Gevrey and analytic local models and normal forms for diffeomorphisms and vector fields

F. Verstringe

May 26, 2011

Acknowledgements

Many people have helped me either directly or indirectly in order to make it possible to finish this text. Before they are forgotten, I would like to begin by thanking everybody who did so and is not explicitly listed below.

First of all I would to express my gratitude to my supervisor, P. Bonckaert, who introduced and guided me through the subject of this thesis. He was always prepared to read and correct my texts that contained a lot of errors and encouraged me to give more precision to the parts that were difficult to read. The help of my co-supervisor, P. De Maesschalck, has also been invaluable: I learned a lot from him through various discussions on a variety of mathematical subjects.

Contents

Nederlandse samenvatting

In deze doctoraatsthesis bestuderen we het lokaal dynamisch gedrag van analytische diffeomorfismen (lokaal inverteerbare functies) en analytische vectorvelden in de omgeving van een fixpunt. Na het eventueel toepassen van een verschuiving, mogen we aannemen dat dergelijk fixpunt zich bevindt in de oorsprong van ons gekozen assenstelsel. Laten we ons bijvoorbeeld focussen op analytische diffeomorfismen $F(x)$. In een omgeving rond de oorsprong hebben dergelijke functies een convergente Taylorreeksuitdrukking waarbij de constante term verdwijnt. De functie wordt aldus gegeven door $F(x) = Ax + f_2(x) + f_3(x) + \ldots$, waarbij $f_k(x)$ een homogene polynoom van graad k is en $Ax = Df(0)x$. Aangezien voor reële of complexe vectoren x die voldoende dicht bij de oorsprong liggen $|x|^2$ veel kleiner is dan $|x|$, hoeft het ons niet te verbazen dat het dynamisch gedrag, t.t.z. is de studie van de rij $x, F(x), F(F(x)), F(F(F(x))), \ldots$, in vele gevallen in hoofdzaak bepaald wordt door het lineair deel Ax, voor zover men voldoende dicht bij de oorsprong blijft. Het lijkt dan ook logisch om dergelijk dynamisch systeem te vergelijken met het dynamisch systeem gevormd door het lineair deel $L(x) = Ax$, en een analytische coördinatentransformatie $U(x) = x + u(x)$ te bepalen die $L(x)$ in verband brengt met $F(x)$. Heel expliciet bedoelen we dat $U^{-1} \circ F \circ U(x) = L(x)$. Dergelijke coördinatentransformatie bestaat helaas niet altijd; we leggen twee obstructies uit. Ten eerste is het niet altijd mogelijk om een formele uitdrukking te bepalen voor de transformatie $U_k(x) = x + u_2(x) + u_3(x) + \ldots + u_k(x)$ waarbij $u_k(x)$ een homogene polynoom van graad k is die op een inductieve manier gekozen wordt zodat $U_k^{-1} \circ F \circ U_k(x) - L(x) = 0$. De reden hiervoor is dat het lineair deel A van de functie F soms eigenwaarden bezit die aan bepaalde numerieke relaties voldoen en die ervoor zorgen dat bepaalde termen niet kunnen worden weggewerkt. We noemen dit fenomeen resonantie. Ten tweede is het mogelijk dat dergelijke formele uitdrukking wel bestaat, maar dat de gevonden transformatie een reeks vormt waarvan de oneindige som niet convergent is voor bijna elke waarde van x. De reden hiervoor is dat bepaalde universele uitdrukkingen in termen van de eigenwaarden zodanig groot kunnen zijn dat aan de oneindige som $U(x) = x + u_2(x) + u_3(x) + \ldots$ geen betekenis meer kan gegeven worden, zelfs niet voor waarden van x waarvoor $|x|$ relatief klein is. We noemen dit fenomeen quasi-resonantie. In de literatuur wordt uitgelegd dat het fenomeen resonantie en quasi-resonantie 'bijna nooit' optreedt, waarbij 'bijna nooit' gekarakteriseerd wordt in termen van de eigenwaarden van het lineair deel A. Het is verleidelijk om te denken dat de randgevallen waar resonantie of quasi-resonantie optreedt min of meer verwaarloosbaar zijn en we dus een bijna compleet overzicht hebben. Niets is minder waar! Voor een vast, bijvoorbeeld diagonaal, lineair deel met eigenwaarden $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ bestaan er op een willekeurig dichte afstand altijd eigenwaarden $(\mu_1, \mu_2, \ldots, \mu_n)$ waarvoor wel resonantie of quasi-resonantie optreedt. In wiskundige termen zeggen we dat de eigenwaarden waarvoor resonantie en/of quasi-resonantie optreedt dicht liggen.

In veel gevallen is men geïnteresseerd in de verandering van een gegeven dynamisch systeem onder de invloed van een parameter. Onderstel bijvoorbeeld dat $F_\varepsilon(x) = A_\varepsilon x + f_\varepsilon(x)$ een dergelijk analytisch systeem is en we geïnteresseerd zijn in het dynamisch gedrag van deze functie voor waarden van de parameter $\varepsilon \approx \varepsilon_0$. De opmerking omtrent de dichtheid van de resonante en quasi-resonante eigenwaarden suggereert dat voor waarden van ε in een dicht deel D in een omgeving van de parameter ε_0 het overeenkomstig lineair deel A_ε een verzameling van eigenwaarden heeft die quasi-resonant en/of resonant is. Het valt dus te verwachten dat de hierboven beschreven linearizatieprocedure een 'erg discontinu' gedrag vertoont als functie van de parameter.

In Hoofdstuk 2 van deze tekst stellen we dan ook een partiële oplossing voor met betrekking tot dit probleem. In plaats van het systeem te linearizeren, zullen we het systeem reduceren met behulp van een coördinatentransformatie naar een eenvoudigere vorm genaamd normaalvorm. Eenvoudiger dient begrepen te worden in termen van de Taylorreeks van de uiteindelijk gevonden gereduceerde vorm. Het gereduceerde parameterafhankelijke systeem noemen we een 'normaalvorm in een kegel', genaamd naar de kegelstructuur die verscholen zit in de Taylorreeks van het gereduceerde systeem, die een cruciale rol speelt in het bewijs. Vooruitlopend op de exacte formulering kunnen we het hoofdresultaat van Hoofdstuk 2 schetsen als

Stelling *Voor een gegeven parameterafhankelijk diffeomorfisme $F_\varepsilon(x) = A_\varepsilon x + f_\varepsilon(x)$ bestaat er een coördinatentransformatie van de vorm $U_\varepsilon(x) = x + u_\varepsilon(x)$, zodanig dat het gereduceerde systeem $G_\varepsilon(x) = U_\varepsilon^{-1} \circ F_\varepsilon \circ U_\varepsilon(x) = Ax + g_\varepsilon(x)$ eenvoudiger is in de zin dat de Taylorreeks van G_ε enkel niet-nul termen bevat van de vorm $ax_1^{k_1} \ldots x_n^{k_n}$ met overeenkomstige indices $k_1, \ldots, k_n$ die in een deelverzameling K van $\mathbb{N}^n$ liggen die een kegel-structuur heeft. De richting en opening van de kegel zijn gerelateerd aan de eigenwaarden van het niet-lineair deel en aan de variatie van de eigenwaarden als functie van de parameter.*

Als gevolg van dit hoofdresultaat kunnen we in een tweedimensionale context heel expliciete uitdrukkingen geven voor de gevonden normaalvormen; ik verwijs voor de geïnteresseerde lezer door naar het einde van Hoofdstuk 2. Gelijkaardige resultaten bestonden reeds voor vectorvelden, zie bijvoorbeeld [3]. In Hoofdstuk 3 worden deze resultaten in een tweedimensionale context herhaald met een specifiek doel voor ogen. Zoals reeds uitgelegd, kunnen we ook voor vectorvelden het concept 'Normaalvorm in een kegel' definiëren. De richting van deze kegel is bepaald door de twee eigenwaarden (λ_1, λ_2) van het lineair deel van het vectorveld, en de opening van deze kegel kan willekeurig klein gekozen worden; en de vraag i.v.m. convergentie dringt zich op wanneer we de kegel sluiten met behulp van een limietprocedure. We noemen een tweedimensionaal vectorveld waarvan het lineair deel een negatieve en een positieve eigenwaarde heeft, een zadel. We onderstellen vanaf nu dat het lineair deel niet afhangt van een externe parameter. Zoals reeds eerder uitgelegd, bestaan er getaltheoretische condities op de eigenwaarden die moeten voldaan zijn opdat linearizatie mogelijk is. In twee dimensies werd deze conditie o.a. door [7] herschreven in termen van kettingbreuken van de verhouding van de eigenwaarden $-\frac{\lambda_1}{\lambda_2}$. We gebruiken expliciet deze kettingbreuken om de limietprocedure waarbij de kegel gesloten wordt te voltooien en aldus linearizatie te bekomen van het gegeven systeem. De uitwerking hiervan is —ook in wiskundige termen— zeer technisch en iedere stap in de limietprocedure vereist erg precieze afschattingen. Gelukkig kunnen we een aantal van de gemaakte afschattingen recupereren door gebruik te maken van een zogenaamde renormalisatieprocedure, die in ons geval essentieel bestaat uit een herschalingstransformatie.

Hoofdstukken 4–5–6 zijn wiskundig technisch van aard. Ik zal in deze korte samenvatting in vage bewoordingen uitleggen wat de hoofdideeën zijn, en het valt aan te raden de bijhorende Engelse inleiding van de desbetreffende hoofdstukken te lezen voor preciezere formuleringen.

In Hoofdstuk 4 onderzoeken we analytische vectorvelden met een nilpotent lineair deel. Het lineair deel van deze vectorvelden heeft enkel eigenwaarden die 0 zijn, en de linearizatietechniek hierboven beschreven dient te worden aangepast aangezien er steeds resonantie optreedt. Onder andere in het artikel [37] wordt resonantie in een veel ruimere context geplaatst, en een ‘algemenere normaalvormprocedure’ omschreven in het desbetreffende artikel leert ons, dat onder bepaalde getaltheoretische voorwaarden op de ‘veralgemeende resonantierelaties’ we mogen besluiten dat er een bovengrens bestaat op de groei van de coefficienten van de gevonden ‘veralgemeende normaalvorm’. Deze voorwaarden werden in [34] expliciet berekend door directe berekening voor twee- en driedimensionale nilpotente lineaire delen. In hoofdstuk 4 worden ze uitgewerkt voor iedere dimensie door een indirecte methode die steunt op de representatietheorie van Lie-algebra’s. Dit lijkt op het eerste zicht misschien abstract, maar in essentie betreft het de berekening van de eigenwaarden van een aftelbaar aantal gegeven matrices $(M_i)_{i\in\mathbb{N}}$, waarbij M_i willekeurig groot wordt in dimensie naarmate de index i stijgt. Dergelijk probleem oplossen is hopeloos indien de opgegeven matrices niet aan bepaalde extra eigenschappen voldoen. Deze extra eigenschappen worden in dit geval beschreven met representatietheorie.

In Hoofdstuk 5 werken we in eerste instantie een ‘algemenere normaalvormprocedure’ uit voor diffeomorfismen (van quasi-graad 0). Daarna bepalen we voor diffeomorfismen met een diagonaal lineair deel een bovengrens voor de groei van de coefficienten van de overgangstransformatie en de normaalvorm indien we alle niet-resonante termen wegwerken. De hoop is uiteraard dat de overgangstransformatie en de gevonden normaalvorm convergeren voor voldoende kleine waarden van $|x|$. We vinden dat een bovengrens van de k-de term in de Taylorreeks van de overgangstransformatie en de normaalvorm van de vorm $k!R^k$. Reeksen waarvan de k-de term aan deze afschatting voldoet, noemen we Gevrey-reeksen. Ze zijn speciaal omdat voor dergelijke reeksen in combinatie met het gegeven vectorveld of diffeomorfisme soms betekenis kan worden gegeven aan de divergente reeks door een speciale sommatieprocedure.

In Hoofdstuk 6 wordt uitgelegd dat voor driedimensionale zadels met een niet-diagonalizeerbaar lineair deel de klassieke linearizatieprocedure bijna altijd divergent is. We leggen uit dat linearizatie nooit binnen eenzelfde Gevrey-klasse kan gebeuren, hetgeen een uitbreiding is van de resultaten omschreven in [30] en [49].

Chapter 1

Introduction

The theme of this dissertation is situated in the study of dynamical systems. Dynamical systems are a very broad and general subject in mathematics and are used as a tool to solve problems in many sciences as physics, economics, biology, chemistry, ... In all these applications one is interested in the quantitative or qualitative behaviour of one or more variables that undergo a dynamical process. This process may be continuous or with discrete steps, and we focus in this dissertation on deterministic dynamical systems, i.e. the influence of noise and stochastic distortions is not considered. We will deal with two types of dynamical systems, *vector fields* and *diffeomorphisms* (on $\mathbb{C}^n$).

A vector field X is basically a function defined on an open set $U \subset \mathbb{C}^n$ with range $\mathbb{C}^n$. The associated dynamics, the flow of the vector field is the collection of solutions of the initial value problem $\dot{\phi}(t) = X \circ \phi(t)$, $\phi(0) = x_0$. It consists of paths that have a tangent vector that coincides with the vector field in each point of the path. Using this interpretation, a vector field is merely a section of the tangent bundle. It is this geometric interpretation that is commonly used in the definition of vector fields on manifolds. Because we treat only local problems in this text, there is no need to give a precise definition of a vector field on manifolds.

A vector field can be time independent: the vector field is at a fixed position independent of the time or time dependent: the vector field itself changes at a fixed position while the time evolves. We will consider only vector fields that do not evolve in time. Although, mathematically spoken, a time-dependent vector field can be interpreted as a time independent one if one introduces an extra dimension. The starting point is usually a smooth vector field. Here 'smooth' can mean analytic, C^∞, C^k, $(k < \infty)$, or sometimes Gevrey. These conditions are not too restrictive in most problems but they simplify the situation in a lot of cases, see e.g. [21] for an introduction to what can happen if one loses smoothness in some points.

Locally, in a small neighbourhood of a certain point p, the study of a vector field

X is not too difficult for most of the points. Indeed, if at a certain point p the vector field is non-zero, it can be shown that there exists a coordinate transformation, that has the same smoothness as X, that is defined on some neighbourhood of that point p, and that conjugates the flow with a straight flow. In a neighbourhood of a point where the vector field is zero, the flow has richer dynamics. A well-known theorem of Hartman [28] and Grobmann [27] tells us that if the linear part of a C^1 vector field has no eigenvalues with zero real part, then there exists a continuous coordinate transformation that linearizes the flow. Hence the study of the flow of a vector field around such a singular point is equivalent to study of the flow of a linear vector field up to a continuous coordinate transformation. This allows to distract a lot of qualitative information for the local situation. This result, that can be applied to very general situations, is however not always satisfactory: in order to study more global phenomena, one sometimes needs more regular or more precise coordinate transformations. In this text we will mainly focus on the local study of analytic vector fields in a small neighbourhood of its singular points, by means of analytic coordinate transformations.

The second type of dynamical systems we discuss, diffeomorphisms, are essentially invertible maps from an open subset $V_1 \subset \mathbb{C}^n$ to another open subset $V_2 \subset \mathbb{C}^n$, having a certain regularity. From the dynamical point of view, one wants to analyze what happens if one starts to iterate the diffeomorphisms, i.e. one is interested in the evolution of the series x, $f(x)$, $f(f(x))$, $\ldots$; provided such series is well-defined. Such systems are seen a lot in practical situations for problems that are of a more discrete nature. For example the description of interacting populations (the so called Lotka-Volterra equations, see [55] for an overview in two dimensions), the yearly based interest calculations used by financial institutions, and so on. The concept is also useful in mathematical applications; for example the Poincaré map of a periodic orbit of a vector field and the time one map of a vector field are interesting diffeomorphisms to study.

For diffeomorphisms it is again interesting to study the behaviour in the neighbourhood of a fixed point. This is a point for which $F(p) = p$. Up to a translation, we may suppose that such a point is situated at the origin. As for vector fields, the main tool we use is a coordinate transformation and we will study the problem mainly in the analytic class. A coordinate transformation U, that is locally defined near the fixed point, reduces the local study near this point of the initial diffeomorphism to the study of the conjugation $G = U^{-1} \circ F \circ U$. In 'a lot of cases' the transformation U can be chosen in such a way that G becomes an invertible linear function, we briefly say that U linearizes F. We will quantify this later. We explain the obstructions to linearization in Section 1.1 and Section 1.2. The dynamics of linear functions are well-understood. Hence they provide a good model for the local situation.

