Discretizing Elementary Bifurcations

PhD Thesis

Lóczi Lajos

Thesis Advisor: Garay Barnabás

Budapest University of Technology and Economics
Department of Differential Equations

December 2006

Contents

List of the most frequent symbols

Symbol	Meaning, properties or examples
$\mathbb{N}$	The set of nonnegative integers
$\mathbb{N}^+$	The set of positive integers
$\#A$	Cardinality of the finite set A
$\lfloor\cdot\rfloor$	The floor function, that is the greatest integer function
$\lceil\cdot\rceil$	The ceiling function, that is the least integer function
$[\{a,b\}]$	For a, $b \in \mathbb{R}$, the closed interval between elements of the set $\{a,b\}$, that is $[\{a,b\}] := [\min(a,b), \max(a,b)]$
id	The identity function of $\mathbb{R}$
$g^{[k]}$	The k^{th} iterate ($k \in \mathbb{Z}$) of the function $g(\cdot)$. When k is negative, g is assumed to be invertible, so, for example, $g^{[-1]}$ denotes the inverse function of g.
smooth function	A C^k function (in all variables) with sufficiently large $k \in \mathbb{N}^+$ and the last derivative bounded
h_0, ε_0, α_0	Sufficiently small positive constants. Upper bounds on them in terms of c, p and K (see below) have been computed during the closeness estimates.
h	The discretization stepsize with $0 < h \leq h_0$
x	The space variable for the maps considered with $\lvert x\rvert \leq \varepsilon_0$
α, β, $\widetilde{\alpha}$, $\widetilde{\beta}$	Bifurcation parameters and their transformed counterparts. It is always assumed that $\lvert\alpha\rvert \leq \alpha_0$.
$x \mapsto \Phi(h,x,\alpha)$	The time-h-map of the solution flow generated by the differential equation
$x \mapsto \varphi(h,x,\alpha)$	The one-step discretization of the above map with stepsize h and of order $p \geq 1$
$\mathcal{N}_\Phi(h,x,\alpha)$	The normal form of $\Phi(h,x,\alpha)$
$\mathcal{N}_\varphi(h,x,\alpha)$	The normal form of $\varphi(h,x,\alpha)$
$x \mapsto J(h,x,\alpha)$	The conjugacy map, that is a homeomorphism satisfying a certain functional equation. It is generally not unique.
x_n, y_n, z_n	Iterates of one of the normal forms with suitable starting values. The dependence on h and α of the sequences is often suppressed. These sequences define the fundamental domains. Sometimes the names $\widetilde{x}_n$, $\widetilde{y}_n$, p_n, q_n are also used.

Symbol	Meaning, properties or examples
c	The particular positive constant in the estimate of the distance of the normal forms. It is independent of h, x and α. Its value is fixed within a Chapter.
p	The order of the discretization method, $p \in \mathbb{N}^+$
K	A positive uniform bound on the moduli of the functions $\widehat{\eta}_3$, $\widetilde{\eta}_3$ or $\widehat{\eta}_4$, $\widetilde{\eta}_4$ together with their first and second derivatives. (These functions appear in the tails of the normal forms.) The value of K is assumed to be fixed within a Chapter.
κ	The "cutting level" in the fold bifurcation
c_i $(i \in \mathbb{N})$	Generic positive constants, independent of h, x and α.
$const_i$ $(i \in \mathbb{N})$	Generic positive constants, independent of h, x and α.
$const$	Generic positive constant, independent of h, x and α. It may denote different positive numbers at different appearances.
$\omega_{\Phi,-}$, $\omega_{\Phi,0}$, $\omega_{\Phi,+}$	Certain fixed points of the maps $\Phi(h,\cdot,\alpha)$. The fixed points depend on h and α.
$\omega_{\varphi,-}$, $\omega_{\varphi,0}$, $\omega_{\varphi,+}$	The same as above, but with $\varphi(h,\cdot,\alpha)$
f_x, $\Phi_{hx\alpha}$, $\dots$	Various (mixed) partial derivatives with respect to the corresponding variables
$\mathcal{N}_\Phi^E$, J^E, $\dots$	Evaluation at general parameter values h and α, so J^E, for example, abbreviates the function $J(h,\cdot,\alpha)$
f^B, f_x^B, $\dots$	Evaluation at the bifurcation point, so f_x^B, for example, stands for $f_x(0,0)$
I_{110}, I_{031}, $\dots$	The remainders in the integral form in the multivariate Taylor series expansion of $\Phi(h,x,\alpha)$ about the bifurcation point. The three indices refer to the position of the remainder in the expansion with respect to h, x and α, respectively.
$\widetilde{\mathrm{I}}_{110}$, $\widetilde{\mathrm{I}}_{031}$, $\dots$	The analogues of I_{110}, I_{031}, but with $\varphi(h,x,\alpha)$ instead of $\Phi(h,x,\alpha)$

Chapter 1

Introduction

A fundamental problem of numerical analysis is to estimate the error between the exact solution to an ordinary differential equation and its numerical approximations. Usually, these estimates hold only on finite time intervals and are valid for solutions starting from specific initial points.

With the evolution of the abstract notion of discretization methods it became possible to view numerical methods as dynamical systems. During the last two decades this led to the emergence of a new branch of mathematics: *numerical dynamics.*

The principal aim of numerical dynamics is to compare the dynamics induced by the original equation with that induced by the discretization method. This time, however, we simultaneously consider solutions starting from all possible initial values on a long, possibly infinite time interval. In other words, we wish to compare the original phase portrait with its discretized counterpart to see what properties are preserved upon discretizations, and, vice versa, if a discretized phase portrait appears in a computer calculation with roundoff errors, how the original one might look like. For ordinary differential equations, these questions are treated in [43], while [20] deals with functional differential equations. This latter survey article contains, for example, the abstract definition and basic properties of a discretization method. General aspects of dynamical systems can be found, for example, in [1] and [38].

A discrete dynamical system is said to be structurally stable if it is equivalent to all neighbouring dynamical systems. Closeness of dynamical systems here is usually understood in the C^1-topology, and they are C^0-equivalent if there is a C^0-conjugacy between them. If two dynamical systems are conjugate, then their phase portraits are topologically the same. Often the conjugacy can only be locally defined, meaning that only a small portion of the phase portraits can be identified. This will be the case in our work, too. We remark that these conjugacies provide a topological classification and are generally not unique. It is often convenient to think of them as nonlinear coordinate transformations. In contrast to discrete dynamical systems, time reparametrization is usually required in continuous systems to obtain equivalence. General conjugacy results and the question of structural stability are described, for example, in [38].

Discretizations of a given dynamical system lie close to the original one if the stepsize is small enough, and it might be the case that the original system is equivalent to all of its nearby discretizations. In such a situation, the original system is said to be *numerically structurally stable.* For ordinary differential equations, and in the vicinity of the origin, this can be formulated as follows. Let us consider the autonomous equation $\dot{x} = f(x)$ and fix a sufficiently small number $h > 0$. Let $\Phi(h, x)$ denote the solution to this equation starting from x after time h. The function $\Phi(h, \cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is called the solution operator (or the time-h-map) of the equation. Let $\varphi(h, \cdot) : \mathbb{R}^n \to \mathbb{R}^n$ denote a stepsize-h discretization of the above map, where φ is a one-step numerical method of order p. This latter condition essentially means that φ satisfies the approximation property near the origin, that is, the inequality

$$|\Phi(h,x)-\varphi(h,x)| \leq const \cdot h^{p+1}$$

holds for any $h \in (0,h_0]$ and $|x| \leq \varepsilon_0$ with suitable $h_0 > 0$, $\varepsilon_0 > 0$ and $const > 0$. The functions f and φ are assumed to be sufficiently smooth. Iterates of $\Phi(h,\cdot)$ and $\varphi(h,\cdot)$ define two discrete time dynamical systems. The dynamical system corresponding to $\Phi(h,\cdot)$ is then numerically structurally stable, if $\Phi(h,\cdot)$ and $\varphi(h,\cdot)$ are locally conjugate, that is, if there exist neighbourhoods $0 \in U \subset \mathbb{R}^n$ and $0 \in V \subset \mathbb{R}^n$ and a function $J : [0,h_0] \times U \to V$ such that $J(h,\cdot)$ is a homeomorphism for all $h \in [0,h_0]$, and further the conjugacy equation

$$\Phi(h,J(h,x)) = J(h,\varphi(h,x))$$

is satisfied for all $x \in U$ whenever $\varphi(h,x) \in U$.

Numerical structural stability is a qualitative property of the dynamical system $\Phi(h,\cdot)$, which can be made *quantitative* if, for example, the distance between the conjugacy and the identity is estimated. In what follows, estimates for $|J(h,x)-x|$ will be called *closeness estimates.* These estimates constitute the heart of the present thesis.

The question of numerical structural stability has been addressed and solved in general [17], [30], [31] for ordinary differential equations satisfying various hyperbolicity conditions (for example, for gradient-like Morse-Smale systems or systems satisfying "Axiom A" and the strong transversality condition). Numerical counterparts of classical results on ordinary differential equations have been established—such as the numerical flow box theorem (around a non-equilibrium point) or the numerical Grobman-Hartman Lemma (around a hyperbolic equilibrium), stating that the original dynamics $\Phi(h,\cdot)$ and its discretization $\varphi(h,\cdot)$ are conjugate, moreover, $\mathcal{O}(h^p)$ closeness estimates hold.

It is thus natural to compare exact and discretized dynamics in the simplest *non-hyperbolic* cases—for example, in one-parameter families of ODEs with hyperbolicity violated at a single value of the parameter. Such points are called *bifurcation points* and we will focus on them in the thesis, so our work also fits into the framework of the theory of numerical bifurcations. The behaviour and properties of numerical methods near bifurcation points of ODEs are discussed in detail in [22], while [6], containing more than 200 references, gives a good account of various convergence questions.

The original dynamics and its discretization do not need to be conjugate near a general non-hyperbolic equilibrium, as illustrated by the explicit Euler-method near a planar center. (Indeed, consider the system $\dot{x} = y, \dot{y} = -x$, whose solution curves are concentric circles around the origin. These invariant curves are perturbed into spirals, however small the stepsize $h > 0$, implying that the two dynamics are not conjugate.) See, for example, [8] and [15].

Nevertheless, the existence of a conjugacy can be proved near certain bifurcation points. Consider a bifurcating family of ODEs of the form $\dot{x} = f(x,\alpha)$ where α denotes the bifurcation parameter. Let $\Phi(h,\cdot,\alpha)$ and $\varphi(h,\cdot,\alpha)$ denote the solution operator and its stepsize-h discretization, and assume for simplicity that the bifurcation point is the origin $x = 0$ and bifurcation takes place at $\alpha = 0$. In [7] Gyula Farkas (1972–2002) has constructed a conjugacy between the time-1-map $\Phi(1,\cdot,\alpha)$ of the ODE and the N^{th} iterate of its stepsize $h = 1/N$ discretization $\varphi^{[N]}(h,\cdot,\alpha)$ (with $N \in \mathbb{N}^+$ sufficiently large) in the vicinity of a fold bifurcation point. He also "showed" that the constructed conjugacy is $\mathcal{O}(h^p)$-close to the identity on the center manifold. Quotation marks have been used in the previous sentence because the proof of the main estimate in [7] contains some gaps in the $\alpha \leq 0$ case, and, more importantly, the symmetry argument applied in the $\alpha > 0$ case breaks down, so this case can not be considered as proven: the main technical difficulty went unnoticed and remains unresolved.

Li [32] shows again that a fold bifurcation point is numerically structurally stable by comparing the same family of maps as Gyula Farkas did. Li's proof is largely based on the earlier work of Sotomayor [40] and [41] on structural stability of generic bifurcations. However, no closeness estimates appear in [32].

There has been extensive research on how the most common invariant sets (*e.g.* equilibria or periodic orbits) are transformed by discretizations. For example, [45] examines the effect of discretizations on hyperbolic equilibria and on some bifurcation points, regarding consistency. The authors show that the Euler-method is bifurcationally consistent near fold, transcritical, pitchfork or Hopf bifurcation points: they essentially prove that the bifurcation point of the original system can not be shifted much by the numerical method (depending on its order). These results are, however, weaker than conjugacy results. Appearance of the Hopf bifurcation is considered in [26] and [37] when the Euler or Runge-Kutta methods are applied. Due to the existence of periodic solutions, time reparametrization is needed here. *Conjugacy* questions about the Hopf bifurcation are not found in the literature.

1.1 Summary of the results

The aim of the present thesis is to examine the numerical structural stability of some of the most elementary bifurcation points—the fold, transcritical and pitchfork points—qualitatively by constructing a conjugacy, and quantitatively by proving some closeness estimates.

Since the center manifold is one-dimensional in each of the three cases, the original equation $\dot{x} = f(x, \alpha)$ becomes a one-dimensional ($x \in \mathbb{R}$, $\alpha \in \mathbb{R}$) system after center manifold reduction. Hence, we assume in the following that $\Phi(h, \cdot, \alpha) : \mathbb{R} \to \mathbb{R}$ and $\varphi(h, \cdot, \alpha) : \mathbb{R} \to \mathbb{R}$ denote the corresponding maps on their center manifolds. The construction of conjugacies and closeness estimates then generates one-dimensional problems with dependence on two parameters h and α. A reasonable goal is that these estimates be uniform in the parameters.

The primary aim of the thesis is to extend the above mentioned conjugacy result on the fold bifurcation, by comparing—for any fixed $h \in [0, h_0]$—the one-parameter family of time-h-maps $\Phi(h, \cdot, \alpha)$ of the original flow with the one-parameter family $\varphi(h, \cdot, \alpha)$ of discretizations. The degeneracy at the bifurcation point is coupled with the fact that members of both families tend to the identity map as $h \to 0^+$—a map often troublesome in perturbation theory.

Our primary task can be formulated more formally as follows. Suppose that the origin is a fold, transcritical or pitchfork bifurcation point for a sufficiently smooth one-parameter family of ODEs $\dot{x} = f(x, \alpha)$. Further, suppose that the approximation property

$$|\Phi(h, x, \alpha) - \varphi(h, x, \alpha)| \leq \mathit{const} \cdot h^{p+1} \tag{1.1}$$

holds between the original and p^{th} order discretized dynamics for any $h \in (0, h_0]$, $|x| \leq \varepsilon_0$ and $|\alpha| \leq \alpha_0$ with suitable constants $h_0 > 0$, $\varepsilon_0 > 0$, $\alpha_0 > 0$ and *const* > 0.

Then, for any $h \in (0, h_0]$ and $\alpha \in [-\alpha_0, \alpha_0]$, a homeomorphism $J(h, \cdot, \alpha) : [-\varepsilon_0, \varepsilon_0] \to \mathbb{R}$ is to be constructed such that the conjugacy equation

$$J(h, \Phi(h, x, \alpha), \alpha) = \varphi(h, J(h, x, \alpha), \widetilde{\alpha}) \tag{1.2}$$

is satisfied with a suitable number $\widetilde{\alpha}$ depending on α, for any $x \in [-\varepsilon_0, \varepsilon_0]$. Aligning the bifurcation parameter α with $\widetilde{\alpha}$ in this functional equation is usually necessary, since numerical methods may shift the original bifurcation point $x = 0$. (We have proved, however, that the displacement $|\alpha - \widetilde{\alpha}|$ is always within the desired order $\mathcal{O}(h^p)$, hence it does not undermine the closeness estimates later.)

Besides proving existence, our second, equally important and challenging aim is to estimate the distance between the constructed conjugacy $J(h, \cdot, \alpha)$ and the identity. The main difficulty of the problem lies in obtaining these estimates.

1.1.1 Construction of the conjugacies

First, as usual in bifurcation theory, we perform some normal form transformations, which bring $\Phi(h,\cdot,\alpha)$ and $\varphi(h,\cdot,\alpha)$ into their canonical form. The normal form reductions are preparatory conjugacies: suitable nonlinear coordinate transformations that make computations later more concrete and transparent. In the vicinity of the bifurcation points, the two bifurcating families in question have the following normal forms. (Subscripts of f in the theorems below denote partial differentiation.)

Theorem 1 (Fold bifurcation) Suppose that the origin $(x,\alpha)=(0,0)$ is a fold bifurcation point for the family of ODEs $\dot{x}=f(x,\alpha)$ with a sufficiently smooth right-hand side, that is $f\in C^{p+6}$, $f(0,0)=0$, $f_x(0,0)=0$, $f_{xx}(0,0)\neq 0$ and $f_\alpha(0,0)\neq 0$. Then there exists a smooth, invertible coordinate and parameter transformation bringing the map

$$x\mapsto \Phi(h,x,\alpha)$$

into its canonical form

$$x\mapsto \mathcal{N}_\Phi(h,x,\alpha):=h\alpha+x+s\cdot hx^2+hx^3\cdot\eta(h,x,\alpha).$$

Similarly, there exists a smooth, invertible coordinate and parameter transformation bringing the map

$$x\mapsto \varphi(h,x,\alpha)$$

into its canonical form

$$x\mapsto \mathcal{N}_\varphi(h,x,\alpha):=h\alpha+x+s\cdot hx^2+hx^3\cdot\widetilde{\eta}(h,x,\alpha).$$

The sign $s=\pm1$ above is the same for Φ and φ, and η and $\widetilde{\eta}$ are sufficiently smooth functions. The normal form transformations for Φ and φ above are $\mathcal{O}(h^p)$-close to each other, further $|\eta-\widetilde{\eta}|=\mathcal{O}(h^p)$. ■

Theorem 2 (Transcritical bifurcation) Suppose that the origin $(x,\alpha)=(0,0)$ is a transcritical bifurcation point for the family of ODEs $\dot{x}=f(x,\alpha)$, that is $f\in C^{p+6}$, $f(0,\alpha)=0$ (for all $|\alpha|\leq\alpha_0$), $f_x(0,0)=0$, $f_{xx}(0,0)\neq 0$ and $f_{x\alpha}(0,0)\neq 0$. Suppose further that the discretization φ satisfies $\varphi(h,0,\alpha)=0$ for all $h\in(0,h_0]$ and $|\alpha|\leq\alpha_0$. Then there exists a smooth, invertible coordinate and parameter transformation bringing the map

$$x\mapsto \Phi(h,x,\alpha)$$

into its canonical form

$$x\mapsto \mathcal{N}_\Phi(h,x,\alpha):=(1+h\alpha)x+s\cdot hx^2+hx^3\cdot\eta(h,x,\alpha).$$

Similarly, there exists a smooth, invertible coordinate and parameter transformation bringing the map

$$x\mapsto \varphi(h,x,\alpha)$$

into its canonical form

$$x\mapsto \mathcal{N}_\varphi(h,x,\alpha):=(1+h\alpha)x+s\cdot hx^2+hx^3\cdot\widetilde{\eta}(h,x,\alpha).$$

The sign $s=\pm1$ above is the same for Φ and φ, and η and $\widetilde{\eta}$ are sufficiently smooth functions. The normal form transformations for Φ and φ above are $\mathcal{O}(h^p)$-close to each other, further

$|\eta - \widetilde{\eta}| = \mathcal{O}(h^p)$. ■

Theorem 3 (Pitchfork bifurcation) Suppose that the origin $(x, \alpha) = (0, 0)$ is a pitchfork bifurcation point for the family of ODEs $\dot{x} = f(x, \alpha)$, that is $f \in C^{p+7}$, $f(0, \alpha) = 0$ (for all $|\alpha| \leq \alpha_0$), $f_x(0, 0) = 0$, $f_{xx}(0, 0) = 0$, $f_{xxx}(0, 0) \neq 0$ and $f_{x\alpha}(0, 0) \neq 0$. Suppose further that the discretization φ satisfies $\varphi(h, 0, \alpha) = 0$, $\varphi_x(h, 0, 0) = 1$ and $\varphi_{xx}(h, 0, 0) = 0$ for all $h \in (0, h_0]$ and $|\alpha| \leq \alpha_0$. Then there exists a smooth, invertible coordinate and parameter transformation bringing the map

$$x \mapsto \Phi(h, x, \alpha)$$

into its canonical form

$$x \mapsto \mathcal{N}_\Phi(h, x, \alpha) := (1 + h\alpha)x + s \cdot hx^3 + hx^4 \cdot \eta(h, x, \alpha).$$

Similarly, there exists a smooth, invertible coordinate and parameter transformation bringing the map

$$x \mapsto \varphi(h, x, \alpha)$$

into its canonical form

$$x \mapsto \mathcal{N}_\varphi(h, x, \alpha) := (1 + h\alpha)x + s \cdot hx^3 + hx^4 \cdot \widetilde{\eta}(h, x, \alpha).$$

The sign $s = \pm 1$ above is the same for Φ and φ, and η and $\widetilde{\eta}$ are sufficiently smooth functions. The normal form transformations for Φ and φ above are $\mathcal{O}(h^p)$-close to each other, further $|\eta - \widetilde{\eta}| = \mathcal{O}(h^p)$. ■

Corollary 4 The approximation property (1.1) implies that the corresponding normal forms satisfy

$$|\mathcal{N}_\Phi(h, x, \alpha) - \mathcal{N}_\varphi(h, x, \alpha)| \leq c \cdot h^{p+1}|x|^\omega,$$

where $\omega = 3$ for the fold and transcritical bifurcation, and $\omega = 4$ for the pitchfork bifurcation, further, the constant $c > 0$ is independent of h, x and α. ■

The proofs are along the lines of the corresponding sections of [35], but with the discretization parameter h suitably built into its computations. Appropriate smoothness has always been assumed on the right-hand side f and the discretization, so Φ and φ can be expanded into multivariate Taylor series, further η and $\widetilde{\eta}$ in the tails of the normal forms can be differentiated twice with the last derivative bounded. (This latter property is used later.) The adjective "smooth" in the thesis hence means finite smoothness. Basic results on discretizations are found in [16] and [18]. Estimates in our theorems above are based on quantitative versions of the inverse and implicit function theorems, see, *e.g.* [47], and they pave the way for later closeness estimates.

The variable x and the bifurcation parameter α are both transformed during the normal form transformations (the discretization parameter h is left intact, however). After these normal form reductions we may assume, among others, that the bifurcation point for both families $\mathcal{N}_\Phi(h, \cdot, \alpha)$ and $\mathcal{N}_\varphi(h, \cdot, \alpha)$ is the origin.

An interesting by-product of our computations is that *Runge-Kutta methods* (see, *e.g.* [24]) of order at least 1 completely preserve the n-dimensional conditions for fold and cusp bifurcations, and also conditions for the transcritical and pitchfork bifurcations in 1 dimension—provided that the stepsize h is sufficiently small. We do not formulate these technical results more precisely here. The description of bifurcations in n-dimensions is found, for example, in [3]. As [4] points out, the higher order chain rule we use is inconsistently formulated in some basic references.

These results extend [9]–[14], where the author investigates some properties preserved by Runge-Kutta methods near bifurcation points. Our results also imply that some of the normal form transformations are unnecessary when $\varphi(h,\cdot,\alpha)$ comes from a Runge-Kutta method: for example, $\widetilde{\alpha}=\alpha$ in (1.2) is appropriate. It is also true that conditions in Theorems 2 and 3 on φ of Runge-Kutta type are automatically satisfied.

The "sufficiently small stepsize h" is an essential assumption above, because, in many cases, numerical methods are known to produce spurious solutions for "large" stepsizes, see [23] and [42] concerning Runge-Kutta methods.

Now let us comment a bit more on Theorems 1–3 above. Some books, *e.g.* [46] or [21] contain inconsistencies concerning transcritical conditions for maps. We indicate, by example, that transcritical bifurcation does not necessarily occur under these conditions.

For a pitchfork bifurcation it is often required in the literature that the right-hand side f be odd, that is $f(x,\alpha)=-f(-x,\alpha)$ should hold. Theorem 3 points out that this assumption is not essential: our theorem guarantees the existence of an *asymmetric* pitchfork bifurcation. In the proof we are confronted with the following question: if a sufficiently smooth one-parameter family of real maps $g(x,\alpha)$ is given, and the origin $x=0$ at parameter value $\alpha=0$ is a root of $g(x,\alpha)$ with multiplicity k, then what can be said about the other real roots near $(x,\alpha)=(0,0)$. The Preparation Theorems of Weierstrass and Malgrange, or the Division Theorem of Mather (see [39] or [1]) or their generalizations handle these cases when g is a complex or real analytic, C^∞ or C^k ($k\in\mathbb{N}^+$) function. These results are typically used in bifurcation analysis or in singularity theory. One of our theorems is therefore a special case of these general results: we have examined how the triple root of a certain map g of finite smoothness can be perturbed near $\alpha=0$.

Observe that in Theorems 2 and 3 some properties on the discretization φ are assumed. We illustrate by simple examples that these assumptions are necessary: $\varphi(h,\cdot,\alpha)$ does not necessarily undergo a transcritical or pitchfork bifurcation otherwise.

After these steps, the question of conjugacy between the corresponding normal forms reduces to solving the following functional equation

$$J(h,\mathcal{N}_\Phi(h,x,\alpha),\alpha)=\mathcal{N}_\varphi(h,J(h,x,\alpha),\alpha). \tag{1.3}$$

We have established the following results:

Theorem 5 Under the assumptions of Theorems 1–3 and $h\in(0,h_0]$, $|x|\le\varepsilon_0$ and $|\alpha|\le\alpha_0$ sufficiently small, equation (1.3) has a solution such that $J(h,\cdot,\alpha)$ is homeomorphism. ■

The conjugacies are constructed by using the technique of *fundamental domains*: for fixed h and α, $J(h,\cdot,\alpha)$ is prescribed on a suitable interval and extended recursively by using a rearrangement of (1.3).

The method of fundamental domains is described, for example in [35]. The classical work of [34] gives an account of functional equations in general, while functional and conjugacy equations of the theory of dynamical systems are solved in different smoothness classes in [2]. Contemporary aspects of one-dimensional non-hyperbolic dynamical systems are surveyed in [36].

Summarizing our results so far, the family $\Phi(h,\cdot,\alpha)$ is numerically structurally stable near a fold bifurcation point, but *unstable* near transcritical and pitchfork points, as far as general discretizations are concerned. If, however, the allowed φ discretizations are restricted to those in Theorems 2 and 3, numerical structural stability is recovered near these bifurcation points too.

1.1.2 The closeness estimates

After the conjugacies J have been constructed, the quantity $|\,x - J(h,x,\alpha)|$ is to be estimated near the bifurcation points. From a technical point of view, this question belongs to the quantitative theory of functional equations.

We have illustrated by simple examples that fixed points of $\Phi(h,\cdot,\alpha)$ and $\varphi(h,\cdot,\alpha)$ can be estimated from below and above by $\mathcal{O}(h^p)$. Since a conjugacy necessarily maps fixed points into fixed points, these imply that better estimates than $\mathcal{O}(h^p)$ for $|\,x - J(h,x,\alpha)|$ generally can not be expected.

Theorem 6 (Transcritical and pitchfork bifurcation) Suppose that conditions of Theorems 2 and 3 hold. Then the constructed conjugacies J satisfy *optimal* closeness estimates near *transcritical* and *pitchfork* bifurcation points. In other words, there exists a positive constant *const* > 0 such that for all $h \in (0,h_0]$, $|x| \leq \varepsilon_0$ and $|\alpha| \leq \alpha_0$

$$|\,x - J(h,x,\alpha)| \leq const \cdot h^p,$$

further, the conjugacies $J(h,\cdot,\alpha)$ *also* depend continuously on their first and third variables. ■

In the proof, using the recursive definitions of the conjugacies, we show that the closeness estimates basically depend on discrete Gronwall-type estimates. That is, expressions of the form

$$\left(h \sum_{i=0}^{n} |x_i|^{\omega} \left(\prod_{j=i}^{n} \left(\frac{d}{dx} \mathcal{N}_{\Phi} \right) (h, x_i, \alpha) \right) \right) \cdot h^p \tag{1.4}$$

need to be estimated, where $\omega = 3$ for the transcritical, and $\omega = 4$ for the pitchfork bifurcation. The sequence $x_i \equiv x_i(h,\alpha)$ is defined by the iterates of the normal form $\mathcal{N}_{\varphi}(h,\cdot,\alpha)$, with x_0 chosen suitably.

On one hand, the difficulty in estimating (1.4) is that the derivatives near the bifurcation point cluster around 1, so the contribution of the product is not easily established. On the other hand, the terms of the sequence x_i are only implicitly defined by nonlinear recursions, and so a closed form for them can hardly be expected.

At this point, the clever parametric *model functions* of Thorsten Hüls (see [27], [28] and [29]) proved to be an invaluable tool. This is a special family of nonlinear recursions with closed form solution available: if $a > 0$ and $q \in \mathbb{N}^+$ are arbitrary parameters and $z_0 > 0$ is a sufficiently small starting value, then the recursion

$$z_{n+1} := \frac{z_n}{(1 + aqz_n^q)^{1/q}}$$

has closed form solution as

$$z_n = \frac{z_0}{(1 + naqz_0^q)^{1/q}}.$$

These nice formulae combined with the symbolic and numeric power of *Mathematica* made it possible to prove estimates for the sequences x_i, which in turn led to the optimal results of Theorem 6.

Let us finally consider closeness estimates concerning the fold bifurcation. Due to symmetry, we can clearly assume in Theorem 1 that $s = 1$, corresponding to the presence of two branches of fixed points for $\alpha < 0$ in the normal forms, merging together at the bifurcation point at $\alpha = 0$ and disappearing for $\alpha > 0$.

Theorem 7 (Fold bifurcation, $\alpha \leq 0$ case) Under the assumptions of Theorem 1 and near a fold bifurcation point, optimal closeness estimates hold in the half plane *containing the fixed points.* That is, there exists a positive constant *const* > 0 such that for all $h \in (0, h_0]$, $|x| \leq \varepsilon_0$ and $\alpha \leq 0$

$$|\, x - J(h, x, \alpha)| \leq const \cdot h^p,$$

further, the conjugacies $J(h, \cdot, \alpha)$ also depend continuously on their first and third variables. ■

Theorem 8 (Fold bifurcation, $\alpha > 0$ case) Under the assumptions of Theorem 1 and near a fold bifurcation point in the *fixed-point-free* half plane, the following singular estimate holds: with a suitable positive constant *const* > 0

$$|\, x - J(h, x, \alpha)| \leq const \cdot \ln \frac{1}{\alpha} \cdot h^p$$

is satisfied for all $h \in (0, h_0]$, $|x| \leq \varepsilon_0$ and $\alpha > 0$. ■

The closeness estimates in these theorems also depend on estimating expressions like (1.4). The difficulty in the fold bifurcation case is that no good model function is known for $\alpha \neq 0$. So instead of estimating convergence speed of the iterates x_i in the $\alpha < 0$ case directly, we carry out an inductive proof involving fractional powers of the discretization parameter h, while estimates deduced from the model function at $\alpha = 0$ are refined further to obtain information on the growth of iterates in the $\alpha > 0$ case. Convergence speed of nonlinear recursions of similar type is discussed in [44], although in a less explicit form. One of the important estimates is analyzed more deeply in [5] from the viewpoint of asymptotic analysis, however, this stronger result is not directly applicable in our situation.

The absence of fixed points poses another difficulty in the $\alpha > 0$ case: iterates x_i stay in a fixed neighbourhood of the bifurcation point only for a finite time, and estimating this number of iteration steps does not seem to be easy. (These types of estimates for iteration numbers appear in [33], where a special quadratic recursion in the definition of the Mandelbrot set is analyzed. The corresponding phenomenon is called *intermittency* and is described and depicted in [25]. The "scaling law" for intermittency is heuristically proved in [21]. A simple observation in connection with these estimates is explained in [19] in more detail.)

Yet another problem in Theorem 8 is that continuity of the homeomorphism $J(h, \cdot, \alpha)$ with respect to the third variable is not apparent.

We remark that in all of the conjugacy constructions so far the natural choice $J(h, 0, \alpha)$ $:= 0$ has been adopted. The "power" of Theorem 8 is supported by the following results. First, careful numerical tests down to the level $\alpha \approx 10^{-14}$ convincingly show that the distance between the identity and J in our natural construction is bounded from *below* by $\mathcal{O}(\ln \frac{1}{\alpha} \cdot h^p)$, indicating that the closeness estimate in Theorem 8 in the limiting case $\alpha \to 0^+$ is optimal. On the other hand, we have proved that if $\omega = 3$ were replaced by any $\omega > 3$ in the distance of normal forms in Corollary 4, then $|\, x - J(h, x, \alpha)| \leq const \cdot h^p$ would hold also in the $\alpha > 0$ case.

We remark that with the original $\omega = 3$ exponent and any *fixed* $\delta > 0$, $\mathcal{O}(h^p \cdot \alpha^{-\delta})$ closeness estimates for $|J(h, x, \alpha) - x|$ are proved without much effort. It is also true that under the assumptions of Theorem 8, uniform $\mathcal{O}(h^p)$ closeness estimate holds on a shrinking parabola-shaped domain in the (α, x)-plane for $\alpha > 0$.

Using the natural construction of the conjugacies J, the closeness estimates imply the following: near the three bifurcation points the discretized orbit and the original one are *uniformly* close to each other, provided that the domain contains fixed points.

Corollary 9 There is a positive number *const* > 0 such that for all sufficiently small $h \in (0, h_0]$, $|x_0|$ and

- (for transcritical and pitchfork bifurcations) $|\alpha| \leq \alpha_0$
- (for fold bifurcation) $-\alpha_0 \leq \alpha \leq 0$,

we have that for any $n \in \mathbb{N}$

$$|\mathcal{N}_{\Phi}^{[n]}(h, x_0, \alpha) - \mathcal{N}_{\varphi}^{[n]}(h, x_0, \alpha)| \leq const \cdot h^p$$

holds, where $\mathcal{N}^{[n]}(h, x, \alpha)$ is the n^{th} iterate of the map $\mathcal{N}(h, \cdot, \alpha)$ evaluated at x. ■

A notable feature of all of our estimates is that they are fairly *explicit*, meaning that h_0, ε_0, α_0 and the *const* > 0 numbers in the closeness estimates are all expressed in terms of p, c and K, where $p \geq 1$ is the order of the discretization method, $c > 0$ is the constant in Corollary 4, and $K > 0$ is a common uniform bound on the moduli of the functions $x \mapsto \eta(h, x, \alpha)$ and $x \mapsto \widetilde{\eta}(h, x, \alpha)$ (appearing in the normal forms) together with their first and second derivatives on $[0, h_0] \times [-\varepsilon_0, \varepsilon_0] \times [-\alpha_0, \alpha_0]$. In other words, we specify many ”sufficiently small” quantities in terms of the initial data.

Albeit the choice $J(h, 0, \alpha) := 0$ in the definition of the conjugacies is natural, it is by no means necessary, so the question arises whether a ”trickier” construction in the $\alpha > 0$ case could yield better estimates.

Using the monotonicity and convexity of the normal forms in the fold bifurcation case, we have constructed two grids for $\alpha > 0$ in the (α, x) half-plane and proved the following.

Theorem 10 (Fold bifurcation, $\alpha > 0$ case) The conjugacy defined by the ”grid construction” satisfies optimal $\mathcal{O}(h^p)$ closeness estimates in the grid points and $\mathcal{O}(h)$ closeness estimates otherwise. Further, the conjugacies $J(h, x, \alpha)$ are continuous in their third variable along the grid sequence $\alpha = \alpha_N \to 0^+$. ■

It is currently an unresolved question whether the closeness estimate and the continuity property above can be improved further.

1.2 Acknowledgements

First of all, I would like to express my sincere gratitude to my thesis advisor, Barna Garay. During our work together I had many an insight into interesting topics of numerical dynamics, including bifurcation phenomena and the one-dimensional iterations that attracted me in my secondary school years.

In the autumns of 2002 and 2003 I had the opportunity to research at the University of Bielefeld. I owe much to Wolf-Jürgen Beyn and Thorsten Hüls for their hospitality and helpful discussions during my stay in Bielefeld.

I am grateful to my teachers and colleagues for their help and guidance at the Department of Applied Analysis and the Department of Numerical Analysis of the Eötvös Loránd University, and also at the Department of Differential Equations of the Budapest University of Technology and Economics. I am especially grateful to László Czách, Miklós Farkas, János Karátson, Gusztáv Reményi, László Simon and János Tóth, who definitely influenced my mathematical career and provided me with valuable advice during the past years.

I owe a great deal of gratitude also to my Family for their continuous support and patience. I thank my uncle, Tibor Csörföly for his prescient introduction to the computer program *Mathematica* in the early 1990s. Both the numeric and symbolic capabilities of this system have proved to be indispensable in establishing the three key estimates of the thesis.

Without these people the present work could not have been completed.

Chapter 2

Conjugacy in the discretized fold bifurcation

Summary. In Section 2.1 we recall some basic properties of discretizations of ordinary differential equations and the notion of a fold bifurcation point. In Section 2.2 we apply quantitative inverse and implicit function theorems and some smooth invertible coordinate and parameter changes to transform the time-h-maps $x \mapsto \Phi(h, x, \alpha)$ and their discretizations $x \mapsto \varphi(h, x, \alpha)$ into their normal forms. The normal forms turn out to be sufficiently close to each other. In Section 2.3 a conjugacy between families $\Phi(h, \cdot, \alpha)$ and $\varphi(h, \cdot, \widetilde{\alpha})$ is constructed, if $\alpha \leq 0$. The parameter shift is $|\alpha - \widetilde{\alpha}|$ is bounded by $\mathcal{O}(h^p)$. Section 2.4 uses some fractional powers of h to prove inductively an optimal closeness estimate. Section 2.5 derives the $\frac{3}{2}$-Lemma, being the main tool in the $\alpha > 0$ case, then this Lemma is applied in Section 2.6 to give a singular logarithmic closeness estimate. We also prove that if additional closeness on the normal forms is assumed, then the singularity disappears. In Section 2.7 we illustrate by careful numerical tests that the singular estimate seems to be sharp for our particular construction of the conjugacy. Finally, in Section 2.8 a modified construction of the conjugacy is given to overcome difficulties due to the lack of fixed points.

Dedicated to the memory of Gyula Farkas (1972–2002)

2.1 Introduction

Consider the ordinary differential equation

$$\dot{x} = f(x, \alpha) \tag{2.1}$$

together with its discretization

$$X_{n+1} := \varphi(h, X_n, \alpha), \qquad n = 0, 1, 2, \dots, \tag{2.2}$$

where $\alpha \in \mathbb{R}$ is a scalar bifurcation parameter, $h > 0$ is the step-size of the sufficiently smooth one-step method $\varphi : \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of order $p \geq 1$, and the function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is of class C^{p+k+1} with $k \geq 5$ and uniformly bounded derivatives.

By the definition of the order of the method, we have that

$$|\Phi(h,x,\alpha)-\varphi(h,x,\alpha)| \le const\cdot h^{p+1}, \quad \forall h\in[0,h_0], \forall|x|\le\varepsilon_0, \forall|\alpha|\le\alpha_0, \tag{2.3}$$

where $\Phi(h,\cdot,\alpha):\mathbb{R}\to\mathbb{R}$ is the time-h-map of the solution flow induced by (2.1) at parameter value α, further h_0, ε_0 and α_0 are some (small) positive constants. (Throughout the thesis, symbols *const* will denote generic positive constants in the estimates, with dependence only on f.)

Suppose that the origin $x=0$, $\alpha=0$ is an equilibrium as well as a *fold-bifurcation* point for (2.1), that is the following conditions hold

$$f^B=0, \quad f_x^B=0, \quad f_{xx}^B\ne 0, \quad f_\alpha^B\ne 0, \tag{2.4}$$

where—and throughout the thesis also—subscripts h, x (or z), and α denote partial differentiation with respect to their corresponding variables, while superscript ^{B} denotes *evaluation at the bifurcation point*: that is, evaluation at $x=0$ and $\alpha=0$. (This evaluation operator is understood, of course, to have the lowest precedence, *i.e.*, it is performed *after* taking all partial derivatives.)

Some more notation is finally introduced. The superscript ^{E} will denote *function evaluation at h and α* for functions from $\mathbb{R}^3$ to $\mathbb{R}$, that is, for example, J^E stands for the function $J(h,\cdot,\alpha)$. The range of parameters h and α will be clear from the context. The symbol $f^{[-1]}$ means the *inverse* of a real function f. Similarly, $f^{[k]}$ is the k^{th} *iterate* ($k\in\mathbb{Z}$) of $f:\mathbb{R}\to\mathbb{R}$. The symbol *id* denotes the identity function on $\mathbb{R}$. Symbols $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$, as usual, denote the *floor* and *ceiling* functions, that is the greatest integer and least integer functions, respectively (for a reference and origin of their usage, see, *e.g.*, the online encyclopedia at `http://mathworld.wolfram.com/IntegerPart.html`). The set of nonnegative integers is denoted by $\mathbb{N}$. $\#A$ will denote the number of elements of the (finite) set A. Finally, for any $a,b\in\mathbb{R}$, the symbol $[\{a,b\}]$ represents the closed *interval between* the elements of the set $\{a,b\}$, that is $[\{a,b\}]:=[\min(a,b),\max(a,b)]$.

2.2 Construction of the normal forms

In this section, we compute normal forms for the maps

$$x\mapsto\Phi(h,x,\alpha) \tag{2.5}$$

and

$$x\mapsto\varphi(h,x,\alpha) \tag{2.6}$$

near the fold-bifurcation point. Since now—as opposed to [7]—they both depend also on h, this extra parameter together with uniform estimates on $[0,h_0]$ should be built into the computations [35] we follow.

The properties of the solution flow together with (2.3) imply for $h\ge 0$, $|x|\le\varepsilon_0$ and $|\alpha|\le\alpha_0$ that

$$\Phi(h,0,0)=0, \tag{2.7}$$

$$\varphi(0,x,\alpha)=\Phi(0,x,\alpha)=x, \tag{2.8}$$

$$\Phi_h(h,x,\alpha)=f(\Phi(h,x,\alpha),\alpha), \tag{2.9}$$

$$\varphi_h(0,x,\alpha)=\Phi_h(0,x,\alpha). \tag{2.10}$$

Instead of (2.9), the shorter $\Phi_h = f \circ \Phi$ form will be used. We remark that the property $\varphi(h,0,0) = 0$ is *not* assumed here; nevertheless it often holds for discretizations, see Chapter 5.

Lemma 2.2.1 *Under the assumptions above and for $h \in [0, h_0]$, $|x| \le \varepsilon_0$, $|\alpha| \le \alpha_0$, we have that*

$$\Phi(h,x,\alpha) = f_0(h,\alpha) + f_1(h,\alpha)x + f_2(h,\alpha)x^2 + \psi_3(h,x,\alpha)x^3,$$

where

$$\begin{aligned} f_0(h,\alpha) &= \Phi^B_{h\alpha} \cdot h\alpha + h\alpha^2 \cdot \psi_0(h,\alpha), \qquad \Phi^B_{h\alpha} \neq 0, \\ f_1(h,\alpha) &\equiv 1 + g(h,\alpha) = 1 + h\alpha \cdot \psi_1(h,\alpha), \\ f_2(h,\alpha) &= \frac{1}{2}\Phi^B_{hxx} \cdot h + h\alpha \cdot \psi_2(h,\alpha), \qquad \Phi^B_{hxx} \neq 0, \\ \psi_3(h,x,\alpha) &= h \cdot \widehat{\psi_3}(h,x,\alpha) \end{aligned}$$

hold with some smooth functions ψ_0, ψ_1, ψ_2 and $\widehat{\psi_3}$.

Proof. We expand Φ in a multivariate Taylor series at the equilibrium with the remainders in the integral form, that is, the following representation is used recursively in all normal form transformations throughout the thesis:

$$\mathcal{F}(x) = \mathcal{F}(0) + \mathcal{F}'(0)\frac{x}{1!} + \ldots + \mathcal{F}^{(n)}(0)\frac{x^n}{n!} + \frac{x^{n+1}}{n!}\int_0^1 \mathcal{F}^{(n+1)}(\tau x)(1-\tau)^n \mathrm{d}\tau.$$

For f_0 we have that

$$f_0(h,\alpha) = \Phi^B + \alpha \cdot \mathrm{I}_{001}(\alpha) + h \cdot \mathrm{I}_{100}(h) + h\alpha \cdot \Phi^B_{h\alpha} +$$
$$h\alpha^2 \cdot \mathrm{I}_{102}(\alpha) + h^2\alpha \cdot \mathrm{I}_{201}(h) + h^2\alpha^2 \cdot \mathrm{I}_{202}(h,\alpha),$$

where—taking into account (2.4) and (2.7)–(2.9) repeatedly— we get that $\Phi^B = 0$, and

$$\mathrm{I}_{001}(\alpha) = \int_0^1 \Phi_\alpha(0,0,\tau\alpha)\mathrm{d}\tau \equiv 0,$$

$$\mathrm{I}_{100}(h) = \int_0^1 \Phi_h(\tau h,0,0)\mathrm{d}\tau \equiv 0.$$

Further, since

$$\Phi_{hh\alpha} = (f \circ \Phi)_{h\alpha} = ((f_x \circ \Phi) \cdot \Phi_h)_\alpha = (f_x \circ \Phi)_\alpha \cdot \Phi_h + (f_x \circ \Phi) \cdot \Phi_{h\alpha}$$

and

$$\Phi^B_{h\alpha} = (f \circ \Phi)^B_\alpha = (f_x \circ \Phi)^B \cdot \Phi^B_\alpha + (f_\alpha \circ \Phi)^B = 0 + f^B_\alpha \neq 0,$$

so we also have

$$\mathrm{I}_{201}(h) = \int_0^1 (1-\tau)\Phi_{hh\alpha}(\tau h,0,0)\mathrm{d}\tau \equiv 0,$$

and $\Phi^B_{h\alpha} \neq 0$. The explicit form of the smooth functions

$$\mathrm{I}_{102}(\alpha) = \int_0^1 (1-\tau)\Phi_{h\alpha\alpha}(0,0,\tau\alpha)\mathrm{d}\tau$$

and

$$\mathrm{I}_{202}(h,\alpha) = \int_0^1\int_0^1 (1-\tau)(1-\sigma)\Phi_{hh\alpha\alpha}(\tau h,0,\sigma\alpha)\mathrm{d}\sigma\mathrm{d}\tau$$

will not play any role in the following, hence grouping together these remaining terms into ψ_0 gives the desired expression for f_0.

As for f_1, one gets that

$$f_1(h,\alpha) = \Phi_x^B + \alpha \cdot \mathrm{I}_{011}(\alpha) + h \cdot \mathrm{I}_{110}(h) + h\alpha \cdot \mathrm{I}_{111}(h,\alpha),$$

where $\Phi_x^B = 1$,

$$\mathrm{I}_{011}(\alpha) = \int_0^1 \Phi_{x\alpha}(0,0,\tau\alpha)\mathrm{d}\tau \equiv 0,$$

$$\mathrm{I}_{110}(h) = \int_0^1 \Phi_{hx}(\tau h,0,0)\mathrm{d}\tau \equiv 0,$$

because $\Phi_{hx} = (f \circ \Phi)_x = (f_x \circ \Phi) \cdot \Phi_x$. Finally,

$$\mathrm{I}_{111}(h,\alpha) = \int_0^1 \int_0^1 \Phi_{hx\alpha}(\tau h, 0, \sigma\alpha)\mathrm{d}\sigma\mathrm{d}\tau.$$

In the case of f_2, we see that

$$f_2(h,\alpha) = \frac{1}{2}\left(\Phi_{xx}^B + \alpha \cdot \mathrm{I}_{021}(\alpha) + h \cdot \Phi_{hxx}^B + h^2 \cdot \mathrm{I}_{220}(h) + h\alpha \cdot \mathrm{I}_{121}(h,\alpha)\right),$$

where $\Phi_{xx}^B = 0$ and

$$\mathrm{I}_{021}(\alpha) = \int_0^1 \Phi_{xx\alpha}(0,0,\tau\alpha)\mathrm{d}\tau \equiv 0.$$

However,

$$\Phi_{hxx}^B = (f \circ \Phi)_{xx}^B = (f_{xx} \circ \Phi)^B \cdot \left((\Phi_x)^2\right)^B + (f_x \circ \Phi)^B \cdot \Phi_{xx}^B = f_{xx}^B \cdot 1 + 0 \neq 0.$$

Further,

$$\Phi_{hhxx} = (f_x \circ \Phi)_{xx} \cdot \Phi_h + 2(f_x \circ \Phi)_x \cdot \Phi_{hx} + (f_x \circ \Phi) \cdot \Phi_{hxx},$$

thus

$$\mathrm{I}_{220}(h) = \int_0^1 (1-\tau)\Phi_{hhxx}(\tau h, 0, 0)\mathrm{d}\tau \equiv 0.$$

Finally,

$$\mathrm{I}_{121}(h,\alpha) = \int_0^1 \int_0^1 \Phi_{hxx\alpha}(\tau h, 0, \sigma\alpha)\mathrm{d}\sigma\mathrm{d}\tau.$$

For the remainder ψ_3, the integral formula gives

$$\psi_3(h,x,\alpha) = \frac{1}{2}\int_0^1 (1-\tau)^2 \Phi_{xxx}(h,\tau x,\alpha)\mathrm{d}\tau. \tag{2.11}$$

But

$$\Phi_{xxx}(h,\tau x,\alpha) = \Phi_{xxx}(0,\tau x,\alpha) + h \cdot \int_0^1 \Phi_{hxxx}(\sigma h, \tau x, \alpha)\mathrm{d}\sigma$$

and $\Phi_{xxx}(0,\tau x,\alpha) \equiv 0$, so the lemma is proved. ■

Now let us perform a coordinate shift by introducing a new variable

$$\xi := x + \delta_0,$$

where $\delta_0 \equiv \delta_0(h,\alpha)$ will be defined soon via the implicit function theorem. This shift transforms (2.5) into $\xi \mapsto \Phi(h, \xi - \delta_0, \alpha) + \delta_0$, which—similarly, but more explicitly than in [35]—turns out to be equal to

$$\begin{aligned}\xi \mapsto \Big[f_0(h,\alpha) - g(h,\alpha)\delta_0(h,\alpha) + f_2(h,\alpha)\delta_0^2(h,\alpha) + h\cdot\delta_0^3(h,\alpha)\widehat{\psi}_{30}(h,\alpha,\delta_0)\Big] +& \\ \xi + \xi\cdot\Big[g(h,\alpha) - 2f_2(h,\alpha)\delta_0(h,\alpha) + h\cdot\delta_0^2(h,\alpha)\widehat{\psi}_{31}(h,\alpha,\delta_0)\Big] +& \qquad (2.12) \\ \xi^2\cdot\Big[f_2(h,\alpha) + h\cdot\delta_0(h,\alpha)\widehat{\psi}_{32}(h,\alpha,\delta_0)\Big] + h\cdot\widehat{\psi}_{33}(h,\xi,\alpha,\delta_0)\xi^3&\end{aligned}$$

with some smooth functions $\widehat{\psi}_{30}$, $\widehat{\psi}_{31}$, $\widehat{\psi}_{32}$, and $\widehat{\psi}_{33}$, where

$$\widehat{\psi}_{30}(h,\alpha,\delta) \equiv -\widehat{\psi}_3(h,-\delta,\alpha), \tag{2.13}$$

$$\widehat{\psi}_{31}(h,\alpha,\delta) \equiv 3\widehat{\psi}_3(h,-\delta,\alpha) - \delta\cdot\frac{\mathrm{d}}{\mathrm{d}x}\widehat{\psi}_3(h,-\delta,\alpha), \tag{2.14}$$

$$\widehat{\psi}_{32}(h,\alpha,\delta) \equiv -3\widehat{\psi}_3(h,-\delta,\alpha) + 3\delta\cdot\frac{\mathrm{d}}{\mathrm{d}x}\widehat{\psi}_3(h,-\delta,\alpha) - \frac{\delta^2}{2}\cdot\frac{\mathrm{d}^2}{\mathrm{d}x^2}\widehat{\psi}_3(h,-\delta,\alpha) \tag{2.15}$$

and

$$\widehat{\psi}_{33}(h,\xi,\alpha,\delta) \equiv \frac{1}{2}\int_0^1 (1-\tau)^2\cdot\Phi_{xxx}(h,\tau\xi-\delta,\alpha)\mathrm{d}\tau.$$

In order to annihilate the parameter-dependent linear term in (2.12), define

$$F(h,\alpha,\delta) \equiv \frac{1}{h}\left(g(h,\alpha) - 2f_2(h,\alpha)\delta + h\cdot\delta^2\cdot\widehat{\psi}_{31}(h,\alpha,\delta)\right),$$

where, in the case of $h=0$, the continuous extension of F is used. Since we have that

$$\begin{aligned} F(h,0,0) &= 0 & \forall h\in[0,h_0], \\ \frac{\partial F}{\partial\delta}(h,0,0) = \frac{-2f_2(h,0)}{h} &= -\Phi^B_{hxx} \neq 0 & \forall h\in[0,h_0], \end{aligned}$$

the implicit function theorem provides the local existence and uniqueness of a smooth function $\delta_0(h,\alpha)$, defined on $h\in[0,h_0]$ and $|\alpha|\le\alpha_0$, for which

$$F(h,\alpha,\delta_0(h,\alpha)) \equiv 0$$

holds. From uniqueness, it is seen that this δ_0 also satisfies $\delta_0(h,0)=0$ for $h\in[0,h_0]$, so

$$\delta_0(h,\alpha) = \alpha\cdot\psi_d(h,\alpha) \tag{2.16}$$

holds true for $h\in[0,h_0]$ and $|\alpha|\le\alpha_0$ with some smooth function ψ_d.

As a next step, introduce a new parameter $\mu_0 \equiv \mu_0(h,\alpha)$ by

$$\mu_0(h,\alpha) := \frac{f_0(h,\alpha)}{h} - \frac{g(h,\alpha)\delta_0(h,\alpha)}{h} + \frac{f_2(h,\alpha)\delta_0^2(h,\alpha)}{h} + \delta_0^3(h,\alpha)\widehat{\psi}_{30}(h,\alpha,\delta_0),$$

i.e., as the ξ-independent term of (2.12) divided by h. Since $\mu_0(h,0)=0$ and $\frac{\mathrm{d}}{\mathrm{d}\alpha}\mu_0(h,0) = \Phi^B_{h\alpha} \neq 0$ independently of $h\in[0,h_0]$, the inverse function theorem guarantees the local existence and uniqueness of a smooth inverse function $\overline{\alpha}_0 \equiv \overline{\alpha}_0(h,\mu)$ of $\alpha \mapsto \mu_0(h,\alpha)$. Moreover, the domain of definition of this inverse function is easily seen to contain a neighbourhood of the origin independent of $h\in[0,h_0]$. Further, $\overline{\alpha}_0(h,0)=0$, hence

$$\overline{\alpha}_0(h,\mu) = \mu\cdot\psi_a(h,\mu) \tag{2.17}$$

holds for $h \in [0, h_0]$ and $|\mu|$ small with some smooth function ψ_a.

Therefore (2.12) now reads as

$$\xi \mapsto h \cdot \mu_0 + \xi + h \cdot q(h, \mu_0) \cdot \xi^2 + h \cdot \xi^3 \cdot \widehat{\psi}_{m3}(h, \xi, \mu_0)$$

with $q(h, \mu_0) \equiv \frac{1}{2}\Phi^B_{hxx} + \widehat{\psi}_{m2}(h, \mu_0)$ and some smooth functions $\widehat{\psi}_{m2}$ and $\widehat{\psi}_{m3}$, where

$$\widehat{\psi}_{m2}(h, \mu_0) \equiv \overline{\alpha}_0 \cdot \psi_2(h, \overline{\alpha}_0) + \delta_0(h, \overline{\alpha}_0) \cdot \widehat{\psi}_{32}(h, \overline{\alpha}_0, \delta_0(h, \overline{\alpha}_0))$$

and

$$\widehat{\psi}_{m3}(h, \xi, \mu_0) \equiv \widehat{\psi}_{33}(h, \xi, \overline{\alpha}_0, \delta_0(h, \overline{\alpha}_0)).$$

A final scaling $\eta := |q(h, \mu_0)|\xi$ and $\beta := |q(h, \mu_0)|\mu_0$ with $s := \text{sign}(q(h, 0)) = \pm 1$ (being also independent of $h \in [0, h_0]$) yields the following normal form.

Lemma 2.2.2 *There are smooth invertible coordinate and parameter changes transforming the system*

$$x \mapsto \Phi(h, x, \alpha)$$

into

$$\eta \mapsto h\beta + \eta + s \cdot h\eta^2 + h\eta^3 \cdot \widehat{\eta}_3(h, \eta, \beta)$$

where $\widehat{\eta}_3(h, \eta, \beta) = \widehat{\psi}_{m3}(h, \xi, \mu_0) \cdot |q(h, \mu_0)|^{-2}$ *is a smooth function.*

Now let us consider the discretization map φ. We prove an analogous result to that of Lemma 2.2.1 first.

Lemma 2.2.3 *Under the assumptions of Lemma 2.2.1 and for* $h \in [0, h_0]$, $|x| \le \varepsilon_0$, $|\alpha| \le \alpha_0$, *we have that*

$$\varphi(h, x, \alpha) = \widetilde{f}_0(h, \alpha) + \widetilde{f}_1(h, \alpha)x + \widetilde{f}_2(h, \alpha)x^2 + \chi_3(h, x, \alpha)x^3,$$

where

$$\begin{aligned}
\widetilde{f}_0(h, \alpha) &= h^{p+1} \cdot \chi_{00}(h) + \varphi^B_{h\alpha} \cdot h\alpha + h\alpha \cdot \chi_{01}(h, \alpha), \quad \varphi^B_{h\alpha} = \Phi^B_{h\alpha} \neq 0,\\
\widetilde{f}_1(h, \alpha) &\equiv 1 + \widetilde{g}(h, \alpha) = 1 + h^{p+1} \cdot \chi_{10}(h) + h\alpha \cdot \chi_{11}(h, \alpha),\\
\widetilde{f}_2(h, \alpha) &= h^{p+1} \cdot \chi_{20}(h) + \frac{1}{2}\varphi^B_{hxx} \cdot h + h\alpha \cdot \chi_{21}(h, \alpha), \quad \varphi^B_{hxx} = \Phi^B_{hxx} \neq 0,\\
\chi_3(h, x, \alpha) &= h \cdot \widehat{\chi}_3(h, x, \alpha)
\end{aligned}$$

hold with some smooth functions χ_{00}, χ_{01}, χ_{10}, χ_{11}, χ_{20}, χ_{21} *and* $\widehat{\chi}_3$. *Moreover, for* $h \in [0, h_0]$, $|x| \le \varepsilon_0$ *and for* $|\alpha| \le \alpha_0$,

$$|\psi_3(h, x, \alpha) - \chi_3(h, x, \alpha)| \le const \cdot h^{p+1}. \tag{2.18}$$

Proof. Proceeding similarly as in Lemma 2.2.1, we get that

$$\begin{aligned}
\widetilde{f}_0(h, \alpha) = \varphi^B + \alpha \cdot \widetilde{\mathrm{I}}_{001}(\alpha) + h \cdot \widetilde{\mathrm{I}}_{100}(h) + h\alpha \cdot \varphi^B_{h\alpha} + \\
h\alpha^2 \cdot \widetilde{\mathrm{I}}_{102}(\alpha) + h^2\alpha \cdot \widetilde{\mathrm{I}}_{201}(h) + h^2\alpha^2 \cdot \widetilde{\mathrm{I}}_{202}(h, \alpha),
\end{aligned} \tag{2.19}$$

where the integrals $\widetilde{\mathrm{I}}$'s are defined just as in the proof of Lemma 2.2.1, but with φ instead of Φ. Due to (2.4)–(2.9), here we also have $\varphi^B = 0$ and $\widetilde{\mathrm{I}}_{001}(\alpha) \equiv 0$. From (2.3) at $x = 0$ we infer that for $h \in [0, h_0]$ and for $|\alpha| \le \alpha_0$

$$\left| f_0(h, \alpha) - \widetilde{f}_0(h, \alpha) \right| \le const \cdot h^{p+1}. \tag{2.20}$$

Evaluating this at $\alpha = 0$ shows that $|h \cdot \widetilde{\mathrm{I}}_{100}(h)| \leq \mathit{const} \cdot h^{p+1}$. Further, differentiating (2.10) yields that $\varphi^B_{h\alpha} = \Phi^B_{h\alpha}$.

As for $\widetilde{f}_1$, one has that $\varphi^B_x = 1$ and $\widetilde{\mathrm{I}}_{011}(\alpha) \equiv 0$, hence

$$\widetilde{f}_1(h, \alpha) = 1 + h \cdot \widetilde{\mathrm{I}}_{110}(h) + h\alpha \cdot \widetilde{\mathrm{I}}_{111}(h, \alpha).$$

Since f is at least C^{p+4}, from [18] we obtain that

$$\left| f_1(h, \alpha) - \widetilde{f}_1(h, \alpha) \right| \leq \mathit{const} \cdot h^{p+1}. \tag{2.21}$$

Evaluation at $\alpha = 0$ yields $|h \cdot \widetilde{\mathrm{I}}_{110}(h)| \leq \mathit{const} \cdot h^{p+1}$.

