to construct a normal form for the cubic polynomials with distinct roots, which can be chosen to be $f(x) = (x+1)(x-a)(x-1)$ with $a \in \mathcal{A} = \{x \in \mathbb{C} : |x \pm 1| \leq 2, x \neq \pm 1\}$. This can be achieved by using an affine transformation which maps the roots with the greatest separation to ±1 respectively, the remaining root is closer to the roots and thus has to lie in $\mathcal{A}$. The form is chosen so that deflation of this standard cubic we can arrive at the standard quadratic from the previous example. We construct a subset $\mathcal{E} \subset \mathcal{A}$ such that if $a \in \mathcal{E}$ there exists a point which converges quadratically to all the roots of the cubic polynomial.

Note that $f'(x) = 3x^2 - 2ax - 1$ so that $f'(a) = a^2 - 1 \neq 0$. Given a convex set $D \subset \mathbb{C}$ we thus have

$$|f'(a)^{-1}(f'(x) - f'(y))| = \frac{1}{|1-a^2|}|f'(x) - f'(y)| \leq \frac{\max_{z \in D}|f''(z)|}{|1-a^2|}|x-y|, \quad (4.1)$$

using the mean value theorem 3. Now suppose the convex set D is an open ball $B(a,q) \subset \mathbb{C}$. This allows us to find an affine covariant Lipschitz constant on D (possibly not the best one) as used in the theorem 6,

$$\omega^* = \frac{\max_{z \in D}|6z - 2a|}{|1-a^2|} = \frac{4|a| + 6q}{|1-a^2|},$$

where we have used that the affine transformation $z \mapsto 6z - 2a$ takes $B(a, q)$ to $B(4a, 6q)$. The element of this new ball furthest away from the origin then has modulus $4|a| + 6q$.

In order to use theorem 6 we need to have $B(a, \rho^*) \subset D$ for $\rho^* \leq 2/(3\omega^*)$. Take $q = \rho^*$ and then solving this inequality yields $q \leq \left(\sqrt{|a|^2 + |1 - a^2|} - |a|\right)/3$.

W.l.o.g. assume that $\Re(a) \geq 0$ and let $q = \left(\sqrt{|a|^2 + |1 - a^2|} - |a|\right)/3$, so that $2/(3\omega^*) = q$. To gain quadratic convergence to all three roots it now suffices to show that $B(a, 2/(3\omega^*)) \cap B(1, 2/3) \neq \emptyset$. The condition of intersecting balls from a geometrical point of view yields the condition $q + 2/3 > |1 - a|$. This now defines the set $\mathcal{E}_+ = \{x \in \mathcal{A} : \left(\sqrt{|x|^2 + |1 - x^2|} - |x|\right) + 2 - 3|1 - x| > 0\}$. Consequently, if $a \in \mathcal{E}$ the point $2a/3 + 1/3$ is guaranteed to converge quadratically to firstly $x = a$, then $x = 1$ and finally to $x = -1$. Completely analogous we can define $\mathcal{E}_-$ around $x = -1$ and their union creates the set $\mathcal{E}$, depicted in Figure 4.2.

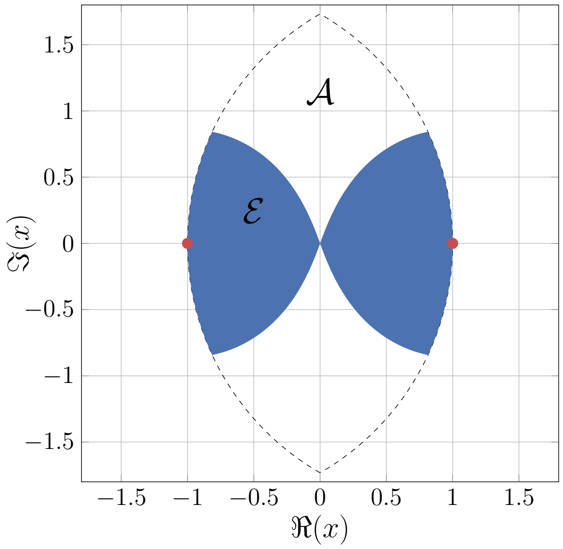


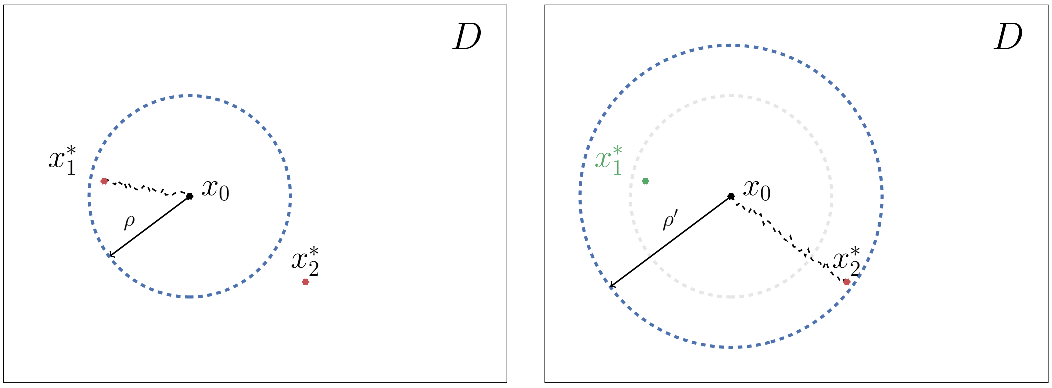
Figure 4.2: Area within dotted lines depicts $\mathcal{A}$ and the shaded area shows $\mathcal{E} \subset \mathcal{A}$. If the root a of $f(x) = (x + 1)(x - a)(x - 1)$ lies within $\mathcal{E}$, there exists a point which converges quadratically to all three roots of the cubic polynomial. The roots at $x = \pm 1$ are indicated by ●.

As we saw from the previous examples there are indeed functions for which the situation sketch made in Figure 2.1 holds. The convergence balls using for example the Rall-Rheinboldt theorem grow after deflation which makes it possible for points to be contained in a convergence ball of a different root at every step of the deflation. This raises the question whether we can formalise sufficiency conditions for this to happen in the case of deflation on general functions.

4.2 Convergence criteria for deflated functions

Having defined a general framework for deflation in Banach spaces we will now look at the applicability of theorems in Chapter 3 in the case of deflation in its most general form.

First we note that we can rule out the use of any theorem which is centred around the initial guess x_0 if we want to prove statements of convergence to multiple roots. This can be understood as follows. Suppose there exist x_1 and x_2 such that $F(x_1) = F(x_2) = 0$ and $x_1 \neq x_2$. Now assume we consider an initial guess x_0 that provably converges to x_1 using a convergence theorem in the x_0 framework. The result is a ρ such that $x_1 \in B(x_0, \rho)$ and $x_2 \notin B(x_0, \rho)$, see Figure 4.3a. Now after deflating F we have to consider $\mathcal{M}(x; x_1)F(x)$. This deflated function in most cases will behave badly at the deflated root. In the case of shifted norm-deflation (2.9) for example there will be a pole or discontinuity in the Fréchet derivative at $x = x_1$. In order to prove convergence to x_2 for the deflated function we would need to be able to find a $\rho' > \rho$ such that $x_2 \in B(x_0, \rho')$, but this would imply that $x_1 \in B(x_0, \rho')$, see Figure 4.3b. Since the convergence theorems require the function to be well-behaved on these convergence balls and prove that the Fréchet derivative is invertible, the ill behaviour at x_1 poses a problem. As a result we will focus in the next section on the theorems of the x^*-framework.



(a) Convergence before deflation

(b) Convergence after deflation

Figure 4.3: Illustration of local convergence of the failure of convergence towards multiple roots in the x_0-framework of the convergence theorems. In order to show convergence to multiple solutions we need the convergence region to grow, i.e. $\rho' > \rho$. This would however imply that the deflated root lies within the new convergence region, which poses regularity problems on the Fréchet derivative of the deflated function in the convergence region.

4.2.1 Deflation on analytic functions

In the case of an analytic function and under regularity conditions on the deflation operator we can show that the deflated operator will remain well-behaved close to the unknown roots.

Theorem 12. *Let $F : X \to Y$ be an analytic function on an open subset $D \subseteq X$. Suppose that $z \in D$ such that $F(z) = 0$ and the Fréchet derivative $F'(z)$ is invertible. Furthermore assume that we have an $x^* \in X$, $x^* \neq z$, such that $F(x^*) = 0$. Given a deflation operator $\mathcal{M}(\cdot; x^*) : D \setminus \{x^*\} \to GL(Y, Z)$ which is analytic on an open subset $E \subseteq D \setminus \{x^*\}$ such that $z \in E$, then the following holds;*

i) the Fréchet derivative of the deflated operator $\mathcal{M}(x; x^*)F(x)$ is invertible at z,

ii) the deflated operator is analytic on the open subset E containing z.

Note that a shifted deflation operator, as defined by 2.9, satisfies these requirements if the norm on X is (locally) analytic and we are close enough to the root x^* in the sense that $0 < \|x - x^*\|^p < 2$, because we want $1/\|x - x^*\|^p$ to be analytic. In that case we can use the composition rule for analytic functions to show that the deflation operator, which is now a composition of analytic operators, is indeed analytic on some open set $E \subseteq D \setminus \{x^*\}$. The question whether a norm is analytic, or whether a Banach space has an equivalent real analytic norm appears to be subject of continuing study in functional analysis and we will not dive into detail here, but just mention that analytic norms do in fact exist [24] and that for separable Hilbert spaces and $L^p[0, 1]$ spaces with p an even integer we can approximate any equivalent norm by real analytic norms [14].

In order to prove Theorem 12 we need a lemma generalising the Cauchy product formula for the product of two series in $\mathbb{C}$ to Banach space, which will allow us to take products and sums of power series.