1.1 Resonances, a formal obstruction to linearization

In this section we explain why it is sometimes impossible to find a formal power series solution U that linearizes a given analytic function F. Let us explain this problem in more detail for a diffeomorphism $F(x) = Ax + f(x)$ of $\mathbb{C}^n$, where A is an invertible linear mapping, and $f(x) = O(||x||^2)$. As is commonly done in literature (see e.g. [56],[15]), a suitable formal coordinate transformation

$$U(x) = x + u(x) = x + u_2(x) + u_3(x) + \ldots,$$

that conjugates $F(x)$ to its linear part Ax is constructed degree by degree; $u_i(x)$ is a homogeneous polynomial of degree i. We now explain why we cannot always find a formal coordinate transformation U that linearizes F, by analyzing the l-th step of this procedure. We make one step in the formal procedure for the construction of the normal form. Therefore, let $F(x) = Ax + f_2(x) + f_3(x) + \ldots + f_l(x) + \ldots$ be the Taylor series expansion of $F(x)$ and let $v_l(x)$ be a homogeneous polynomial of degree l. Define $V_l(x) = x + v_l(x)$, then $V_l(x)$ is a polynomial and is hence defined on $\mathbb{C}^n$, however, V_l is only invertible in a neighbourhood of the origin. This neighbourhood may shrink as l increases. The inverse $V_l^{-1}(x)$ has Taylor series expansion $V_l^{-1}(x) = x - v_l(x) + \ldots$ as is readily verified. Consequently, the Taylor expansion of the conjugation $V_l^{-1} \circ F \circ V_l$ is

$$\begin{aligned} V_l^{-1} \circ F \circ V_l(x) &= (x - v_l(x) + \ldots) \circ (Ax + f_2(x) + \ldots + f_l(x) + \ldots) \circ (x + v_l(x)) \\ &= Ax + f_2(x) + \ldots + f_{l-1}(x) + (f_l(x) + Av_l(x) - v_l(Ax)) + \ldots. \end{aligned}$$

It is hence natural to introduce the linear operator $d_0(v_l)(x) = v_l(Ax) - Av_l(x)$ on the space of polynomials of degree l. If this operator is surjective, then v_l can be chosen such that $f_l(x) = d_0(v_l)(x) = v_l(Ax) - Av_l(x)$, and terms of degree l in the Taylor series expansion of $V_l^{-1} \circ F \circ V_l$ vanish. However, this operator is not always surjective. Let us explain this, for simplicity of the exposition, for diagonal linear parts $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Let $k = (k_1, \ldots, k_n) \in \mathbb{N}^n$, $j \in \{1, \ldots, n\}$. We use the short hand notation $x^k e_j = \Big(0, \ldots, 0, x_1^{k_1} \ldots x_n^{k_n}, 0, \ldots, 0\Big)$. In this case d_0 is clearly a diagonal operator since $d_0(x^k e_j) = (\lambda^k - \lambda_j)x^k e_j$. It is surjective if $(\lambda^k - \lambda_j) \neq 0$ for all k for which $|k| \geq 2$. Each polynomial (of degree l) in the kernel of d_0 is called a *resonant polynomial* (of degree l). It now is clear that if $f_l = \sum_{|k|=l} f_k x^k e_j$, we can choose v_l such that

$$v_l(Ax) - Av_l(x) = \sum_{\substack{|k|=l, \\ \lambda^k - \lambda_j = 0}} f_k x^k e_j.$$

Repeating this procedure recursively starting from $l = 2$, we obtain the following classical theorem, already going back to H. Poincaré.

Theorem 1.1 *Let $F(x) = Ax + f(x)$ be a local analytic diffeomorphism fixing the origin, and suppose that $A = DF(0)$ is its linear part and $f(x) = O(||x||^2)$, $x \to 0$. Then there exists a formal coordinate transform $U(x) = x + u(x)$ such that*

$$U^{-1} \circ F \circ U(x) = Ax + \sum_{j=1}^{n} \sum_{\substack{|k|\geq 2, \\ \lambda^k - \lambda_j = 0}} g_k x^k e_j.$$

Proof: Repeat the above described procedure and define

$$U_l(x) = (x + v_2(x)) \circ (x + v_3(x)) \circ \ldots \circ (x + v_l(x))$$

such that

$$U_l^{-1} \circ F \circ U_l(x) = A.x + \sum_{\substack{2\leq |k| \leq l, \\ \lambda^k - \lambda_j = 0}} g_k x^k e_j + O(||x||^{l+1}).$$

Taking the limit $l \to \infty$ finishes the proof. □

For sure the existence of this formal transformation does not guarantee its convergence. But even at the formal level, there arise problems when the linear part depends on additional parameters. For example, suppose that $A = \mathrm{diag}(\lambda_1(\varepsilon), \ldots, \lambda_n(\varepsilon))$ and $\lambda^k(\varepsilon_0) - \lambda_j(\varepsilon_0) = 0$ for a certain value of ε_0 and a fixed k and j. It is in this case quite unusual that $\lambda^k(\varepsilon) - \lambda_j(\varepsilon) = 0$, (resp. $\lambda^k(\varepsilon) - \lambda_j(\varepsilon) \neq 0$) for all ε close to ε_0. On the contrary if $\lambda^k(\varepsilon_0) - \lambda_j(\varepsilon_0) \neq 0$ for all values of k and j, then it is unusual that $\lambda^k(\varepsilon) - \lambda_j(\varepsilon) \neq 0$ for all ε close to ε_0 and all values of k and j. As a consequence the procedure described above is usually discontinuous: for ε_0 a resonant term will appear while for $\varepsilon \neq \varepsilon_0$, no such term will be present in the Taylor series expansion of the normal form. We want to avoid such discontinuous behaviour. In order to do so, there are at least two options available. The first option is to allow more (non-analytic) transformations by allowing for example a formal expansion that contains logarithmic-like terms for the transformations. This is explained in [5] for vector fields and in [24] for both vector fields and diffeomorphisms. This approach is sometimes preferable to allowing the transformations to be only of type C^k for a finite k, since it allows one to see the very specific nature of the type of C^k-behaviour. The second option is not to remove those terms in the Taylor series expansion where this phenomenon arises. This is the main subject of Chapter 2, where it is extended to the context of Banach spaces. We will call these simplified local models briefly 'normal forms in cones' (or 'normal forms', if there is no confusion possible with the classical notion of a normal form). The name is due to the formal structure: the terms $x^k e_j$ that are possibly not removed in the Taylor series of these local models have indices k that lie in a set K that has a conical structure in $\mathbb{N}^n$, see Section 2.2.5 for a precise definition. These results have been published in [4]. The drawback for these normal forms in cones is obvious: it contains more terms than explained in the formal procedure described in 1.1, but there is also one main advantage: these normal forms in cones, as well as their corresponding coordinate transforms actually converge and the dependence of parameters is allowed (also in the linear part). This is mainly due to the fact that we stay away from the resonant region.

1.2 Convergence versus divergence

The formal transformation U, constructed in Section 1.1, that transforms a local analytic diffeomorphism F into its classical normal form, depends strongly on the linear part of $A = DF(0)$ of this diffeomorphism. If the eigenvalues of the linear part are resonant, there may arise problems concerning the convergence of this normal form transformation. But even if the eigenvalues are non-resonant, there are sometimes convergence problems. We explain first the situation where things are known to converge. It is known that if $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, and the eigenvalues satisfy the Brjuno condition

$$-\sum_{k=1}^{\infty} 2^{-k} \log(\omega(2^k)) < \infty, \text{ where } \omega(m) = \inf\{|\lambda^i - \lambda_j| \mid 2 \leq |i| \leq m, j \in \{1, \ldots, n\}\},$$

then the linearizing formal transformation U converges. We extend these results in Section 5.1.3 in a more general context. See e.g. [7], [65], [13] for related results. As far as we know, it is not known in general whether the Brjuno condition is sharp or not. In [64] it is shown that in the one-dimensional situation the Brjuno condition is sharp.

If the linear part is either not semi-simple, or, if the normal form contains resonant terms, there may arise convergence problems. In Chapter 5 we consider the problem where resonances are involved, but where the linear part A is diagonal. In this case, we show that if the linear part A has eigenvalues that satisfy the Siegel condition of type τ,

$$\exists C > 0, \forall k \in \mathbb{N}^n, |k| \geq 2, \forall j \in \{1, \ldots, n\} : \left|\lambda^k - \lambda_j\right| \geq \frac{C}{|k|^\tau},$$

then there exists a formal Gevrey-$(2 + 2\tau)$ transformation to a normal form that is Gevrey-$(2 + 2\tau)$. By this we mean that the formal transformation $U(x) = x + \sum_{\delta \geq 2} u_\delta(x)$ that is constructed satisfies $||u_\delta(x)||_R \leq (\delta!)^{2+2\tau}$ for a certain $R > 0$. Here $u_\delta(x)$ is a polynomial of degree δ and $||u_\delta(x)||_R = || \sum_{|\alpha|=\delta} u_\alpha x^\alpha ||_R = \sum_{|\alpha|=\delta} |u_\alpha| R^\delta$. We prove this result in Section 5.1.3. Comparable results for vector fields have been proven in [34], however the current proofs are very different. In fact, for vector fields, it is also shown in [34] that there exists an optimal cut-off to stop the normal form procedure. The vector field X is then conjugated to the vector field $Y(x) = Ax + f_2(x) + f_n(x) + R(x)$, where f_i is a polynomial of degree i and the remainder is exponentially small with respect to $||x||$. It would be interesting to see if a similar cut-off also exists for normal forms of diffeomorphisms.

For vector fields with non-resonant linear part $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ a similar Brjuno condition

$$-\sum_{k=1}^{\infty} 2^{-k} \log(\omega(2^k)) < \infty, \text{ where } \omega(m) = \inf_{2 \leq |k| \leq m, 1 \leq j \leq n} | \langle \lambda, k \rangle - \lambda_j |$$

exists. It is not known whether this condition is sharp or not, but the author of [7] shows that if $\limsup_{k \to \infty} \frac{\omega(2^k)}{2^k} = +\infty$, there exist examples of vector fields that have a divergent linearizing transformation.

1.3 Vector fields versus diffeomorphisms

There is a close correspondence between results concerning diffeomorphisms and vector fields. This is usually achieved by considering the time-one diffeomorphism that is constructed from the flow of the vector field in casu. The theorems for vector fields are usually a consequence of the theorems for diffeomorphisms. This correspondence is commonly used for example in the proof of invariant and center manifold theorems, see e.g. [53], the theorems of Hartman and Grobmann in [28], [27], ...

We explain briefly the principle, focusing on the aspects that are relevant with respect to this text. Consider an analytic vector field X, with fixed point at 0 and defined near the origin. One can show, using for example the Picard-Lindelöf iteration, that the flow $\phi_{x_0}(t)$ of this vector field is an analytic function. This flow is the unique solution of the initial value problem $\dot{\phi}(t) = X \circ \phi(t)$, $\phi(0) = x_0$. One can check that the series $\exp(tX)(x_0) = \sum_{n \geq 0} t^n \underbrace{X \circ \ldots \circ X}_{n}(x_0)$ is a formal solution to this initial value problem that converges for small values of t. Hence $\exp(tX)(x_0) = \phi_{x_0}(t)$. The map $\exp(X)$ is called the time-one map of the vector field. We will suppose that it converges (if not, one could consider $\exp(t_0 X)$, for a fixed t_0 small enough). There is a close relation between the study of vector fields and the study of diffeomorphisms by considering this time-one map. Indeed, each transformation U that linearizes the flow of the vector field, will also linearize the corresponding time-one map. It is however not always the case that an analytic diffeomorphism can be *embedded* in an analytic vector field, see e.g. [32], meaning that it is possible to find analytic diffeomorphisms $F(x)$ that cannot be written as $\exp(X)$ with X an analytic vector field. Therefore, one should always be a bit cautious when generalizing theorems that have been proven for vector fields to equivalent theorems concerning diffeomorphisms.

Let me indicate at least two places in this text where we do not have a unified approach for diffeomorphism compared to vector fields:

1. The main result in Chapter 2 could be seen as the 'diffeo version' of the results by the authors of [3], see also the beginning of Chapter 3. However, the corresponding proof for diffeomorphims is in this particular case not a copy of the theorem concerning vector fields: it is more complicated due to the fact that

the main machinery –essentially an application of the implicit function theorem– has to be applied twice, while maintaining the structure.

2. In Section 5.1.3 of Chapter 5 we provide results concerning perturbations of vector-valued polynomials of quasi-homogeneous degree 0. These again could be seen as the 'diffeo version' of the results in [37], but there is some difference: for the corresponding theorems concerning vector field, one can equally well consider perturbations of a quasi-homogeneous vector field of quasi-homogeneous degree s, where s is not necessarily 0. It should also be noted that the proofs concerning Gevrey-results appearing in Section 5.1.3 do not resemble their vector field analogues at all.

Chapter 2

Analytic families of diffeomorphisms and formal cones

2.1 Introduction and statement of the results

In this chapter we explore the limits of analytic simplification, by means of changes of variables (i.e. a conjugacy), of a dynamical system described by a diffeomorphism, in the neighbourhood of a fixed point p. Normal forms obtained through such a conjugacy often simplify the local analysis needed as a starting point for understanding more difficult global phenomena. For instance, if there is a saddle-type fixed point p, it is important to have a good local model in order to study the orbits in the vicinity of the fixed point. In bifurcation theory it is assumed that this diffeomorphism moreover depends on additional parameters. Hence the dependence of the change of variables on the parameter will also be of importance.

We consider a family of diffeomorphisms $F_\mu : \mathbb{C}^n \longmapsto \mathbb{C}^n$ fixing the origin i.e. $F_\mu(0) = 0$ for all values of μ. We look for a (parameter dependent) change of variables U_μ such that $G_\mu = U_\mu^{-1} \circ F_\mu \circ U_\mu$ has as few terms as possible in its Taylor series expansion. Later, we will be more precise on the meaning of 'as few terms as possible' and the dependence on the parameter μ. It is our aim to look for conjugacies in the same smoothness category as the diffeomorphism, whenever possible.

It is well known that the arithmetic relations between the eigenvalues of the linear part at the fixed point determine to a great extent the kind of normal form that can be obtained by a conjugacy. In a generic family these relations may vary significantly if the parameter changes, so this has an influence on the normal form. If we start from an analytic diffeomorphism, we look for a 'simplest possible' analytic local model and corresponding conjugacy. We will refer to such a local model as 'a normal form', although that it does not always correspond to the definitions that are common in the literature: our normal form will contain some non-resonant terms in the ‘flat remainder’. Ideally the normal form would be: the linear part of the diffeomorphism, or at least some polynomial form. Unfortunately even for a general parameter-dependent saddle the existence of such a normal form is highly unexpected, even on the level of formal Taylor series. For example in two real dimensions, if the eigenvalues are λ_1 and λ_2, with $0 < \lambda_1 < 1 < \lambda_2$, the ratio $\log \lambda_1 / \log \lambda_2$ may perturb to rational and irrational values, giving an obstruction for a polynomial analytic normal form, even on the formal level.

One can then reduce analytically to a polynomial normal form up to a ‘finitely flat remainder’, that is: a remainder term of finite order in the space variable. In the context of vector fields this approach was already studied in H. Dulac’s memoir [22] for planar systems; a generalization can be found in [57].

Our methods below allow to give an explicit and sharp expression for this flat remainder, that presumably cannot be improved in the analytic setting that we consider. A first result along this line is published in [6] where we treat normal forms for saddles in the planar case. If there are extra constraints on the system, or if there are no parameters, then a further analytic simplification is sometimes possible. Several contributions have been made, see e.g. [2], [16], [66], [56], [49], [19], [54], [52], [12], [58]. We emphasize that we will only discuss analytic conjugacies and normal forms. The use of finitely smooth (C^k, $k < \infty$) conjugacies in order to eliminate this flat remainder, will not be discussed this here. See e.g. [31], [42]. The method of proof of the principal result closely follows the ideas in [44].

2.1.1 Setting

The conjugacy problem for an analytic diffeomorphism has a long history. Here we give a short overview to provide a context for our results. Let us first introduce some terminology and notation. We will frequently use multi-index notation, that is: $k = (k_1, \ldots, k_m)$, $|k| = k_1 + \cdots + k_m$ and $\lambda^k = \lambda_1^{k_1} \cdots \lambda_m^{k_m}$. Furthermore we denote $e_j = (0, \ldots, 1, \ldots, 0)$, the non-zero component at the j-th place.