Considering $\widetilde{f}_2$, we obtain that $\varphi^B_{xx} = 0$ and $\widetilde{\mathrm{I}}_{021}(\alpha) \equiv 0$, thus

$$\widetilde{f}_2(h, \alpha) = \frac{1}{2} \left(h \cdot \varphi^B_{hxx} + h^2 \cdot \widetilde{\mathrm{I}}_{220}(h) + h\alpha \cdot \widetilde{\mathrm{I}}_{121}(h, \alpha) \right)$$

and again,

$$\left| f_2(h, \alpha) - \widetilde{f}_2(h, \alpha) \right| \leq \mathit{const} \cdot h^{p+1}. \tag{2.22}$$

Evaluating this at $\alpha = 0$, we see that $|h^2 \cdot \widetilde{\mathrm{I}}_{220}(h)| \leq \mathit{const} \cdot h^{p+1}$. Further, differentiating (2.10) again yields that $\varphi^B_{hxx} = \Phi^B_{hxx}$.

For the remainder χ_3, the same argument applies as in the proof of Lemma 2.2.1, together with the estimate

$$|\psi_3(h, x, \alpha) - \chi_3(h, x, \alpha)| \leq \mathit{const} \cdot h^{p+1} \cdot \frac{1}{2} \int_0^1 (1 - \tau)^2 \mathrm{d}\tau,$$

which completes the proof of the lemma. ■

Now applying the corresponding coordinate shift with $\widetilde{\delta}$ instead of δ_0, we arrive at some formulae completely analogous to (2.12)–(2.15), where $\widetilde{\delta}$ is the implicit function defined by (the continuous extension at $h = 0$ of)

$$\widetilde{F}(h, \alpha, \delta) \equiv \frac{1}{h} \left(\widetilde{g}(h, \alpha) - 2\widetilde{f}_2(h, \alpha)\delta + h \cdot \delta^2 \cdot \widetilde{\chi}_{31}(h, \alpha, \delta) \right).$$

However, for the $\mathcal{O}(h^p)$-estimates, we will need a *quantitative* (or parametrized) version of the *implicit function theorem*, see [47]. Instead of its full form (*i.e.* Banach space setting with more parameter-dependence), we cite that result in a simplified form tailored to our needs and using our notation.

Lemma 2.2.4 *Let $\widetilde{F} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ be a C^j mapping. Assume there exist a function $\delta_0 : \mathbb{R}^2 \to \mathbb{R}$ and some constants $\kappa_1 > 0$, $\kappa_2 > 0$ such that for $|\delta - \delta_0(h, \alpha)| \leq r_1$ and $|h|, |\alpha| < r_2$ we have*

$$\left| \frac{\partial \widetilde{F}}{\partial \delta}(h, \alpha, \delta) - \frac{\partial \widetilde{F}}{\partial \delta}(h, \alpha, \delta_0(h, \alpha)) \right| \leq \kappa_2 < \kappa_1 \leq \left| \frac{\partial \widetilde{F}}{\partial \delta}(h, \alpha, \delta_0(h, \alpha)) \right|,$$

$$\left| \widetilde{F}(h, \alpha, \delta_0(h, \alpha)) \right| \leq (\kappa_1 - \kappa_2) \cdot r_1.$$

Then for any $|h|, |\alpha| < r_2$, $\widetilde{F}(h, \alpha, \cdot)$ has a unique C^j-smooth zero $\widetilde{\delta} \equiv \widetilde{\delta}(h, \alpha)$ near $\delta_0(h, \alpha)$, and the following estimate holds

$$\left| \widetilde{\delta}(h, \alpha) - \delta_0(h, \alpha) \right| \leq (\kappa_1 - \kappa_2)^{-1} \cdot |\widetilde{F}(h, \alpha, \delta_0(h, \alpha))|.$$

In order to verify the conditions of this lemma, define $\kappa_1 := \frac{1}{2}|\varphi^B_{hxx}|$ and $\kappa_2 := \frac{1}{2}\kappa_1$. The estimate

$$\left|\frac{\partial \widetilde{F}}{\partial \delta}(h,\alpha,\delta_0(h,\alpha))\right| =$$

$$\left|\frac{-2\widetilde{f}_2(h,\alpha)}{h} + 2\delta_0(h,\alpha)\cdot\widehat{\chi}_{31}(h,\alpha,\delta_0) + \delta_0^2(h,\alpha)\cdot\frac{d}{d\delta}\widehat{\chi}_{31}(h,\alpha,\delta_0)\right| \geq \kappa_1$$

is seen to be valid—due to the form of $\widetilde{f}_2$ and (2.16)—provided that r_2 is small. On the other hand,

$$\left|\frac{\partial \widetilde{F}}{\partial \delta}(h,\alpha,\delta) - \frac{\partial \widetilde{F}}{\partial \delta}(h,\alpha,\delta_0(h,\alpha))\right| \leq$$

$$|2[\delta-\delta_0(h,\alpha)]\widehat{\chi}_{31}(h,\alpha,\delta) + 2\delta_0(h,\alpha)\left[\widehat{\chi}_{31}(h,\alpha,\delta) - \widehat{\chi}_{31}(h,\alpha,\delta_0)\right]| +$$

$$\left|\left[\delta^2 - \delta_0^2(h,\alpha)\right]\frac{\mathrm{d}}{\mathrm{d}\delta}\widehat{\chi}_{31}(h,\alpha,\delta) + \delta_0^2(h,\alpha)\left[\frac{\mathrm{d}}{\mathrm{d}\delta}\widehat{\chi}_{31}(h,\alpha,\delta) - \frac{\mathrm{d}}{\mathrm{d}\delta}\widehat{\chi}_{31}(h,\alpha,\delta_0)\right]\right| \leq \kappa_2,$$

if r_1 and r_2 are sufficiently small. Finally,

$$\left|\widetilde{F}(h,\alpha,\delta_0(h,\alpha))\right| \leq |F(h,\alpha,\delta_0(h,\alpha))| + \left|\widetilde{F}(h,\alpha,\delta_0(h,\alpha)) - F(h,\alpha,\delta_0(h,\alpha))\right| \leq$$

$$0 + \frac{1}{h}\,|\widetilde{g}(h,\alpha) - g(h,\alpha)| + \frac{2|\delta_0(h,\alpha)|}{h}\left|\widetilde{f}_2(h,\alpha) - f_2(h,\alpha)\right| +$$

$$\delta_0^2(h,\alpha)\left|\widehat{\chi}_{31}(h,\alpha,\delta_0) - \widehat{\psi}_{31}(h,\alpha,\delta_0)\right| \leq \mathit{const}\cdot h^p,$$

owing to (2.21), (2.22) and the estimate

$$\begin{gathered}\left|\widehat{\chi}_{31}(h,\alpha,\delta_0) - \widehat{\psi}_{31}(h,\alpha,\delta_0)\right| \leq \\ 3\left|\widehat{\chi}_3(h,-\delta_0,\alpha) - \widehat{\psi}_3(h,-\delta_0,\alpha)\right| + \\ |\delta_0(h,\alpha)|\left|\frac{1}{2h}\int_0^1 (1-\tau)^2\tau\cdot\left(\varphi_{xxxx}(h,-\tau\delta_0,\alpha) - \Phi_{xxxx}(h,-\tau\delta_0,\alpha)\right)\mathrm{d}\tau\right| \leq \\ \mathit{const}\cdot h^p + \mathit{const}\cdot h^p,\end{gathered} \tag{2.23}$$

being valid due to (2.14), (2.11), (2.18) and the fact (see [18] again) that f is at least C^{p+5}.

Therefore, Lemma 2.2.4 proves the local existence and uniqueness of the function $\widetilde{\delta}$ such that

$$\widetilde{F}(h,\alpha,\widetilde{\delta}(h,\alpha)) \equiv 0,$$

and

$$\left|\widetilde{\delta}(h,\alpha) - \delta_0(h,\alpha)\right| \leq \mathit{const}\cdot h^p \tag{2.24}$$

holds for $h\in[0,h_0]$ and $|\alpha|\leq\alpha_0$.

Now let us define a new parameter $\widetilde{\mu}$ in an analogous way as we did before, *i.e.* as the ξ-independent term divided by h, that is

$$\widetilde{\mu}(h,\alpha) := \frac{\widetilde{f}_0(h,\alpha)}{h} - \frac{\widetilde{g}(h,\alpha)\widetilde{\delta}(h,\alpha)}{h} + \frac{\widetilde{f}_2(h,\alpha)\widetilde{\delta}^2(h,\alpha)}{h} + \widetilde{\delta}^3(h,\alpha)\widehat{\chi}_{30}(h,\alpha,\widetilde{\delta}).$$

We see from the analogous expression of (2.13) for $\widehat{\chi}_{30}$, from (2.18), (2.20)–(2.22) and (2.24) that

$$|\widetilde{\mu}(h,\alpha) - \mu_0(h,\alpha)| \leq const \cdot h^p \tag{2.25}$$

holds for $h \in [0, h_0]$ and $|\alpha| \leq \alpha_0$. In order to use a *quantitative inverse function theorem* for $\alpha \mapsto \widetilde{\mu}(h,\alpha)$, we apply Lemma 2.2.4 again, but this time with G instead of $\widetilde{F}$, and $\overline{\alpha}_0$ instead of δ_0, where

$$G(h,\mu,\alpha) := \mu - \widetilde{\mu}(h,\alpha).$$

To check the conditions of the lemma, define $\kappa_1 := \frac{1}{2}|\varphi^B_{h\alpha}|$ and $\kappa_2 := \frac{1}{2}\kappa_1$. We have that

$$\left|\frac{\partial G}{\partial \alpha}(h,\mu,\alpha)\right| =$$

$$\left|\varphi^B_{h\alpha} + \chi_{01}(h,\alpha) + \alpha\frac{\mathrm{d}}{\mathrm{d}\alpha}\chi_{01}(h,\alpha) - \frac{\widetilde{g}(h,\alpha)}{h}\frac{\mathrm{d}}{\mathrm{d}\alpha}\widetilde{\delta}(h,\alpha) + \widetilde{\delta}(h,\alpha)\widetilde{\chi}_G(h,\alpha)\right|$$

holds with a suitable smooth function $\widetilde{\chi}_G$. By (2.24), (2.16) and (2.17), the expression $|\widetilde{\delta}(h, \overline{\alpha}_0(h,\mu))|$ can be made arbitrary small provided that $|h|, |\mu| < r_2$ are small enough. The same is true for $|\frac{1}{h}\widetilde{g}(h, \overline{\alpha}_0(h,\mu))|$. Moreover, the definition (2.19) of χ_{01} shows that $\chi_{01}(0,0) = 0$, so from these we can conclude that

$$\left|\frac{\partial G}{\partial \alpha}(h,\mu,\overline{\alpha}_0(h,\mu))\right| \geq \kappa_1,$$

provided that r_2 is sufficiently small. The other condition

$$\left|\frac{\partial G}{\partial \alpha}(h,\mu,\alpha) - \frac{\partial G}{\partial \alpha}(h,\mu,\overline{\alpha}_0(h,\mu))\right| \leq \kappa_2$$

is seen to hold by continuity if $|\alpha - \overline{\alpha}_0(h,\mu)| \leq r_1$ and r_2 are small enough. Finally, by (2.25),

$$|G(h,\mu,\overline{\alpha}_0(h,\mu))| = |\mu_0(h,\overline{\alpha}_0(h,\mu)) - \widetilde{\mu}(h,\overline{\alpha}_0(h,\mu))| \leq const \cdot h^p.$$

Therefore, we get a unique zero $\widetilde{\alpha}(h,\mu)$ of $G(h,\mu,\cdot)$, which—by the construction of G—is just the inverse function of $\alpha \mapsto \widetilde{\mu}(h,\alpha)$. Furthermore,

$$|\widetilde{\alpha}(h,\mu) - \overline{\alpha}_0(h,\mu)| \leq const \cdot h^p \tag{2.26}$$

holds for $h \in [0, h_0]$ and $|\mu|$ sufficiently small.

As a conclusion, (2.6) becomes

$$\widetilde{\xi} \mapsto h \cdot \widetilde{\mu} + \widetilde{\xi} + h \cdot \widetilde{q}(h,\widetilde{\mu}) \cdot \widetilde{\xi}^2 + h \cdot \widetilde{\xi}^3 \cdot \widehat{\chi}_{m3}(h,\widetilde{\xi},\widetilde{\mu})$$

with $\widetilde{q}(h,\widetilde{\mu}) \equiv \frac{1}{2}\varphi^B_{hxx} + \widehat{\chi}_{m2}(h,\widetilde{\mu})$ and some smooth functions $\widehat{\chi}_{m2}$ and $\widehat{\chi}_{m3}$. We claim that

$$|\widetilde{q}(h,\widetilde{\mu}) - q(h,\mu_0)| \leq const \cdot h^p$$

also holds. Indeed, since

$$|\widetilde{q}(h,\widetilde{\mu}(h,\alpha)) - q(h,\mu_0(h,\alpha))| \leq \left|\widetilde{f}_2(h,\alpha) - f_2(h,\alpha)\right| +$$

$$h\left|\widetilde{\delta}(h,\alpha) \cdot \widehat{\chi}_{32}(h,\alpha,\widetilde{\delta}(h,\alpha)) - \delta_0(h,\alpha) \cdot \widehat{\psi}_{32}(h,\alpha,\delta_0(h,\alpha))\right|,$$

we deduce the desired estimate from the $\mathcal{O}(h^p)$-estimates obtained so far together with some standard (triangle) inequalities, and an estimate similar to (2.23) but with (2.15) and using that

f is at least C^{p+6}.

By applying a final scaling

$$\widetilde{\eta} := |\widetilde{q}(h,\widetilde{\mu})|\widetilde{\xi} \quad \text{and} \quad \widetilde{\beta} := |\widetilde{q}(h,\widetilde{\mu})|\widetilde{\mu}$$

with $s := \operatorname{sign}(\widetilde{q}(h,0)) = \pm 1$ (being independent of $h \in [0,h_0]$), further taking into account the fact that $|\xi - \widetilde{\xi}|$, $|\eta - \widetilde{\eta}|$ and $|\beta - \widetilde{\beta}|$ are all $\mathcal{O}(h^p)$-small, we have derived the following normal form together with the desired closeness estimates.

Theorem 2.2.5 *There are smooth invertible coordinate and parameter changes transforming the system*

$$x \mapsto \varphi(h,x,\alpha)$$

into

$$\widetilde{\eta} \mapsto h\widetilde{\beta} + \widetilde{\eta} + s \cdot h\widetilde{\eta}^2 + h\widetilde{\eta}^3 \cdot \widetilde{\eta}_3(h,\widetilde{\eta},\widetilde{\beta})$$

where $\widetilde{\eta}_3$ is a smooth function.

Moreover, the smooth invertible coordinate and parameter changes above and those in Lemma 2.2.2 are $\mathcal{O}(h^p)$-close to each other, further

$$|\widehat{\eta}_3 - \widetilde{\eta}_3| \leq const \cdot h^p.$$

Finally, we apply a parameter shift $\widetilde{\beta} \mapsto \beta$ to the normal form in the theorem above, being $\mathcal{O}(h^p)$-close to the identity, since $|\beta - \widetilde{\beta}| = \mathcal{O}(h^p)$. So from now on we will use the bifurcation parameter α again instead of β and $\widetilde{\beta}$. To simplify our notation further, instead of variables η and $\widetilde{\eta}$ the letter x will be used.

2.3 Construction of the conjugacy in the $\alpha \leq 0$ case

We have just shown that there exists a constant $c > 0$ such that

$$|\mathcal{N}_\Phi(h,x,\alpha) - \mathcal{N}_\varphi(h,x,\alpha)| \leq c \cdot h^{p+1}|x|^3 \tag{2.27}$$

holds for all sufficiently small $h > 0$, $|x| \geq 0$ and $|\alpha| \geq 0$. Throughout the Chapter, c will denote this particular positive constant.

Our task is to construct a homeomorphism $J(h,\cdot,\alpha)$ in a small neighbourhood of the origin with $h \in (0,h_0]$ and $\alpha \in [-\alpha_0,\alpha_0]$ as parameters such that $J(h,\cdot,\alpha)$ solves the conjugacy equation

$$\mathcal{N}_\Phi(h,J(h,x,\alpha),\alpha) = J(h,\mathcal{N}_\varphi(h,x,\alpha),\alpha). \tag{2.28}$$

We consider first the case $\alpha < 0$. Let us denote the negative fixed point of $\mathcal{N}_\varphi^E$ and $\mathcal{N}_\Phi^E$ near the origin by $\omega_{\varphi,-} \equiv \omega_{\varphi,-}(h,\alpha)$ and $\omega_{\Phi,-} \equiv \omega_{\Phi,-}(h,\alpha)$, respectively.

Lemma 2.3.1 *For every $0 < h \leq h_0$ and $-\alpha_0 \leq \alpha < 0$ we have that*

$$-\sqrt{2}\sqrt{|\alpha|} \leq \omega_{\varphi,-} \leq -\sqrt{\frac{2}{3}}\sqrt{|\alpha|}$$

and

$$-\sqrt{2}\sqrt{|\alpha|} \leq \omega_{\Phi,-} \leq -\sqrt{\frac{2}{3}}\sqrt{|\alpha|},$$

provided that $\alpha_0 \leq \frac{1}{8K^2}$.

Proof. By definition, $\omega_{\varphi,-} < 0$ solves $\alpha + x^2 + x^3 \cdot \widetilde{\eta}_3(h, x, \alpha) = 0$. Since if $|x| \leq \frac{1}{2K}$, then $|x^3\widetilde{\eta}_3| \leq \frac{1}{2}x^2$ and hence

$$\alpha + \frac{x^2}{2} \leq \alpha + x^2 + x^3 \cdot \widetilde{\eta}_3(h, x, \alpha) \leq \alpha + \frac{3x^2}{2}$$

holds, we get the desired estimates provided that $\sqrt{2}\sqrt{|\alpha|} \leq \frac{1}{2K}$, which is true if $|\alpha| \leq \frac{1}{8K^2}$. The proof for $\omega_{\Phi,-}$ is similar. ■

By iterating one of the normal forms, let us define two sequences x_k and y_k. Let $x_k \equiv x_k(h, \alpha)$ be defined as

$$x_{k+1} := \mathcal{N}_\varphi(h, x_k, \alpha), \quad k = 0, 1, 2, \ldots$$

with $x_0 := 0$, further let $y_k \equiv y_k(h, \alpha)$ be defined as

$$y_{k+1} := \mathcal{N}_\varphi(h, y_k, \alpha), \quad k = 0, 1, 2, \ldots \tag{2.29}$$

with $y_0 < \omega_{\varphi,-}$, being independent of both h and α, and $|y_0|$ being chosen appropriately, see below. Note that y_0 is a negative number.

Since, by Lemma 2.3.1, if h and $|\alpha|$ are sufficiently small, $0 < (\mathcal{N}_\varphi^E)'(\omega_{\varphi,-}) < 1$ holds, the fixed point $\omega_{\varphi,-}$ is attracting, hence $\lim_{k\to\infty} x_k(h, \alpha) = \lim_{k\to\infty} y_k(h, \alpha) = \omega_{\varphi,-}$. Moreover, a simple calculation shows that $y_0 < y_1(h, \alpha)$ can also be achieved, for example, $-\frac{1}{4K} \leq y_0 \leq -2\sqrt{\alpha_0}$ suffices, hence it follows by induction that the sequence y_k is monotone increasing. Similarly, it can be assumed that the sequence x_k is monotone decreasing.

We remark that suitable values of h_0, α_0 and y_0 have been built into the conditions of the following lemmas and theorems corresponding to the $\alpha \leq 0$ case. (There is only one constraint which has not been taken into account explicitly: if the domain of definition of the functions $\widehat{\eta}_3$ and $\widetilde{\eta}_3$ is smaller than $(0, h_0] \times [-\varepsilon_0, \varepsilon_0] \times [-\alpha_0, \alpha_0]$ given later, then h_0, ε_0 or α_0 should suitably be decreased further.)

The following figure shows the branch of stable and unstable fixed points of $\mathcal{N}_\varphi^E$ in the (α, x)-plane together with the first few terms of the inner sequence $x_k(h, \alpha)$ and the outer sequence $y_k(h, \alpha)$ with some $h > 0$ and $\alpha < 0$ fixed. The arrows point toward terms of the sequences with larger k indices.

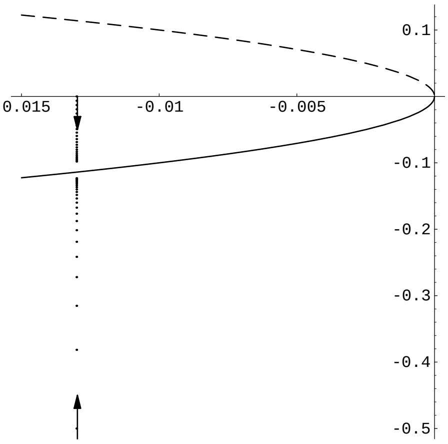


Figure 2.3.1

The intervals $[x_{k+1}, x_k]$ and $[y_k, y_{k+1}]$ ($k \in \mathbb{N}$) constitute the so-called *fundamental domains* on which the homeomorphism J^E is now piecewise defined.

Fix $0 < h \le h_0$ and $-\alpha_0 \le \alpha < 0$ arbitrarily.

Let $J^E(x) := x$ for $x \in [x_1, x_0] \equiv [h\alpha, 0]$. For $n > 1$, set

$$J^E(x_n) := \left(\mathcal{N}_\Phi^E\right)^{[n]}(x_0),$$

and recursively, for $n > 1$ and for $x \in (x_n, x_{n-1})$, let

$$J^E(x) := \left(\mathcal{N}_\Phi^E \circ J^E \circ \left(\mathcal{N}_\varphi^E\right)^{[-1]}\right)(x). \tag{2.30}$$

Here the right hand side has already been defined by the recursion. Finally, set

$$J^E(\omega_{\varphi,-}) := \omega_{\Phi,-}.$$

Then J^E is continuous, strictly monotone increasing on $[\omega_{\varphi,-}, 0]$, as it is a composition of three such functions, and satisfies (2.28).

Fix $-\frac{1}{4K} \le y_0 \le -2\sqrt{\alpha_0}$ as well. Let $J^E(y_0) := y_0$, and for $n > 1$, set

$$J^E(y_n) := \left(\mathcal{N}_\Phi^E\right)^{[n]}(y_0).$$

On the interval $[y_0, y_1]$, extend J^E linearly. Recursively, for $n > 1$ and for $y \in (y_{n-1}, y_n)$, set

$$J^E(y) := \left(\mathcal{N}_\Phi^E \circ J^E \circ \left(\mathcal{N}_\varphi^E\right)^{[-1]}\right)(y).$$

Then J^E is continuous, strictly monotone increasing on $[y_0, \omega_{\varphi,-}]$ and satisfies (2.28).

The same construction is carried out for $\alpha = 0$. This time, however, only the sequence y_k is needed, since the two fixed points merge then disappear as α passes through 0^-. Of course, $J(h, 0, 0) := 0$.

Currently, the construction is halfway ready—the function J has been defined on $(0, h_0] \times [-|y_0|, 0] \times [-\alpha_0, 0]$ so far.

On $(0, h_0] \times [0, |y_0|] \times [-\alpha_0, 0]$, that is in the region of *repelling* fixed points, the *inverses* of the normal forms are iterated. For any $0 < h \le h_0$ and $-\alpha_0 \le \alpha < 0$, set $\widetilde{x}_0 := x_0 = 0$ and for $k = 1, 2, \ldots$

$$\widetilde{x}_k \equiv \widetilde{x}_k(h, \alpha) := \left(\mathcal{N}_\varphi^E\right)^{[-k]}(\widetilde{x}_0),$$

further for any $0 < h \le h_0$ and $-\alpha_0 \le \alpha \le 0$, let $\widetilde{y}_0 := |y_0|$ and for $k = 1, 2, \ldots$

$$\widetilde{y}_k \equiv \widetilde{y}_k(h, \alpha) := \left(\mathcal{N}_\varphi^E\right)^{[-k]}(\widetilde{y}_0).$$

Then for $\alpha < 0$, the monotone increasing sequence $\widetilde{x}_k$ tends to $\omega_{\varphi,+}$, while the monotone decreasing sequence $\widetilde{y}_k$ converges to $\omega_{\Phi,+}$, where $\omega_{\varphi,+}$ and $\omega_{\Phi,+}$ denote the *positive* fixed points of $\mathcal{N}_\varphi^E$ and $\mathcal{N}_\Phi^E$, respectively.

The construction for J^E is analogous: for example, we set

$$J^E(\widetilde{x}_n) := \left(\mathcal{N}_\Phi^E\right)^{[-n]}(\widetilde{x}_0) \quad \text{and} \quad J^E(\widetilde{y}_n) := \left(\mathcal{N}_\Phi^E\right)^{[-n]}(\widetilde{y}_0),$$

but now the relation $J^E = \left(\mathcal{N}_\Phi^E\right)^{[-1]} \circ J^E \circ \mathcal{N}_\varphi^E$ is used in the recursive extensions.

Remark 2.3.1 Notice that our construction is more direct than the one in [7], since the intermediate pure quadratic function $g(x, \alpha)$ as well as the two auxiliary homeomorphisms H and G in [7] are eliminated.

2.4 The closeness estimate in the $\alpha \le 0$ case

2.4.1 The inner region

We now prove that the constructed conjugacy J^E is $\mathcal{O}(h^p)$-close to the identity on the interval $[\omega_{\varphi,-}, 0]$ uniformly for any $h \in (0, h_0]$ and $\alpha \in [-\alpha_0, 0)$.

We mention that $\mathcal{O}(h^{p-1})$-closeness could be proved easily by arguing as [7] with estimates formulated in terms of the sequence x_k itself. Nevertheless, it turns out that restoring this lost order is possible by a different subdivision of $[\omega_{\varphi,-}, 0]$, established by the following preparatory lemma. The sequence s_n defined below successfully bridges the gap between two different orders of magnitude: it connects the "micro" level $\mathcal{O}(h\alpha)$ with the "meso" level $\mathcal{O}(\sqrt{|\alpha|})$. The meso-level and the "macro" level $\mathcal{O}(1)$ will be connected by the sequence y_k in Section 2.4.2.

Lemma 2.4.1 *Let $h_0 \le \frac{1}{16}$ and $\sqrt{\alpha_0} \le \min\left(\frac{1}{2}, \frac{1}{\sqrt{8K}}\right)$. For every $h \in (0, h_0]$ and $\alpha \in [-\alpha_0, 0)$, define*

$$m \equiv m(h) := \lfloor \log_2 \log_2 \frac{1}{h} \rfloor$$

and for $1 \le n \le m$

$$s_n \equiv s_n(h, \alpha) := -\sqrt[2^n]{h}\sqrt{|\alpha|},$$

further let $s_0 := h\alpha \equiv x_1$. Then $m \ge 2$,

$$\omega_{\varphi,-} < -\frac{\sqrt{|\alpha|}}{2} \le s_m \le -\frac{\sqrt{|\alpha|}}{4}$$

and for any $1 \le n < m$ we have that

$$\omega_{\varphi,-} < \mathcal{N}_\varphi^E(s_{n+1}) < s_{n+1} < \mathcal{N}_\varphi^E(s_n) < s_n < \ldots < \mathcal{N}_\varphi^E(s_1) < s_1 < \mathcal{N}_\varphi^E(s_0) < s_0 < 0.$$

Proof. It is seen that $\frac{1}{2} \ge h^{2^{-m}} \ge \frac{1}{4}$ is equivalent to $0 \le -m + \log_2 \log_2 \frac{1}{h} \le 1$, which is always satisfied due to the definition of $m(h)$. $\sqrt{\alpha_0}$ has been chosen so small that Lemma 2.3.1 can be applied, hence $\omega_{\varphi,-} \le -\sqrt{\frac{2}{3}}\sqrt{|\alpha|} < -\frac{1}{2}\sqrt{|\alpha|}$. Also notice that (due to the definition of $\omega_{\varphi,-}$, $\mathcal{N}_\varphi^E(0) < 0$ and continuity) $\omega_{\varphi,-} < \mathcal{N}_\varphi^E(x) < x$ holds for $x \in (\omega_{\varphi,-}, 0]$. It is easy to see that $h\alpha + x \le \mathcal{N}_\varphi^E(x) < 0$, if $|x| \le \frac{1}{K}$. The already shown inequality $-\frac{\sqrt{|\alpha|}}{2} \le s_n$ $(1 \le n \le m)$ and condition $\sqrt{|\alpha|} \le \frac{1}{\sqrt{8K}}$ imply $|s_n| \le \frac{1}{K}$, hence it is sufficient to prove that $s_{n+1} < h\alpha + s_n$ holds for $1 \le n < m$. (The case $n = 0$ can be verified directly.) But this is equivalent to $h^{2^{-n-1}} + h \cdot h^{-(2^{-n-1})}\sqrt{|\alpha|} < 1$. The second term is strictly less than $\frac{1}{2}$, hence $h^{2^{-n-1}} \le \frac{1}{2}$ remains to be shown. However, this reduces to $\log_2 \log_2 \frac{1}{h} \ge n+1$, which is true, since $m > n$. ■

Now the desired closeness is shown to hold on each of the subintervals defined above. However, since the number of these subintervals tends to infinity as $h \to 0^+$, the constants on the right hand sides of the estimates should be controlled carefully. Thus, instead of a generic positive constant *const*, the symbol $c > 0$ being the same as in (2.27) with *fixed value* is used throughout the proof.

Lemma 2.4.2 *Suppose that $h_0 \le \min\left(\frac{1}{16}, \sqrt[p]{\frac{1}{8c}}\right)$ and $\sqrt{\alpha_0} \le \min\left(\frac{1}{2}, \frac{1}{19K}\right)$. Then using the notation of the previous lemma, for every $h \in (0, h_0]$, $\alpha \in [-\alpha_0, 0)$ and $0 \le n < m$ we have the following estimates:*

$$\sup_{[s_0, 0]} |\, id - J^E| = 0, \tag{2.31}$$

$$\sup_{[\mathcal{N}_\varphi^E(s_0),s_0]} |\,id - J^E| \leq c \cdot h^{p+4}|\alpha|^3, \tag{2.32}$$

$$\sup_{[\mathcal{N}_\varphi^E(s_{n+1}),\mathcal{N}_\varphi^E(s_n)]} |\,id - J^E| \leq c \cdot h^{p+2^{-n-1}}\sqrt{|\alpha|}, \tag{2.33}$$

$$\sup_{[\omega_{\varphi,-},\mathcal{N}_\varphi^E(s_m)]} |\,id - J^E| \leq 12c \cdot h^p\sqrt{|\alpha|}. \tag{2.34}$$

Proof. Step 1. h_0 and α_0 have been chosen such that $\max(|\,\omega_{\varphi,-}|, |\,\omega_{\Phi,-}|) \leq \min\left(1, \frac{1}{13K}\right)$ (see Lemma 2.3.1), which implies that $0 < (\mathcal{N}_\Phi^E)' \leq 1 + h \cdot id$ and $(\mathcal{N}_\Phi^E)'$ is monotone increasing (due to $0 < (\mathcal{N}_\Phi^E)''$) on $[\omega_{\varphi,-}, 0] \cup [\omega_{\Phi,-}, 0]$. So the above estimates can be evaluated at any $x \in [\omega_{\varphi,-}, 0]$ and therefore at any $J^E(x) \in [\omega_{\Phi,-}, 0]$. (We remind that, by construction, J^E maps the interval $[\omega_{\varphi,-}, 0]$ onto $[\omega_{\Phi,-}, 0]$.) Hence for any $x \in [\omega_{\varphi,-}, 0]$

$$\sup_{[\{x,J^E(x)\}]} (\mathcal{N}_\Phi^E)' \leq 1 + h \cdot \max(x, J^E(x)) \tag{2.35}$$

holds. (With the $[\{\cdot,\cdot\}]$ notation, both cases $x \leq J^E(x)$ and $J^E(x) < x$ can be treated simultaneously.) Taking also into account that J^E is strictly monotone increasing by construction, further inequality (2.27) and definition (2.30), we get for any $\omega_{\varphi,-} \leq a < b \leq 0$ that

$$\sup_{[\mathcal{N}_\varphi^E(a),\mathcal{N}_\varphi^E(b)]} |\,id - J^E| = \sup_{[\mathcal{N}_\varphi^E(a),\mathcal{N}_\varphi^E(b)]} \left|\mathcal{N}_\varphi^E \circ (\mathcal{N}_\varphi^E)^{[-1]} - \mathcal{N}_\Phi^E \circ J^E \circ (\mathcal{N}_\varphi^E)^{[-1]}\right| \leq$$

$$\sup_{[a,b]} |\,\mathcal{N}_\varphi^E - \mathcal{N}_\Phi^E| + \sup_{[a,b]} |\,\mathcal{N}_\Phi^E - \mathcal{N}_\Phi^E \circ J^E| \leq$$

$$c \cdot h^{p+1}|a|^3 + \sup_{x\in[a,b]}\left(\left(\sup_{[\{x,J^E(x)\}]} (\mathcal{N}_\Phi^E)'\right)|x - J^E(x)|\right) \leq$$

$$c \cdot h^{p+1}|a|^3 + \left(1 + h \cdot \max\left(b, J^E(b)\right)\right)\sup_{[a,b]} |\,id - J^E|. \tag{2.36}$$

Step 2. $\sup_{[s_0,0]} |\,id - J^E| = \sup_{[x_1,x_0]} |\,id - J^E| = 0$, since $J^E = id$ on $[x_1, x_0]$ by construction.
Step 3. By (2.36) and the previous step,

$$\sup_{[\mathcal{N}_\varphi^E(s_0),s_0]} |\,id - J^E| = \sup_{[\mathcal{N}_\varphi^E(x_1),\mathcal{N}_\varphi^E(x_0)]} |\,id - J^E| \leq$$

$$c \cdot h^{p+1}|x_1|^3 + \left(1 + h \cdot \max\left(x_0, J^E(x_0)\right)\right)\sup_{[x_1,x_0]} |\,id - J^E| = c \cdot h^{p+4}|\alpha|^3.$$

Step 4. By (2.36), we have that

$$\sup_{[\mathcal{N}_\varphi^E(s_1),\mathcal{N}_\varphi^E(s_0)]} |\,id - J^E| \leq c \cdot h^{p+1}|s_1|^3 + \left(1 + h \cdot \max\left(s_0, J^E(s_0)\right)\right)\sup_{[s_1,s_0]} |\,id - J^E|. \tag{2.37}$$

Here $J^E(s_0) = J^E(x_1) = -h|\alpha| = s_0$, $|s_1|^3 = h^{\frac{3}{2}}|\alpha|^{\frac{3}{2}}$ and narrowing the interval in the supremum on the left hand side yields that

$$\sup_{[s_1,\mathcal{N}_\varphi^E(s_0)]} |\,id - J^E| \leq$$

$$c \cdot h^{p+1+\frac{3}{2}}|\alpha|^{\frac{3}{2}} + \left(1 - h^2|\alpha|\right)\max\left(\sup_{[s_1,\mathcal{N}_\varphi^E(s_0)]} |\,id - J^E|, \sup_{[\mathcal{N}_\varphi^E(s_0),s_0]} |\,id - J^E|\right).$$

If the maximum is attained on the second term, the estimate from Step 3 is used (together with $h \le 1$ and $\sqrt{|\alpha|} \le \frac{1}{2}$), while if the maximum is attained on the first term, the resulting inequality is solved. In any case, we can establish that

$$\sup_{[s_1, \mathcal{N}^E_\varphi(s_0)]} |\, id - J^E| \le c \cdot h^{p+\frac{1}{2}} \sqrt{|\alpha|}, \tag{2.38}$$

with the same c as before. Now, turning to (2.37) again, but this time also using Step 3 and (2.38), we get that

$$\sup_{[\mathcal{N}^E_\varphi(s_1), \mathcal{N}^E_\varphi(s_0)]} |\, id - J^E| \le$$
$$c \cdot h^{p+2+\frac{1}{2}} |\alpha|^{\frac{3}{2}} + \left(1 - h^2|\alpha|\right) \max\left(c \cdot h^{p+\frac{1}{2}} \sqrt{|\alpha|}, c \cdot h^{p+4}|\alpha|^3\right).$$

Again, it is easy to see that in any case the right hand side can not be greater than $c \cdot h^{p+\frac{1}{2}} \sqrt{|\alpha|}$.

Step 5. Repeating inductively, we get for $1 \le n < m$ that

$$\sup_{[\mathcal{N}^E_\varphi(s_n), \mathcal{N}^E_\varphi(s_{n-1})]} |\, id - J^E| \le c \cdot h^{p+2^{-n}} \sqrt{|\alpha|}.$$

By (2.36),

$$\sup_{[\mathcal{N}^E_\varphi(s_{n+1}), \mathcal{N}^E_\varphi(s_n)]} |\, id - J^E| \le$$
$$c \cdot h^{p+1}|s_{n+1}|^3 + \left(1 + h \cdot \max\left(s_n, J^E(s_n)\right)\right) \sup_{[s_{n+1}, s_n]} |\, id - J^E|. \tag{2.39}$$

Here $|s_{n+1}|^3 = h^{3 \cdot 2^{-n-1}} |\alpha|^{\frac{3}{2}}$. Further, since $s_n \in [\mathcal{N}^E_\varphi(s_n), \mathcal{N}^E_\varphi(s_{n-1})]$, by the induction hypothesis we have that

$$J^E(s_n) - s_n \le |J^E(s_n) - s_n| \le c \cdot h^{p+2^{-n}} \sqrt{|\alpha|},$$

from which it is easy to deduce that

$$J^E(s_n) \le -\frac{\sqrt{|\alpha|}}{2} h^{2^{-n}}$$

using that $h^p \le \frac{1}{8c} \le \frac{1}{2c}$ by assumption. Obviously, $s_n \le -\frac{\sqrt{|\alpha|}}{2} h^{2^{-n}}$ holds as well. So (2.39) yields that

$$\sup_{[\mathcal{N}_\varphi(s_{n+1}), \mathcal{N}^E_\varphi(s_n)]} |\, id - J^E| \le c \cdot h^{p+1+3\cdot 2^{-n-1}} |\alpha|^{\frac{3}{2}} +$$
$$\left(1 - \frac{\sqrt{|\alpha|}}{2} h^{1+2^{-n}}\right) \max\left(\sup_{[s_{n+1}, \mathcal{N}^E_\varphi(s_n)]} |\, id - J^E|, \sup_{[\mathcal{N}^E_\varphi(s_n), s_n]} |\, id - J^E|\right). \tag{2.40}$$

Clearly, the supremum on the left hand side is not increased if it is taken only on $[s_{n+1}, \mathcal{N}^E_\varphi(s_n)]$. Evaluating the first case in the maximum we have that

$$\sup_{[s_{n+1}, \mathcal{N}^E_\varphi(s_n)]} |\, id - J^E| \le \frac{c \cdot h^{p+1+3\cdot 2^{-n-1}} |\alpha|^{\frac{3}{2}}}{\frac{\sqrt{|\alpha|}}{2} h^{1+2^{-n}}} = 2c \cdot h^{p+2^{-n-1}} |\alpha| \le c \cdot h^{p+2^{-n-1}} \sqrt{|\alpha|},$$

since $\sqrt{|\alpha|} \le \frac{1}{2}$, and similarly, evaluating the second case in the maximum on the right hand side of (2.40) (and using the induction hypothesis also) yields the same, since for $1 \le n < m$

$$c \cdot h^{p+1+3\cdot 2^{-n-1}} |\alpha|^{\frac{3}{2}} + c \cdot h^{p+2^{-n}} \sqrt{|\alpha|} \le c \cdot h^{p+2^{-n-1}} \sqrt{|\alpha|}. \tag{2.41}$$

Therefore, we have shown that

$$\sup_{[s_{n+1},\mathcal{N}_\varphi^E(s_n)]} |\, id - J^E| \leq c \cdot h^{p+2^{-n-1}} \sqrt{|\alpha|}.$$

Now with this additional information substituted back into the right hand side of (2.40) together with the induction hypothesis, we see as in (2.41) that

$$\sup_{[\mathcal{N}_\varphi^E(s_{n+1}),\mathcal{N}_\varphi^E(s_n)]} |\, id - J^E| \leq c \cdot h^{p+2^{-n-1}} \sqrt{|\alpha|}.$$

The induction is complete.

Step 6. Finally, by using (2.36) we get that

$$\sup_{[\omega_{\varphi,-},\mathcal{N}_\varphi^E(s_m)]} |\, id - J^E| \leq c \cdot h^{p+1}|\omega_{\varphi,-}|^3 + \left(1 + h \cdot \max\left(s_m, J^E\left(s_m\right)\right)\right) \sup_{[\omega_{\varphi,-},s_m]} |\, id - J^E|,$$

since $\omega_{\varphi,-}$ is a fixed point of $\mathcal{N}_\varphi^E$. Now we use inequality $|\,\omega_{\varphi,-}| \leq \sqrt{2}\sqrt{|\alpha|}$ from Lemma 2.3.1, further inequality $\mathcal{N}_\varphi^E(s_m) \leq s_m \leq -\frac{\sqrt{|\alpha|}}{4}$ from Lemma 2.4.1 and (2.33) with $n = m-1$ together with the assumption $h^p \leq \frac{1}{8c}$ to obtain $J^E\left(s_m\right) \leq -\frac{\sqrt{|\alpha|}}{4} + c \cdot h^p \sqrt{|\alpha|} \leq -\frac{\sqrt{|\alpha|}}{8}$ and $s_m \leq -\frac{\sqrt{|\alpha|}}{8}$, finally the decomposition $[\omega_{\varphi,-}, s_m] = [\omega_{\varphi,-}, \mathcal{N}_\varphi^E(s_m)] \cup [\mathcal{N}_\varphi^E(s_m), s_m]$ in the supremum on the right hand side to point out in the first case that

$$\sup_{[\omega_{\varphi,-},\mathcal{N}_\varphi^E(s_m)]} |\, id - J^E| \leq \frac{\sqrt{8}c \cdot h^{p+1}|\alpha|^{\frac{3}{2}}}{h\frac{\sqrt{|\alpha|}}{8}} = 16\sqrt{2}c \cdot h^p|\alpha| \leq 8\sqrt{2}c \cdot h^p\sqrt{|\alpha|},$$

while in the second case—using (2.33) again—that

$$\sup_{[\omega_{\varphi,-},\mathcal{N}_\varphi^E(s_m)]} |\, id - J^E| \leq 2c \cdot h^p \sqrt{|\alpha|}.$$

Now the proof of the lemma is complete. ∎

Remark 2.4.1 At $\omega_{\varphi,-}$, we can obtain a slightly better estimate in terms of α. Namely, we have

$$|\, id - J^E|(\omega_{\varphi,-}) = |\,\omega_{\varphi,-} - \omega_{\Phi,-}| \leq c \cdot h^{p+1}|\omega_{\varphi,-}|^3 + \left(\sup_{[\{\,\omega_{\varphi,-},\omega_{\Phi,-}\}]} (\mathcal{N}_\Phi^E)'\right) |\,\omega_{\varphi,-} - \omega_{\Phi,-}|,$$

which—since the positive supremum is at most $1 - \frac{h}{2}\sqrt{|\alpha|}$ together with Lemma 2.3.1—implies that

$$|\,\omega_{\varphi,-} - \omega_{\Phi,-}| \leq 2c \cdot h^p \frac{|\omega_{\varphi,-}|^3}{\sqrt{|\alpha|}} \leq 4\sqrt{2}c \cdot h^p|\alpha|.$$

Remark 2.4.2 on optimality. The following explicit example illustrates that the distance of fixed points of functions satisfying (2.27) may be bounded from *below* by $\mathcal{O}(h^p)$ $(h \to 0)$. Hence the fact that fixed points must be mapped into nearby fixed points by the conjugacy J^E implies that better estimates than $\mathcal{O}(h^p)$ of $|\, id - J^E|$ generally can not be expected.

Indeed, set $\widehat{\eta}_3(h, x, \alpha) := 0$ and $\widetilde{\eta}_3(h, x, \alpha) := h^p \cdot x$. Then $\mathcal{N}_\Phi(h, x, \alpha) = h\alpha + x + hx^2$ and $\mathcal{N}_\varphi(h, x, \alpha) = h\alpha + x + hx^2 + h^{p+1}x^4$ satisfy (2.27) in a neighbourhood of the origin, further, $\omega_{\Phi,-} = -\sqrt{|\alpha|}$ and $\omega_{\varphi,-} = -\sqrt{\frac{1}{2h^p}\left(\sqrt{1+4h^p|\alpha|} - 1\right)}$. Using inequality $1 + \frac{t}{2} - \frac{t^2}{8} \leq \sqrt{1+t} \leq 1 + \frac{t}{2} - \frac{t^2}{8} + \frac{t^3}{16}$ for $0 \leq t \leq 1$, one can show that

$$|\,\omega_{\varphi,-} - \omega_{\Phi,-}| \geq h^p \left(\frac{1}{2}|\alpha|^{\frac{3}{2}} - h^p|\alpha|^{\frac{5}{2}}\right)$$

holds, if, for example, $h \leq 1$ and $\sqrt{|\alpha|} \leq \frac{1}{2}$.

2.4.2 The outer region

In the following lemma—also interesting in itself—we first estimate the growth of iterates of the normal form $\mathcal{N}_\varphi^E$, *i.e.* the convergence speed of $y_k(h,\alpha)$.

Lemma 2.4.3 *Suppose that the positive numbers h_0 and α_0 are small enough, further $|y_0|$ has been chosen appropriately. Then for $-\alpha_0 \le \alpha \le 0$, $0 < h \le h_0$ and $k \ge 0$ we have that*

$$-|y_0| \le y_k(h,\alpha) \le 0,$$

while for $-\alpha_0 \le \alpha \le 0$, $0 < h \le h_0$ and $k \ge \lfloor \frac{1}{h} \rfloor + 1$ we have that

$$-\sqrt{2|\alpha|} - \frac{2}{kh} \le y_k(h,\alpha) < 0. \tag{2.42}$$

Proof. The first estimate follows from the fact that the sequence y_k is monotone increasing, as we have seen (for example, the condition $-\frac{1}{4K} \le y_0 \le -2\sqrt{\alpha_0}$ guarantees this).

As for the second estimate, we prove by induction on k. For $k = \lfloor \frac{1}{h} \rfloor + 1$, $kh \le 2$ holds if h is small enough, hence $-\sqrt{2|\alpha|} - \frac{2}{kh} \le -1 \le y_0 \le y_k < 0$, if $|y_0|$ is small enough.

For $k > \lfloor \frac{1}{h} \rfloor + 1$, we have that $y_{k+1} = \mathcal{N}_\varphi(h, y_k, \alpha) \ge h\alpha + y_k + \frac{h}{2}y_k^2$, if $|y_k|$ is small enough (for example, if $|y_k| \le |y_0| \le \frac{1}{2K}$). Hence it is sufficient to prove

$$h\alpha + y_k + \frac{h}{2}y_k^2 \ge -\sqrt{2|\alpha|} - \frac{2}{(k+1)h}. \tag{2.43}$$

To this end, notice that the function $x \mapsto h\alpha + x + \frac{h}{2}x^2$ is monotone increasing provided that $x > -\frac{1}{h}$. It is easy to see that $-\sqrt{2|\alpha|} - \frac{2}{kh} > -\frac{1}{h}$, if $k > \lfloor \frac{1}{h} \rfloor + 1$, further h and $|\alpha|$ are small enough. Then, by the induction hypothesis, we see that

$$h\alpha + \left(-\sqrt{2|\alpha|} - \frac{2}{kh}\right) + \frac{h}{2}\left(-\sqrt{2|\alpha|} - \frac{2}{kh}\right)^2 \ge -\sqrt{2|\alpha|} - \frac{2}{(k+1)h} \tag{2.44}$$

implies (2.43). However, since now $|\alpha| = -\alpha$, (2.44) is equivalent to $hk\sqrt{2|\alpha|} \ge -\frac{1}{k+1}$. Therefore, the induction is complete. ■

Remark 2.4.3 The precise conditions for h_0, α_0 and $|y_0|$ are collected in the next lemma.

Remark 2.4.4 Estimate (2.42) has been devised by superimposing the following pieces of information: on one hand, the convergence speed of the sequence $y_k(h,0) = \mathcal{O}(\frac{1}{hk})$ $(k \to \infty)$ can be inferred from [28], while, on the other hand, we know from Lemma 2.3.1 that $\lim_{k\to\infty} y_k(h,\alpha) = \mathcal{O}(\sqrt{|\alpha|})$, if $-\alpha_0 \le \alpha < 0$ is small.

Our estimate of y_k is simpler and more explicit than the corresponding one in [7], in which a majorizing sequence z_k containing a fractional power of k is used. We will only need fractional powers in the finer analyses in Sections 2.5 and 2.6 for $\alpha > 0$.

Now it is proved that the conjugacy J^E is $\mathcal{O}(h^p)$-close to the identity on the interval $[-|y_0|, \omega_{\varphi,-})$ for any $h \in (0, h_0]$ and $\alpha \in [-\alpha_0, 0)$, as well as on the interval $[-|y_0|, 0]$ for any $h \in (0, h_0]$ when $\alpha = 0$.

Lemma 2.4.4 *Suppose that $h_0 \le \frac{1}{5}$, $\sqrt{\alpha_0} \le \min\left(\frac{1}{2}, \frac{1}{26K}\right)$ and $\max\left(-1, -\frac{1}{13K}\right) \le y_0 \le -2\sqrt{\alpha_0}$. Then for each $h \in (0, h_0]$ and $\alpha \in [-\alpha_0, 0)$ we have that*

$$\sup_{[y_0, \omega_{\varphi,-})} |id - J^E| \le c\left(3y_0^2 + 4\sqrt{\alpha_0} + \frac{4}{1-h_0} + 12\right)h^p, \tag{2.45}$$

and similarly, for $h \in (0, h_0]$ and $\alpha = 0$ the estimate

$$\sup_{[y_0,0]} |\, id - J^E| \leq c \left(3y_0^2 + \frac{4}{1-h_0} \right) h^p \tag{2.46}$$

holds.

Proof. Step 1. The assumptions have been set up such that Lemma 2.3.1 and Lemma 2.4.3 are both applicable (hence $\omega_{\varphi,-} < -\frac{\sqrt{|\alpha|}}{2}$ holds for example, when $\alpha < 0$), further $0 < (\mathcal{N}_\Phi^E)' \leq 1 + h \cdot id$ and $(\mathcal{N}_\Phi^E)'$ is monotone increasing on $[-|y_0|, 0]$.

Step 2a. Consider the case $\alpha \in [-\alpha_0, 0)$ first. It is clear that

$$\sup_{[y_0,\omega_{\varphi,-})} |\, id - J^E| = \sup_{n\in\mathbb{N}} \sup_{[y_n,y_{n+1}]} |\, id - J^E|. \tag{2.47}$$

Step 2b. Now since $J^E(y_0) = y_0$ and J^E is linear on $[y_0, y_1]$, we get that

$$\sup_{[y_0,y_1]} |\, id - J^E| = |\, y_1 - J^E(y_1)| = |\mathcal{N}_\varphi^E(y_0) - \mathcal{N}_\Phi^E(y_0)| \leq c \cdot h^{p+1} y_0^2,$$

by a weaker form of (2.27).

Step 2c. For $n \geq 1$, similarly to (2.36), we obtain that

$$\begin{gathered}
\sup_{[y_n,y_{n+1}]} |\, id - J^E| \leq \sup_{[y_n,y_{n+1}]} \left| \mathcal{N}_\varphi^E \circ (\mathcal{N}_\varphi^E)^{[-1]} - \mathcal{N}_\Phi^E \circ (\mathcal{N}_\varphi^E)^{[-1]} \right| + \\
\sup_{[y_n,y_{n+1}]} \left| \mathcal{N}_\Phi^E \circ (\mathcal{N}_\varphi^E)^{[-1]} - \mathcal{N}_\Phi^E \circ J^E \circ (\mathcal{N}_\varphi^E)^{[-1]} \right| = \\
\sup_{[y_{n-1},y_n]} \left| \mathcal{N}_\varphi^E - \mathcal{N}_\Phi^E \right| + \sup_{[y_{n-1},y_n]} \left| \mathcal{N}_\Phi^E - \mathcal{N}_\Phi^E \circ J^E \right| \leq \\
\sup_{[y_{n-1},y_n]} \left| \mathcal{N}_\varphi^E - \mathcal{N}_\Phi^E \right| + \sup_{y\in[y_{n-1},y_n]} \left(\left(\sup_{[\{y,J^E(y)\}]} (\mathcal{N}_\Phi^E)' \right) |\, y - J^E(y)| \right) \leq \\
c \cdot h^{p+1} y_{n-1}^2 + \left(1 - \frac{h}{2}\sqrt{|\alpha|} \right) \sup_{[y_{n-1},y_n]} |\, id - J^E|,
\end{gathered}$$

using the fact that for $y \in [y_0, \omega_{\varphi,-}]$ the inclusion $[\{y, J^E(y)\}] \subset [y_0, \omega_{\varphi,-}] \cup [y_0, \omega_{\Phi,-}]$ holds, further, $\sup_{[y_0,\max(\omega_{\varphi,-},\omega_{\Phi,-})]} (\mathcal{N}_\Phi^E)' \leq 1 + h \cdot \max(\omega_{\varphi,-}, \omega_{\Phi,-}) \leq 1 - \frac{h}{2}\sqrt{|\alpha|}$ by Step 1.

Step 2d. Repeating inductively, for any $n \geq 1$ we have that

$$\begin{gathered}
\sup_{[y_n,y_{n+1}]} |\, id - J^E| \leq \\
1 \cdot \sup_{[y_0,y_1]} |\, id - J^E| + c \cdot h^{p+1} \sum_{i=0}^{n-1} \left(1 - \frac{h}{2}\sqrt{|\alpha|} \right)^{n-1-i} y_i^2
\end{gathered}$$

with c being the same constant as in (2.27). Hence in order to show (2.45), it is sufficient to verify—by virtue of Step 2a and 2b—that

$$\sup_{h\in(0,h_0]} \sup_{\alpha\in[-\alpha_0,0)} \sup_{k\in\mathbb{N}} \left(h \sum_{i=0}^{k} \left(1 - \frac{h}{2}\sqrt{|\alpha|} \right)^{k-i} y_i^2(h,\alpha) \right) \leq const \tag{2.48}$$

holds with a suitable $const > 0$.

Step 2e. We first estimate (2.48) for $0 \leq k \leq \lfloor \frac{1}{h} \rfloor$, using the first estimate of Lemma 2.4.3.

$$h\sum_{i=0}^{k}\left(1-\frac{h}{2}\sqrt{|\alpha|}\right)^{k-i} y_i^2 \leq h\sum_{i=0}^{k} y_i^2 \leq h\sum_{i=0}^{\lfloor \frac{1}{h}\rfloor} y_0^2 \leq$$

$$h\left(\frac{1}{h}+1\right) y_0^2 \leq 2y_0^2.$$

Step 2f. We can now estimate (2.48) for $k \geq \lfloor \frac{1}{h} \rfloor + 1$ by making use of the second estimate of Lemma 2.4.3 and Step 2e.

$$h\left(\sum_{i=0}^{\lfloor \frac{1}{h}\rfloor} + \sum_{i=\lfloor \frac{1}{h}\rfloor+1}^{k}\right)\left(1-\frac{h}{2}\sqrt{|\alpha|}\right)^{k-i} y_i^2 \leq$$

$$2y_0^2 + h\sum_{i=\lfloor \frac{1}{h}\rfloor+1}^{k}\left(1-\frac{h}{2}\sqrt{|\alpha|}\right)^{k-i}\left(2|\alpha| + \frac{4\sqrt{2|\alpha|}}{ih} + \frac{4}{i^2h^2}\right) \leq \dots$$

Now we use $ih \geq \left(\lfloor \frac{1}{h}\rfloor + 1\right)h > \frac{1}{h}\cdot h = 1$, and $\sum_{i=\lfloor \frac{1}{h}\rfloor+1}^{k}\left(1-\frac{h}{2}\sqrt{|\alpha|}\right)^{k-i} \leq \frac{1}{1-\left(1-\frac{h}{2}\sqrt{|\alpha|}\right)}$ to proceed.

$$\dots \leq 2y_0^2 + h\frac{1}{1-\left(1-\frac{h}{2}\sqrt{|\alpha|}\right)}\left(2|\alpha| + 4\sqrt{2|\alpha|}\right) + h\sum_{i=\lfloor \frac{1}{h}\rfloor+1}^{k} 1^{k-i}\frac{4}{i^2h^2} \leq$$

$$2y_0^2 + 2\left(2\sqrt{|\alpha|} + 4\sqrt{2}\right) + \frac{4}{h}\int_{\lfloor \frac{1}{h}\rfloor}^{\infty}\frac{1}{i^2}\mathrm{d}i \leq$$

$$2y_0^2 + 4(\sqrt{\alpha_0} + 2\sqrt{2}) + \frac{4}{h}\frac{1}{\frac{1}{h}-1} \leq$$

$$2y_0^2 + 4\sqrt{\alpha_0} + \frac{4}{1-h_0} + 12,$$

which is a suitable choice for *const* in (2.48). This estimate, substituted back into (2.47), yields (2.45). The proof of the lemma in the case $\alpha \in [-\alpha_0, 0)$ is complete.

Step 3. Consider now the case $\alpha = 0$. Then the estimate $0 < (\mathcal{N}_\Phi^E)' \leq 1$ can be used on $[y_0, 0]$. As in Step 2d, we arrive at the following condition

$$h\sum_{i=0}^{k} y_i^2 \leq const \tag{2.49}$$

to be proved, uniformly in h and α for all values of $k \in \mathbb{N}$. But (2.49) can be proved along the lines of Step 2e and 2f, being now much simpler, thanks to the last two estimates of Lemma 2.4.3 at $\alpha = 0$. ■

2.4.3 Conclusion and further remarks

Taking into account Lemma 2.4.2 and Lemma 2.4.4, we have thus proved the following theorem.

Theorem 2.4.5 *Suppose that* $h_0 \leq \min\left(\frac{1}{16}, \sqrt[p]{\frac{1}{8c}}\right)$ *and* $\sqrt{\alpha_0} \leq \min\left(\frac{1}{2}, \frac{1}{26K}\right)$*, further* $\max\left(-1, -\frac{1}{13K}\right) \leq y_0 \leq -2\sqrt{\alpha_0}$*. Then, for every* $h \in (0, h_0]$ *and* $\alpha \in [-\alpha_0, 0]$*, the conjugacy defined in Section 2.3 satisfies*

$$\sup_{[y_0,0]} |\, id - J^E| \leq 22c \cdot h^p,$$

where $c > 0$ *is the same as in (2.27).*

Now a similar task has to be carried out to acquire the appropriate estimates on $[0, |y_0|]$ as well, for any $h \in (0, h_0]$ and $-\alpha_0 \leq \alpha \leq 0$. These proofs are however a bit more technical due to the ubiquitous *inverses* of the normal forms. We only illustrate how *some* of the estimates can be derived in this case by showing two fragments of the proof. The case $-\alpha_0 \leq \alpha < 0$ is considered now and attention is focused only near the boundary of the interval $[0, \omega_{\varphi,+}]$.

1. We begin proving the counterpart of Lemma 2.4.2. Let us formulate two basic inequalities first.

$$|\, id - J^E|(x) = \left|\left(\mathcal{N}_\Phi^E\right)^{[-1]} \circ \mathcal{N}_\Phi^E - \left(\mathcal{N}_\Phi^E\right)^{[-1]} \circ J^E \circ \mathcal{N}_\varphi^E\right|(x) \leq$$
$$\left(\sup_{[\{\mathcal{N}_\Phi^E(x), J^E \circ \mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]}\right)'\right) \left(\left|\mathcal{N}_\Phi^E - \mathcal{N}_\varphi^E\right|(x) + \left|\mathcal{N}_\varphi^E - J^E \circ \mathcal{N}_\varphi^E\right|(x)\right) \leq$$
$$\left(\sup_{[\{\mathcal{N}_\Phi^E(x), J^E \circ \mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]}\right)'\right) \left(c \cdot h^{p+1}|x|^3 + |\, id - J^E| \left(\mathcal{N}_\varphi^E(x)\right)\right). \tag{2.50}$$

On the other hand, by using that $\left(\left(\mathcal{N}_\Phi^E\right)^{[-1]}\right)' \leq 2$ is valid on a small neighbourhood of the origin, inequality

$$\left|\left(\mathcal{N}_\Phi^E\right)^{[-1]}(x) - \left(\mathcal{N}_\varphi^E\right)^{[-1]}(x)\right| = \left|\left(\mathcal{N}_\Phi^E\right)^{[-1]}(x) - \left(\mathcal{N}_\Phi^E\right)^{[-1]} \circ \mathcal{N}_\Phi^E \circ \left(\mathcal{N}_\varphi^E\right)^{[-1]}(x)\right| \leq$$
$$\left(\sup_{[\{x, \mathcal{N}_\Phi^E \circ \left(\mathcal{N}_\varphi^E\right)^{[-1]}(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]}\right)'\right) \left|\mathcal{N}_\varphi^E \circ \left(\mathcal{N}_\varphi^E\right)^{[-1]}(x) - \mathcal{N}_\Phi^E \circ \left(\mathcal{N}_\varphi^E\right)^{[-1]}(x)\right| \leq$$
$$2c \cdot h^{p+1} \left|\left(\mathcal{N}_\varphi^E\right)^{[-1]}(x)\right|^3 \tag{2.51}$$

is also at our disposal.