Lemma 6 (Cauchy product formula [25])**.** *Let X, Y, Z be Banach spaces and $\alpha : X \times Y \to Z$ a continuous bilinear map. Suppose that the series $\sum_{n=0}^{\infty} x_n$ and $\sum_{n=0}^{\infty} y_n$ are absolutely convergent in X and Y and denote their sum by $x \in X$ and $y \in Y$ respectively. Then*

$$\alpha(x, y) = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \alpha(x_k, y_{n-k}). \tag{4.2}$$

Proof of Theorem 12. By the product rule for differentiation we have

$$[\mathcal{M}(x; x^*)F(x)]' = \mathcal{M}(x; x^*)F'(x) + \mathcal{M}'(x; x^*)F(x). \tag{4.3}$$

Since $z \in X$ is a root of the original functional F, the Fréchet derivative of the deflated functional at z becomes $\mathcal{M}(z; x^*)F'(z)$. For any $x \in D$ the deflation operator $\mathcal{M}(x; x^*) \in GL(Y, Z)$ and thus it is invertible. The deflated functional evaluated z is thus invertible and

$$\left([\mathcal{M}(x; x^*)F(x)]'\right)^{-1}\Big|_z = F'(z)^{-1}\mathcal{M}(z; x^*)^{-1}. \tag{4.4}$$

As both F and $\mathcal{M}(\cdot; x^*)$ are analytic on an open set around z we can expand them in convergent Taylor series around z,

$$\mathcal{M}(x;x^*) = \sum_{k=0}^{\infty} \frac{1}{k!}\mathcal{M}^{(k)}(z;x^*)(x-z)^k = \sum_{k=0}^{\infty} Q_k(x-z), \tag{4.5}$$

$$F(x) = \sum_{k=0}^{\infty} \frac{1}{k!}F^{(k)}(z)(x-z)^k = \sum_{k=0}^{\infty} P_k(x-z), \tag{4.6}$$

where $Q_k(x), P_k(x)$ are continuous k-homogeneous polynomials. Now we can apply the Cauchy product formula to the bilinear form $\alpha(x,y) = xy$ to find a power series expansion of the deflated operator

$$\mathcal{M}(x;x^*)F(x) = \sum_{n=0}^{\infty}\sum_{k=0}^{n} Q_k(x-z)P_{n-k}(x-z) = \sum_{n=0}^{\infty} R_n(x-z), \tag{4.7}$$

where $R_n(x)$ is a n-homogeneous polynomial.

The above argument can be repeated for any $u \in E$ as both F and $\mathcal{M}(\cdot;x^*)$ are analytic and can be expanded in power series. There exists therefore an open neighbourhood U around all $x \in E$ such that the deflated operator can be expanded into a convergent power series. As the power series expansion of an operator is uniquely defined it follows from theorem 9 that the deflated operator is analytic on E. □

A direct consequence of the above theorem is that Smale's γ remains well-defined after deflation.

Corollary 2. *Under the assumptions in theorem 12, Smale's gamma-function, given by (3.42), is well-defined at $z \in X$ for both the original function $F(x)$ and the deflated function $\mathcal{M}(x;x^*)F(x)$. In this case denote $\tilde{\gamma}(x)$ as Smale's gamma-function applied to the deflated function $\mathcal{M}(x;x^*)F(x)$.*

Although it is not clear from the above results what the relation is between $\tilde{\gamma}(z)$ after deflation and $\gamma(z)$ before deflation, we can conclude that after deflation with suitable smooth operators that in principle at every stage of deflation there exist open neighbourhoods around the undiscovered roots for which quadratic convergence with Newton is guaranteed.

Note, though, that deflating a function can in fact result in $\tilde{\gamma}(z) < \gamma(z)$ which results in larger convergence balls after deflation compared with the situation before deflation, as the following example shows.

Example 3 (Shrinking Smale's gamma after deflation)**.** We consider again the standard cubic polynomial $f(x) = (x+1)(x-a)(x-1)$ with $a \in \mathcal{A} = \{x \in \mathbb{C} : |x \pm 1| \leq 2, x \neq \pm 1\}$ where $x \in \mathbb{C}$. Explicitly calculating the Smale gamma-functions yields

$$\gamma(1) = \max\left\{\frac{|3-a|}{2|1-a|}, \frac{1}{\sqrt{2|1-a|}}\right\} \geq \frac{1}{2},$$

$$\gamma(-1) = \max\left\{\frac{|3+a|}{2|1+a|}, \frac{1}{\sqrt{2|1+a|}}\right\} \geq \frac{1}{2},$$

where the inequality is derived from $a \in \mathcal{A}$, which implies $|1-a| \leq 2$ and $|1+a| \leq 2$, and the second term over which we take the maximum. Equality for both inequalities at the same time could only happen then if $2 = |1-a| = |1+a|$, i.e. $a = \pm i\sqrt{3}$, but in that case $\gamma(\pm 1) = \sqrt{3}/2 > 1/2$. So at least one of the inequalities is strict.

After deflation of the root a we can use Example 1 to find that $\tilde{\gamma}(\pm 1) = 1/2$. Therefore we see that at least one of the gamma functions shrinks after deflation.

Combining Theorem 12 with the idea of overlapping convergence balls as in Figure 2.1 we get the first sufficient conditions for convergence to multiple roots from a single initial guess using deflation and Newton's method.

Theorem 13 (Deflated Smale's γ-theorem)**.** *Let $F : X \to Y$ be an analytic function on an open subset $D \subseteq X$. Suppose that $z_1, z_2 \in D$ such that $F(z_1) = F(z_2) = 0$ and the Fréchet derivatives $F'(z_1)$ and $F'(z_2)$ are invertible. Given a deflation operator $\mathcal{M}(\cdot; z_1) : D \setminus \{z_1\} \to GL(Y, Z)$ which is analytic on an open subset $E \subset D \setminus \{z_1\}$ such that $z_2 \in E$, then define*

i) $\tilde{\gamma}(x)$ as Smale's gamma-function defined by (3.42) applied to $\mathcal{M}(x; z_1)F(x)$,

ii) $\gamma_1^* = \gamma(z_1)$ and $\rho_1^* = \frac{5-\sqrt{17}}{4\gamma_1^*}$,

iii) $\gamma_2^* = \tilde{\gamma}(z_2)$ and $\rho_2^* = \frac{5-\sqrt{17}}{4\gamma_2^*}$.

If $\|z_1 - z_2\| < \rho_1 + \rho_2$ for some $\rho_1 \leq \rho_1^*$ and $\rho_2 \leq \rho_2^*$ such that $\mathcal{B}_1 = B(z_1, \rho_1) \subset D$ and $\mathcal{B}_2 = B(z_2, \rho_2) \subset E$, then there exists an $x_0 \in \mathcal{B}_1 \cap \mathcal{B}_2$. Starting from this x_0 with Newton's method we first converge to $z_1 \in D$ and then after deflation using $\mathcal{M}(\cdot; z_1)$ we converge to $z_2 \in E$.

Note that by Theorem 12 all definitions in Theorem 13 are well-defined. A proof applies Theorem 12, Theorem 10 and an overlapping condition on convergence balls before and after deflation as depicted in Figure 2.1.

As the above results are restricted to the case of analytic functions one could ask a similar type of question for general functions, as treated in Section 3.2.

4.2.2 Deflation on general functions

We now consider functions which are not necessarily analytic, but for which we can prove local convergence of Newton's method. As a crucial ingredient of the affine covariant theorems in Section 3.2 is the affine covariant Lipschitz continuity of the Fréchet derivative F', we first state a lemma on the product of Lipschitz continuous functions.

Lemma 7 (Product of Lipschitz continuous functions)**.** *Let X, Y, Z be Banach spaces and $G : X \to L(Y, Z)$ and $F : X \to Y$ be Lipschitz continuous functions on the open subset $D \subseteq X$ with Lipschitz constants ω_F and ω_G respectively. Assume furthermore that F is bounded on D, i.e. there exist $N_F, N_G \in \mathbb{R}$ such that for all $x \in D$ we have $\|F(x)\| < N_F$ and $\|G(x)\| < N_G$, as G is bounded by definition. Then the product $GF : X \to Z$ is bounded and Lipschitz continuous on D with Lipschitz constant $(N_F\omega_G + N_G\omega_F)$.*

For a proof see Appendix A.2. Note that the given Lipschitz constant might not be the best Lipschitz constant.

This now sets conditions for our deflation operator such that we can apply the Rall-Rheinboldt convergence theorem to deflated functions.

Theorem 14. *Let $F : D \to Y$ be a continuously Fréchet differentiable function on the open convex subset $D \subseteq X$. Suppose that $z \in D$ such that $F(z) = 0$ and the Fréchet derivative $F'(z)$ is invertible. Furthermore assume that we have an $x^* \in D$, $x^* \neq z$, such that $F(x^*) = 0$. Suppose then that we have a deflation operator $\mathcal{M}(\cdot; x^*) : D \setminus \{x^*\} \to GL(Y, Z)$ and an open bounded convex subset $E \subseteq D \setminus \{x^*\}$ with $z \in E$ such that the following conditions hold:*

i) F satisfies the two Rall-Rheinboldt conditions around $z \in D$ (see theorem 6),

ii) $\mathcal{M}(x; x^*)$ is continuously Fréchet differentiable for all $x \in E$,

iii) $\|\mathcal{M}'(x; x^*) - \mathcal{M}'(y; x^*)\| \leq \omega_{\mathcal{M}'} \|x - y\|$ for all $x, y \in E$.

Then there exists a $\rho > 0$ such that if we start at $x_0 \in \mathcal{B} = B(z, \rho)$ the Newton sequence defined by (3.1) on the deflated function $\mathcal{M}(x; x^*)F(x)$ is well-defined, remains in $\mathcal{B}$ and converges to $z \in \mathcal{B}$.