We consider a family $F_\mu : \mathbb{C}^m \mapsto \mathbb{C}^m$ of local analytic diffeomorphisms, depending on μ in some set of parameters Λ, with $F_\mu(0) = 0$ for all parameter values $\mu \in \Lambda$. For example for hyperbolic fixed points it is not restrictive to assume that the fixed point is at the origin for all μ near some given parameter value $\mu = \mu_0$. The Taylor series of F_μ which converges on some polydisc. Let $F_\mu(z) = A_\mu z + f_\mu(z)$ where $A_\mu = D_z F_\mu(0)$ is the linear part of F_μ at zero and $f_\mu(z) := F_\mu(z) - A_\mu z$, so that $D_z f_\mu(0) = 0$. In order to explain the ideas we assume, for simplicity, that A_μ is semi-simple, although this hypothesis will not be necessary in the principal result in Section 2.1.3 . Then there is a μ-dependent basis such that $A_\mu = \operatorname{diag}(\lambda_1(\mu), \cdots, \lambda_m(\mu))$. Let us fix the parameter at $\mu = \mu_0$ for this moment, then the eigenvalues λ of A are called *resonant* if there exist $(k, j) \in \mathbb{N}^m \times \{1, \ldots, n\}$ with $|k| > 1$ and $R(\lambda, k, j) = 0$, where the function R is defined as

$$R(\lambda, k, j) = \lambda_j - \lambda^k. \tag{2.1}$$

Conversely, if for all $(k, j) \in \mathbb{N}^m \times \{1, \dots, n\}$ one has $R(\lambda, k, j) \neq 0$ then the eigenvalues are called *non-resonant.* A term $x^k e_j$ in the Taylor series of F, is called *resonant* if $R(\lambda, k, j) = 0$ and *non-resonant* if $R(\lambda, k, j) \neq 0$. Thus only for resonant eigenvalues, resonant terms exist. The diffeomorphism G is called a *normal form* of F if the Taylor series of g in $G = A + g$ consists of resonant terms only.

Classical results for the conjugacy problem with a fixed parameter are theorems by Poincaré and Siegel, see for example [1]. The theorem by Poincaré, in the real case, assumes that the eigenvalues of A are located either all inside the unit circle or all outside the unit circle. Then F is locally linearizable, that is $G = A$, by an analytic change of coordinates in the absence of resonance. In the presence of resonance there is an analytic transformation which conjugates F to a polynomial $G = A + p$ containing resonant terms only. In case of parameter dependency, this was studied in [10] for vector fields and in [25] for diffeomorphisms.

The assumptions of the theorem by Poincaré exclude a number of interesting cases like reversible and Hamiltonian maps. These cases are included in the theorem of Siegel which states that F is locally linearizable by an analytic transformation if the eigenvalues of A satisfy a diophantine condition: there exist C and τ such that for all (k, j) with $|k| > 1$

$$|R(\lambda, k, j)| \geq C|k|^{-\tau}. \tag{2.2}$$

This diophantine condition clearly excludes resonant eigenvalues. Later these results have been refined in [7] yielding the Brjuno condition

$$-\sum_{k=1}^{\infty} 2^{-k} \log(\omega(2^k)) < \infty,$$

where $\omega(m) = \inf_{2 \leq |i| \leq m, 1 \leq j \leq n} |\lambda^i - \lambda_j|$. Recent results on this matter where obtained by [48].

Another approach using the diophantine conditions on eigenvalues is KAM-theory. Here one obtains linearizability of systems depending on parameters. But since eigenvalues vary when parameters are varied, results in KAM theory hold on Cantor sets of parameter values. The subject is much broader than we wish to describe here, for an overview see [9]. Similarly the Brjuno condition only holds on Cantor sets of eigenvalues and thus of parameter values. However, here we wish to consider open sets of parameter values. This means that eigenvalues will vary in open sets. If one assumes that the parameter only affects the nonlinear part, one can find results for example in [50]. Since every open set of eigenvalues contains resonant eigenvalues we have to take resonance into account.

There are only a few results for the complementary situation. That is, parameter dependent systems with parameters in an open set, where eigenvalues are located on either side of the unit circle and possibly resonant. Here we mention [45] and [18] where analytic normal forms are presented for particular two and four dimensional systems. Also see [20] and references therein.

Although comparable results in the analytic category, like the ones in this section, are already known for *vector fields* [3], we have experienced that the usual passage from 'the vector fields case' to the 'diffeomorphisms case' is not at all as classical as could be expected; particular issues appear such as in Section 2.2.5.

To our knowledge there are two ways to proceed in general. Either enlarge the transformation group or allow a more general normal form. Enlarging the transformation group will almost inevitably mean losing smoothness. In our approach, we will keep analyticity and allow a 'slight tolerance' to the formal normal form.

2.1.2 An example

In order to fix the main ideas we sketch a simple example. We consider an analytic family of saddles in $\mathbb{R}^2$ passing through a $1:-1$ resonance. By this we mean a family

$$F(x,\mu) = \left(x_1 \left(\lambda_1(\mu) + \sum_{|k|\geq 1} f_k^1(\mu) x^k \right), x_2 \left(\lambda_2(\mu) + \sum_{|k|\geq 1} f_k^2(\mu) x^k \right) \right).$$

Remind that $x = (x_1, x_2)$ and $k = (k_1, k_2)$. We will suppose for simplicity that the numbers $\lambda_1(\mu), \lambda_2(\mu)$ are real and positive, for each value of the parameter λ. Suppose that for $\mu = \mu_0$ the condition

$$\frac{\log(\lambda_1(\mu_0))}{\log(\lambda_2(\mu_0))} = -1 \tag{2.3}$$

holds and that the family depends analytically on the parameter μ. When the parameter is fixed at $\mu = \mu_0$, we see that the resonant terms $f_k^1(\mu_0)x^k$, $f_k^2(\mu_0)x^k$ of $F(x,\mu_0)$ correspond to those k for which $k_1 = k_2$. We claim the following: given any big $N \in \mathbb{N}$ there exists an analytic change of variables, depending moreover analytically on the parameter μ near μ_0, conjugating $F(x,\mu)$ to

$$\begin{aligned} G(x,\mu) = \Bigg(& x_1 \left(\lambda_1(\mu) + b_0^1(\mu,u) + \sum_{s\geq 1} u^{Ns} (x_2^s b_s^1(\mu,u) + x_1^s b_s^2(\mu,u)) \right), \\ & x_2 \left(\lambda_2(\mu) + b_0^2(\mu,u) + \sum_{s\geq 1} u^{Ns} (x_2^s c_s^1(\mu,u) + x_1^s c_s^2(\mu,u)) \right) \Bigg), \end{aligned} \tag{2.4}$$

where $u = x_1 x_2$ and where all the occurring functions are analytic. Note that, if we put $\mu = \mu_0$ and if we truncate the foregoing expression to $\hat{G}(x,\mu_0) = (x_1(\lambda_1(\mu_0) + b_0^1(\mu_0,u)), x_2(\lambda_2(\mu_0) + b_0^2(\mu_0,u)))$, then we have the usual normal form at $\mu = \mu_0$.

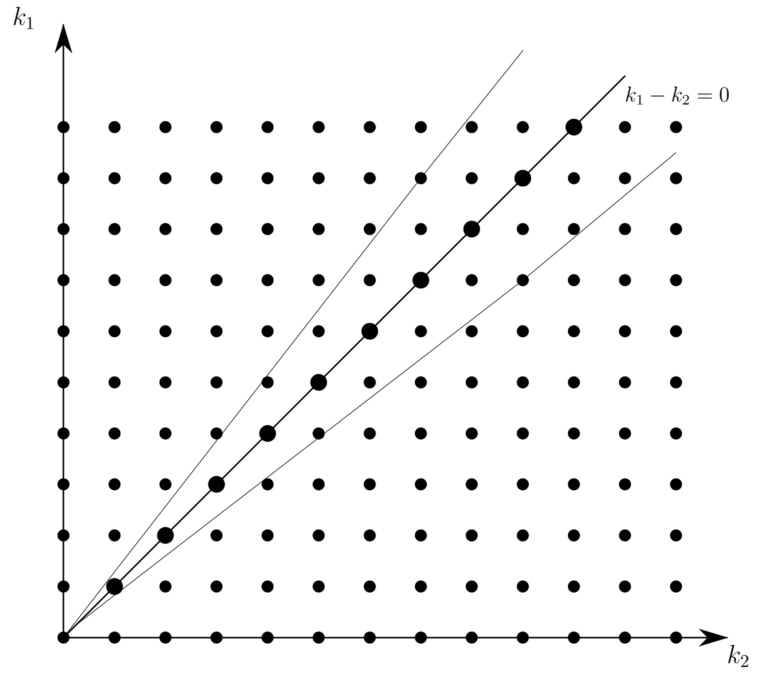


Figure 2.1: The cone B of terms which cannot be removed.

Moreover the 'remainder' $R(x,\mu) = G(x,\mu) - \hat{G}(x,\mu)$ is N-flat in u and we obtain an explicit form for this remainder R.

Let us explain, on this example, some ideas that we will use in the next sections in a general framework. Therefore we fix the direction $j = 1$ or $j = 2$. We have that all the resonant terms in the j-th component of (2.4) correspond to dots in the first quadrant of the plane on the line $k_1 = k_2$. Moreover, when the parameter varies slightly to μ, the ratio $\log(\lambda_1(\mu_0))/\log(\lambda_2(\mu_0))$ varies only slightly to $-1+\epsilon(\mu)$. For this parameter value μ we draw the points that correspond to the possible resonant eigenvalues, we see that they lie on the line $k_2 = -\dfrac{\ln(\lambda_1(\mu))}{\ln(\lambda_2(\mu))}k_1$. This line has a slope that is close to the slope of $k_1 = k_2$ if μ is close to μ_0. Remark that resonance occurs each time the ratio $-\dfrac{\ln(\lambda_1(\mu))}{\ln(\lambda_2(\mu))}$ is rational. Since at a fixed resonant parameter value, we know that in general we cannot remove those terms analytically (not even formally), it is reasonable to expect that, when varying the parameter, the best parameter dependent normal form one can obtain, is a normal form which contains no terms outside the cone B that consists of all the resonant terms for all values of the parameter. We will call G the set outside the cone B the set of 'good terms', since it contains the indices corresponding to terms that we are able to remove in our normal form. We call the complementary set B the 'bad set'. Remark that the set B can be described as (we use a notation consistent with the next sections):

$$B_{D,D} = \left\{ k \in \mathbb{N}^2 | \, \lambda_1(\mu_0)^{k_1} \lambda_2(\mu_0)^{k_2} > D^k \text{ and } \lambda_1(\mu_0)^{k_1} \lambda_2(\mu_0)^{k_2} > \frac{1}{D^k} \right\},$$

for a certain $0 < D < 1$. Indeed: this is equivalent to

$$B_{D,D} = \left\{ k \in \mathbb{N}^2 | \, - \frac{\log(\lambda_1(\mu_0)) + \log(D)}{\log(\lambda_2(\mu_0)) + \log(D)} > \frac{k_2}{k_1} > - \frac{\log(\lambda_1(\mu_0)) - \log(D)}{\log(\lambda_2(\mu_0)) - \log(D)} \right\},$$

We make a diagram of the situation in Figure 2.1. The grid of dots represent the possible terms in the Taylor series expansion: (k_1, k_2) corresponds to the term $x^k x_j e_j$. The resonant terms in the direction j are clearly $(x_1 x_2)^k x_j e_j$ and correspond to couples (k, k) on the figure. The exponents (k_1, k_2) that are in $B_{D,D}$ (the terms that are possibly not removed from the Taylor series) is shown on the diagram in figure 2.1: $B_{D,D}$ contains the tuples of natural numbers in the cone determined by the lines $k_2 = -k_1 \frac{\log(\lambda_1(\mu_0)) + \log(D)}{\log(\lambda_2(\mu_0)) + \log(D)}$ and $k_2 = -k_1 \frac{\log(\lambda_1(\mu_0)) - \log(D)}{\log(\lambda_2(\mu_0)) - \log(D)}$. One of these lines has a slope smaller than 1, one has a slope bigger than 1.

It is the subject of this chapter to show that for every $0 < D < 1$ there exists a neighborhood of μ_0 and a parameter dependent normal form not containing any term outside $B_{D,D}$. When D is close to 1, the cone of terms becomes very narrow, and a better normal form is obtained (but the neighborhood of μ_0 in the parameter space and the radius of convergence of the normal form and the normal form transformation may shrink). Hence, for a fixed D, we find the existence of a normal form

$$G(x, \mu) = \left(x_1 \left(\lambda_1(\mu) + \sum_{\substack{k \in B_{D,D}, \\ |k| \geq 1}} f_k^1(\mu) x^k \right), x_2 \left(\lambda_2(\mu) + \sum_{\substack{k \in B_{D,D}, \\ |k| \geq 1}} f_k^2(\mu) x^k \right) \right). \tag{2.5}$$

An explicit characterization of the coefficients in $B_{D,D}$ allows us to relate 2.5 to (2.4); we refer to Section 2.6.1.

2.1.3 Statement of the results

The example in the previous section can be generalized to higher dimensions, but it can equally well be generalized to maps on Banach spaces E. We have of course $E = \mathbb{C}^n$ or $\mathbb{R}^n$ in mind as main cases. It is hence natural to present the results in the most general framework. This will not complicate the exposition. Suppose that an analytic function fixing the origin $F : U \subset E \rightarrow E$ is given, where $E = E^1 \oplus \ldots \oplus E^n$ is a direct sum of Banach spaces, and let A be the linear part of F. Suppose also that each E_i is an invariant subspace for F (we will comment on this assumption later).

We use the usual formalism of symmetric multi-linear maps on direct sums of vector spaces, including multi-index notation: see Section 2.2.4 for details.

We suppose that F is analytic near 0, that is: for a certain $r > 0$, the Taylor series of F converges to F for all $||x|| \leq r$ and, due to the invariance of the splitting, we can write:

$$F(x) = \sum_{j=1}^{n} \sum_{k \in \mathbb{N}^n} F^j_{k+e_j}(x^{k+e_j}).$$

Furthermore, it follows that A is block-diagonal with respect to the direct sum splitting of E, i.e. $A = A^1 \oplus \ldots \oplus A^n$, with each A^i a continuous and invertible linear map $A^i : E^i \to E^i$. Put $\lambda_i = ||A^i||$, $\widetilde{\lambda_i} = ||(A^i)^{-1}||$, $\rho = \max_{i=1}^{n}\{\lambda_i.\widetilde{\lambda_i}\}$ and let $D, C \in \mathbb{R}$ be fixed, such that $0 < D\rho < 1$ and $0 < C\rho < 1$. We introduce the good set as

$$G_{D,C} = \{k \in \mathbb{N}^n |\, \lambda^k \leq D^{|k|} \text{ or } \widetilde{\lambda}^k \leq C^{|k|}\},$$

and the bad set as its complement

$$B_{D,C} = \mathbb{N}^n \setminus G_{D,C}.$$

We give some brief comments on the value of ρ. If $E = \mathbb{C}^n$ and $A^i = [a_i]$, $i = 1, \ldots, n$ (i.e. A is a diagonal matrix) then $\rho = 1$; this is also true if each A^i is a multiple of the identity map. In the case that A^i is a Jordan block we can assume, up to a linear change of variables, that ρ is arbitrarily close to 1. On the other hand, if the variation of the spectrum of A^i is large, then ρ can be large compared to 1. In the case of matrices the factor ρ is known as the condition number.

Our main result is the following:

Theorem 2.1 *Suppose E is a Banach space that admits a direct sum decomposition $E = E^1 \oplus \ldots \oplus E^n$. Suppose that $F : E \to E$ is an analytic function for which its Taylor series converges to F for all $||x|| \leq r$. Suppose that each E^i is an invariant subspace for F, and let A be the linear part of F. We suppose that A is linear and continuous. Then there exists an $\widetilde{r} > 0$ and an analytic near identity transformation U convergent for each $||x|| \leq \widetilde{r}$ such that*

i) U contains only terms in the good set, i.e.

$$U(x) = id + \sum_{j=1}^{n} \sum_{\substack{k \in G_{D,C}, \\ |k| \geq 1}} u^j_{k+e_j}(x^{k+e_j}).$$

ii) The conjugation $G = U^{-1} \circ F \circ U$ contains only terms in the bad set, i.e.

$$G(x) = Ax + \sum_{j=1}^{n} \sum_{\substack{k \in B_{D,C}, \\ |k| \geq 1}} g^j_{k+e_j}(x^{k+e_j}).$$

Moreover, $G(x)$ is convergent on $\{x |\, ||x|| \leq \widetilde{r}\}$.