Now using (2.50) and the definitions $\widetilde{x}_1 \equiv \left(\mathcal{N}_\varphi^E\right)^{[-1]}(0)$, further $J^E(0) = 0$, we establish that

$$\sup_{[0,\widetilde{x}_1]} |\, id - J^E| \leq \left(\sup_{[\{\mathcal{N}_\Phi^E(\widetilde{x}_1), 0\}]} \left((\mathcal{N}_\Phi^E)^{[-1]}\right)'\right) \cdot c \cdot h^{p+1}|\widetilde{x}_1|^3 \leq 2c \cdot h^{p+1}|\widetilde{x}_1|^3.$$

Now it is verified that $\widetilde{x}_1$ has the correct order of magnitude in terms of h and α. To this end, use the fact that, say, $\frac{1}{2} \leq \left(\left(\mathcal{N}_\varphi^E\right)^{[-1]}\right)' \leq 2$ holds on a small neighbourhood of the origin to get

$$\frac{1}{2}h|\alpha| \leq \left(\inf_{[h\alpha,0]} \left((\mathcal{N}_\varphi^E)^{[-1]}\right)'\right) |h\alpha| = \left(\inf_{[\{0,\mathcal{N}_\varphi^E(0)\}]} \left((\mathcal{N}_\varphi^E)^{[-1]}\right)'\right) \left|0 - \mathcal{N}_\varphi^E(0)\right| \leq$$
$$\left|\left(\mathcal{N}_\varphi^E\right)^{[-1]}(0) - \left(\mathcal{N}_\varphi^E\right)^{[-1]}\left(\mathcal{N}_\varphi^E(0)\right)\right| \equiv |\widetilde{x}_1| \leq$$
$$\left(\sup_{[\{0,\mathcal{N}_\varphi^E(0)\}]} \left((\mathcal{N}_\varphi^E)^{[-1]}\right)'\right) \left|0 - \mathcal{N}_\varphi^E(0)\right| = \left(\sup_{[h\alpha,0]} \left((\mathcal{N}_\varphi^E)^{[-1]}\right)'\right) |h\alpha| \leq 2h|\alpha|.$$

Thus, $\sup_{[0,\widetilde{x}_1]} |\, id - J^E| \leq 16c \cdot h^{p+4} |\alpha|^3$. (However, with a little analysis, similar to (2.52) below, one can show that $\sup_{[h\alpha,0]} \left((\mathcal{N}_\varphi^E)^{[-1]}\right)' = 1$, hence estimate $\sup_{[0,\widetilde{x}_1]} |\, id - J^E| \leq 2c \cdot h^{p+4}|\alpha|^3$ is closer to the truth.) The rest of the proof can be carried over similarly.

2. Secondly, it is shown that the repelling fixed points are sufficiently close to each other, *i.e.* the conjugacy J^E is $\mathcal{O}(h^p)$-close to the identity also at $\omega_{\varphi,+}$. By (2.50) at $x = \omega_{\varphi,+}$ we have that

$$|\, \omega_{\varphi,+} - \omega_{\Phi,+}| = |\, id - J^E|(\omega_{\varphi,+}) \leq$$

$$\left(\sup_{[\{\mathcal{N}_\Phi^E(\omega_{\varphi,+}), \omega_{\Phi,+}\}]} \left((\mathcal{N}_\Phi^E)^{[-1]}\right)' \right) \left(c \cdot h^{p+1} \cdot \omega_{\varphi,+}^3 + |\, \omega_{\varphi,+} - \omega_{\Phi,+}| \right).$$

Now let us examine this supremum. We observe that

$$\sup_{[\{\mathcal{N}_\Phi^E(\omega_{\varphi,+}), \omega_{\Phi,+}\}]} \left((\mathcal{N}_\Phi^E)^{[-1]}\right)' = \sup_{[\{\mathcal{N}_\Phi^E(\omega_{\varphi,+}), \omega_{\Phi,+}\}]} \frac{1}{(\mathcal{N}_\Phi^E)' \circ (\mathcal{N}_\Phi^E)^{[-1]}} =$$

$$\sup_{[\{\omega_{\varphi,+}, \omega_{\Phi,+}\}]} \frac{1}{(\mathcal{N}_\Phi^E)'} \leq \frac{1}{(\mathcal{N}_\Phi^E)'(\min(\omega_{\varphi,+}, \omega_{\Phi,+}))} \leq$$

$$\frac{1}{(\mathcal{N}_\Phi^E)'\left(\frac{\sqrt{|\alpha|}}{2}\right)} \leq \frac{1}{1 + \frac{h}{2}\sqrt{|\alpha|}} \leq 1 - \frac{h}{4}\sqrt{|\alpha|}, \tag{2.52}$$

by taking into account that the positive function $(\mathcal{N}_\Phi^E)'$ is monotone increasing, the corresponding estimates (cf. Lemma 2.3.1) $\frac{\sqrt{|\alpha|}}{2} \leq \sqrt{\frac{2}{3}}\sqrt{|\alpha|} \leq \omega_{\varphi,+}, \omega_{\Phi,+} \leq \sqrt{2}\sqrt{|\alpha|}$ for both positive fixed points of the normal forms, further the fact that $(\mathcal{N}_\Phi^E)'(x) \geq 1 + hx$, if $0 \leq x$ is sufficiently small, finally the inequality $\frac{1}{1+x} \leq 1 - \frac{x}{2}$, if $0 \leq x \leq 1$. From these we express the desired quantity to get

$$|\, \omega_{\varphi,+} - \omega_{\Phi,+}| \cdot \frac{h}{4}\sqrt{|\alpha|} \leq \left(1 - \frac{h}{4}\sqrt{|\alpha|}\right) c \cdot h^{p+1} \cdot \omega_{\varphi,+}^3$$

which in turn results in the inequality

$$|\, \omega_{\varphi,+} - \omega_{\Phi,+}| \leq 8\sqrt{2}\, c \cdot h^p \cdot |\alpha|.$$

2.5 Preparation: Growth of the iterates in the $\alpha > 0$ case

For the construction of a conjugacy and the corresponding closeness estimates in the $\alpha > 0$ case, the current preparatory section analyzes some properties of the orbit of 0 under mappings of the form $x \mapsto h\alpha + x + hx^2 + hx^3 \cdot \eta(h, x, \alpha)$ with η from a suitable function class. (Then, of course, η will be replaced either by $\widehat{\eta}_3$ or $\widetilde{\eta}_3$.)

Let $p_n \equiv p_n(h, \alpha)$ denote any sequence satisfying $p_0 = 0$ and

$$p_{n+1} = \mathcal{N}_\eta(h, p_n, \alpha) \tag{2.53}$$

for $n \in \mathbb{N}$, where $\mathcal{N}_\eta(h, x, \alpha) := h\alpha + x + hx^2 + hx^3 \cdot \eta(h, x, \alpha)$ and η is any smooth function with $|\eta|$, $|\frac{\mathrm{d}}{\mathrm{d}x}\eta|$ and $|\frac{\mathrm{d}}{\mathrm{d}x^2}\eta|$ bounded, again, by some $K > 0$ uniformly for all $h \in (0, h_0]$, $x \in [-\varepsilon_0, \varepsilon_0]$ and $\alpha \in (0, \alpha_0]$, when $h_0 > 0$, $\varepsilon_0 > 0$ and $\alpha_0 > 0$ are sufficiently small. In what follows, we fix parameters $h \in (0, h_0]$ and $\alpha \in (0, \alpha_0]$ arbitrarily.

First notice that the asymptotic behaviour of p_n can be qualitatively different for different choices of η—for example, p_n can be unbounded, but can tend to a finite limit as well. In order to make its behaviour uniform, we will cut p_n at some suitable value $\kappa > 0$ and consider only the terms of the sequence below this *cutting level.*

Lemma 2.5.1 *Let* $\kappa := \min\left(\frac{3}{8}, (13\widetilde{K})^{-1}\right)$ *with* $\widetilde{K} := K + 3h_0$. *Then the sequence* p_n *reaches level* κ *at some* n*, further, for* $p_n \le \kappa$ *the sequence is strictly monotone increasing, and for* $0 < x \le \kappa$ *both* $\left(\mathcal{N}_\eta^E\right)'(x)$ *and* $\left(\mathcal{N}_\eta^E\right)''(x)$ *are positive.*

Proof. Since $0 < x \le 1$, we have $\left(\mathcal{N}_\eta^E\right)'(x) \ge 1 + hx(2 - 3Kx - Kx)$ being positive due to $x \le (4K)^{-1}$. Similarly, $\left(\mathcal{N}_\eta^E\right)''(x) \ge h(2 - 6Kx - 6Kx - Kx) > 0$ because of $x \le (13K)^{-1}$. Strict monotonicity of p_n follows easily from $p_{n+1} - p_n > hp_n^2(1 - Kp_n) > 0$. Finally, since $(p_{n+2} - p_{n+1}) - (p_{n+1} - p_n) = h \cdot (t(p_{n+1}) - t(p_n))$ with $t(x) := x^2 + x^3 \cdot \eta^E(x)$, and $t(p_{n+1}) - t(p_n) = t'(\xi) \cdot (p_{n+1} - p_n)$ with some $\xi \in (p_n, p_{n+1})$, further $t'(\xi) \ge \xi(2 - 3\xi K - \xi K) > 0$, the proof is complete. ■

Remark 2.5.1 The role of $\widetilde{K}$ will be explained by Lemma 2.5.2, while that of $\frac{3}{8}$ by Lemma 2.5.8.

Having assured the strict monotonicity of p_n, the rest of this section will be devoted to devising suitable upper estimates for the sequence.

The behaviour of p_n under the level κ is the juxtaposition of two, qualitatively different phases.

In the interval $[0, \sqrt{\alpha}]$, the sequence is mainly determined by the term $h\alpha$ in the recursive definition (2.53), hence here $p_n \approx nh\alpha$, see (2.59) in the proof of Lemma 2.5.6.

However, after the level $\mathcal{O}(\sqrt{\alpha})$ has been passed, higher order terms begin to dominate and the linear growth suddenly turns into a steep increase.

Therefore, splitting our investigations into two is natural: the "trivial" linear part, and the tail part of the sequence will be treated separately. In this latter region—due to the fact that higher order terms are only "weakly" α-dependent, as α is present only in η—it is reasonable to expect some similarities between the $\alpha > 0$ ($\alpha \to 0^+$) and the $\alpha = 0$ cases—and indeed, we will explicitly exploit this phenomenon. Hence, an essential part of the proofs in this section will not contain α. It seems hard, however, to control the growth rate of p_n effectively as n *increases*, that is, to devise suitable *global* estimates of p_n in terms of n *and* to say something meaningful about the index where the sequence reaches level κ, see Proposition 2.7.1 and the subsequent remarks. Nevertheless, a "backward" approach will work, that is, properties of the *inverse-iteration* can be grasped better by considering p_{N-k} as k increases, where N is chosen such that $p_N \approx \kappa$.

In exploring quantitative properties of this sequence described by the current Section, the program *Mathematica* has been heavily relied upon.

We first obtain an *a priori* inverse estimate for one term of the sequence in terms of its successor.

Lemma 2.5.2 *Suppose* $h_0 \le \frac{1}{3}$, $h\alpha \le 1$, *further* κ *and* $\widetilde{K}$ *are as above. Then for all* $n \ge 1$ *satisfying* $p_n \le \kappa$ *we have that*

$$p_{n-1} \le p_n - h\alpha + h^2\alpha - hp_n^2 + h\widetilde{K}p_n^3. \tag{2.54}$$

Proof. Substituting $n-1$ (instead of n) into (2.53), rearranging, and using the upper and lower bounds of η together with the fact that p_{n-1} is nonnegative, we see that

$$p_n - h\alpha - hp_{n-1}^2 - hKp_{n-1}^3 \leq p_{n-1} \leq p_n - h\alpha - hp_{n-1}^2 + hKp_{n-1}^3. \tag{2.55}$$

From the left hand side inequality—since p_n is monotone increasing and positive—we get

$$p_n - h\alpha - hp_n^2 - hKp_n^3 \leq p_{n-1}. \tag{2.56}$$

Now we first show that the left hand side of (2.56) is nonnegative for $n \geq 2$. Using $p_n \leq \frac{1}{2}$, $Kp_n \leq \frac{1}{2}$ and $h \leq \frac{1}{3}$, we have $h(\alpha+p_n^2+Kp_n^3) \leq h(\alpha+\frac{1}{2}p_n+\frac{1}{4}p_n) \leq h\alpha+\frac{1}{4}p_n$. From this we get that the left hand side of (2.56) is nonnegative if $\frac{4}{3}h\alpha \leq p_n$. But since $p_1 = h\alpha$ and $p_2 > 2h\alpha$, condition $\frac{4}{3}h\alpha \leq p_n$ is implied by $n \geq 2$. So we temporarily assume $n \geq 2$.

Rearrangement of (2.56) thus yields

$$-(p_n - h\alpha - hp_n^2 - hKp_n^3)^2 \geq -p_{n-1}^2$$

for $n \geq 2$. Now let us combine this with the right hand side inequality of (2.55), showing for $n \geq 2$ that

$$p_{n-1} \leq p_n - h\alpha - h(p_n - h\alpha - hp_n^2 - hKp_n^3)^2 + hKp_{n-1}^3.$$

Now we will simplify the right hand side here to arrive at the desired result. To this end, first replace the term hKp_{n-1}^3 with hKp_n^3 by monotonicity, then expand the square to get

$$p_{n-1} \leq p_n - h\alpha - hp_n^2 + hKp_n^3 + 2h^2p_n(\alpha + p_n^2 + Kp_n^3) - h^3(\alpha + p_n^2 + Kp_n^3)^2. \tag{2.57}$$

Let us examine the last two terms above. The last negative term can safely be omitted, so we are left with estimating $2h^2p_n(\alpha+p_n^2+Kp_n^3)$ from above. But using again $p_n \leq \frac{1}{2}$ and $Kp_n \leq \frac{1}{2}$, we get $2h^2p_n(\alpha + p_n^2 + Kp_n^3) = 2p_nh^2\alpha + 2h^2p_n^3(1 + Kp_n) \leq h^2\alpha + hp_n^3 3h$.

Hence, by suitable upper estimates, (2.57) has been transformed into (2.54), with $\widetilde{K} := K + 3h_0$ for $n \geq 2$.

Finally, we directly verify (2.54) for $n = 1$. A direct substitution $p_0 = 0$ and $p_1 = h\alpha$ yields that it is sufficient to have $h^2\alpha - hp_1^2 + h\widetilde{K}p_1^3 \geq 0$, which is, however, implied by $h\alpha \leq 1$. ■

Remark 2.5.2 Inequalities (2.54) and (2.56) quantitatively express the natural fact that the inverse mapping of (identity + higher order terms), *i.e.* of (2.53), has the form (identity − perturbed higher order terms). Of course, it is not apparent at the first sight, how large this perturbation can be in terms of the parameters h and α.

Let us denote by $N \equiv N(h, \alpha)$ the unique index where the sequence p_n *passes level* κ, that is determine $N \in \mathbb{N}$ in such a way that $p_N \leq \kappa$ but $p_{N+1} > \kappa$.

Although $p_N \leq \kappa$, it will be important later to exclude the possibility of p_N being too small as $h, \alpha \to 0^+$. The next Lemma shows that, under appropriate conditions, p_N is uniformly separated from 0.

Lemma 2.5.3 *Suppose that the conditions of Lemma 2.5.2 hold and $\alpha_0 \leq \kappa$. Then $p_N \geq \frac{\kappa}{2}$ for all $h \in (0, h_0]$ and $\alpha \in (0, \alpha_0]$.*

Proof. Suppose, to the contrary, that $p_N < \frac{\kappa}{2}$ holds. Then by $K\kappa \leq 1$, $3h_0 \leq 1$ and $\kappa \leq 1$, one would get

$$\kappa < p_{N+1} = h\alpha + p_N + hp_N^2 + hp_N^3 \cdot \eta^E(p_N) \leq h\alpha + \frac{\kappa}{2} + h\frac{\kappa^2}{4} + h\frac{\kappa^3}{8}K \leq$$

$$h\alpha+\frac{\kappa}{2}+h\frac{\kappa^2}{4}+h\frac{\kappa^2}{8}=\frac{\kappa}{2}+h\alpha+\frac{3h\kappa^2}{8}\leq\frac{\kappa}{2}+\frac{\alpha_0}{3}+\frac{\kappa^2}{8}\leq\frac{\kappa}{2}+\kappa\left(\frac{1}{3}+\frac{1}{8}\right)<\kappa,$$

a contradiction. ■

In a similar fashion, we can replace the level κ to be passed by $\sqrt{\alpha}$. This type of result will also be needed later.

Lemma 2.5.4 *Suppose that the conditions of Lemma 2.5.2 hold and* $0<\alpha_0\leq\kappa^2$. *If m is the index such that* $p_m\leq\sqrt{\alpha}$, *but* $p_{m+1}>\sqrt{\alpha}$, *then* $p_m\geq\frac{\sqrt{\alpha}}{2}$ *for any* $h\in(0,h_0]$ *and* $\alpha\in(0,\alpha_0]$.

Proof. Suppose, to the contrary, that $p_m<\frac{\sqrt{\alpha}}{2}$. Then by $p_mK\leq\kappa K\leq 1$, $3h_0\leq 1$ and $\alpha\leq 1$, we would get

$$\sqrt{\alpha}<p_{m+1}=h\alpha+p_m+hp_m^2+hp_m^3\cdot\eta^E(p_m)\leq h\alpha+p_m+2hp_m^2\leq$$

$$h\alpha+\frac{\sqrt{\alpha}}{2}+h\frac{\alpha}{2}\leq\frac{\sqrt{\alpha}}{2}+\frac{3}{2}h\alpha\leq\frac{\sqrt{\alpha}}{2}+\frac{\sqrt{\alpha}}{2}=\sqrt{\alpha},$$

a contradiction. ■

For $k\in\mathbb{N}$ sufficiently large, we now deduce a rough, but direct auxiliary estimate for p_{N-k} based on Lemma 2.5.2. However, it will be required later to have estimates not only for p_{N-k}, but also for p_{N^*-k}, where $0<N^*\leq N$, so we have to prove a bit more general statement.

Lemma 2.5.5 *Suppose that the conditions of Lemma 2.5.2 hold. Set* $k_1:=\frac{1}{h\kappa}$ *and let* $N^*\in\mathbb{N}^+$ *be arbitrary with* $N^*\leq N$. *Then for* $k_1\leq k\leq N^*$

$$p_{N^*-k}\leq\frac{2}{hk}. \tag{2.58}$$

Proof. Due to the monotone increasing property of p_n and the fact that we are dealing with upper estimates, it is sufficient to prove everything for N instead of $N^*\leq N$.

The proof is by induction on k. Since $\kappa h\leq 1$, it is clear for $k=\lceil k_1\rceil$ that $p_{N-k}\leq p_N\leq\kappa\leq\frac{2}{h(\frac{1}{h\kappa}+1)}\leq\frac{2}{hk}$.

So assume (2.58) is true for some $k\geq k_1$. Then by (2.54)—omitting the nonpositive $-h\alpha+h^2\alpha$ due to $h\leq 1$—we see that

$$p_{N-(k+1)}\leq p_{N-k}-hp_{N-k}^2+h\widetilde{K}p_{N-k}^3.$$

Since the function $p\mapsto p-hp^2+h\widetilde{K}p^3$ is monotone increasing for $0<p\leq\kappa<\frac{1}{2h}$, it is enough to show

$$\frac{2}{hk}-h\frac{4}{h^2k^2}+h\widetilde{K}\frac{8}{h^3k^3}\leq\frac{2}{h(k+1)}$$

to finish the induction. But the above is equivalent to

$$\frac{8\widetilde{K}}{h^2k^3}\leq\frac{4}{hk^2}-\frac{2}{hk(k+1)},$$

which is a bit more strengthened if its right hand side is decreased by writing $\frac{4}{hk^2}-\frac{2}{hk^2}$. So it is sufficient to establish

$$\frac{8\widetilde{K}}{h^2k^3}\leq\frac{2}{hk^2},$$

being equivalent to $4\widetilde{K} \leq hk$, which latter is however implied by $4\widetilde{K} \leq \frac{1}{\kappa} = hk_1 \leq hk$ due to the definition of κ and k_1. ■

Remark 2.5.3 The lemma above can be considered as a counterpart of Lemma 2.4.3.

Remark 2.5.4 In this elementary argument one could replace the 2 in the numerator in (2.58) by $1+\delta$, with δ being an arbitrarily small positive number. However, the limiting process $\delta \to 0^+$ is not allowed, because this would shift k_1 to infinity (or the cutting level κ to zero).

On the other hand, by scanning the proofs of (2.72) and (2.78), it can be seen that a constant strictly greater than 1 in the numerator in (2.58) would destroy the order of magnitude of the upper estimate (2.78): instead of the logarithmic singularity $\ln \frac{1}{\alpha}$, one would get only $\left(\frac{1}{\alpha}\right)^{\varepsilon}$, with a suitable $\varepsilon > 0$ as $\alpha \to 0^+$. So, for the sake of a sharper result, we are going to analyze deeper the growth rate of p_n.

First, we prove that the maximal index $N \equiv N(h,\alpha)$ is $\mathcal{O}(\frac{1}{h\sqrt{\alpha}})$, as $h, \alpha \to 0^+$.

Lemma 2.5.6 *Suppose that the conditions of Lemma 2.5.2 hold and* $0 < \alpha_0 \leq \kappa^2$. *Then we have*

$$\begin{aligned} \#\{p_n | p_n \in [0, \sqrt{\alpha}]\} &\leq \frac{1}{h\sqrt{\alpha}}, \\ \#\{p_n | p_n \in [\sqrt{\alpha}, \kappa]\} &\leq \frac{2}{h\sqrt{\alpha}} + \frac{1}{h\kappa}, \end{aligned}$$

and hence

$$N \leq \frac{3}{h\sqrt{\alpha}} + \frac{1}{h\kappa} \leq \frac{4}{h\sqrt{\alpha}}.$$

Proof. Since $0 \leq p_n \leq \kappa$, again $hp_n^2(1 + p_n \cdot \eta^E(p_n)) \geq 0$, so for $0 \leq n \leq N$ we get the *trivial lower estimate* from (2.53)

$$p_n \geq nh\alpha, \tag{2.59}$$

which yields the upper estimate to the number of elements of the first set in the Lemma. To get the second estimate, apply (2.58) but noticing that at most $\frac{1}{h\kappa}$ elements should be counted separately due to the restriction on the starting index k in (2.58). ■

Remark 2.5.5 It is possible to prove $N \leq \frac{2+\delta}{h\sqrt{\alpha}}$, with $\delta > 0$ being arbitrary small. But, as remarked previously, such an improvement would be of no help for the following estimates.

Returning to our "backward approach", we develop two lemmas—the refined counterparts of Lemma 2.5.2 and Lemma 2.5.5.

The first subtle step is to conceal cubic terms in the inverse iteration by introducing a new sequence s_k.

Lemma 2.5.7 *Suppose that the conditions of Lemma 2.5.2 hold, and again,* $\frac{1}{h\kappa} =: k_1 \leq k \leq N^*$ *with some* $N^* \leq N$. *Then*

$$p_{N^*-(k+1)} \leq p_{N^*-k} - hs_k p_{N^*-k}^2, \tag{2.60}$$

where $s_k \equiv s_k(h,\kappa) := 1 - \frac{1}{hk\kappa}$ *for* $k \geq k_1$.

Proof. Inequality (2.54) implies that $p_{N^*-(k+1)} \leq p_{N^*-k} - hp_{N^*-k}^2(1 - \widetilde{K}p_{N^*-k})$. Now use (2.58) and the definition of κ to obtain

$$1 - \widetilde{K}p_{N^*-k} \geq 1 - \widetilde{K}\frac{2}{hk} = \frac{hk - 2\widetilde{K}}{hk} \geq \frac{hk - \frac{1}{\kappa}}{hk} = s_k \geq 0.$$

These yield (2.60). ■

Now we can state and prove our main tool in the $\alpha > 0$ case. The Lemma below yields additional information on the convergence speed of the backward iteration, being fundamental to the final closeness estimates.

Lemma 2.5.8 (The $\frac{3}{2}$-Lemma) *Suppose that the conditions of Lemma 2.5.2 hold, but now* $\frac{1}{h\kappa^2} \le k \le N^*$ *with some* $N^* \le N$. *Then*

$$p_{N^*-k} \le \frac{1}{hk} + \frac{1/\kappa}{(hk)^{3/2}}. \tag{2.61}$$

Proof. Again, it is enough to prove everything for N instead of N^*.

We prove by induction on k. The induction can be started because for $k = \frac{1}{h\kappa^2} \ge \frac{1}{h\kappa}$, (2.58) yields that $p_{N-\lceil k\rceil} \le 2(h\lceil \frac{1}{h\kappa^2}\rceil)^{-1} \le 2\kappa^2 = \frac{1}{hk} + \frac{1/\kappa}{(hk)^{3/2}}$.

Let us now introduce the abbreviation $P(h,k,\kappa) := \frac{1}{hk} + \frac{1/\kappa}{(hk)^{3/2}}$ and suppose that $p_{N-k} \le P(h,k,\kappa)$ holds for some $k \ge \frac{1}{h\kappa^2}$. Then using (2.60) together with the monotonicity of the function $p \mapsto p - hs_k p^2$ (being true since $p_{N-k} \le \kappa < \frac{1}{2hs_k}$), we get

$$p_{N-(k+1)} \le p_{N-k} - hs_k p_{N-k}^2 \le P(h,k,\kappa) - hs_k P^2(h,k,\kappa).$$

Hence, clearly, in order to finish the induction, it is sufficient to establish that

$$Q(h,k,\kappa) := P(h,k+1,\kappa) - P(h,k,\kappa) + hs_k P^2(h,k,\kappa) \ge 0,$$

for all $h \in (0,1)$, $\kappa \in (0,\frac{3}{8}]$ and $k \ge \frac{1}{h\kappa^2}$.

To this end, we first shift the second argument in Q by setting $\ell := k - \frac{1}{h\kappa^2}$, then introduce a new variable $A := 1 + h\kappa^2\ell$ to get

$$Q(h,k,\kappa) \equiv Q\left(h, \frac{1}{h\kappa^2} + \ell, \kappa\right) \equiv Q\left(h, \frac{A}{h\kappa^2}, \kappa\right).$$

(So $\ell \ge 0$ is arbitrary, thus $A \ge 1$ is also arbitrary.) Albeit the expressions above are mathematically equivalent, yet, from a structural point of view, they are substantially different: the last form can be simplified to

$$Q\left(h, \frac{A}{h\kappa^2}, \kappa\right) \equiv$$

$$\kappa^2\left(\frac{1}{A+h\kappa^2} - \frac{1}{A} + \frac{1}{(A+h\kappa^2)^{3/2}} - \frac{1}{A^{3/2}} + \frac{(1+\sqrt{A})^2(A-\kappa)h\kappa^2}{A^4}\right),$$

where a new parameter $\nu := h\kappa^2$ is immediately introduced. Also dropping the positive factor κ^2 outside, and noticing that the whole expression is not increased if the only explicitly remaining κ is replaced by its maximal value $\frac{3}{8}$ in $(A-\kappa)$, we arrive at the following inequality in *two* variables

$$0 \le \frac{1}{A+\nu} - \frac{1}{A} + \frac{1}{(A+\nu)^{3/2}} - \frac{1}{A^{3/2}} + \frac{(1+\sqrt{A})^2(A-\frac{3}{8})\nu}{A^4} \tag{2.62}$$

to be shown for all $A \ge 1$ and (even for all) $\nu \ge 0$.

Let us abbreviate the right hand side of (2.62) by $R(A,\nu)$ and notice that $R(A,0) = 0$ for all $A \ge 1$. Furthermore, notice that for $\nu > 0$, the partial derivative $\partial_\nu R(A,\nu)$ satisfies

$$\partial_\nu R(A,\nu) = \frac{(1+\sqrt{A})^2(A-\frac{3}{8})}{A^4} - \frac{3}{2(A+\nu)^{5/2}} - \frac{1}{(A+\nu)^2} >$$

$$\frac{(1+\sqrt{A})^2(A-\frac{3}{8})}{A^4} - \frac{3}{2A^{5/2}} - \frac{1}{A^2} = \frac{(\sqrt{A}-1)(4A+9\sqrt{A}+3)}{8A^4} \geq 0.$$

The proof is complete. ■

Remark 2.5.6 The exponent in (2.61) has been postulated to be $\frac{3}{2}$, because it is the "simplest" number between 1 and 2. Numerical tests suggest that this fractional order is necessary, since if the exponent $\frac{3}{2}$ was replaced by 2, then—according to the tests—nonnegativity of the counterpart of $Q\left(h, \frac{A}{h\kappa^2}, \kappa\right)$ would not hold uniformly, *i.e.* for any small $\kappa > 0$ in the factor $(A-\kappa)$, it is possible to choose $A \gg 1$ and $0 < \nu \ll 1$ such that the corresponding Q-expression is negative.

The very same fact is indicated in [28] as well, when studying the recursion

$$u_{k+1} = g(u_k),$$

with $g(x) \equiv x - bx^{q+1} + \mathcal{O}(x^{q+2})$ and $b > 0$, $q \in \mathbb{N}$ being fixed parameters. If $u_0 > 0$ is *sufficiently small*, then the sequence u_k tends to 0 as $k \to \infty$. The convergence speed of this iteration in terms of b and q together with strict lower and upper bounds for u_k are given in [28], provided that k is *large enough*. We remark that the same limiting relation in the $q = 1$ case—though with a different proof—also appears in [44]. It is seen that the iteration u_k with $q = 1$ is of the same type as our backward iteration p_{N-k} when $\alpha = 0$. The upper bounds given in [28] have a similar fractional order structure as (2.61). However, it will be important for us to have *explicit* estimates from a starting index of the form $\frac{\text{const}}{h}$; estimates only from a sufficiently large and unspecified starting index k would be insufficient.

We add that the sharpest estimate concerning this class of iteration we know about is contained in [5]. Define, similarly as above,

$$u_{k+1} = u_k - u_k^2 + \mathcal{O}(u_k^3)$$

and suppose that $u_0 > 0$ is chosen so small such that $u_k \to 0$. Then [5] sketches the proof of

$$u_k = \frac{1}{k} + \mathcal{O}\left(\frac{\log k}{k^2}\right).$$

However, the above sharper convergence rate is not yet an explicit estimate, and even if it was, it would not make our later closeness estimates better.

Remark 2.5.7 After the form of inequality (2.61) to be proved was set, the maximal value $\frac{3}{8}$ of κ became sharp—it originates from the Taylor series expansion of $Q\left(h, \frac{1}{h\kappa^2}, \kappa\right)$ about the origin:

$$\lim_{h\to 0} \frac{\mathrm{d}}{\mathrm{d}h}\left(Q\left(h, \frac{1}{h\kappa^2}, \kappa\right)\right) = \frac{\kappa^4}{2}(3-8\kappa).$$

This necessary condition $\kappa \leq \frac{3}{8}$ for nonnegativity of Q turned out to be sufficient as well, further $\kappa = \frac{3}{8}$ allows the nice factorization in the lower estimate of $\partial_\nu R(A,\nu)$.

During the search for the proof of $Q \geq 0$, the combined symbolic, numeric and graphical capabilities of *Mathematica* proved to be indispensable. The main source of problems has been the fact that the function Q with two parameters fixed often exhibits unimodality. The simple structural manipulations described in the Lemma above successfully eliminate unimodality as well as reduce the number of parameters by suitably grouping them together.

2.6 The conjugacy and closeness estimates in the $\alpha > 0$ case

Let us consider our mappings

$$x \mapsto \mathcal{N}_\Phi(h,x,\alpha) \equiv h\alpha + x + hx^2 + hx^3 \cdot \widehat{\eta}_3(h,x,\alpha) \tag{2.63}$$

and

$$x \mapsto \mathcal{N}_\varphi(h,x,\alpha) \equiv h\alpha + x + hx^2 + hx^3 \cdot \widetilde{\eta}_3(h,x,\alpha) \tag{2.64}$$

for any fixed $h \in (0,h_0]$ and $\alpha \in (0,\alpha_0]$. Both $\widehat{\eta}_3$ and $\widetilde{\eta}_3$ satisfy the assumptions at the beginning of Section 2.5, that is they are smooth functions with a *common* uniform bound $K > 0$. Suppose, that they are sufficiently close, that is there exists a positive constant $c > 0$ such that

$$|\mathcal{N}_\Phi(h,x,\alpha) - \mathcal{N}_\varphi(h,x,\alpha)| \le c \cdot h^{p+1}|x|^\omega \tag{2.65}$$

holds for all $h \in (0,h_0]$, $x \in [-\varepsilon_0,\varepsilon_0]$ and $\alpha \in (0,\alpha_0]$, where the exponent is assumed to be $\omega = 3$ (what we have proved in Section 2.2) or $\omega = 4$ (an additional assumption).

For every fixed $h \in (0,h_0]$ and $\alpha \in (0,\alpha_0]$ we construct a conjugacy between (2.63) and (2.64), that is a strictly monotone increasing map $x \mapsto J(h,x,\alpha)$ in a neighbourhood $[-\varepsilon_0,\varepsilon_0]$ of the origin such that

$$\mathcal{N}_\Phi^E \circ J^E = J^E \circ \mathcal{N}_\varphi. \tag{2.66}$$

We will deal only with the case $x \in [0,\varepsilon_0]$, the negative part $x \in [-\varepsilon_0,0]$ (using the appropriate inverse mappings) is similar.

To this end, suppose—again as in Section 2.5—that the sequence p_n is the orbit of 0 under (2.63), while the sequence q_n is the orbit of 0 under (2.64), with $p_0 = q_0 \equiv 0$. Hence, all the results of Section 2.5 can be applied to both p_n and q_n—quantities κ, K and $\widetilde{K}$ are the same in both cases, however, of course, a clear distinction should be made between the cutting indices: let us denote by N_p the index where $p_{N_p} \le \kappa$ but $p_{N_p+1} > \kappa$, and similarly, by N_q the index where $q_{N_q} \le \kappa$ but $q_{N_q+1} > \kappa$. Since we are going to work with p_n and q_n simultaneously, they both should be kept below κ for the results of Section 2.5 to work, so a common cutting index N^* is now defined as

$$N^* := \min(N_p, N_q).$$

The following figure shows the first few (but same number of) terms of the sequences $q_n(h,\alpha_1)$ and $q_n(h,\alpha_2)$ in the (α,x)-plane with some $\alpha_1 > 0$, $\alpha_2 > 0$ and $h > 0$ fixed. Condensation of the sequences near the horizontal axis is clearly visible, however, for any $\alpha > 0$, due to the absence of the fixed points, they will eventually pass this axis, then begin increasing rapidly. (Note that for the sake of a better comparison, the value of q_0 has been redefined on this plot as $q_0 := -\frac{1}{2}$. The branch of stable and unstable fixed points of $\mathcal{N}_\varphi^E$ are also displayed. Again, the arrows point toward terms of the sequences with larger n indices.)

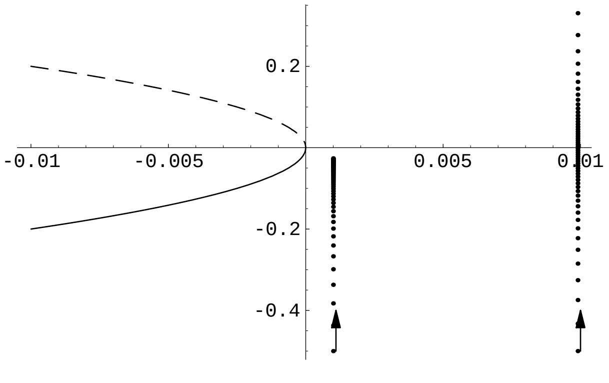


Figure 2.6.1

The figure below depicts part of the global dynamics of the map $\mathcal{N}_{\varphi}^{E}$ near the bifurcation point in the (α, x)-plane. The same n_0 $(0 \le n \le n_0)$ number of terms of the sequences $y_n(h, \alpha)$, $y_0 := -\frac{1}{2}$ and $q_n(h, \alpha)$, $q_0 := -\frac{1}{2}$ are displayed together (cf. Figure 2.3.1 and Figure 2.6.1), with $h > 0$ fixed and α running from -0.01 to 0.01 on an equidistant grid.

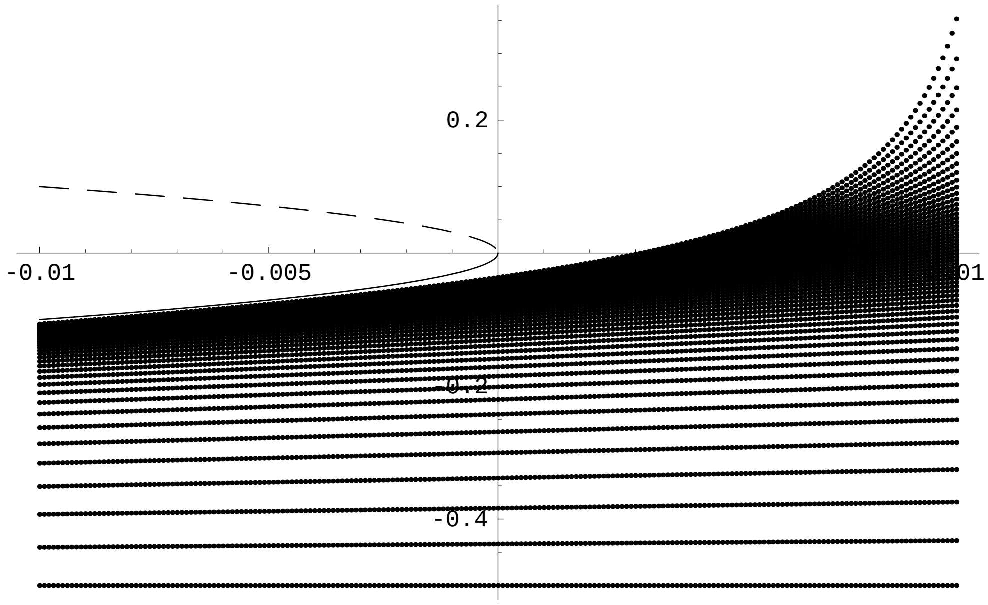


Figure 2.6.2

Now fix $h \in (0, h_0]$ and $\alpha \in (0, \alpha_0]$. Set

$$J^E(0) := 0. \tag{2.67}$$

Then (2.66) recursively forces the definition of J^E at q_n to be

$$J^E(q_n) := \left(\mathcal{N}_{\Phi}^{E}\right)^{[n]} (J^E(0)) \equiv \left(\mathcal{N}_{\Phi}^{E}\right)^{[n]} (0) \equiv p_n.$$

Define further $J^E(x) := x$ for $x \in [0, q_1]$. This will be a compatible extension, since $J^E(q_1) = p_1$ by definition, but $q_1 = p_1 \equiv h\alpha$. Then using these, together with (2.66) recursively, we can extend J^E homeomorphically in an (upper semi-)neighbourhood of the origin. We have thus proved the following theorem.

Theorem 2.6.1 *For every fixed $h \in (0, h_0]$ and $\alpha \in (0, \alpha_0]$, there exists a conjugacy $J(h, \cdot, \alpha)$ between (2.63) and (2.64) defined in a uniform neighbourhood $[-\varepsilon_0, \varepsilon_0]$ of the origin with $\varepsilon_0 := \kappa > 0$.*

Our aim now will be to measure the distance of J^E from the identity.

Remark 2.6.1 Due to the monotonicity of the mapping J^E, its growth rate and its distance from the identity can not be affected by the chosen extension on $[0, q_1]$, hence the only degree of freedom in the construction is prescribing the value of $J^E(0)$.

First we present an auxiliary estimate, similar to (2.36) before. Suppose $0 < a < b$. Using (2.66), Lemma 2.5.1 for the monotonicity and (2.65), we get

$$\sup_{[\mathcal{N}_\varphi^E(a),\mathcal{N}_\varphi^E(b)]} |\, id - J^E| = \sup_{[\mathcal{N}_\varphi^E(a),\mathcal{N}_\varphi^E(b)]} \left| \mathcal{N}_\varphi^E \circ (\mathcal{N}_\varphi^E)^{[-1]} - \mathcal{N}_\Phi^E \circ J^E \circ (\mathcal{N}_\varphi^E)^{[-1]} \right| \leq$$

$$\sup_{[a,b]} \left| \mathcal{N}_\varphi^E - \mathcal{N}_\Phi^E \right| + \sup_{[a,b]} \left| \mathcal{N}_\Phi^E - \mathcal{N}_\Phi^E \circ J^E \right| \leq$$

$$c \cdot h^{p+1} b^\omega + \sup_{x\in[a,b]} \left(\left(\sup_{[\{x,J^E(x)\}]} (\mathcal{N}_\Phi^E)' \right) |x - J^E(x)| \right) \leq$$

$$c \cdot h^{p+1} b^\omega + (\mathcal{N}_\Phi^E)' \left(\max \left(b, J^E(b)\right)\right) \cdot \sup_{[a,b]} |\, id - J^E|. \tag{2.68}$$

For $n \in \mathbb{N}^+$, let us abbreviate the supremum by

$$S_n \equiv S_n(h,\alpha) := \sup_{[q_{n-1},q_n]} |\, id - J^E|$$

and the derivative by

$$D_n \equiv D_n(h,\alpha) := (\mathcal{N}_\Phi^E)' \left(\max \left(q_n, J^E(q_n)\right)\right).$$

With $n \leq N^*$, $a = q_{n-1}$ and $b = q_n$, (2.68) therefore becomes

$$S_{n+1} \leq D_n S_n + c \cdot h^{p+1} q_n^\omega.$$

Applying this recursively, we construct the upper estimate for $n \leq N^*$

$$S_{n+1} \leq \left(\prod_{i=1}^n D_i \right) S_1 + c \cdot h^{p+1} \left(\sum_{i=1}^n \left(\prod_{j=i+1}^n D_j \right) q_i^\omega \right), \tag{2.69}$$

where, of course, the product $\prod_{j=n+1}^n D_j$ is understood to be 1.

The first term on the right hand side vanishes, since $S_1 \equiv 0$ by construction. The second term, however, is monotone increasing in $n \leq N^*$, so we get

$$\sup_{[0,q_{N^*+1}]} |\, id - J^E| \leq c \cdot h^p \left(h \sum_{i=1}^{N^*} \left(\prod_{j=i+1}^{N^*} D_j \right) q_i^\omega \right). \tag{2.70}$$

In order to be able to estimate the right hand side of (2.70), we prove an important estimate concerning the sum of powers of μ_{N^*-k}, where

$$\mu_n \equiv \mu_n(h,\alpha) := \max(q_n, p_n),$$

for $0 \leq n \leq N^*$.

Lemma 2.6.2 *Suppose that the conditions of Lemma 2.5.2 hold and $0 < \alpha_0 \leq \kappa^2$. Then there exist positive constants* $\mathrm{const}_1(\kappa) > 0$ *and* $\mathrm{const}_2(\kappa) > 0$*, depending only on κ, such that for any index $i \in \{0, 1, \ldots, N^*\}$ and for any $h \in (0, h_0]$ and $\alpha \in (0, \alpha_0]$ we have that*

$$2h \sum_{k=0}^i \mu_{N^*-k} \leq \mathrm{const}_1(\kappa), \tag{2.71}$$

provided that $0 \le i \le \lfloor \frac{1}{h\kappa^2} \rfloor$, *and*

$$2h \sum_{k=0}^{i} \mu_{N^*-k} \le \mathrm{const}_2(\kappa) + 2\ln(h\,i), \tag{2.72}$$

provided that $\lfloor \frac{1}{h\kappa^2} \rfloor + 1 \le i \le N^*$.

Further, for any $\delta > 0$ *there exists a positive constant* $\mathrm{const}_3(\delta, \kappa) > 0$, *depending on* δ *and* κ, *such that for any* $h \in (0, h_0]$ *and* $\alpha \in (0, \alpha_0]$ *we have that*

$$h \sum_{k=0}^{N^*} (\mu_{N^*-k})^{1+\delta} \le \mathrm{const}_3(\delta, \kappa). \tag{2.73}$$

Proof. By the definition of N^* and μ_n, $0 \le \mu_n \le \kappa$ for any $n \in \{0, 1, \ldots, N^*\}$. If $i \le \lfloor \frac{1}{h\kappa^2} \rfloor$, then

$$2h \sum_{k=0}^{i} \mu_{N^*-k} \le 2h \sum_{k=0}^{\lfloor \frac{1}{h\kappa^2} \rfloor} \kappa \le 2h\kappa \left(\frac{1}{h\kappa^2} + 1 \right) \le \frac{2}{\kappa} + 2h_0\kappa < \frac{2}{\kappa} + \kappa,$$

which shows (2.71).

Assume now that $N^* > \lfloor \frac{1}{h\kappa^2} \rfloor + 1$ and $\lfloor \frac{1}{h\kappa^2} \rfloor + 1 \le i \le N^*$. (Case $N^* = \lfloor \frac{1}{h\kappa^2} \rfloor + 1$ is just like (2.71).) From (2.71) and Lemma 2.5.8 we deduce that

$$2h \sum_{k=0}^{i} \mu_{N^*-k} \le \left(\frac{2}{\kappa} + \kappa \right) + \kappa + 2h \sum_{k=\lfloor \frac{1}{h\kappa^2} \rfloor + 2}^{i} \frac{1}{hk} + \frac{1/\kappa}{(hk)^{3/2}} \le$$

$$\left(\frac{2}{\kappa} + 2\kappa \right) + 2h \int_{\lfloor \frac{1}{h\kappa^2} \rfloor + 1}^{i} \left(\frac{1}{hx} + \frac{1/\kappa}{(hx)^{3/2}} \right) \mathrm{d}x \le$$

$$\left(\frac{2}{\kappa} + 2\kappa \right) + 2h \int_{\frac{1}{h\kappa^2}}^{i} \left(\frac{1}{hx} + \frac{1/\kappa}{(hx)^{3/2}} \right) \mathrm{d}x =$$

$$\left(\frac{2}{\kappa} + 2\kappa \right) + 4 - \frac{4}{\kappa\sqrt{h\,i}} + 4\ln\kappa + 2\ln(h\,i) \le \left(\frac{2}{\kappa} + 2\kappa + 4 + 4\ln\kappa \right) + 2\ln(h\,i),$$

proving (2.72). (We remark that keeping the term $-\frac{4}{\kappa\sqrt{h\,i}}$ would not make (2.72) any sharper, since $-4 \le -\frac{4}{\kappa\sqrt{h\,i}} \le 0$.)

Finally, for (2.73) use $N^* \le \frac{4}{h\sqrt{\alpha}}$ from Lemma 2.5.6. Then it suffices to turn to the weaker estimate (2.58) to get that

$$h \sum_{k=0}^{N^*} (\mu_{N^*-k})^{1+\delta} \le h \sum_{k=0}^{\lfloor \frac{1}{h\kappa} \rfloor + 1} \kappa^{1+\delta} + h \sum_{k=\lfloor \frac{1}{h\kappa} \rfloor + 2}^{N^*} \left(\frac{2}{hk} \right)^{1+\delta} \le$$

$$\kappa^{1+\delta} h \left(\frac{1}{h\kappa} + 2 \right) + h \int_{\frac{1}{h\kappa}}^{\frac{4}{h\sqrt{\alpha}}} \left(\frac{2}{hx} \right)^{1+\delta} \mathrm{d}x =$$

$$\left(\kappa^{\delta} + 2h\kappa^{1+\delta} \right) + \frac{2^{1+\delta}\kappa^{\delta}}{\delta} - \frac{2^{1-\delta}\alpha^{\delta/2}}{\delta} \le \kappa^{\delta} + \kappa^{1+\delta} + \frac{2^{1+\delta}\kappa^{\delta}}{\delta},$$

completing the proof.

It is seen that the choices for the constants $\mathrm{const}_1(\kappa) := \frac{2}{\kappa} + \kappa$, $\mathrm{const}_2(\kappa) := \frac{2}{\kappa} + 2\kappa + \frac{1}{10}$ (due to $4 + 4\ln\kappa \le 4 + 4\ln\frac{3}{8} < \frac{1}{10}$) and $\mathrm{const}_3(\delta,\kappa) := \kappa^\delta + \kappa^{1+\delta} + \frac{2^{1+\delta}\kappa^\delta}{\delta}$ are appropriate. ■

Now let us examine the product $\prod_{j=i+1}^{N^*} D_j$ in (2.70) for $i \in \{1, 2, \ldots, N^* - 1\}$. Computing $D_j \equiv (\mathcal{N}_\Phi^E)'(\mu_j)$ and using $1 + x \le e^x$ $(x \in \mathbb{R})$, we get that

$$\prod_{j=i+1}^{N^*} D_j \le \exp\left(2h\sum_{j=i+1}^{N^*}\mu_j + 3hK\sum_{j=i+1}^{N^*}\mu_j^2 + hK\sum_{j=i+1}^{N^*}\mu_j^3\right),$$

but taking into account (2.73), the right hand side can be simplified further to get

$$\prod_{j=i+1}^{N^*} D_j \le \mathrm{const}_4 \cdot \exp\left(2h\sum_{j=i+1}^{N^*}\mu_j\right), \tag{2.74}$$

with a suitable positive constant $\mathrm{const}_4 > 0$, uniformly in h and α.

Using the value of $\mathrm{const}_3(\delta,\kappa)$ set at the very end of the proof of Lemma 2.6.2, $\kappa \le 1$, $h_0 \le \frac{1}{3}$ and $\kappa K \le \frac{1}{13}$ we see that

$$3hK\sum_{j=i+1}^{N^*}\mu_j^2 + hK\sum_{j=i+1}^{N^*}\mu_j^3 \le 3hK\mathrm{const}_3(1,\kappa) + hK\mathrm{const}_3(2,\kappa) \le$$

$$3hK(5\kappa+\kappa^2) + hK(5\kappa^2+\kappa^3) \le (5\kappa+\kappa^2)\,4hK \le 24h_0\kappa K \le \frac{8}{13},$$

hence $e^{8/13} < 2 =: \mathrm{const}_4$ is a possible choice.

Substituting this into (2.70), we arrive at the estimate

$$\sup_{[0,q_{N^*+1}]} |id - J^E| \le 2c \cdot h^p\left(h\sum_{i=1}^{N^*}\exp\left(2h\sum_{j=i+1}^{N^*}\mu_j\right)\cdot q_i^\omega\right), \tag{2.75}$$

where c is the same as in (2.65).

Remark 2.6.2 Since $e^{x-x^2/2} \le 1 + x \le e^x$ $(x \in \mathbb{R}^+)$, it is seen that

$$\mathrm{const}\cdot\exp\left(2h\sum_{j=i+1}^{N^*}\mu_j\right) \le \prod_{j=i+1}^{N^*} D_j \le 2\exp\left(2h\sum_{j=i+1}^{N^*}\mu_j\right)$$

also holds with a suitable uniform constant $\mathrm{const} > 0$.

Now we are prepared to prove the following Theorem.

Theorem 2.6.3 *Suppose that κ has been defined as in Lemma 2.5.1. Suppose further, that $h_0 \le \min\left(\frac{1}{3}, \sqrt[p]{\frac{\exp(-2/\kappa)}{128c}}\right)$ and $0 < \alpha_0 \le \kappa^2$, with $c > 0$ being the same as in (2.65). If $\omega = 4$ in (2.65), then the conjugacy J^E defined between (2.63) and (2.64) satisfies*

$$\sup_{[0,\kappa/4]} |id - J^E| \le \left(12\,c\,e^{2/\kappa}\right)h^p, \tag{2.76}$$

uniformly in $h \in (0, h_0]$ and $\alpha \in (0, \alpha_0]$.

Proof. The choice of h_0 and α_0 satisfy the assumptions of all Lemmas listed so far.

First, by using $q_0 \equiv 0$ and reindexing the sums, we show that the complicated part of (2.75)

$$h\sum_{i=0}^{N^*-1}\exp\left(2h\sum_{j=i+1}^{N^*}\mu_j\right)\cdot q_i^4 \equiv h\sum_{i=1}^{N^*}\exp\left(2h\sum_{j=0}^{i-1}\mu_{N^*-j}\right)\cdot q_{N^*-i}^4 \le$$

$$h\sum_{i=1}^{N^*}\exp\left(2h\sum_{j=0}^{i}\mu_{N^*-j}\right)\cdot q_{N^*-i}^4$$

is uniformly bounded. Applying (2.71), the trivial estimate $q_{N^*-i} \le \kappa$, (2.72) and (2.58), further inequalities $\mathrm{const}_2(\kappa) \ge \mathrm{const}_1(\kappa)$ from the end of the proof of Lemma 2.6.2 and $N^* \le \frac{4}{h\sqrt{\alpha}}$ from Lemma 2.5.6, we have that

$$h\sum_{i=1}^{N^*}\exp\left(2h\sum_{j=0}^{i}\mu_{N^*-j}\right)\cdot q_{N^*-i}^4 =$$

$$h\left(\sum_{i=1}^{\lfloor\frac{1}{h\kappa^2}\rfloor} + \sum_{i=\lfloor\frac{1}{h\kappa^2}\rfloor+1}^{N^*}\right)\exp\left(2h\sum_{j=0}^{i}\mu_{N^*-j}\right)\cdot q_{N^*-i}^4 \le$$

$$h\sum_{i=1}^{\lfloor\frac{1}{h\kappa^2}\rfloor} e^{\mathrm{const}_1(\kappa)}\cdot\kappa^4 + h\sum_{i=\lfloor\frac{1}{h\kappa^2}\rfloor+1}^{N^*} e^{\mathrm{const}_2(\kappa)+2\ln(h\,i)}\cdot\left(\frac{2}{h\,i}\right)^4 \le$$

$$e^{\mathrm{const}_2(\kappa)}\left(h\kappa^4\frac{1}{h\kappa^2} + h\sum_{i=\lfloor\frac{1}{h\kappa^2}\rfloor+1}^{N^*} h^2 i^2\frac{16}{h^4 i^4}\right) =$$

$$e^{\mathrm{const}_2(\kappa)}\left(\kappa^2 + \frac{16}{h\left(\lfloor\frac{1}{h\kappa^2}\rfloor+1\right)^2} + 16\sum_{i=\lfloor\frac{1}{h\kappa^2}\rfloor+2}^{N^*}\frac{1}{hi^2}\right) \le$$

$$e^{\mathrm{const}_2(\kappa)}\left(\kappa^2 + 16h\kappa^4 + 16\int_{\frac{1}{h\kappa^2}}^{\frac{4}{h\sqrt{\alpha}}}\frac{1}{hx^2}\mathrm{d}x\right) =$$

$$e^{\mathrm{const}_2(\kappa)}\left(\kappa^2 + 16h\kappa^4 - 4\sqrt{\alpha} + 16\kappa^2\right) \le e^{\mathrm{const}_2(\kappa)}\left(17\kappa^2 + 16h_0\kappa^4\right) \le$$

$$\kappa\, e^{2/\kappa+2\kappa+1/10}\left(17\kappa + \frac{16}{3}\kappa^3\right) < \kappa\, e^{2/\kappa}\cdot 16.$$

Hence we have proved so far that

$$\sup_{[0,q_{N^*+1}]} |\,id - J^E| \le 32c\kappa\, e^{2/\kappa}\cdot h^p \le 12c\, e^{2/\kappa}\cdot h^p, \tag{2.77}$$

uniformly in $h \in (0, h_0]$ and $\alpha \in (0, \alpha_0]$.

Finally we show, that the interval on which the supremum is taken is uniformly large. There are two possibilities: $N^* = N_q$ or $N^* = N_p$. In the first case, Lemma 2.5.3 applied to the sequence q_n (with its own cutting index) yields that $[0, q_{N^*+1}] \supset [0, q_{N_q}] \supset [0, \kappa/2]$. In the second case however, when $N^* = N_p$, we can turn to the left inequality in (2.77) *itself* together with the fact that $J^E(q_{N^*}) = p_{N^*}$ by construction, to establish relation

$$|q_{N_p} - p_{N_p}| = |q_{N_p} - J^E(q_{N_p})| \leq \sup_{[0,q_{N^*+1}]} |\,id - J^E| \leq 32c\kappa\, e^{2/\kappa} \cdot h_0^p \leq \frac{\kappa}{4},$$

if $h_0 \leq \sqrt[p]{\frac{\exp(-2/\kappa)}{128c}}$. But the result of Lemma 2.5.3 is again that $p_{N_p} \geq \frac{\kappa}{2}$, so $q_{N_p} \geq \frac{\kappa}{4}$ must be true. Therefore $[0, q_{N^*+1}] \supset [0, q_{N_p}] \supset [0, \kappa/4]$ and the Theorem is proved. ■

Remark 2.6.3 Since, by definition, $\kappa \leq \frac{1}{13K}$, that is $26K \leq \frac{2}{\kappa}$, we see that if the common uniform bound K in the mappings (2.63) and (2.64) is increased, then the upper estimate (2.76) *and* the upper bounds on h_0 and α_0 become worse.

For the case $\omega = 3$ in (2.65), the situation seems to be not so "uniform".

Theorem 2.6.4 *Suppose that κ has been defined as in Lemma 2.5.1. Suppose further, that $h_0 \leq \frac{1}{3}$, $0 < \alpha_0 \leq \kappa^2$, and $c > 0$ is the same as in (2.65). If $\omega = 3$ in (2.65), then the conjugacy J^E defined between (2.63) and (2.64) satisfies*

$$\sup_{[0,q_{N^*+1}]} |\,id - J^E| \leq c\left(\text{const}_5(\kappa) + \text{const}_6(\kappa)\ln\frac{1}{\alpha}\right) \cdot h^p. \tag{2.78}$$

Proof. Estimate (2.75) will be used with $\omega = 3$. We apply the same type of manipulations as in the proof of Theorem 2.6.3 to get

$$h\sum_{i=1}^{N^*} \exp\left(2h\sum_{j=i+1}^{N^*} \mu_j\right) \cdot q_i^3 \leq h\sum_{i=1}^{N^*} \exp\left(2h\sum_{j=0}^{i} \mu_{N^*-j}\right) \cdot q_{N^*-i}^3$$

$$h\sum_{i=1}^{\lfloor \frac{1}{h\kappa^2}\rfloor} e^{\text{const}_1(\kappa)} \cdot \kappa^3 + h\sum_{i=\lfloor \frac{1}{h\kappa^2}\rfloor+1}^{N^*} e^{\text{const}_2(\kappa)+2\ln(h\,i)} \cdot \left(\frac{2}{h\,i}\right)^3 \leq$$

$$e^{\text{const}_2(\kappa)}\left(h\kappa^3\frac{1}{h\kappa^2} + h\sum_{i=\lfloor \frac{1}{h\kappa^2}\rfloor+1}^{N^*} h^2 i^2 \frac{8}{h^3 i^3}\right) =$$

$$e^{\text{const}_2(\kappa)}\left(\kappa + \frac{8}{\lfloor \frac{1}{h\kappa^2}\rfloor + 1} + 8\sum_{i=\lfloor \frac{1}{h\kappa^2}\rfloor+2}^{N^*} \frac{1}{i}\right) \leq$$

$$e^{\text{const}_2(\kappa)}\left(\kappa + 8h\kappa^2 + 8\int_{\frac{1}{h\kappa^2}}^{\frac{4}{h\sqrt{\alpha}}} \frac{1}{x}\,\mathrm{d}x\right) =$$

$$e^{\text{const}_2(\kappa)}\left(\kappa + 8h\kappa^2 + 8\ln 4 + 16\ln\kappa + 4\ln\frac{1}{\alpha}\right). \quad ■$$

Remark 2.6.4 Unfortunately, estimate (2.78) is singular as $\alpha \to 0^+$. Besides this, we can not control the interval $[0, q_{N^*}]$ in the supremum, so it may shrink too much if $N^* = N_p$ as $\alpha \to 0^+$.

A positive result in the $\omega = 3$ case we have is that on a special shrinking domain, namely on a parabola-shaped domain in the (α, x)-plane, a better closeness result holds.