Note that a shifted deflation operator, as defined by 2.9, satisfies these requirements under the condition that the norm on X is (locally) at least twice continuously differentiable on E (this implies a Lipschitz continuous Fréchet derivative as E is bounded). In that case we can use the composition rule for differentiable functions to show that the deflation operator, which is now a composition of differentiable operators, is indeed twice continuously differentiable E. The L^p spaces for $1 < p < \infty$ and Sobolev spaces ($W^{k,p}$ for $k \in \mathbb{N}$ and $1 < p < \infty$) are reflexive spaces and therefore admit a continuously differentiable norm on any open subset not containing zero, see for example [20]. The question whether a continuously twice differentiable norm exists seems to be a more involved question in functional analysis. We point out that for $2 \leq p < \infty$ the L^p standard norm is at least twice continuously differentiable [46, Theorem 8] and that Banach spaces that are isomorphic to a Hilbert space can be equipped with twice differentiable norms as well [29, 17].

This statement in essence states conditions for which the Rall-Rheinboldt theorem is not only applicable to the original function, but to the deflated function as well.

Proof. To start with we know by the same reasoning as in theorem 12 that the Fréchet derivative of $\mathcal{M}(x; x^*)F(x)$ is invertible at z.

As $F(x)$ and $\mathcal{M}(x; x^*)$ are continuously Fréchet differentiable on E we know that they are Lipschitz continuous as well. Being Lipschitz continuous also implies that the operators are bounded on E and thus $F(x), F'(x), \mathcal{M}(x; x^*)$ and $\mathcal{M}'(x; x^*)$ are all bounded on E.

As a result we know that the Fréchet derivative of the deflated operator

$$(\mathcal{M}(x; x^*)F(x))' = \mathcal{M}(x; x^*)F'(x) + \mathcal{M}'(x; x^*)F(x) \tag{4.8}$$

is Lipschitz continuous by use of the triangle inequality and Lemma 7. Therefore there exists some (affine covariant) $\tilde{\omega} > 0$ such that

$$\left\| \left([\mathcal{M}(z; x^*)F(z)]' \right)^{-1} [(\mathcal{M}(x; x^*)F(x))' - (\mathcal{M}(y; x^*)F(y))'] \right\| \leq \tilde{\omega} \|x - y\|, \quad (4.9)$$

for all $x, y \in E$. This means that the affine covariant Rall-Rheinboldt theorem can be applied to both $F(x)$ and $\mathcal{M}(x; x^*)F(x)$, proving the claim in the theorem. □

Now we can state sufficient conditions for a general function to convergence to two solutions by using deflation and Newton's method. The proof simply applies the previous theorem and the Rall-Rheinboldt theorem 6.

Theorem 15 (Deflated Rall-Rheinboldt)**.** *Let $F : D \to Y$ be a continuously Fréchet differentiable function on an open subset $D \subseteq X$. Suppose that $z_1, z_2 \in D$ such that $F(z_1) = F(z_2) = 0$. Let E_1 be an open convex subset E_1 such that $E_1 \subset D \setminus \{z_2\}$ and $z_1 \in E_1$. Furthermore let E_2 be an open bounded convex subset E_2 such that $E_2 \subset D \setminus \{z_1\}$ and $z_2 \in E_2$. Additionally let $\mathcal{M}(\cdot; z_1) : D \setminus \{z_1\} \to GL(Y, Z)$ be a deflation operator such that the following conditions hold*

i) $F'(z_1)^{-1}$ *and* $F'(z_2)^{-1}$ *exist,*

ii) $\|F'(z_1)^{-1} (F'(x) - F'(y)) \| \leq \omega_1^* \|x - y\|$ *for all* $x, y \in E_1$,

iii) $\|F'(z_2)^{-1} (F'(x) - F'(y)) \| \leq \omega_2^* \|x - y\|$ *for all* $x, y \in E_2$,

iv) $\mathcal{M}(x; z_1)$ *is continuously Fréchet differentiable for all* $x \in E_2$,

v) $\|\mathcal{M}'(x; z_1) - \mathcal{M}'(y; z_1)\| \leq \omega_{\mathcal{M}'} \|x - y\|$ *for all* $x, y \in E_2$.

Then there exists a $\tilde{\omega}_2 > 0$ such that for all $x, y \in E_2$ there holds

$$\left\| \left([\mathcal{M}(z_2; z_1)F(z_2)]' \right)^{-1} [(\mathcal{M}(x; z_1)F(x))' - (\mathcal{M}(y; z_1)F(y))'] \right\| \leq \tilde{\omega}_2 \|x - y\|. \quad (4.10)$$

If $\|z_1 - z_2\| < \rho_1 + \rho_2$ for some $\rho_1 \leq 2/(3\omega_1^*)$ and $\rho_2 \leq 2/(3\tilde{\omega}_2)$ such that we have $\mathcal{B}_1 = B(z_1, \rho_1) \subset E_1$ and $\mathcal{B}_2 = B(z_2, \rho_2) \subset E_2$, then there exists an $x_0 \in \mathcal{B}_1 \cap \mathcal{B}_2$. Starting from this x_0 with Newton's method we first converge to $z_1 \in E_1$ and then after deflation using $\mathcal{M}(\cdot; z_1)$ we converge to $z_2 \in E_2$.

Note that a crude upper bound for $\tilde{\omega}_2$ could be derived using Lemma 7, the triangle inequality and bounds on the norms of $F(x), \mathcal{M}(x; z_1)$ and their Fréchet derivatives. The Lipschitz constant derived in this way is, however, likely to be a gross overestimation of the optimal Lipschitz constant and therefore might not be useful in practice.

We conclude with the observation that in order to derive sufficient conditions for convergence we imposed the same regularity conditions on the deflation operator as those imposed on the original function. This seems natural as we want their product, the deflated function, to satisfy similar regularity conditions in order to apply the local convergence theorems from Chapter 3.

5. Bifurcation diagrams and deflation

Now we turn our attention to an application of the method of deflation in finding multiple solutions, namely the construction of bifurcation diagrams. We will first give a short introduction to the numerical computation of bifurcation diagrams using so-called continuation techniques. Then we will highlight where deflation can complement the traditional numerical bifurcation toolbox. In practice we observe that for a range of illustrative problems with deflation and continuation combined we can trace out bifurcation diagrams where continuation alone fails.

Part of the code used for this thesis is publicly available at https://bitbucket.org/CasperBeentjes/files-msc-dissertation.

5.1 Numerical bifurcation techniques

Branch continuation

We return to the original problem

$$F(u, \lambda) = 0, \tag{5.1}$$

and suppose for the moment that we are initially given a root $(\tilde{u}, \tilde{\lambda})$. A natural question now is whether we can solve 5.1 for $u(\lambda)$, i.e. whether we can find a relation between the output of the model and the controls. Sufficient conditions for such a curve to exist are given by the implicit function theorem [26, Theorem 13.22]. They mainly require an invertible partial Fréchet derivative with respect to u, denoted F_u, at the given roots and result in an open neighbourhood around the given root for which $u(\lambda)$ such that $u(\tilde{\lambda}) = \tilde{u}$ is uniquely defined and a root to (5.1). As these conditions are satisfied away from bifurcation points we can hope to trace out these curves by making small steps along the curve, an idea which was already known to Poincaré [38], and is known under the name continuation.

In this classical setting of continuation the solution curve is parametrised by the natural occurring parameter for the problem, namely λ. This parametrisation works well in regions where F_u is non-singular so that we are guaranteed to have a unique parametrisation by the implicit function theorem. Note, however, that exactly at the bifurcation points we see that F_u becomes singular and we lose uniqueness which results in a breakdown of our parametrisation. There are multiple ways to elude this problem and one of the most widespread options is to change the parametrisation of the solution branch such that at bifurcation points the partial Fréchet derivative is non-singular. A natural way to parametrise a curve is the arclength parametrisation which is depicted in Figure 5.1. If we denote the arclength parameter by s we now look for $u(s)$ and $\lambda(s)$ and have to add to (5.1) the arclength constraint to get a closed system of equations. Doing so we arrive at

$$F\left(u(s), \lambda(s)\right) = 0, \tag{5.2a}$$

$$\left\| \left(\frac{du(s)}{ds}, \frac{d\lambda(s)}{ds} \right) \right\|_{U\times\Lambda} = 1. \tag{5.2b}$$

With this added equation we are now able to continue branches through bifurcation points and we compute a bifurcation curve parametrised by s instead of λ.

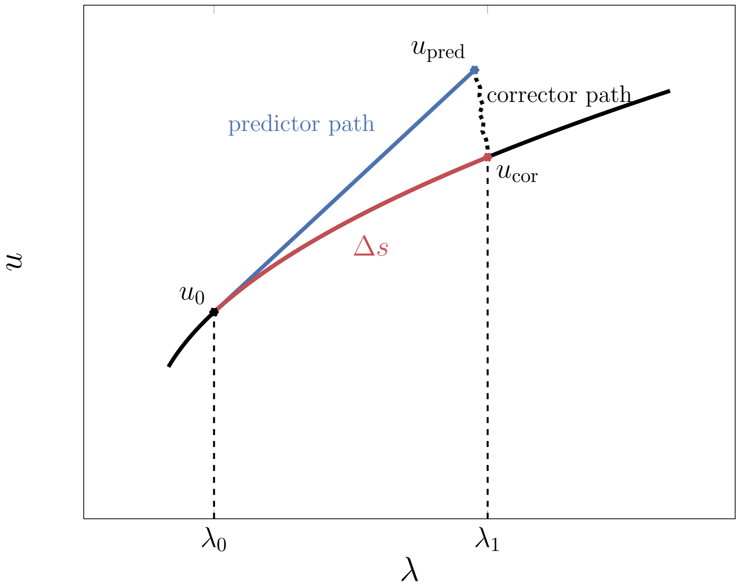


Figure 5.1: Illustration of a predictor-corrector step for a scalar problem, where we have used a tangent predictor to make an arclength step of length Δs. The predictor step is shown in blue and serves as an initialisation for the correction step, whose path is shown by the dashed line.