As a consequence we obtain the following theorem:

Theorem 2.2 *Suppose that $F : \mathbb{C}^n(\Lambda) \to \mathbb{C}^n(\Lambda)$ is a parameter dependent analytic function leaving invariant each coordinate axis, so F is of the form*

$$F(x) = \sum_{j=1}^{n} \sum_{k\in\mathbb{N}^n} F^j_{k+e_j}(\mu) x^{k+e_j} e_j, \tag{2.6}$$

Suppose that F depends continuously (resp. C^k, C^∞, C^ω) on the parameter and that its Taylor series converges to F for all $||x|| \leq r$, i.e.

$$\|F\|_r := \max_{j=1}^{n} \left(\sum_{k\in\mathbb{N}^n} \sup_{\mu\in\Lambda} ||F^j_{k+e_j}(\mu)|| r^{k+1} \right) < \infty.$$

From expression 2.6 automatically we have that the linear part A of f is semi-simple i.e. $A = \operatorname{diag}(\lambda_1(\mu), \ldots, \lambda_n(\mu))$. Define the good set as

$$G_{D,C} = \{k \in \mathbb{N}^n | \, |\lambda(\mu_0)|^k \leq D^{|k|} \text{ or } \frac{1}{|\lambda(\mu_0)|^k} \leq C^{|k|}\},$$

and the bad set $B_{D,C}$ as its complement.

Then there exists an $\widetilde{r} > 0$, a neighborhood $\widetilde{\Lambda}$ of μ_0 and a near identity transformation U that is analytic in x and depends continuously (resp. C^k, C^∞, C^ω) on the parameter such that

i) U contains only terms in the good set, i.e.

$$U(x) = id + \sum_{j=1}^{n} \sum_{\substack{k\in G_{D,C},\\ |k|\geq 1}} u^j_{k+e_j}(\mu) x^{k+e_j} e_j.$$

and

$$\|U\|_{\widetilde{r}} := \max_{j=1}^{n} \left(\widetilde{r} + \sum_{\substack{k\in G_{D,C},\\ |k|\geq 1}} \sup_{\mu\in\widetilde{\Lambda}} ||u^j_{k+e_j}(\mu)|| \widetilde{r}^{k+1} \right) < \infty.$$

ii) The conjugation $G = U^{-1} \circ F \circ U$ does not contain any term in the good set, i.e.

$$G(x) = Ax + \sum_{j=1}^{n} \sum_{\substack{k\in B_{D,C},\\ |k|\geq 1}} g^j_{k+e_j}(\mu) x^{k+e_j} e_j$$

and

$$\|G\|_{\widetilde{r}} := \max_{j=1}^{n} \left(r + \sum_{\substack{k\in B_{D,C},\\ |k|\geq 1}} \sup_{\mu\in\widetilde{\Lambda}} ||g^j_{k+e_j}(\mu)|| \widetilde{r}^{k+1} \right) < \infty.$$

2.1.4 Method of proof of Theorem 2.1

Write $F = A + f$ where A is the linear part and f is the nonlinear part. We assume that A is already in some standard form so we do not perform linear transformations, that is we let $U = \mathrm{id} + u$ be a near identity transformation. Thus A is also the linear part of G and we write $G = A + g$.

Inspired by [44] we use the following approach. We write the conjugacy problem as

$$0 = F \circ U - U \circ G = A \circ u - u \circ (A + g) + f \circ (\mathrm{id} + u) - g. \tag{2.7}$$

With appropriate open parts of Banach spaces V, W, X and Z, to be defined in Section 2.3, we introduce the functional

$$\mathcal{F} : V \times W \times X \to Z : (f, g, u) \mapsto A \circ u - u \circ (A + g) + f \circ (\mathrm{id} + u) - g \tag{2.8}$$

and we solve $\mathcal{F}(f, g, u) = 0$ for (g, u) given the map f. We will do this by an application of the implicit function theorem. The main difficulty in applying this theorem is to prove that $\mathcal{F}$ is well defined between appropriate function spaces, and is C^1 in (f, g, u). In order to achieve this result we need some machinery which will be reviewed in Section 2.2. In Section 2.3 we prove Theorem 2.1 and in Section 2.5 we will prove Theorem 2.2 .

2.2 Analytic functions on Banach spaces

Here we review some properties of analytic maps between Banach spaces, focusing on sums, products and compositions. Our approach is based on that of [44]. The main reason for this approach is that it provides good estimates for the functional $\mathcal{F}$ and its derivatives. We need these in order to solve the reformulated conjugacy problem in equation (2.7) by applying the implicit function theorem to the functional equation $\mathcal{F}(f, g, u) = 0$.

An important aspect is the following problem: for A, B and C being certain carefully chosen Banach spaces, the composition operator $O : A \times B \to C : (f, g) \mapsto f \circ g$ should be well defined and continuously differentiable.

2.2.1 Local analytic functions

Suppose that E and F are Banach spaces and that $f : E \to F$ is a C^∞ function near $x = 0$. Using Taylor's theorem we obtain

$$f(x) = f(0) + \sum_{k=1}^{n} f_k(x, \ldots, x) + O(||x||^{n+1}),$$

where $f_k : E^k \to F$ is a k-multi-linear symmetric mapping. Let us be more precise about our notion of an analytic function from E to F. A function $f : E \to F$ is analytic at 0 when its Taylor series converges to f at least on a small disk

$$B_E(0;r) = \{x \in E|\, ||x|| \leq r\}$$

around the origin.

Definition 2.3 *We define $\mathcal{L}^k(E,F)$ to be the space of k-multi-linear symmetric mappings $f_k : E^k \to F : (x_1, x_2, \ldots, x_k) \to f_k(x_1, x_2, \ldots, x_k)$, i.e.*

$$f_k(x_1, \ldots, x_i, \ldots, x_k) = f_k(x_{\varphi(1)}, \ldots, x_{\varphi(i)}, \ldots, x_{\varphi(k)})$$

for all $x_i \in E$ and all permutations $\varphi \in S_k$.

Equipped with the norm

$$||f_k|| := \sup_{x \in E} \frac{||f_k(x, \ldots, x)||}{||x||^k},$$

it is a standard result that $\mathcal{L}^k(E,F)$ is a Banach space.

We introduce the analogue of formal power series for Banach spaces, and define analytic functions as those power series that converge absolutely on a certain neighbourhood of the origin.

Definition 2.4 *We define formal power series and convergent power series $E \to F$.*

i) We denote by $\mathcal{P}(E,F)$ the set of formal power series $f = \sum_{k \geq 0} f_k$, where $f_k \in \mathcal{L}^k(E,F)$.

ii) $\mathcal{A}(E,F)$ is the set of formal power series $f = \sum_{k \geq 0} f_k$, where $f_k \in \mathcal{L}^k(E,F)$ are such that there exists a $r > 0$ for which $\sum_{k \geq 0} ||f_k|| r^k < \infty$. (Note that this condition is equivalent with $\overline{\lim}_{k \to \infty} \sqrt[k]{||f_k||} < \infty$.) We will refer to $\mathcal{A}(E,F)$ as the set of convergent power series from E to F.

iii) $\mathcal{A}_r(E,F)$ is the subset of $\mathcal{A}(E,F)$ for which $||f||_r := \sum_{k \geq 0} ||f_k|| r^k < \infty$, for some $r > 0$. We will refer to $\mathcal{A}_r(E,F)$ as the set of convergent power series with radius of convergence at least r.

Note that when $f \in \mathcal{A}(E,F)$, for each $x \in E$, with $||x|| \leq r$, the power series $\sum\limits_{k \geq 0} f_k(x, \ldots, x)$ converges absolutely since

$$\sum_{k \geq 0} ||f_k(x, \ldots, x)|| \leq \sum_{k \geq 0} ||f_k|| ||x||^k \leq ||f||_r.$$

Hence we can define for each power series $f = \sum_{k \geq 0} f_k \in \mathcal{A}(E,F)$ the associated the analytic function

$$f : B_E(0;r) \to F : x \mapsto \sum_{k \geq 0} f_k(x, \ldots, x).$$

Consequently there is a one-to-one relation between analytic functions $f : B_E(0;r) \to F$ and elements of $\mathcal{A}_r(E,F)$. We will most often regard elements $f \in \mathcal{A}_r(E,F)$ as formal power series, but will sometimes refer to those f as being analytic functions. It is standard to show that $\mathcal{A}_r(E,F)$, $||.||_r$ is a Banach space.

2.2.2 Derivatives

In this section we derive some properties of derivatives and compositions of elements in $\mathcal{A}_r(E, F)$. Suppose that $f : E \to F$ is an arbitrary local analytic function defined at least on a small disk with radius r around the origin, identified with an element of $\mathcal{A}_r(E, F)$.

The next calculation serves as an introduction to the definition of derivatives on the space of formal power series. According to Taylor's theorem, we have that for $||x||$ and $||y||$ small enough

$$f(x+y) = \sum_{k\geq 0} \frac{1}{k!} D^k f_x.(y, \ldots, y)$$

on one hand, and on the other hand

$$\begin{aligned} f(x+y) &= \sum_{k\geq 0} \frac{1}{k!} D^k f_0.(x+y, \ldots, x+y) \\ &= \sum_{k\geq 0} f_k(x+y, \ldots, x+y) \\ &= \sum_{k\geq 0} \sum_{i=0}^{k} \binom{k}{i} f_k(\underbrace{x, \ldots, x}_{k-i}, \underbrace{y, \ldots, y}_{i}) \\ &= \sum_{k\geq 0} \sum_{i=0}^{k} \binom{k}{i} \widetilde{f}_k(\underbrace{x, \ldots, x}_{k-i})(\underbrace{y, \ldots, y}_{i}) \\ &= \sum_{i\geq 0} \left(\sum_{k\geq i} \binom{k}{i} \widetilde{f}_k(\underbrace{x, \ldots, x}_{k-i}) \right) (\underbrace{y, \ldots, y}_{i}). \end{aligned}$$

This suggest the following definition for the formal derivative.

Definition 2.5 *Let $f = \sum_{k\geq 0} f_k \in \mathcal{P}(E, F)$, then we define its formal i-th derivative as*

$$D^i f = \sum_{k\geq i} i! \binom{k}{i} \widetilde{f}_k \in \mathcal{P}(E, \mathcal{L}^i(E, F)),$$

where $\widetilde{f}_k$ is the element of $\mathcal{L}^{k-i}(E, \mathcal{L}^i(E, F))$ for which

$$\widetilde{f}_k(\underbrace{x, \ldots, x}_{k-i})(\underbrace{y, \ldots, y}_{i}) = f_k(\underbrace{x, \ldots, x}_{k-i}, \underbrace{y, \ldots, y}_{i}), \ \forall x, y \in E.$$

In the sequel we will omit the tilde when no confusion is possible.

Using Cauchy's inequality one obtains estimates on the derivatives of elements of $\mathcal{A}_r(E,F)$.

Lemma 2.6 *Let $a_i \geq 0$ and $M = \sum_{k\geq 0} a_k r^k < \infty$. Then, for $i \in \mathbb{N}$ and any $0 < \rho < r$ we have the estimate $\sum_{k\geq i} \frac{k!}{(k-i)!} a_k \rho^{k-i} \leq \frac{i!M}{(r-\rho)^i}$.*

Proof: This follows by applying Cauchy's inequality to the function $g(z) := \sum_{k\geq 0} a_k z^k$ which is analytic for $|z| < r$. We obtain that $|g^i(z)| \leq \frac{i! \sup_{|z|=r} |g(z)|}{(r-\rho)^i} \leq \frac{i!M}{(r-\rho)^i}$.
□

Lemma 2.7 *Let $f \in \mathcal{A}_r(E,F)$ and let $\rho \in \mathbb{R}$ be such that $0 < \rho < r$, then its ith derivative $D^i f$ is an element of $\mathcal{A}_\rho\left(E, \mathcal{L}^i\left(E,F\right)\right)$. Furthermore we obtain the estimate*

$$||D^i f||_\rho \leq \frac{i!||f||_r}{(r-\rho)^i}.$$

Proof: Since $D^i f = \sum_{k\geq i} i!\binom{k}{i} f_k$ it follows that

$$||D^i f||_\rho = \sum_{k\geq i} i!\binom{k}{i} ||f_k||\rho^k = \sum_{k\geq i} \frac{k!||f_k||\rho^k}{(k-i)!} \overset{(*))}{\leq} \frac{i! \sum_{k\geq 0} ||f_k||r^k}{(r-\rho)^i} = \frac{i!||f||_r}{(r-\rho)^i},$$

where we used Cauchy's inequality in $(*)$. □

2.2.3 Compositions

Suppose that $f, g \in \mathcal{A}_r(E,E)$ and $g(0) = 0$. The Taylor series of their composition can be expanded as

$$(f \circ g)(x) = \sum_{k\geq 0} \frac{1}{k!} D^k (f \circ g)(x, \ldots, x) = \sum_{k\geq 0} (f \circ g)_k (x, \ldots, x),$$

on one hand. On the other hand it can also be expanded as

$$\begin{aligned} f\circ g(x) &= (\sum_{k\geq 0} f_k(\cdot,\dots,\cdot))\circ(\sum_{l\geq 0} g_l(x,\dots,x)) \\ &= \sum_{k\geq 0} f_k\left(\sum_{l_1\geq 0} g_{l_1}(x,\dots,x),\dots,\sum_{l_k\geq 0} g_{l_k}(x,\dots,x)\right) \\ &= \sum_{k\geq 0}\sum_{l_1\geq 0}\dots\sum_{l_k\geq 0} f_k\left(g_{l_1}(x,\dots,x),\dots,g_{l_k}(x,\dots,x)\right) \\ &= \sum_{k\geq 0}\sum_{n\geq 0}\sum_{l_1+\ldots+l_n=k} f_n\left(g_{l_1}(x,\dots,x),\dots,g_{l_n}(x,\dots,x)\right). \end{aligned}$$

This suggests the following definition for the composition of formal power series:

Definition 2.8 *Let $f=\sum_{k\geq 0} f_k\in\mathcal{P}(E,F)$, $g=\sum_{k\geq 1} g_k\in\mathcal{P}(D,E)$, then we define the formal composition $f\circ g$ as*

$$f\circ g=\sum_{k\geq 0}\sum_{n\geq 0}\left(\sum_{l_1+\ldots+l_n=k} f_n(g_{l_1},\dots,g_{l_n})\right).$$

We make some useful estimates on the composition of functions.

Lemma 2.9 *Let D, E and F be Banach spaces over $\mathbb{C}$, $f\in\mathcal{A}_r(E,F)$ and $g\in\mathcal{A}_\eta(D,E)$ with $||g||_\eta\leq r$, $g(0)=0$. Then $f\circ g\in\mathcal{A}_\eta(D,F)$ and $||f\circ g||_\eta\leq||f||_r$.*

Proof: Let $f(x)=\sum_{k\geq 0} a_k(x^k)$ and $g(x)=\sum_{l\geq 1} b_l(x^l)$. Their composition is defined as

$$(f\circ g)=\sum_{k\geq 0}\sum_{n\geq 0}\left(\sum_{l_1+\ldots+l_n=k} a_n(b_{l_1},b_{l_2},\dots,b_{l_n})\right)$$

Since

$$\begin{aligned} ||a_n\left(b_{l_1},b_{l_2},\dots,b_{l_n}\right)(\underbrace{x,\dots,x}_{k})|| &= ||a_n(b_{l_1}(\underbrace{x,\dots,x}_{l_1}),b_{l_2}(\underbrace{x,\dots,x}_{l_2}),\dots,b_{l_n}(\underbrace{x,\dots,x}_{l_n}))|| \\ &\leq ||a_n||\,||b_{l_1}||\dots||b_{l_n}||\,||x||^k. \end{aligned}$$

Hence

$$||a_n\left(b_{l_1},b_{l_2},\dots,b_{l_n}\right)||\leq||a_n||\,||b_{l_1}||\dots||b_{l_n}||.$$

It follows that

$$\begin{aligned}||f\circ g||_\eta &\le \sum_{k\ge 0}\left(\sum_{n\ge 0}\sum_{l_1+\ldots+l_n=k}||a_n||||b_{l_1}||||b_{l_2}||\ldots||b_{l_n}||\right)\eta^k\\ &=\sum_{k\ge 0}\sum_{n\ge 0}\sum_{l_1+\ldots+l_n=k}||a_n||\left(||b_{l_1}||\eta^{l_1}||b_{l_2}||\eta^{l_2}\ldots||b_{l_n}||\eta^{l_n}\right)\\ &=\sum_{n\ge 0}||a_n||\left(\sum_{l\ge 0}||b_l||\eta^l\right)^n \le \sum_{n\ge 0}||a_n||r^n=||f||_r.\end{aligned}$$

□

Lemma 2.10 *The composition operator*

$$O:\mathcal{A}_r(E,F)\times B_r\to\mathcal{A}_\eta(D,F):(f,g)\mapsto f\circ g$$

is continuous. Here $B_r:=\{g\in\mathcal{A}_\eta(D,E)|\,g(0)=0$ *and* $||g||_\eta<r\}$

Proof: We have

$$||O(f+f',g)-O(f,g)||_\eta=||O(f',g)||_\eta\le||f'||_r$$

which proves uniform continuity of O in its first argument. It is now sufficient to prove that it is continuous in its second argument. Therefore take $f\in\mathcal{A}_r(E,F))$ and $g\in B_r$. Suppose that $||g||_\eta=\alpha<r$. Define $\beta:=(r-\alpha)/3$. Then we know by Lemma 2.7 that $D^kf\in\mathcal{A}_{r-2\beta}\left(E,\mathcal{L}^k\left(E,F\right)\right)$ and $||D^kf||_{r-2\beta}\le k!||f||_r(2\beta)^{-k}$. Let $h\in B_r$ with $||h||_\eta<\beta$; then

$$\begin{aligned}||O(f,g+h)-O(f,g)||_\eta &=||\sum_{k\ge 1}\sum_{n\ge 0}\sum_{l_1+\ldots+l_n=k}f_n(g_{l_1},\ldots,g_{l_n})\\ &\qquad-\sum_{k\ge 1}\sum_{n\ge 0}\sum_{l_1+\ldots+l_n=k}f_n(g_{l_1}+h_{l_1},\ldots,g_{l_n}+h_{l_n})||_\eta\\ &=||\sum_{k\ge 1}\sum_{n\ge 0}\sum_{j=0}^{n}\sum_{|l|+|l'|=k}\binom{n}{j}f_n(g_{l_1},\ldots,g_{l_{n-j}},h_{l'_1},\ldots,h_{l'_j})||_\eta\\ &=||\sum_{k\ge 1}\frac{D^kf\circ g}{k!}(\underbrace{h,\ldots,h}_{k})||_\eta.\end{aligned}$$

Hence it follows that

$$||O(f,g+h)-O(f,g)||_\eta\le\frac{||f||_r||h||_\eta}{\beta}.$$

□

Proposition 2.11 *Let $B_r := \{g \in \mathcal{A}_\eta(D,E) |\, g(0) = 0$ and $||g||_\eta < r\}$, then the composition operator*

$$O : \mathcal{A}_r(E,F) \times B_r \to \mathcal{A}_\eta(D,F) : (f,g) \mapsto f \circ g$$

is C^1.