Theorem 2.6.5 *Suppose that κ has been defined as in Lemma 2.5.1. Suppose further, that $h_0 \le \min\left(\frac{1}{3}, \sqrt[p]{\frac{1}{16ce^2\kappa}}\right)$ and $0 < \alpha_0 \le \kappa^2$, with $c > 0$ being the same as in (2.65). If $\omega = 3$ in (2.65), then the conjugacy J^E defined between (2.63) and (2.64) satisfies*

$$\sup_{[0,\sqrt{\alpha}/4]} |\, id - J^E| \le \left(4ce^2\alpha\right) h^p.$$

Proof. Similarly to the cutting indices N_q, N_p and N^*, let us define $N_q(\sqrt{\alpha})$ to be the index such that $q_{N_q(\sqrt{\alpha})} \le \sqrt{\alpha}$, but $q_{N_q(\sqrt{\alpha})+1} > \sqrt{\alpha}$. Let us denote by $N_p(\sqrt{\alpha})$ the corresponding index for the sequence p_n. Further, let $N^*(\sqrt{\alpha}) := \min(N_q(\sqrt{\alpha}), N_p(\sqrt{\alpha}))$. Then it is easy to reconsider that all formulae (2.70)–(2.75) are still valid if N^* is replaced by this (not greater) $N^*(\sqrt{\alpha})$. So, as a starting point, we have

$$\sup_{[0,q_{N^*(\sqrt{\alpha})+1}]} |\, id - J^E| \le 2c \cdot h^p \left(h \sum_{i=1}^{N^*(\sqrt{\alpha})} \exp\left(2h \sum_{j=i+1}^{N^*(\sqrt{\alpha})} \mu_j\right) \cdot q_i^3 \right).$$

But, by the definition of $N^*(\sqrt{\alpha})$, further using Lemma 2.5.6 to get $N^*(\sqrt{\alpha}) \le \frac{1}{h\sqrt{\alpha}}$, we see that

$$h \sum_{i=1}^{N^*(\sqrt{\alpha})} \exp\left(2h \sum_{j=i+1}^{N^*(\sqrt{\alpha})} \mu_j\right) \cdot q_i^3 \le h \sum_{i=1}^{N^*(\sqrt{\alpha})} \exp\left(2h \sum_{j=2}^{N^*(\sqrt{\alpha})} \sqrt{\alpha}\right) \cdot \sqrt{\alpha}^3 \le$$

$$h\, e^2 \sum_{i=1}^{N^*(\sqrt{\alpha})} \sqrt{\alpha}^3 \le e^2\alpha \left(\frac{1}{h\sqrt{\alpha}} + 1\right) h\sqrt{\alpha} \le 2e^2\alpha.$$

Hence we know that

$$\sup_{[0,q_{N^*(\sqrt{\alpha})+1}]} |\, id - J^E| \le 4ce^2\alpha \cdot h^p. \tag{2.79}$$

Now, similarly to the end of the proof of Theorem 2.6.3, we show that the domain of the supremum contains $[0, \sqrt{\alpha}/4]$, uniformly in $h \in (0, h_0]$. If $N^*(\sqrt{\alpha}) = N_q(\sqrt{\alpha})$, then by Lemma 2.5.4 with $m = N_q(\sqrt{\alpha})$ we see that $q_m \ge \frac{\sqrt{\alpha}}{2}$, while if $N^*(\sqrt{\alpha}) = N_p(\sqrt{\alpha})$, then by (2.79)

$$|q_{N_p(\sqrt{\alpha})} - p_{N_p(\sqrt{\alpha})}| = |q_{N_p(\sqrt{\alpha})} - J^E(q_{N_p(\sqrt{\alpha})})| \le \sup_{[0,q_{N^*+1}]} |\, id - J^E| \le 4ce^2\alpha \cdot h_0^p \le \frac{\sqrt{\alpha}}{4},$$

if, for example, $h_0 \le \sqrt[p]{\frac{1}{16ce^2\kappa}}$. But again, by Lemma 2.5.4 with $m = N_p(\sqrt{\alpha})$, $p_m \ge \frac{\sqrt{\alpha}}{2}$, so $q_m \ge \frac{\sqrt{\alpha}}{4}$. ■

Remark 2.6.5 We have tacitly assumed (especially in Lemma 2.6.2) that $N^* \ge \frac{1}{h\kappa^2}$. However, this is not a real restriction, since otherwise every estimate is *ab ovo* trivial—just as the proof of (2.71) in Lemma 2.6.2—and we would have uniform boundedness in the Theorems.

2.7 Further results for the $\alpha > 0$ case

2.7.1 The tangent estimate

Although the following nice proposition finally has not been used in the closeness estimates, we still include it, because it reveals some information about the behaviour of the direct iteration p_n, and, together with the subsequent remarks, served as a motivation for the "backward" approach. (The number N, of course, is N_p here.)

Proposition 2.7.1 (The Tan-Estimate) *Suppose that the conditions of Lemma 2.5.2 hold and* $0 < \alpha_0 \le \kappa$. *Then for* $0 \le n < \min\left(N, \frac{\pi}{2h\sqrt{2\alpha}} - 1\right)$ *we have*

$$p_n \le \sqrt{\frac{\alpha}{2}}\tan(\sqrt{2\alpha}\,hn).$$

Proof. We prove by induction on n. The case $n = 0$ is trivial. If $n = 1$, then

$$p_1 \equiv h\alpha \le \sqrt{\frac{\alpha}{2}}\tan(\sqrt{2\alpha}\,h)$$

is equivalent to $\sqrt{2\alpha}\,h \le \tan(\sqrt{2\alpha}\,h)$, but this latter is true since $x \le \tan x$ (if, *e.g.*, $0 \le x \le 1$), and $\sqrt{2\alpha}\,h \le \sqrt{2\alpha_0}h_0 \le \sqrt{2\kappa}h_0 \le \sqrt{2\cdot\frac{3}{8}\cdot\frac{1}{3}} < 1$.

So suppose the induction hypothesis is true for some $n \ge 1$. Then

$$p_{n+1} \le h\alpha + \sqrt{\frac{\alpha}{2}}\tan(\sqrt{2\alpha}\,hn) + 2h\left(\sqrt{\frac{\alpha}{2}}\tan(\sqrt{2\alpha}\,hn)\right)^2$$

holds, since $p_{n+1} = \mathcal{N}_\eta(h, p_n, \alpha) \le h\alpha + p_n + 2hp_n^2$, if, *e.g.*, $0 \le p_n \le \kappa \le \frac{1}{K}$ (implied by $n < N$). In order to finish the induction, it is sufficient to establish

$$h\alpha + \sqrt{\frac{\alpha}{2}}\tan(\sqrt{2\alpha}\,hn) + 2h\left(\sqrt{\frac{\alpha}{2}}\tan(\sqrt{2\alpha}\,hn)\right)^2 \le \sqrt{\frac{\alpha}{2}}\tan(\sqrt{2\alpha}\,h(n+1)).$$

By using the abbreviation $x := \sqrt{2\alpha}\,h$, the inequality above can be rewritten as

$$x + \tan(nx) + x\tan^2(nx) \le \tan((n+1)x). \tag{2.80}$$

Since $n < \frac{\pi}{2x} - 1$ by assumption, we know that

$$0 < \tan((n+1)x) = \frac{\tan x + \tan(nx)}{1 - \tan x \cdot \tan(nx)}.$$

But due to $0 < x < 1$ and $n < \frac{\pi}{2x}$, both $\tan x$ and $\tan(nx)$ are positive, so the denominator above is also positive. Hence, instead of (2.80), it is enough to prove

$$\left(x + \tan(nx) + x\tan^2(nx)\right)(1 - \tan x \cdot \tan(nx)) - \tan x - \tan(nx) \le 0.$$

However, the left hand side can be factored to get

$$-\left(1 + \tan^2(nx)\right)(\tan x - x + x\tan x \cdot \tan(nx)),$$

so it must be nonpositive, because $\tan x - x \ge 0$ and $x, \tan x, \tan(nx) \ge 0$. ■

Remark 2.7.1 The tangent estimate, *i.e.* the function $t \mapsto \sqrt{\frac{\alpha}{2}}\tan(\sqrt{2\alpha}\,ht)$, has been obtained as the solution of the initial value problem $\dot{X} = h\alpha + 2hX^2$, $X(0) = 0$. (Of course, the multiplying constant 2 could be replaced by $1 + \delta$, for any positive δ, but the limit $\delta \to 0^+$ is not allowed.) This ordinary differential equation has been chosen because its stepsize-1 explicit Euler discretization is just the sequence p_n with a slightly modified definition $p_{n+1} := h\alpha + p_n + 2hp_n^2$. Thus, we have proved in the Proposition above that for this particular equation the explicit Euler discretization is a lower approximation to the true solution—although, of course, from our viewpoint the roles are reversed: the known true solution is an upper estimate for the more implicit sequence p_n. The previous observation can be extended to a general class of ordinary differential equations: it can be shown that under a simple assumption on the sign of the right hand side and its derivative of the ordinary differential equation, the explicit/implicit Euler discretization is a lower/upper approximation to the exact solution, provided that the discretization stepsize is sufficiently small, see in [19]. This more general result however can not be directly applied to prove Proposition 2.7.1, because here the stepsize is 1. A fundamental and very interesting question would be to determine classes of equations where—or explain, in our case, why—the discretization is such a surprisingly sharp estimate of the true solution even with so large stepsizes.

Remark 2.7.2 The tangent estimate is "nearly global": it is a very good upper estimate of p_n as long as the tangent function is defined and not "too large". Of course, when the tangent reaches its first singularity, it becomes a useless estimate of p_n. This is exactly the main difficulty with the "direct" approach: estimating p_n as n increases is hard in the region when the tangent estimate is no longer valid but still $p_n < \kappa$ for some time.

Remark 2.7.3 We mention [33] as a peculiar result concerning the forward and backward iterates of the sequence $w_{n+1} = w_n^2 + \frac{1}{4} + \alpha$, $w_0 = \frac{1}{2}$. These recursions appear several times in the literature in connection with the phenomenon of intermittency, but probably this is the first paper containing a proof of the following observation. If $S(\alpha)$ denotes the number of steps needed for w_n to reach, say, 1, then [33] shows that $\lim_{\alpha\to 0^+} \sqrt{\alpha}\, S(\alpha) = \frac{\pi}{2}$. The calculations in the proof are elementary, but quite involved—the basic idea is to compare the difference equation with the corresponding differential equation similar to the one mentioned in Remark 2.7.1 above, and prove that the leading coefficients in the series expansion of their solutions satisfy the same type of recursive relations. This asymptotic relation in the case of p_n with $\eta \equiv 0$ in (2.53)—simply being a shifted version of w_n above—would mean that $\lim_{\alpha\to 0^+} h\sqrt{\alpha}\, N(h,\alpha) = \frac{\pi}{2}$.

2.7.2 Numerical test results

The optimality of Theorems 2.6.3–2.6.5 above—*under assumption* (2.67)—will now be illustrated by some numerical tests.

The following setting has been chosen: for $n \in \mathbb{N}^+$, let

$$q_{n+1} := h\alpha + q_n + hq_n^2$$

denote the pure quadratic iteration with $q_0 := 0$, while

$$p_{n+1} := h\alpha + p_n + hp_n^2 + \frac{1}{2}h^{p+1}p_n^{\omega}$$

with $p_0 := 0$ is a perturbed sequence. Choice of the cutting level $\kappa := \frac{3}{8}$ conforms to the requirements of Lemma 2.5.1.

What we measure in every case is the quantity

$$\text{dist} := \frac{|p_{N^*} - q_{N^*}|}{h^p}$$

under different choices of the exponents $p \in \mathbb{N}^+$ and $\omega \in \{3,4\}$, further, the parameters h and α. The quantity $\text{dist}\cdot h^p$ is clearly a numerical *lower estimate* of $\sup_{[0,q_{N^*+1}]} |\, id - J^E|$, see, *e.g.*, at the end of the proof of Theorem 2.6.3.

For the sake of comparison, we will also indicate the value of N^*. Since $p_n \geq q_n \geq 0$, we have $N^* = N_p$.

Due to its simplicity and elegance, we include the actual *Mathematica 5* code devised to perform the computations.

After fixing the values of p and ω, the following definition

```
perturbedsequence=Compile[{h,α},NestWhile[
   {h α+#[[1]]+h #[[1]]^2+1/2 h^(p+1)#[[1]]^ω,Last[#]+1}&,{0.,1},#[[1]]<3/8&]]
```

will yield $\{p_{N_p}, N_p\}$ in a list, while

```
quadraticsequence=Compile[{h,α,iternumber},Nest[
   {h α+#[[1]]+h #[[1]]^2,Last[#]+1}&,{0.,1},iternumber-1]]
```

will determine $\{q_{N_p}, N_p\}$, with `iternumber`$:=N_p$. Now—with h_1 and α_1 representing concrete numerical quantities—evaluate the following three commands

```
   perturbedsequence[h₁,α₁]
   quadraticsequence[h₁,α₁,Last[%]]
   Abs[First[%]-First[%%]]/h₁^p
```

to obtain finally the value of ”dist”.

Remark 2.7.4 The code for `perturbedsequence` and `quadraticsequence` given above uses machine precision numbers (see the `Compile` commands and the dots behind the 0's), since this substantially reduces the time needed to obtain ”dist” when α is very small. For large and medium values of α, the moderate computing time made it possible to exploit *Mathematica*'s arbitrary precision arithmetic as well. At $h = 10^{-1}$ and 10^{-2}, we have experienced total agreement between calculations based on machine precision and arbitrary precision—providing a good reliability check. However, for $h = 10^{-5}$, and $p = 3$, $\alpha = 10^{-3}$, for example, machine precision turned out to be insufficient, so arbitrary precision has been applied, since in this case $|p_{N^*} - q_{N^*}| \le 1.5 \cdot 10^{-16}$.

The actual output—together with a graphical representation some of the data—is listed below. The arrangement of these tables is explained by the fact that in this way it is a bit easier for the eye to compare pairs of α-exponents and recognize the logarithmic law.

$\omega = 3,\ p = 2,\ h = 10^{-1}$

α	10^{-1}	10^{-2}	10^{-4}	10^{-8}
dist	$1.602 \cdot 10^{-2}$	$6.814 \cdot 10^{-2}$	$2.169 \cdot 10^{-1}$	$5.239 \cdot 10^{-1}$
N^*	29	134	1549	$1.5706 \cdot 10^{5}$

α	10^{-3}	10^{-6}	10^{-9}	10^{-12}
dist	$1.397 \cdot 10^{-1}$	$3.650 \cdot 10^{-1}$	$6.066 \cdot 10^{-1}$	$8.292 \cdot 10^{-1}$
N^*	474	15688	$4.9671 \cdot 10^{5}$	$1.5708 \cdot 10^{7}$

α	10^{-5}	10^{-10}
dist	$3.061 \cdot 10^{-1}$	$6.835 \cdot 10^{-1}$
N^*	4947	$1.5708 \cdot 10^{6}$

α	10^{-7}	10^{-14}
dist	$4.748 \cdot 10^{-1}$	1.0096
N^*	49655	$1.5708 \cdot 10^{8}$

The relation between α and "dist" is illustrated graphically by the following logarithmic plot: on the horizontal axis, values of $\log_{10}\left(\frac{1}{\alpha}\right)$ are displayed against the values of "dist" on the vertical axis. For the sake of convenience, linear interpolation has been used between the discrete points.

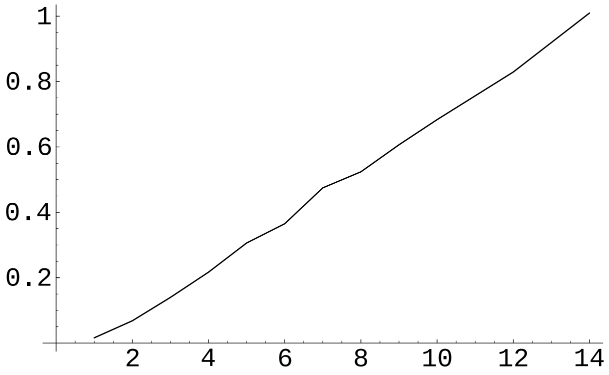


Figure 2.7.1

$\omega = 3,\ p = 3,\ h = 10^{-1}$

α	10^{-1}	10^{-2}	10^{-4}	10^{-8}
dist	$1.601 \cdot 10^{-2}$	$6.799 \cdot 10^{-2}$	$2.155 \cdot 10^{-1}$	$5.579 \cdot 10^{-1}$
N^*	29	134	1549	$1.5706 \cdot 10^{5}$

α	10^{-3}	10^{-6}	10^{-9}	10^{-12}
dist	$1.392 \cdot 10^{-1}$	$3.909 \cdot 10^{-1}$	$6.444 \cdot 10^{-1}$	$8.745 \cdot 10^{-1}$
N^*	474	15689	$4.9671 \cdot 10^{5}$	$1.5708 \cdot 10^{7}$

α	10^{-5}	10^{-10}
dist	$3.036 \cdot 10^{-1}$	$7.243 \cdot 10^{-1}$
N^*	4947	$1.5708 \cdot 10^{6}$

α	10^{-7}	10^{-14}
dist	$4.690 \cdot 10^{-1}$	1.060
N^*	49655	$1.5708 \cdot 10^{8}$

The corresponding graph is quite similar to the one above:

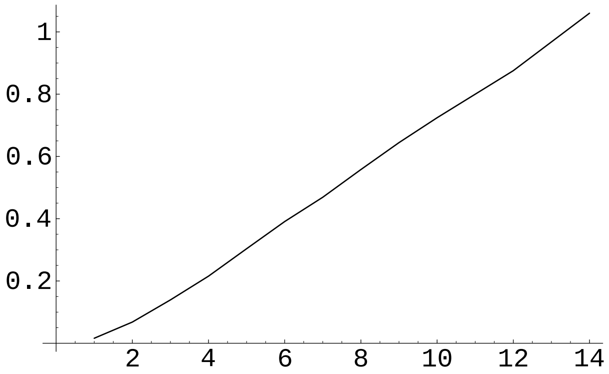


Figure 2.7.2

$\omega = 3$, $p = 3$, **varying** h

α	10^{-3}				
h	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
dist	$1.392 \cdot 10^{-1}$	$1.399 \cdot 10^{-1}$	$1.403 \cdot 10^{-1}$	$1.402 \cdot 10^{-1}$	$1.402 \cdot 10^{-1}$
N^*	474	4705	47017	$4.701 \cdot 10^5$	$4.701 \cdot 10^6$

α	10^{-4}			
h	10^{-1}	10^{-2}	10^{-3}	10^{-4}
dist	$2.155 \cdot 10^{-1}$	$2.189 \cdot 10^{-1}$	$2.199 \cdot 10^{-1}$	$2.198 \cdot 10^{-1}$
N^*	1549	15446	154419	$1.5441 \cdot 10^6$

α	10^{-5}			
h	10^{-1}	10^{-2}	10^{-3}	10^{-4}
dist	$3.036 \cdot 10^{-1}$	$3.017 \cdot 10^{-1}$	$3.006 \cdot 10^{-1}$	$3.006 \cdot 10^{-1}$
N^*	4947	49413	$4.9407 \cdot 10^6$	$4.9406 \cdot 10^7$

$\omega = 4$, $p = 2$, $h = 10^{-1}$

α	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
dist	$2.145 \cdot 10^{-2}$	$2.412 \cdot 10^{-2}$	$2.619 \cdot 10^{-2}$	$2.708 \cdot 10^{-2}$	$2.681 \cdot 10^{-2}$
N^*	474	1549	4947	15689	49655

α	10^{-8}	10^{-9}	10^{-10}	10^{-11}	10^{-12}
dist	$2.731 \cdot 10^{-2}$	$2.753 \cdot 10^{-2}$	$2.734 \cdot 10^{-2}$	$2.746 \cdot 10^{-2}$	$2.662 \cdot 10^{-2}$
N^*	157063	496714	$1.5708 \cdot 10^6$	$4.9672 \cdot 10^6$	$1.5708 \cdot 10^7$

The logarithmic plot this time is completely different:

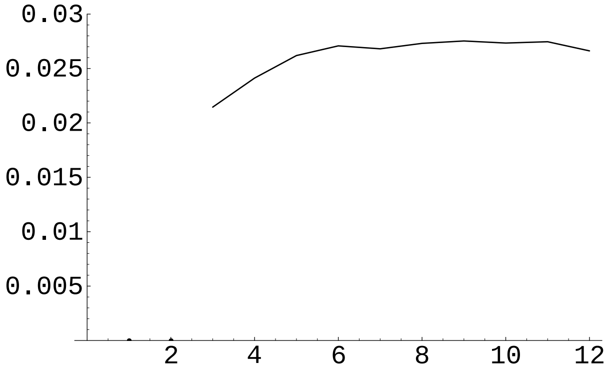


Figure 2.7.3

2.7.3 Conclusions of the numerical tests

Case $\omega = 3$. From the *linear* graphs of Figure 2.7.1 and 2.7.2, it is seen that the quantity "dist" grows like const $\cdot \ln \frac{1}{\alpha}$, as $\alpha \to 0^+$. Since dist $\cdot\, h^p$ is a numerical lower estimate of $\sup_{[0,q_{N^*+1}]} |\, id - J^E|$, this numerical evidence—together with the right hand side of estimate (2.78)—gives a convincing argument that *if the crucial (but natural) assumption $J^E(0) := 0$ is made, then the distance of the constructed conjugacy and the identity map indeed shows a logarithmic singularity as $\alpha \to 0^+$.*

Further, "dist" seems to be more or less independent of p, as Figure 2.7.1 and 2.7.2 are nearly the same, moreover, values of "dist" show stabilization as $h \to 0^+$ and $\alpha > 0$ is fixed.

Case $\omega = 4$. Numerical results together with Figure 2.7.3 clearly show uniform boundedness of "dist", which, of course, has been proved in Theorem 2.6.3.

In all cases, the values of N^* very closely follow the asymptotic formula $N^* \approx \frac{\pi}{2h\sqrt{\alpha}} \approx \frac{1.5708}{h\sqrt{\alpha}}$ $(\alpha \to 0^+)$ stated in Remark 2.7.3 in Section 2.7.1.

2.7.4 Open questions

1. Attempts to transform the normal forms further. It can be asked whether it is possible to eliminate the cubic term in (2.63) (or in (2.64)). For simplicity, set $\widehat{\eta}_3 \equiv a \in \mathbb{R}$, $h = 1$ and $\alpha = 0$. Then we aim to find a near-identity transform *trans* : $x \mapsto x + bx^\nu$ with suitable b and ν such that it brings our mapping *map* : $x \mapsto x + x^2 + ax^3$ into a mapping with the cubic term eliminated. In other words, we would like to find b and ν such that *elimmap* $:=$ *trans*$^{[-1]} \circ$ *map* $\circ$ *trans* contains no cubic terms.

The actual computations were performed again in *Mathematica.* If the value of ν is set, then the following command computes the series expansion of *elimmap* about the origin up to order 10:

```
ComposeSeries[InverseSeries[x+bx^ν+O[x]^10],x+x^2+ax^3+O[x]^10,
                x+bx^ν+O[x]^10]//Simplify
```

Substituting different values of ν into the above expression, the following pattern emerges. With $2 \le \nu \in \mathbb{N}$ set, it is possible to choose b (also depending on a) such that *elimmap* contains no terms of order 2ν. This suggests trying *Puiseaux-series* instead of *Taylor-series*, however,

$\nu = \frac{3}{2}$ leads to *elimmap* $\equiv x + x^2 + \frac{1}{2}bx^{5/2} + \left(a + \frac{b^2}{4}\right)x^3 + \ldots$, so an unwanted term of order $\frac{5}{2}$ enters. It was also in vain to try *trans* $\equiv x + bx^\nu + cx^\mu$ with various choices of ν and μ.

Therefore, we conclude that—at least with these type of transformations—it does not seem to be possible to convert the general $\omega = 3$ case into the $\omega = 4$ case.

2. The other question is the continuity of the conjugacy mapping. In our construction, we have assured that $x \mapsto J(h, x, \alpha)$ is a homeomorphism, for every fixed h and α. Continuity of $h \mapsto J(h, x, \alpha)$ $(0 < h \le h_0)$ is also seen to hold. However, the map $\alpha \mapsto J(h, x, \alpha)$ does not seem to be continuous at the critical bifurcation value $\alpha = 0$ and $x \ge 0$. The reason for this discrepancy at $\alpha = 0$, $x \ge 0$ is that while in the fixed point-free $\alpha > 0$ case the conjugacy equation (2.66) extends $J(h, \cdot, \alpha)$ to the whole $[-\varepsilon_0, \varepsilon_0]$ interval if it is defined on *one* fundamental domain, in the case of $\alpha = 0$—due to the presence of the fixed point at $x = 0$—two fundamental domains are needed on each half-line. Regarding this continuity problem, among others, see the next Section.

2.8 An improved approach in the $\alpha > 0$ case: the grid construction

In this section we present a modified construction which will enable us to prove $\mathcal{O}(h^p)$-closeness between the conjugacy and the identity on some set of grid points. For any $h > 0$ and $\alpha > 0$ $(0 < h \le h_0, 0 < \alpha \le \alpha_0)$, the conjugacy $J(h, \cdot, \alpha)$ will be defined on a domain containing the interval $[-\kappa, \kappa]$, where $\kappa > 0$ (the cutting level) is sufficiently small and independent of h and α.

Let us define two sequences. For any $h > 0$ and $\alpha > 0$ set

$$x_0(h, \alpha) := y_0(h, \alpha) := -\kappa,$$

and for $n = 0, 1, 2, \ldots$, define

$$x_{n+1}(h, \alpha) := \mathcal{N}_\varphi(h, x_n(h, \alpha), \alpha), \qquad y_{n+1}(h, \alpha) := \mathcal{N}_\Phi(h, y_n(h, \alpha), \alpha).$$

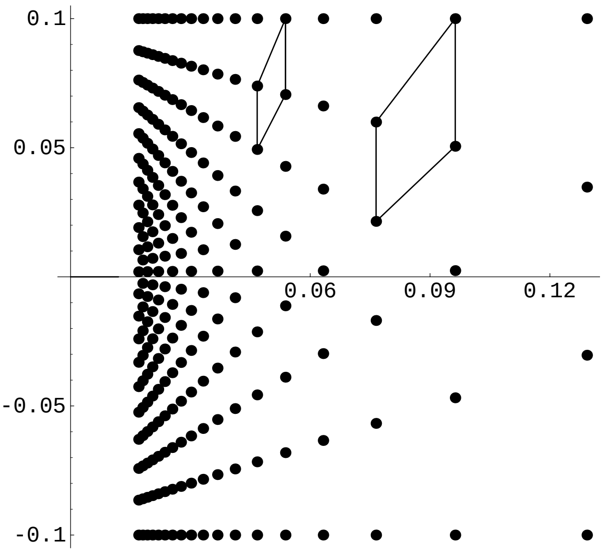
We can assume that both normal forms have bounded derivatives with respect to α also. Then the functions $\alpha \mapsto \mathcal{N}_\varphi(h, x, \alpha)$ and $\alpha \mapsto \mathcal{N}_\Phi(h, x, \alpha)$ are strictly monotone increasing for small $\alpha > 0$, $|x|$ and $h > 0$, so there exist two sequences, $\alpha_N(h)$ and $\beta_N(h)$ $(N \in \mathbb{N}^+)$, converging to 0^+ as $N \to +\infty$ and h is fixed, such that

$$x_N(h, \alpha_N(h)) = \kappa = y_N(h, \beta_N(h)),$$

see also Figure 2.6.2. Thus, for any $h > 0$, we have defined two sets of grid points

$$(\alpha_N(h), x_n(h, \alpha_N(h))) \quad \text{and} \quad (\beta_N(h), y_n(h, \beta_N(h))) \quad (n \in \{0, 1, \ldots, N\}, N \in \mathbb{N}^+).$$

Notice however that for any fixed positive integer N, $\alpha_N(h)$ and $\beta_N(h)$ tend to infinity as $h \to 0^+$, so in order to keep these grid points within $(0, \alpha_0] \times [-\kappa, \kappa]$, it is enough to consider $\alpha_N(h)$ and $\beta_N(h)$ only for $N \ge N_0(h)$, with suitable positive integers $N_0(h)$.



Some grid points, grid lines and domains of the α-grid

Now let us make a natural correspondence between these two sets of grid points. For any $h > 0$, define the conjugacy mapping on the "α-grid points" for $N \in \mathbb{N}^+, N \geq N_0(h), n = 0, 1, \ldots, N$ as

$$J(h, x_n(h, \alpha_N(h)), \alpha_N(h)) := y_n(h, \beta_N(h)).$$

Our aim is to show that corresponding grid points are $\mathcal{O}(h^p)$-close to each other, that is, it is possible to eliminate the $\ln \frac{1}{\alpha}$ term from the closeness estimates on the grid points. This improvement is made possible by imposing the above "two-sided boundary condition" $x_0(\alpha_N) = y_0(\beta_N)$ and $x_N(\alpha_N) = y_N(\beta_N)$ on the defining sequences x_n and y_n. In other words, we will prove that for each $\mathcal{N}_\varphi$-orbit we can find a uniformly close $\mathcal{N}_\Phi$-orbit by "synchronization" using a suitable shift in the parameter α also. (We remark that an $\mathcal{O}(h^p)$-shift in α has already been introduced, when the normal forms $\mathcal{N}_\Phi$ and $\mathcal{N}_\varphi$ were derived via inverse and implicit function theorems.) As opposed to this, the previous construction (2.67) with sequences p_n and q_n contained only a "one-sided boundary condition" (the direct normal forms were iterated starting simultaneously from 0 until the faster sequence reached level κ) with "loose upper ends", which—according to our estimates and numerical test results—necessarily drew away logarithmically from each other near κ. It turns out that if *both* sides of the sequences x_n and y_n are kept fixed, a symmetry argument becomes feasible, where direct iterates are considered from $-\kappa$ to 0 upward, but inverse ones from κ to 0 downward. Taking this approach balances out the *absence* of fixed points. (We remark that construction of the conjugacy in the $\alpha \leq 0$ case can also be considered as a construction with two-sided boundary condition: on one side we let the two orbits run from the same starting point, while the *presence* of the fixed point forces an "asymptotic boundary condition" on the other side. In that situation the same α-value could be used.)

Since we will have to keep track of the positive constants appearing in the inequalities carefully, let $c > 0$ again denote a particular, but henceforth fixed constant for which $|\mathcal{N}_\Phi(h, x, \alpha) - \mathcal{N}_\varphi(h, x, \alpha)| \leq c \cdot h^{p+1}|x|^3$ is uniformly true, further, as before, let $K > 0$ denote a uniform bound on the derivatives of $\widehat{\eta}_3$ and $\widetilde{\eta}_3$ near the origin, as in (2.53). Other positive constants will be denoted by indexed letters c_i. A general positive constant is denoted by *const*. All of these numbers are independent of the parameters h and α.

In order to keep notation simple, we fix some $h > 0$ and $N \geq N_0(h)$ and write x_n and y_n instead of $x_n(h, \alpha_N(h))$ and $y_n(h, \beta_N(h))$. Then

$$y_{n+1} - x_{n+1} = \mathcal{N}_\Phi(h, y_n, \beta) - \mathcal{N}_\Phi(h, x_n, \beta) + \tag{2.81}$$

$$\mathcal{N}_\Phi(h, x_n, \beta) - \mathcal{N}_\Phi(h, x_n, \alpha) + \mathcal{N}_\Phi(h, x_n, \alpha) - \mathcal{N}_\varphi(h, x_n, \alpha) =$$

$$(y_n - x_n) \cdot \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_n, \beta) + (\beta - \alpha) \cdot \frac{\mathrm{d}}{\mathrm{d}\alpha}\mathcal{N}_\Phi(h, x_n, \gamma) + hx_n^3(\widehat{\eta}_3(h, x_n, \alpha) - \widetilde{\eta}_3(h, x_n, \alpha))$$

for any $n = 0, 1, 2, \ldots, N-1$, with some $\xi_n(h, N) \equiv \xi_n \in [\{x_n, y_n\}]$ and $\gamma(h, N) \equiv \gamma \in [\{\alpha_N, \beta_N\}]$. Since $\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathcal{N}_\Phi(h, x_n, \gamma) = h \cdot (1 + x_n^3 \cdot \frac{\mathrm{d}}{\mathrm{d}\alpha}\widehat{\eta}_3(h, x_n, \gamma)) =: h \cdot A_n(h, N)$, if we choose $\kappa > 0$ so small, such that

$$\kappa \leq \min\left(1, \frac{1}{2K}\right),$$

then

$$h \cdot A_n \in \left(\frac{h}{2}, 2h\right).$$

Further, if $\kappa \leq 1$ and $h_0 \leq \frac{1}{12K}$, then

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_n, \beta) \in \left(\frac{1}{2}, 2\right)$$

uniformly.

Now we prove that the shift in the parameter α is again only $\mathcal{O}(h^p)$, that is, corresponding vertical grid lines are $\mathcal{O}(h^p)$-close to each other.

Lemma 2.8.1 *For every $N \in \mathbb{N}^+$ and $h \in (0, h_0]$ we have that*

$$|\beta_N(h) - \alpha_N(h)| \leq 2ch^p.$$

Proof. By inductively applying (2.81), our omnipresent discrete Gronwall estimate, now with an extra term containing $\beta_N - \alpha_N \equiv \beta - \alpha$, reads as

$$y_n - x_n = \tag{2.82}$$

$$\sum_{i=0}^{n-1} \left[h(\beta - \alpha)A_i + hx_i^3(\widehat{\eta}_3(h, x_i, \alpha) - \widetilde{\eta}_3(h, x_i, \alpha))\right] \prod_{j=i+1}^{n-1} \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_j, \beta)$$

for $n = 1, 2, \ldots, N$, where, as usual, $\prod_{j=n}^{n-1}(\cdot) := 1$. If we substitute $n = N$ into (2.82) and use that $y_N - x_N = 0$, we can express the desired quantity explicitly as

$$|\beta - \alpha| =$$

$$\left| - \frac{\sum_{i=0}^{N-1}\left[hx_i^3(\widehat{\eta}_3(h, x_i, \alpha) - \widetilde{\eta}_3(h, x_i, \alpha))\right] \prod_{j=i+1}^{N-1} \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_j, \beta)}{\sum_{i=0}^{N-1} h\, A_i \prod_{j=i+1}^{N-1} \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_j, \beta)} \right| \leq$$

$$\frac{h\kappa^3 \cdot ch^p \sum_{i=0}^{N-1} \prod_{j=i+1}^{N-1} \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_j, \beta)}{\frac{1}{2}h \sum_{i=0}^{N-1} \prod_{j=i+1}^{N-1} \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_j, \beta)} \leq 2ch^p. \quad \blacksquare$$

Second proof. We may assume that, say, $\beta \geq \alpha$. Suppose, to the contrary, that $\beta - \alpha \geq 3ch^p$. Then, by using (2.81) and $y_0 - x_0 = 0$, we have inductively for $n = 0, 1, \dots, N-1$ that

$$y_{n+1} - x_{n+1} \geq (y_n - x_n) \cdot \frac{1}{2} + 3ch^p \cdot \frac{h}{2} - h\kappa^3 ch^p \geq 0 + \frac{3}{2}ch^{p+1} - ch^{p+1} \geq \frac{1}{2}ch^{p+1}.$$

But then $y_N - x_N \geq \frac{1}{2}ch^{p+1}$ would contradict to the definition $y_N - x_N = 0$. (The case $\alpha \geq \beta$ is symmetric, because then $\alpha - \beta \geq 3ch^p$ would lead to a contradiction $0 = x_N - y_N > 0$.) ■

Since we are going to utilize some estimates from previous sections, we now collect the relevant results here. It is easy to see that if, for example, $\kappa \leq \min(1, \frac{1}{2K})$, $h_0 \leq \frac{1}{2}$ and $\frac{1}{h} \leq n \leq N$, then, by Lemma 2.4.3 we have for $\alpha > 0$ and $\beta > 0$ that

$$-\frac{2}{hn} \leq x_n(\alpha), y_n(\beta). \tag{2.83}$$

(The estimate is *a fortiori* true in the present $\alpha > 0$ and $\beta > 0$ case, since $\alpha \mapsto \mathcal{N}_\varphi(h, x, \alpha)$ and $\alpha \mapsto \mathcal{N}_\Phi(h, x, \alpha)$ are increasing functions, so iterates with $\alpha > 0$ run upwards faster than iterates at $\alpha = 0$, if all are started from $-\kappa$.)

Let us denote by $\widetilde{x_n} := x_{N-n}$ $(n = 0, 1, \dots, N)$. The grid construction is quite convenient, since now the inverse iteration is nothing else than a simple *relabeling*. The meaning of $\widetilde{y_n}$ is similar. If the conditions of Lemma 2.5.5 are satisfied (that is, if h_0, α_0 and κ are sufficiently small), then there exists a constant $c_1 > 0$ such that

$$\widetilde{x_n}(\alpha), \widetilde{y_n}(\beta) \leq \frac{2}{hn} \tag{2.84}$$

whenever $\frac{c_1}{h} \leq n \leq N$. (If $N = \frac{const}{h}$, then all the proofs below would become trivial.)

From these estimates (see Lemma 2.5.6 also) we get that there exist a constant $c_2 > 0$ such that

$$\begin{aligned} \#\{x_n | x_n \in [-\kappa, 0]\} &\leq \frac{c_2}{h\sqrt{\alpha}}, \\ \#\{\widetilde{x_n} | \widetilde{x_n} \in [0, \kappa]\} &\leq \frac{c_2}{h\sqrt{\alpha}}. \end{aligned}$$

The same estimates hold for y_n and $\widetilde{y_n}$, with α replaced by β on the right hand sides.

As a final preparation, we give counterparts of (2.81) and (2.82) for the inverses. From (2.81) we express $y_n - x_n$ $(n = 0, 1, \dots, N-1)$ as

$$y_n - x_n = (y_{n+1} - x_{n+1}) \cdot \left(\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_n, \beta)\right)^{-1} + h(\alpha - \beta)A_n \cdot \left(\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_n, \beta)\right)^{-1} -$$

$$hx_n^3(\widehat{\eta}_3(h, x_n, \alpha) - \widetilde{\eta}_3(h, x_n, \alpha)) \cdot \left(\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_n, \beta)\right)^{-1}.$$

Now for $n = 0, 1, \dots, N-1$ set $\widetilde{\xi_n} := \xi_{N-n}$, $\widetilde{A_n} := A_{N-n}$ and $k := N - n - 1$. Then the above expression for $k = 0, 1, \dots, N-1$ is simply

$$\widetilde{y_{k+1}} - \widetilde{x_{k+1}} = (\widetilde{y_k} - \widetilde{x_k}) \cdot \left(\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \widetilde{\xi_{k+1}}, \beta)\right)^{-1} + h(\alpha - \beta)\widetilde{A_{k+1}} \cdot \left(\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \widetilde{\xi_{k+1}}, \beta)\right)^{-1}$$

$$-h\widetilde{x_{k+1}}^3(\widehat{\eta}_3(h,\widetilde{x_{k+1}},\alpha)-\widetilde{\eta}_3(h,\widetilde{x_{k+1}},\alpha))\cdot\left(\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h,\widetilde{\xi_{k+1}},\beta)\right)^{-1}. \tag{2.85}$$

From this we get the following formula for $k=1,2,\ldots,N$.

$$\widetilde{y_k}-\widetilde{x_k}= \tag{2.86}$$

$$\sum_{i=1}^{k}\left[h(\alpha-\beta)\widetilde{A_i}-h\widetilde{x}_i^{\,3}(\widehat{\eta}_3(h,\widetilde{x}_i,\alpha)-\widetilde{\eta}_3(h,\widetilde{x}_i,\alpha))\right]\prod_{j=i}^{k}\left(\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h,\widetilde{\xi}_j,\beta)\right)^{-1}.$$

Now we have come to the main point of the present section and are able to prove that x_n and y_n are uniformly close to each other.

Lemma 2.8.2 *Suppose that h_0, α_0 and κ are sufficiently small. Then there exists a constant $c_3>0$ such that for every $h\in(0,h_0]$, $N\geq N_0(h)$ and $n=0,1,\ldots,N$ we have that*

$$|x_n(h,\alpha_N)-y_n(h,\beta_N)|\leq c_3h^p.$$

Proof. We will present the detailed proof assuming that $\beta\geq\alpha$.

Claim 1. First we show the following. There exists a constant $c_4>0$ such that if $\beta-\alpha>c_4h^p\sqrt{\alpha}$, then y_n ($n=0,1,\ldots$, travelling upwards) leaves the interval $[-\kappa,0]$ sooner (at 0 from below) than x_n, *and* $\widetilde{y_n}$ ($n=0,1,\ldots$, running downwards) exits the interval $[0,\kappa]$ sooner (at 0 from above) than $\widetilde{x_n}$. This clearly contradicts to the grid construction.

Claim 2. So we may assume that $0\leq\beta-\alpha\leq c_4h^p\sqrt{\alpha}$. Then we directly prove that there exists a constant $c_3>0$ such that $|x_n-y_n|\leq c_3h^p$.

In Claim 1 we will argue along the lines of the "second proof" of Lemma 2.8.1, while in Claim 2 we use the expressions from "first proof" of Lemma 2.8.1 and the fact that the number of x_n iterates in $[-\kappa,\kappa]$ is at most $\frac{c_2}{h\sqrt{\alpha}}$.

In the converse situation, that is, if $\alpha\geq\beta$, the proof above carries through if we make the following modifications.

1′. One can prove that there exists a constant $c_4>0$ such that if $\alpha-\beta>c_4h^p\sqrt{\beta}$, then x_n leaves the interval $[-\kappa,0]$ sooner than y_n and $\widetilde{x_n}$ leaves the interval $[0,\kappa]$ sooner than $\widetilde{y_n}$—again, a contradiction. In the proof, (2.81) should be replaced by the analogous formula

$$x_{n+1}-y_{n+1}=$$

$$(x_n-y_n)\cdot\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\varphi(h,\xi_n,\alpha)+(\alpha-\beta)\cdot\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathcal{N}_\varphi(h,y_n,\gamma)+hy_n^3(\widetilde{\eta}_3(h,y_n,\beta)-\widehat{\eta}_3(h,y_n,\beta)).$$

2′. When $0\leq\alpha-\beta\leq c_4h^p\sqrt{\beta}$, one uses the analogue of (2.82), now reading as

$$x_n-y_n=$$

$$\sum_{i=0}^{n-1}\left[h(\alpha-\beta)A_i+hy_i^3(\widetilde{\eta}_3(h,y_i,\beta)-\widehat{\eta}_3(h,y_i,\beta))\right]\prod_{j=i+1}^{n-1}\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\varphi(h,\xi_j,\alpha),$$

and the fact that the number of y_n iterates in $[-\kappa,\kappa]$ is at most $\frac{c_2}{h\sqrt{\beta}}$ to prove $|x_n-y_n|\leq c_3h^p$.

Now let us turn to the actual proof of Claim 1. Our first aim is to prove that the quantity y_n-x_n is strictly positive when the faster sequence reaches 0 from below. Hence we study the behaviour of the function $n\mapsto y_n-x_n$ and perform a "worst case" analysis.

Step 1. We choose a number $c_{40} \geq 1$ for a start and suppose that $\beta - \alpha > c_{40} h^p \sqrt{\alpha}$. Later, we will set $c_{40} \geq 1$ sufficiently large. Define $c_{41} := \sqrt[3]{\frac{c_{40}}{2c}}$ and $c_{42} := \sqrt[3]{\frac{1}{2c}}$. It is readily seen that if $x_n \in [-c_{41}\sqrt[6]{\alpha}, 0]$, then

$$(\beta - \alpha) \cdot \frac{\mathrm{d}}{\mathrm{d}\alpha}\mathcal{N}_\Phi(h, x_n, \gamma) + h x_n^3 (\widehat{\eta}_3(h, x_n, \alpha) - \widetilde{\eta}_3(h, x_n, \alpha)) \geq$$

$$c_{40} h^p \sqrt{\alpha}\frac{h}{2} - h|x_n|^3 c h^p \geq 0.$$

Hence if $x_n \in [-\kappa, 0]$, then $(\beta - \alpha) h A_n + h x_n^3 (\widehat{\eta}_3 - \widetilde{\eta}_3)$ can only be negative when $x_n \in [-\kappa, -c_{41}\sqrt[6]{\alpha}]$. (This interval is nondegenerate, if $\alpha \leq \alpha_0$ is small enough.) But $[-\kappa, -c_{41}\sqrt[6]{\alpha}] \subset [-\kappa, -c_{42}\sqrt[6]{\alpha}]$, because of the choice of c_{40}.

Step 2. It is elementary to see that if $\max(x_n, y_n) \leq 0$, then $\xi_n \leq 0$, so if κ is small enough, then $\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_n, \beta) \in \left(\frac{1}{2}, 1\right]$. Let us denote by I_1 the index set $\{n \in \mathbb{N} : x_n \in [-\kappa, -c_{41}\sqrt[6]{\alpha}]$ and $y_n \leq 0\}$, while set $I_2 := \{n \in \mathbb{N} : x_n \in [-\kappa, -c_{42}\sqrt[6]{\alpha}]$ and $y_n \leq 0\}$. Then $I_1 \subset I_2$, and—due to monotonicity of x_n and y_n—both are intervals in $\mathbb{N}$. For $n \in I_2$ we have by (2.81) that

$$y_{n+1} - x_{n+1} \geq -|y_n - x_n| \cdot \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_n, \beta) + (\beta - \alpha) h A_n + h x_n^3 (\widehat{\eta}_3 - \widetilde{\eta}_3) \geq$$

$$-|y_n - x_n| \cdot 1 + 0 - h|x_n|^3 c h^p = -|y_n - x_n| - c h^{p+1} |x_n|^3,$$

Step 3. Now we can easily prove by induction that for $1 \leq n \in I_2$

$$y_{n+1} - x_{n+1} \geq -\sum_{k=0}^{n} c h^{p+1} |x_k|^3.$$

This implies that the minimal value of the function $n \mapsto y_n - x_n$ on $n \in I_1$ for any $c_{40} \geq 1$ can not be less than $-\sum_{n \in I_2} c h^{p+1} |x_n|^3$. We now show that there exists a constant $c_5 > 0$ such that $-\sum_{n \in I_2} c h^{p+1} |x_n|^3 \geq -c_5 h^p$. To this end, define $c_{51} := \frac{2}{h c_{42}\sqrt[6]{\alpha}}$. Suppose α has been chosen so small such that $c_{42}\sqrt[6]{\alpha} \leq 2$, then $c_{51} \geq \frac{1}{h}$. So by (2.83) we have for $n > c_{51}$ that $x_n \geq -\frac{2}{hn} > -c_{42}\sqrt[6]{\alpha}$, thus $n \notin I_2$. In other words, $n \in I_2$ implies $n \leq c_{51}$. Then, using that $\kappa \leq 1$,

$$\sum_{n \in I_2} c h^{p+1} |x_n|^3 \leq \sum_{n=0}^{\lfloor c_{51} \rfloor} c h^{p+1} |x_n|^3 \leq \sum_{n=0}^{\lceil 1/h \rceil} c h^{p+1} \kappa^3 + \sum_{n=\lceil 1/h \rceil}^{\lfloor c_{51} \rfloor} c h^{p+1} \left(\frac{2}{hn}\right)^3 \leq$$

$$c h^{p+1}\left(\frac{1}{h} + 2\right) + c h^{p+1} \int_{1/h}^{c_{51}} \frac{8}{h^3 n^3} \mathrm{d}n = c h^p (1 + 2h) + c h^p \cdot 4 - c h^p c_{42}^2 \sqrt[3]{\alpha},$$

so $5 + 2h_0$ is an appropriate choice for c_5.

Step 4. At this point we know that for any $c_{40} \geq 1$ and $n \in I_1$ we have $y_n - x_n \geq -c_5 h^p$. Let us increase n further and consider the "complementary" index set $I_3 := \{n \in \mathbb{N} : x_n \in [-c_{41}\sqrt[6]{\alpha}, 0]$ and $y_n \leq 0\}$. Due to (2.81), Step 1 and estimate $-\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_n, \beta) \geq -1$, we see that if $y_n - x_n \leq 0$ for some $n \in I_3$, then $y_n - x_n \leq y_{n+1} - x_{n+1}$. Also notice that if $y_n - x_n$ is positive for some $n = n_0 \in I_3$, then $y_n - x_n$ remains positive for *all* $I_3 \ni n > n_0$ and we are ready, because then y_n leaves the interval $[-\kappa, 0]$ first.

Now set $I_4 := \{n \in \mathbb{N} : x_n \in [-\sqrt{\alpha}, -\frac{1}{2}\sqrt{\alpha}]\}$. We show that $\#I_4 > \frac{1}{10h\sqrt{\alpha}}$. First we verify that if h_0, α_0 and κ are small enough, then there exists an $x_n \in [-\sqrt{\alpha}, -\frac{3}{4}\sqrt{\alpha}]$. Indeed, at least one member of the x-sequence must lie in an interval $[-\sqrt[2^{\ell+1}]{\alpha}, -\sqrt[2^{\ell}]{\alpha}]$ for some $\ell \in \mathbb{N}^+$. (For example, if $\alpha < \kappa < 1$, then $\ell = \lfloor -\log_2 \frac{\ln \kappa}{\ln \alpha} \rfloor$ is appropriate.) Then we start a "backward induction" on ℓ. Suppose that $x_n \in [-\sqrt[2^{k+1}]{\alpha}, -\sqrt[2^k]{\alpha}]$ for some $1 \le k \le \ell$ and $n \in \mathbb{N}$. Then $h \le \frac{1}{10}$ implies that

$$x_{n+1} = \mathcal{N}_\varphi(h, x_n, \alpha) \le h\alpha + x_n + \frac{3}{2}x_n^2 \le h\alpha - \sqrt[2^k]{\alpha} + \frac{3}{2}h\sqrt[2^k]{\alpha} \le -\frac{3}{4}\sqrt[2^k]{\alpha}.$$

Thus, when x_n leaves interval $[-\sqrt[2^{k+1}]{\alpha}, -\sqrt[2^k]{\alpha}]$ (sooner or later it has to), it can jump only into $[-\sqrt[2^k]{\alpha}, -\frac{3}{4}\sqrt[2^k]{\alpha}]$. But this latter interval is contained for all $1 \le k \le \ell$ in $[-\sqrt[2^k]{\alpha}, -\sqrt[2^{k-1}]{\alpha}]$, for example, if $\kappa \le \frac{3}{4}$, so the induction can be continued. Finally, we get some $x_n \in [-\sqrt{\alpha}, -\frac{3}{4}\sqrt{\alpha}]$. Now with this x_n in hand, we can estimate from below the number of elements of I_4. Since the x-sequence in $[-\sqrt{\alpha}, -\frac{1}{2}\sqrt{\alpha}]$ travels through a line segment of length at least $\sqrt{\alpha}/4$, and the increment in each step is at most $h\alpha + \frac{3}{2}h(\sqrt{\alpha})^2 = \frac{5}{2}h\alpha$, we get that $\#I_4 > \frac{\sqrt{\alpha}/4}{5h\alpha/2} = \frac{1}{10h\sqrt{\alpha}}$. (The existence of an $x_n \in [-\sqrt{\alpha}, -\frac{3}{4}\sqrt{\alpha}]$ excludes the possibility of the numerator being too small.)

Step 5. Now if $n \in I_4$ and α is chosen such that $\alpha \le \frac{1}{4c}$, then (due to $c_{40} \ge 1$)

$$(\beta - \alpha) \cdot \frac{\mathrm{d}}{\mathrm{d}\alpha}\mathcal{N}_\Phi(h, x_n, \gamma) + hx_n^3(\widehat{\eta}_3(h, x_n, \alpha) - \widetilde{\eta}_3(h, x_n, \alpha)) \ge \frac{c_{40}}{4}h^{p+1}\sqrt{\alpha}.$$

Let $n_1 := \min I_4$ and $n_2 := \max I_4$. If $y_{n_2} - x_{n_2} > 0$ or $y_{n_2} > 0$, then we are ready by the first part of Step 4. Otherwise, if $y_{n_2} - x_{n_2} \le 0$ and $y_{n_2} \le 0$, then $y_{n_1+n} - x_{n_1+n} \le 0$ and $y_{n_1+n} \le 0$ for all $n \in \mathbb{N}$ satisfying $n_1 + n \le n_2$. Then (using $-\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \xi_n, \beta) \ge -1$ again) we can easily verify by induction on n that

$$y_{n_1+n} - x_{n_1+n} \ge -c_5h^p + n\frac{c_{40}}{4}h^{p+1}\sqrt{\alpha}.$$

This means that if $y_{n_2} - x_{n_2} \le 0$ and $y_{n_2} \le 0$, then

$$y_{n_2} - x_{n_2} \ge -c_5h^p + (n_2 - n_1)\frac{c_{40}}{4}h^{p+1}\sqrt{\alpha}.$$

But in the second part of Step 4 we have shown that $n_2 - n_1 \ge \frac{1}{10h\sqrt{\alpha}}$ (since I_4 is also an interval in $\mathbb{N}$). If we now choose $c_{40} := 40c_5 + 1$, then it is easy to see that $y_{n_2} - x_{n_2} > 0$, contradicting to our last assumption $y_{n_2} - x_{n_2} \le 0$. We have thus proved that if c_{40} is large enough and $\beta - \alpha \ge c_{40}h^p\sqrt{\alpha}$, then, in any case, y_n leaves the interval $[-\kappa, 0]$ sooner than x_n.

We can argue similarly to prove that there exists a constant $\widetilde{c_{40}} > 0$, such that if $\beta - \alpha \ge \widetilde{c_{40}}h^p\sqrt{\alpha}$, then $\widetilde{y_k}$ $(k = 0, 1, \ldots)$ exits the interval $[0, \kappa]$ sooner at 0 from above than $\widetilde{x_k}$. (Of course, we prove that the maximal value of the function $k \mapsto \widetilde{y_{k+1}} - \widetilde{x_{k+1}}$ is at most $\widetilde{c}_5h^p$, but this "advantage" of $\widetilde{x_k}$ over $\widetilde{y_k}$ gradually decreases as $\widetilde{x_k}$ runs through $[\frac{1}{2}\sqrt{\alpha}, \sqrt{\alpha}]$ and finally, $\widetilde{y_k}$ "takes the lead" anyway and reaches 0 first. We apply (2.85) instead of (2.81), and (2.84) instead of (2.83). In estimating (2.85), $\left(\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}_\Phi(h, \widetilde{\xi_{k+1}}, \beta)\right)^{-1}$ is estimated from above by 1 (provided that $\min(\widetilde{x_k}, \widetilde{y_k}) \ge 0$) at its first appearance in (2.85), while it is simply merged into $\widetilde{A_{k+1}}$ and c as an absolute constant at its second and third appearance.) Setting $c_4 := \max(c_{40}, \widetilde{c_{40}})$ proves Claim 1.

As for the proof of Claim 2, we suppose that $0 \le \beta - \alpha \le c_4h^p\sqrt{\alpha}$ and set $M_n := \max(x_n, y_n)$ and $m_n := \min(x_n, y_n)$.

Step I. Let us consider first those $n = 1, 2, \ldots$ indices for which $M_n \le 0$. As we have seen in Step 2, the derivatives in this case in (2.82) can not exceed 1, and we also know that $n \le N \le \frac{c_2}{h\sqrt{\alpha}}$. Then

$$|y_n - x_n| \le$$

$$\sum_{i=0}^{n-1}\left(c_4 h^p\sqrt{\alpha}\cdot 2h + h|x_i|^3\cdot ch^p\right)\prod_{j=i+1}^{n-1} 1 \le 2c_2c_4h^p + ch^p\sum_{i=0}^{n-1} h|x_i|^3.$$

But, as in Step 3, for example, we can estimate the last sum from above by an absolute positive constant. We have thus proved that, for some $c_{31} > 0$, $|y_n - x_n| \le c_{31}h^p$ whenever $M_n \le 0$.

Step II. Similarly, since "tilde counterparts" of formulae in Step I are completely analogous, we get the existence of a constant $c_{32} > 0$ such that $|y_n - x_n| \le c_{32}h^p$ if $m_n \ge 0$. (Applying a "tildeless" notation now suits better the next step.)

Step III. We are left with estimating $|y_n - x_n|$ during the transition of the sequences from the negative side to the positive one, that is, if the index n satisfies $m_n \le 0 \le M_n$. Set $c_{33} := \max(c_{31}, c_{32})$ and define the following partitioning intervals $P_1 := [-\kappa, -2c_{33}h^p]$, $P_2 := (-2c_{33}h^p, -c_{33}h^p]$, $P_3 := (-c_{33}h^p, 0]$, $\widetilde{P_3} := (0, c_{33}h^p]$, $\widetilde{P_2} := (c_{33}h^p, 2c_{33}h^p]$ and $\widetilde{P_1} := (2c_{33}h^p, \kappa]$.

Let us consider the final index in Step I, that is the *maximal* n_I such that $M_{n_I} \le 0$. There are two possibilities:

Case a) $M_{n_I} \in P_1 \cup P_2$

Case b) $M_{n_I} \in P_3$. (Of course, by Step I, $m_{n_I} \in P_2 \cup P_3$ then.)

Let us clarify Case a) first. Due to the maximality of n_I, M_{n_I+1} must lie in $(0, \kappa]$. So we have to check that M_n and m_n will be close to each other also for $n \in I_5 := \{k \in \mathbb{N} : k \ge n_I+1,\ m_k < 0\}$. The idea is simple: because M_n had at least one "big" jump (from $P_1 \cup P_2$ into $(0, \kappa]$), and α and β are close to each other, then m_n, too, should move quickly, implying that $\#I_5 \le c_{60}$, with some $c_{60} > 0$ being independent of the parameters h, α and β. With such an estimate of $\#I_5$, we see that it is enough to apply (2.81) only *finitely* many times (independently of the parameters) to link Step I and Step II together, meaning that $|y_n - x_n| \le c_{34}h^p$ holds in the transition phase $m_n \le 0 \le M_n$ with a suitable constant $c_{34} > 0$.

Now let us construct such a constant c_{60} to finish the proof of Case a). We know that $M_{n_I} \le -c_{33}h^p$ and $M_{n_I+1} > 0$. We may assume, say, that parameter α corresponds to the sequence M_n at its jump point. Then from the normal form we see that $0 < M_{n_I+1} \le h\alpha + \frac{1}{2}M_{n_I} = h\alpha - \frac{1}{2}|M_{n_I}|$ (if κ, hence $|M_{n_I}|$ is small enough), meaning that $\frac{1}{2h}|M_{n_I}| \le \alpha$. But then by Lemma 2.8.1, $\beta \ge \frac{1}{4h}|M_{n_I}|$, if $\frac{1}{8ch}|M_{n_I}| \ge h^p$, which—on account of condition $|M_{n_I}| \ge c_{33}h^p$—is true if $0 < h \le \frac{c_{33}}{8c}$.

Step I implies that $m_{n_I} \ge M_{n_I} - c_{31}h^p \ge M_{n_I} - c_{33}h^p \ge 2M_{n_I}$. On the other hand, if κ is small enough then the sum of the quadratic and cubic terms in the normal form is nonnegative, so

$$m_{n_I+1} \ge h\ \min(\alpha, \beta) + m_{n_I} + 0 \ge h\cdot|M_{n_I}|\frac{1}{4h} + m_{n_I}.$$

Applying this recursively, we get that $m_{n_I+\ell} \ge \frac{\ell}{4}|M_{n_I}| + m_{n_I}$. Combining estimates in this paragraph we see that $m_{n_I+\ell} \ge \frac{\ell}{4}|M_{n_I}| - 2|M_{n_I}|$. The right hand side is positive if $\ell \ge 8$, hence $\#I_5 \le 8$.

Now let us turn to Case b). There are three disjoint subcases here for $n > n_I$.

Case b1) M_n is staying within $\widetilde{P_3} \cup \widetilde{P_2}$ until m_n first becomes positive. Then we are ready, since for these n indices $|y_n - x_n| = M_n - m_n \le (2+2)c_{33}h^p$, and we are back in Step II.

Case b2) M_n has already left $\widetilde{P_3} \cup \widetilde{P_2}$ before m_n enters $(0, \kappa]$, *and* m_n enters $(0, \kappa]$ at $\widetilde{P_3}$. We can immediately rule out this dangerous possibility, because it would contradict to Step II.

Case b3) M_n has already left $\widetilde{P_3} \cup \widetilde{P_2}$ before m_n enters $(0, \kappa]$, but m_n jumps from $[-\kappa, 0]$ into $\widetilde{P_2} \cup \widetilde{P_1}$. Then we have to verify that M_n and m_n stayed close to each other also "in the past", that is, for $n \in I_6 := \{k \in \mathbb{N} : M_k \in (0, \kappa],\ m_k \in P_2 \cup P_3\}$. Just as in Case a), Case b3) is finished if we establish $\#I_6 \le c_{61}$, with a suitable absolute constant $c_{61} > 0$.