We make use of a standard framework in numerical bifurcation analysis, predictor-corrector methods, which we sketch in Figure 5.1. One of the most used predictor steps is based on the observation that along the curve, parametrised by λ, we have

$$0 = F_u(u, \lambda)u_\lambda + F_\lambda(u, \lambda), \tag{5.3}$$

known as the Davidenko equation. If given (u, λ) we can solve this for the tangent vector u_λ. Note that in the natural parametrisation this equation breaks down at the bifurcation due to the Jacobian F_u becoming singular. To overcome this problem we again use the arclength parametrisation (5.2) and solve for a tangent vector, depending on s, in this augmented system, which is done by solving

$$F_u \frac{du(s)}{ds} + F_\lambda \frac{d\lambda(s)}{ds} = 0, \tag{5.4a}$$

$$\left\| \left(\frac{du(s)}{ds}, \frac{d\lambda(s)}{ds} \right) \right\|_{U\times\Lambda} = 1. \tag{5.4b}$$

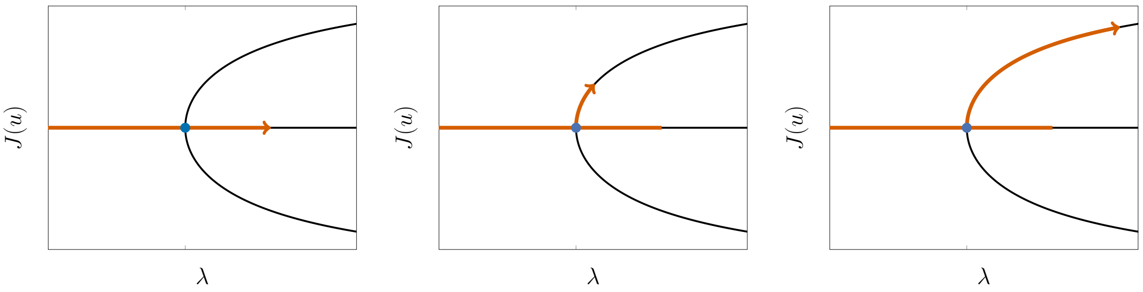
One constructs a predictor for a new solution in this tangent direction and uses this as an initial guess for the computation of the corrected solution which should lie on our bifurcation curve again as depicted in Figure 5.1.

Therefore under the assumption of a starting point on the branch, or at least a point sufficiently close for Newton to converge, we have a numerical procedure to trace out the solution branch. For a more detailed treatment of arclength continuation and different continuation techniques see, for example, [44].

Branch detecting and branch switching

In addition to continuation techniques we need methods which can detect bifurcation points and a device that gives us the option to change solution branch at these bifurcation points. The detection of a bifurcation is often done by looking at a test function, which is constructed such that it has a zero at bifurcation points. The specific form of the test function relies heavily on the type of bifurcation one wants to detect. The type of bifurcation points is determined by the behaviour of the eigenvalues of the Jacobian F_u and in practice test functions therefore rely often on calculation of either eigenvalues or determinants of the Jacobian, which is numerically expensive. This makes this approach not scalable to large-scale systems, such as those arising in discretisations of PDEs, as the cost of the test-function becomes prohibitive. We will not discuss test functions in more detail here, but instead refer the reader to [28, 44].

Having detected a bifurcation point one now knows that multiple solutions have to exist close to this point. Different techniques, dependent on the type of bifurcation, have been devised to switch to different branches at bifurcation points and we refer for details again to [28, 44]. Having switched branches we can, once more, use continuation techniques to trace out these branches fully. A sketch of how the aforementioned techniques are combined is given in Figure 5.2.



(a) Detect bifurcation point (b) Switch to new branch (c) Continuation new branch

Figure 5.2: Illustration of standard numerical bifurcation algorithm. We start out by continuation of a branch as in Figure 5.2a. If we detect a bifurcation point we can use specialised techniques to switch branch at the bifurcation point, depicted in Figure 5.2b. If we have a point on the new branch we can use standard continuation to trace out the new branch as in the rightmost figure.

5.2 Deflation and continuation combined

As mentioned in the introduction there is a flaw in the current numerical bifurcation algorithms in the sense that they are not able to detect different branches in the absence of a bifurcation point. A standard situation in which this might arise is in the case of bifurcation branches which are disconnected from the initial branch such as we saw in Figure 1.3. The numerical continuation techniques are capable of tracing out branches if and only if we can initialise them with a point which actually lies on the solution branch. The problem of tracing out diagrams with disconnected branches, such as in Figure 5.3, now asks for a method to calculate multiple solution points for a given parameter value. The deflation technique as discussed in the previous chapters can provide an answer to this problem and we will now shortly discuss two novel applications of augmenting numerical bifurcation algorithms with deflation.

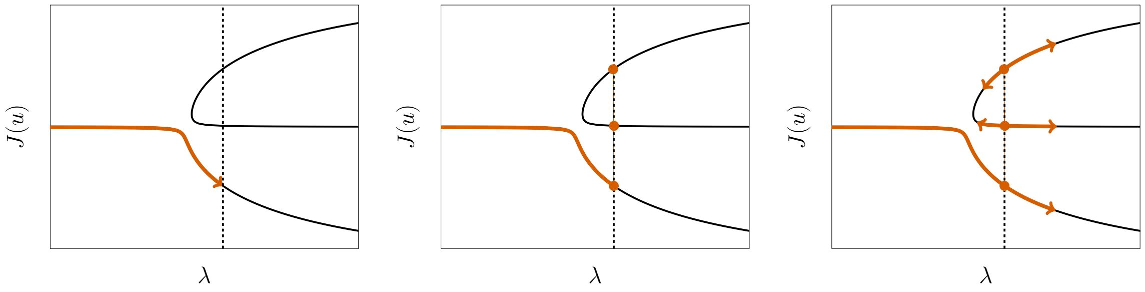
Detecting (disconnected) branches

As deflation is able to compute multiple solutions from one single initial guess, we propose the augmentation of numerical bifurcation analysis with deflation in the following way; we start with a given initial solution to trace out an initial branch. If the parameter λ reaches some pre-determined value $\tilde{\lambda}$ we stop the continuation and we thus have some solution $(\tilde{u}, \tilde{\lambda})$ to (5.1). Instead of tracing a branch we now look for a solution to (5.1) at $\tilde{\lambda} + \Delta\lambda$ based on our knowledge of the solution at $\tilde{\lambda}$, i.e. we use $\tilde{u}$ as an initial guess for the problem at parameter $\tilde{\lambda} + \Delta\lambda$. If $\Delta\lambda$ is sufficiently small, this will at least yield the continuation of the solution $\tilde{u}$ on the same branch. Having found this continuation we can then employ deflation, which will prevent the solver to converge to the same branch again. If we now retry with $\tilde{u}$ as an initial guess we are guaranteed not to converge to the same branch again. Under suitable conditions we can now find different solutions to (5.1) for the same parameter value, which thus must lie on a different branch. In this way we can generate initial points on multiple branches and as we now have good initial points we can again use continuation techniques to fully trace out these branches. For a sketch of this method see Figure 5.3.

Note that this approach does not rely on either the presence or proximity of a bifurcation point and can thus be applied to find disconnected branches.

Branch switching

The previous framework provides a method to find disconnected branches by using deflation at some specific parameter values, without any extra knowledge on the bifurcation structure of the problem. If we, however, do know more details about the specific bifurcation structure in the form of the approximate location of a bifurcation point we can use this to our advantage. Close to a bifurcation point, the solutions which emanate from this point will be close to each other. This means that it is likely that an initial guess close to the bifurcation point will be able to yield multiple solutions using deflation. Deflation can therefore be used to perform branch switching.



(a) Initial continuation

(b) Deflate to find multiple branches

(c) Continuation of branches

Figure 5.3: Illustration of the use of deflation in numerical bifurcation analysis. We start out by continuation of a branch as in Figure 5.3a. Along our branch we stop and fix the parameter λ and attempt a deflation step to find multiple solutions for this parameter value as in Figure 5.3b. If the deflation step is successful we have a points on multiple branches and we can use standard continuation to trace out these branches as in the rightmost figure.

Although many different branch switching techniques do already exist, deflation is useful in this context for two reasons.

First of all its generality, as branch switching with deflation is not dependent on the bifurcation type and can be used for a variety of bifurcations. The only requirement that we need for deflation is that the problem can be formulated as a system of equations of the form (5.1).

More importantly, branch switching with deflation can be made scalable. The classical branch switching techniques depend heavily on detection of the bifurcation point, which is computationally expensive. By contrast, if a good preconditioner is available for the underlying problem Farrell et al. [18] showed that the deflated systems can be solved with the same computational efficiency as the underlying undeflated system. Deflation for branch switching is thus the first *scalable* bifurcation technique.

5.3 Comparison with existing software

AUTO-07P

A large collection of software packages is available for numerical bifurcation analysis and an extensive list is described in [16]. One of the most widely used packages is AUTO-07P [15], which is written in Fortran 95 and partially equipped with a Python wrapper. The program is known for its reliability and offers both continuation and branch switching algorithms for various types of bifurcations. AUTO-07P can apply its bifurcation analysis tools to two different classes of equations. First of all it can do a limited analysis on algebraic systems of the form

$$0 = F(u, \lambda), \quad u \in \mathbb{R}^n, \lambda \in \mathbb{R}^m. \tag{5.5}$$

This is for example useful in order to track bifurcations of steady state solutions to a system of ODEs as this naturally results in a system of the form (5.5). The second class of problems for which it can be used are one-dimensional boundary value problems (BVPs) of the form

$$u'(\tau) = F(u(\tau), \lambda), \quad u(\tau) \in \mathbb{R}^n, \lambda \in \mathbb{R}^m, \tau \in [a, b], \tag{5.6}$$

with boundary conditions and possibly integral constraints. Besides its use in looking at bifurcation diagrams for BVPs this also allows one to investigate the stationary states of certain parabolic PDEs describing reaction-diffusion equations of the form

$$u_\tau = Du_{xx} + F(u, \lambda), \quad u(x) \in \mathbb{R}^n, \lambda \in \mathbb{R}^m, x \in [a, b], \tag{5.7}$$

where D is a diagonal $\mathbb{R}^{n\times n}$ matrix containing the diffusion constants. The steady states to this problem can be rewritten to (5.6).