Proof: We show that O has continuous partial derivatives of first order. Therefore, let $f \in \mathcal{A}_r(E,F)$ and $g \in B_\eta$. Suppose that $||g||_\eta = \alpha < r$. Define $\beta := (r-\alpha)/3$ and suppose that $||h||_\eta < \beta$. A similar calculation as in the previous lemma shows that

$$O(f,g+h) - O(f,g) - (D^1 f \circ g)(h) = \sum_{k \geq 2} \frac{D^k f \circ g}{k!} (\underbrace{h, \ldots, h}_{k}),$$

From the estimate

$$||D^k f||_{r-2\beta} \leq \frac{k! ||f||_r}{(2\beta)^k},$$

it follows that

$$||O(f,g+h) - O(f,g) - (Df \circ g).h||_\eta \leq \sum_{k \geq 2} \frac{||f||_r ||h||_\eta^k}{(2\beta)^k} \leq \frac{2||f||_r ||h||_\eta^2}{(2\beta)^2} \leq \frac{||f||_r ||h||_\eta^2}{(\beta)^2}.$$

Hence $Df \circ g$ is the partial derivative of O with respect to its second variable. Its continuity follows as a consequence of Lemma 2.10. The partial derivative of O with respect to its first variable is easier since O is linear in this variable. Hence it follows from

$$(f + f') \circ g - f \circ g = f' \circ g.$$

that $D_1(O)(f,g).\widetilde{f} = O(\widetilde{f}, g)$. The continuity of this partial derivative follows from Lemma 2.10. Since O has continuous first order partial derivatives, it is C^1. □

2.2.4 Direct sum splitting of an analytic function.

Let X be a Banach space and let E be a Banach space that is a direct sum of the Banach spaces E_1, E_2, …, E_n. Then an element x of $E = E_1 \oplus \cdots \oplus E_n$ can be written in a unique way as $x = \pi_1(x) + \ldots + \pi_n(x) = x_1 + \ldots + x_n$, with $x_i \in E_i$, and where $\pi_i : E \to E_i$ is the projection on the i-th component. Let $f_k \in \mathcal{L}^k(E,X)$. Like in $\mathbb{C}^n$ we develop f_k in homogeneous polynomials of degree k. With the use of the multinomium of Newton, it is readily verified that

$$f_k((x_1 + \ldots + x_n)^k) = \sum_{\substack{l \in \mathbb{N}^n, \\ |l| = k}} \binom{k}{l} f_k(x_1^{l_1} x_2^{l_2} \ldots x_n^{l_n}),$$

where $\binom{k}{l} = \frac{k!}{l_1! \dots l_n!}$ are the multinomial coefficients and $|l| = l_1 + \dots + l_n$. Note that in the formula above we used the power notations $x^k = (\underbrace{x, \dots, x}_{k})$ for $k \in \mathbb{N}$ and

$$x^l = x_1^{l_1} \dots x_n^{l_n} = (\underbrace{x_1, \dots, x_1}_{l_1}, \dots, \underbrace{x_n, \dots, x_n}_{l_n})$$

for $l = (l_1, \dots, l_n) \in \mathbb{N}^n$. Define now for each $l = (l_1, \dots, l_n) \in \mathbb{N}^n$

$$f_l := \binom{|l|}{l} f_{|l|} \circ (\underbrace{\pi_1, \dots, \pi_1}_{l_1}, \dots, \underbrace{\pi_n, \dots, \pi_n}_{l_n}),$$

then clearly $f_l \in \mathcal{L}^{|l|}(E, X)$. Furthermore $f_k = \sum_{\substack{l \in \mathbb{N}^n, \\ |l|=k}} f_l$ and a general $f = \sum_{k \in \mathbb{N}} f_k \in \mathcal{P}(E, X)$, can be decomposed as $f = \sum_{l \in \mathbb{N}^n} f_l$. If X also admits a direct sum splitting $X_1 \oplus \dots \oplus X_m$, then we can further split this function into its components, then this formula becomes

$$f = \sum_{i=1}^{m} \sum_{l \in \mathbb{N}^n} f_l^i,$$

where $f_l^i = \pi_i \circ f_l$. As an analogy to the situation in $\mathbb{C}^n$, we will refer to f_l^i as a term (monomial) in x^l or as a term (monomial) with degree l.

2.2.5 A class of formal (semi-)groups in $\mathcal{P}(E, E)$

Suppose that E is a Banach space with direct sum splitting $E = E_1 \oplus \dots \oplus E_n$ and suppose that $f \in \mathcal{P}(E, E)$. As explained in Section 2.2.4, we can decompose f as

$$f = \sum_{i=1}^{n} \sum_{k \in \mathbb{N}^n} f_k^i.$$

Let $K \subset \mathbb{N}^n$, and define $\mathcal{P}_K(E, E)$, the set of formal series adapted to K, as

$$\mathcal{P}_K(E, E) := \{f \in \mathcal{P}(E, E) \mid f = \sum_{i=1}^{n} \sum_{k \in K} f_{e_i+k}^i\},$$

where $e_i = (0, \dots, 1, \dots, 0)$ is the i-th unit vector. Intuitively, the term $f_{e_i+k}^i$, where $k \in K$, corresponds to a term $x_i x^k = x_i x_1^{k_1} \dots x_n^{k_n}$ in the classical Taylor series. Note that this implies that if $l = (\dots, 0, \dots)$, with a zero at the i-th entry, then $f_l^i = 0$ or, equivalently, each E_i is an invariant subspace.

With respect to the composition of maps it is natural to require that the subset K of $\mathbb{N}^n$ is a semi-group, i.e. for every $k_1, k_2 \in K$ also $k_1 + k_2 \in K$. We shall call a semi-group in $\mathbb{N}^n$ a *cone*, cf. figure 2.1.

Lemma 2.12 *Let $K \subset \mathbb{N}^n$ be a cone. Then $\mathcal{P}_K(E,E)$ forms a semi-group under composition.*

Proof: Let K be a cone and let g and h be elements of $\mathcal{P}_K(E,E)$. We show that their composition $g \circ h$ remains in $\mathcal{P}_K(E,E)$. Since on the formal level the composition is defined as

$$\sum_{i=1}^{n} \sum_{k \in \mathbb{N}^n} g_k^i \circ \sum_{i=1}^{n} \sum_{k \in \mathbb{N}^n} h_k^i = \sum_{i=1}^{n} \sum g_k^i(h^1_{l^1_1}, \ldots, h^1_{l^1_{k_1}}, \ldots, h^n_{l^n_1}, \ldots, h^n_{l^n_{k_1}}),$$

where the sum ranges over all indices k and l_i^j for which $l^i_1 + \ldots + l^i_{k_n} = k_i$, for each $1 \leq i \leq n$. We consider a general term in the composition. Such a term appearing in the formal composition looks like:

$$g_k^i(h^1_{l^1_1}, \ldots, h^1_{l^1_{k_1}}, h^2_{l^2_1}, \ldots, h^2_{l^2_{k_2}}, \ldots, h^n_{l^n_1}, \ldots, h^n_{l^n_{k_n}}) \,. \tag{2.9}$$

Since $g \in \mathcal{P}_K(E,E)$, it follows that $k = e_i + \tilde{k}$; where $\tilde{k} \in K$. Furthermore, since also $h \in \mathcal{P}_K(E,E)$; it follows that for each $h^\alpha_{l^\alpha_\beta}$ we have that $l^\alpha_\beta = e_\alpha + m^\alpha_\beta$ where $m^\alpha_\beta \in K$. The term given by (2.9) is clearly a term of degree

$$\begin{aligned} l^1_1 + \ldots + l^n_{k_n} &= e_1 + m^1_1 + \ldots + e_1 + m^1_{k_1} + \ldots + m^n_{k_n} \\ &= \underbrace{e_1 + \ldots + e_1}_{k_1} + \ldots + \underbrace{e_n + \ldots + e_n}_{k_n} + m^1_1 + \ldots + m^1_{k_1} + \ldots + m^n_{k_n} \\ &= k + \gamma, \end{aligned}$$

where $\gamma = m^1_1 + \ldots m^n_{k_n} \in K$ since K is a semi-group and $k = e_i + \tilde{k}$. Hence $k + \gamma = e_i + \tilde{k} + \gamma = e_i + \hat{k}$, where $\tilde{k} + \gamma = \hat{k} \in K$. Since this is an arbitrary term, it follows that the composition $g \circ h \in \mathcal{P}_K(E,E)$.

□

Let $\mathcal{D}_K(E,E)$ be the subset of $\mathcal{P}_K(E,E)$ consisting of elements that have an invertible linear part, we show that this set forms a group if $K \neq \emptyset$.

Lemma 2.13 *Let $K \subset \mathbb{N}^n$, $K \neq \emptyset$ be a cone, then $\mathcal{D}_K(E,E)$ forms a group under composition.*

Proof: We only have to prove that if $h \in \mathcal{D}_K(E,E)$ then also $h^{-1} \in \mathcal{D}_K(E,E)$. Suppose that

$$h = \sum_{i=1}^{n} \sum_{l \in \mathbb{N}^n} h^i_l \in \mathcal{D}_K(E,E)$$

is given, we define its formal inverse

$$g = \sum_{i=0}^{n} \sum_{l \in \mathbb{N}^n} g_l^1$$

by induction on $|l|$. For $|l| = 1$, we define $g_1 := h_1^{-1}$; this defines g_l^i with $|l| = 1$. Suppose that all g_l^i are defined for $|l| \leq N-1$ and satisfy the property $l \notin e_i + K \Rightarrow g_l^i = 0$. We show that the same is true for $|l| = N$.

Therefore suppose $l \in \mathbb{N}^n$ and $i \in \{1, \dots, n\}$ are fixed with $|l| = N$. As in Lemma 2.12, we observe that any term with degree l appearing in the formal composition has the form

$$g_k^i(h^1_{l^1_1}, \dots, h^1_{l^1_{k_1}}, h^2_{l^1_1}, \dots, h^2_{l^1_{k_2}}, \dots, h^n_{l^n_1}, \dots, h^2_{l^n_{k_n}}), \tag{2.10}$$

where $l^1_1 + \ldots + l^n_{k_n} = l$. Hence, apart from the term,

$$g_l^i(h^1_{e_1}, \dots, h^n_{e_n}),$$

all terms with degree l are already defined: it contains only h's and g_i^k with $|k| < |l|$. Define g_l^i in such a way that the inversion relation $g \circ h = \mathrm{id}$ is satisfied. More precisely

$$g_l^i(h^1_{e_1}, \dots, h^n_{e_n}) = -\sum_{\substack{(l^1_1,\dots,l^n_{k_n}) \\ \neq (e_1,\dots,e_n)}} g_k^i(h^1_{l^1_1}, \dots, h^1_{l^1_{k_1}}, h^2_{l^1_1}, \dots, h^2_{l^1_{k_2}}, \dots, h^n_{l^n_1}, \dots, h^2_{l^n_{k_n}}). \tag{2.11}$$

Suppose now that $l \notin e_i + K$ and consider an arbitrary term from the sum in the right hand side of (2.11). Such a term is non-zero if and only if each $h^\alpha_{l^\alpha_\beta} \neq 0$ and $g_k^i \neq 0$. Since

$$\begin{aligned} h^\alpha_{l^\alpha_\beta} \neq 0 &\implies l^\alpha_\beta = m^\alpha_\beta + e_\alpha \in e_\alpha + K \\ g_k^i \neq 0 &\implies k \in e_i + K, \end{aligned}$$

the corresponding term has index

$$\begin{aligned} l = l^1_1 + \ldots + l^n_{k_n} &= e_1 + m^1_1 + \ldots + e_1 + m^1_{k_1} + \ldots + m^n_{k_n} \\ &= \underbrace{e_1 + \ldots + e_1}_{k_1} + \ldots + \underbrace{e_n + \ldots + e_n}_{k_n} + m^1_1 + \ldots + m^1_{k_1} + \ldots + m^n_{k_n} \\ &= k + \gamma, \end{aligned}$$

where $\gamma \in K$ and $k = e_i + \widetilde{k} \in e_i + K$. It follows, using the semi-group property of K that $l \in e_i + K$, a contradiction. □

We now define for each cone K the subspaces $\mathcal{D}_{K,r}(E,E) = \mathcal{A}_r(E,E) \cap \mathcal{D}_K(E,E)$ of $\mathcal{A}_r(E,E)$. If the cone $K = \mathbb{N}^n$, then we use the notation $\mathcal{D}_r(E,E)$. Note that each $F \in \mathcal{D}_{K,r}(E,E)$ has a linear part A which can be split as in Subsection 2.2.4, i.e.

$$A = A^1 \oplus \cdots \oplus A^n, \tag{2.12}$$

then the same holds for compositions in $\mathcal{D}_{K,r}(E,E)$.

2.3 Proof of the main results

We use notations from Section 2.2, in particular we will use $\mathcal{D}_{K,r}(E,E)$ and $\mathcal{D}_K(E,E)$ introduced there. The main result, Theorem 2.1 is a consequence of the following proposition.

Proposition 2.14 *Let $r > 0$, K be a cone and let $F \in \mathcal{D}_{K,r}(E,E)$ be an analytic diffeomorphism on the Banach space E. Suppose that E admits a direct sum decomposition $E = E_1 \oplus \ldots \oplus E_n$ and suppose that each E_i is an invariant subspace for F. Then an $\widetilde{r} > 0$ and an analytic near-identity coordinate transform $U \in \mathcal{D}_{K \cap G_{D,C}, \widetilde{r}}(E,E)$ exist, such that $G = U^{-1} \circ F \circ U \in \mathcal{D}_{K \cap B_{D,C}, \widetilde{r}}(E,E)$.*

The proof of Proposition 2.14 consists of two steps. The first step serves to remove terms in a somewhat smaller good set

$$G_D = \{k \in \mathbb{N}^m \mid ||A||^k \leq D^{|k|}\}, \tag{2.13}$$

where $0 < \rho D < 1$. The corresponding bad set is

$$B_D = \{k \in \mathbb{N}^m \mid ||A||^k > D^{|k|}\}. \tag{2.14}$$

The removal of 'bad terms' is reflected in the following proposition:

Proposition 2.15 *Let F be as in Proposition 2.14. Then there exists an $\widetilde{r} > 0$ and an analytic near-identity coordinate transform $U \in \mathcal{D}_{K \cap G_D, \widetilde{r}}(E,E)$, such that $G = U^{-1} \circ F \circ U \in \mathcal{D}_{K \cap B_D, \widetilde{r}}(E,E)$.*

The second step consist of repeating the same idea for F^{-1}, and requires that at the formal level 'we do not introduce already removed terms'. This is actually the entire idea of Section 2.2.5.

As already explained in Section 2.1.4, our intention is to solve $\mathcal{F}(f,g,u) = 0$ for g and u for a given map f, where the functional $\mathcal{F}$ was defined in (2.8) and which we recall for the convenience of the reader.

$$\mathcal{F} : V \times W \times X \to Z : (f,g,u) \mapsto A \circ u - u \circ (A + g) + f \circ (\mathrm{id} + u) - g$$

To solve this functional equation we use an appropriate version of the implicit function theorem, which we state.