We can reason similarly as in Case a). Let n_{II} denote the starting index in Step II, that is the one for which $m_{n_{II}} \leq 0$, but $m_{n_{II}+1} > 0$. We may assume, say, that now parameter β corresponds to the sequence m_n at n_{II}. What Case b3) means is that $c_{33}h^p < m_{n_{II}+1} \leq h\beta + m_{n_{II}} - \frac{1}{2}m_{n_{II}} \leq h\beta$, where in the $\leq$-estimates we have used that the absolute value of the sum of quadratic and cubic terms in the actual normal form is at most $\frac{1}{2}|m_{n_{II}}|$ (if κ is small enough), further, that $m_{n_{II}}$ is nonpositive. But then $\beta \geq c_{33}h^{p-1}$, and due to Lemma 2.8.1, $\alpha \geq \frac{1}{2}c_{33}h^{p-1}$, if h is chosen from $(0, \frac{c_{33}}{4c})$. So both parameters are again large enough: $\min(\alpha, \beta) \geq \frac{1}{2}c_{33}h^{p-1}$. This means that m_n could not stay too long in $P_2 \cup P_3$. Indeed, similarly to Case a) we see from the normal form that $m_{k+1} \geq h \cdot \frac{1}{2}c_{33}h^{p-1} + m_k + 0$, so $m_{k+\ell} \geq \frac{\ell}{2}c_{33}h^p + m_k$. Thus, if $m_k \leq -2c_{33}h^p$ for some index k, and $\frac{\ell}{2} > 2$, then $m_{k+\ell} > 0$, yielding $\#I_6 \leq 5$ and also completing the proof. ■

Remark 2.8.1 First we were able to prove Claim 2, that is, closeness of the grid points under an additional hypothesis $\beta - \alpha \leq c_4 h^p \sqrt{\alpha}$. (This choice was motivated by the fact that the number of iterates in $[-\kappa, \kappa]$ is at most $\frac{c_2}{h\sqrt{\alpha}}$.) It turned out only later—as the proof of Claim 1 evolved—that $\beta - \alpha \geq c_4 h^p \sqrt{\alpha}$ actually can not hold within the grid structure. This fact is worth highlighting separately, since it strengthens the result of Lemma 2.8.1.

Corollary 2.8.3 *Suppose that h_0, α_0 and κ are sufficiently small. Then there exists a constant $c_4 > 0$ such that for every $N \geq N_0(h)$ we have that*

$$|\beta_N(h) - \alpha_N(h)| \leq c_4 \max\left(\sqrt{\alpha_N(h)}, \sqrt{\beta_N(h)}\right) h^p. \quad ■$$

2.8.1 Continuity at $\alpha = 0$ along the grid sequence

In the previous section we have proved optimal $\mathcal{O}(h^p)$ closeness estimates on the points of the "α-grid". In this section, we extend the definition of the conjugacy from this "skeleton" to the vertical α-grid lines and address the problem of continuity of the conjugacy at $\alpha = 0^+$, $x > 0$. Finally, the conjugacy is extended further for all $0 \leq |\alpha| \leq \alpha_0$, $|x| \leq \kappa$, and $\mathcal{O}(h)$ closeness estimates are proved.

We assume in the following that $h > 0$ is fixed and use the notation of the previous section. The letter $\mathcal{N}$ will denote either of the normal forms $\mathcal{N}_\varphi$ or $\mathcal{N}_\Phi$. For the rest of the section, fix, say, $\mathcal{N} := \mathcal{N}_\varphi$. For simplicity, the dependence on h is suppressed, thus $\mathcal{N}(x,\alpha) := \mathcal{N}(h,x,\alpha)$. For $k \in \mathbb{Z}$, set $\mathcal{N}^{[k]}(x,\alpha) := \left(\mathcal{N}(h,\cdot,\alpha)\right)^{[k]}(x)$. The dependence of the sequence $x_k(\alpha) := \mathcal{N}^{[k]}(x_0,\alpha)$ ($\alpha \geq 0$, $k \in \mathbb{N}$) on h is also suppressed. We set $x_0 \equiv x_0(\alpha) := -\kappa$ and $\widetilde{x}_0 \equiv \widetilde{x}_0(\alpha) := \kappa$ for any $\alpha \geq 0$. Of course, $\widetilde{x}_k(\alpha) := \mathcal{N}^{[-k]}(\widetilde{x}_0,\alpha)$ ($\alpha \geq 0$, $k \in \mathbb{N}$). The abbreviation $\mathcal{N}'(x,\alpha) := (\mathcal{N}(\cdot,\alpha))'(x)$ is also used. We will frequently use that $\alpha_N \to 0^+$ is equivalent to saying that $N \to \infty$.

Since the key theme of this section is the *preservation of ratios inside iterates of the fundamental domains*, let us define for $x \in [x_0(\alpha), x_1(\alpha)]$, $\alpha \geq 0$ and $k \in \mathbb{N}$ the quotients

$$Q_k(x,\alpha) := \frac{u_k(x,\alpha) - x_k(\alpha)}{x_{k+1}(\alpha) - x_k(\alpha)} := \frac{\mathcal{N}^{[k]}(x,\alpha) - \mathcal{N}^{[k]}(x_0,\alpha)}{\mathcal{N}^{[k+1]}(x_0,\alpha) - \mathcal{N}^{[k]}(x_0,\alpha)}.$$

It turns out that these fundamental objects will have nice properties: as we know the *length* of intervals $[x_k(\alpha_N), x_{k+1}(\alpha_N)]$ ($k = 0, 1, \ldots, N-1$) exhibits unimodality (*i.e.* the length first decreases, then increases as x_k passes through 0), however, the *ratio* in Q_k is monotone in k.

Obviously, $Q_k(x_0, \alpha) = 0$ and $Q_k(x_1(\alpha), \alpha) = 1$, further—since $\mathcal{N}$ is strictly increasing, strictly convex and continuous—so is $Q_k(\cdot, \alpha)$ ($\alpha \geq 0$, $k \in \mathbb{N}$) on $[x_0(\alpha), x_1(\alpha)]$. Monotonicity and convexity of the normal form $\mathcal{N}$ allow us to prove the following interesting lemma concerning the $\alpha = 0$ case.

Lemma 2.8.4 *For any $x \in [x_0(0), x_1(0)]$ the limit*

$$Q_\infty(x) := \lim_{k\to\infty} Q_k(x, 0) \equiv \lim_{k\to\infty} \frac{\mathcal{N}^{[k]}(x, 0) - x_k(0)}{x_{k+1}(0) - x_k(0)}$$

exists and finite, further $Q_\infty : [x_0(0), x_1(0)] \to [0, 1]$ is an increasing homeomorphism.

Similarly, for $\widetilde{x} \in [\widetilde{x}_1(0), \widetilde{x}_0(0)]$ the map

$$Q_{-\infty}(\widetilde{x}) := \lim_{k\to\infty} \frac{\mathcal{N}^{[-k]}(\widetilde{x}, 0) - \widetilde{x}_{k+1}(0)}{\widetilde{x}_k(0) - \widetilde{x}_{k+1}(0)}$$

is well defined, and $Q_{-\infty} : [\widetilde{x}_1(0), \widetilde{x}_0(0)] \to [0, 1]$ is also an increasing homeomorphism.

Proof. For simplicity, throughout the proof we write x_k instead of $x_k(0)$. Fix any $x \in [x_0, x_1]$. For any $k \in \mathbb{N}^+$, the mean value theorem implies that

$$Q_k(x, 0) = \frac{\mathcal{N}'(\chi_{k-1}(x), 0)}{\mathcal{N}'(\xi_{k-1}, 0)} Q_{k-1}(x, 0),$$

or, recursively, that

$$Q_k(x, 0) = \frac{x - x_0}{x_1 - x_0} \prod_{i=0}^{k-1} \frac{\mathcal{N}'(\chi_i(x), 0)}{\mathcal{N}'(\xi_i, 0)}, \tag{2.87}$$

with some suitable $\chi_i(x) \in [x_i, \mathcal{N}^{[i]}(x, 0)] \subset [x_i, x_{i+1}]$ and $\xi_i \in [x_i, x_{i+1}]$ for $i = 0, 1, \ldots, k-1$. Since $\mathcal{N}(\cdot, 0)$ is convex, it is easily seen that

$$x_i \leq \chi_i(x) \leq \xi_i \leq x_{i+1} < 0$$

for $x \in [x_0, x_1]$. But $\mathcal{N}'(\cdot, 0)$ is increasing and positive, so $0 < \frac{\mathcal{N}'(\chi_i(x),0)}{\mathcal{N}'(\xi_i,0)} \leq 1$. This implies that for any $x \in [x_0, x_1]$, the function $k \mapsto Q_k(x, 0) \in [0, 1]$ is monotone decreasing, hence $Q_\infty(x) \in [0, 1]$ exists. Obviously, $Q_\infty(x_0) = 0$ and $Q_\infty(x_1) = 1$.

Now it is time for some estimates. First we rewrite (2.87) as

$$Q_k(x, 0) = \frac{x - x_0}{x_1 - x_0} \exp\left(\sum_{i=0}^{k-1} \ln\left(\mathcal{N}'(\chi_i(x), 0)\right) - \ln\left(\mathcal{N}'(\xi_i, 0)\right)\right).$$

Using the boundedness of the η term and its x-derivative in the normal form $\mathcal{N}$, further inequality $t - t^2 \leq \ln(1 + t) \leq t$ (*e.g.* for $t \in [-\frac{1}{2}, \frac{1}{2}]$), we can estimate the sum in the exponent from above by

$$\sum_{i=0}^{k-1} \left(2h(\chi_i(x) - \xi_i) + h \cdot \mathcal{O}(\chi_i^2(x), \xi_i^2)\right).$$

A similar lower estimate for the exponent also holds. As we know, Lemma 2.4.3 in the $\alpha = 0$ case shows that $h \sum_{i=0}^{\infty} x_i^2 < \infty$ (uniformly in h). But $\xi_i^2 \leq \chi_i^2(x) \leq x_i^2$ and $\chi_i(x) - \xi_i \leq 0$, further—by the definition of $\mathcal{N}$—we have $|\chi_i(x) - \xi_i| \leq |x_{i+1} - x_i| = h \cdot \mathcal{O}(x_i^2)$, hence

$$\sum_{i=0}^{k-1} \left|\ln\left(\mathcal{N}'(\chi_i(\cdot), 0)\right) - \ln\left(\mathcal{N}'(\xi_i, 0)\right)\right|$$

is uniformly bounded on $[x_0, x_1]$ and also in k and h. This tells us that there exists a constant $c_1 > 0$ such that for each $k \in \mathbb{N}$

$$c_1 \cdot \frac{x - x_0}{x_1 - x_0} \leq Q_k(x, 0) \leq \frac{x - x_0}{x_1 - x_0}. \tag{2.88}$$

On the other hand, uniform boundedness of the exponent also implies that

$$\sum_{i=0}^{\infty} \ln \left(\frac{\mathcal{N}'(\chi_i(\cdot), 0)}{\mathcal{N}'(\xi_i, 0)} \right)$$

converges uniformly on $[x_0, x_1]$. This—taking into account the continuity of each term of the series—proves that the limit function, $Q_\infty(\cdot)$ is also continuous.

Since $Q_k(\cdot, 0)$ is strictly increasing, we have for any $x_0 \leq x < y \leq x_1$ that $Q_\infty(x) \leq Q_\infty(y)$. Suppose to the contrary that for some $x_0 < x < y \leq x_1$ we have $Q_\infty(x) = Q_\infty(y)$. (The case $x = x_0$—which would otherwise hinder the proof, since the statement is false in that case—is excluded by (2.88)). Then clearly $Q_\infty(\cdot) \equiv c_2 > 0$ holds on $[x, y]$ with some constant c_2. But for fixed x the convergence of $Q_k(x, 0)$ in k is monotone decreasing, so $Q_k(x, 0) \geq Q_\infty(x)$. We also know that for each $\varepsilon > 0$ there exists a $k \in \mathbb{N}^+$ such that $\sup_{[x,y]} |Q_k(\cdot, 0) - Q_\infty(\cdot)| \leq \varepsilon$, that is $c_2 \leq Q_k(\cdot, 0) \leq c_2 + \varepsilon$ on $[x, y]$. Since $Q_k(\cdot, 0)$ is strictly convex, its graph lies above its tangent line at x. But the slope of this line is at most $\frac{(c_2+\varepsilon)-c_2}{y-x}$, so for $\varepsilon > 0$ sufficiently small the tangent is nearly horizontal, hence $Q_k(x_0) = 0$ can not hold. This contradiction proves that $Q_\infty(\cdot)$ is strictly increasing.

The proofs for $Q_{-\infty}$ are similar. ■

The next lemma shows boundedness of the derivative of some high iterates of the normal form on the first fundamental domain.

Lemma 2.8.5

$$\left| \left(\mathcal{N}^{[N-1]} \right)' (x, \alpha_N) \right|$$

is uniformly bounded in $x \in [x_0(\alpha_N), x_1(\alpha_N)]$, $h > 0$ and $N \in \mathbb{N}^+$.

Proof. On the one hand, by using our earlier definition $u_k(x, \alpha) := \mathcal{N}^{[k]}(x, \alpha)$ and the chain rule, we get for any $x \in [x_0(\alpha_N), x_1(\alpha_N)]$ and $N \in \mathbb{N}^+$ that

$$\left(\mathcal{N}^{[N-1]} \right)' (x, \alpha_N) = \prod_{k=0}^{N-2} \mathcal{N}'(u_k(x, \alpha_N), \alpha_N).$$

Clearly, $u_k(x, \alpha_N) \in [x_k(\alpha_N), x_{k+1}(\alpha_N)]$. On the other hand, by iterated application of the mean value theorem we have that

$$x_N(\alpha_N) - x_{N-1}(\alpha_N) = (x_1(\alpha_N) - x_0) \prod_{k=0}^{N-2} \mathcal{N}'(\xi_k(\alpha_N), \alpha_N),$$

with some $\xi_k(\alpha_N) \in [x_k(\alpha_N), x_{k+1}(\alpha_N)]$. But $\mathcal{N}'(\cdot, \alpha_N)$ is monotone increasing and positive, so

$$\prod_{k=0}^{N-2} \mathcal{N}'(u_k(x, \alpha_N), \alpha_N) \leq \left(\prod_{k=1}^{N-2} \mathcal{N}'(\xi_k(\alpha_N), \alpha_N) \right) \cdot \mathcal{N}'(u_{N-2}(x, \alpha_N), \alpha_N).$$

Combining these we have that

$$\left(\mathcal{N}^{[N-1]} \right)' (x, \alpha_N) \leq \frac{x_N(\alpha_N) - x_{N-1}(\alpha_N)}{x_1(\alpha_N) - x_0} \cdot \frac{\mathcal{N}'(u_{N-2}(x, \alpha_N), \alpha_N)}{\mathcal{N}'(\xi_0(\alpha_N), \alpha_N)} \leq$$

$$\frac{\kappa - \widetilde{x}_1(\alpha_N)}{x_1(\alpha_N) - (-\kappa)} \cdot \frac{\mathcal{N}'(\kappa, \alpha_N)}{\mathcal{N}'(-\kappa, \alpha_N)},$$

since $x_N(\alpha_N) = \kappa$ and $x_0 = -\kappa$. But we know that $\frac{1}{2} \leq |\mathcal{N}'(-\kappa, \alpha_N)| \leq |\mathcal{N}'(\kappa, \alpha_N)| \leq 2$, if h is small enough, further, that $c_1 h \leq \kappa - \widetilde{x}_1(\alpha_N) \leq c_2 h$ and $c_1 h \leq x_1(\alpha_N) - (-\kappa) \leq c_2 h$, with some $c_1 > 0$ and $c_2 > 0$ being independent of N. ■

Remark 2.8.2 We remark that by using the same argument, one can show for $\alpha > 0$ that

$$\sum_{k \in \mathbb{N}\,:\, x_k \in [-\kappa, \kappa]} h \cdot x_k(h, \alpha)$$

is also bounded, uniformly in $h > 0$ and $\alpha > 0$, while

$$\sum_{k \in \mathbb{N}\,:\, x_k \in [0, \kappa]} h \cdot x_k(h, \alpha) = \mathcal{O}\left(\frac{1}{\alpha}\right).$$

Now we have come to the main technical lemma of this section.

Lemma 2.8.6 *For each $x \in [x_0(0), x_1(0)] \subset [x_0(\alpha), x_1(\alpha)]$ ($\alpha > 0$) there exists an $x^* \in [\widetilde{x}_1(0), \widetilde{x}_0(0)] \subset [\widetilde{x}_1(\alpha), \widetilde{x}_0(\alpha)]$ such that*

$$\lim_{N \to \infty} \mathcal{N}^{[N-1]}(x, \alpha_N) = x^*.$$

Moreover, $x^* = Q_{-\infty}^{[-1]}(Q_\infty(x))$, so the mapping $x \mapsto x^*$ is an increasing homeomorphism.

Proof. Let us fix $x \in [x_0(0), x_1(0)]$ arbitrarily. We will show that $\mathcal{N}^{[N-1]}(x, \alpha_N)$ is a Cauchy-sequence as $N \to \infty$, and once we know it converges, a "diagonal argument" will prove that $x^* = Q_{-\infty}^{[-1]}(Q_\infty(x))$. Lemma 2.8.4 then yields that $x \mapsto x^*$ is an increasing homeomorphism.

Notice first, that the statement of the present lemma is true for $x = x_0$, when $x^* = \widetilde{x}_1(0)$, since $\mathcal{N}^{[N-1]}(x_0, \alpha_N) = \mathcal{N}^{[-1]}(\kappa, \alpha_N)$ and $\mathcal{N}^{[-1]}(\kappa, \cdot)$ is continuous. So we suppose that $x \in (x_0(0), x_1(0)]$—this will exclude a "division by zero" later.

Step 1. The central idea of the proof is to analyze the behaviour of $\mathcal{N}^{[N-1]}(x, \alpha_N)$ in terms of some $Q_k(x, \alpha_N)$ with k chosen suitably. So let us fix a small $\varepsilon > 0$, choose some small $\alpha(\varepsilon) > 0$ and consider only those N indices for which $0 < \alpha_N \leq \alpha(\varepsilon)$. We also fix some large index $L(\varepsilon) \in \mathbb{N}$, such that $0 < L(\varepsilon) < N - L(\varepsilon) < N$. It is vital that $L(\varepsilon)$ is independent of N. During the proof we will choose $\alpha(\varepsilon) > 0$ sufficiently small and $L(\varepsilon)$ sufficiently large (in fact, $L(\varepsilon) \to \infty$ as $\varepsilon \to 0^+$, however, if ε is fixed, then $L(\varepsilon)$ is fixed, too).

Now let us consider the following identity:

$$Q_{N-L(\varepsilon)}(x, \alpha_N) = \frac{Q_{N-L(\varepsilon)}(x, \alpha_N)}{Q_{L(\varepsilon)}(x, \alpha_N)} \cdot \frac{Q_{L(\varepsilon)}(x, \alpha_N)}{Q_{L(\varepsilon)}(x, 0)} \cdot \frac{Q_{L(\varepsilon)}(x, 0)}{Q_\infty(x)} \cdot Q_\infty(x).$$

The denominators are separated from 0 (since $x \in (x_0(0), x_1(0)]$ is fixed). But if $\alpha(\varepsilon)$ is small enough, then for *all* $\alpha_N \leq \alpha(\varepsilon)$ we have that

$$\frac{Q_{L(\varepsilon)}(x, \alpha_N)}{Q_{L(\varepsilon)}(x, 0)} \in [1 - \varepsilon, 1 + \varepsilon],$$

because $\mathcal{N}^{[L(\varepsilon)]}(x, \cdot)$ is continuous, since $L(\varepsilon)$ is fixed. Similarly, by the definition of Q_∞,

$$\frac{Q_{L(\varepsilon)}(x, 0)}{Q_\infty(x)} \in [1 - \varepsilon, 1 + \varepsilon]$$

if $L(\varepsilon)$ is large enough. For any $\gamma > 0$ and $I := [a, b]$ $(a, b > 0)$, let γI represent the interval $[\gamma a, \gamma b]$. Then, with a suitable function $\tau_1(\varepsilon)$ (with $\tau_1(\varepsilon) \to 0^+$ as $\varepsilon \to 0^+$) we get for all $\alpha_N \leq \alpha(\varepsilon)$ that

$$Q_{N-L(\varepsilon)}(x, \alpha_N) \in \frac{Q_{N-L(\varepsilon)}(x, \alpha_N)}{Q_{L(\varepsilon)}(x, \alpha_N)} \cdot Q_\infty(x) \cdot [1 - \tau_1(\varepsilon), 1 + \tau_1(\varepsilon)]. \tag{2.89}$$

Now let us turn to the remaining fraction on the right-hand side of (2.89). A small rearrangement in the definition of $Q_k(x, \alpha_N)$ yields that

$$\frac{Q_{N-L(\varepsilon)}(x, \alpha_N)}{Q_{L(\varepsilon)}(x, \alpha_N)} = \frac{\frac{u_{N-L(\varepsilon)}(x,\alpha_N) - x_{N-L(\varepsilon)}(\alpha_N)}{u_{L(\varepsilon)}(x,\alpha_N) - x_{L(\varepsilon)}(\alpha_N)}}{\frac{x_{N-L(\varepsilon)+1}(\alpha_N) - x_{N-L(\varepsilon)}(\alpha_N)}{x_{L(\varepsilon)+1}(\alpha_N) - x_{L(\varepsilon)}(\alpha_N)}} = \dots$$

where, we recall that $u_k(x, \alpha_N) := \mathcal{N}^{[k]}(x, \alpha_N) \in [x_k(\alpha_N), x_{k+1}(\alpha_N)]$. The next step, of course, is to appeal to the mean value theorem several times to obtain that

$$\dots = \frac{\frac{x - x_0}{x - x_0} \cdot \frac{\prod_{k=0}^{N-L(\varepsilon)-1} \mathcal{N}'(\chi_k(x,\alpha_N),\alpha_N)}{\prod_{k=0}^{L(\varepsilon)-1} \mathcal{N}'(\chi_k(x,\alpha_N),\alpha_N)}}{\frac{x_1(\alpha_N) - x_0}{x_1(\alpha_N) - x_0} \cdot \frac{\prod_{k=0}^{N-L(\varepsilon)-1} \mathcal{N}'(\xi_k(\alpha_N),\alpha_N)}{\prod_{k=0}^{L(\varepsilon)-1} \mathcal{N}'(\xi_k(\alpha_N),\alpha_N)}} = \dots$$

where—due to convexity—we have $x_k(\alpha_N) \leq \chi_k(x, \alpha_N) \leq \xi_k(\alpha_N) \leq x_{k+1}(\alpha_N)$ with suitable $\chi_k(x, \alpha_N)$ and $\xi_k(\alpha_N)$. Then

$$\dots = \prod_{k=L(\varepsilon)}^{N-L(\varepsilon)-1} \frac{\mathcal{N}'(\chi_k(x, \alpha_N), \alpha_N)}{\mathcal{N}'(\xi_k(\alpha_N), \alpha_N)}.$$

But $x_k(\alpha_N) \leq \chi_k(x, \alpha_N) \leq \xi_k(\alpha_N) \leq x_{k+1}(\alpha_N) \leq \chi_{k+1}(x, \alpha_N)$, and $0 < \mathcal{N}'(\cdot, \alpha_N)$ is monotone increasing, so

$$\frac{\mathcal{N}'(x_{L(\varepsilon)}(\alpha_N), \alpha_N)}{\mathcal{N}'(x_{N-L(\varepsilon)}(\alpha_N), \alpha_N)} \leq \frac{\mathcal{N}'(\chi_{L(\varepsilon)}(x, \alpha_N), \alpha_N)}{\mathcal{N}'(\xi_{N-L(\varepsilon)-1}(\alpha_N), \alpha_N)} \leq \prod_{k=L(\varepsilon)}^{N-L(\varepsilon)-1} \frac{\mathcal{N}'(\chi_k(x, \alpha_N), \alpha_N)}{\mathcal{N}'(\xi_k(\alpha_N), \alpha_N)} \leq 1.$$

Since for fixed $L(\varepsilon)$ both functions $\mathcal{N}^{[L(\varepsilon)]}(x_0, \cdot)$ and $\mathcal{N}^{[-L(\varepsilon)]}(\kappa, \cdot)$ are continuous, we have that $x_{L(\varepsilon)}(\alpha_N) \to x_{L(\varepsilon)}(0)$ and $x_{N-L(\varepsilon)}(\alpha_N) = \mathcal{N}^{[-L(\varepsilon)]}(x_N(\alpha_N), \alpha_N) = \mathcal{N}^{[-L(\varepsilon)]}(\kappa, \alpha_N) \to \mathcal{N}^{[-L(\varepsilon)]}(\kappa, 0) = \widetilde{x}_{L(\varepsilon)}(0)$ as $N \to \infty$. However, $x_{L(\varepsilon)}(0)$ and $\widetilde{x}_{L(\varepsilon)}(0)$ are arbitrarily close to 0, if $L(\varepsilon)$ is sufficiently large (and $\alpha_N \leq \alpha(\varepsilon)$ is sufficiently small so $0 < L(\varepsilon) < N - L(\varepsilon) < N$ can hold). Further, $\mathcal{N}'(0, \alpha_N) = 1$ and $\mathcal{N}'(x, \cdot)$ is continuous, so we get that if $\alpha(\varepsilon)$ is sufficiently small and $L(\varepsilon)$ is sufficiently large, but both quantities are fixed as N varies, then for every $\alpha_N \leq \alpha(\varepsilon)$

$$\frac{Q_{N-L(\varepsilon)}(x, \alpha_N)}{Q_{L(\varepsilon)}(x, \alpha_N)} \in [1 - \varepsilon, 1].$$

This, together with (2.89) imply that for every sufficiently small $\varepsilon > 0$ there exist a sufficiently small $\alpha(\varepsilon) > 0$, a sufficiently large $L(\varepsilon) \in \mathbb{N}^+$ and a function $\tau_2(\varepsilon)$ (with $\tau_2(\varepsilon) \to 0^+$ as $\varepsilon \to 0^+$) such that for every $\alpha_N \leq \alpha(\varepsilon)$

$$Q_{N-L(\varepsilon)}(x, \alpha_N) \in Q_\infty(x) \cdot [1 - \tau_2(\varepsilon), 1 + \tau_2(\varepsilon)]. \tag{2.90}$$

In this construction we can clearly choose $L(\varepsilon)$ such that $L(\varepsilon) \to \infty$ as $\varepsilon \to 0^+$, moreover, we can assume that $L(\varepsilon)$ strictly increases as ε decreases. We remark that these estimates are also valid if $\alpha_N \le \alpha(\varepsilon)$ is replaced by $\alpha \le \alpha(\varepsilon)$, so α can tend to 0^+ not only along the grid values. In what follows, however, we will exploit that $\alpha = \alpha_N$ is jumping from grid point to grid point.

Now let us return to (2.90). The definition of Q_k yields that with a suitable $\tau_3(\varepsilon)$

$$\mathcal{N}^{[N-L(\varepsilon)]}(x, \alpha_N) \in$$

$$x_{N-L(\varepsilon)}(\alpha_N) + (x_{N-L(\varepsilon)+1}(\alpha_N) - x_{N-L(\varepsilon)}(\alpha_N)) \cdot Q_\infty(x) \cdot [1-\tau_3(\varepsilon), 1+\tau_3(\varepsilon)].$$

We have seen that the right-hand side of "$\in$" converges to

$$\widetilde{x}_{L(\varepsilon)}(0) + (\widetilde{x}_{L(\varepsilon)-1}(0) - \widetilde{x}_{L(\varepsilon)}(0)) \cdot Q_\infty(x) \cdot [1-\tau_3(\varepsilon), 1+\tau_3(\varepsilon)]$$

as $N \to \infty$, if $L(\varepsilon)$ is fixed. So with a suitable $\tau_4(\varepsilon)$ we have that for all $\alpha_N \le \alpha(\varepsilon)$

$$\mathcal{N}^{[N-L(\varepsilon)]}(x, \alpha_N) \in (\widetilde{x}_{L(\varepsilon)}(0) + (\widetilde{x}_{L(\varepsilon)-1}(0) - \widetilde{x}_{L(\varepsilon)}(0)) \cdot Q_\infty(x)) \cdot [1-\tau_4(\varepsilon), 1+\tau_4(\varepsilon)]. \quad (2.91)$$

Since $\mathcal{N}^{[N-1]}(x, \alpha_N) = \mathcal{N}^{[L(\varepsilon)-1]}(\mathcal{N}^{[N-L(\varepsilon)]}(x, \alpha_N), \alpha_N)$, $L(\varepsilon)-1$ is fixed, and $\mathcal{N}^{[L(\varepsilon)-1]}(y, \alpha_N) \to \mathcal{N}^{[L(\varepsilon)-1]}(y, 0)$ as $N \to \infty$ and y is fixed, we obtain that

$$\mathcal{N}^{[N-1]}(x, \alpha_N) \in \mathcal{N}^{[L(\varepsilon)-1]}(\mathcal{N}^{[N-L(\varepsilon)]}(x, \alpha_N), 0) \cdot [1-\tau_5(\varepsilon), 1+\tau_5(\varepsilon)] \quad (2.92)$$

with $\tau_5(\varepsilon) \to 0^+$ as $\varepsilon \to 0^+$, for every $\alpha_N \le \alpha(\varepsilon)$. But $\mathcal{N}^{[L(\varepsilon)-1]}(\cdot, 0)$ is continuous, so (2.91) and (2.92) imply that for all $\alpha_N \le \alpha(\varepsilon)$

$$\mathcal{N}^{[N-1]}(x, \alpha_N) \in w_x(L(\varepsilon)) \cdot [1-\tau_6(\varepsilon), 1+\tau_6(\varepsilon)] \quad (2.93)$$

with a suitable $\tau_6(\varepsilon) \to 0^+$ (as $\varepsilon \to 0^+$) and with

$$w_x(k) := \mathcal{N}^{[k-1]}(\widetilde{x}_k(0) + (\widetilde{x}_{k-1}(0) - \widetilde{x}_k(0)) \cdot Q_\infty(x), 0).$$

Since we can make the interval on the right-hand side of (2.93) as narrow as we wish, we get finally that $\mathcal{N}^{[N-1]}(x, \alpha_N)$ is a Cauchy-sequence, so it converges to some x^* as $N \to \infty$.

Step 2. Let us proceed further. The convergence $\mathcal{N}^{[N-1]}(x, \alpha_N) \to x^*$ with (2.93) also implies that

$$x^* \in w_x(L(\varepsilon)) \cdot [1-\tau_6(\varepsilon), 1+\tau_6(\varepsilon)],$$

or, in other words that

$$w_x(L(\varepsilon)) \in x^* \cdot [1-\tau_7(\varepsilon), 1+\tau_7(\varepsilon)],$$

with a suitable $\tau_7(\varepsilon)$. But $L(\varepsilon) \to \infty$ and $\tau_7(\varepsilon) \to 0^+$ as $\varepsilon \to 0^+$, hence—by fixing a strictly decreasing sequence $\varepsilon_k \to 0^+$ as $k \to \infty$—we get that

$$\lim_{k\to\infty} w_x(L(\varepsilon_k)) = x^*.$$

However, we will need a bit more.

Step 2a. For $\widetilde{x} \in [\widetilde{x}_1(0), \widetilde{x}_0(0)]$ and $M \in \mathbb{N}^+$ we will use the abbreviation

$$Q_{-M}(\widetilde{x}, 0) := \frac{\mathcal{N}^{[-M]}(\widetilde{x}, 0) - \widetilde{x}_{M+1}(0)}{\widetilde{x}_M(0) - \widetilde{x}_{M+1}(0)}.$$

Then, by the definition of $Q_{-\infty}$

$$Q_{-\infty}(\widetilde{x}) = \lim_{M\to\infty} Q_{-M}(\widetilde{x}, 0),$$

and it is easy to see that this convergence is *monotone increasing* in M, because $\mathcal{N}^{[-1]}(\cdot,0)$ is a monotone increasing concave function (*cf.* the proof of Lemma 2.8.4).

Step 2b. We now show that for any $x \in (x_0(0), x_1(0)]$ and $k \in \mathbb{N}^+$

$$w_x(k+1) \leq w_x(k).$$

Indeed, since $\mathcal{N}(\cdot,0)$ is convex, monotone increasing and lies above the identity map, we know (*cf.* the proof of Lemma 2.8.4 again) that for any $y \in [y_0, \mathcal{N}(y_0,0)]$ (with $|y_0|$ sufficiently small)

$$\frac{\mathcal{N}(y,0)-\mathcal{N}(y_0,0)}{\mathcal{N}^{[2]}(y_0,0)-\mathcal{N}(y_0,0)} \leq \frac{y-y_0}{\mathcal{N}(y_0,0)-y_0}.$$

Here we set $y_0 := \widetilde{x}_{k+1}(0)$ and $y := \widetilde{x}_{k+1}(0) + (\widetilde{x}_k(0) - \widetilde{x}_{k+1}(0)) \cdot Q_\infty(x) \in [y_0, \mathcal{N}(y_0,0)]$ (because $Q_\infty(x) \in [0,1]$) to get that

$$\frac{\mathcal{N}(\widetilde{x}_{k+1}(0) + (\widetilde{x}_k(0) - \widetilde{x}_{k+1}(0)) \cdot Q_\infty(x), 0) - \widetilde{x}_k(0)}{\widetilde{x}_{k-1}(0) - \widetilde{x}_k(0)} \leq Q_\infty(x),$$

since now $\frac{y-y_0}{\mathcal{N}(y_0,0)-y_0} = Q_\infty(x)$. But this means that

$$\mathcal{N}(\widetilde{x}_{k+1}(0) + (\widetilde{x}_k(0) - \widetilde{x}_{k+1}(0)) \cdot Q_\infty(x), 0) \leq \widetilde{x}_k(0) + (\widetilde{x}_{k-1}(0) - \widetilde{x}_k(0)) \cdot Q_\infty(x).$$

However, k is now fixed and applying the monotone increasing function $\mathcal{N}^{[k-1]}(\cdot,0)$ to both sides we get that $w_x(k+1) \leq w_x(k)$.

Step 2c. We now study

$$\mathcal{F}_x(k,M) := Q_{-M}(w_x(L(\varepsilon_k)),0)$$

for $k \in \mathbb{N}^+$ and $M \in \mathbb{N}^+$, where we recall that by construction $L(\varepsilon_k)$ is strictly increasing and tends to ∞ as $k \to \infty$. (Observe that the definition of $\mathcal{F}_x$ makes sense since $Q_\infty(x) \in [0,1]$, so $w_x(L(\varepsilon_k)) \in [\widetilde{x}_1(0), \widetilde{x}_0(0)]$.)

We know that

$$\lim_{k\to\infty} \mathcal{F}_x(k,M) = Q_{-M}(x^*,0)$$

if M is fixed, since $Q_{-M}(\cdot,0)$ is continuous and $\lim_{k\to\infty} w_x(L(\varepsilon_k)) = x^*$ by Step 2. Moreover, by using Step 2b and the fact that $Q_{-M}(\cdot,0)$ is monotone increasing, we get that $\mathcal{F}_x(\cdot,M)$ is monotone decreasing.

We also know that

$$\lim_{M\to\infty}\lim_{k\to\infty} \mathcal{F}_x(k,M) = Q_{-\infty}(x^*)$$

by the definition of Q_∞.

On the other hand if k is fixed then

$$\lim_{M\to\infty} \mathcal{F}_x(k,M) = Q_{-\infty}(w_x(L(\varepsilon_k)))$$

and $\mathcal{F}_x(k,\cdot)$ is monotone increasing by Step 2a. Further, since $Q_{-\infty}(\cdot)$ is continuous

$$\lim_{k\to\infty}\lim_{M\to\infty} \mathcal{F}_x(k,M) = Q_{-\infty}(x^*).$$

What this means pictorially is that $\mathcal{F}_x$ has a "saddle point at (∞,∞)", moreover, both iterated limits exist and are equal.

Step 2d. Let us show that the double limit also exists and, of course,

$$\lim_{\substack{k\to\infty\\ M\to\infty}} \mathcal{F}(k,M) = Q_{-\infty}(x^*).$$

Suppose to the contrary that there exist subsequences $k_n \to \infty$ and $M_n \to \infty$ as $n \to \infty$ such that

$$\lim_{n\to\infty} \mathcal{F}_x(k_n, M_n) = \Omega$$

with some $\Omega \neq Q_{-\infty}(x^*)$. Consider the case $\Omega < Q_{-\infty}(x^*)$ first and set $\delta := \frac{1}{2}(Q_{-\infty}(x^*) - \Omega) > 0$. Then there exists an index $\nu(\delta) \in \mathbb{N}^+$ such that for all $n > \nu(\delta)$

$$\mathcal{F}_x(k_n, M_n) \leq Q_{-\infty}(x^*) - \delta.$$

But $\mathcal{F}_x(k, M_n) \leq \mathcal{F}_x(k_n, M_n)$ if $k \geq k_n$, so if $n > \nu(\delta)$ then

$$Q_{-M_n}(x^*, 0) = \lim_{k\to\infty} \mathcal{F}_x(k, M_n) \leq Q_{-\infty}(x^*) - \delta,$$

so

$$Q_{-\infty}(x^*) = \lim_{n\to\infty} Q_{-M_n}(x^*, 0) \leq Q_{-\infty}(x^*) - \delta,$$

a contradiction. The other case $\Omega > Q_{-\infty}(x^*)$ can be treated similarly with natural modifications.

Step 2e. This last step is now easy, because we are entitled to evaluate the double limit along the "diagonal" and get

$$\lim_{n\to\infty} \mathcal{F}_x(n, L(\varepsilon_n) - 1) = Q_{-\infty}(x^*).$$

Observe, however, that

$$\mathcal{F}_x(n, L(\varepsilon_n) - 1) = \frac{\mathcal{N}^{[-(L(\varepsilon_n)-1)]}(w(L(\varepsilon_n)), 0) - \widetilde{x}_{L(\varepsilon_n)}(0)}{\widetilde{x}_{L(\varepsilon_n)-1}(0) - \widetilde{x}_{L(\varepsilon_n)}(0)} = Q_\infty(x),$$

meaning that $Q_\infty(x) = Q_{-\infty}(x^*)$, so, indeed, $x^* = Q_{-\infty}^{[-1]}(Q_\infty(x))$. ■

Corollary 2.8.7 *Suppose that z_N $(-\kappa \leq z_N)$ is a sequence converging to some $x \in [x_0(0), x_1(0)]$ as $N \to \infty$. Then*

$$\lim_{N\to\infty} \mathcal{N}^{[N-1]}(z_N, \alpha_N) = x^*.$$

Proof. We have

$$\left|\mathcal{N}^{[N-1]}(z_N, \alpha_N) - x^*\right| \leq \left|\mathcal{N}^{[N-1]}(z_N, \alpha_N) - \mathcal{N}^{[N-1]}(x, \alpha_N)\right| + \left|\mathcal{N}^{[N-1]}(x, \alpha_N) - x^*\right|.$$

The second term converges to 0 as $N \to \infty$ by Lemma 2.8.6, while the first term is estimated as

$$\left|\mathcal{N}^{[N-1]}(z_N, \alpha_N) - \mathcal{N}^{[N-1]}(x, \alpha_N)\right| \leq |z_N - x| \cdot \sup_{[\{z_N, x\}]} \left|\left(\mathcal{N}^{[N-1]}\right)'(\cdot, \alpha_N)\right|.$$

The right-hand side again converges to 0, because the supremum is bounded by an absolute constant according to Lemma 2.8.5. (Only a minor consideration is needed when $z_N \to x_1(0)$, so $z_N > x_1(0)$ may occur. Nevertheless, if N is large enough, then $-\kappa = x_0(0) < z_N \in [x_0(0), x_2(0)] \subset [x_0(\alpha_N), x_2(\alpha_N)]$, and it is clear that Lemma 2.8.5 is valid on this slightly larger interval too.) ■

After these preparations, let us prescribe $J(h, x, \alpha)$ on the vertical grid lines $(x, \alpha) \in [-\kappa, \kappa] \times \{\alpha_N : N \in \mathbb{N}^+\}$.

For some fixed $N \in \mathbb{N}^+$ and $x \in [x_0(\alpha_N), x_1(\alpha_N)]$, we set

$$J(h, x, \alpha_N) := \text{lin}(h, x, \alpha_N),$$

where $\text{lin}(h, \cdot, \alpha_N) : [x_0(\alpha_N), x_1(\alpha_N)] \to [y_0(\beta_N), y_1(\beta_N)]$ is the increasing linear homeomorphism between the first fundamental domains. (Of course, lin does not need to be linear at all—at this point we have some freedom, possibly leading to better closeness estimates in the future: see below.) The desired conjugacy equation

$$J(h, \mathcal{N}_\varphi(h, x, \alpha), \alpha) = \mathcal{N}_\Phi(h, J(h, x, \alpha), \beta)$$

now forces for any $k = 1, 2, \ldots, N-1$ and $x \in [x_k(\alpha_N), x_{k+1}(\alpha_N)]$ that

$$J(h, x, \alpha_N) := \mathcal{N}_\Phi^{[k]}\left(h, \text{lin}\left(h, \mathcal{N}_\varphi^{[-k]}\left(h, x, \alpha_N\right), \alpha_N\right), \beta_N\right),$$

where $\mathcal{N}_\Phi^{[k]}(h, x, \beta)$, for example, of course abbreviates $\left(\mathcal{N}_\Phi(h, \cdot, \beta)\right)^{[k]}(x)$. This definition extends the earlier one given in Section 2.8.

The conjugacy $J(h, x, 0)$ has already been defined for all $h \in [0, h_0]$ and $x \in [-\kappa, \kappa]$ in the previous sections (preceding the grid approach). From these constructions it is not hard to see that for any fixed $h > 0$ and $x < 0$, $J(h, x, \alpha_N) \to J(h, x, 0)$ as $N \to \infty$. (Essentially this depends on the fact that for any $x < 0$ we have $x \in [x_k(0), x_{k+1}(0)]$ for some fixed $k = k(h, x)$ and here both $\mathcal{N}_\Phi^{[k]}(h, x, \cdot)$ and $\mathcal{N}_\varphi^{[-k]}(h, x, \cdot)$ are continuous.)

Now let us consider a fixed $x \in [\widetilde{x}_1(0), \widetilde{x}_0(0)]$. If Lemma 2.8.6 is applied to the function $\mathcal{N} := \mathcal{N}_\varphi^{[-1]}$ (being monotone increasing and concave), we get that $\mathcal{N}_\varphi^{[-(N-1)]}(h, x, \alpha_N) \to x_*$ $(N \to \infty)$, where $x_* := Q_\infty^{[-1]}(Q_{-\infty}(x))$. Then, by Corollary 2.8.7, if $N \to \infty$, then

$$\mathcal{N}_\Phi^{[N-1]}\left(h, \text{lin}\left(h, \mathcal{N}_\varphi^{[-(N-1)]}\left(h, x, \alpha_N\right), \alpha_N\right), \beta_N\right) \to (\text{lin}(h, x_*, 0))^*,$$

where, of course, $\text{lin}(h, \cdot, 0)$ is the increasing homeomorphism between the intervals $[x_0(0), x_1(0)]$ and $[y_0(0), y_1(0)]$. Therefore, if we redefine $J(h, x, 0) := (\text{lin}(h, x_*, 0))^*$ for $x \in [\widetilde{x}_1(0), \widetilde{x}_0(0)]$, then $J(h, x, \alpha_N) \to J(h, x, 0)$ as $N \to \infty$. By using the conjugacy equation recursively, we can extend the definition of $J(h, x, 0)$ for any $x \in [\widetilde{x}_{k+1}(0), \widetilde{x}_k(0)]$ with a suitable and fixed $k = k(h, x)$, implying the continuity of J in the third variable along the grid sequence $\alpha = \alpha_N \to 0^+$, for any fixed $x > 0$. Finally, using strict monotonicity, it follows that $J(h, 0, \alpha_N)$ has to converge to $J(h, 0, 0) = 0$ as $N \to \infty$.

Since the values of $J(h, x, 0)$ $(x > 0)$ have been redefined, we should make a compatible extension of J in the left half-plane $(\alpha < 0)$ above the repelling fixed points. For example, the values of $J(h, x, \alpha)$ can be defined via a linear expansion (similar to the one given above) if $\alpha < 0$ and $x \in [\widetilde{x}_1(\alpha), \widetilde{x}_0(\alpha)] = [\widetilde{x}_1(\alpha), \kappa]$. Then for $x \in [\widetilde{x}_{k+1}(\alpha), \widetilde{x}_k(\alpha)]$ the conjugacy equation itself is used recursively to define J. (Notice that $|\widetilde{x}_1(\alpha) - \widetilde{y}_1(\alpha)| = \mathcal{O}(h^p)$, so this new definition will affect neither the earlier closeness estimates in the outer region of the $\alpha < 0$ half-plane, nor the continuity of J in the third variable for $x > 0$ and $\alpha \to 0^-$.)

Finally, we define $J(h, x, \alpha)$ between the grid lines, that is for $\alpha \in (\alpha_{N+1}, \alpha_N)$ $(N \in \mathbb{N}^+)$: a similar linear transformation is used on the first fundamental domain $[x_0(\alpha), x_1(\alpha)]$, then a recursive extension on $[x_k(\alpha), x_{k+1}(\alpha)]$ follows. At this point we again have great freedom in the definition: using any strictly monotone correspondence between $\alpha \in (\alpha_{N+1}, \alpha_N)$ and $\beta \in (\beta_{N+1}, \beta_N)$, we have that if $\alpha \in (\alpha_{N+1}, \alpha_N)$ and $x \in [x_k(\alpha), x_{k+1}(\alpha)] \subset [x_k(\alpha_{N+1}), x_{k+1}(\alpha_N)]$ (since $x_k(h, \cdot)$ is monotone increasing) for some k and N, then $J(h, x, \alpha) \in [y_k(\beta_{N+1}), y_{k+1}(\beta_N)]$.

For a closeness estimate, it is enough to consider case $x > 0$, $\alpha > 0$ (because it is seen from the proof of Lemma 2.8.2 that $|J(h,x,\alpha) - x| = \mathcal{O}(h^p)$ if $x \leq 0$ and $\alpha > 0$). But by analyzing the derivative of $\mathcal{N}_\Phi(h,y,\cdot)$, one sees that if $y > 0$, then

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\left(\mathcal{N}_\Phi^{[k]}\right)(h,y,\beta) \leq \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\mathcal{N}_\Phi^{[k+1]}\right)(h,y,\beta),$$

implying that $y_k(\beta_N) - y_k(\beta_{N+1}) \geq y_{k-1}(\beta_N) - y_{k-1}(\beta_{N+1})$. Now by rearranging we get that $y_k(\beta_N) - y_{k-1}(\beta_N) \geq y_k(\beta_{N+1}) - y_{k-1}(\beta_{N+1})$. However, since $\left(\frac{\mathrm{d}}{\mathrm{d}y}\mathcal{N}_\Phi\right)(h,y,\beta) \geq 1$ for $y > 0$, we obtain that $y_{k+1}(\beta_N) - y_k(\beta_N) \geq y_k(\beta_N) - y_{k-1}(\beta_N)$. From these we conclude that $y_{k+1}(\beta_N) - y_k(\beta_N) \geq y_k(\beta_{N+1}) - y_{k-1}(\beta_{N+1})$, so $y_{k+1}(\beta_N) - y_k(\beta_{N+1}) \geq y_k(\beta_N) - y_{k-1}(\beta_{N+1})$, or, recursively, $y_{k+1}(\beta_N) - y_k(\beta_{N+1}) \leq y_N(\beta_N) - y_{N-1}(\beta_{N+1})$. But $y_{N-1}(\beta_{N+1}) \geq y_{N-2}(\beta_N)$ and $y_N(\beta_N) - y_{N-2}(\beta_N) \leq c_1 h$ with an absolute constant $c_1 > 0$, hence $y_N(\beta_N) - y_{N-1}(\beta_{N+1}) \leq y_N(\beta_N) - y_{N-2}(\beta_N)$, so finally we see that

$$y_{k+1}(\beta_N) - y_k(\beta_{N+1}) \leq c_1 h.$$

A similar argument shows that

$$x_{k+1}(\alpha_N) - x_k(\alpha_{N+1}) \leq c_1 h.$$

But by Lemma 2.8.2 we see that $|y_{k+1}(\beta_N) - x_{k+1}(\alpha_N)| = \mathcal{O}(h^p)$ and $|y_k(\beta_{N+1}) - x_k(\alpha_{N+1})| = \mathcal{O}(h^p)$, and we know that $x \in [x_k(\alpha_{N+1}), x_{k+1}(\alpha_N)]$ and $J(h,x,\alpha) \in [y_k(\beta_{N+1}), y_{k+1}(\beta_N)]$. Therefore, we can conclude that if $x > 0$ and $\alpha > 0$, then

$$|J(h,x,\alpha) - x| = \mathcal{O}(h).$$

It is currently investigated how to construct $J(h,x,\alpha)$ for $\alpha \in [\alpha_{N+1}, \alpha_N)$ ($N \in \mathbb{N}^+$) such that $J(h,x,\alpha) \to J(h,x,0)$ if $\alpha \to 0^+$ arbitrarily and not only along the grid values, and secondly, how to refine the $\mathcal{O}(h)$ closeness estimate.

Chapter 3

Conjugacy in the discretized transcritical bifurcation

SUMMARY. THE PRESENT CHAPTER CONTAINS ANOTHER CASE STUDY, ON DISCRETIZATIONS NEAR A TRANSCRITICAL BIFURCATION POINT. UNLIKE IN THE FOLD CASE, THE ORIGIN IS NOW SURROUNDED BY FIXED POINTS, HENCE NO EXTRA DIFFICULTIES WILL ARISE. IN SECTION 3.1, TRANSCRITICAL CONDITIONS ARE EXAMINED, THEN IN SECTION 3.2 THE NECESSARY AND SUFFICIENT CONDITION ON THE DISCRETIZATION MAP $x \mapsto \varphi(h, x, \alpha)$ TO UNDERGO A TRANSCRITICAL BIFURCATION IS DISCUSSED AND NORMAL FORM TRANSFORMATIONS WITH CLOSENESS ESTIMATES FOR $x \mapsto \Phi(h, x, \alpha)$ AND $x \mapsto \varphi(h, x, \alpha)$ ARE CARRIED OUT. IN SECTION 3.3 A CONJUGACY BETWEEN FAMILIES $\Phi(h, \cdot, \alpha)$ AND $\varphi(h, \cdot, \widetilde{\alpha})$ IS CONSTRUCTED NEAR THE ORIGIN $x = 0$ FOR ANY FIXED h. THE BIFURCATION PARAMETER IS SHIFTED, SINCE DISCRETIZATIONS CAN RELOCATE THE BIFURCATION POINT, HOWEVER, ONLY BY $\mathcal{O}(h^p)$. FINALLY, IN SECTION 3.4 WE PROVE THAT THE DISTANCE BETWEEN THE CONJUGACY AND THE IDENTITY IS BOUNDED BY $\mathcal{O}(h^p)$ AND THIS ESTIMATE IS OPTIMAL.

3.1 Introduction

Suppose we have a one-dimensional ordinary differential equation

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

and its one-step discretization

$$X_{n+1} := \varphi(h, X_n, \alpha), \qquad n = 0, 1, 2, \dots, \tag{3.2}$$

where $\alpha \in \mathbb{R}$ is a scalar bifurcation parameter, $h > 0$ is the step-size of the sufficiently smooth one-step method $\varphi : \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of order $p \geq 1$, and the function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is of class C^{p+k+1} with $k \geq 5$ and uniformly bounded derivatives.

Since the numerical method is of order p, we have that

$$|\Phi(h, x, \alpha) - \varphi(h, x, \alpha)| \leq const \cdot h^{p+1}, \quad \forall\, h \in [0, h_0], \forall\, |x| \leq \varepsilon_0, \forall\, |\alpha| \leq \alpha_0, \tag{3.3}$$

where $\Phi(h, \cdot, \alpha) : \mathbb{R} \to \mathbb{R}$ is the time-h-map of the solution flow induced by (3.1) at parameter value α, further h_0, ε_0 and α_0 are some small positive constants.

Suppose that the origin $x = 0$, $\alpha = 0$ is an equilibrium as well as a *transcritical bifurcation point* for (3.1), that is the following conditions hold

$$f(0, \alpha) = 0, \quad \forall\, |\alpha| \leq \alpha_0,$$

$$f_x^B = 0, \quad f_{xx}^B \neq 0, \quad f_{x\alpha}^B \neq 0, \tag{3.4}$$

where subscripts x and α denote partial differentiation with respect to their corresponding variables, while superscript ^{B} abbreviates *evaluation at the bifurcation point*, that is, evaluation at $x = 0$ and $\alpha = 0$. (The evaluation is performed *after* taking all partial derivatives.)

The evaluation operator ^{B} will also be used for functions of three variables—h, x and α—when we evaluate a function at $h = 0$, $x = 0$ and $\alpha = 0$, as in $\Phi_{hx\alpha}^B$ abbreviating $\Phi_{hx\alpha}(0,0,0)$. (Here subscript h, of course, again stands for partial differentiation.)

For functions of three variables h, x and α, the evaluation operator ^{E} denotes *evaluation at general parameter values* h and α, where the dependence of ^{E} on h and α is suppressed. (Values of the parameters $h \in [0, h_0]$ and $\alpha \in [-\alpha_0, \alpha_0]$ can be arbitrary but fixed.) Thus, for example, the function $J(h, \cdot, \alpha)$ is abbreviated to J^E, if $J : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

Remark 3.1.1 Besides (3.4), there are other ways to formulate transcritical conditions. Since we are going to deal with maps induced by the differential equation, for the sake of the present remark, we first rewrite condition (3.4) for maps (this will be done in more detail in the next section): for a map $x \mapsto g(x, \alpha)$ to undergo a transcritical bifurcation near the origin it is sufficient that

$$g(0, \alpha) = 0, \quad \forall\, |\alpha| \leq \alpha_0,$$

$$g_x^B = 1, \quad g_{xx}^B \neq 0, \quad g_{x\alpha}^B \neq 0.$$

Notice that we have imposed $g(0, \alpha) = 0$, $\forall\, |\alpha| \leq \alpha_0$, which is *not* a *point condition* on g at the bifurcation point $(0, 0)$. The question naturally arises whether this latter condition can be relaxed.

The answer is affirmative, however, a little care should be taken. Let us have a look, for example, at [46], where, instead of $g(0, \alpha) = 0$, $\forall\, |\alpha| \leq \alpha_0$, the map g is simply required to satisfy

$$g^B = 0, \quad g_\alpha^B = 0, \quad g_x^B = 1, \quad g_{xx}^B \neq 0, \quad g_{x\alpha}^B \neq 0.$$

However, this is insufficient as illustrated by the map $x_{n+1} := g(x_n, \alpha)$ with

$$g(x, \alpha) := \alpha^2 + (1 + \alpha)x + x^2.$$

Since $(x, \alpha) = (0, 0)$ is the *only* fixed point of this map, clearly no bifurcation of fixed points can occur here. (The proof in [46] is correct, but it proves a slightly different proposition: [46] tacitly assumes that x can be factored out from g.) It is unfortunate that some other mathematical works or teaching materials also try to define transcritical bifurcation imperfectly as in [46].

Nevertheless, transcritical bifurcation for maps can be guaranteed via point conditions only. In [21], for example, a kind of discriminant condition is used: conditions

$$g^B = 0, \quad g_\alpha^B = 0, \quad g_x^B = 1, \quad g_{xx}^B \neq 0, \quad \left(g_{x\alpha}^B\right)^2 - g_{xx}^B \cdot g_{\alpha\alpha}^B > 0$$

imply a transcritical bifurcation near the origin. (It is a pity that instead of the last one-sided inequality, a "$\neq$" sign stands there in [21], but this is really just a typographical error as seen from the context.)

To summarize, condition $g(0, \alpha) = 0$ we have adopted is not the weakest one, but simple enough and still retains all the essential features of the problem.

Finally, it is instructive to compare this remark with its counterpart in the pitchfork bifurcation case.

3.2 Construction of the normal forms

In this section, we compute normal forms for the maps

$$x \mapsto \Phi(h, x, \alpha) \tag{3.5}$$

and

$$x \mapsto \varphi(h, x, \alpha) \tag{3.6}$$

near the equilibrium being also a transcritical bifurcation point.

To ensure that *the origin $x = 0$ is a fixed point also for the discretization map* (3.6), we assume that

$$\varphi(h, 0, \alpha) = 0 \tag{3.7}$$

holds for sufficiently small $h \geq 0$ and $|\alpha|$.

This condition is necessary for (3.6) to undergo a transcritical bifurcation near the origin, as illustrated by the following example.

Example 3.2.1 Suppose we have a map

$$\varphi(h, x, \alpha) := h^{2p+1} + (1 + h\alpha)x + h\,x^2.$$

For this φ, condition $\varphi(h, 0, \alpha) = 0$ does not hold, but φ satisfies (3.3) with

$$\Phi(h, x, \alpha) := (1 + h\alpha)x + h\,x^2.$$

This Φ undergoes a transcritical bifurcation, however, φ does not: its fixed points are given by $x_{\pm} = \frac{1}{2}\left(-\alpha \pm \sqrt{\alpha^2 - 4h^{2p}}\right)$, from which it is seen that φ does not have any fixed points in the interval $\alpha \in (-2h^p, 2h^p)$, so this map can not have a transcritical bifurcation near the origin.

Remark 3.2.1 It is well-known that all Runge-Kutta methods preserve equilibria, hence (3.7) is automatically satisfied for these discretizations.

The properties of the solution flow together with (3.3)–(3.4) imply for $h \geq 0$, $|x| \leq \varepsilon_0$ and $|\alpha| \leq \alpha_0$ that

$$\Phi(h, 0, \alpha) = 0, \quad \forall\, |\alpha| \leq \alpha_0, \tag{3.8}$$

$$\varphi(0, x, \alpha) = \Phi(0, x, \alpha) = x, \tag{3.9}$$

$$\Phi_h(h, x, \alpha) = f(\Phi(h, x, \alpha), \alpha), \tag{3.10}$$

$$\varphi_h(0, x, \alpha) = \Phi_h(0, x, \alpha). \tag{3.11}$$

Instead of (3.10), the shorter form $\Phi_h = f \circ \Phi$ will be used.

Lemma 3.2.1 *Under the assumptions above and for $h \in [0, h_0]$, $|x| \leq \varepsilon_0$, $|\alpha| \leq \alpha_0$, we have that*

$$\Phi(h, x, \alpha) = f_0(h, \alpha) + f_1(h, \alpha)x + f_2(h, \alpha)x^2 + \psi_3(h, x, \alpha)x^3,$$

where

$$\begin{aligned}
f_0(h, \alpha) &\equiv 0, \\
f_1(h, \alpha) &\equiv 1 + h\alpha \cdot f^B_{x\alpha} + h\alpha^2 \cdot \psi_1(h, \alpha), \qquad f^B_{x\alpha} \neq 0, \\
f_2(h, \alpha) &= \frac{1}{2} h \cdot f^B_{xx} + h\alpha \cdot \psi_2(h, \alpha), \qquad f^B_{xx} \neq 0, \\
\psi_3(h, x, \alpha) &= h \cdot \widehat{\psi}_3(h, x, \alpha)
\end{aligned}$$

hold with some smooth functions ψ_1, ψ_2 and $\widehat{\psi}_3$.

Proof. We expand Φ in a multivariate Taylor series about the equilibrium with the remainders in the integral form.

Since $f(0,\alpha)=0$ for all $|\alpha|$ sufficiently small, we have (3.8), hence $f_0(h,\alpha)$ should vanish.

As for f_1, we get that

$$f_1(h,\alpha)=\Phi^B_x+\alpha\cdot \mathrm{I}_{011}(\alpha)+h\cdot \mathrm{I}_{110}(h)+h\alpha\cdot\Phi^B_{hx\alpha}+$$

$$h\alpha^2\cdot \mathrm{I}_{112}(\alpha)+h^2\alpha\cdot \mathrm{I}_{211}(h)+h^2\alpha^2\cdot \mathrm{I}_{212}(h,\alpha),$$

where $\Phi^B_x=1$,

$$\mathrm{I}_{011}(\alpha)=\int_0^1\Phi_{x\alpha}(0,0,\tau\alpha)\mathrm{d}\tau\equiv 0,$$

$$\mathrm{I}_{110}(h)=\int_0^1\Phi_{hx}(\tau h,0,0)\mathrm{d}\tau\equiv 0,$$

because $\Phi_{hx}=(f\circ\Phi)_x=(f_x\circ\Phi)\cdot\Phi_x$.

It is easy to verify that $\Phi^B_{hx\alpha}=f^B_{x\alpha}$. Indeed, we have that

$$\Phi^B_{hx\alpha}=(f\circ\Phi)^B_{x\alpha}=((f_x\circ\Phi)_\alpha\cdot\Phi_x+(f_x\circ\Phi)\cdot\Phi_{x\alpha})^B=(f_x\circ\Phi)^B_\alpha,$$

because $\Phi^B_{x\alpha}=0$ and $\Phi^B_x=1$. But

$$(f_x\circ\Phi)^B_\alpha=f_{xx}(\Phi^B,0)\cdot\Phi^B_\alpha+f_{x\alpha}(\Phi^B,0)=f^B_{x\alpha},$$

since $\Phi_\alpha(0,x,\alpha)\equiv 0$.