Practical implementation deflation and continuation

We compare the bifurcation diagrams computed with AUTO-07P with our implementation of deflation and continuation techniques. Our implementation was written in FEniCS [31], an automated programming environment for solving differential equations using the finite element method (FEM), which we control using a Python interface. We need to supply the equations in variational form, see [31], and the FEM discretisation to be used. FEniCS then assembles the discretised system of FEM equations which can be solved using numerical linear algebra software, in this case the PETSc toolbox [4, 5, 10] in combination with the MUMPS package [1, 2].

To carry out the numerical bifurcation analysis we use an arclength continuation and deflation class written for FEniCS by P.E. Farrell [19] augmented by a branch switching method purely based on deflation and a bifurcation detector based on the determinant test function. Our implementation of branch switching is thus purely based on deflation in contrast to other software such as AUTO-07P. The algorithm gives the option to only switch branches at bifurcation points or to safeguard against missed branches by using deflation away from bifurcation points at points specified by the user.

5.3.1 Steady state continuation

We consider the problem of steady state continuation of ODEs, a BVP and a nonlinear diffusion equation so that we can make use of AUTO-07P. We assume that we are initially given a steady state by either an explicit construction, a homotopy [44] or by integrating the time-dependent problem.

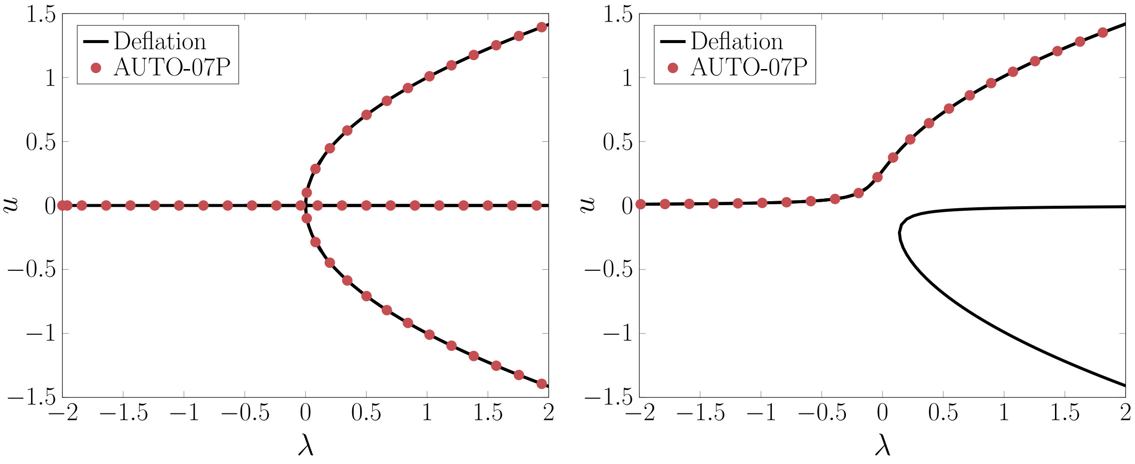
Algebraic problems

The first problem that we consider is the cusp, the second of the elementary catastrophes in catastrophe theory, which can depend on two parameters λ and μ. The ODE that we consider is

$$u' = u(\lambda - u^2) + \mu, \tag{5.8}$$

where now $u(t)$ and u' means differentiation with respect to t. We are interested in the steady states of this ODE as a function of the parameters, which can be achieved by setting $u' = 0$ and solving the algebraic equation.

We start with the case $\mu = 0$ and allow λ to vary. This is now the normal form of a pitchfork bifurcation in which two non-trivial (symmetrical) solutions emanate from a trivial branch in the bifurcation point. Since the bifurcation point connects all branches, both AUTO-07P and our deflation implementation trace out the full bifurcation diagram, see Figure 5.4a.



(a) $\mu = 0$

(b) $\mu = 0.1$

Figure 5.4: Bifurcation diagram for (5.8) as computed by AUTO-07P and the deflation technique. For AUTO-07P we plot a data point every 2 steps. In the case of $\mu = 0$ there is a supercritical pitchfork bifurcation which connects all branches, whereas for $\mu = 0.1$ we see that the symmetry of the pitchfork bifurcation has been broken and the branches are disconnected. In this case AUTO-07P can only find one branch.

If we now set $\mu \neq 0$ the reflection symmetry $u \mapsto -u$ gets broken and we arrive at the imperfect pitchfork bifurcation. This is an example of a bifurcation diagram with disconnected branches. We have one trivial branch and two branches which arise from a fold bifurcation and do not connect with the trivial branch. As a result AUTO-07P is not able to compute the full bifurcation diagram if we start from the trivial branch, see Figure 5.4b.

The next problem is inspired by Rosenblat and Davis [41] who studied bifurcations at infinity as a possible explanation of observations in stability analysis of Hagen-Poiseuille flow and Couette flow. We start with the ODE

$$u' = -\lambda u + u^2 - u^3, \tag{5.9}$$

and again set $u' = 0$. Now the bifurcation structure consists of a trivial branch and two solutions emanating from a subcritical fold bifurcation or saddle-node bifurcation at $\lambda = 1/4$. The non-trivial branches connect with the zero branch in a transcritical bifurcation at $\lambda = 0$ and the bifurcation diagram is thus fully connected. Deflation and AUTO-07P therefore correctly compute the full bifurcation diagram, see Figure 5.5.

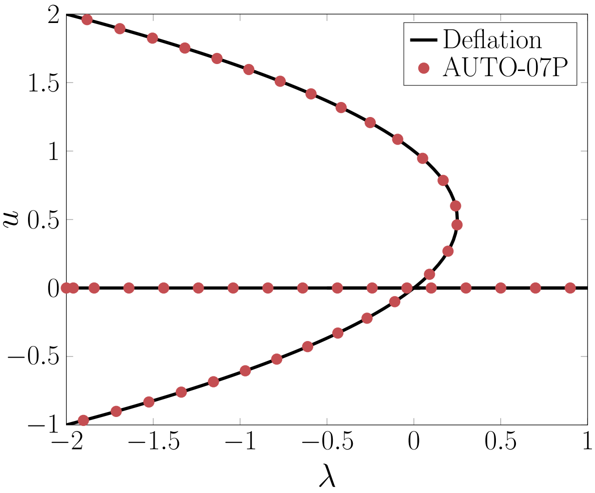


Figure 5.5: Bifurcation diagram for (5.9) as computed by AUTO-07P and the deflation technique. For AUTO-07P we plot a data point every 2 steps. There are two bifurcation points, a transcritical bifurcation at $\lambda = 0$ and a subcritical fold bifurcation at $\lambda = 1/4$, which connect all the branches. Both AUTO-07P and deflation compute the full bifurcation diagram.

If we now, however, make the transformation $\mu = 1/\lambda$ we drastically change the bifurcation diagram. We consider,

$$u' = -\frac{u}{\mu} + u^2 - u^3, \tag{5.10}$$

for $\mu > 0$. We still have a trivial branch and two non-trivial branches originating in a supercritical fold bifurcation at $\mu = 4$. However, we see now that the transcritical bifurcation at $\lambda = 0$ becomes a bifurcation at $\mu = \infty$ and as a result the branches only connect at infinity, which leaves them uncomputable to AUTO-07P, see Figure 5.6.

The seemingly disconnected branches are important to the understanding of the problem. A linear stability analysis would tell that the trivial branch is stable, but for sufficiently large μ any positive perturbation would make the solution switch to one of the disconnected branches and this indicates that a linear stability analysis can yield misleading results.

The previous examples show that deflation can trace out complete diagrams where AUTO-07P fails in the case of algebraic problems. For the preceding examples this result could have been anticipated based on the results in Section 4.1 and the observation that the considered problems were in fact all cubic polynomials. We will now show that in the case of more complicated problems deflation still gives better results than AUTO-07P.

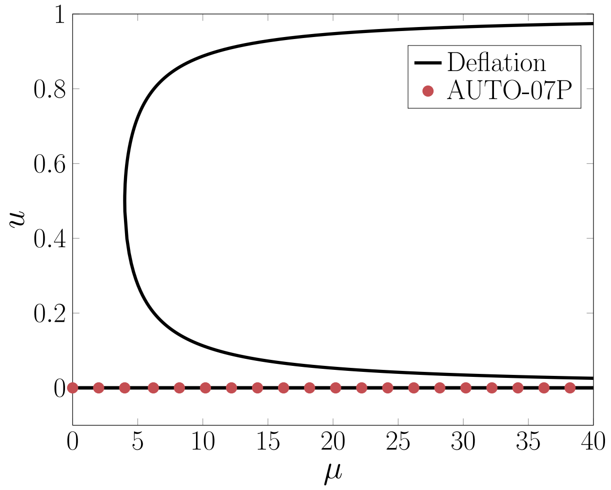


Figure 5.6: Bifurcation diagram for (5.10) as computed by AUTO-07P and the deflation technique. For AUTO-07P we plot a data point every 20 steps. There is one finite bifurcation point, a supercritical fold bifurcation at $\mu = 4$. The branches are however disconnected for all finite values of μ. One of the branches that emanates from the fold bifurcation connects to the trivial branch at infinity. AUTO-07P does not detect a bifurcation point and thus only finds the trivial branch by continuation, whereas deflation traces out the full bifurcation diagram.