Theorem 2.16 (Implicit Function Theorem) *Let V, W, X be open neighbourhoods of the origin in the Banach spaces $\overline{V}$, $\overline{W}$, $\overline{X}$ and let $\overline{Z}$ be a Banach space. Suppose that $\mathcal{F} : V \times W \times X \to Z$ is C^1, $\mathcal{F}(0,0,0) = 0$ and that*

$$D_{(g,u)}\mathcal{F} : \overline{W} \times \overline{X} \to \overline{Z} : (g,u) \mapsto D\mathcal{F}(0,0,0).(0,g,u)$$

is an isomorphism of Banach spaces. Then there exists open neighbourhoods $V_1 \subset V$, $W_1 \subset W$, $X_1 \subset X$ of the origin, such that for each $f \in V_1$ there exists a unique $(g,u) \in W_1 \times X_1$ with $\mathcal{F}(f,g,u) = 0$.

Let us introduce appropriate Banach spaces and well chosen open subsets of them.

Definition 2.17 *The Banach spaces $\overline{V}$, $\overline{W}$, $\overline{X}$ and $\overline{Z}$ and their corresponding open parts V, W, X and Z are defined as follows*

$$\begin{aligned}
\overline{V} = V &= \Big\{ f \in \mathcal{A}_{K,2r}(E,E) | f_0 = 0, f_1 = 0 \Big\}, \\
\overline{W} &= \Big\{ g \in \mathcal{A}_{K,r}(E,E) | g = \sum_{j=1}^{n} \sum_{k \in B_D, |k| \geq 1} g^j_{k+e_j}, \Big\}, \\
W &= \Big\{ g \in \overline{W} \,|\, \left(\frac{||g^j||_r}{||A^j||} \right) < (1-D)r, \; \textit{for each } j = 0,1,\ldots,n \Big\}, \\
\overline{X} &= \Big\{ u \in \mathcal{A}_{K,r}(E,E) | u = \sum_{j=1}^{n} \sum_{k \in G_D, |k| \geq 1} u^j_{k+e_j}, \Big\}, \\
X &= \Big\{ u \in \overline{X} \,|\, ||u||_r < r \Big\}, \\
\overline{Z} = Z &= \Big\{ h \in \mathcal{A}_{K,r}(E,E) | h_0 = 0, h_1 = 0 \Big\}.
\end{aligned}$$

Three crucial points in the proof are: (1) the fact that $\mathcal{F}$ is well defined, (2) the continuous differentiability of the functional $\mathcal{F}$ and (3) the fact that its derivative is an isomorphism. We state these points as lemmas and prove them.

Lemma 2.18 *The functional $\mathcal{F}$ is C^1, in particular its Gâteaux derivatives are continuous.*

Proof

Since $||A \circ u|| \leq ||A||||u||$, it follows that the part $(f,g,u) \mapsto A \circ u$ is C^1, it is also clear that the part $(f,g,u) \mapsto -g$ is C^1. Because $||id + u||_r \leq ||id||_r + ||u||_r < 2r$, it follows directly from Proposition 2.11 that $(f,g,u) \mapsto f \circ (\text{id} + u)$ is C^1. The part

$(f, g, u) \mapsto u \circ (A + g)$ is more difficult. First let's make a short calculation:

$$\begin{aligned} u^j \circ (A+g) &= \sum_{k \in G_D} u^j_{k+e_j} (A+g)^{k+e_j} \\ &= \sum_{(k,j) \in G_D} u^j_{k+e_j} \left((A^j + g^j), (A^1+g^1)^{k_1}, \ldots, (A^n+g^n)^{k_n}\right) \\ &= \sum_{(k,j) \in G_D} \frac{||A^j|| ||A^1||^{k_1} \ldots ||A^n||^{k_n}}{D^{|k|}} \times \end{aligned}$$

$$u^j_{k+e_j} \left(\left(\frac{DA^j}{||A^j||} + \frac{D}{||A^j||} g^j \right), \left(\frac{DA^1}{||A^1||} + \frac{D}{||A^1||} g^1 \right)^{k_1}, \ldots, \left(\frac{DA^n}{||A^n||} + \frac{D}{||A^n||} g^n \right)^{k_n} \right).$$

The map

$$u = \sum_{j=1}^{n} \sum_{k \in G_D} u^j_{k+e_j} \mapsto u' := \sum_{j=1}^{n} \sum_{k \in G_D} \frac{||A^j|| ||A^1||^{k_1} \ldots ||A^n||^{k_n}}{D^{|k|}} u^j_{k+e_j}$$

is clearly linear. It is also continuous since

$$\sum_{k \in G_D} \frac{||A^j|| ||A^1||^{k_1} \ldots ||A^n||^{k_n}}{D^{|k|}} ||u^j_{k+e_j}|| r^k \leq \sum_{k \in G_D} ||A||_{\sup} ||u^j_{k+e_j}|| r^k \leq ||A||_{\sup} ||u^j||_r.$$

Here $||A||_{\sup} := \max_{j \in \{1,\ldots,n\}} ||A^j||$. We use Proposition 2.11 a second time, finding that

$$(u'^j, \left(\sum_{i=1}^{n} \frac{DA^i}{||A^i||} x + \frac{D}{||A^i||} g^i \right)) \mapsto u'^j \circ \left(\sum_{i=1}^{n} \frac{DA^i}{||A^i||} x + \frac{D}{||A^i||} g^i \right)$$

is C^1. This is justified since

$$|| \left(\frac{DA^i}{||A^i||} + \frac{D}{||A^i||} g^i \right) ||_r < Dr + (1-D)r = r.$$

Hence this mapping is C^1. Adding the individual C^1 pieces finishes the proof. □

We calculate the Gâteaux derivatives and find:

$$\begin{aligned} D_u \mathcal{F}(0,0,0).u &= \lim_{t \to 0} \frac{A \circ tu - tu \circ A}{t} = A \circ u - u \circ A, \\ D_f \mathcal{F}(0,0,0).f &= \lim_{t \to 0} \frac{tf \circ \mathrm{id}}{t} = f, \\ D_g \mathcal{F}(0,0,0).g &= \lim_{t \to 0} \frac{-tg}{t} = -g. \end{aligned} \tag{2.15}$$

Using these derivatives, we are ready to prove

Lemma 2.19 $D_{(g,u)}\mathcal{F}(0,0,0) : \overline{W} \times \overline{X} \to \overline{Z} : (g,u) \mapsto D\mathcal{F}(0,0,0).(0,g,u)$ *is an isomorphism.*

Proof We split $D_{(g,u)}\mathcal{F}(0,0,0)$ in its 'good' and its 'bad' part. Since $A^j \circ u = \sum_{k\in G_D} A^j \circ u^j_{k+e_j}$ and $u^j \circ A = \sum_{k\in G_D} u^j_{k+e_j} \circ (A,\ldots,A)$, it follows that the projection on the good and bad cone yield $\pi_{G_D}(A\circ u - u\circ A - g) = A\circ u - u\circ A$ and $\pi_{B_D}(A\circ u - u\circ A - g) = -g$. Hence, in order to show that $D_{(g,u)}\mathcal{F}(0,0,0)$ is an isomorphism, it is sufficient to show that $\mathcal{G}_1 : W \to W : g \mapsto -g$ and $\mathcal{G}_2 : X \to X : u \mapsto (A\circ u - u \circ A)$ are isomorphisms. It is clear that $\mathcal{G}_1$ is an isomorphism. It remains to show that $\mathcal{G}_2$ is an isomorphism. Now

$$\begin{aligned} A\circ u - u\circ A &= \sum_{i=1}^n A^i \circ u^i - u^i \circ A = \sum_{i=1}^n A^i \left(u^i - (A^i)^{-1}\circ u^i \circ A\right) \\ &= A^i \sum_{i=1}^n (\mathrm{id} - R_i)\,(u^i), \end{aligned}$$

where $R_i : \overline{X} \to \overline{X} : u \mapsto (A^i)^{-1} \circ u^i \circ A$. If we can show that $||R_i|| < 1$, then it follows that $id - R_i$ and hence also $A^i\,(\mathrm{id} - R_i)$ is an isomorphism, which completes the proof. It remains to show that $||R_i|| < 1$. This is true since

$$\begin{aligned} ||(A^i)^{-1}\circ u^i \circ A||_r &= \sum_{k\in G_D} ||(A^i)^{-1} \circ u^i_{k+e_i}(A^i, \underbrace{A^1,\ldots,A^1}_{k_1},\ldots,\underbrace{A^n,\ldots,A^n}_{k_n})||r^{|k|+1} \\ &\le \sum_{k\in G_D} ||(A^i)^{-1}||\,||u^i_{k+e_i}||\,||A^i||\,||A^1||^{k_1}\ldots||A^n||^{k_n} r^{|k|+1} \\ &\le \sum_{k\in G_D} \rho D^k ||u^i_{k+e_i}||r^{|k|+1} \le \rho D \sum_{k\in G_D} ||u^i_{k+e_i}||r^{|k|+1} \le \rho D ||u^i||_r, \end{aligned}$$

and since, by assumption, $\rho D < 1$. □

We are in a position to prove Proposition 2.15.

Proof of Proposition 2.15 According to Theorem 2.16, with the help of Lemma 2.19, there exists a $r > 0$ such that this theorem is true for all $F = A + f$ with $||f||_{2r} < r$. Suppose that $||f||_{2r} \ge r$. We apply classical rescaling. Choose $0 < \gamma < 1$ such that $\widetilde{f} = \gamma^{-1} f \circ (\gamma \mathrm{id}) = \gamma^{-1} \sum_{(k,j)\in\mathbb{N}^2\times\{1,\ldots,n\}} \gamma^{|k|} f^j_k$ has a norm $||\widetilde{f}|| < r$. Let $\widetilde{u}$, $\widetilde{g}$ be the solution of the equation $\mathcal{F}(\widetilde{f},\widetilde{g},\widetilde{u}) = 0$ and define $u := \gamma\widetilde{u} \circ (\gamma^{-1}\mathrm{id})$ and $g := \gamma\widetilde{g}\circ(\gamma^{-1}\mathrm{id})$. Then it is clear that

$$0 = \gamma\mathcal{F}(\widetilde{f},\widetilde{g},\widetilde{u})\circ(\gamma^{-1}\mathrm{id}) = \mathcal{F}(f,g,u).$$

This concludes the proof of this proposition. □

As a corollary we can complete the proof of our main result.

Proof of Proposition 2.14 The proof is done in two steps.

Step 1 We first invert F and then apply Theorem 2.15 to F^{-1}, with $K = \mathbb{N}^n$. Note that F^{-1} corresponds to the same factor ρ as F, since reversing the roles of A and A^{-1} does not alter the value of ρ. Hence we know that the reduction G does not contain any term outside the cone

$$B_C = \{k \in \mathbb{N}^m \mid ||A^{-1}||^k > C^{|k|}\}.$$

Using Lemma 2.13, we see that the same is true for G^{-1}, since B_C is conic.
Step 2 We rename G^{-1}, our previous reduction from step 1, again to F. Then F contains only terms in the cone $K = B_C$, it follows that there exists a reduction to a certain G containing only terms in the cone $K \cap B_D = B_C \cap B_D = B_{D,C}$. □

2.4 An invariant manifold theorem

Using similar techniques as in the previous section, we obtain the well-known stable and unstable manifold theorems for analytic diffeomorphisms, as well as the smooth dependence on possible parameters. It is the topic of this section to explain these ideas.

Let us first describe the situation in $\mathbb{C}^2$ in order to sketch the ideas for the reader. Suppose that $F(x, y) = (\lambda_1 x, \lambda_2 y) + O(||(x, y)||^2)$ is given, where $|\lambda_1| < 1$, $|\lambda_2| > 1$, and we want to find a stable manifold for F. We could then try to find a coordinate transform $U = \mathrm{id} + O(||(x, y)||^2)$ such that in new coordinates $G(x, y) = U^{-1} \circ F \circ U(x, y)$ leaves $y = 0$ invariant. This is equivalent to

$$G(x, 0) = \begin{pmatrix} \lambda_1 x + O(||(x, y)||^2) \\ 0 \end{pmatrix}.$$

The inverse image of $y = 0$ is then an invariant (stable) manifold of F. This is precisely what we will do in a slightly more general context.

Let $E = E_1 \oplus E_2$ be a direct sum of Banach spaces and $F = A + \sum_{k \geq 2} F_k \in \mathcal{A}_r(E, E)$ with diagonal linear part A. Hence, using the notations of Section 2.2.4, $A = F_1 = F^1_{(1,0)} + F^2_{(0,1)} = A^1 + A^2$. Suppose that $||A^1|| < 1$ and $||(A^2)^{-1}|| < 1$. Choose $||A^1|| < D < 1$ and define the bad set

$$BS := \{(k, j) \in \mathbb{N}^2 \times \{1, 2\}, |k| = k_1 + k_2 \geq 2 \,|\, (k, j) \neq ((k_1, 0), 2)\},$$

i.e. if $(k, 2) \in BS$, then $k_2 \geq 1$; and the good set, the set of terms that we intend to remove (see below for more details),

$$GS := \{(k, j) \in \mathbb{N}^2 \times \{1, 2\}, |k| = k_1 + k_2 \geq 2 \,|\, (k, j) = ((k_1, 0), 2)\}.$$

We will look for a coordinate transform $U = id + \sum_{(k,j) \in GS} u^j_k$ containing only good terms, that conjugates F to $G = U^{-1} \circ F \circ U$, such that $G = A + \sum_{(k,j) \in BS} g^j_k$ contains only bad terms. Here the set of bad terms is chosen exactly as the set of terms that are still left (i.e. unremoved) in the Taylor expansion of G, thus note that when G contains only bad terms, then, for $x_1 \in E_1$

$$\pi_2 \circ G(x_1) = 0,$$

because $k_2 \geq 1$ if $j = 2$. Hence G leaves E_1 invariant. As explained in the introduction, this problem is equivalent to finding a zero of the functional equation

$$\mathcal{F} : V \times W \times X \to Z : (f, g, u) \mapsto A \circ u - u \circ (A + g) + f \circ (\mathrm{id} + u) - g,$$

a problem that we can solve in a similar way as in Section 2.3. We denote

$$\begin{aligned}
\overline{V} &= V = \{f \in \mathcal{A}_{2r}(E, E) | f_0 = 0, f_1 = 0\} \\
\overline{W} &= \{g \in \mathcal{A}_r(E, E) | g = \sum_{(k,j) \in BS} u_k^j\} \\
W &= \{g \in \overline{W} | \, ||g||_r < (1 - D)r\} \\
\overline{X} &= \{u \in \mathcal{A}_r(E, E) | u = \sum_{(k,j) \in GS} u_k^j\} \\
X &= \{u \in \overline{X} | \, ||u||_r < r\} \\
\overline{Z} &= Z = \{h \in \mathcal{A}_r(E, E) | h_0 = 0, h_1 = 0\}
\end{aligned}$$

Lemma 2.20 *$\mathcal{F}$ is C^1.*

Proof: We use the same technique as in Lemma 2.18.
Since $||A \circ u|| \leq ||A|| ||u||$, it follows that the part $(f, g, u) \mapsto A \circ u$ is C^1, it is also clear that the part $(f, g, u) \mapsto -g$ is C^1. Because $||id + u||_r \leq ||id||_r + ||u||_r < 2r$, it follows directly from Proposition 2.11 that $(f, g, u) \mapsto f \circ (\mathrm{id} + u)$ is C^1. We take a closer look at the composition

$$\begin{aligned}
u \circ (A + g) &= \sum_{(k,1) \in GS} u_k^1(\underbrace{A + g, \ldots, A + g}_{|k|}) + \sum_{(k,2) \in GS} u_k^2(\underbrace{A + g, \ldots, A + g}_{|k|}) \\
&= \sum_{(k,2) \in GS} u_k^2(\underbrace{A + g, \ldots, A + g}_{|k|}) = \sum_{k_1 \geq 2} u_{(k_1,0)}^2(\underbrace{A^1 + g^1, \ldots, A^1 + g^1}_{k_1}).
\end{aligned}$$

Since $||A^1 + g^1||_r \leq ||A^1||r + ||g^1||_r < (D + (1 - D))r = r$, we can use Proposition 2.11 to conclude that $u^2 \circ (A^1 + g^1)$ is C^1. Since the projections $u \mapsto u^2$ and $g \mapsto g^1$ are C^1, it follows that the composition $(u, g) \mapsto (u^2, g^1) \mapsto u^2 \circ (A^1 + g^1) = u \circ (A + g)$ is also C^1. By adding the individual C^1 pieces we obtain a C^1 function F which finishes the proof. □

Lemma 2.21 *We use the same technique as in Lemma 2.19.*
$D_{(g,u)}\mathcal{F}(0, 0, 0) : \overline{W} \times \overline{X} \to Z : (g, u) \mapsto D\mathcal{F}(0, 0, 0).(0, g, u)$ is an isomorphism of Banach spaces.