The last three integrals read

$$\mathrm{I}_{112}(\alpha)=\int_0^1(1-\tau)\Phi_{hx\alpha\alpha}(0,0,\tau\alpha)\mathrm{d}\tau,$$

$$\mathrm{I}_{211}(h)=\int_0^1(1-\tau)\Phi_{hhx\alpha}(\tau h,0,0)\mathrm{d}\tau$$

and

$$\mathrm{I}_{212}(h,\alpha)=\int_0^1\int_0^1(1-\tau)(1-\sigma)\Phi_{hhx\alpha\alpha}(\tau h,0,\sigma\alpha)\mathrm{d}\tau\mathrm{d}\sigma.$$

We now show that $\mathrm{I}_{211}(h)$ vanishes, or, more precisely, that $\Phi_{hhx\alpha}(h,0,0)\equiv 0$ for every small $h\geq 0$. By direct differentiation we obtain that

$$\Phi_{hhx\alpha}=(f_{xx}\circ\Phi)_\alpha\cdot\Phi_x\cdot\Phi_h+(f_{xx}\circ\Phi)\cdot\Phi_{x\alpha}\cdot\Phi_h+$$

$$(f_{xx}\circ\Phi)\cdot\Phi_x\cdot\Phi_{h\alpha}+(f_x\circ\Phi)_\alpha\cdot\Phi_{hx}+(f_x\circ\Phi)\cdot\Phi_{hx\alpha}.$$

Here $\Phi_h(h,0,0)=f(\Phi(h,0,0),0)=f(0,0)=0$, so the first two terms above vanish. The third term is also zero, since

$$\Phi_{h\alpha}(h,0,0)=f_x(\Phi(h,0,0),0)\cdot\Phi_\alpha(h,0,0)+f_\alpha(\Phi(h,0,0),0)$$

but $\Phi(h,0,0)=0$ and $f_x(0,0)=0=f_\alpha(0,0)$. The fourth term is zero, because

$$\Phi_{hx}(h,0,0)=f_x(\Phi(h,0,0),0)\cdot\Phi_x(h,0,0)=0\cdot\Phi_x(h,0,0).$$

Finally, the fifth term vanishes due to the factor $f_x(\Phi(h,0,0),0)=0$.

By defining the smooth function $\psi_1(h,\alpha):=\mathrm{I}_{112}(\alpha)+h\cdot \mathrm{I}_{212}(h,\alpha)$, f_1 has the form stated above.

In the case of f_2, we have that

$$f_2(h,\alpha) = \frac{1}{2}\left(\Phi^B_{xx} + \alpha\cdot \mathrm{I}_{021}(\alpha) + h\cdot\Phi^B_{hxx} + h^2\cdot\mathrm{I}_{220}(h) + h\alpha\cdot\mathrm{I}_{121}(h,\alpha)\right),$$

where $\Phi^B_{xx} = 0$ and

$$\mathrm{I}_{021}(\alpha) = \int_0^1 \Phi_{xx\alpha}(0,0,\tau\alpha)\mathrm{d}\tau \equiv 0.$$

However,

$$\Phi^B_{hxx} = (f\circ\Phi)^B_{xx} = (f_{xx}\circ\Phi)^B\cdot\left((\Phi_x)^2\right)^B + (f_x\circ\Phi)^B\cdot\Phi^B_{xx} = f^B_{xx}\cdot 1 + 0 \neq 0.$$

Further,

$$\Phi_{hhxx} = (f_x\circ\Phi)_{xx}\cdot\Phi_h + 2(f_x\circ\Phi)_x\cdot\Phi_{hx} + (f_x\circ\Phi)\cdot\Phi_{hxx},$$

thus

$$\mathrm{I}_{220}(h) = \int_0^1 (1-\tau)\Phi_{hhxx}(\tau h,0,0)\mathrm{d}\tau \equiv 0.$$

Finally,

$$\mathrm{I}_{121}(h,\alpha) = \int_0^1\int_0^1 \Phi_{hxx\alpha}(\tau h,0,\sigma\alpha)\mathrm{d}\sigma\mathrm{d}\tau.$$

Thus, $\psi_2(h,\alpha) := \frac{1}{2}\mathrm{I}_{121}(h,\alpha)$ defines the desired smooth function.

For the remainder ψ_3, the integral formula gives

$$\psi_3(h,x,\alpha) = \frac{1}{2}\int_0^1 (1-\tau)^2\Phi_{xxx}(h,\tau x,\alpha)\mathrm{d}\tau. \tag{3.12}$$

But

$$\Phi_{xxx}(h,\tau x,\alpha) = \Phi_{xxx}(0,\tau x,\alpha) + h\cdot\int_0^1 \Phi_{hxxx}(\sigma h,\tau x,\alpha)\mathrm{d}\sigma$$

and $\Phi_{xxx}(0,\tau x,\alpha)\equiv 0$, so the lemma is proved. ■

Now we introduce a new parameter $\beta\equiv\beta(h,\alpha)$ by

$$\beta(h,\alpha) := \alpha\cdot f^B_{x\alpha} + \alpha^2\cdot\mathrm{I}_{112}(\alpha) + h\alpha^2\cdot\mathrm{I}_{212}(h,\alpha),$$

i.e., $\beta(h,\alpha) = \frac{f_1(h,\alpha)-1}{h}$.

We notice that $\beta(h,0) = 0$ and $\frac{\mathrm{d}}{\mathrm{d}\alpha}\beta(h,0) = f^B_{x\alpha} \neq 0$ independently of $h\in[0,h_0]$, thus the inverse function theorem guarantees the local existence and uniqueness of a smooth inverse function $\overline{\alpha}_0 \equiv \overline{\alpha}_0(h,\beta)$ of $\alpha\mapsto\beta(h,\alpha)$. Moreover, it is easy to see that the domain of definition of this inverse function contains a neighbourhood of the origin independent of $h\in[0,h_0]$. Further, $\overline{\alpha}_0(h,0)=0$, hence

$$\overline{\alpha}_0(h,\beta) = \beta\cdot\psi_a(h,\beta) \tag{3.13}$$

holds for $h\in[0,h_0]$ and $|\beta|$ small with some smooth function ψ_a.

Therefore (3.5) is transformed into the map

$$x\mapsto(1+h\beta)x + h\cdot q(h,\beta)x^2 + h\cdot\widehat{\psi}_3(h,x,\overline{\alpha}_0(h,\beta))x^3$$

with $q(h,\beta)\equiv\frac{1}{2}f^B_{xx} + \frac{1}{2}\overline{\alpha}_0(h,\beta)\cdot\mathrm{I}_{121}(h,\overline{\alpha}_0(h,\beta))$.

A final scaling $\eta := |q(h,\beta)|x$ with $s := \mathrm{sign}(q(h,0)) = \pm 1$ (being also independent of $h\in[0,h_0]$) yields the following normal form.

Lemma 3.2.2 *There are smooth invertible coordinate and parameter changes transforming the system*

$$x \mapsto \Phi(h, x, \alpha)$$

into

$$\eta \mapsto (1 + h\beta)\eta + s \cdot h\eta^2 + h\eta^3 \cdot \widehat{\eta}_3(h, \eta, \beta)$$

where $\widehat{\eta}_3(h, \eta, \beta) = \widehat{\psi}_3(h, x, \overline{\alpha}_0(h, \beta)) \cdot |q(h, \beta)|^{-2}$ *is a smooth function.* ■

Now let us consider the discretization map φ. We prove an analogous result to that of Lemma 3.2.1 first.

Lemma 3.2.3 *Under the assumptions of Lemma 3.2.1 together with (3.7) and for* $h \in [0, h_0]$, $|x| \le \varepsilon_0$, $|\alpha| \le \alpha_0$, *we have that*

$$\varphi(h, x, \alpha) = \widetilde{f}_0(h, \alpha) + \widetilde{f}_1(h, \alpha)x + \widetilde{f}_2(h, \alpha)x^2 + \chi_3(h, x, \alpha)x^3,$$

where

$$\begin{aligned} \widetilde{f}_0(h, \alpha) &= 0, \\ \widetilde{f}_1(h, \alpha) &= 1 + h\alpha \cdot f^B_{x\alpha} + h^{p+1} \cdot \chi_{10}(h) + h\alpha \cdot \chi_{11}(h, \alpha), \\ \widetilde{f}_2(h, \alpha) &= \frac{1}{2} h \cdot f^B_{xx} + h^{p+1} \cdot \chi_{20}(h) + h\alpha \cdot \chi_{21}(h, \alpha), \\ \chi_3(h, x, \alpha) &= h \cdot \widetilde{\chi}_3(h, x, \alpha) \end{aligned}$$

hold with some smooth functions χ_{10}, χ_{11}, χ_{20}, χ_{21} *and* $\widetilde{\chi}_3$. *Moreover, for* $h \in [0, h_0]$, $|x| \le \varepsilon_0$ *and for* $|\alpha| \le \alpha_0$,

$$|\psi_3(h, x, \alpha) - \chi_3(h, x, \alpha)| \le const \cdot h^{p+1}. \tag{3.14}$$

Proof. By (3.7), we have that $\widetilde{f}_0(h, \alpha) \equiv 0$.

The remainders of the Taylor series are also represented by integrals and denoted—analogously to the proof of Lemma 3.2.1—by $\widetilde{\mathrm{I}}$'s. These integrals, of course, now always contain φ instead of Φ.

As for $\widetilde{f}_1$, by (3.9) one has that $\varphi^B_x = 1$ and $\widetilde{\mathrm{I}}_{011}(\alpha) \equiv 0$, further, we get that $\varphi^B_{hx\alpha} = \Phi^B_{hx\alpha} = f^B_{x\alpha} \neq 0$, hence

$$\widetilde{f}_1(h, \alpha) = 1 + h \cdot \widetilde{\mathrm{I}}_{110}(h) + h\alpha \cdot f^B_{x\alpha} + \\ h\alpha^2 \cdot \widetilde{\mathrm{I}}_{112}(\alpha) + h^2\alpha \cdot \widetilde{\mathrm{I}}_{211}(h) + h^2\alpha^2 \cdot \widetilde{\mathrm{I}}_{212}(h, \alpha).$$

Since f is at least C^{p+4}, from [18] we obtain that

$$\left| f_1(h, \alpha) - \widetilde{f}_1(h, \alpha) \right| \le const \cdot h^{p+1}. \tag{3.15}$$

Evaluating this at $\alpha = 0$ yields $|h \cdot \widetilde{\mathrm{I}}_{110}(h)| \le const \cdot h^{p+1}$. The smooth functions χ_{10} and χ_{11} are defined as

$$\chi_{10}(h) := \frac{h \cdot \widetilde{\mathrm{I}}_{110}(h)}{h^{p+1}}$$

and

$$\chi_{11}(h, \alpha) := \alpha \cdot \widetilde{\mathrm{I}}_{112}(\alpha) + h \cdot \widetilde{\mathrm{I}}_{211}(h) + h\alpha \cdot \widetilde{\mathrm{I}}_{212}(h, \alpha).$$

(It can be easily proved that $\widetilde{\mathrm{I}}_{112}(\alpha) \equiv \mathrm{I}_{112}(\alpha)$, but this property will not be needed later.)

Considering $\widetilde{f}_2$, we obtain that $\varphi^B_{xx} = 0$ and $\widetilde{\mathrm{I}}_{021}(\alpha) \equiv 0$. By differentiating (3.11) we see that $\varphi^B_{hxx} = \Phi^B_{hxx} = f^B_{xx} \neq 0$, thus

$$\widetilde{f}_2(h, \alpha) = \frac{1}{2} \left(h \cdot f^B_{xx} + h^2 \cdot \widetilde{\mathrm{I}}_{220}(h) + h\alpha \cdot \widetilde{\mathrm{I}}_{121}(h, \alpha) \right),$$

and again, using $f \in C^{p+5}$ and [18]

$$\left| f_2(h,\alpha) - \widetilde{f}_2(h,\alpha) \right| \leq const \cdot h^{p+1}. \tag{3.16}$$

Evaluating this at $\alpha = 0$, we see that $|h^2 \cdot \widetilde{\mathrm{I}}_{220}(h)| \leq const \cdot h^{p+1}$, so we can set

$$\chi_{20}(h) := \frac{1}{2} \cdot \frac{h^2 \cdot \widetilde{\mathrm{I}}_{220}(h)}{h^{p+1}}$$

and

$$\chi_{21}(h,\alpha) := \frac{1}{2} \cdot \widetilde{\mathrm{I}}_{121}(h,\alpha)$$

to obtain two smooth functions.

To prove the product form of the remainder χ_3, we use the same argument as in (3.12). Finally, for (3.14) we take into account $f \in C^{p+6}$ and [18] again to get

$$|\psi_3(h,x,\alpha) - \chi_3(h,x,\alpha)| = \left| \frac{1}{2} \int_0^1 (1-\tau)^2 \left(\Phi_{xxx}(h,\tau x,\alpha) - \varphi_{xxx}(h,\tau x,\alpha) \right) \mathrm{d}\tau \right| \leq$$

$$const \cdot h^{p+1} \cdot \frac{1}{2} \int_0^1 (1-\tau)^2 \mathrm{d}\tau,$$

completing the proof of the lemma. ■

Now we the introduce the analogue of parameter β. Set

$$\widetilde{\beta} \equiv \widetilde{\beta}(h,\alpha) := \widetilde{\mathrm{I}}_{110}(h) + \alpha \cdot f^B_{x\alpha} + \alpha^2 \cdot \widetilde{\mathrm{I}}_{112}(\alpha) + h\alpha \cdot \widetilde{\mathrm{I}}_{211}(h) + h\alpha^2 \cdot \widetilde{\mathrm{I}}_{212}(h,\alpha).$$

We will show that the function $\widetilde{\beta}(h,\cdot)$ is locally invertible at the origin for every $h \geq 0$ small enough, and its inverse function, $\widetilde{\alpha}(h,\cdot)$ is $\mathcal{O}(h^p)$-close to $\overline{\alpha}_0(h,\cdot)$, *i.e.* to the inverse of $\beta(h,\cdot)$. As in Section 2.2, we will use the same quantitative inverse function theorem, see Lemma 2.2.4. (Now a letter G will play the role of $\widetilde{F}$ in that lemma.) We set

$$G(h,\beta,\alpha) := \beta - \widetilde{\beta}(h,\alpha).$$

In order to check the conditions of the lemma, define $\kappa_1 := \frac{1}{2}|f^B_{x\alpha}| > 0$ and $\kappa_2 := \frac{1}{2}\kappa_1$. We have that

$$\frac{\partial G}{\partial \alpha}(h,\beta,\alpha) = f^B_{x\alpha} + 2\alpha \cdot \widetilde{\mathrm{I}}_{112}(\alpha) + \alpha^2 \frac{\mathrm{d}}{\mathrm{d}\alpha} \widetilde{\mathrm{I}}_{112}(\alpha) +$$

$$h \cdot \widetilde{\mathrm{I}}_{211}(h) + 2h\alpha \cdot \widetilde{\mathrm{I}}_{212}(h,\alpha) + h\alpha^2 \frac{\mathrm{d}}{\mathrm{d}\alpha} \widetilde{\mathrm{I}}_{212}(h,\alpha).$$

Thus

$$\left| \frac{\partial G}{\partial \alpha}(h,\beta,\alpha) - \frac{\partial G}{\partial \alpha}(h,\beta,\overline{\alpha}_0(h,\beta)) \right| \leq \kappa_2$$

holds by smoothness of the functions $\widetilde{\mathrm{I}}$'s provided that $|\alpha - \overline{\alpha}_0(h,\beta)| \leq r_1$ and $h < r_2$ are small enough. It is also seen that

$$\left| \frac{\partial G}{\partial \alpha}(h,\beta,\overline{\alpha}_0(h,\beta)) \right| \geq \kappa_1,$$

if $h, |\beta| < r_2$ are small enough, taking also into account (3.13). Finally, using that $\overline{\alpha}_0(h,\cdot)$ is the inverse function of $\beta(h,\cdot)$, we get that

$$|G(h,\beta,\overline{\alpha}_0(h,\beta))| = \left| \beta - \widetilde{\beta}(h,\overline{\alpha}_0(h,\beta)) \right| = \left| \beta(h,\overline{\alpha}_0(h,\beta)) - \widetilde{\beta}(h,\overline{\alpha}_0(h,\beta)) \right|.$$

But (3.15) implies that

$$|\beta(h,\alpha) - \widetilde{\beta}(h,\alpha)| \le const \cdot h^p, \tag{3.17}$$

hence $|G(h,\beta,\overline{\alpha}_0(h,\beta))| \le const \cdot h^p$ and also $|G(h,\beta,\overline{\alpha}_0(h,\beta))| \le (\kappa_1 - \kappa_2)\cdot r_1$ if $h < r_2$ is small enough.

Therefore, Lemma 2.2.4 is applicable in our situation and we get a unique zero $\widetilde{\alpha}(h,\beta)$ of $G(h,\beta,\cdot)$, which—by the construction of G—is the inverse function of $\alpha \mapsto \widetilde{\beta}(h,\alpha)$. Furthermore,

$$|\widetilde{\alpha}(h,\beta) - \overline{\alpha}_0(h,\beta)| \le const \cdot h^p \tag{3.18}$$

holds for $h \in [0,h_0]$ and $|\beta|$ sufficiently small.

As a conclusion, (3.6) becomes

$$x \mapsto (1 + h\widetilde{\beta})x + h\cdot \widetilde{q}(h,\widetilde{\beta})x^2 + h\cdot \widetilde{\chi}_3(h,x,\widetilde{\alpha}(h,\widetilde{\beta}))x^3$$

with $\widetilde{q}(h,\widetilde{\beta}) \equiv \frac{1}{2}\left(f^B_{xx} + h\cdot \widetilde{\mathrm{I}}_{220}(h) + \widetilde{\alpha}(h,\widetilde{\beta})\cdot \widetilde{\mathrm{I}}_{121}(h,\widetilde{\alpha}(h,\widetilde{\beta}))\right)$.

We claim that

$$\left|\widetilde{q}(h,\widetilde{\beta}) - q(h,\beta)\right| \le const \cdot h^p \tag{3.19}$$

also holds. But this is a consequence of inequalities (3.18), (3.16) and the smoothness (and boundedness) of the functions I_{121} and $\widetilde{\mathrm{I}}_{121}$ when combined with standard triangle inequalities and the mean value theorem.

By applying a final scaling

$$\widetilde{\eta} := |\widetilde{q}(h,\widetilde{\beta})|x$$

with $s := \mathrm{sign}(\widetilde{q}(h,0)) = \pm 1$ (being independent of $h \in [0,h_0]$ for h_0 small enough, due to (3.18) evaluated at $\beta = 0$, (3.13) and the boundedness of the function $\widetilde{\mathrm{I}}_{121}$) and defining

$$\widetilde{\eta}_3(h,\widetilde{\eta},\widetilde{\beta}) := \widetilde{\chi}_3(h,x,\widetilde{\alpha}(h,\widetilde{\beta}))\cdot |\widetilde{q}(h,\widetilde{\beta})|^{-2},$$

we have derived a normal form for (3.6) in the theorem below.

For the closeness estimates in the theorem, we should only verify that

$$\left|\widehat{\eta}_3(h,\eta,\beta) - \widetilde{\eta}_3(h,\widetilde{\eta},\widetilde{\beta})\right| \le const \cdot h^p.$$

This estimate, however, is a simple consequence of (3.19) and the fact that

$$\left|\widehat{\psi}_3(h,x,\overline{\alpha}_0(h,\beta)) - \widetilde{\chi}_3(h,x,\widetilde{\alpha}(h,\widetilde{\beta}))\right| \le const \cdot h^p.$$

(For this last inequality, (3.14), the smoothness of $\widehat{\psi}_3$, a standard triangle inequality and the mean value theorem suffice.)

Theorem 3.2.4 *Suppose that conditions (3.1)–(3.4) and (3.7) hold. Then there are smooth invertible coordinate and parameter changes transforming the system*

$$x \mapsto \varphi(h,x,\alpha)$$

into

$$\widetilde{\eta} \mapsto (1 + h\widetilde{\beta})\widetilde{\eta} + s\cdot h\widetilde{\eta}^2 + h\widetilde{\eta}^3\cdot \widetilde{\eta}_3(h,\widetilde{\eta},\widetilde{\beta})$$

where $\widetilde{\eta}_3$ is a smooth function.

Moreover, the smooth invertible coordinate and parameter changes above and those in Lemma 3.2.2 are $\mathcal{O}(h^p)$-close to each other, further

$$|\widehat{\eta}_3 - \widetilde{\eta}_3| \le const \cdot h^p \qquad \blacksquare$$

Finally, we apply a parameter shift $\widetilde{\beta} \mapsto \beta$ to the normal form in the theorem above, being $\mathcal{O}(h^p)$-close to the identity due to (3.17). So from now on we will use the bifurcation parameter α again instead of β and $\widetilde{\beta}$. To simplify our notation further, instead of η and $\widetilde{\eta}$ the letter x will be used.

3.3 Construction of the conjugacy

We have thus the following normal forms

$$\mathcal{N}_\Phi(h,x,\alpha) = (1+h\alpha)x + s\cdot hx^2 + hx^3\,\widehat{\eta}_3(h,x,\alpha) \tag{3.20}$$

$$\mathcal{N}_\varphi(h,x,\alpha) = (1+h\alpha)x + s\cdot hx^2 + hx^3\,\widetilde{\eta}_3(h,x,\alpha) \tag{3.21}$$

with $s=1$ or $s=-1$, where $\widehat{\eta}_3$ and $\widetilde{\eta}_3$ are smooth functions. Let $K>0$ denote a uniform bound on $\left|\frac{\mathrm{d}^i}{\mathrm{d}x^i}\,\eta(h,\cdot,\alpha)\right|$ $(i\in\{0,1,2\}, \eta\in\{\widehat{\eta}_3,\widetilde{\eta}_3\})$ in a neighbourhood of the origin for any small $h>0$ and $|\alpha|$, as well as a uniform bound on $\left|\frac{\mathrm{d}}{\mathrm{d}\alpha}\,\eta(h,x,\cdot)\right|$ $(\eta\in\{\widehat{\eta}_3,\widetilde{\eta}_3\})$ in a neighbourhood of the origin for any small $h>0$ and $|x|$. We also have that there exists a constant $c>0$ such that

$$|\mathcal{N}_\Phi(h,x,\alpha) - \mathcal{N}_\varphi(h,x,\alpha)| \le c\cdot h^{p+1}|x|^3 \tag{3.22}$$

holds for all sufficiently small $h>0$, $|x|\ge 0$ and $|\alpha|\ge 0$. Throughout the section, c will denote this particular positive constant. (Other generic constants, if needed, are denoted by *const.*)

We will consider the case $s=1$, the other one is similar. Then it is easy to see that $\omega_{\Phi,0}(h,\alpha)\equiv 0$ is an attracting fixed point of the map $\mathcal{N}_\Phi(h,\cdot,\alpha)$ for $\alpha<0$, and repelling for $\alpha>0$. For any fixed $h>0$ and $\alpha\in[-\alpha_0,\alpha_0]\setminus\{0\}$, this map possesses another fixed point, denoted by $\omega_{\Phi,+}\equiv\omega_{\Phi,+}(h,\alpha)>0$ (if $\alpha<0$) and $\omega_{\Phi,-}\equiv\omega_{\Phi,-}(h,\alpha)<0$ (if $\alpha>0$). It is seen that $\omega_{\Phi,+}$ is repelling and $\omega_{\Phi,-}$ is attracting. The two branches of fixed points, $\omega_{\Phi,0}(h,\alpha)$ and $\omega_{\Phi,\pm}(h,\alpha)$ merge at $\alpha=0$.

Analogous results hold, of course, for the map $\mathcal{N}_\varphi(h,\cdot,\alpha)$. Its fixed points are denoted by $\omega_{\varphi,0}$ and $\omega_{\varphi,-}$ (or $\omega_{\varphi,+}$).

We will construct a conjugacy in a natural way and prove optimal closeness estimates in the $x\le 0$ region—the $x>0$ case is similar due to symmetry.

In what follows, we suppose that

$$\begin{gathered} 0<h\le h_0 := \frac{1}{5}, \\ |x|\le\varepsilon_0 := \min\left(\frac{1}{25},\frac{1}{25K}\right) \text{ and} \\ |\alpha|\le\alpha_0 := \min\left(\frac{1}{51},\frac{1}{51K}\right). \end{gathered} \tag{3.23}$$

With these values of h_0, ε_0 and α_0, all constructions and proofs below can be carried out. (There is only one constraint which has not been taken into account explicitly: if the domain of definition of the functions $\widehat{\eta}_3$ and $\widetilde{\eta}_3$ is smaller than $(0,h_0]\times[-\varepsilon_0,\varepsilon_0]\times[-\alpha_0,\alpha_0]$ given above, then h_0, ε_0 or α_0 should be decreased further suitably.)

Lemma 3.3.1 *For every* $0<h\le h_0$ *and* $0<\alpha\le\alpha_0$ *we have that*

$$\{\omega_{\varphi,-},\omega_{\Phi,-}\}\subset\left(-\frac{3}{2}\alpha,-\frac{6}{7}\alpha\right).$$

Proof. By definition, $\omega_{\varphi,-}$ solves $\alpha+x+x^2\cdot\widetilde{\eta}_3(h,x,\alpha)=0$. But $|x|\le\frac{1}{6K}$ implies $\frac{2}{3}\le 1+x\,\widetilde{\eta}_3\le\frac{7}{6}$, so

$$-\frac{3\alpha}{2}\le\omega_{\varphi,-}=\frac{-\alpha}{1+\omega_{\varphi,-}\cdot\widetilde{\eta}_3(h,\omega_{\varphi,-},\alpha)}\le-\frac{6\alpha}{7}.$$

The proof for $\omega_{\Phi,-}$ is similar. ■

By iterating one of the normal forms, say $\mathcal{N}_\varphi(h,\cdot,\alpha)$, let us define three sequences x_n, y_n and z_n. For $\alpha>0$, let $x_n\equiv x_n(h,\alpha)$ be defined as

$$x_{n+1}:=\mathcal{N}_\varphi(h,x_n,\alpha),\quad n=0,1,2,\ldots$$

with $x_0:=-\frac{\alpha}{3}$, further, let $y_n\equiv y_n(h,\alpha)$ be defined as

$$y_n:=\left(\mathcal{N}_\varphi^E\right)^{[-n]}(x_0),\quad n=0,1,2,\ldots,$$

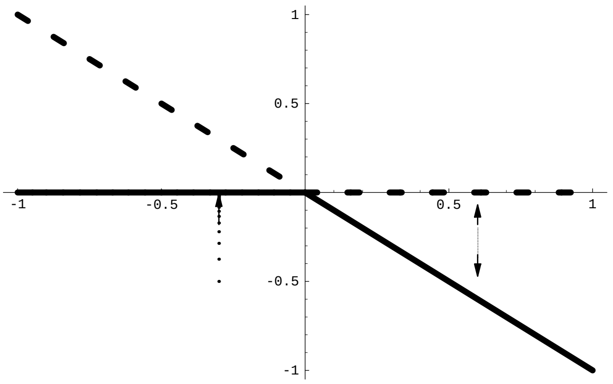
so $y_0:=x_0$, and set $y_{-1}:=x_1$. Finally, for all $\alpha\in[-\alpha_0,\alpha_0]$ define $z_n\equiv z_n(h,\alpha)$ as

$$z_n:=\left(\mathcal{N}_\varphi^E\right)^{[n]}(z_0),\quad n=0,1,2,\ldots,$$

with $z_0<0$ being independent of h and α such that $2\alpha_0<|z_0|<\frac{1}{2K}$ holds. An appropriate choice for z_0 is, *e.g.*, $z_0:=-\varepsilon_0$.

Simple calculations show that, for example, under conditions (3.23), both $\mathcal{N}_\varphi^E$ and $\mathcal{N}_\Phi^E$ (together with their inverses) are monotone increasing, further $|\alpha|<\frac{6}{K}$ implies $x_0(\alpha)>x_1(h,\alpha)$ and $2\alpha_0<|z_0|<\frac{1}{2K}$ implies $z_0<z_1(h,\alpha)$. This means that x_n is monotone decreasing, y_n is monotone increasing (if $\alpha>0$ and $n\ge0$), and $\lim_{n\to\infty}x_n(h,\alpha)=\omega_{\varphi,-}$, while $\lim_{n\to\infty}y_n(h,\alpha)=\omega_{\varphi,0}$. Moreover, z_n is monotone increasing, further, for $\alpha>0$, $\lim_{n\to\infty}z_n(h,\alpha)=\omega_{\varphi,-}$ and for $\alpha\le0$, $\lim_{n\to\infty}z_n(h,\alpha)=\omega_{\varphi,0}$.

The following figure shows the branch of stable and unstable fixed points of $\mathcal{N}_\varphi^E$ in the (α,x)-plane together with the first few terms of the inner sequences ($x_n(h,\alpha)$ and $y_n(h,\alpha)$), and the outer sequence $z_n(h,\alpha)$ with some $h>0$ and α fixed. The arrows indicate the direction of the sequences.



A homeomorphism J^E satisfying the conjugacy equation

$$J^E\circ\mathcal{N}_\varphi^E=\mathcal{N}_\Phi^E\circ J^E\tag{3.24}$$

is now piecewise defined on the fundamental domains, *i.e.* on $[x_{n+1},x_n]$, $[y_n,y_{n+1}]$ and $[z_n,z_{n+1}]$ ($n\in\mathbb{N}$), for any fixed $0<h\le h_0$ and $-\alpha_0\le\alpha\le\alpha_0$.

We first consider the region between the fixed points for $0 < \alpha \leq \alpha_0$.

Let $J^E(x_0) := x_0$ and $J^E(x_1) := \mathcal{N}_\Phi^E(x_0)$. For $x \in [x_1, x_0]$ extend J^E linearly. For $n \geq 1$ and $x \in [x_{n+1}, x_n]$, we recursively set

$$J^E(x) := \left(\mathcal{N}_\Phi^E \circ J^E \circ \left(\mathcal{N}_\varphi^E\right)^{[-1]}\right)(x),$$

while for $n \geq 0$ and $x \in [y_n, y_{n+1}]$, we let

$$J^E(x) := \left(\left(\mathcal{N}_\Phi^E\right)^{[-1]} \circ J^E \circ \mathcal{N}_\varphi^E\right)(x).$$

(Since $[y_{-1}, y_0] \equiv [x_1, x_0]$, these two definitions are compatible.) Finally, set

$$J^E(\omega_{\varphi,-}) := \omega_{\Phi,-}$$

and

$$J^E(\omega_{\varphi,0}) := \omega_{\Phi,0}.$$

Then J^E is continuous, strictly monotone increasing on $[\omega_{\varphi,-}, 0]$, since it is a composition of three such functions, and satisfies (3.24).

In the outer region, *i.e.* below the fixed points, fix $z_0 < 0$ ($2\alpha_0 < |z_0| < \frac{1}{2K}$), then for $\alpha \in [-\alpha_0, \alpha_0]$ the construction of J^E is analogous to the construction above with the sequence x_n: this time z_n plays the role of x_n. (Of course, now the counterpart of the sequence y_n is not needed.) Then the function J^E becomes continuous, strictly monotone increasing on $[z_0, \omega_{\varphi,-}]$ ($0 < \alpha \leq \alpha_0$) and $[z_0, \omega_{\varphi,0}]$ (for $-\alpha_0 \leq \alpha \leq 0$), and satisfies (3.24).

The construction of J^E—with the appropriate and natural modifications—in the upper half-plane $x > 0$ is analogous to the one presented above.

3.4 The closeness estimate for the conjugacy

3.4.1 Optimality at the fixed points

We first prove that the constructed conjugacy J^E is $\mathcal{O}(h^p\,\alpha^2)$-close to the identity at the fixed points $\omega_{\varphi,-}(h,\alpha)$, further, an explicit example will show that this estimate is optimal in h and α.

Since fixed points must be mapped into nearby fixed points by the conjugacy and we are going to prove $\mathcal{O}(h^p)$-closeness in the whole domain, the result above means that our estimates of $|\,id - J^E|$ near a transcritical bifurcation point are optimal in h.

The following auxiliary estimate will frequently be used.

Lemma 3.4.1 *For any $0 < h \leq h_0$, $-\varepsilon_0 \leq x < 0$ and $-\alpha_0 \leq \alpha \leq \alpha_0$, we have that*

$$(\mathcal{N}_\Phi^E)'(x) \leq 1 + h\alpha + \frac{7}{4}hx.$$

Proof. The conditions in (3.23) have been set up to imply this inequality, too. ■

Lemma 3.4.2 *For any $0 < h \leq h_0$ and $0 < \alpha \leq \alpha_0$ (satisfying (3.23)), we have that*

$$|\,\omega_{\varphi,-} - \omega_{\Phi,-}| \leq \frac{27}{4}c \cdot h^p\,\alpha^2.$$

Proof.

$$|\,id - J^E|(\omega_{\varphi,-}(h,\alpha)) \le |\mathcal{N}^E_\varphi(\omega_{\varphi,-}) - \mathcal{N}^E_\Phi(\omega_{\varphi,-})| + |\mathcal{N}^E_\Phi(\omega_{\varphi,-}) - \mathcal{N}^E_\Phi(\omega_{\Phi,-})| \le$$

$$c \cdot h^{p+1}|\omega_{\varphi,-}|^3 + \left(\sup_{[\{\,\omega_{\varphi,-},\omega_{\Phi,-}\}]} (\mathcal{N}^E_\Phi)' \right) |\,\omega_{\varphi,-} - \omega_{\Phi,-}| \le$$

$$\frac{27}{8} c \cdot h^{p+1}\alpha^3 + \left(1 - \frac{h\alpha}{2}\right) |\,\omega_{\varphi,-} - \omega_{\Phi,-}|,$$

by Lemma 3.3.1, (3.22) and Lemma 3.4.1. Solving the above inequality for $|\,\omega_{\varphi,-} - \omega_{\Phi,-}| \equiv |\,id - J^E|(\omega_{\varphi,-})$ yields the desired result. ■

Remark 3.4.1 on optimality. The next example shows that the distance of fixed points of normal forms satisfying (3.22) can be bounded from *below* by $\mathcal{O}(h^p)$ $(h \to 0)$.

Indeed, set $\mathcal{N}_\Phi(h,x,\alpha) := (1+h\alpha)x + hx^2$ and $\mathcal{N}_\varphi(h,x,\alpha) := (1+h\alpha)x + hx^2 + h^{p+1}x^3$. Then these maps satisfy (3.22) in a neighbourhood of the origin, further, $\omega_{\Phi,-} = -\alpha$ and $\omega_{\varphi,-} = \frac{-1+\sqrt{1-4h^p\,\alpha}}{2h^p}$. Using inequality $1 + \frac{t}{2} - \frac{t^2}{4} \le \sqrt{1+t} \le 1 + \frac{t}{2} - \frac{t^2}{8}$ for $-\frac{1}{2} \le t \le 0$, one sees that

$$|\,\omega_{\varphi,-} - \omega_{\Phi,-}| \ge h^p\,\alpha^2,$$

if, for example, $h \le 1$ and $\alpha \le \frac{1}{8}$.

3.4.2 The inner region

Now the closeness estimate in $(\omega_{\varphi,-}, x_0]$ is proved for any fixed $0 < h \le h_0$ and $0 < \alpha \le \alpha_0$. It is clear that $\sup_{(\omega_{\varphi,-},x_0]} |\,id - J^E| = \sup_{n\in\mathbb{N}} \sup_{[x_{n+1},x_n]} |\,id - J^E|$.

Since $x_0 = J^E(x_0)$, we have that

$$\sup_{[x_1,x_0]} |\,id - J^E| = |x_1 - J^E(x_1)| = |\mathcal{N}^E_\varphi(x_0) - \mathcal{N}^E_\Phi(x_0)| \le$$

$$c \cdot h^{p+1}|x_0|^3 = \frac{c}{27} h^{p+1}\alpha^3,$$

while for $n \ge 1$

$$\sup_{[x_{n+1},x_n]} |\,id - J^E| \le \sup_{[x_{n+1},x_n]} \left| \mathcal{N}^E_\varphi \circ (\mathcal{N}^E_\varphi)^{[-1]} - \mathcal{N}^E_\Phi \circ (\mathcal{N}^E_\varphi)^{[-1]} \right| +$$

$$\sup_{[x_{n+1},x_n]} \left| \mathcal{N}^E_\Phi \circ (\mathcal{N}^E_\varphi)^{[-1]} - \mathcal{N}^E_\Phi \circ J^E \circ (\mathcal{N}^E_\varphi)^{[-1]} \right| =$$

$$\sup_{[x_n,x_{n-1}]} \left| \mathcal{N}^E_\varphi - \mathcal{N}^E_\Phi \right| + \sup_{[x_n,x_{n-1}]} \left| \mathcal{N}^E_\Phi - \mathcal{N}^E_\Phi \circ J^E \right| \le$$

$$\sup_{[x_n,x_{n-1}]} \left| \mathcal{N}^E_\varphi - \mathcal{N}^E_\Phi \right| + \sup_{x\in[x_n,x_{n-1}]} \left(\left(\sup_{[\{x,J^E(x)\}]} (\mathcal{N}^E_\Phi)' \right) |\,x - J^E(x)| \right) \le$$

$$c \cdot h^{p+1}|x_n|^3 + \left(1 + h\alpha + \frac{7}{4} h \max\left(x_{n-1}, J^E(x_{n-1})\right)\right) \sup_{[x_n,x_{n-1}]} |\,id - J^E|,$$

the last inequality being true due to

$$\sup_{[\{x,J^E(x)\}]} (\mathcal{N}^E_\Phi)' \le \sup_{[\{x,J^E(x)\}]} (1 + h\alpha + \frac{7}{4} h \cdot id) \le 1 + h\alpha + \frac{7}{4} h \max\left(x, J^E(x)\right)$$

taking into account Lemma 3.4.1, then using the fact that the functions id and J^E are increasing.

From these we have for $n \geq 1$ that

$$\sup_{[x_{n+1},x_n]} |id - J^E| \leq c \cdot h^{p+1} \sum_{i=0}^{n} |x_i|^3 \prod_{j=i}^{n-1} \left(1 + h\alpha + \frac{7}{4} h \max\left(x_j, J^E(x_j)\right)\right),$$

where $\prod_{j=n}^{n-1}$ is understood to be 1.

So in order to prove that the conjugacy J^E is $\mathcal{O}(h^p)$-close to the identity on the interval $(\omega_{\varphi,-}, x_0]$ for any $h \in (0, h_0]$ and $\alpha \in (0, \alpha_0]$, it is enough to show that

$$\sup_{h\in(0,h_0]} \sup_{\alpha\in(0,\alpha_0]} \sup_{n\in\mathbb{N}} h \sum_{i=0}^{n} |x_i|^3 \prod_{j=i}^{n-1} \left(1 + h\alpha + \frac{7}{4} h \max\left(x_j, J^E(x_j)\right)\right) \leq const \tag{3.25}$$

holds with a suitable $const \geq 0$.

First an explicit estimate of the sequence $\max\left(x_n, J^E(x_n)\right)$ is given.

Lemma 3.4.3 *For $n \geq 0$, set*

$$a_n(h, \alpha) := -\frac{3}{4}\alpha \cdot \frac{(1 + h\alpha)^{n+1}}{2 + (1 + h\alpha)^n},$$

then we have that $x_n \in (\omega_{\varphi,-}, a_n)$ and $J^E(x_n) \in (\omega_{\Phi,-}, a_n)$.

Proof. It is easily checked that, due to assumptions (3.23),

$$\max\left(\omega_{\varphi,-}, \omega_{\Phi,-}\right) < a_n$$

for $n \geq 0$, so the intervals in the lemma are non-degenerate. We proceed by induction.

$a_0 = -\frac{\alpha}{4}(1 + h\alpha) > x_0 \equiv J^E(x_0) \equiv -\frac{\alpha}{3}$ is equivalent to $h\alpha < \frac{1}{3}$, being true by assumptions (3.23) on h_0 and α_0.

So suppose that the statement is true for some $n \geq 0$. Since $\mathcal{N}_\varphi^E(x) < (1 + h\alpha)x + \frac{6}{5}h\, x^2$ is implied by $|x| \leq \varepsilon_0 < \frac{1}{5K}$, and $\mathcal{N}_\varphi^E$ is monotone increasing, we get that

$$x_{n+1} = \mathcal{N}_\varphi^E(x_n) < \mathcal{N}_\varphi^E(a_n) < (1 + h\alpha)a_n + \frac{6}{5} h\, a_n^2,$$

thus it is enough to prove that the right-hand side above is smaller than a_{n+1}. But

$$a_{n+1} - \left((1 + h\alpha)a_n + \frac{6}{5} h\, a_n^2\right) =$$

$$-\frac{3h\alpha^2(1 + h\alpha)^{2+2n}\left(-2 + (1 + h\alpha)^n(-1 + 9h\alpha)\right)}{40\left(2 + (1 + h\alpha)^n\right)^2\left(2 + (1 + h\alpha)^{n+1}\right)} > 0$$

is equivalent to $-2 + (1 + h\alpha)^n(-1 + 9h\alpha) < 0$, which is implied by $h\alpha < \frac{1}{9}$.

Of course, the above inequalities remain true, if $\mathcal{N}_\varphi$ is replaced by $\mathcal{N}_\Phi$, also noticing that, by construction, $J^E(x_{n+1}) = \mathcal{N}_\Phi^E(J^E(x_n))$, so the induction is complete. ■

Remark 3.4.2.1 The induction would fail, if, in estimate $\mathcal{N}_\varphi^E(x) < (1 + h\alpha)x + \frac{6}{5}h\, x^2$, the constant $\frac{6}{5}$ was replaced by, say, $\frac{7}{5}$. (The explanation resides in the particular choice of the constant $\frac{3}{4}$ in the definition of a_n, since $\frac{3}{4}\cdot\frac{6}{5} < 1 < \frac{3}{4}\cdot\frac{7}{5}$.)

Remark 3.4.2.2 The upper estimate a_n in our first main lemma has been found by computer experiments with *Mathematica* based on the parametrized model function in [29].

In order to prove the boundedness of (3.25), the sum $\sum_{i=0}^{n}$ will be split into two. An appropriate index to split at is $\lceil \frac{const}{h\alpha}\rceil$, as established by the following lemma.

Lemma 3.4.4 *Suppose that* $n > \lceil \frac{6}{h\alpha}\rceil$. *Then*

$$\max\left(x_n, J^E(x_n)\right) < -\frac{2}{3}\alpha,$$

hence

$$1 + h\alpha + \frac{7}{4}h\max\left(x_n, J^E(x_n)\right) < 1 - \frac{h\alpha}{6}$$

holds for $n > \lceil \frac{6}{h\alpha}\rceil$.

Proof. By Lemma 3.4.3 it is sufficient to show that $n > \lceil \frac{6}{h\alpha}\rceil$ implies $a_n < -\frac{2}{3}\alpha$. This latter inequality is equivalent to $(1+h\alpha)^n(1+9h\alpha) > 16$. But if $n > \lceil \frac{6}{h\alpha}\rceil$, then

$$(1+h\alpha)^n > (1+h\alpha)^{\lceil \frac{6}{h\alpha}\rceil} = \left(1+\frac{1}{\frac{1}{h\alpha}}\right)^{(1+\frac{1}{h\alpha})\cdot\frac{h\alpha}{1+h\alpha}\cdot\lceil\frac{6}{h\alpha}\rceil}.$$

However, it is known that $\left(1+\frac{1}{A}\right)^{A+1} > e$, if $A \geq 1$, and it is easy to see that $\frac{B}{1+B}\cdot\lceil\frac{6}{B}\rceil > 3$, if $0 < B < 1$. Since $e^3 > 16$, the proof is complete. ■

Now we can turn to (3.25). Fix $h \in (0, h_0]$, $\alpha \in (0, \alpha_0]$ and $n \in \mathbb{N}^+$. (If $n \leq \lceil \frac{6}{h\alpha}\rceil$, then the sums $\sum_{i=\lceil \frac{6}{h\alpha}\rceil+1}^{n}$ below are, of course, not present, making the proof even simpler.) Since now $\omega_{\varphi,-} < x_i < 0$, by Lemma 3.3.1 $|x_i| \leq \frac{3}{2}\alpha$, and by monotonicity $\max\left(x_j, J^E(x_j)\right) \leq x_0 \equiv J^E(x_0) \equiv -\frac{\alpha}{3}$, further, by using Lemma 3.4.4, assumption $h\alpha < 1$ from (3.23) and inequality $(1+\frac{1}{A})^A \leq e$ (if $A \geq 1$), we get that

$$h\sum_{i=0}^{n}|x_i|^3\prod_{j=i}^{n-1}\left(1+h\alpha+\frac{7}{4}h\max\left(x_j, J^E(x_j)\right)\right) \leq$$

$$\frac{27h\alpha^3}{8}\sum_{i=0}^{\lceil\frac{6}{h\alpha}\rceil}\prod_{j=1}^{\lceil\frac{6}{h\alpha}\rceil-1}\left(1+h\alpha-\frac{7}{4}\cdot\frac{h\alpha}{3}\right) + \frac{27h\alpha^3}{8}\sum_{i=\lceil\frac{6}{h\alpha}\rceil+1}^{n}\prod_{j=i}^{n-1}\left(1-\frac{h\alpha}{6}\right) \leq$$

$$\frac{27h\alpha^3}{8}\left(1+\frac{5}{12}h\alpha\right)^{\frac{6}{h\alpha}}\left(\left\lceil\frac{6}{h\alpha}\right\rceil+1\right) + \frac{27h\alpha^3}{8}\sum_{i=\lceil\frac{6}{h\alpha}\rceil+1}^{n}\left(1-\frac{h\alpha}{6}\right)^{n-i} \leq$$

$$\frac{27h\alpha^3}{8}\left(1+\frac{5}{12}h\alpha\right)^{\frac{12}{5h\alpha}\cdot\frac{5h\alpha}{12}\cdot\frac{6}{h\alpha}}\left(\frac{6+2h\alpha}{h\alpha}\right) + \frac{27h\alpha^3}{8}\sum_{i=0}^{\infty}\left(1-\frac{h\alpha}{6}\right)^{i} \leq$$

$$\frac{27h\alpha^3}{8}\cdot e^{\frac{30}{12}}\cdot\frac{8}{h\alpha} + \frac{27h\alpha^3}{8}\cdot\frac{6}{h\alpha} \leq 350\,\alpha^2.$$

Therefore, $\sup_{[x_{n+1},x_n]}|\,id - J^E| \leq 350c\cdot h^p\alpha^2$ for any $h \in (0, h_0]$, $\alpha \in (0, \alpha_0]$ and $n \geq 1$, further, as we have seen, $\sup_{[x_1,x_0]}|\,id - J^E| \leq \frac{c}{27}h^{p+1}\alpha^3$, which yield the following lemma.

Lemma 3.4.5 *Under assumption (3.23)*

$$\sup_{(\omega_{\varphi,\_},x_0]} |\,id - J^E| \le 350c \cdot h^p\alpha^2.$$

Now the closeness estimate is proved in the interval $(y_0, \omega_{\varphi,0})$. Recall that $y_0 = x_0 = J^E(x_0) \equiv -\frac{\alpha}{3}$ and $\omega_{\varphi,0} = \omega_{\Phi,0} \equiv 0$.

Suppose that $n \ge 1$. (The case $n = 0$ will be examined later.) Then

$$\sup_{[y_n,y_{n+1}]} |\,id - J^E| = \sup_{[y_n,y_{n+1}]} \left| \left(\mathcal{N}_\Phi^E\right)^{[-1]} \circ \mathcal{N}_\Phi^E - \left(\mathcal{N}_\Phi^E\right)^{[-1]} \circ J^E \circ \mathcal{N}_\varphi^E \right| \le$$

$$\sup_{x\in[y_n,y_{n+1}]} \left[\left(\sup_{[\{\mathcal{N}_\Phi^E(x), J^E\circ\mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]} \right)' \right) \left(\left|\mathcal{N}_\Phi^E - \mathcal{N}_\varphi^E\right|(x) + \left|\mathcal{N}_\varphi^E - J^E \circ \mathcal{N}_\varphi^E\right|(x) \right) \right]$$

$$\le \left[\sup_{x\in[y_n,y_{n+1}]} \sup_{[\{\mathcal{N}_\Phi^E(x), J^E\circ\mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]} \right)' \right] \left[c \cdot h^{p+1}|y_n|^3 + \sup_{[y_{n-1},y_n]} |\,id - J^E| \right],$$

provided that $\sup_{[\{\mathcal{N}_\Phi^E(x), J^E\circ\mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]} \right)'$ is nonnegative.

Lemma 3.4.6 *Suppose that $n \ge 1$, then under assumption (3.23) we have that*

$$\sup_{x\in[y_n,y_{n+1}]} \sup_{[\{\mathcal{N}_\Phi^E(x), J^E\circ\mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]} \right)' \le 1 - \frac{h\alpha}{8}.$$

Proof.

$$\sup_{x\in[y_n,y_{n+1}]} \sup_{[\{\mathcal{N}_\Phi^E(x), J^E\circ\mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]} \right)' = \sup_{x\in[y_n,y_{n+1}]} \sup_{[\{\mathcal{N}_\Phi^E(x), J^E\circ\mathcal{N}_\varphi^E(x)\}]} \frac{1}{(\mathcal{N}_\Phi^E)' \circ (\mathcal{N}_\Phi^E)^{[-1]}}$$

$$= \sup_{x\in[y_n,y_{n+1}]} \sup_{[\{x, (\mathcal{N}_\Phi^E)^{[-1]}\circ J^E\circ\mathcal{N}_\varphi^E(x)\}]} \frac{1}{(\mathcal{N}_\Phi^E)'} = \dots$$

But, by definition, $(\mathcal{N}_\Phi^E)^{[-1]} \circ J^E \circ \mathcal{N}_\varphi^E(x) = J^E(x)$, if $x \in [y_n, y_{n+1}]$, and $[\{x, J^E(x)\}] = [\min(x, J^E(x)), \max(x, J^E(x))]$, further, by the monotonicity of id and J^E we obtain that

$$\dots = \sup_{[\min(y_n, J^E(y_n)), \max(y_{n+1}, J^E(y_{n+1}))]} \frac{1}{(\mathcal{N}_\Phi^E)'} \le \dots$$

By construction, however, $[\min(y_n, J^E(y_n)), \max(y_{n+1}, J^E(y_{n+1}))] \subset (y_0, 0) = (-\frac{\alpha}{3}, 0)$ and $(\mathcal{N}_\Phi^E)'$ is nonnegative here by assumption (3.23), justifying the computations just above the lemma. We now continue the proof of the lemma.

$$\dots \le \sup_{(-\frac{\alpha}{3},0)} \frac{1}{(\mathcal{N}_\Phi^E)'} \le \dots$$

It is easy to see that assumption (3.23) together with $x < 0$ imply that $(\mathcal{N}_\Phi^E)'(x) \ge 1 + h\alpha + \frac{9}{4}hx \ge 0$. So

$$\dots \le \sup_{x\in(-\frac{\alpha}{3},0)} \frac{1}{1 + h\alpha + \frac{9}{4}hx} \le \frac{1}{1 + h\alpha + \frac{9}{4}h\left(-\frac{\alpha}{3}\right)} = \frac{1}{1 + \frac{1}{4}h\alpha} \le 1 - \frac{h\alpha}{8},$$

since $\frac{1}{1+A} \leq 1 - \frac{A}{2}$, if $A \in [0,1]$. ∎

We have thus proved (also using $|y_n| \leq \frac{\alpha}{3}$) that for $n \geq 1$

$$\sup_{[y_n,y_{n+1}]} |\, id - J^E| \leq \left(1 - \frac{h\alpha}{8}\right) \left[\frac{c}{27} \cdot h^{p+1}\alpha^3 + \sup_{[y_{n-1},y_n]} |\, id - J^E|\right] \tag{3.26}$$

For $n = 0$, similarly as before, we get that

$$\sup_{[y_0,y_1]} |\, id - J^E| \leq \left[\sup_{x\in[y_0,y_1]} \sup_{[\{\mathcal{N}_\Phi^E(x), J^E\circ\mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]}\right)'\right] \left[c \cdot h^{p+1}|y_0|^3 + \sup_{[y_{-1},y_0]} |\, id - J^E|\right].$$

But $[y_{-1}, y_0] \equiv [x_1, x_0]$, so the second factor $[\ldots]$ is bounded by $2 \cdot \frac{c}{27}h^{p+1}\alpha^3$. As for the first factor $[\ldots]$, we notice that $y_0 < (\mathcal{N}_\Phi^E)^{[-1]}(y_0)$ (since this is equivalent to $x_1 < x_0$), which implies that

$$\sup_{x\in[y_0,y_1]} \sup_{[\{\mathcal{N}_\Phi^E(x), J^E\circ\mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]}\right)' = \sup_{x\in[y_0,y_1]} \sup_{[\{x, (\mathcal{N}_\Phi^E)^{[-1]}\circ J^E\circ\mathcal{N}_\varphi^E(x)\}]} \frac{1}{(\mathcal{N}_\Phi^E)'} =$$

$$\sup_{[y_0,y_1]\cup[y_0,(\mathcal{N}_\Phi^E)^{[-1]}(y_0)]} \frac{1}{(\mathcal{N}_\Phi^E)'} \leq \sup_{[y_0,0)} \frac{1}{(\mathcal{N}_\Phi^E)'} \leq 1,$$

therefore

$$\sup_{[y_0,y_1]} |\, id - J^E| \leq 2 \cdot \frac{c}{27}h^{p+1}\alpha^3. \tag{3.27}$$

Repeated application of (3.26), further (3.27) yield for $n \geq 1$ that

$$\sup_{[y_n,y_{n+1}]} |\, id - J^E| \leq \left(1 - \frac{h\alpha}{8}\right)^n \sup_{[y_0,y_1]} |\, id - J^E| + \frac{c}{27}h^{p+1}\alpha^3 \sum_{i=1}^{n} \left(1 - \frac{h\alpha}{8}\right)^i \leq$$

$$1 \cdot 2 \cdot \frac{c}{27}h^{p+1}\alpha^3 + \frac{c}{27}h^{p+1}\alpha^3 \cdot \frac{8}{h\alpha} \leq \frac{c}{3}h^p\alpha^2,$$

due to $h\alpha \leq \frac{1}{2}$ by (3.23). The same upper estimate is valid for $n = 0$, so we have proved the following result.

Lemma 3.4.7 *Under assumption (3.23)*

$$\sup_{(x_0,0)} |\, id - J^E| \leq \frac{c}{3}h^p\alpha^2.$$

3.4.3 The outer region

In this section, we first prove an $\mathcal{O}(h^p)$ closeness-estimate in the interval $[z_0, \omega_{\varphi,-})$ for $\alpha > 0$. Then, in the second part, the closeness is proved on $[z_0, \omega_{\Phi,0}) \equiv [z_0, 0)$ for $\alpha \leq 0$.

The derivation of the following formulae is similar to their counterparts in the inner region, with the difference that—since this time the sequence z_n is increasing—an extra term and an index-shift occur.

For $n \geq 1$ (also using (3.23)) we have that

$$\sup_{[z_n,z_{n+1}]} |\, id - J^E| \leq c \cdot h^{p+1}|z_0|^3 \prod_{j=1}^{n} \left(1 + h\alpha + \frac{7}{4}h \max\left(z_j, J^E(z_j)\right)\right) +$$

$$c \cdot h^{p+1} \sum_{i=0}^{n-1} |z_i|^3 \prod_{j=i+2}^{n} \left(1 + h\alpha + \frac{7}{4} h \max\left(z_j, J^E(z_j)\right)\right), \tag{3.28}$$

where, again $\prod_{j=n+1}^{n}$ above is 1, and

$$\sup_{[z_0,z_1]} | id - J^E| \leq c \cdot h^{p+1} |z_0|^3.$$

The following main lemma, as a counterpart of Lemma 3.4.3, gives a lower estimate of the sequence z_n, if $\alpha > 0$.

Lemma 3.4.8 *For $n \geq 0$, set*

$$b_n(h, \alpha) := -2\alpha \cdot \frac{(1 + h\alpha)^{n+1}}{-1 + \alpha + (1 + h\alpha)^n},$$

then $b_n \leq \min\left(z_n, J^E(z_n)\right)$.

Proof. $b_0 = -2 - 2h\alpha < -2 \leq -1 \leq -\varepsilon_0 \leq z_0 = J^E(z_0)$ holds due to assumption (3.23). Suppose that the statement is true for some $n \geq 0$. Since $\mathcal{N}_\varphi^E(x) \geq (1 + h\alpha)x + \frac{3}{5} h\, x^2$ follows from $|x| \leq \varepsilon_0 < \frac{2}{5K}$, further $(1 + h\alpha)id + \frac{3}{5} h\, id^2$ is monotone increasing (which is implied by, *e.g.*, $|x| \leq \frac{5}{6h}$, but it is easy to see that $h \leq \frac{5}{18}$ and $-3 < b_n < 0$ follows from (3.23), hence $|b_n| \leq \frac{5}{6h}$), so we obtain that

$$z_{n+1} = \mathcal{N}_\varphi^E(z_n) \geq (1 + h\alpha) z_n + \frac{3}{5} h\, z_n^2 \geq (1 + h\alpha) b_n + \frac{3}{5} h\, b_n^2,$$

thus it is sufficient to show that

$$(1 + h\alpha) b_n + \frac{3}{5} h\, b_n^2 \geq b_{n+1}.$$

However, this is equivalent to

$$0 \leq \frac{2h\alpha^2 (1 + h\alpha)^{2+2n}}{5\left(-1 + \alpha + (1 + h\alpha)^n\right)^2} \cdot \frac{-1 + \alpha + (1 + h\alpha)^n (1 + 6h\alpha)}{-1 + \alpha + (1 + h\alpha)^{n+1}},$$

which is true since $\alpha > 0$ and $h > 0$.

The proof remains valid if $\mathcal{N}_\varphi$ is replaced by $\mathcal{N}_\Phi$ (and $J^E(z_n)$ is written instead of z_n), hence $b_n \leq J^E(z_n)$ also holds. ■

Now, since $z_j < \omega_{\varphi,-}$ and $J^E(z_j) < \omega_{\Phi,-}$, by Lemma 3.3.1 we get that the right-hand side of (3.28) is at most

$$c \cdot h^{p+1} |z_0|^3 \prod_{j=1}^{n} \left(1 - \frac{h\alpha}{2}\right) + c \cdot h^{p+1} \sum_{i=0}^{n-1} |z_i|^3 \prod_{j=i+2}^{n} \left(1 - \frac{h\alpha}{2}\right) \leq$$

$$c \cdot h^{p+1} |z_0|^3 + c \cdot h^{p+1} \sum_{i=0}^{n-1} |z_i|^3 \left(1 - \frac{h\alpha}{2}\right)^{n-1-i}.$$

We will verify that $h \sum_{i=0}^{n} |z_i|^3 \left(1 - \frac{h\alpha}{2}\right)^{n-i}$ is uniformly bounded for any $n \geq 0$, $0 < h \leq h_0$ and $0 < \alpha \leq \alpha_0$.

If $n \geq \lceil \frac{1}{h\alpha} \rceil$, then by Lemma 3.4.8 (also using that $h\alpha \leq \frac{1}{9}$ and $z_j < 0$)

$$h \sum_{i=\lceil \frac{1}{h\alpha} \rceil}^{n} |z_i|^3 \left(1 - \frac{h\alpha}{2}\right)^{n-i} \leq h \sum_{i=\lceil \frac{1}{h\alpha} \rceil}^{n} |b_i|^3 \left(1 - \frac{h\alpha}{2}\right)^{n-i} \leq$$

$$11h\alpha^3 \sum_{i=\lceil \frac{1}{h\alpha} \rceil}^{n} \left(\frac{(1+h\alpha)^i}{-1+\alpha+(1+h\alpha)^i}\right)^3 \left(1 - \frac{h\alpha}{2}\right)^{n-i} \leq \ldots$$

for these i indices however $\frac{(1+h\alpha)^i}{-1+\alpha+(1+h\alpha)^i} \leq 3$ holds (since this is implied by $\frac{3}{2} \leq (1+h\alpha)^i$, being true by $(1+h\alpha)^i \geq (1+h\alpha)^{\frac{1}{h\alpha}} \geq 1 + \frac{1}{h\alpha} \cdot h\alpha > \frac{3}{2}$), thus

$$\ldots \leq 27 \cdot 11\alpha^2 h\alpha \sum_{i=0}^{\infty} \left(1 - \frac{h\alpha}{2}\right)^i = 594\alpha^2.$$

On the other hand, if $n < \lceil \frac{1}{h\alpha} \rceil$, then (using that $|z_i| \leq 1$ and $h\alpha \leq \frac{1}{9}$ again)

$$h \sum_{i=0}^{n} |z_i|^3 \left(1 - \frac{h\alpha}{2}\right)^{n-i} \leq h \sum_{i=0}^{n} |z_i|^2 \left(1 - \frac{h\alpha}{2}\right)^{n-i} \leq \tag{3.29}$$

$$5h \sum_{i=0}^{n} \left(\frac{\alpha(1+h\alpha)^i}{-1+\alpha+(1+h\alpha)^i}\right)^2 \left(1 - \frac{h\alpha}{2}\right)^{n-i} \leq \ldots$$

now using inequalities $e^{\frac{x}{2}} \leq 1 + x$ ($x \in [0,1]$) and $1 + x \leq e^x$ ($x \in \mathbb{R}$) we get that $(1+h\alpha)^{2i} \leq e^{h\alpha 2i} \leq e^{h\alpha 2n} \leq e^2 < 8$, further, $\left(1 - \frac{h\alpha}{2}\right)^{n-i} \leq e^{-\frac{h\alpha}{2}(n-i)}$ and $e^{\frac{h\alpha}{2}i} \leq (1+h\alpha)^i$, therefore

$$\ldots \leq 40h \sum_{i=0}^{n} \left(\frac{\alpha e^{-\frac{h\alpha}{4}(n-i)}}{-1+\alpha+e^{\frac{h\alpha}{2}i}}\right)^2. \tag{3.30}$$

Set $g_{h,\alpha,n}(x) \equiv g(x) := \left(\frac{\alpha \exp\left(-\frac{1}{4}h\alpha(n-x)\right)}{-1+\alpha+\exp\left(\frac{1}{2}h\alpha x\right)}\right)^2$, if $x \in [0, \infty)$. Notice that g is bounded at $x = 0$. For this function we have that

$$g'(x) = -\frac{1}{2}h\alpha^3\, e^{-\frac{1}{2}h\alpha(n-x)} \cdot \frac{1 - \alpha + e^{\frac{1}{2}hx\alpha}}{\left(-1+\alpha+e^{\frac{1}{2}hx\alpha}\right)^3},$$

meaning that g is strictly monotone decreasing, if $\alpha < 1$. Hence

$$40h \sum_{i=0}^{n} \left(\frac{\alpha e^{-\frac{h\alpha}{4}(n-i)}}{-1+\alpha+e^{\frac{h\alpha}{2}i}}\right)^2 = 40h + 40h \sum_{i=1}^{n} g_{h,\alpha,n}(i) \leq$$

$$40h + 40h \int_0^n g_{h,\alpha,n}(x)\mathrm{d}x = 40h + 40h \left[-2\alpha \frac{\exp\left(-\frac{1}{2}h\alpha n\right)}{h\left(-1+\alpha+\exp\left(\frac{1}{2}h\alpha x\right)\right)}\right]_{x=0}^{n} =$$

$$40h + 40h \left(\frac{2\left(1 - \exp\left(-\frac{1}{2}h\alpha n\right)\right)}{h\left(\exp\left(\frac{1}{2}h\alpha n\right) - 1 + \alpha\right)}\right) \leq 40h + 80 \left(\frac{1 - \exp\left(-\frac{1}{2}h\alpha n\right)}{\exp\left(\frac{1}{2}h\alpha n\right) - 1}\right) =$$

$$40h + 80e^{-\frac{1}{2}h\alpha n} \leq 120,$$

since $h \le 1$.