Boundary value problems

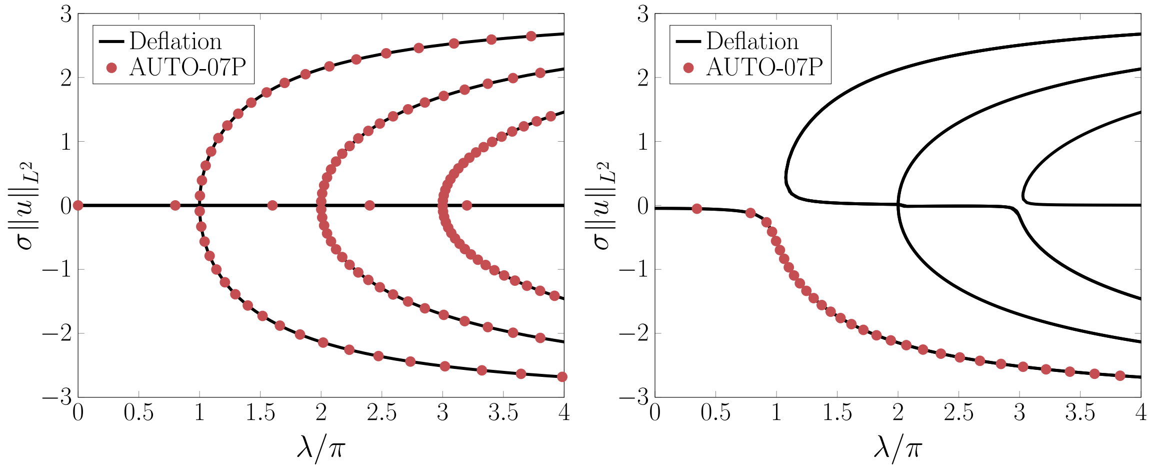
We now return to the problem sketched in the introduction of a slender beam under a loading, where we assume the loading to be in the longitudinal direction as in Figure 1.2. This problem has a long mathematical history tracing back to Galileo, the Bernoulli family and Euler [30]. It was the latter who showed that the deformation of the beam can be described in terms of the angle θ made relative to the vertical axis as a function of the arclength s. If the beam is subject to a transversal force the steady states of the beam are governed by Eulers elastica equation

$$\theta_{ss} + \lambda^2 \sin(\theta) = \mu, \quad 0 \le s \le 1, \tag{5.11a}$$

$$\theta(0) = \theta(1) = 0, \tag{5.11b}$$

where λ is a non-dimensional load and μ depicts a non-dimensional transversal force. In the case $\mu = 0$, i.e. in the absence of a transversal force, the initially straight solution, $\theta \equiv 0$, forms the trivial branch and is valid for all s. A series of pitchfork bifurcations at $\lambda = m\pi$ for $m \in \mathbb{N}$ result in the buckled modes emanating from the trivial branch. Both deflation and AUTO-07P therefore compute a complete bifurcation diagram, see Figure 5.7a.

If we introduce a transversal force by taking $\mu \neq 0$ we destroy a reflection symmetry, just as we saw for the cusp bifurcation. The initial branch now disconnects from all the other branches in a similar fashion as we observed before. The difference now, however, is that there is an infinite number of branches which originally were connected to the initial branch and which now become disconnected. The result is



(a) $\mu = 0$

(b) $\mu = 0.5$

Figure 5.7: Bifurcation diagram for (5.11) as computed by AUTO-07P and the deflation technique, where we take the scalar measure $J(u) = \sigma\|u\|_{L^2}$, a signed norm. Here σ is given by $\sigma = u_x(0)$. For AUTO-07P we plot a data point every 8 steps. In the case of $\mu = 0$ a series of supercritical pitchfork bifurcations at $\lambda = m\pi$ for $m \in \mathbb{N}$ connects all branches. For $\mu = 0.1$ we see that the symmetry of the pitchfork bifurcation has been broken and the branches are disconnected. In this case AUTO-07P can only find one branch.

that those branches are not found by continuation in AUTO-07P, yielding a very unsatisfactory representation of the bifurcation structure. Deflation on the other hand is able to trace out disconnected branches as we see in Figure 5.7b.

Lastly we return to the work by Rosenblat and Davis [41] on bifurcations from infinity. We consider the PDE

$$u_t = \frac{u_{xx}}{\mu} + u^2 - u^3, \qquad 0 \le x \le 1, \mu > 0, \tag{5.12a}$$

$$u(0,t) = u(1,t) = 0, \tag{5.12b}$$

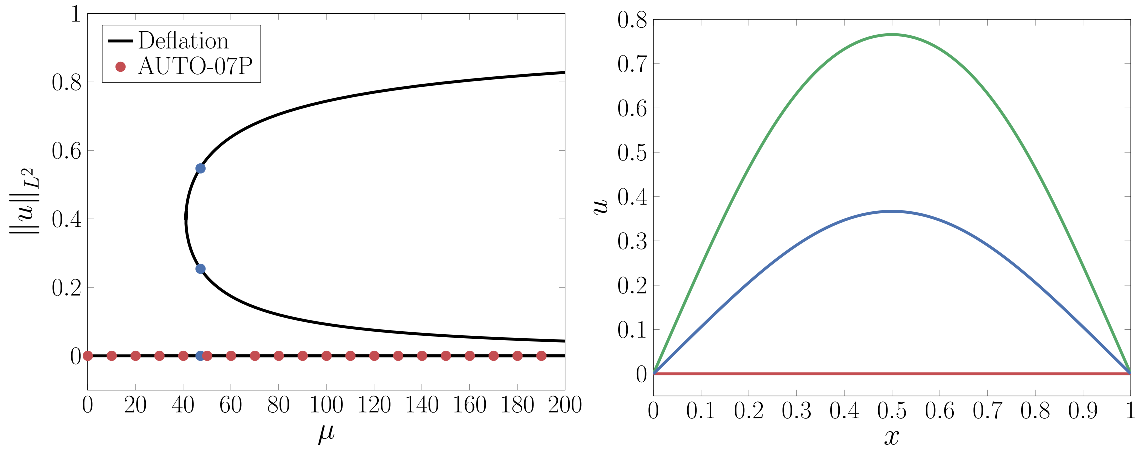
$$u(x,0) = u_0(x), \tag{5.12c}$$

and look for its steady state solutions. As one can see, the trivial state is a solution for all μ. Analogous to (5.10) Rosenblat and Davis showed using singular perturbation theory that there are two non-trivial solutions which emanate from a supercritical fold bifurcation and one of these branches connects with the trivial branch at $\mu = \infty$. These two non-trivial branches can not be found by AUTO-07P using continuation and bifurcation tools, see Figure 5.8.

This PDE can be used to show that there exist systems in which an infinite number of periodic solutions of small norm bifurcate from the trivial solution at infinity. This has implications for the use of standard analytic methods to compute bifurcation solutions [41].

If we transform the system to a form similar to (5.9) by identifying $\lambda = 1/\mu$,

$$u_t = \lambda u_{xx} + u^2 - u^3, \qquad 0 \le x \le 1, \tag{5.13}$$

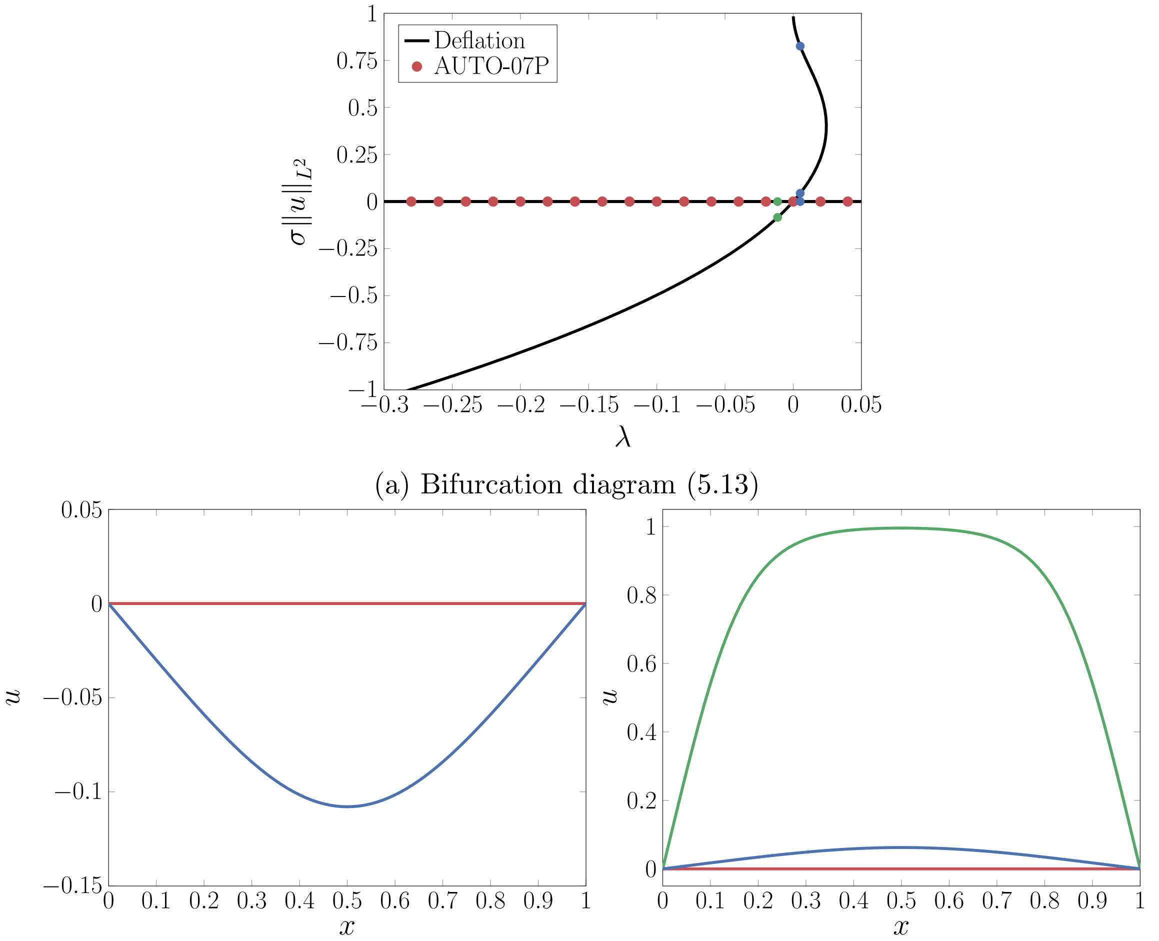


(a) Bifurcation diagram (5.12)

(b) Solutions (5.12) at $\mu \approx 47.303$

Figure 5.8: Bifurcation diagram for (5.12) as computed by AUTO-07P and the deflation technique, where we take the scalar measure $J(u) = \|u\|_{L^2}$. Solutions at ● are depicted in Figure 5.8b.
For AUTO-07P we plot a data point every 100 steps. There is one finite bifurcation point, a supercritical fold bifurcation at $\mu \approx 41.157$. One of the branches which emanates from this fold bifurcation connects to the trivial branch at infinity. AUTO-07P does not detect a bifurcation point and thus only finds the trivial branch by continuation, whereas deflation traces out the full bifurcation diagram.

with the same boundary and initial conditions, one would expect to move the bifurcation at infinity to $\lambda = 0$. However, we do not detect a bifurcation point on the trivial branch using AUTO-07P, see Figure 5.9. A possible explanation would be that the point $\lambda = 0$ is at the same time a singular point for the equation and AUTO-07P might not be suited to solve such problems. Deflation with the bifurcation detector is able to trace out the full diagram as shown in Figure 5.9.