Proof: The differential $D_{(g,u)}\mathcal{F}(0,0,0)$ is given by the same formulas as in (2.15). We split $D_{(g,u)}\mathcal{F}(0,0,0)$ in its good and its bad part. Since $A\circ u = \sum_{(k,j)\in GS} A^j \circ u_k^j$ and $u\circ A = \sum_{(k,j)\in GS} u_k^j \circ (A,\dots,A)$, it follows that k_2 remains 0 in the second components of these parts. Hence

$$\pi_{GS}(A\circ u - u\circ A - g) = A\circ u - u\circ A,$$
$$\pi_{BS}(A\circ u - u\circ A - g) = -g.$$

Hence, in order to show that $D_{(g,u)}\mathcal{F}(0,0,0)$ is an isomorphism, it is sufficient to show that $\mathcal{G}_1 : \overline{W} \to \overline{W} : g \mapsto -g$ and $\mathcal{G}_2 : \overline{X} \to \overline{X} : u \mapsto (A\circ u - u\circ A)$ are isomorphisms. It is clear that $\mathcal{G}_1$ is an isomorphism. It remains to show that $\mathcal{G}_2$ is an isomorphism. Because $u \in X$, it follows that $u = \sum_{(k,1)\in G} u_k^1 + \sum_{(k,2)\in G} u_k^2 = \sum_{k_1\geq 2} u_{(k_1,0)}^2 = u^2$. Hence

$$A\circ u - u\circ A = A^2\circ u^2 - u^2\circ A^1 = (A^2)(u^2 - (A^2)^{-1}\circ u^2\circ A^1) = (A^2)\circ(\mathrm{id} - M)(u^2),$$

where $M : \overline{X} \to \overline{X} : u^2 \mapsto (A^2)^{-1}\circ u^2 \circ A^1$. Because, by Lemma 2.9, $||u^2\circ A^1|| \leq ||u^2||$ for any $u^2 \in \overline{X}$, it follows that

$$\frac{||M(u^2)||}{||u^2||} = \frac{||(A^2)^{-1}\circ u^2\circ A^1||}{||u^2||} \leq \frac{||(A^2)^{-1}||\,||u^2\circ A^1||}{||u^2||} \leq ||(A^2)^{-1}||.$$

Hence

$$||M|| = \sup_{u^2} \frac{||M(u^2)||}{||u^2||} \leq ||(A^2)^{-1}||,$$

where $||(A^2)^{-1}|| < 1$ and it follows that $\mathrm{id} + M$ is an isomorphism. Hence $\mathcal{G}_2$ is also an isomorphism. □

The proof of the next corollary is similar to that of Proposition 2.15.

Corollary 2.22 *Let $F : E \to E$ be an analytic diffeomorphism, $F(0) = 0$, with diagonal linear part $F^1 = A^1 + A^2$. Suppose that $||A^1|| < 1$ and $||(A^2)^{-1}|| < 1$, then there exists a coordinate transform $U : E \to E$, $U = id + u$ with $u = O(||x||^2)$, such that $G = U^{-1}\circ F\circ U$ has the E_1 plane as an invariant manifold or equivalently $G = \sum_{(k,j)\in BS} g_k^j$.*

2.5 Diffeomorphisms in $\mathbb{C}^n$

We explain how Theorem 2.2 follows from Theorem 2.1. We work with parameter dependent analytic power series, where the parameter varies in an open set Λ. More precisely:

Definition 2.23 *We define $\mathbb{C}^n(\Lambda)$ to be the space of power series $\sum_{l\geq 0} f_l(\mu)x^l$, where $f_l(\mu)$ is an analytic (resp. continuous) function of μ in a neighbourhood $\Lambda \subset \mathbb{R}^n$ of μ_0, such that*

$$\sum_{n\geq 0} ||f_n||_\infty r^n < \infty. \tag{2.16}$$

for a certain $r > 0$. We define $\mathbb{C}^n(\Lambda)_r$ as the subset of $\mathbb{C}^n(\Lambda)$ for which (2.16) holds.

It is clear that $\mathbb{C}^n(\Lambda)_r$ is a Banach space.

We only need to check is that for each $0 < D < 1$ there exists a neighbourhood $\widetilde{\Lambda}$ of μ_0 such that $\rho < \frac{1}{D}$. This follows since $\rho = \max_{i=1}^{n} \left\{ \sup_{\lambda\in\widetilde{\Lambda}} |\lambda_i(\mu)|, \sup_{\lambda\in\widetilde{\Lambda}} \frac{1}{|\lambda_i(\mu)|} \right\}$, and the eigenvalues λ_i depend continuously on μ.

2.6 Examples

2.6.1 A $1:-1$ resonant saddle

We reconsider the example from Section 2.1.2 and explain how expression (2.4) can be obtained from the main result. We consider a family F_μ passing through a $1:-1$ resonance in modulus; by this we mean a family

$$F_\mu(x) = \begin{cases} x_1(\lambda_1(\mu) + \sum_{|k|\geq 2} f_k^1(\mu)x^k) \\ x_2(\lambda_2(\mu) + \sum_{|k|\geq 2} f_k^2(\mu)x^k), \end{cases}$$

where

$$\frac{\log(|\lambda_1(\mu_0)|)}{\log(|\lambda_2(\mu_0)|)} = -1, \tag{2.17}$$

and the series are convergent on a sufficient small neighbourhood around the origin. Remark that condition (2.17) concerns the moduli of the eigenvalues: this is necessary in order to apply our main result; omitting the modulus in (2.17) would lead us to questions of a completely different nature, like for example in the case of elliptic fixed points. Using Theorem 2.2, we see that for any $0 < D < 1$ we can conjugate F in an analytic way to a form

$$G(x) = \left(x_1(\lambda_1(\mu) + \sum_{k\in K} g_k^1(\mu)x^k), x_2(\lambda_2(\mu) + \sum_{k\in K} g_k^2(\mu)x^k) \right)$$

where $K = B_{D,D}$ is a cone containing the resonant line. The closer D is chosen to 1, the smaller the cone. Note also that if $f_k^i(\lambda)$ is continuous (resp. analytic on a neighbourhood with fixed radius) and the supremum norm is considered, then also the coefficients $g_k^i(\mu)$ are continuous (resp. analytic on a neighbourhood with fixed

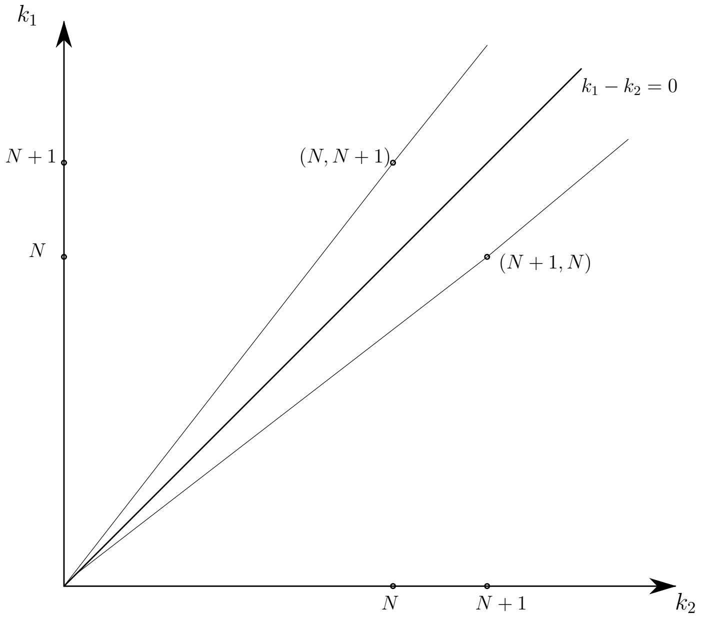


Figure 2.2: The cone of terms which possibly cannot be removed.

radius). Since we supposed a $1 : -1$ resonance in modulus, the main resonant equation at μ_0, is given by

$$(|\lambda_1(\mu_0)|, |\lambda_2(\mu_0|)^k = 1 \Leftrightarrow k_1 - k_2 = 0.$$

Hence we can choose the cones as in Figure 2.2. The only thing we still need to do is describing the terms inside the cone determined by $(N, N+1)$ and $(N+1, N)$. Note that the terms in the upper part of this cone determined by $(N, N+1)$ and $(1, 1)$ correspond to linear combinations of these two vectors

$$r(N, N+1) + s(1, 1) = (A, B),$$

such that r, s are positive real numbers and A, B are natural numbers. Since $A - B = (rN + r + s) - (rN + s) = r$, it follows that r is a natural number. Hence it follows that also $s = A - rN$ is a natural number. It follows that any couple $(A, B) \in \mathbb{N}^2$ in the upper part of this cone can be expressed as $r(N, N+1) + s(1, 1)$, where r, s are natural numbers. In the same way in can be shown that any $(A, B) \in \mathbb{N}^2$ in the lower cone determined by $(N+1, N)$ and $(1, 1)$ can be expressed as $r(N+1, N) + s(1, 1)$, where r, s are natural numbers. Hence

$$G(x) = \left(x_1\left[\lambda_1(\mu) + x_1x_2b_0^1(\mu, x_1x_2) + T_1\right], x_2\left[\lambda_2(\mu) + x_1x_2b_0^2(\mu, x_1x_2) + T_2\right]\right),$$
$$\text{where } T_1 = \sum_{s\geq 1,\, r\geq 0}\left(g^1_{(r,s)}(\mu)(x_1x_2)^r(x_1^Nx_2^{N+1})^s + h^1_{(r,s)}(\mu)(x_1x_2)^r(x_1^{N+1}x_2^N)^s\right)$$
$$\text{and } T_2 = \sum_{s\geq 1,\, r\geq 0}\left(g^2_{(r,s)}(\mu)(x_1x_2)^r(x_1^Nx_2^{N+1})^s + h^2_{(r,s)}(\mu)(x_1x_2)^r(x_1^{N+1}x_2^N)^s\right).$$

or, when putting $u = x_1x_2$, $b^i_s(\mu, u) = \sum_{r\geq 0} g^i_{(r,s)}(\mu)u^r$ and $c^i_s(\mu, u) = \sum_{r\geq 0} h^i_{(r,s)}(\mu)u^r$, we obtain

$$\begin{aligned} G(x,\mu) = \Bigg(& x_1\left[\lambda_1(\mu) + ub_0^1(\mu,u) + \sum_{s\geq 1} u^{Ns}\left(x_2^sb_s^1(\mu,u) + x_1^sb_s^2(\mu,u)\right)\right], \\ & x_2\left[\lambda_2(\mu) + ub_0^2(\mu,u) + \sum_{s\geq 1}\left(u^{Ns}(x_2^sc_s^1(\mu,u) + x_1^sc_s^2(\mu,u)\right)\right]\Bigg). \end{aligned}$$

2.6.2 A $p : -q$ resonant saddle

The situation is quite similar to the one described above. It is however somewhat more technical. Since a similar theorem for vector fields exists, and we give a thorough description in Chapter 3, we shall give a few forward references to lemmas from Chapter 3. This is the main reason to switch temporarily 'κ' for the notation of the eigenvalues to avoid conflict with the 'λ' appearing in the corresponding theorems for vector fields.

We obtain the following theorem:

Theorem 2.24 *Let $F(x, y) = (\kappa_1(\mu)x, \kappa_2(\mu)y) + f_\mu(x, y)$ be an analytic local diffeomorphism, depending on a parameter that varies continuously (resp. analytically) in an open set U containing μ_0. Suppose that $-\frac{\ln(|\kappa_2(\mu_0)|)}{\ln(|\kappa_1(\mu_0)|)} = \frac{p}{q} \in \mathbb{Q}$ and $\frac{p}{q} > 0$ (i.e. we are in a saddle case). Define $(r_0, s_0), (r_1, s_1)$ the unique tuples in $\mathbb{N}^2$ for which $qr_1 - ps_1 = -1$, $qr_0 - ps_0 = 1$, $(r_1, s_1) = (p, q) - (r_0, s_0)$ and $0 \leq r_0 \leq p$. Then for any $N \in \mathbb{N}$ there is a neighbourhood of the parameter U_N, $\mu_0 \in U_N$ and an analytic normal form that can be written as*

$$\begin{aligned} G(x,y,\mu) = \Bigg(& x\left[\kappa_1(\mu) + ub_{l,1}(u,\mu) + \sum_{l\geq 1} u^{Nl}\left(x^{lr_0}y^{ls_0}g_{l,1}(u,\mu) + x^{lr_1}y^{ls_1}h_{l,1}(u,\mu)\right)\right], \\ & y\left[\kappa_2(\mu) + ub_{l,2}(u,\mu) + \sum_{l\geq 1} u^{Nl}\left(x^{lr_0}y^{ls_0}g_{l,2}(u,\mu) + x^{lr_1}y^{ls_1}h_{l,2}(u,\mu)\right)\right]\Bigg). \end{aligned}$$

in this formula $u = x^p y^q$ and the coefficients $g_{l,1}(u,\mu)$, $g_{l,2}(u,\mu)$, $h_{l,1}(u,\mu)$, $h_{l,2}(u,\mu)$, $b_{l,2}(u,\mu)$, $b_{l,1}(u,\mu)$ are analytic functions in u with coefficients depending continuously (resp. analytically) on a parameter μ.

Proof: Upon applying the parameter dependent invariant manifold theorem (e.g. Corollary 2.22), we may suppose that $F(x,y)=(\kappa_1(\mu)x+xf^1_\mu(x,y), \kappa_2(\mu)y+yf^2_\mu(x,y))$. Hence we can for each $D>1$ apply Theorem 2.2, and find a neighbourhood of the parameter V_D, $\mu_0 \in V_D$ for which the normal form is given by

$$G(x,y,\mu) = (\lambda_1(\mu)x, \lambda_2(\mu)y) + (xg^1_\mu(x,y), yg^2_\mu(x,y)),$$

where $g^i_\mu(x) = \sum_{k\in B_{D,D}} g^i_k(\mu)x^k$, for $i=1$ and $i=2$. Using Lemma 3.3 with $\lambda_1 = \ln(|\kappa_1(\mu_0)|), \lambda_2 = \ln(|\kappa_2(\mu_0)|)$, $\epsilon = -\ln(D)$, it follows that $\widetilde{\mathbf{B}}_\epsilon = B_{D,D}$ is contained in $\mathbf{B}_{N,1} \cup \mathbf{B}_{N,2}$, and if D is close to 1, then ϵ is close to zero. Hence N can be chosen arbitrary large.

□

2.6.3 An irrational resonant saddle

Theorem 2.25 *Let $\left(\frac{p_n}{q_n}\right)_{n\in\mathbb{N}}$ be the continued fraction expansion of $-\frac{\ln(|\kappa_2(\mu_0)|)}{\ln(|\kappa_1(\mu_0)|)}$. For every $N\in\mathbb{N}$ there exists a neighbourhood V of μ_0 such that the normal form can be written as*

$$G(x,y,\mu) = \left(\kappa_1(\mu)x + x\sum_{(a,b)\in\mathbb{N}^2_0} u_1^a u_2^b h_{(a,b),1}(\mu),\ \kappa_2(\mu)y + y\sum_{(a,b)\in\mathbb{N}^2_0} u_1^a u_2^b h_{(a,b),2}(\mu)\right),$$

where $u_1 = x^{p_{2N}}y^{q_{2N}}$, $u_2 = x^{p_{2N+1}}y^{q_{2N+1}}$, $\mathbb{N}^2_0 = \mathbb{N}^2 \setminus \{(0,0)\}$, and this normal form is defined for all values of $\mu \in V$. In this formula the coefficients $h_{(a,b),1}(\mu)$ and $h_{(a,b),2}(\mu)$ are analytic functions in u with coefficients depending continuously (resp. analytically) on a parameter μ.

Proof: Following the proof of the corresponding Theorem 3.18 for vector fields, where $\lambda_1 = \ln(|\kappa_1(\mu_0)|), \lambda_2 = \ln(|\kappa_2(\mu_0)|)$, $\epsilon = -\ln(D)$, we obtain that

$$\widetilde{\mathbf{B}}_\epsilon = B_{D,D} = \left\{k\in\mathbb{N}^2 |\, |\kappa_1(\mu_0)|^{k_1}|\kappa_2(\mu_0)|^{k_2} > D^k \text{ and } |\kappa_1(\mu_0)|^{k_1}|\kappa_2(\mu_0)|^{k_2} > \frac{1}{D^k}\right\},$$

is contained in

$$C_N := \left\{(k_1,k_2)\in\mathbb{N}^2 |\, \exists a,b\in\mathbb{N},\ (k_1,k_2) = a(p_{2N},q_{2N}) + b(p_{2N+1},q_{2N+1})\right\},$$

if $\epsilon := -\ln(D)$ is small enough. The proof is finished by expanding the sum over all indices in C_N.