Now combining all the estimates so far in the section, under assumption (3.23) we get that if $\alpha > 0$, then

$$\sup_{[z_0,\omega_{\varphi,-})} |\, id - J^E| = \sup_{n\in\mathbb{N}} \sup_{[z_n,z_{n+1}]} |\, id - J^E| \le$$

$$\sup_{n\in\mathbb{N}} \max\left(c\cdot h^{p+1}|z_0|^3,\, c\cdot h^{p+1}|z_0|^3 + c\cdot h^{p+1}\sum_{i=0}^{n}|z_i|^3\left(1-\frac{h\alpha}{2}\right)^{n-i}\right) \le$$

$$c\cdot h^{p+1}|z_0|^3 + c\cdot h^p\cdot(120+594\alpha^2) \le 130c\cdot h^p.$$

Remark 3.4.3.1 If, in (3.29), the exponent of $|z_i|$ had not been changed to 2, then the integral of g would have been significantly more complicated. (Interestingly, similar complication occurs, if one considers simply $|z_i|$ instead of $|z_i|^2$.) The rational pair $\frac{1}{4}$ and $\frac{1}{2}$ in the definition of g has also been a fortunate choice: when working with the numbers $\frac{1}{5}$ and $\frac{1}{2}$ instead, for example, *Mathematica* produced so complicated integrals that were practically useless from the viewpoint of further analysis.

Remark 3.4.3.2 An alternative approach to analyze sum (3.30) is to estimate $e^{-\frac{h\alpha}{4}(n-i)}$ above by 1. However, the resulting integral would not be much simpler in that case either. (Then we would use the boundedness of $\alpha\ln\alpha$ for $\alpha\in(0,\alpha_0]$.) Compare the above calculations with their counterparts in the pitchfork case.

Finally, we prove a closeness estimate on $[z_0, 0)$ for $\alpha \le 0$. We begin with a simple observation on monotonicity of the sequence $z_n \equiv z_n(\alpha)$. (As before, for brevity, the dependence on h is still suppressed.)

Lemma 3.4.9 *Suppose that $\alpha \le 0$ and assumption (3.23) hold. Then for any $0 < h \le h_0$, $-\alpha_0 \le \alpha \le \beta \le 0$ and $n \in \mathbb{N}$ we have that*

$$0 > z_n(\alpha) \ge z_n(\beta).$$

Proof. By definition, we have that $z_0(\alpha) = z_0(\beta) = z_0$, so suppose that for some n we already know that $z_n(\alpha) \ge z_n(\beta)$. Then, by the definition of the sequence z_n, further by the facts that the function $z \mapsto \mathcal{N}_\varphi(h,z,\alpha)$ is monotone *increasing* and the function $\alpha \mapsto \mathcal{N}_\varphi(h,z,\alpha)$ is monotone *decreasing*, we get that

$$z_{n+1}(\alpha) = \mathcal{N}_\varphi(h,z_n(\alpha),\alpha) \ge \mathcal{N}_\varphi(h,z_n(\beta),\alpha) \ge \mathcal{N}_\varphi(h,z_n(\beta),\beta) = z_{n+1}(\beta),$$

which completes the induction. ■

This means that $0 > z_n(\alpha) \ge z_n(0)$ holds for $\alpha \le 0$, hence it is enough to give a lower estimate for $z_n(0)$. But such an estimate has been constructed in Lemma 2.4.3, namely we recall the following.

Lemma 3.4.10 *Under assumption (3.23), we have for $n \in \mathbb{N}$ that*

$$z_n(0) \ge z_0$$

and for $n \ge \lfloor\frac{1}{h}\rfloor + 1$

$$z_n(0) \ge -\frac{2}{nh}.$$

Then we can simply estimate (3.28) for $\alpha \leq 0$ as follows. Supposing that $n \geq 1$ we get that

$$\sup_{[z_n, z_{n+1}]} |\, id - J^E| \leq c \cdot h^{p+1} |z_0|^3 \prod_{j=1}^{n} \left(1 + h\alpha + \frac{7}{4} h \max\left(z_j, J^E(z_j)\right)\right) +$$

$$c \cdot h^{p+1} \sum_{i=0}^{n-1} |z_i|^3 \prod_{j=i+2}^{n} \left(1 + h\alpha + \frac{7}{4} h \max\left(z_j, J^E(z_j)\right)\right) \leq$$

$$c \cdot h^{p+1} |z_0|^3 \cdot 1 + c \cdot h^p \cdot h \sum_{i=0}^{n} |z_i(0)|^3 \cdot 1 \leq$$

$$c \cdot h^p \left(h|z_0|^3 + h \sum_{i=0}^{\lfloor \frac{1}{h} \rfloor} |z_i(0)|^2 + h \sum_{i=\lfloor \frac{1}{h} \rfloor + 1}^{n} |z_i(0)|^2 \right),$$

where, of course, for $n \leq \lfloor \frac{1}{h} \rfloor$, the sum above $\sum_{i=\lfloor \frac{1}{h} \rfloor + 1}^{n}$ should be omitted. But

$$h \sum_{i=0}^{\lfloor \frac{1}{h} \rfloor} |z_i(0)|^2 \leq (h+1) \cdot z_0^2 = 2z_0^2,$$

and

$$h \sum_{i=\lfloor \frac{1}{h} \rfloor + 1}^{n} |z_i(0)|^2 \leq h \sum_{i=\lfloor \frac{1}{h} \rfloor + 1}^{n} \frac{4}{i^2 h^2} \leq \frac{4}{h} \int_{\frac{1}{h}-1}^{\infty} \frac{1}{i^2} = \frac{4}{1-h} \leq 8.$$

We have thus proved that

$$\sup_{[z_0, 0)} |\, id - J^E| \leq 12c \cdot h^p.$$

Chapter 4

Conjugacy in the discretized pitchfork bifurcation

Summary. In Section 4.1, some conditions on the right hand side of the differential equation to define a pitchfork bifurcation for the induced map $x \mapsto \Phi(h, x, \alpha)$ are examined and one set of conditions is chosen for further study. In Section 4.2, the necessary and sufficient conditions on the discretization map $x \mapsto \varphi(h, x, \alpha)$ to undergo a pitchfork bifurcation are identified, then normal forms of $x \mapsto \Phi(h, x, \alpha)$ and $x \mapsto \varphi(h, x, \alpha)$ together with appropriate closeness estimates are derived. In Section 4.3, a conjugacy between the exact $\Phi(h, \cdot, \alpha)$ and discretized $\varphi(h, \cdot, \widetilde{\alpha})$ families is defined. Notice that a parameter shift is needed, but $|\alpha - \widetilde{\alpha}| = \mathcal{O}(h^p)$. Finally, in Section 4.4 we show that the constructed conjugacy is $\mathcal{O}(h^p)$-close to the identity and this estimate is optimal.

4.1 Introduction

Suppose we have a one-dimensional ordinary differential equation depending on a scalar bifurcation parameter $\alpha \in \mathbb{R}$

$$\dot{x} = f(x, \alpha) \tag{4.1}$$

and its one-step discretization

$$X_{n+1} := \varphi(h, X_n, \alpha), \qquad n = 0, 1, 2, \ldots, \tag{4.2}$$

where $h > 0$ is the step-size of the sufficiently smooth one-step method $\varphi : \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of order $p \geq 1$, and the function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is of class C^{p+k+1} with $k \geq 6$ and the last derivatives uniformly bounded.

The order of the numerical method means that

$$|\Phi(h, x, \alpha) - \varphi(h, x, \alpha)| \leq \mathit{const} \cdot h^{p+1}, \quad \forall\, h \in [0, h_0], \forall\, |x| \leq \varepsilon_0, \forall\, |\alpha| \leq \alpha_0. \tag{4.3}$$

Here $\Phi(h, \cdot, \alpha) : \mathbb{R} \to \mathbb{R}$ is, as always, the time-h-map of the solution flow induced by (4.1) at parameter value α, further h_0, ε_0 and α_0 are some small positive constants.

Suppose that the origin $x = 0$, $\alpha = 0$ is an equilibrium as well as a *pitchfork bifurcation point* for (4.1), that is the following conditions hold

$$f(0, \alpha) = 0, \quad \forall\, |\alpha| \leq \alpha_0,$$

$$f_x^B = 0, \quad f_{xx}^B = 0, \quad f_{xxx}^B \neq 0, \quad f_{x\alpha}^B \neq 0. \tag{4.4}$$

We apply the same notation as in the transcritical case, including, for example, evaluation ^{B} at the bifurcation point $x = 0$ and $\alpha = 0$, and later, evaluation ^{E} at general parameter values h and α.

Remark 4.1.1 In this context f is usually assumed to be odd, that is $f(x, \alpha) = -f(-x, \alpha)$ for $|x| \leq \varepsilon_0$ and $|\alpha| \leq \alpha_0$. The above (4.4) *asymmetric pitchfork* conditions are thus weaker than the usual ones. If symmetry were assumed, the normal form transformations in the next section would be much easier. Symmetry, however, is not essential here.

We remark that these pitchfork conditions can even be slightly weakened (*cf.* the corresponding remark in the transcritical case). For *maps* of form $x \mapsto g(x, \alpha)$, [46], for example, uses again point conditions at the bifurcation point $(0, 0)$, formulated in Theorem 4.1.1 below. However, since the corresponding point conditions in [46] proved to be insufficient to define a transcritical bifurcation, we investigated with care whether these point conditions do really define a pitchfork now. Interestingly, this time the answer is affirmative, as shown by Theorem 4.1.1.

Summarizing, as in the transcritical case, we have not adopted the most general conditions for a pitchfork bifurcation in (4.4), nevertheless symmetry is not assumed.

Definition A smooth map $x \mapsto g(x, \alpha)$ defined near the origin $(0, 0)$ and depending smoothly on a parameter α is said to have a *pitchfork bifurcation* at the origin, if there exists a neighbourhood of $(0, 0)$ in which there are exactly three distinct branches of fixed points $\rho_0(\alpha)$ and $\rho_-(\alpha) < 0 < \rho_+(\alpha)$ of the map g for $\alpha < 0$, while there is a unique branch of fixed points $\rho_0(\alpha)$ of g for $\alpha \geq 0$. (Of course, the converse situation—with inequalities $\alpha < 0$ and $\alpha \geq 0$ exchanged—is also called a pitchfork bifurcation.)

Theorem 4.1.1 (Asymmetric pitchfork bifurcation) *Assume we have a smooth map $x \mapsto g(x, \alpha)$ defined near the origin $(0, 0)$ and depending smoothly on a parameter α such that*

$$g(0,0) = 0,\ g_x(0,0) = 1,\ g_{xx}(0,0) = 0,\ g_{xxx}(0,0) \neq 0,\ g_\alpha(0,0) = 0 \ \textit{ and } \ g_{x\alpha}(0,0) \neq 0.$$

Suppose that $g_{xxx}(0,0) \cdot g_{x\alpha}(0,0) > 0$. Then the map g undergoes a pitchfork bifurcation locally at the origin, moreover, there exist positive constants $c_0 > 0, c_2 > c_1 > 0$ such that $|\rho_0(\alpha)| \leq c_0|\alpha|$ (for $|\alpha| \leq \alpha_0$) and $c_1|\alpha|^{1/2} \leq |\rho_\pm(\alpha)| \leq c_2|\alpha|^{1/2}$ (for $-\alpha_0 \leq \alpha < 0$) hold with some $\alpha_0 > 0$ sufficiently small.

The case $g_{xxx}(0,0) \cdot g_{x\alpha}(0,0) < 0$ yields the "mirror-symmetrical" counterpart: there are three branches of fixed points for $\alpha > 0$ and a unique branch for $\alpha \leq 0$ with similar estimates.

Proof. Notation introduced here is understood to be "local", not to interfere with any notation outside this proof. An appropriate degree of smoothness of g is assumed in order for all derivatives in the proof to exist. Let us define $G(x, \alpha) := g(x, \alpha) - x$. We are then interested in the roots of G. By multiplying G with a suitable constant, we can assume that $G_{x\alpha}(0, 0) = 1$. Then conditions on g imply that G has a Taylor expansion

$$G(x, \alpha) = \alpha^2 s_0(\alpha) + (\alpha + \alpha^2 s_1(\alpha))x + \alpha s_2(\alpha)x^2 + (r + \alpha s_3(\alpha))x^3 + s_4(x, \alpha)x^4 \qquad (4.5)$$

valid locally near the origin with an $r \neq 0$ constant and s_i being smooth functions. Assume, say, that $r > 0$ (corresponding to $g_{xxx}(0,0) \cdot g_{x\alpha}(0,0) > 0$), the other case is symmetrical.

Let $K > 0$ be such that $|s_0| \leq K$ and set $c_0 := K + 1$. Then, if $\alpha \neq 0$, it is elementary to see that $G(c_0|\alpha|, \alpha)/\alpha^2$ and $G(-c_0|\alpha|, \alpha)/\alpha^2$ have opposite signs for every α, $0 \neq |\alpha| \leq \alpha_0$ with a suitably small $\alpha_0 > 0$, since the functions s_i are bounded. So the intermediate value theorem applies and we get a branch of zeros $|\rho_0(\alpha)| \leq c_0|\alpha|$ of $G(\cdot, \alpha)$ for $0 \neq |\alpha| \leq \alpha_0$. But at $\alpha = 0$

it is seen that for $x = 0$, $G(x, 0) = 0$ and—due to the boundedness of s_4—this zero is unique, provided that we are focusing on a small enough neighbourhood of the origin.

Now consider the $-\alpha_0 \leq \alpha < 0$ case and set $c_1 := \frac{1}{2\sqrt{r}}$ and $c_2 := 4c_1$. It is again elementary to see that $G(c_1|\alpha|^{1/2}, \alpha)/|\alpha|^{3/2} < 0$, while $G(c_2|\alpha|^{1/2}, \alpha)/|\alpha|^{3/2} > 0$, if α_0 is small enough. Hence $G(\cdot, \alpha)$ has a branch $\rho_+(\alpha)$ of zeros in the interval $[c_1|\alpha|^{1/2}, c_2|\alpha|^{1/2}]$ for every $-\alpha_0 \leq \alpha < 0$. Similarly, $G(-c_1|\alpha|^{1/2}, \alpha)/|\alpha|^{3/2} > 0$ and $G(-c_2|\alpha|^{1/2}, \alpha)/|\alpha|^{3/2} < 0$ for every $-\alpha_0 \leq \alpha < 0$ with α_0 being small enough, yielding another branch $\rho_-(\alpha)$ of zeros in $[-c_2|\alpha|^{1/2}, -c_1|\alpha|^{1/2}]$.

So far we have located three *distinct* zeros of $G(\cdot, \alpha)$ for every $\alpha < 0$, and one zero for every $\alpha \geq 0$ close to zero. We have also seen that the zero at $\alpha = 0$ is unique, if we are in a sufficiently small neighbourhood of the origin $(0, 0)$.

It is easily seen that $G_{xxx}(x, \alpha) > 0$ if $|x|$ and $|\alpha|$ are small enough, since the derivatives of s_4 (with respect to x and up to order 3) are bounded. This means that for every fixed α with $|\alpha| \leq \alpha_0$ the function $x \mapsto G_x(x, \alpha)$ is strictly convex, hence it has at most 2 zeros in the vicinity of the origin. So $G(\cdot, \alpha)$ can have at most 2 local extrema, thus 3 zeros at most.

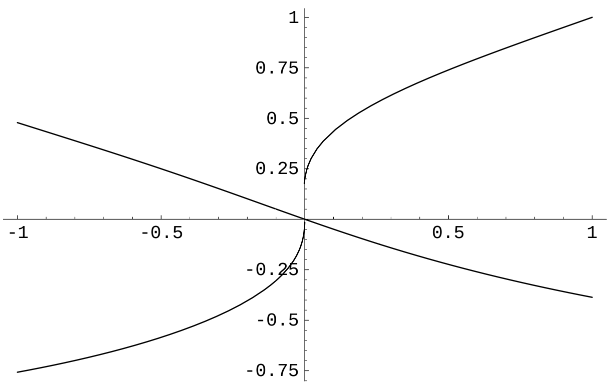
Finally, observe that $G_{x\alpha}(x, \alpha) > 0$, too, for all small $|x|$ and $|\alpha|$ values, implying that $\alpha \mapsto G_x(x, \alpha)$ is strictly increasing for any such x's, hence the graph of the convex function $x \mapsto G_x(x, \alpha)$ "moves upwards" if α increases, so the number of zeros of $G_x(\cdot, \alpha)$ can only decrease as α increases—but this means that the number of zeros of $G(\cdot, \alpha)$ is also a nonincreasing function of α.

We have thus proved that $G(\cdot, \alpha)$ has exactly 3 zeros for every negative α, and exactly one zero for any nonnegative α, in a neighbourhood of the origin, therefore, fixed points of our original map g form a local pitchfork near the origin. ■

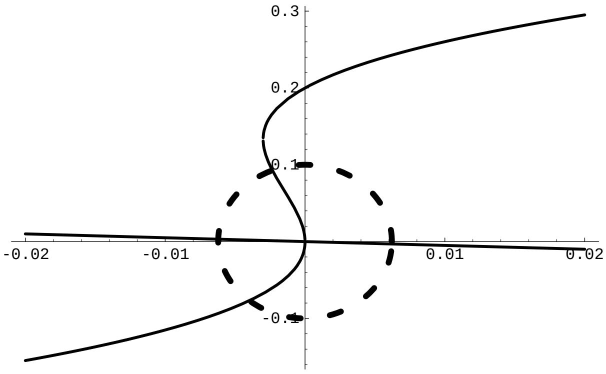
Remark 4.1.2 Let us illustrate the theorem by depicting the fixed points of the map

$$x \mapsto \alpha^2 + (1 + 2\alpha)\, x + \left(1 + \alpha^2\right) x^3 - 5\, x^4,$$

with the α-axis being horizontal:



Upon having a closer look at the origin, we locally recognize the missing pitchfork.

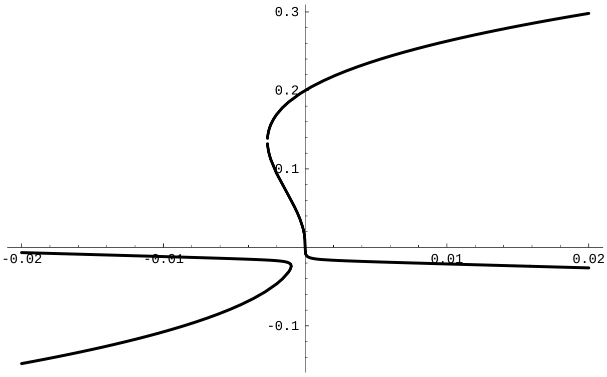


Remark 4.1.3 The function $x \mapsto G(x,\alpha)$ in (4.5) can be viewed as a perturbation of the "normal form" $x \mapsto N(x,\alpha) := \alpha x + r\,x^3$, so Theorem 4.1.1 says that there is a one-to-one correspondence between zeros of $N(\cdot,\alpha)$ and $G(\cdot,\alpha)$ for every value of the parameter α near 0, moreover, their branches of zeros have the same order of magnitude in terms of α near 0. Generalizations of these questions constitute the content of the *preparation theorems of Weierstrass* [39] or *Malgrange* [1], often encountered in singularity theory and in the theory of differential operators. How the "product structure" of a (complex or real analytic, C^∞ or C^k) function over the complex or real numbers persists under perturbations is investigated in different versions of these theorems. (As [39] puts it: "In many cases the investigation of an arbitrary function holomorphic at a point (z_0, w_0) can be reduced to the investigation of a function which is a polynomial with respect to one of the variables z, w.") Our elementary proof is thus a special case of these general theorems.

Remark 4.1.4 In the previous remark we have not specified which perturbations are allowed. If the perturbation of $N(\cdot,\alpha)$ becomes "large", for example, condition $g_\alpha(0,0) = 0$ is dropped in Theorem 4.1.1, then the structure of the branches of zeros near the origin topologically changes. As an example, consider the modified map

$$x \mapsto \frac{1}{30}\alpha + \alpha^2 + (1+2\alpha)\,x + \left(1+\alpha^2\right)x^3 - 5\,x^4.$$

The fixed points of this map are plotted below: a pitchfork is about to be born.



4.2 Construction of the normal forms

In this section, we compute normal forms for the maps

$$x \mapsto \Phi(h,x,\alpha) \tag{4.6}$$

and

$$x \mapsto \varphi(h, x, \alpha) \tag{4.7}$$

near the equilibrium being also a pitchfork bifurcation point.

To ensure that the origin $x = 0$ is a fixed point also for the discretization map (4.7), we assume that

$$\varphi(h, 0, \alpha) = 0 \tag{4.8}$$

holds for sufficiently small $h \geq 0$ and $|\alpha|$. Suppose further that

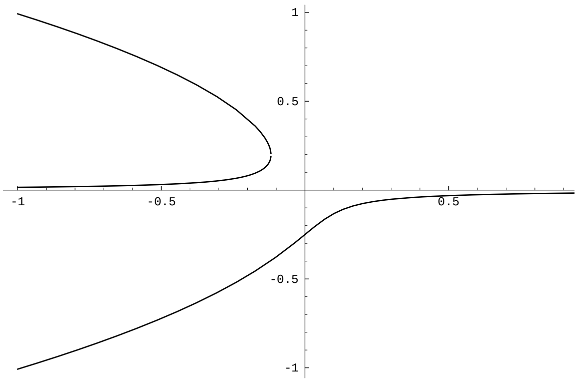
$$\varphi_x(h, 0, 0) = 1 \quad \text{and} \quad \varphi_{xx}(h, 0, 0) = 0 \tag{4.9}$$

hold for all, sufficiently small $h \geq 0$. These conditions are necessary and sufficient for (4.7) with property (4.3) to undergo a pitchfork bifurcation near the origin. Necessity is illustrated by three examples below, while sufficiency is proved by the normal form transformations themselves in the rest of the section.

Example 4.2.1 Suppose we are given a map

$$\varphi(h, x, \alpha) := h^{3p+1} + (1 + h\alpha)x + h\,x^3.$$

For this map φ, condition $\varphi(h, 0, \alpha) = 0$ does not hold, but φ satisfies (4.3) with $\Phi(h, x, \alpha) := (1 + h\alpha)x + h\,x^3$. This latter map has a pitchfork bifurcation at the origin, φ however does not, since it can be shown (for example, by computing the discriminant $-h^4\left(27h^{6p} + 4\alpha^3\right)$ of the cubic polynomial $\varphi(h, x, \alpha) - x$, or even by determining its roots exactly) that at $\alpha = \frac{-3h^{2p}}{2^{2/3}}$ the map φ has exactly two fixed points, thus it can not have a pitchfork bifurcation. The figure below depicts the solution of $\varphi(h, x, \alpha) - x = 0$ (the α-axis is horizontal and $h > 0$ is fixed).



Example 4.2.2 Suppose now we have a map

$$\varphi(h, x, \alpha) := (1 + h\alpha)x + h^{p+1}x^2 + h\,x^3.$$

Clearly, this φ violates condition $\varphi_{xx}(h, 0, 0) = 0$, but φ satisfies (4.3) with $\Phi(h, x, \alpha) := (1 + h\alpha)x + h\,x^3$. The fixed points of φ are given by $x = 0$ and $x_\pm = \frac{1}{2}\left(-h^p \pm \sqrt{h^{2p} - 4\alpha}\right)$, so at $\alpha = 0$ it has exactly two distinct fixed points, which is impossible in a pitchfork scenario.

Example 4.2.3 Suppose finally that the discretized map has the form

$$\varphi(h, x, \alpha) := (1 + h\alpha - h^{p+1})x + h\alpha x^2 + h\,x^3.$$

This φ does not satisfy $\varphi_x(h,0,0)=1$, but is sufficiently close to $\Phi(h,x,\alpha) := (1+h\alpha)x+h\alpha x^2+h\,x^3$. It is easily seen that $x \mapsto \Phi(h,x,\alpha)$ has a pitchfork at the origin. The fixed points of φ are now $x=0$ and $x_\pm = \frac{1}{2}\left(-\alpha \pm \sqrt{4\,h^p - 4\,\alpha+\alpha^2}\right)$. But, for example, at $\alpha = 2\,\left(1-\sqrt{1-h^p}\right)$, it again has exactly two fixed points, so φ can not have a pitchfork.

We remark that all Runge-Kutta methods satisfy the three requirements (4.8)-(4.9) above. The following lemma would fit naturally into Chapter 5, however, now we work in one dimension.

Lemma 4.2.1 *Suppose that $f^B=0$, $f^B_x=0$, $f^B_{xx}=0$ and let φ be a general s-stage Runge-Kutta method with step-size h, that is*

$$\varphi(h,x,\alpha) \equiv x + h\sum_{i=1}^{s}\gamma_i\cdot k_i(h,x,\alpha),$$

with every function k_i $(i=1,2,\ldots,s)$ satisfying the (implicit) equation

$$k_i(h,x,\alpha) = f(x+h\sum_{j=1}^{s}\beta_{ij}\cdot k_j(h,x,\alpha),\alpha),$$

where γ_i and β_{ij} $(i,j=1,2,\ldots,s)$ are given real parameters. Then for every $h\geq 0$ sufficiently small we have that $\varphi_x(h,0,0)=1$ and $\varphi_{xx}(h,0,0)=0$.

Proof. Since $f^B=0$, from unique solvability we get for all i that $k_i(h,0,0)\equiv 0$, implying the well-known property (4.8). On the other hand, differentiating the implicit defining equation we get that

$$(k_i)_x(h,0,0) = \left(1+h\sum_{j=1}^{s}\beta_{i,j}\,(k_j)_x(h,0,0)\right)\,f_x\left(h\sum_{j=1}^{s}\beta_{i,j}\,k_j(h,0,0),0\right).$$

But $k_j(h,0,0)\equiv 0$ and $f_x(0,0)=0$, hence $(k_i)_x(h,0,0)\equiv 0$, so $\varphi_x(h,0,0)\equiv 1$.

Differentiating the defining equation again we see that

$$(k_i)_{xx}(h,0,0) = h\sum_{j=1}^{s}\beta_{i,j}\,(k_j)_{xx}(h,0,0)\,f_x\left(h\sum_{j=1}^{s}\beta_{i,j}\,k_j(h,0,0),0\right)+$$

$$\left(1+h\sum_{j=1}^{s}\beta_{i,j}\,(k_j)_x(h,0,0)\right)^2\,f_{xx}\left(h\sum_{j=1}^{s}\beta_{i,j}\,k_j(h,0,0),0\right),$$

but since $f_x(0,0)=f_{xx}(0,0)=0$, so $(k_i)_{xx}(h,0,0)\equiv 0$ and $\varphi_{xx}(h,0,0)\equiv 0$. ■

Remark 4.2.1 It is remarkable that $f^B=0$, $f^B_x=0$ and $f^B_{xx}=0$ imply many other vanishing quantities for Runge-Kutta methods. It can be proved in a similar, recursive fashion as above, that, for example,

$$(k_i)_h(h,0,0)\equiv(k_i)_{hh}(h,0,0)\equiv(k_i)_{hx}(h,0,0)\equiv(k_i)_{hxx}(h,0,0)\equiv(k_i)_{hhxx}(h,0,0)\equiv 0.$$

These proofs require considerably more computations. A consequence, for example, is that

$$\varphi_{hx}(h,0,0)\equiv\varphi_{hhxx}(h,0,0)\equiv 0.$$

However, these latter two formulae—needed later—directly follow from (4.9) as well, so they hold not only for Runge-Kutta methods. ■

Now let us begin with the normal form transformations, first for the time-h-map (4.6), then for the discretized map (4.7).

Since the pitchfork conditions (4.4) are a special case of the corresponding transcritical conditions, we start the normal form transformation just as in the transcritical case (but truncate Taylor-expansion at fourth order instead of third). Also using the same notation and introducing the same parameter β, we get that the map $x \mapsto \Phi(h, x, \alpha)$ takes the form

$$\Phi(h, x, \alpha) = (1 + h\beta)x + hx^2 \left(\frac{1}{2}f^B_{xx} + \frac{1}{2}\overline{\alpha}_0(h, \beta) \cdot \mathrm{I}_{121}(h, \overline{\alpha}_0(h, \beta))\right) +$$

$$\frac{1}{6}x^3 \left(\Phi^B_{xxx} + \alpha \cdot \mathrm{I}_{031}(\alpha) + h \cdot \Phi^B_{hxxx} + h\alpha \cdot \mathrm{I}_{131}(h, \alpha) + h^2 \cdot \mathrm{I}_{230}(h)\right) + x^4\psi_4(h, x, \alpha)$$

with some smooth function ψ_4, and integrals I_{031}, I_{131}, I_{230} defined analogously as the other I's.

Here however $f^B_{xx} = 0$ by (4.4), and take into account that $\overline{\alpha}_0(h, \beta) = \beta \cdot \psi_a(h, \beta)$, as we have seen in the transcritical case. It is easy to see that $\Phi^B_{xxx} = 0$, $\Phi^B_{hxxx} = f^B_{xxx}$ and

$$\mathrm{I}_{031}(\alpha) = \int_0^1 \Phi_{xxx\alpha}(0, 0, \tau\alpha)\mathrm{d}\tau \equiv 0.$$

Now h can be factored out again from the last term, that is $\psi_4(h, x, \alpha) = h \cdot \widehat{\psi}_4(h, x, \alpha)$ holds with a smooth function $\widehat{\psi}_4$. With these considerations we can further rewrite $\Phi(h, x, \alpha)$ as

$$\Phi(h, x, \alpha) = (1 + h\beta)x + \frac{1}{2}hx^2 \left(\beta \cdot \psi_a(h, \beta) \cdot \mathrm{I}_{121}(h, \overline{\alpha}_0(h, \beta))\right) +$$

$$\frac{1}{6}hx^3 \left(f^B_{xxx} + \beta \cdot \psi_a(h, \beta) \cdot \mathrm{I}_{131}(h, \overline{\alpha}_0(h, \beta)) + h \cdot \mathrm{I}_{230}(h)\right) + hx^4\widehat{\psi}_4(h, x, \overline{\alpha}_0(h, \beta)).$$

Thus, by appropriately defining the smooth functions s_2, s_3 and s_4, we have transformed (4.6) into

$$x \mapsto \Psi(h, x, \beta) := (1 + h\beta)x + \frac{1}{2}hx^2\,\beta\, s_2(h, \beta) + \frac{1}{6}hx^3\, s_3(h, \beta) + hx^4\, s_4(h, x, \beta). \tag{4.10}$$

Notice that, due to condition $f^B_{xxx} \neq 0$, the smooth function s_3 is nonzero and has a constant sign, provided that h and $|\beta|$ are small. This fact will be important in the final scaling.

We are now concerned with eliminating the quadratic term from (4.10) with—as usual—a smooth and invertible transformation. (If f^B_{xx} were nonzero, as, for example, in the transcritical or the fold case, such transformation would not exist.) We will employ the same type of nonlinear transformation as [35] used in the case of the flip bifurcation to remove the quadratic term. (As opposed to [35] however, we fortunately have a multiplier β in front of s_2 in (4.10). This is crucial, since otherwise that transformation would be "truly" singular. Anyway, we will work with singular expressions, but these will only have removable singularities.) Moreover, we should carefully keep track of the extra parameter h, which will make our task harder.

So let us define a smooth invertible near-identity transformation T for any $h \geq 0$ and β, sufficiently close to zero, as

$$x \mapsto T^E(x) \equiv T(h, x, \beta) := x + \frac{s_2(h, \beta)}{2(1 + h\beta)}x^2. \tag{4.11}$$

(The evaluation operator ^{E} now evaluates at (h, x, β).) Then the inverse function of T^E reads as

$$(T^E)^{[-1]}(x) = \frac{-1 - h\beta + \sqrt{1+h\beta}\,\sqrt{1+h\beta+2\,x\,s_2(h,\beta)}}{s_2(h,\beta)}, \tag{4.12}$$

if $s_2(h,\beta) \neq 0$ and $(T^E)^{[-1]}(x) = x$, if $s_2(h,\beta) = 0$.

Later it will be convenient to know how $(T^E)^{[-1]}$ looks like in terms of x. Standard Taylor expansion tells us that it has the form

$$(T^E)^{[-1]}(x) = \tag{4.13}$$

$$x - \frac{s_2(h,\beta)}{2(1+h\,\beta)}x^2 + \frac{s_2(h,\beta)^2}{2(1+h\,\beta)^2}x^3 - \frac{x^4}{3!}\int_0^1 \frac{15\sqrt{1+h\beta}\,s_2(h,\beta)^3}{(1+h\beta+2\tau x\,s_2(h,\beta))^{\frac{7}{2}}}\cdot(1-\tau)^3 \mathrm{d}\tau\,.$$

With this T in hand we transform map (4.6) into

$$x \mapsto \Xi^E(x) := \left((T^E)^{[-1]} \circ \Psi^E \circ T^E\right)(x).$$

We now prove that this new map Ξ has the desired property, that is, it no longer contains any quadratic term. However, it will be also important for us that h's are preserved in front of the cubic and quartic terms, just as in the expansion of Ψ. Fortunately this can be proved, though it is not apparent at all: neither $(T^E)^{[-1]}$ nor T contain a factor h (see Remark 4.2.2 below).

The notation and abbreviations of the following lemma should be considered only as "local", not to conflict with any existing notation outside the lemma. (The function Ψ is the same as before, but the dependence on parameter β is now irrelevant, hence suppressed in the lemma.)

Lemma 4.2.2 *Suppose we have a near-identity, invertible, quadratic change of coordinates Q depending on a parameter h, that is $Q(h,x) := x + b(h)x^2$ with some bounded function b. For any fixed $h \geq 0$ let $(Q^E)^{[-1]}$ denote the inverse function of $x \mapsto Q(h,x)$ and set $\Pi(h,x) := \left((Q^E)^{[-1]} \circ \Psi^E \circ Q^E\right)(x)$. Then*

$$\Pi(h,0) = 0,$$

$$\left(\frac{\mathrm{d}}{\mathrm{d}x}\Pi\right)(h,0) = 1 + h\beta,$$

$$\left(\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Pi\right)(h,0) = h\beta\,\left(-2\,(1+h\,\beta)\,b(h) + s_2(h,\beta)\right),$$

$$\left(\frac{\mathrm{d}^3}{\mathrm{d}x^3}\Pi\right)(h,0) = h\left(12\beta\,(1+h\,\beta)^2\,b(h)^2 - 6h\beta^2\,b(h)\,s_2(h,\beta) + s_3(h,\beta)\right),$$

$$\left(\frac{\mathrm{d}^4}{\mathrm{d}x^4}\Pi\right)(h,x) = \frac{-24\,b^3\,{\psi_1}^2}{D^3} + \frac{12\,b^2\,\psi_2}{D} + \frac{144\,b^3\,{\psi_1}^3\,q^2}{D^5} - \frac{72\,b^2\,\psi_1\,\psi_2\,q^2}{D^3} +$$

$$\frac{12\,b\,\psi_3\,q^2}{D} - \frac{120\,b^3\,{\psi_1}^4\,q^4}{D^7} + \frac{72\,b^2\,{\psi_1}^2\,\psi_2\,q^4}{D^5} - \frac{6\,b\,{\psi_2}^2\,q^4}{D^3} - \frac{8\,b\,\psi_1\,\psi_3\,q^4}{D^3} + \frac{\psi_4\,q^4}{D},$$

where $D := (Q^E)'\left((Q^E)^{[-1]}(\Psi(h,Q^E(x)))\right)$, $\psi_i := \left(\frac{\mathrm{d}^i}{\mathrm{d}x^i}\Psi\right)(h,Q^E(x))$ ($i = 1,2,3,4$), $b := b(h)$ and $q := (Q^E)'(x)$.
Finally,

$$\left(\frac{\mathrm{d}^4}{\mathrm{d}x^4}\Pi\right)(0,x) = \frac{-24\,{b(0)}^3}{{D_0}^3} + \frac{144\,{b(0)}^3\,{q_0}^2}{{D_0}^5} - \frac{120\,{b(0)}^3\,{q_0}^4}{{D_0}^7} \equiv 0,$$

where D_0 and q_0 are D and q above, respectively, evaluated at $h = 0$.

Proof. The proof is nothing else but formal differentiation of the composition Π.

For expressions $\left(\frac{\mathrm{d}^i}{\mathrm{d}x^i}\,\Pi\right)(h,0)$ $(i=0,1,2,3)$ we have taken into account *only the following* pieces of information—yielding some further generalizations of the lemma: $Q^E(0)=0$, $(Q^E)^{[-1]}(0)=0$, $(Q^E)'(0)=1$, $(Q^E)''(0)=2b(h)$, $(Q^E)'''(0)=0$, $(Q^E)''''(0)=0$, $\Psi(h,0)=0$, $\left(\frac{\mathrm{d}}{\mathrm{d}x}\,\Psi\right)(h,0)=1+h\beta$, $\left(\frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\Psi\right)(h,0)=h\beta\, s_2(h,\beta)$, $\left(\frac{\mathrm{d}^3}{\mathrm{d}x^3}\,\Psi\right)(h,0)=h\, s_3(h,\beta)$. Beyond these requirements the concrete form of $Q(h,x)$ and $\Psi(h,x)$ is irrelevant.

For $\left(\frac{\mathrm{d}^4}{\mathrm{d}x^4}\,\Pi\right)(h,x)$ we have *only* used properties $(Q^E)''(\cdot)=2b(h)$, $(Q^E)'''(\cdot)=0$ and $(Q^E)''''(\cdot)=0$, where "$\cdot$" stands for any argument. (All of these are easily implemented in *Mathematica* as replacement rules.)

Finally, for $\left(\frac{\mathrm{d}^4}{\mathrm{d}x^4}\,\Pi\right)(0,x)$ one takes into consideration *only* that $\left(\frac{\mathrm{d}}{\mathrm{d}x}\,\Psi\right)(0,Q^{E_0}(x))=1$, $\left(\frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\Psi\right)(0,Q^{E_0}(x))=\left(\frac{\mathrm{d}^3}{\mathrm{d}x^3}\,\Psi\right)(0,Q^{E_0}(x))=\left(\frac{\mathrm{d}^4}{\mathrm{d}x^4}\,\Psi\right)(0,Q^{E_0}(x))=0$, $(Q^{E_0})''(\cdot)=2b(0)$, $(Q^{E_0})'''(\cdot)=0$, $(Q^{E_0})''''(\cdot)=0$, where superscript ^{E_0} denotes evaluation at $h=0$. The fact that $(Q^{E_0})^{[-1]}(\Psi(0,Q^{E_0}(x)))=x$ means that $D_0=q_0$, so the last expression is identically zero indeed. ■

By Taylor expansion

$$\Xi^E(x)=$$

$$\Xi^E(0)+(\Xi^E)'(0)x+\frac{1}{2!}(\Xi^E)''(0)x^2+\frac{1}{3!}(\Xi^E)'''(0)x^3+\frac{x^4}{3!}\int_0^1(\Xi^E)''''(\tau x)\cdot(1-\tau)^3\mathrm{d}\tau,$$

where the evaluation operator ^{E}, as before, evaluates at general h and β values.

From the lemma we conclude that the constant term above is always 0 and the linear term has coefficient $(1+h\beta)$, for any choice of $b(h)$ in (4.11). We also get that the quadratic term of Ξ can be eliminated if and only if

$$b(h)=\frac{s_2(h,\beta)}{2(1+h\beta)},$$

justifying the definition of T. But beyond this, the above form of $b(h)$ again does *not* play any role: as established by the lemma, a multiplier h *always* appears in the coefficient of x^3 *and* in that of x^4, regardless of the choice of the bounded function $b(h)$.

Remark 4.2.2 In the lemma above we have used the *exact and full* form of the inverse function $(Q^E)^{[-1]}$ (or $(T^E)^{[-1]}$). It is interesting that this seems necessary: we could *not* prove the presence of a factor h in the coefficient of x^4 in the expansion of Ξ when we worked with a formula like $(T^E)^{[-1]}(x)=x-\frac{s_2(h,\beta)}{2(1+h\,\beta)}x^2+\frac{s_2(h,\beta)^2}{2(1+h\,\beta)^2}x^3+x^4\,s^*(h,x,\beta)$ (*cf.* (4.13)) without any further assumption on the structure of the smooth function s^*.

An alternative (and more general) approach to establish that h is present in the coefficient of x^4 would be to compare coefficients of x on both sides of the equation $T^E\circ\Xi^E=\Psi^E\circ T^E$. This avoids computing the inverse of T. The knowledge of the exact form of this inverse, however, will ease the normal form transformations for (4.7).

We proceed further. Let us choose $b(h)$ as in (4.11). Then by simple substitution into the third derivative in the lemma above, we get that

$$\Xi^E(x)=(1+h\beta)x+h\cdot\widehat{s}_3(h,\beta)x^3+h\cdot\widehat{s}_4(h,x,\beta)x^4$$

with $\widehat{s}_3(h,\beta):=\frac{(1+h\beta)\,s_3(h,\beta)+3\beta\,s_2(h,\beta)^2}{6+6h\beta}$ and some smooth function $\widehat{s}_4$.

Now we apply a final scaling $\eta := \frac{x}{\sqrt{|\widehat{s}_3(h,\beta)|}}$. Then it is easy to see that $x \mapsto \Xi^E(x)$ becomes

$$\eta \mapsto (1+h\beta)\eta + h \cdot s \cdot \eta^3 + h \cdot \widehat{\eta}_4(h,\eta,\beta)\eta^4 \tag{4.14}$$

with a smooth $\widehat{\eta}_4(h,\eta,\beta) := \frac{\widehat{s}_4(h,\eta\sqrt{|\widehat{s}_3(h,\beta)|},\beta)}{|\widehat{s}_3(h,\beta)|^{3/2}}$ and $s := \operatorname{sign}(\widehat{s}_3(h,\beta)) = \pm 1$, the sign being independent of $h \in [0,h_0]$ and β sufficiently close to zero, since for such parameter values

$$\operatorname{sign}(\widehat{s}_3(h,\beta)) \equiv \operatorname{sign}(s_3(h,\beta)) \equiv \operatorname{sign}(f^B_{xxx})$$

by (4.10).

Lemma 4.2.3 *There are smooth invertible coordinate and parameter changes transforming the system $x \mapsto \Phi(h,x,\alpha)$ into (4.14).* ■

Now let us consider the discretization map φ. Again, the normal form transformation begins exactly as in the transcritical case and we will use notation of the corresponding lemma in the transcritical case.

It is easy to see (*cf.* Remark 4.2.1) that conditions (4.9) imply $\widetilde{\mathrm{I}}_{110}(h) \equiv 0$ and $\widetilde{\mathrm{I}}_{220}(h) \equiv 0$, so $\chi_{10}(h) \equiv 0$ and $\chi_{20}(h) \equiv 0$. Taking into account that now $f^B_{xx} = 0$, we get that

$$\varphi(h,x,\alpha) = (1 + h\alpha \cdot f^B_{x\alpha} + h\alpha \cdot \chi_{11}(h,\alpha))x + h\alpha \cdot \chi_{21}(h,\alpha)x^2 +$$

$$\frac{1}{6}x^3\left(\varphi^B_{xxx} + \alpha \cdot \widetilde{\mathrm{I}}_{031}(\alpha) + h \cdot \varphi^B_{hxxx} + h\alpha \cdot \widetilde{\mathrm{I}}_{131}(h,\alpha) + h^2 \cdot \widetilde{\mathrm{I}}_{230}(h)\right) + x^4\chi_4(h,x,\alpha).$$

Now observe that $\varphi^B_{xxx} = 0$ and $\varphi^B_{hxxx} = \Phi^B_{hxxx} = f^B_{xxx}$. It is also true that $\varphi_{xxx\alpha}(0,0,\alpha) = \Phi_{xxx\alpha}(0,0,\alpha) \equiv 0$, so $\widetilde{\mathrm{I}}_{031}(\alpha) \equiv 0$. Finally, $\varphi_{xxxx}(0,\tau x,\alpha) \equiv 0$ and

$$\varphi_{xxxx}(h,\tau x,\alpha) = \varphi_{xxxx}(0,\tau x,\alpha) + h \cdot \int_0^1 \varphi_{hxxxx}(\sigma h,\tau x,\alpha)\mathrm{d}\sigma,$$

so $\chi_4(h,x,\alpha) = h \cdot \widetilde{\chi}_4(h,x,\alpha)$ with some smooth function $\widetilde{\chi}_4$. This means that

$$\varphi(h,x,\alpha) = (1 + h\alpha \cdot f^B_{x\alpha} + h\alpha \cdot \chi_{11}(h,\alpha))x + h\alpha \cdot \chi_{21}(h,\alpha)x^2 +$$

$$\frac{1}{6}h\left(f^B_{xxx} + \alpha \cdot \widetilde{\mathrm{I}}_{131}(h,\alpha) + h \cdot \widetilde{\mathrm{I}}_{230}(h)\right)x^3 + h \cdot \widetilde{\chi}_4(h,x,\alpha)x^4.$$

The function $\widetilde{\beta}(h,\alpha)$ is defined analogously as in the transcritical case. Since $\widetilde{\mathrm{I}}_{110}(h) \equiv 0$, $\widetilde{\beta}(h,0) = 0$, so the existing inverse $\widetilde{\alpha}(h,\cdot)$ of $\widetilde{\beta}(h,\cdot)$ can be factored as $\widetilde{\alpha}(h,\beta) = \beta \cdot \widetilde{\psi}_a(h,\beta)$ with some smooth $\widetilde{\psi}_a$. Therefore (4.7) has been transformed into

$$x \mapsto \widetilde{\Psi}(h,x,\widetilde{\beta}) := (1+h\widetilde{\beta})x + \frac{1}{2}hx^2\,\widetilde{\beta}\,\widetilde{s}_2(h,\widetilde{\beta}) + \frac{1}{6}hx^3\,\widetilde{s}_3(h,\widetilde{\beta}) + hx^4\,\widetilde{s}_4(h,x,\widetilde{\beta})$$

with some smooth functions $\widetilde{s}_2$, $\widetilde{s}_3$ and $\widetilde{s}_4$. The same arguments as in the transcritical case show that $|\beta(h,\alpha) - \widetilde{\beta}(h,\alpha)| \le const \cdot h^p$, $|\widetilde{\alpha}(h,\beta) - \overline{\alpha}_0(h,\beta)| \le const \cdot h^p$, $|s_i(h,\beta) - \widetilde{s}_i(h,\beta)| \le const \cdot h^p$ $(i = 2,3)$ and $|s_4(h,x,\beta) - \widetilde{s}_4(h,x,\beta)| \le const \cdot h^p$.

The analogue of transformation T is defined now as

$$x \mapsto \widetilde{T}(h,x,\widetilde{\beta}) := x + \frac{\widetilde{s}_2(h,\widetilde{\beta})}{2(1+h\widetilde{\beta})}x^2.$$

Since T and $\widetilde{T}$ have the same form, Lemma 4.2.2 guarantees that transformation

$$x \mapsto \widetilde{\Xi}(h,x,\widetilde{\beta}) := \left(\left(\widetilde{T}(h,\cdot,\widetilde{\beta}) \right)^{[-1]} \circ \widetilde{\Psi}(h,\cdot,\widetilde{\beta}) \circ \widetilde{T}(h,\cdot,\widetilde{\beta}) \right)(x)$$

has the desired structure, that is, it does not contain any quadratic term, further, a factor h is present in the coefficients of x^3 and x^4. Therefore, with some smooth functions $\widetilde{\widehat{s_3}}$ and $\widetilde{\widehat{s_4}}$ (*cf.* the definition of $\widehat{s_3}$ and $\widehat{s_4}$ after Remark 4.2.2) we have that

$$\widetilde{\Xi}(h,x,\widetilde{\beta}) = (1+h\widetilde{\beta})x + h \cdot \widetilde{\widehat{s_3}}(h,\widetilde{\beta})x^3 + h \cdot \widetilde{\widehat{s_4}}(h,x,\widetilde{\beta})x^4.$$

After a final scaling (*cf.* (4.14)) and sign consideration, we have proved that (4.7) can be transformed into

$$\widetilde{\eta} \mapsto (1+h\widetilde{\beta})\widetilde{\eta} + h \cdot s \cdot \widetilde{\eta}^3 + h \cdot \widetilde{\eta}_4(h,\widetilde{\eta},\widetilde{\beta})\widetilde{\eta}^4$$

with the same $s = \pm 1$ as in (4.14) and a smooth function $\widetilde{\eta}_4$.

So far, the *structure* of the transformation has been considered. As for the *closeness* estimates, one uses a series of standard triangle inequalities and mean value theorems to obtain $\mathcal{O}(h^p)$-closeness. More precisely, since all the necessary derivatives are assumed to be bounded, and we know that $|\beta - \widetilde{\beta}| = \mathcal{O}(h^p)$ and $|s_i - \widetilde{s_i}| = \mathcal{O}(h^p)$ $(i = 2,3,4)$, we get that

$$\left| T(h,x,\beta) - \widetilde{T}(h,x,\widetilde{\beta}) \right| = \mathcal{O}(h^p),$$

$$\left| \Psi(h,x,\beta) - \widetilde{\Psi}(h,x,\widetilde{\beta}) \right| = \mathcal{O}(h^p),$$

and from (4.12) or (4.13) that

$$\left| (T(h,\cdot,\beta))^{[-1]} - \left(\widetilde{T}(h,\cdot,\widetilde{\beta}) \right)^{[-1]} \right| (x) = \mathcal{O}(h^p),$$

so

$$\left| \Xi(h,x,\beta) - \widetilde{\Xi}(h,x,\widetilde{\beta}) \right| = \mathcal{O}(h^p),$$

and

$$\left| \widehat{s_3}(h,\beta) - \widetilde{\widehat{s_3}}(h,\widetilde{\beta}) \right| = \mathcal{O}(h^p),$$

$$\left| \widehat{s_4}(h,x,\beta) - \widetilde{\widehat{s_4}}(h,x,\widetilde{\beta}) \right| = \mathcal{O}(h^p),$$

and

$$\left| \widehat{\eta}_4(h,\eta,\beta) - \widetilde{\eta}_4(h,\eta,\widetilde{\beta}) \right| = \mathcal{O}(h^p).$$

The following theorem has been established.

Theorem 4.2.4 *There are smooth invertible coordinate and parameter changes transforming the system*

$$x \mapsto \varphi(h,x,\alpha)$$

into

$$\widetilde{\eta} \mapsto (1+h\widetilde{\beta})\widetilde{\eta} + s \cdot h\widetilde{\eta}^3 + h\widetilde{\eta}^4 \cdot \widetilde{\eta}_4(h,\widetilde{\eta},\widetilde{\beta})$$

where $\widetilde{\eta}_4$ is a smooth function.

Moreover, the smooth invertible coordinate and parameter changes above and those in Lemma 4.2.3 are $\mathcal{O}(h^p)$-close to each other, further

$$|\widehat{\eta}_4 - \widetilde{\eta}_4| \leq const \cdot h^p \qquad \blacksquare$$

To finish the normal form transformations, we can apply a parameter shift $\widetilde{\beta} \mapsto \beta$ (being $\mathcal{O}(h^p)$-close to the identity, as we have seen) to the normal form of the discretization mapping above. So from now on the bifurcation parameter α is used again instead of β and $\widetilde{\beta}$. To simplify our notation further, instead of dummy variables η and $\widetilde{\eta}$ the letter x will be used, and subscript $_4$ in $\widehat{\eta}_4$ and $\widetilde{\eta}_4$ is dropped.

4.3 Construction of the conjugacy

We have derived that the normal forms in the pitchfork case are

$$\mathcal{N}_\Phi(h,x,\alpha) = (1+h\alpha)x + s\cdot hx^3 + hx^4\,\widehat{\eta}(h,x,\alpha) \tag{4.15}$$

$$\mathcal{N}_\varphi(h,x,\alpha) = (1+h\alpha)x + s\cdot hx^3 + hx^4\,\widetilde{\eta}(h,x,\alpha) \tag{4.16}$$

with $s=1$ or $s=-1$, where $\widehat{\eta}$ and $\widetilde{\eta}$ are smooth functions. Let $K>0$ denote a uniform bound on $\left|\frac{\mathrm{d}^i}{\mathrm{d}x^i}\,\eta(h,\cdot,\alpha)\right|$ $(i\in\{0,1,2\}, \eta\in\{\widehat{\eta},\widetilde{\eta}\})$ in a neighbourhood of the origin for any small $h>0$ and $|\alpha|$, as well as a uniform bound on $\left|\frac{\mathrm{d}}{\mathrm{d}\alpha}\,\eta(h,x,\cdot)\right|$ $(\eta\in\{\widehat{\eta},\widetilde{\eta}\})$ in a neighbourhood of the origin for any small $h>0$ and $|x|$. We also have that there exists a constant $c>0$ such that

$$|\mathcal{N}_\Phi(h,x,\alpha) - \mathcal{N}_\varphi(h,x,\alpha)| \le c\cdot h^{p+1}x^4 \tag{4.17}$$

holds for all sufficiently small $h>0$, $|x|\ge 0$ and $|\alpha|\ge 0$. Throughout the section, c will denote this particular positive constant.

We will consider the case $s=-1$, the other one is similar due to symmetry. Then it is easy to see that $\omega_{\Phi,0}(h,\alpha)\equiv 0$ is an attracting fixed point of the map $\mathcal{N}_\Phi(h,\cdot,\alpha)$ for $\alpha\le 0$, and repelling for $\alpha>0$. For any fixed $h>0$ and $\alpha>0$, this map possesses another two attracting fixed points, denoted by $\omega_{\Phi,+}\equiv\omega_{\Phi,+}(h,\alpha)>0$ and $\omega_{\Phi,-}\equiv\omega_{\Phi,-}(h,\alpha)<0$. The three branches of fixed points, $\omega_{\Phi,0}(h,\alpha)$ and $\omega_{\Phi,\pm}(h,\alpha)$ merge at $\alpha=0$.

Analogous results hold for the map $\mathcal{N}_\varphi(h,\cdot,\alpha)$. Its fixed points are denoted by $\omega_{\varphi,0}$ and $\omega_{\varphi,-}$ (or $\omega_{\varphi,+}$).

The construction of a conjugacy is completely analogous to the transcritical case (the braches of fixed points of the pitchfork and the transcritical bifurcation in the lower half-plane $x\le 0$ look topologically the same), there will be only one minor difference in a starting value. The development of the closeness estimates is also similar in character, however, the proofs of the two key lemmas will be a bit more technical. Here, again, we formulate and prove estimates only in the $x\le 0$ region—the $x>0$ case is symmetrical.

In what follows, suppose that

$$\begin{aligned} 0<h\le h_0 &:= \min\left(\frac{1}{10}, 8K^2\right), \\ |x|\le \varepsilon_0 &:= \min\left(\frac{1}{10}, \frac{1}{5K}\right) \text{ and} \\ |\alpha|\le\alpha_0 &:= \min\left(\frac{1}{288}, \frac{1}{72K^2}\right). \end{aligned} \tag{4.18}$$

(However, if the domain of definition of the functions $\widehat{\eta}$ and $\widetilde{\eta}$ is smaller than $(0,h_0]\times[-\varepsilon_0,\varepsilon_0]\times[-\alpha_0,\alpha_0]$ given above, then h_0, ε_0 or α_0 should be decreased further.)

Lemma 4.3.1 *For every* $0 < h \le h_0$ *and* $0 < \alpha \le \alpha_0$ *we have that*

$$\{\omega_{\varphi,-}, \omega_{\Phi,-}\} \subset \left(-\sqrt{2\alpha}, -\frac{4}{5}\sqrt{\alpha}\right) \subset \left(-\sqrt{2\alpha}, -\sqrt{\frac{3}{5}\alpha}\right).$$

Proof. By definition, $\omega_{\varphi,-} < 0$ solves $\alpha - x^2 + x^3 \cdot \widetilde{\eta}(h,x,\alpha) = 0$, that is

$$\omega_{\varphi,-} = -\sqrt{\frac{\alpha}{1 - \omega_{\varphi,-} \cdot \widetilde{\eta}(h, \omega_{\varphi,-}, \alpha)}}.$$

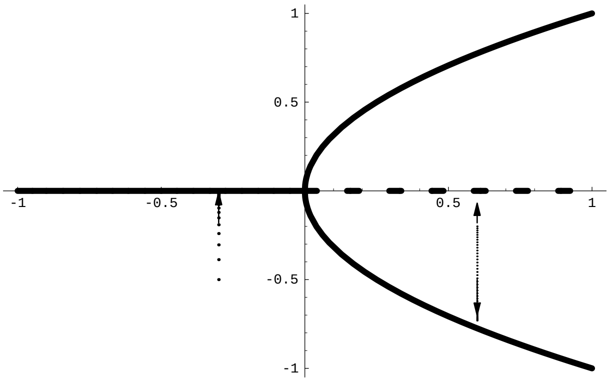
But $|x\,\widetilde{\eta}| \le \varepsilon_0 K \le \frac{1}{2}$ implies, for example, $\frac{1}{2} \le 1 - \omega_{\varphi,-} \cdot \widetilde{\eta}(h, \omega_{\varphi,-}, \alpha) \le \frac{25}{16}$, completing the proof. (In some calculations, the weaker upper bound $-\sqrt{\frac{3}{5}\alpha}$ will yield a numerically simpler result.) The proof for $\omega_{\Phi,-}$ is similar. ■

The construction of the homeomorphism J^E is completely analogous to that in the transcritical case, hence is omitted here. The only difference is that we set

$$x_0 := -\sqrt{\frac{\alpha}{8}} \qquad (\alpha > 0)$$

as a starting value in the inner region. (Set further $z_0 := -\varepsilon_0$. Then conditions (4.18) imply that $\mathcal{N}_\varphi^E$ and $\mathcal{N}_\Phi^E$ (together with their inverses) are monotone increasing, further $|\alpha| < \frac{392}{K^2}$ implies $x_0(\alpha) > x_1(h,\alpha)$ and $\sqrt{2\alpha_0} < |z_0| < \frac{1}{2K}$ implies $z_0 < z_1(h,\alpha)$. This means that x_n is monotone decreasing, y_n is monotone increasing (if $\alpha > 0$ and $n \ge 0$), and $\lim_{n\to\infty} x_n(h,\alpha) = \omega_{\varphi,-}$, while $\lim_{n\to\infty} y_n(h,\alpha) = \omega_{\varphi,0}$. Moreover, z_n is monotone increasing, further, for $\alpha > 0$, $\lim_{n\to\infty} z_n(h,\alpha) = \omega_{\varphi,-}$ and for $\alpha \le 0$, $\lim_{n\to\infty} z_n(h,\alpha) = \omega_{\varphi,0}$.)

The following figure shows the branch of stable and unstable fixed points of $\mathcal{N}_\varphi^E$ in the (α,x)-plane together with the first few terms of the inner sequences ($x_n(h,\alpha)$ and $y_n(h,\alpha)$), and the outer sequence $z_n(h,\alpha)$ with some $h > 0$ and α fixed. The arrows indicate the direction of the sequences.