(a) Bifurcation diagram (5.13)

(b) Solutions (5.13) at $\lambda = -0.01$

(c) Solutions (5.13) at $\lambda = 0.005$

Figure 5.9: Bifurcation diagram for (5.13) as computed by AUTO-07P and the deflation technique, where we take the scalar measure $J(u) = \sigma\|u\|_{L^2}$, a signed norm. Here σ denotes the sign of the function, determined by $u(0.5)$ in this case. Solutions at ● and ● are depicted in Figure 5.9b and Figure 5.9c respectively.
For AUTO-07P we plot a data point every 2 steps. There are two bifurcation points, a transcritical bifurcation at $\lambda = 0$ and a supercritical fold bifurcation at $\lambda \approx 0.024296$. AUTO-07P does not detect a bifurcation point, possibly because the point is also a singular point of the equation, and thus only finds the trivial branch by continuation, whereas deflation traces out the full bifurcation diagram.

6. Conclusions

In this thesis we studied the application of deflation techniques to robustly tracing out bifurcation diagrams, even in cases where solution branches are disconnected in the parameter domain of study.

The first part was concerned with deriving sufficient conditions which guarantee that an initial guess x_0 will converge to multiple solutions using deflation and Newton's method, a result which is not specific to the computation of bifurcation diagrams.

In Chapter 2 an introduction to Newton's method and deflation was given. We derived a new result on the existence of infinitely many points which converge to all the roots of a scalar polynomial with real roots using Wilkinson deflation.

In the quest for more general conditions on functions on Banach spaces we reviewed local convergence theorems for Newton's method in Chapter 3. As the idea for sufficiency conditions is that an initial guess lies in different convergence regions of different roots before and after deflation we made a direct comparison between the theorems from this perspective. Such a review has not been published before and we fill in some details and gaps of the existing review literature. We found that in order to derive sufficiency conditions we need to employ Newton convergence theorems which center their convergence regions around the roots of the equation instead of the initial guess, thereby limiting the theorems that we can use.

In Chapter 4 we showed some example convergence regions for specific functions illustrating that using convergence balls of local theorems we can in fact show convergence to multiple solutions in some cases. Two of the local convergence theorems, the affine covariant Rall-Rheinboldt and Smale's γ-theorem, are then used to derive the first known sufficiency conditions for convergence towards multiple roots using deflation in Banach spaces. These conditions extend the conditions that are assumed to hold for the original function by the original convergence theorems so that the same conditions hold for the deflation operator as well. The deflated function, which is a product of the deflation operator and the original function, can then be shown to satisfy similar conditions to the original function and therefore we can apply the convergence theorems before and after deflation.

In the last part of this thesis we investigated the practical use of deflation in robustly tracing out bifurcation diagrams. We tested an implementation of a deflation and continuation algorithm with one of the best available standard software packages for numerical bifurcation analysis, AUTO-07P. Results in Chapter 5 show that for a range of test problems, containing disconnected branches or bifurcations at infinity, AUTO-07P fails in tracing out the complete bifurcation diagrams, whereas the combination of deflation and continuation succeeds in correctly computing the full bifurcation diagram.

Further work

Both the theoretical and numerical results leave ample space for future research. Deflation as a technique to find multiple solutions has a wide range of possible applications as pointed out in [18], however we will focus here on the computation of bifurcation diagrams using deflation. We will highlight some possible extensions which we believe would be fruitful:

1. The result on Wilkinson deflation on polynomials with real roots relies heavily on the interplay between convexity of the function and properties of Newton's method. Some results on extending the concept of convex functions to finite dimensional real vector spaces have been made in [35, Chapter 13]. Can this be used to extend the scalar case result on Wilkinson deflation?

2. The local convergence theorems considered in this thesis do not make use of any special structure that could be present in bifurcation problems. For example, deflation is expected to work particularly well in regions where the solutions are close to the initial guess, such as is the case for bifurcation points. Does adding information from bifurcation theory yield extra insight?

3. The Banach space framework covers PDEs and integral equations as well and as showed in [18] we can use deflation to trace out connected bifurcation diagrams for some PDE examples. Can we find examples of disconnected branches (for example in two-dimensional elastic beam buckling) or bifurcations at infinity in PDEs or integral equations and can we compute, using deflation, their complete bifurcation diagrams?

4. Deflation can make the switching of solution branches scalable. However, we can often not compute the location of bifurcation points using classical test functions for large scale systems, which would mean that regions of good performance of deflation can be missed and heuristic methods for the application of deflation have to be used. We can therefore ask whether we can devise scalable indicator methods which can hint at the proximity of a bifurcation point? For example, for medium-sized problems, where we can use a LU-decomposition of the Jacobian whilst solving the Newton iterations, the determinant test function can be calculated cheaply. A possible extension of such a technique to the case of solving large-scale systems using preconditioned GMRES is given in [22].

A. Proofs of auxiliary results

A.1 Convergence Newton's method for convex real functions

Lemma. *Let $f : \mathcal{I} \subseteq \mathbb{R} \to \mathbb{R}$ be a convex, differentiable function where $\mathcal{I}$ is an open interval such that there exist precisely one $x^* \in \mathcal{I}$ with $f(x^*) = 0$. Starting from $x_0 \in \mathcal{I}$ with $f'(x_0) \neq 0$, Newton's method converges to x^* if $x_1 \in \mathcal{I}$.*

Proof. Since f is convex and differentiable we know that for all $x, y \in \mathcal{I}$ we have

$$f(x) \geq f(y) + f'(y)(x - y), \tag{A.1}$$

which states that the function must lie above its tangents. In addition one knows that $f'(x)$ is monotonically increasing on $\mathcal{I}$.

First of all we note that by (A.1) and the construction of the Newton iterates that $f(x_1) \geq f(x_0) - f(x_0) = 0$ regardless of the initial guess.

Now let us assume that $f'(x_0) > 0$. Suppose that $f(x_0) < 0$, then by construction of the Newton iterates we have $x_1 > x_0$. As f' is monotonically increasing on $\mathcal{I}$ this implies $f'(x_1) > 0$ as well. We thus have $x_1 \in \mathcal{I}$ (by assumption of the lemma), $f(x_1) \geq 0$ and $f'(x_1) > 0$. If we assume that $f(x_0) = 0$ we are done, as then we would have $x_0 = x^*$. If instead we assumed $f(x_0) > 0$ we know that $x^* < x_0$. Since the function has to lie above its tangent at x_0 we conclude that $x^* \leq x_1 < x_0$ and thus $f'(x_1) \geq 0$ as f' is monotonically increasing and $f'(x^*) \geq 0$. Therefore we know that under the assumptions of the theorem and $f'(x_0) > 0$ there holds $x_1 \in \mathcal{I}, f(x_1) \geq 0$ and $f'(x_1) \geq 0$. If $f'(x_1) = 0$ we conclude that by convexity we have $0 = f(x^*) \geq f(x_1) \geq 0$ and thus $f(x_1) = 0$ and $x_1 = x^*$.

Assume that for all $1 \leq n \leq N$ that $x_n \in \mathcal{I}$, $f(x_n) \geq 0$ and $f'(x_n) > 0$. Then it follows from the construction of the Newton sequence that $x_{n+1} \leq x_n$ for all $n \leq N$. Suppose that $x_{N+1} \notin \mathcal{I}$, then it follows from (A.1) that $f(x) > 0$ for all $x \in \mathcal{I}$, but then f cannot have a root in $\mathcal{I}$, which contradicts our assumption. Therefore it must be that $x_{N+1} \in \mathcal{I}$, from which we can immediately conclude that $f(x_{N+1}) \geq 0$.

It remains to prove that $f'(x_{N+1}) > 0$. If $f'(x_{N+1}) = 0$ we find by a similar argument as before that $x_{N+1} = x^*$ and we are done. Suppose thus, again by contradiction, that $f'(x_{N+1}) < 0$. As the function f is strictly positive on $[x_{N+1}, \infty) \cap \mathcal{I}$ and f' is negative on $(-\infty, x_{N+1}] \cap \mathcal{I}$ by monotonicity we know that f is strictly positive on $\mathcal{I}$ and thus cannot have a root, which contradicts our assumptions. Therefore $f'(x_{N+1}) > 0$.

By induction it then follows that either $x_n = x^*$ for all $n \geq M$ for some $M \in \mathbb{N}$ or for all $n \in N_{\geq 1}$ that $x_n \in \mathcal{I}$, $f(x_n) \geq 0$ and $f'(x_n) > 0$.

In the latter case we see that as the sequence $\{x_n\}$ is constructed by Newton's method it furthermore follows that $x_n \leq x_{n+1}$ for all $n \in N_{\geq 1}$, and thus it is a decreasing sequence. By the monotone convergence theorem it then follows that the sequence $\{x_n\}$ converges, i.e. $\lim_{n\to\infty} x_n = \bar{x}$. Suppose $\bar{x} \notin \mathcal{I}$, then it must hold that f is strictly positive on $\mathcal{I}$, but this contradicts the assumption on x^*. Therefore $\bar{x} \in \mathcal{I}$ and by construction of the Newton sequence we must have that $\bar{x} = x^*$.