4.4 The closeness estimate for the conjugacy

4.4.1 Optimality at the fixed points

We first prove that the constructed conjugacy J^E is $\mathcal{O}(h^p\,\alpha)$-close to the identity at the fixed points $\omega_{\varphi,-}(h,\alpha)$, further, an explicit example shows that this estimate is optimal in h and α.

This means that our $\mathcal{O}(h^p)$-closeness estimates for $|\,id - J^E|$ near a pitchfork bifurcation point on $(0, h_0] \times [-\varepsilon_0, \varepsilon_0] \times [-\alpha_0, \alpha_0]$ are optimal in h.

The following auxiliary estimate will frequently be used.

Lemma 4.4.1 *For any* $0 < h \le h_0$, $-\varepsilon_0 \le x < 0$ *and* $-\alpha_0 \le \alpha \le \alpha_0$, *we have that*

$$(\mathcal{N}_\Phi^E)'(x) \le 1 + h\alpha - \frac{5}{2}hx^2.$$

Proof. The conditions in (4.18) have been set up to imply this inequality. ■

Lemma 4.4.2 *For any* $0 < h \le h_0$ *and* $0 < \alpha \le \alpha_0$ *(satisfying (4.18)), we have that*

$$|\,\omega_{\varphi,-} - \omega_{\Phi,-}| \le 8c \cdot h^p\,\alpha.$$

Proof. (*cf.* its transcritical counterpart) By Lemma 4.3.1, (4.17) and Lemma 4.4.1 we have that

$$|\,\omega_{\varphi,-} - \omega_{\Phi,-}| \le 4c \cdot h^{p+1}\alpha^2 + \left(1 - \frac{h\alpha}{2}\right)|\,\omega_{\varphi,-} - \omega_{\Phi,-}|,$$

which yields the desired result. ■

Remark 4.4.1 on optimality. The example below shows a situation when the distance of fixed points of normal forms satisfying (4.17) is bounded from *below* by $\mathcal{O}(h^p)$ $(h \to 0)$. Since now we are going to deal with cubic polynomials, we do not attempt to construct their explicit solutions (as we did with the quadratic polynomials in the transcritical case), but we give a more general, yet simpler argument.

Set $\mathcal{N}_\Phi(h, x, \alpha) := (1 + h\alpha)x - hx^3$ and $\mathcal{N}_\varphi(h, x, \alpha) := (1 + h\alpha)x - hx^3 + h^{p+1}x^4$. Then these maps satisfy (4.17) in a neighbourhood of the origin, further, $\omega_{\Phi,-} = -\sqrt{\alpha}$. As for $\omega_{\varphi,-}$, we see that $\omega_{\varphi,-} = -\sqrt{\frac{\alpha}{1-h^p\,\omega_{\varphi,-}}}$. Then by Lemma 4.3.1 we get that

$$\omega_{\varphi,-} \in \left(-\sqrt{\frac{\alpha}{1 + h^p\,\frac{4}{5}\sqrt{\alpha}}}, -\sqrt{\frac{\alpha}{1 + h^p\,\sqrt{2\alpha}}}\right)$$

so $\omega_{\varphi,-} > -\sqrt{\alpha} = \omega_{\Phi,-}$, yielding

$$|\,\omega_{\varphi,-} - \omega_{\Phi,-}| \ge \left|-\sqrt{\alpha} + \sqrt{\frac{\alpha}{1 + h^p\,\frac{4}{5}\sqrt{\alpha}}}\right| = \sqrt{\alpha}\left|1 - \frac{1}{\sqrt{1+t}}\right|$$

with $t := h^p\,\frac{4}{5}\sqrt{\alpha}$. Then, by (4.18), $t \in (0, 1)$. But for any such t

$$\left|1 - \frac{1}{\sqrt{1+t}}\right| = \frac{t}{\sqrt{1+t}(1 + \sqrt{1+t})} \ge \frac{t}{4},$$

hence

$$|\,\omega_{\varphi,-} - \omega_{\Phi,-}| \ge \frac{1}{5}h^p\,\alpha.$$

4.4.2 The inner region

The closeness estimate in $(\omega_{\varphi,-}, x_0]$ is proved for any fixed $0 < h \le h_0$ and $0 < \alpha \le \alpha_0$ in a similar way as in the transcritical case, hence most intermediate steps and inequalities are left out or only sketched. However, the key Lemmas 4.4.3 and 4.4.8 are carefully examined.

Now we have that

$$\sup_{[x_1,x_0]} |\,id - J^E| = \frac{c}{64} h^{p+1}\alpha^2,$$

while for $n \ge 1$

$$\sup_{[x_{n+1},x_n]} |\,id - J^E| \le \sup_{[x_n,x_{n-1}]} \left|\mathcal{N}_\varphi^E - \mathcal{N}_\Phi^E\right| + \sup_{x\in[x_n,x_{n-1}]} \left(\left(\sup_{[\{x,J^E(x)\}]} (\mathcal{N}_\Phi^E)'\right) |\,x - J^E(x)|\right)$$

$$\le c\cdot h^{p+1}x_n^4 + \left(1 + h\alpha - \frac{5}{2}h \max\left(x_{n-1}, J^E(x_{n-1})\right)^2\right) \sup_{[x_n,x_{n-1}]} |\,id - J^E|,$$

where we have used Lemma 4.4.1, the fact that the functions id and J^E are increasing, further inequality

$$\sup_{[\{x,J^E(x)\}]} (\mathcal{N}_\Phi^E)' \le \sup_{[\{x,J^E(x)\}]} (1 + h\alpha - \frac{5}{2}h\cdot id^2) \le 1 + h\alpha - \frac{5}{2}h \max\left(x, J^E(x)\right)^2.$$

In order to prove that the conjugacy J^E is $\mathcal{O}(h^p)$-close to the identity on $(\omega_{\varphi,-}, x_0]$ for any $h \in (0, h_0]$ and $\alpha \in (0, \alpha_0]$, we will show that

$$\sup_{h\in(0,h_0]}\ \sup_{\alpha\in(0,\alpha_0]}\ \sup_{n\in\mathbb{N}} h \sum_{i=0}^{n} x_i^4 \prod_{j=i}^{n-1} \left(1 + h\alpha - \frac{5}{2}h \max\left(x_j, J^E(x_j)\right)^2\right) \le const \tag{4.19}$$

holds with a suitable $const \ge 0$ (where $\prod_{j=n}^{n-1}$ is understood to be 1).

First an explicit upper estimate of the sequence $\max\left(x_n, J^E(x_n)\right)$ is given.

Lemma 4.4.3 *For $n \ge 0$, set*

$$a_n(h,\alpha) := -\frac{4}{5}\sqrt{\alpha}\cdot\frac{(1+h\alpha)^n}{\sqrt{5 + (1+h\alpha)^{2n}}},$$

then we have that $x_n \in (\omega_{\varphi,-}, a_n)$ and $J^E(x_n) \in (\omega_{\Phi,-}, a_n)$.

Proof. Due to assumptions (4.18), $\max\left(\omega_{\varphi,-}, \omega_{\Phi,-}\right) < a_n$ for $n \ge 0$, so the intervals in the lemma are non-degenerate. We proceed by induction. $a_0 > x_0 \equiv J^E(x_0) \equiv -\sqrt{\frac{\alpha}{8}}$ is always satisfied.

So suppose that the statement is true for some $n \ge 0$. Condition $|x| < \frac{1}{3K}$ implies $\mathcal{N}_\varphi^E(x) < (1+h\alpha)x - \frac{4}{3}h\,x^3$, further, by monotonicity of $\mathcal{N}_\varphi^E$ we get that

$$x_{n+1} = \mathcal{N}_\varphi^E(x_n) < \mathcal{N}_\varphi^E(a_n) < (1+h\alpha)a_n - \frac{4}{3}h\,a_n^3, \tag{4.20}$$

thus it is enough to prove that

$$a_{n+1} - (1+h\alpha)a_n + \frac{4}{3}h\,a_n^3 > 0. \tag{4.21}$$

For brevity, we set $\lambda := h\alpha > 0$. Then (4.21) is equivalent to

$$\frac{4}{375}\left(-A+B-C\right)\sqrt{\alpha}\left(1+\lambda\right)^n > 0,$$

where $A := \frac{64\,\lambda\,(1+\lambda)^{2\,n}}{\left(5+(1+\lambda)^{2\,n}\right)^{\frac{3}{2}}}$, $B := \frac{75\,(1+\lambda)}{\sqrt{5+(1+\lambda)^{2\,n}}}$ and $C := \frac{75\,(1+\lambda)}{\sqrt{5+(1+\lambda)^{2+2n}}}$. We will show that

$$-A+B-C > 0.$$

First put $B-C$ over a common denominator. Then, to eliminate square roots from its numerator, multiply it by $\frac{\sqrt{5+(1+\lambda)^{2\,n}}+\sqrt{5+(1+\lambda)^{2+2\,n}}}{\sqrt{5+(1+\lambda)^{2\,n}}+\sqrt{5+(1+\lambda)^{2+2\,n}}}$. After these manipulations, the product $\lambda(1+\lambda)^{2n} > 0$ can be factored out from all three terms. Hence $-A+B-C>0$ becomes

$$\frac{-64}{\left(5+(1+\lambda)^{2\,n}\right)^{\frac{3}{2}}}+$$

$$\frac{75\;(1+\lambda)\;(2+\lambda)}{\sqrt{5+(1+\lambda)^{2\,n}}\,\sqrt{5+(1+\lambda)^{2+2\,n}}\,\left(\sqrt{5+(1+\lambda)^{2\,n}}+\sqrt{5+(1+\lambda)^{2+2\,n}}\right)} > 0.$$

The left hand side of the above inequality is decreased, if the denominator of the second term is replaced by $2\sqrt{5+(1+\lambda)^{2\,n}}\,\left(5+(1+\lambda)^{2+2\,n}\right)$. Then we can multiply the expression by $\sqrt{5+(1+\lambda)^{2\,n}}$ and all square roots are got rid of: we are to verify

$$\frac{-64}{5+(1+\lambda)^{2\,n}}+\frac{75\;(1+\lambda)\;(2+\lambda)}{2\,\left(5+(1+\lambda)^{2+2\,n}\right)} > 0.$$

By condition (4.18), $(1+\lambda)^2 < \frac{75}{64}$, so it is enough to show that

$$\frac{-64}{5+(1+\lambda)^{2\,n}}+\frac{75\;(1+\lambda)\;(2+\lambda)}{2\,\left(5+\frac{75\,(1+\lambda)^{2\,n}}{64}\right)} > 0.$$

However, the left hand side above can be factored to yield

$$\frac{32\,\left(22+225\,\lambda+75\,\lambda^2+45\,\lambda\,(1+\lambda)^{2\,n}+15\,\lambda^2\,(1+\lambda)^{2\,n}\right)}{\left(5+(1+\lambda)^{2\,n}\right)\,\left(64+15\,(1+\lambda)^{2\,n}\right)},$$

which is clearly positive.

The proof for the sequence $J^E(x_n)$ is the completely similar: by construction of J, the beginning of (4.20) should (and can) be replaced by $J^E(x_{n+1}) = \mathcal{N}_\Phi^E(J^E(x_n)) < \mathcal{N}_\Phi(a_n)$, but then everything is unchanged. ■

Remark 4.4.2 Attempts to approximate subexpressions of the form $(a+bt)^\gamma$ with their series expansions (up to third order) turned out to be insufficient to complete the proof. To find the above "purely algebraical" manipulations, *Mathematica* has been extensively used. The definition of a_n is again based on the beautiful parametrized model function of [29].

The sum $\sum_{i=0}^n$ in (4.19) is split into two at $\lceil \frac{6}{h\alpha} \rceil$. This choice is motivated by the following lemma.

Lemma 4.4.4 *Suppose that* $n > \lceil \frac{6}{h\alpha} \rceil$. *Then*

$$\max\left(x_n, J^E(x_n)\right) < -\sqrt{\frac{3}{5}\alpha},$$

hence

$$1 + h\alpha - \frac{5}{2}h\max\left(x_n, J^E(x_n)\right)^2 < 1 - \frac{h\alpha}{2}$$

holds for $n > \lceil \frac{6}{h\alpha} \rceil$.

Proof. By virtue of Lemma 4.4.3 it is enough to show that $n > \lceil \frac{6}{h\alpha} \rceil$ implies $a_n < -\sqrt{\frac{3}{5}\alpha}$. But this latter inequality is equivalent to $(1+h\alpha)^n > \sqrt{75}$. However, $e^3 > \sqrt{75}$, so the corresponding proof given in the transcritical case suffices here, too. ■

Let us turn directly to (4.19) now and fix $h \in (0, h_0]$, $\alpha \in (0, \alpha_0]$ and $n \in \mathbb{N}^+$ arbitrarily. (If $n \leq \lceil \frac{6}{h\alpha} \rceil$, then the sums $\sum_{i=\lceil \frac{6}{h\alpha} \rceil + 1}^{n}$ below are absent and the proof is simpler.) Using $\omega_{\varphi,-} < x_i < 0$, $|x_i| \leq \sqrt{2\alpha}$ (by Lemma 4.3.1), and $\max\left(x_j, J^E(x_j)\right) \leq x_0 \equiv J^E(x_0) \equiv -\sqrt{\frac{\alpha}{8}}$ (by monotonicity), further, Lemma 4.4.4, assumption $h\alpha < 1$ from (4.18) and inequality $(1+\frac{1}{A})^A \leq e$ (for $A \geq 1$), we see that

$$h\sum_{i=0}^{n} x_i^4 \prod_{j=i}^{n-1}\left(1 + h\alpha - \frac{5}{2}h\max\left(x_j, J^E(x_j)\right)^2\right) \leq$$

$$4h\alpha^2 \sum_{i=0}^{\lceil \frac{6}{h\alpha} \rceil} \prod_{j=1}^{\lceil \frac{6}{h\alpha} \rceil - 1}\left(1 + h\alpha - \frac{5}{2}\cdot\frac{h\alpha}{8}\right) + 4h\alpha^2 \sum_{i=\lceil \frac{6}{h\alpha} \rceil + 1}^{n} \prod_{j=i}^{n-1}\left(1 - \frac{h\alpha}{2}\right) \leq$$

$$4h\alpha^2\left(1 + \frac{11}{16}h\alpha\right)^{\frac{6}{h\alpha}}\left(\left\lceil \frac{6}{h\alpha} \right\rceil + 1\right) + 4h\alpha^2 \sum_{i=\lceil \frac{6}{h\alpha} \rceil + 1}^{n}\left(1 - \frac{h\alpha}{2}\right)^{n-i} \leq$$

$$4h\alpha^2\left(1 + \frac{11}{16}h\alpha\right)^{\frac{16}{11h\alpha}\cdot\frac{11h\alpha}{16}\cdot\frac{6}{h\alpha}}\left(\frac{6 + 2h\alpha}{h\alpha}\right) + 4h\alpha^2 \sum_{i=0}^{\infty}\left(1 - \frac{h\alpha}{2}\right)^i \leq$$

$$4h\alpha^2 \cdot e^{\frac{66}{16}} \cdot \frac{8}{h\alpha} + 4h\alpha^2 \cdot \frac{2}{h\alpha} \leq 1988\,\alpha.$$

Therefore, $\sup_{[x_{n+1}, x_n]} |\,id - J^E| \leq 1988c \cdot h^p\alpha$ for any $h \in (0, h_0]$, $\alpha \in (0, \alpha_0]$ and $n \geq 1$, further, as we have seen, $\sup_{[x_1, x_0]} |\,id - J^E| = \frac{c}{64}h^{p+1}\alpha^2$, which yield the following lemma.

Lemma 4.4.5 *Under assumption (4.18)*

$$\sup_{(\omega_{\varphi,-}, x_0]} |\,id - J^E| \leq 1988c \cdot h^p\alpha.$$

Now the closeness estimate is proved in the interval $(y_0, \omega_{\varphi,0})$. Recall that $y_0 = x_0 = J^E(x_0) \equiv -\sqrt{\frac{\alpha}{8}}$ and $\omega_{\varphi,0} = \omega_{\Phi,0} \equiv 0$.

Suppose first that $n \geq 1$. (The case $n = 0$ will be examined later.) Then we proceed exactly as in the transcritical case, so we will only list the differences. We get that

$$\sup_{[y_n, y_{n+1}]} |\, id - J^E| \leq$$

$$\left[\sup_{x \in [y_n, y_{n+1}]} \sup_{[\{\mathcal{N}_\Phi^E(x), J^E \circ \mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]} \right)' \right] \left[c \cdot h^{p+1} y_n^4 + \sup_{[y_{n-1}, y_n]} |\, id - J^E| \right].$$

The following lemma gives an upper bound on the first term above (and shows a motivation for the choice of $x_0 = -\sqrt{\frac{\alpha}{8}}$).

Lemma 4.4.6 *Suppose that $n \geq 1$, then under assumption (4.18) we have that*

$$\sup_{x \in [y_n, y_{n+1}]} \sup_{[\{\mathcal{N}_\Phi^E(x), J^E \circ \mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]} \right)' \leq 1 - \frac{h\alpha}{4}.$$

Proof. As in the transcritical case, we have that

$$\sup_{x \in [y_n, y_{n+1}]} \sup_{[\{\mathcal{N}_\Phi^E(x), J^E \circ \mathcal{N}_\varphi^E(x)\}]} \left((\mathcal{N}_\Phi^E)^{[-1]} \right)' \leq \sup_{(-\sqrt{\frac{\alpha}{8}}, 0)} \frac{1}{(\mathcal{N}_\Phi^E)'} \leq \ldots$$

But assumption (4.18) together with $x < 0$ imply that $(\mathcal{N}_\Phi^E)'(x) \geq 1 + h\alpha - 4hx^2 \geq 0$. So

$$\ldots \leq \sup_{x \in (-\sqrt{\frac{\alpha}{8}}, 0)} \frac{1}{1 + h\alpha - 4hx^2} \leq \frac{1}{1 + h\alpha - 4h\left(-\sqrt{\frac{\alpha}{8}}\right)^2} = \frac{1}{1 + \frac{1}{2}h\alpha} \leq 1 - \frac{h\alpha}{4}. \quad \blacksquare$$

We have thus proved (using $|y_n| \leq \sqrt{\frac{\alpha}{8}}$ also) that for $n \geq 1$

$$\sup_{[y_n, y_{n+1}]} |\, id - J^E| \leq \left(1 - \frac{h\alpha}{4}\right) \left[\frac{c}{64} \cdot h^{p+1} \alpha^2 + \sup_{[y_{n-1}, y_n]} |\, id - J^E| \right].$$

For $n = 0$, similarly as in the transcritical case, we get that

$$\sup_{[y_0, y_1]} |\, id - J^E| \leq 2 \cdot \frac{c}{64} h^{p+1} \alpha^2,$$

and for $n \geq 1$ that

$$\sup_{[y_n, y_{n+1}]} |\, id - J^E| \leq \left(1 - \frac{h\alpha}{4}\right)^n \sup_{[y_0, y_1]} |\, id - J^E| + \frac{c}{64} h^{p+1} \alpha^2 \sum_{i=1}^{n} \left(1 - \frac{h\alpha}{4}\right)^i \leq$$

$$1 \cdot 2 \cdot \frac{c}{64} h^{p+1} \alpha^2 + \frac{c}{64} h^{p+1} \alpha^2 \cdot \frac{4}{h\alpha} \leq \frac{c}{8} h^p \alpha,$$

using $h\alpha \leq 1$ by (4.18). Since the same upper estimate is valid for $n = 0$, too, we have proved the following lemma.

Lemma 4.4.7 *Under assumption (4.18)*

$$\sup_{(x_0, 0)} |\, id - J^E| \leq \frac{c}{8} h^p \alpha.$$

4.4.3 The outer region

In this section, we first prove an $\mathcal{O}(h^p)$ closeness-estimate in the interval $[z_0, \omega_{\varphi,-})$ for $\alpha > 0$. Then the closeness is proved on $[z_0, \omega_{\Phi,0}) \equiv [z_0, 0)$ for $\alpha \leq 0$.

We are well familiar with the inequalities below (*cf.* the transcritical case).

For $n \geq 1$ we have that

$$\sup_{[z_n, z_{n+1}]} |\, id - J^E| \leq c \cdot h^{p+1} z_0^4 \prod_{j=1}^{n} \left(1 + h\alpha - \frac{5}{2} h \max\left(z_j, J^E(z_j)\right)^2\right) +$$

$$c \cdot h^{p+1} \sum_{i=0}^{n-1} z_i^4 \prod_{j=i+2}^{n} \left(1 + h\alpha - \frac{5}{2} h \max\left(z_j, J^E(z_j)\right)^2\right), \tag{4.22}$$

where $\prod_{j=n+1}^{n}$ is, as always, 1, and

$$\sup_{[z_0, z_1]} |\, id - J^E| \leq c \cdot h^{p+1} z_0^4.$$

The lemma gives a lower estimate of the sequence z_n for $\alpha > 0$.

Lemma 4.4.8 *For $n \geq 0$, set*

$$b_n(h, \alpha) := -2\sqrt{\alpha} \cdot \frac{(1 + h\alpha)^n}{\sqrt{\alpha - 1 + (1 + h\alpha)^{2n}}},$$

then $b_n \leq \min\left(z_n, J^E(z_n)\right)$.

Proof. We prove by induction. $b_0 = -2 < z_0 = J^E(z_0)$ holds due to assumption (4.18). Suppose that the statement is true for some $n \geq 0$.

We have that $\mathcal{N}_\varphi^E(x) \geq (1 + h\alpha)x - \frac{3}{4} h\, x^3$ (by $x < 0$ and $|x| \leq \varepsilon_0 < \frac{1}{4K}$), further, that $(1 + h\alpha) id - \frac{3}{4} h\, id^3$ is monotone increasing (which is implied by, for example, if $|x| \leq \frac{2}{3\sqrt{h}}$, but $|b_n| \leq 2$ is easily seen, and due to $h \leq \frac{1}{10}$ we get $|b_n| \leq \frac{2}{3\sqrt{h}}$ also), so we obtain that

$$z_{n+1} = \mathcal{N}_\varphi^E(z_n) \geq (1 + h\alpha) z_n - \frac{3}{4} h\, z_n^3 \geq (1 + h\alpha) b_n - \frac{3}{4} h\, b_n^3, \tag{4.23}$$

thus it is sufficient to show that

$$(1 + h\alpha) b_n - \frac{3}{4} h\, b_n^3 \geq b_{n+1},$$

being the same as

$$0 \leq 2\sqrt{\alpha}\, (1 + \lambda)^n \left(\widetilde{A} - \widetilde{B} + \widetilde{C}\right),$$

with $\widetilde{A} := \frac{3\,\lambda\,(1+\lambda)^{2\,n}}{\left(-1+\alpha+(1+\lambda)^{2\,n}\right)^{\frac{3}{2}}}$, $\widetilde{B} := \frac{1+\lambda}{\sqrt{-1+\alpha+(1+\lambda)^{2\,n}}}$ and $\widetilde{C} := \frac{1+\lambda}{\sqrt{-1+\alpha+(1+\lambda)^{2+2\,n}}}$, further, with $\lambda := h\alpha > 0$.

Now proceeding just as in Lemma 4.4.3, we first get

$$0 \leq \frac{3}{\left(-1 + \alpha + (1 + \lambda)^{2\,n}\right)^{\frac{3}{2}}}$$

$$-\frac{\widetilde{B}\,(2+\lambda)}{\sqrt{-1+\alpha+(1+\lambda)^{2+2\,n}}\,\left(\sqrt{-1+\alpha+(1+\lambda)^{2\,n}}+\sqrt{-1+\alpha+(1+\lambda)^{2+2\,n}}\right)}$$

to verify. Then multiply the inequality by $\sqrt{-1+\alpha+(1+\lambda)^{2\,n}}$ to get

$$0\le\frac{3}{-1+\alpha+(1+\lambda)^{2\,n}}$$

$$-\frac{(1+\lambda)\,(2+\lambda)}{\sqrt{-1+\alpha+(1+\lambda)^{2+2\,n}}\,\left(\sqrt{-1+\alpha+(1+\lambda)^{2\,n}}+\sqrt{-1+\alpha+(1+\lambda)^{2+2\,n}}\right)}.$$

A sufficient condition for this is

$$0\le\frac{3}{-1+\alpha+(1+\lambda)^{2\,n}}-\frac{(1+\lambda)\,(2+\lambda)}{2\,\left(-1+\alpha+(1+\lambda)^{2\,n}\right)},$$

but the right hand side is equal to

$$\frac{(1-\lambda)\,(4+\lambda)}{2\,\left(-1+\alpha+(1+\lambda)^{2\,n}\right)},$$

which is positive, if $0<\lambda\equiv h\alpha<1$.

When $\mathcal{N}_\Phi$ and $J^E(z_n)$ are written instead of $\mathcal{N}_\varphi$ and z_n, respectively, the considerations above remain valid, implying $b_n\le J^E(z_n)$. ■

Remark 4.4.3 Notice the subtle difference between the chain of inequalities (4.20) and (4.23). Unlike $\mathcal{N}_\varphi^E(a_n)$, quantity $\mathcal{N}_\varphi^E(b_n)$ is not necessarily defined, since b_n may lie outside the domain of definition of $\mathcal{N}_\varphi^E$.

Now, since $z_j<\omega_{\varphi,-}<0$ and $J^E(z_j)<\omega_{\Phi,-}<0$, by Lemma 4.3.1 we get that the right-hand side of (4.22) is at most

$$c\cdot h^{p+1}z_0^4\prod_{j=1}^{n}\left(1-\frac{h\alpha}{2}\right)+c\cdot h^{p+1}\sum_{i=0}^{n-1}z_i^4\prod_{j=i+2}^{n}\left(1-\frac{h\alpha}{2}\right)\le$$

$$c\cdot h^{p+1}z_0^4+c\cdot h^{p+1}\sum_{i=0}^{n-1}z_i^4\left(1-\frac{h\alpha}{2}\right)^{n-1-i}.$$

We will show that $h\sum_{i=0}^{n}z_i^4\left(1-\frac{h\alpha}{2}\right)^{n-i}$ is uniformly bounded for any $n\ge0$, $0<h\le h_0$ and $0<\alpha\le\alpha_0$.

If $n\ge\lceil\frac{1}{h\alpha}\rceil$ and $i\ge n$, then it is easy to see that $(1+h\alpha)^i\ge(1+h\alpha)^{\frac{1}{h\alpha}}\ge1+\frac{1}{h\alpha}\cdot h\alpha=2$ implies $\frac{(1+h\alpha)^i}{\sqrt{\alpha-1+(1+h\alpha)^{2i}}}\le2$, hence by Lemma 4.4.8

$$h\sum_{i=\lceil\frac{1}{h\alpha}\rceil}^{n}z_i^4\left(1-\frac{h\alpha}{2}\right)^{n-i}\le h\sum_{i=\lceil\frac{1}{h\alpha}\rceil}^{n}b_i^4\left(1-\frac{h\alpha}{2}\right)^{n-i}\le$$

$$16h\alpha^2\sum_{i=\lceil\frac{1}{h\alpha}\rceil}^{n}\left(\frac{(1+h\alpha)^i}{\sqrt{\alpha-1+(1+h\alpha)^{2i}}}\right)^4\left(1-\frac{h\alpha}{2}\right)^{n-i}\le256h\alpha^2\sum_{i=0}^{\infty}\left(1-\frac{h\alpha}{2}\right)^i=512\alpha.$$

On the other hand, if $n < \lceil \frac{1}{h\alpha} \rceil$, then—by inequalities $e^{\frac{x}{2}} \leq 1+x$ ($x \in [0,1]$) and $1+x \leq e^x$ ($x \in \mathbb{R}$)—we have that

$$h\sum_{i=0}^{n} z_i^4 \left(1-\frac{h\alpha}{2}\right)^{n-i} \leq 16h\alpha^2 \sum_{i=0}^{n} \left(\frac{(1+h\alpha)^i}{\sqrt{\alpha-1+(1+h\alpha)^{2i}}}\right)^4 \left(1-\frac{h\alpha}{2}\right)^{n-i} \leq$$

$$16h\alpha^2 \sum_{i=0}^{n} \frac{e^{4h\alpha i}}{\left(\alpha-1+(1+h\alpha)^{2i}\right)^2} \cdot 1 \leq 874\, h\alpha^2 \sum_{i=0}^{n} \frac{1}{\left(\alpha-1+e^{h\alpha i}\right)^2}.$$

Set $g_{h,\alpha}(x) \equiv g(x) := \frac{h\alpha^2}{\left(\alpha-1+e^{h\alpha x}\right)^2}$, if $x \in [0,\infty)$. Notice that g is bounded at $x=0$, and strictly decreasing on $[0,\infty)$, because

$$g'(x) = \frac{-2h^2\alpha^3 e^{h\alpha x}}{\left(\alpha-1+e^{h\alpha x}\right)^3} < 0.$$

So, since $0<\alpha<1$, we see that

$$874\, h\alpha^2 \sum_{i=0}^{n} \frac{1}{\left(\alpha-1+e^{h\alpha i}\right)^2} = 874h + 874\sum_{i=1}^{n} g_{h,\alpha}(i) \leq 874h + 874\int_0^{\frac{1}{h\alpha}} g_{h,\alpha}(x)\mathrm{d}x =$$

$$874h + 874\left[\frac{h\alpha^2 x}{(\alpha-1)^2} + \frac{\alpha}{(\alpha-1)\ (\alpha-1+e^{h\alpha x})} - \frac{\alpha\ \ln(\alpha-1+e^{h\alpha x})}{(\alpha-1)^2}\right]_{x=0}^{\frac{1}{h\alpha}} =$$

$$874h + 874\left(\frac{e-1+\alpha^2+\alpha\ (e-1+\alpha)\ (\ln\alpha - \ln(e-1+\alpha))}{(\alpha-1)^2\ (e-1+\alpha)}\right) \leq$$

$$874h + 874 \cdot \frac{e-1+\alpha^2}{(\alpha-1)^2\ (e-1+\alpha)} \leq 874h + \frac{874}{(\alpha-1)^2} < 3584,$$

since $h \leq \frac{1}{10}$ and $\alpha \leq \frac{1}{2}$, so $\alpha^2 \leq \alpha$.

Now combining all the estimates so far in the section, under assumption (4.18) we get that if $\alpha > 0$, then

$$\sup_{[z_0,\omega_{\varphi,-})} |\,id - J^E| = \sup_{n\in\mathbb{N}} \sup_{[z_n,z_{n+1}]} |\,id - J^E| \leq$$

$$\sup_{n\in\mathbb{N}} \max\left(c\cdot h^{p+1} z_0^4,\ c\cdot h^{p+1} z_0^4 + c\cdot h^{p+1}\sum_{i=0}^{n} z_i^4\left(1-\frac{h\alpha}{2}\right)^{n-i}\right) \leq$$

$$c\cdot h^{p+1} z_0^4 + c\cdot h^p \cdot (3584 + 512\alpha) \leq 3841c\cdot h^p.$$

Finally, a closeness estimate on $[z_0, 0)$ for $\alpha \leq 0$ is proved. The proof of the next lemma is identical to its counterpart in the transcritical case.

Lemma 4.4.9 *Suppose that $\alpha \leq 0$ and assumption (4.18) hold. Then for any $0 < h \leq h_0$, $-\alpha_0 \leq \alpha \leq \beta \leq 0$ and $n \in \mathbb{N}$ we have that*

$$0 > z_n(\alpha) \geq z_n(\beta).$$

So $0 > z_n(\alpha) \geq z_n(0)$ holds for $\alpha \leq 0$ and it is enough to give a lower estimate of $z_n(0)$.

Lemma 4.4.10 *Under assumption (4.18), we have for $n \in \mathbb{N}$ that*

$$z_n(0) \ge z_0 \ge -\frac{1}{2K}$$

and for $n \ge \lfloor\frac{16K^2}{h}\rfloor$

$$z_n(0) \ge -\frac{2}{\sqrt{nh}}.$$

Proof. Monotonicity of the sequence $z_n(0)$ and (4.18) imply the first part. The nontrivial second inequality is proved by induction. Since $z_{\lfloor\frac{16K^2}{h}\rfloor}(0) \ge -\frac{1}{2K} = -\frac{2}{\sqrt{16K^2}} \ge -\frac{2}{\sqrt{h\lfloor\frac{16K^2}{h}\rfloor}}$, the induction can be started. So suppose that $n \ge \lfloor\frac{16K^2}{h}\rfloor$. The function $z \mapsto \mathcal{N}_\varphi(h,z,0)$ is increasing, so, by the induction hypothesis we have that $z_{n+1}(0) = \mathcal{N}_\varphi(h, z_n(0), 0) \ge \mathcal{N}_\varphi(h, -\frac{2}{\sqrt{nh}}, 0)$. It is enough to show that $\mathcal{N}_\varphi(h, -\frac{2}{\sqrt{nh}}, 0) \ge -\frac{2}{\sqrt{(n+1)h}}$. Multiplying this with $\sqrt{h}$ and rearranging, it suffices to prove that

$$\frac{4}{n\sqrt{n}} + \frac{8\widetilde{\eta}}{n^2\sqrt{h}} \ge \frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}}$$

holds, where is $\widetilde{\eta}$ from (4.16). But $\frac{1}{2n\sqrt{n}} \ge \frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}}$, so it is enough to verify $4 + \frac{8\widetilde{\eta}}{\sqrt{nh}} \ge \frac{1}{2}$. But it is easy to see that conditions $n \ge \lfloor\frac{16K^2}{h}\rfloor \ge \frac{16K^2}{h} - 1$ and $h \le 8K^2$ (from (4.18)) together with the definition of K imply $\left|\frac{8\widetilde{\eta}}{\sqrt{nh}}\right| \le 3$. ■

Then we can simply estimate (4.22) for $\alpha \le 0$ as follows. Supposing that $n \ge 1$ we get that

$$\sup_{[z_n, z_{n+1}]} |id - J^E| \le c \cdot h^{p+1} z_0^4 \prod_{j=1}^{n} \left(1 + h\alpha - \frac{5}{2} h \max\left(z_j, J^E(z_j)\right)^2\right) +$$

$$c \cdot h^{p+1} \sum_{i=0}^{n-1} z_i^4 \prod_{j=i+2}^{n} \left(1 + h\alpha - \frac{5}{2} h \max\left(z_j, J^E(z_j)\right)^2\right) \le$$

$$c \cdot h^{p+1} z_0^4 \cdot 1 + c \cdot h^p \cdot h \sum_{i=0}^{n} z_i(0)^4 \cdot 1 \le$$

$$c \cdot h^p \left(h z_0^4 + h \sum_{i=0}^{\lfloor\frac{16K^2}{h}\rfloor} z_i(0)^4 + h \sum_{i=\lfloor\frac{16K^2}{h}\rfloor+1}^{n} z_i(0)^4 \right),$$

where, of course, for $n \le \lfloor\frac{16K^2}{h}\rfloor$, the sum above $\sum_{i=\lfloor\frac{16K^2}{h}\rfloor+1}^{n}$ is not present. But

$$h \sum_{i=0}^{\lfloor\frac{16K^2}{h}\rfloor} z_i(0)^4 \le h \cdot \left(\frac{16K^2}{h} + 1\right) \left(-\frac{1}{2K}\right)^4 \le \frac{2}{K^2}$$

by $h \le 8K^2$, and

$$h \sum_{i=\lfloor\frac{16K^2}{h}\rfloor+1}^{n} z_i(0)^4 \le h \sum_{i=\lfloor\frac{16K^2}{h}\rfloor+1}^{n} \frac{16}{i^2 h^2} \le \frac{16}{h} \int_{\frac{16K^2}{h}}^{\infty} \frac{1}{i^2} = \frac{1}{K^2}.$$

We have thus proved that

$$\sup_{[z_0, 0)} |id - J^E| \le c \cdot h^{p+1} z_0^4 + c \cdot h^p \left(h z_0^4 + \frac{3}{K^2}\right) \le c\left(2 + \frac{3}{K^2}\right) h^p.$$

Chapter 5

Preservation of bifurcations under Runge-Kutta methods

Summary. In this chapter we show that conditions for the fold bifurcation and for the cusp bifurcation in N dimensions are completely preserved by Runge-Kutta methods. A similar one-dimensional result has already been proved in Lemma 4.2.1.

5.1 The fold bifurcation in N dimensions

Consider the ordinary differential equation

$$\dot{z} = f(z, \alpha) \tag{5.1}$$

depending on a parameter $\alpha \in \mathbb{R}$. Suppose that the smooth function $f : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$ has a fold bifurcation [3] at the equilibrium $z = 0$, $\alpha = 0$, that is the following conditions are satisfied:

- $f^B = 0$,
- dim null$(f_z^B) = 1$,
- $f_\alpha^B \notin$ ran(f_z^B),
- $f_{zz}^B(v, v) \notin$ ran(f_z^B), where null(f_z^B)=span(v),

where throughout the chapter null(A) and ran(A) denote the *null space* and the *range* of the linear operator A, respectively. The evaluation operator ^{B} evaluates functions at $z = 0$, $\alpha = 0$, and also at $h > 0$, when it applies. Finally, the $N \times N$ identity matrix is denoted by I_N.

Now consider a discretization $\varphi(h, z, \alpha)$ of the above equation, with the function $\varphi : \mathbb{R}^+ \times \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$ coming from an *s-stage Runge-Kutta method with step-size* $h > 0$, that is

$$Z_{n+1} := \varphi(h, Z_n, \alpha), \qquad n = 0, 1, 2, \ldots \tag{5.2}$$

where

$$\varphi(h, z, \alpha) \equiv z + h \sum_{i=1}^{s} \gamma_i \cdot k_i(h, z, \alpha),$$

and every function k_i $(i = 1, 2, \ldots, s)$ satisfies the (implicit) equation

$$k_i(h, z, \alpha) = f(z + h \sum_{j=1}^{s} \beta_{ij} \cdot k_j(h, z, \alpha), \alpha) \tag{5.3}$$

with some γ_i, β_{ij} $(i, j = 1, 2, \ldots, s)$ given real constants.

The origin $z = 0$, $\alpha = 0$ is a fold bifurcation point for the map $\varphi(h, \cdot, \cdot)$, if the following conditions hold:

- $\varphi^B \equiv \varphi(h, 0, 0) = 0$,
- $\dim \operatorname{null}(\varphi_z^B - I_N) = 1$,
- $\varphi_\alpha^B \notin \operatorname{ran}(\varphi_z^B - I_N)$,
- $\varphi_{zz}^B(v, v) \notin \operatorname{ran}(\varphi_z^B - I_N)$, where $\operatorname{null}(\varphi_z^B - I_N)$=span$(v)$.

Proposition 5.1.1 *Suppose that the equation (5.1) has a fold bifurcation at the equilibrium $z = 0$, $\alpha = 0$, and $\Gamma := \sum_{i=1}^{s} \gamma_i \neq 0$. Then the map (5.2) also has a fold bifurcation at $z = 0$, $\alpha = 0$ for $h > 0$ sufficiently small.*

Remark 5.1 It is well known that the condition $\Gamma = 1$ is necessary for a Runge-Kutta method to be of order at least one, hence the above assumption on Γ is natural.

Proof of the proposition. Step 1. The first, well-known property follows from the fact [24] that for h small enough, there is a locally unique solution to the defining system of equations (5.3) for the functions k_i, which is seen to be $k_i^B \equiv k_i(h, 0, 0) = f^B = 0$ for every $i = 1, 2, \ldots, s$.

Step 2. Next we show that $\operatorname{null}(f_z^B) \subset \operatorname{null}\big((k_i)_z^B\big)$ for all $i = 1, 2, \ldots, s$. To this end, choose $0 \neq v \in \operatorname{null}(f_z^B)$ and use (5.3) to obtain for every i that

$$(k_i)_z^B v = f_z^B \left(I_N + h \sum_{j=1}^{s} \beta_{ij} \cdot (k_j)_z^B \right) v = f_z^B h \sum_{j=1}^{s} \beta_{ij} \cdot (k_j)_z^B v,$$

that is

$$(k_i)_z^B v - h \sum_{j=1}^{s} \beta_{ij} \cdot f_z^B (k_j)_z^B v = 0, \qquad i = 1, 2, \ldots, s.$$

Notice that these s equations can be represented by a single matrix equation as

$$\big(I_{N \cdot s} - h \cdot \beta \otimes f_z^B\big) \begin{pmatrix} (k_1)_z^B v \\ (k_2)_z^B v \\ \vdots \\ (k_s)_z^B v \end{pmatrix} = 0 \in \mathbb{R}^{N \cdot s},$$

where we have used the Kronecker product $\otimes$ of the matrices $\beta := [\beta_{ij}] \in \mathbb{R}^{s \times s}$ and f_z^B. However, for small h, the matrix $I_{N \cdot s} - h \cdot \beta \otimes f_z^B$ is invertible, hence $(k_i)_z^B v = 0$ for every $i = 1, 2, \ldots, s$, and the assertion follows.

Step 3. The previous step also proves that $\text{null}(f_z^B) \subset \text{null}(\varphi_z^B - I_N)$, since for any $v \in \text{null}(f_z^B)$ we have that

$$(\varphi_z^B - I_N)v = \left(h \sum_{i=1}^{s} \gamma_i \cdot f_z^B \left(I_N + h \sum_{j=1}^{s} \beta_{ij} \cdot (k_j)_z^B \right) \right) v =$$

$$h\Gamma \cdot f_z^B v + h^2 \cdot f_z^B \sum_{i=1}^{s} \sum_{j=1}^{s} \gamma_i \beta_{ij} \cdot (k_j)_z^B v = 0.$$

Step 4. In order to prove that $\text{null}(\varphi_z^B - I_N)$ is in fact one dimensional, choose an arbitrary nonzero vector w from this subspace. A similar rearrangement as in the previous step shows that

$$0 = (\varphi_z^B - I_N)w = h \cdot f_z^B Aw,$$

where we have used the abbreviation

$$A \equiv \Gamma \cdot I_N + h \sum_{i=1}^{s} \sum_{j=1}^{s} \gamma_i \beta_{ij} \cdot (k_j)_z^B.$$

Therefore, $Aw \in \text{null}(f_z^B) \subset \text{null}\big((k_j)_z^B\big)$ for all $j = 1, 2, \ldots, s$, which implies that

$$AAw = \Gamma \cdot Aw + h \sum_{i=1}^{s} \sum_{j=1}^{s} \gamma_i \beta_{ij} \cdot (k_j)_z^B Aw = \Gamma \cdot Aw.$$

But A is invertible, because $\Gamma \neq 0$ and h is small, so we have

$$Aw = \Gamma \cdot w,$$

which shows that $w \in \text{null}(f_z^B)$, and also that $\text{null}(\varphi_z^B - I_N) = \text{null}(f_z^B)$.

Step 5. As for the first range condition $\varphi_\alpha^B \notin \text{ran}(\varphi_z^B - I_N)$, suppose to the contrary that there exists a vector $w \in \mathbb{R}^N$ such that $\varphi_\alpha^B = (\varphi_z^B - I_N)w$ holds. This is equivalent to saying that

$$h \sum_{i=1}^{s} \gamma_i \cdot \left(f_z^B \left(h \sum_{j=1}^{s} \beta_{ij} \cdot (k_j)_\alpha^B \right) + f_\alpha^B \right) = h \cdot f_z^B Aw,$$

which is just

$$f_\alpha^B = \frac{1}{\Gamma} \cdot f_z^B \left(Aw - h \sum_{i=1}^{s} \sum_{j=1}^{s} \gamma_i \beta_{ij} \cdot (k_j)_\alpha^B \right).$$

This means, however, that $f_\alpha^B \in \text{ran}(f_z^B)$, a contradiction.

Step 6. Finally, to prove the second range condition, one has to work with the bilinear forms representing the second derivatives. Suppose again, to the contrary, that there exists a vector $w \in \mathbb{R}^N$ such that $\varphi_{zz}^B(v, v) = (\varphi_z^B - I_N)w$, where $v \in \text{null}(f_z^B) = \text{null}(\varphi_z^B - I_N)$. Since

$$\varphi_{zz}^B(v, v) = h \sum_{i=1}^{s} \gamma_i \cdot (k_i)_{zz}^B(v, v),$$

we first need to compute $(k_i)_{zz}^B(v, v)$. To accomplish this, introduce the functions

$$F(z) \equiv f(z, \alpha)$$

and for any $i = 1, 2, \ldots, s$

$$G_i(z) \equiv z + h \sum_{j=1}^{s} \beta_{ij} \cdot k_j(h, z, \alpha).$$

Now $k_i = F \circ G_i$, so according to the higher-order chain rule [4], we get that

$$(k_i)_{zz}^B(v,v) = (F_{zz} \circ G_i)^B \left((G_i)_z^B v, (G_i)_z^B v\right) + (F_z \circ G_i)^B \left((G_i)_{zz}^B(v,v)\right) =$$

$$f_{zz}^B \left(v + h \sum_{j=1}^{s} \beta_{ij} \cdot (k_j)_z^B v, v + h \sum_{j=1}^{s} \beta_{ij} \cdot (k_j)_z^B v \right) + f_z^B \left((G_i)_{zz}^B(v,v)\right).$$

But $v \in \mathrm{null}(f_z^B) \subset \mathrm{null}\big((k_j)_z^B\big)$ for every j, hence

$$(k_i)_{zz}^B(v,v) = f_{zz}^B(v,v) + f_z^B \left((G_i)_{zz}^B(v,v)\right).$$

If $\varphi_{zz}^B(v,v) = (\varphi_z^B - I_N)w$ were true, then

$$\Gamma \cdot f_{zz}^B(v,v) + f_z^B \left(\sum_{i=1}^{s} \gamma_i \cdot (G_i)_{zz}^B(v,v) \right) = f_z^B A w$$

would hold, in other words

$$f_{zz}^B(v,v) = \frac{1}{\Gamma} \cdot f_z^B \left(Aw - \sum_{i=1}^{s} \gamma_i \cdot (G_i)_{zz}^B(v,v) \right),$$

which would clearly violate our original assumption $f_{zz}^B(v,v) \notin \mathrm{ran}(f_z^B)$. ■

5.2 The cusp bifurcation in N dimensions

As for the cusp case, consider (5.1) again, but this time with $\alpha \in \mathbb{R}^2$. The smooth function $f : \mathbb{R}^N \times \mathbb{R}^2 \to \mathbb{R}^N$ has a cusp bifurcation [3] at the equilibrium $z = 0$, $\alpha = 0$, if

- $f^B = 0$,
- $\dim \mathrm{null}(f_z^B) = 1$,
- $f_{zz}^B(v,v) \in \mathrm{ran}(f_z^B)$, where $\mathrm{null}(f_z^B)$=span(v),
- $f_{zzz}^B(v,v,v) + 3 f_{zz}^B(v,x) \notin \mathrm{ran}(f_z^B)$, where v is as above and x is any solution to the equation $f_z^B x = -f_{zz}^B(v,v)$.

Remark 5.2 One can make x unique assuming one extra condition, but we will not make use of this property.

Consider the corresponding Runge-Kutta discretization $\varphi : \mathbb{R}^+ \times \mathbb{R}^N \times \mathbb{R}^2 \to \mathbb{R}^N$. The equilibrium $z = 0$, $\alpha = 0$ is a cusp bifurcation point for the map $\varphi(h, \cdot, \cdot)$, if the following conditions hold:

- $\varphi(h, 0, 0) = 0$,

- $\dim \operatorname{null}(\varphi_z^B - I_N) = 1$,

- $\varphi_{zz}^B(v, v) \in \operatorname{ran}(\varphi_z^B - I_N)$, where $\operatorname{null}(\varphi_z^B - I_N)$=span$(v)$,

- $\varphi_{zzz}^B(v, v, v) + 3\varphi_{zz}^B(v, y) \notin \operatorname{ran}(\varphi_z^B - I_N)$, where v is as above and y is any solution to the equation $(\varphi_z^B - I_N)y = -\varphi_{zz}^B(v, v)$.

Proposition 5.2.1 *Suppose that the equation (5.1) has a cusp bifurcation at the equilibrium $z = 0$, $\alpha = 0$, and $\Gamma := \sum_{i=1}^s \gamma_i \neq 0$. Then the corresponding Runge-Kutta discretization map also has a cusp bifurcation at $z = 0$, $\alpha = 0$ for $h > 0$ sufficiently small.*

Proof. Due to Proposition 5.1.1, only the last two conditions have to be checked.
Step 1. Suppose that $f_{zz}^B(v, v) = f_z^B u$ holds with some $u \in \mathbb{R}^N$ and $0 \neq v \in \operatorname{null}(f_z^B)$. Set

$$w := A^{-1}\left(\Gamma u + \sum_{i=1}^s \gamma_i \cdot (G_i)_{zz}^B(v, v)\right),$$

where the linear operator A and the functions G_i $(i = 1, 2, \ldots, s)$ are as in the proof of Proposition 5.1.1, see Step 4 and 6 there. Then we have that

$$(\varphi_z^B - I_N)w = h \cdot f_z^B A w = h \cdot f_z^B \left(\Gamma u + \sum_{i=1}^s \gamma_i \cdot (G_i)_{zz}^B(v, v)\right) =$$

$$h\sum_{i=1}^s \gamma_i \cdot f_z^B u + h\sum_{i=1}^s \gamma_i \cdot f_z^B \left((G_i)_{zz}^B(v, v)\right) =$$

$$h\sum_{i=1}^s \gamma_i \left(f_{zz}^B(v, v) + f_z^B\left((G_i)_{zz}^B(v, v)\right)\right) = h\sum_{i=1}^s \gamma_i \cdot (k_i)_{zz}^B(v, v) = \varphi_{zz}^B(v, v).$$

Step 2. Suppose to the contrary that there exists a vector $w \in \mathbb{R}^N$ such that

$$\varphi_{zzz}^B(v, v, v) + 3\varphi_{zz}^B(v, y) = (\varphi_z^B - I_N)w \tag{5.4}$$

holds with $0 \neq v \in \operatorname{null}(\varphi_z^B - I_N) = \operatorname{null}(f_z^B)$ and y being any solution to the equation $(\varphi_z^B - I_N)y = -\varphi_{zz}^B(v, v)$.

In order to compute the trilinear and the bilinear forms here, we appeal again to the higher-order chain rule [4] (with the same notation as in Step 6 in the proof of Proposition 5.1.1) to get for every $i = 1, 2, \ldots, s$ that

$$(k_i)_{zzz}^B(v, v, v) = (F_{zzz} \circ G_i)^B \left((G_i)_z^B v, (G_i)_z^B v, (G_i)_z^B v\right) +$$

$$3\left(F_{zz} \circ G_i\right)^B \left((G_i)_{zz}^B(v, v), (G_i)_z^B v\right) + (F_z \circ G_i)^B \left((G_i)_{zzz}^B(v, v, v)\right),$$

where symmetry of the bilinear forms has also been taken into account. Performing some of the evaluations, we arrive at the following formula

$$(k_i)_{zzz}^B(v, v, v) = f_{zzz}^B(v, v, v) + 3f_{zz}^B\left((G_i)_{zz}^B(v, v), v\right) + f_z^B\left((G_i)_{zzz}^B(v, v, v)\right).$$

In a similar manner, we have that

$$(k_i)_{zz}^B(v, y) = f_{zz}^B\left(v, y + h\sum_{j=1}^s \beta_{ij} \cdot (k_j)_z^B y\right) + f_z^B\left((G_i)_{zz}^B(v, y)\right).$$

Now (5.4) is equivalent to the following

$$h\sum_{i=1}^{s}\gamma_i\{f^B_{zzz}(v,v,v)+3f^B_{zz}\left((G_i)^B_{zz}(v,v),v\right)+f^B_z\left((G_i)^B_{zzz}(v,v,v)\right)+$$

$$3f^B_{zz}\left(v,y+h\sum_{j=1}^{s}\beta_{ij}\cdot(k_j)^B_z y\right)+3f^B_z\left((G_i)^B_{zz}(v,y)\right)\}=h\cdot f^B_z Aw,$$

that is

$$f^B_{zzz}(v,v,v)+3f^B_{zz}\left(v,\frac{1}{\Gamma}\left(\sum_{i=1}^{s}\gamma_i\cdot(G_i)^B_{zz}(v,v)+\Gamma y+h\sum_{i=1}^{s}\sum_{j=1}^{s}\gamma_i\beta_{ij}\cdot(k_i)^B_z y\right)\right)=$$

$$\frac{1}{\Gamma}\cdot f^B_z\left(Aw-\sum_{i=1}^{s}\gamma_i\cdot(G_i)^B_{zzz}(v,v,v)-3\sum_{i=1}^{s}\gamma_i\cdot(G_i)^B_{zz}(v,y)\right),$$

using again the symmetry of the bilinear forms.

The desired contradiction will immediately follow as soon as we have shown that

$$x:=\frac{1}{\Gamma}\left(\sum_{i=1}^{s}\gamma_i\cdot(G_i)^B_{zz}(v,v)+\Gamma y+h\sum_{i=1}^{s}\sum_{j=1}^{s}\gamma_i\beta_{ij}\cdot(k_i)^B_z y\right)$$

solves

$$f^B_z x=-f^B_{zz}(v,v).$$

But we know that y satisfies

$$(\varphi^B_z-I_N)(-y)=\varphi^B_{zz}(v,v),$$

which—by the last part of Step 6 in the proof of Proposition 5.1.1—implies that

$$f^B_{zz}(v,v)=\frac{1}{\Gamma}\cdot f^B_z\left(A(-y)-\sum_{i=1}^{s}\gamma_i\cdot(G_i)^B_{zz}(v,v)\right).$$

By the definition of x and A, the right hand side is just $f^B_z(-x)$, so the proof is complete. ■

Bibliography

[1] D. K. Arrowsmith, C. M. Place, *An introduction to dynamical systems*, Cambridge University Press, 1990.

[2] G. Belitskii, V. Tkachenko, *One-dimensional functional equations*, Birkhäuser, Basel, 2003.

[3] W.-J. Beyn, A. Champneys, E. Doedel, W. Govaerts, Y. A. Kuznetsov, B. Sandstede, *Numerical continuation, and computation of normal forms*, In: Handbook of Dynamical Systems (Ed. by B. Fiedler), Vol. 2, Elsevier, 2002.

[4] W.-J. Beyn, W. Kless, *Numerical Taylor expansions of invariant manifolds in large dynamical systems*, Numer. Math. **80**, No. 1 (1998) 1–38.

[5] N. G. de Bruijn, *Asymptotic Methods in Analysis* (Chapter *Iterated Functions*), North Holland, Amsterdam, 1961.

[6] K. A. Cliffe, A. Spence, S. J. Tavener, *The numerical analysis of bifurcation problems with application to fluid mechanics*, Acta Numerica **9** (2000) 39–131.

[7] G. Farkas, *Conjugacy in the discretized fold bifurcation*, Computers and Mathematics with Applications **43** (2002) 1027–1033.

[8] L. Z. Fishman, *Stability conditions for a fixed point of point maps in the critical case of a pair of complex conjugate roots on the unit circle*, Math. Notes **52**, No. 6 (1992) 1265–1271.

[9] L. Z. Fishman, *Preservation of stability and bifurcations in discretization of nonlinear differential equations*, Differ. Equations **31**, No. 4 (1995) 569–576.

[10] L. Z. Fishman, *On the preservation of properties of differential equations under discretization*, Dokl. Math. **53**, No. 1 (1996) 73–75.

[11] L. Z. Fishman, *Preservation of equilibrium states and their stability for discrete Runge-Kutta approximations of continuous systems*, Math. Notes **59**, No. 5 (1996) 568–571.

[12] L. Z. Fishman, *On conservation of stability and bifurcations upon discretization*, Autom. Remote Control **57**, No. 9 (1996) 1311–1315.

[13] L. Z. Fishman, *Preservation of the properties of continuous systems under discretization by the Runge-Kutta and Adams methods*, Autom. Remote Control **58**, No. 10 (1997) 1640–1646.

[14] L. Z. Fishman, *Properties of differential and approximate finite-difference equations*, Differ. Equ. **36**, No. 3 (2000) 403–408.

[15] L. Z. Fishman, *Preservation of stability of differential equations under discretization*, Differ. Equ. **39**, No. 4 (2003) 607–608.

[16] B. M. Garay, *Discretization and some qualitative properties of ordinary differential equations about equilibria*, Acta Math. Univ. Comen., New Ser. **62**, No. 2 (1993) 249–275.

[17] B. M. Garay, *On stuctural stability of ordinary differential equations with respect to discretization methods*, Numer. Math. **72**, No. 4 (1996) 449–479.

[18] B. M. Garay, *On C^j-closeness between the solution flow and its numerical approximation*, J. Difference Equ. Appl. **2**, No. 1 (1996) 67–86.

[19] B. M. Garay, L. Lóczi, *Monotone Delay Equations and Runge-Kutta Discretizations*, Special Issue of Functional Differential Equations **11**, No. 1–2 (2004) 59–67.

[20] B. M. Garay, *A brief survey on the numerical dynamics for functional differential equations*, Int. J. Bifurcation Chaos Appl. Sci. Eng. **15**, No. 3 (2005) 729–742.

[21] P. Glendinning, *Stability, Instability and Chaos*, Cambridge Univ. Press, Cambridge, 1994.

[22] W. J. F. Govaerts, *Numerical methods for bifurcations of dynamical equilibria*, SIAM Philadelphia, 2000.

[23] D. F. Griffiths, P. K. Sweby, H. C. Yee, *On spurious asymptotic numerical solutions of explicit Runge-Kutta methods*, IMA J. Numer. Anal. **12**, No. 3 (1992) 319–338.

[24] E. Hairer, S.P. Nørsett, G. Wanner, *Solving Ordinary Differential Equations I.*, 2nd Edition, Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[25] J. E. Hirsch, B. A. Huberman, D. J. Scalapino, *Theory of intermittency*, Physical Review A **25**, No. 1 (1982) 519–532.

[26] J. Hofbauer, G. Iooss, *A Hopf bifurcation theorem for difference equations approximating a differential equation*, Monatsh. Math. **98**, No. 2 (1984) 99–113.

[27] T. Hüls, Y. Zou, *Polynomial estimates and discrete saddle-node homoclinic orbits*, J. Math. Anal. Appl. **256**, No. 1 (2001) 115–126.

[28] T. Hüls, *Numerische Approximation nicht-hyperbolischer heterokliner Orbits*, PhD Thesis, University of Bielefeld, 2002.

[29] T. Hüls, *A model function for polynomial rates in discrete dynamical systems*, Appl. Math. Lett. **17**, No. 1 (2004) 1–5.

[30] M. C. Li, *Structural stability of Morse-Smale gradient-like flows under discretizations*, SIAM J. Math. Anal. **28**, No. 2 (1997) 381–388.

[31] M. C. Li, *Structural stability of flows under numerics*, J. Differ. Equations **141**, No. 1 (1997) 1–12.

[32] M. C. Li, *Stability of a saddle node bifurcation under numerical approximations*, Computers and Mathematics with Applications **49**, No. 11–12 (2005) 1849–1852.

[33] A. Klebanoff, *π in the Mandelbrot set*, Fractals **9**, No. 4 (2001) 393–402.

[34] M. Kuczma, *Functional equations in a single variable*, Monografie matematyczne, Tom. 46, PWN – Polish Scientific Publishers, Warszawa, 1968.

[35] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 2004.

[36] W. de Melo, S. van Strien, *One-dimensional dynamics*, Springer-Verlag, 1993.

[37] G. W. Reddien, *On the stability of numerical methods of Hopf points using backward error analysis*, Computing **55**, No. 2 (1995) 163–180.

[38] C. Robinson, *Dynamical systems*, Stability, symbolic dynamics, and chaos. Second Edition, CRC Press, Boca Raton, 1999.

[39] S. Saks, A. Zygmund, *Analytic functions*, Monografie Matematyczne, Tom XXVIII. Polskie Towarzystwo Matematyczne, Warszawa-Wrocław, 1952.

[40] J. Sotomayor, *Generic bifurcations of dynamical systems*, Dynamical Systems, Ed. by M. M. Peixoto (Proc. of a Symp. at Univ. Bahia, 1971), Academic Press, New York, 1973, 561–582.

[41] J. Sotomayor, *Generic one-parameter families of vector fields on two-dimensional manifolds*, Publ. Math. Inst. Hautes Etudes Sci. **43** (1974) 5–46.

[42] O. Stein, *Bifurcations of hyperbolic fixed points for explicit Runge-Kutta methods*, IMA J. Numer. Anal. **17**, No. 2 (1997) 151–175.

[43] A. M. Stuart, A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, 1998.

[44] G. Szekeres, *Regular iteration of real and complex functions*, Acta Math. **100** (1958) 203–258.

[45] X. Wang, E. K. Blum, Q. Li, *Consistency of local dynamics and bifurcation of continuous-time dynamical systems and their numerical discretizations*, J. Difference Equ. Appl. **4**, No. 1 (1998) 29–57.

[46] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 2003.

[47] Y.-K. Zou, W.-J. Beyn, *On manifolds of connecting orbits in discretizations of dynamical systems*, Nonlinear Analysis TMA **52**, No. 5 (A) (2003) 1499–1520.