The assumption $f'(x_0) > 0$ leads to a monotonic decreasing sequence to $x^* \in \mathcal{I}$. A similar argument with $f'(x_0) < 0$ results in a monotonic increasing sequence to $x^* \in \mathcal{I}$. □

A.2 Product of Lipschitz continuous functions

Lemma. *Let X, Y, Z be Banach spaces and $G : X \to L(Y, Z)$ and $F : X \to Y$ be Lipschitz continuous functions on the open subset $D \subseteq X$ with Lipschitz constants ω_F and ω_G respectively. Assume furthermore that F is bounded on D, i.e. there exist $N_F, N_G \in \mathbb{R}$ such that for all $x \in D$ we have $\|F(x)\| \leq N_F$ and $\|G(x)\| \leq N_G$, as G is bounded by definition. Then the product $GF : X \to Z$ is bounded and Lipschitz continuous on D with Lipschitz constant $(N_F\omega_G + N_G\omega_F)$.*

Proof. Let $x, y \in D$ be arbitrary points. As both F and G are bounded on D their product is bounded as well

$$\|G(x)F(x)\| \leq \|G(x)\|\|F(x)\| \leq N_G N_F < \infty.$$

From the assumptions we can furthermore derive that

$$\begin{aligned}
\|G(x)F(x) - G(y)F(y)\| &= \|G(x)F(x) - G(x)F(y) + G(x)F(y) - G(y)F(y\| \\
&\leq \|G(x)F(x) - G(x)F(y)\| + \|G(x)F(y) - G(y)F(y)\| \\
&= \|G(x)\|\|F(x) - F(y)\| + \|G(x) - G(y)\|\|F(y)\| \\
&\leq N_G\|F(x) - F(y)\| + N_F\|G(x) - G(y)\| \\
&\leq N_G\omega_F\|x - y\| + N_F\omega_G\|x - y\| \\
&= (N_F\omega_G + N_G\omega_F)\|x - y\|,
\end{aligned}$$

which proves the claim. □

Bibliography

1. Amestoy, P. R., Duff, I. S., L'Excellent, J.-Y. & Koster, J. A Fully Asynchronous Multifrontal Solver Using Distributed Dynamic Scheduling. *SIAM Journal on Matrix Analysis and Applications* **23,** 15–41 (2001).
2. Amestoy, P. R., Guermouche, A, L'Excellent, J.-Y. & Pralet, S. Hybrid scheduling for the parallel solution of linear systems. *Parallel Computing* **32,** 136–156 (2006).
3. Arnold, V. I., Afrajmovich, V. S., Ilyashenko, Y. S. & Shilnikov, L. P. *Bifurcation Theory and Catastrophe Theory, Encyclopaedia of Mathematical Sciences, Vol. 5* (Springer-Verlag Berlin Heidelberg, 1994).
4. Balay, S., Gropp, W. D., McInnes, L. C. & Smith, B. F. *Efficient Management of Parallelism in Object Oriented Numerical Software Libraries* in *Modern Software Tools in Scientific Computing* (eds Arge, E, Bruaset, A. M. & Langtangen, H. P.) (Birkhäuser Press, 1997), 163–202.
5. Balay, S. *et al. PETSc Users Manual* tech. rep. ANL-95/11 - Revision 3.6 (Argonne National Laboratory, 2015).
6. Birkisson, A. *Numerical Solution of Nonlinear Boundary Value Problems for Ordinary Differential Equations in the Continuous Framework* PhD thesis (University of Oxford, 2014).
7. Blum, S., Cucker, L., Shub, F. & Smale, S. *Complexity and Real Computation* (Springer-Verlag New York, 1997).
8. Brown, K. M. & Gearhart, W. B. Deflation Techniques for the Calculation of Further Solutions of a Nonlinear System. *Numerische Mathematik* **16,** 334–342 (1971).
9. Ciarlet, P. G. & Mardare, C. On the Newton-Kantorovich Theorem. *Analysis and Applications* **10,** 249–269 (2012).
10. Dalcin, L. D., Paz, R. R., Kler, P. A. & Cosimo, A. Parallel distributed computing using Python. *Advances in Water Resources* **34,** 1124–1139 (2011).
11. Deuflhard, P. & Heindl, G. Affine Invariant Convergence Theorems for Newtons Method and Extensions to Related Methods. *SIAM Journal on Numerical Analysis* **16,** 1–10 (1979).
12. Deuflhard, P. *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms* (Springer-Verlag Berlin Heidelberg, 2011).
13. Deuflhard, P. & Potra, F. A. Asymptotic Mesh Independence of Newton-Galerkin Methods via a Refined Mysovskii Theorem. *SIAM Journal on Numerical Analysis* **29,** 1395–1412 (1992).
14. Deville, R., Fonf, V. & Hájek, P. Analytic and Cˆk approximations of norms in separable Banach spaces. *Studia Math* **120,** 61–74 (1996).

15. Doedel, E. *et al. AUTO-07P: Continuation and Bifurcation Software for Ordinary Differential Equations* (2008).

16. *Dynamical Systems Web: Software* <http://www.dynamicalsystems.org/sw/sw/> (visited on Aug. 24, 2015).

17. Fabian, M., Habala, P., Hájek, P., Montesinos, V. & Zizler, V. *Banach Space Theory: The Basis for Linear and Nonlinear Analysis* (Springer-Verlag New York, 2011).

18. Farrell, P. E., Birkisson, A. & Funke, S. W. Deflation Techniques for Finding Distinct Solutions of Nonlinear Partial Differential Equations. *SIAM Journal on Scientific Computing* **37,** A2026–A2045 (2015).

19. Farrell, P. E. *Personal communication* 2015.

20. Fry, R. & McManus, S. Smooth Bump Functions and the Geometry of Banach Spaces. *Expositiones Mathematicae* **20,** 143–183 (2002).

21. Galántai, A. The theory of Newton's method. *Journal of Computational and Applied Mathematics* **124,** 25–44 (2000).

22. García-Archilla, B., Sánchez, J. & Simó, C. Krylov methods and determinants for detecting bifurcations in one parameter dependent partial differential equations. *BIT Numerical Mathematics* **46,** 731–757 (2006).

23. Gutiérrez, J. A new semilocal convergence theorem for Newton's method. *Journal of Computational and Applied Mathematics* **79,** 131–145 (1997).

24. Hájek, P. & Troyanski, S. Analytic norms in Orlicz spaces. *Proceedings of the American Mathematical Society* **129,** 713–717 (2001).

25. Hilgert, J. & Neeb, K.-H. *Structure and Geometry of Lie Groups* (Springer-Verlag New York, 2011).

26. Hunter, J. K. & Nachtergaele, B. *Applied Analysis* (World Scientific Pub Co Inc, 2001).

27. Kantorovich, L. On Newtons Method for Functional Equations. *Doklady Akademii Nauk SSSR* **59,** 1237–1249 (In Russian) (1948).

28. Keller, H. B. *Lectures on Numerical Methods in Bifurcation Problems* (Published for the Tata Institute of Fundamental Research by Springer-Verlag, 1987).

29. Leonard, E. & Sundaresan, K. A note on smooth Banach spaces. *Journal of Mathematical Analysis and Applications* **43,** 450–454 (1973).

30. Levien, R. *The elastica: a mathematical history* tech. rep. UCB/EECS-2008-103 (EECS Department, University of California, Berkeley, 2008).

31. *Automated Solution of Differential Equations by the Finite Element Method* (eds Logg, A., Mardal, K.-A. & Wells, G.) (Springer-Verlag Berlin Heidelberg, 2012).

32. Mysovskikh, I. P. On the Convergence of Newton's Method. *Trudy Matematicheskogo Instituta imeni V.A. Steklova* **28,** 145–147 (In Russian) (1949).

33. Nachbin, L. *Topology on Spaces of Holomorphic Mappings* (Springer-Verlag Berlin Heidelberg, 1969).

34. Ortega, J. M. The Newton-Kantorovich Theorem. *The American Mathematical Monthly* **75,** 658–660 (1968).

35. Ortega, J. M. & Rheinboldt, W. *Iterative Solution of Nonlinear Equations in Several Variables* (Society for Industrial and Applied Mathematics, 2000).

36. Peters, G. & Wilkinson, J. H. Practical Problems Arising in the Solution of Polynomial Equations. *IMA Journal of Applied Mathematics* **8,** 16–35 (1971).

37. *Point Estimation of Root Finding Methods* (ed Petković, M.) (Springer-Verlag Berlin Heidelberg, 2008).

38. Poincaré, H. *Les Méthodes Nouvelles de la Méchanique Celeste* (Gauthier-Villars, Paris, 1892).

39. Rall, L. B. A Note on the Convergence of Newtons Method. *SIAM Journal on Numerical Analysis* **11,** 34–36 (1974).

40. Rheinboldt, W. C. *An adaptive continuation process for solving systems of nonlinear equations* in *Mathematical Models and Numerical Methods* **3** (Banach Center Publications, 1978), 129–142.

41. Rosenblat, S. & Davis, S. H. Bifurcation from Infinity. *SIAM Journal on Applied Mathematics* **37,** 1–19 (1979).

42. Rynne, B. & Youngson, M. *Linear Functional Analysis* (Springer-Verlag London, 2008).

43. Scott, A. *Encyclopedia of Nonlinear Science* (Routledge, 2005).

44. Seydel, R. *Practical Bifurcation and Stability Analysis* (Springer-Verlag New York, 2010).

45. Smale, S. in *The Merging of Disciplines: New Directions in Pure, Applied, and Computational Mathematics SE - 13* (eds Ewing, R., Gross, K. & Martin, C.) 185–196 (Springer-Verlag New York, 1986).

46. Sundaresan, K. Smooth Banach spaces. *Mathematische Annalen* **173,** 191–199 (1967).

47. Wigner, E. P. The Unreasonable Effectiveness of Mathematics in the Natural Sciences. Richard Courant lecture in mathematical sciences delivered at New York University, May 11, 1959. *Communications on Pure and Applied Mathematics* **13,** 1–14 (1960).

48. Wilkinson, J. H. *Rounding Errors in Algebraic Processes* (H.M.S.O. London, 1963).

49. Yamamoto, T. Historical developments in convergence analysis for Newton's and Newton-like methods. *Journal of Computational and Applied Mathematics* **124,** 1–23 (2000).