□

Chapter 3

Vector fields in cones

3.1 Introduction

In this chapter we first explain the results from [3]. These results are the 'vector field equivalent' of the results from Chapter 2, although it should be noted that they also hold in Gevrey-α classes for $0 \leq \alpha \leq 1$. The proofs in [3] are less complicated compared to the corresponding proof of Theorem 2.1 concerning diffeomorphisms: they do not 'suffer' from the need to take the inverse. We refer to Chapter 5 to compare the Gevrey-results for diffeomorphisms proven there to the Gevrey-results obtained for vector fields in [3]. Such normal forms can be used in practical situations to give some precise time and trajectory estimates for the vector field at hand. See e.g. [63] where this is done in a two-dimensional situation.

In a second section of this chapter, we restrict to two-dimensional vector fields that are saddles. By this we mean that the linear part is diagonal and has a ratio of eigenvalues $\frac{\lambda_1}{\lambda_2}$ that is negative. We use the theorems from [3] in order to derive some very explicit expressions for the normal forms; these explicit expressions in the saddle case form a new contribution. We will consider two cases. First we will suppose that we are in a resonant situation i.e. the ratio of eigenvalues $\frac{\lambda_1}{\lambda_2} = -\frac{q}{p} \in \mathbb{Q}$. For the second case we will suppose that we are in the non-resonant case where the ratio of eigenvalues is irrational. In this case it is natural to choose the cones in close relation with the continued fraction expansion of the ratio of the eigenvalues.

In the Section 3.4 of this chapter we explain how the transformation can be chosen to converge if the eigenvalues satisfy a Brjuno type condition. The procedure that we use is different from the classical one. We will iterate the idea of removing terms inside a cone, narrowing the cone at each step. The slopes of the cone are again related to the continued fraction expansion of the ratio of the eigenvalues and determine a numerical condition that is necessary in order to obtain convergence. We will show that the condition we find is equivalent to the Brjuno condition, but the method reveals a close connection between the continued fraction expansion of the eigenvalues of the linear part and the convergence of the normal form. The method can possibly be extended towards systems of larger dimension. In order to do so, one needs a higher dimensional alternative for the continued fractions. Such alternatives have been developed in relation with Hamiltonian dynamics, see e.g. [35]. Moreover, it is interesting to compare the proofs in [36, 38, 35] to our proofs: they use a similar idea of an iteration-renormalization technique in Fourier space. It is also worth to note that the authors of [35] recently succeeded in extending their scheme to a higher dimensional case by using a more dimensional equivalent of continued fractions.

3.2 Results of [3]

We consider vector fields $X(x)$ defined on an open subset U of $\mathbb{C}^n$ with a singularity at the origin that are analytic (resp. formally Gevrey-s). Such vector fields can be written as $X_\mu(x) = A_\mu(x) + f_\mu(x)$, where A_μ is the linear part $DX_\mu(0)$ and $f_\mu(x) = \sum_{|k|\geq 2} f_k(\mu)x^k$, where $f_k(\mu)$ is a continuous (resp. C^k, analytic) function of the parameter. We will assume that the parameter is centered around a parameter value $\mu = \mu_0$. The idea of [3] is similar to Theorem 2.2 for diffeomorphisms in Chapter 2: in general, a parameter dependent vector field cannot be transformed to its linear part, due to the presence of resonances. If the linear part A_μ depends explicitly on the parameter, such resonances cannot be avoided. It is known that the corresponding classical normal form diverges in general, see e.g. [7]. Hence, it is natural to wonder whether it is possible to find a normal form procedure for these vector fields. The idea is to remove fewer terms in the Taylor series development of the 'normal form', while retaining analyticity (resp. the Gevrey-s property). The terms that are not removed are called 'bad terms', those that are removed are called 'good terms'. Since the local model to which the vector field is reduced by means of a coordinate transform is not the classical normal form, it is better to name it differently. We will use the terminology 'normal form in a cone' for this local model, referring to the conical structure of the bad set, but we might sometimes use the short 'normal form' with the same meaning.

Let $s \in [0, 1]$. The value of s will play the role of the Gevrey-order of the normal form transformation and the normal form. Let us first define what we mean by a Gevrey power series.

Definition 3.1 *A formal power series $f_\mu(x) = \sum_{l=0}^{\infty} \sum_{|k|=l} f_k(\mu)x^k$ is said to be Gevrey-α when there exists an $r > 0$ such that*

$$\sum_{|k|=l} ||f_k(\mu)||r^l \leq l!^\alpha,$$

the norm on $f_k(\mu)$ that is chosen may depend on the initial parameter dependence e.g. if continuous dependence on the parameter is supposed, one could use the sup-norm $||f_k(\mu)|| = \sup_{\mu\in\Lambda} |f_k(\mu)|$. Remark that if $\alpha = 0$, then the power series is analytic.

Suppose that the vector field X is Gevrey-s, i.e. each of the components of X is a formal power series that is Gevrey-s. Suppose also that the parameter $\mu \in \Lambda$. Let $\lambda(\mu) = (\lambda_1(\mu), \ldots, \lambda_n(\mu))$ be the parameter dependent eigenvalues of the linear part A_μ. We now introduce the good set $\mathbf{G}_s$ and the bad set $\mathbf{B}_s$.

$$\mathbf{G}_s = \left\{(k,j) \in \mathbb{N}^n \times \{1,\ldots,n\} \middle| \, |\langle \lambda(\mu), k\rangle - \lambda_j(\mu)| \geq C|k|^{1-s}, \forall \mu \in \Lambda \right\}. \tag{3.1}$$

$$\mathbf{B}_s = \left\{(k,j) \in \mathbb{N}^n \times \{1,\ldots,n\} \middle| \, |\langle \lambda(\mu), k\rangle - \lambda_j(\mu)| < C|k|^{1-s}, \forall \mu \in \Lambda \right\}. \tag{3.2}$$

We have now enough information to state one of the main results of [3].

Theorem 3.2 ([3]) *Let $s \in [0,1]$, and let $X_\mu = A_\mu(x) + \sum_{(k,j)} g_{k,j}(\mu)x^k e_j$ be a formal Gevrey-s vector field. Here $A_\mu(x) = DX(0)$, the linear part of X, is in diagonal form. There exists a formal Gevrey-s coordinate transform $U = x + \sum_{(k,j)} u_{k,j}x^k e_j$ containing only terms in the good set, i.e. $u_{k,j} \neq 0 \implies (k,j) \in \mathbf{G}_s$, that transforms X into a formal Gevrey-s vector field $Y = A_\mu + \sum_{(k,j)} g_{k,j}(\mu)x^k e_j$ and Y contains only bad terms i.e. $g_{k,j} \neq 0 \implies (k,j) \in \mathbf{B}_s$. Note that if $s = 0$ then X, U, Y are convergent and correspond to analytic functions.*

Proof:[This is only a short sketch] The formal conjugacy equation $U^*(X) = Y$ can be written as

$$Du_\mu.(A_\mu + g_\mu) - A_\mu.u_\mu + g_\mu - f_\mu(\mathrm{id} + u_\mu) = 0. \tag{3.3}$$

It is hence natural to consider the functional

$$\mathcal{F}(u,g,f) := Du(x).(A+g) - A.u + g - f \circ (\mathrm{id} + u), \tag{3.4}$$

in the variables (u,g,f). It is then shown that the functional is C^1 in a neighbourhood of the $(0,0,0)$ for appropriate Gevrey-s Banach spaces where (u,g,f) are defined. This is proven in a similar way for diffeomorphisms in Chapter 2, the main difference is the need to use a C^1 norm on the Banach space where u is an element of, because of the term containing Du in the definition of the functional $\mathcal{F}$. One proceeds then to calculate the directional derivative $D_{(u,g)}\mathcal{F}(0,0,0)$ and shows that it has a bounded inverse in the appropriate Banach spaces. An application of the implicit function theorem shows then that for small values of $||f||$ there exists $u(f)$ and $v(f)$ such that $\mathcal{F}(u(f), g(f), f) = 0$. The solution for larger values of $||f||$ can be found by rescaling the conjugacy equation $U^*(X) = Y$ first, like we did in the proof of Proposition 2.14 in Chapter 2. □

In order to be able to use an iteration process in case $s = 0$, where, at each step we make the cone more narrow, we will need to apply the above theorem at each step, using a smaller value of C in the definition of the cones $\mathbf{B}_0$ and $\mathbf{G}_0$. We will however also need an explicit estimate on the radius of convergence of the transformation U and the normal form Y. In order to do so, we will prove the above result using a fixed point technique. This is done in Section 3.4.

3.3 An explicit expression for the normal form in two dimensions

The purpose of this section is to give some explicit expressions for the normal forms in cones for a two-dimensional saddle with eigenvalues that are real. Therefore, consider such an analytic vector field X_μ depending on an additional parameter μ near μ_0. After applying Theorem 2.2, we may suppose that such a vector field X_μ is reduced to a vector field of the form

$$Y_\mu(x) = \lambda_1(\mu)x\frac{\partial}{\partial x} + \sum_{(l,1)\in\mathbf{B}_C} g_l^i(\mu)x^{l_1}y^{l_2}\frac{\partial}{\partial x} + \lambda_2(\mu)y\frac{\partial}{\partial y} + \sum_{(l,2)\in\mathbf{B}_C} g_l^2(\mu)x^{l_1}y^{l_2}\frac{\partial}{\partial y}, \tag{3.5}$$

where

$$\begin{aligned}\mathbf{B}_C &= \{(k,j)\in\mathbb{N}^n\times\{1,\dots,n\} \mid |\langle\lambda,k\rangle-\lambda_j| \le C|k| \text{ and } |k|\ge 2\} && (3.6)\\ &= \{(k,j)\in\mathbb{N}^n\times\{1,\dots,n\} \mid |\langle\lambda,k-e_j\rangle| \le C|k| \text{ and } |k|\ge 2\}\,. && (3.7)\end{aligned}$$

Hence

$$\begin{aligned}((k_1,k_2),1)\in\mathbf{B}_C &\Leftrightarrow -C(k_1+k_2)\le\lambda_1(k_1-1)+\lambda_2k_2\le C(k_1+k_2)\\ &\Leftrightarrow -C(k_1+k_2)\le\lambda_1(k_1-1)+\lambda_2k_2\le C(k_1+k_2)\\ &\Leftrightarrow \frac{(\lambda_1-C)k_1-\lambda_1}{(C-\lambda_2)}\le k_2\le\frac{(\lambda_1+C)k_1-\lambda_1}{(-C-\lambda_2)}.\end{aligned}$$

Let $\varepsilon>0$ be fixed, we show that there exists a $C>0$ such that $\mathbf{B}_{C,1}=\{(k,1)\in\mathbf{B}_C\}$ is contained in

$$\widetilde{\mathbf{B}}_{\varepsilon,1}=\left\{(k_1,k_2)\in N^2|\left(-\frac{\lambda_1}{\lambda_2}-\varepsilon\right)(k_1-1)\le k_2\le\left(-\frac{\lambda_1}{\lambda_2}+\varepsilon\right)(k_1-1)\right\}.$$

Note that $\mathbf{B}_{C,1}$ is determined by the lines

$$L_1 : k_2=\frac{(\lambda_1-C)k_1-\lambda_1}{(C-\lambda_2)} \tag{3.8}$$

$$L_2 : k_2=\frac{(\lambda_1+C)k_1-\lambda_1}{(-C-\lambda_2)} \tag{3.9}$$

and $\widetilde{\mathbf{B}}_{\varepsilon,1}$ is determined by the lines

$$M_1 : k_2=\left(-\frac{\lambda_1}{\lambda_2}-\varepsilon\right)(k_1-1) \tag{3.10}$$

$$M_2 : k_2=\left(-\frac{\lambda_1}{\lambda_2}+\varepsilon\right)(k_1-1)\,. \tag{3.11}$$

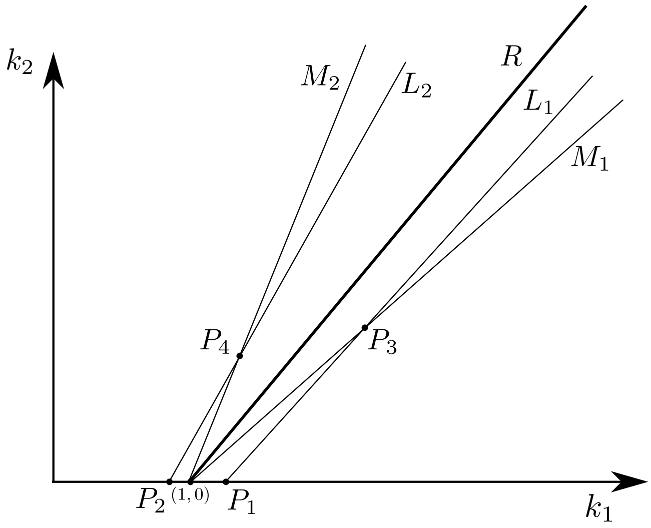


Figure 3.1: The cone B of terms which cannot be removed.

See also Figure 3.1, where the lines M_1, M_2, L_1, L_2 and $R: \lambda_1 k_1 + \lambda_2 k_2 - \lambda_1 = 0$ are drawn. We define $P_3 = M_1 \cap L_1$, $P_4 = M_2 \cap L_2$, $P_1 = L_1 \cap \{y = 0\}$, $P_2 = L_2 \cap \{y = 0\}$. If $C > 0$ is small enough and μ is close enough to μ_0, then the slope of L_2 is smaller then the slope of M_2 and the slope of L_1 is larger than the slope of M_1. It is also straightforward to show that

$$\lim_{C\to 0,\mu\to\mu_0} P_1 = \lim_{C\to 0,\mu\to\mu_0} P_2 = \lim_{C\to 0,\mu\to\mu_0} P_3 = \lim_{C\to 0,\mu\to\mu_0} P_4 = (1,0) = R \cap \{y = 0\}.$$

It follows that if we choose $C > 0$ small enough, and μ close enough to μ_0, then the interior of the triangles determined by $(1,0), P_2, P_4$ and $(1,0), P_1, P_3$ does not contain any point $T = (t_1, t_2)$ for which t_1 or t_2 is an integer. Indeed, each such point $T \neq (1,0)$ lies on a distance smaller then $1/2$ from the point $(1,0)$. We conclude that $\mathbf{B}_{C,1} \subset \widetilde{\mathbf{B}}_{\varepsilon,1}$. The same way we can show that if $C > 0$ is small enough and μ is close enough to μ_0, that $\mathbf{B}_{C,2} \subset \widetilde{\mathbf{B}}_{\varepsilon,2}$, where

$$\widetilde{\mathbf{B}}_{\varepsilon,2} = \left\{ (k_1, k_2) \in \mathbb{N}^2 | \left(-\frac{\lambda_1}{\lambda_2} - \varepsilon \right) k_1 \leq k_2 - 1 \leq \left(-\frac{\lambda_1}{\lambda_2} + \varepsilon \right) k_1 \right\}. \tag{3.12}$$

Let

$$\widetilde{\mathbf{B}}_{\varepsilon} = \left\{ (k_1, k_2) \in \mathbb{N}^2 | \left(-\frac{\lambda_1}{\lambda_2} - \varepsilon \right) k_1 \leq k_2 \leq \left(-\frac{\lambda_1}{\lambda_2} + \varepsilon \right) k_1 \right\}, \tag{3.13}$$

then it is readily verified that $\widetilde{\mathbf{B}}_{\varepsilon,1} = (1,0) + \widetilde{\mathbf{B}}_{\varepsilon}$ and $\widetilde{\mathbf{B}}_{\varepsilon,2} = (0,1) + \widetilde{\mathbf{B}}_{\varepsilon}$.

We prove the following lemma for the description of the cones in case the eigenvalues have a rational ratio i.e. $-\frac{\lambda_1}{\lambda_2} \in \mathbb{Q}$.

Lemma 3.3 *Let $\varepsilon > 0$. Suppose that $-\frac{\lambda_1}{\lambda_2} = \frac{q}{p}$, where q, p are natural numbers without common divisors. There exists unique $(r_0, s_0), (r_1, s_1) \in \mathbb{N}^2$ for which $qr_1 - ps_1 = -1$, $qr_0 - ps_0 = 1$, $0 \le r_0 \le p$, $0 \le r_0 \le q$ and $(r_1, s_1) = (p, q) - (r_0, s_0)$; and there exists a natural number N such that $\widetilde{\mathbf{B}}_{\varepsilon}$ is a subset of $\widehat{\mathbf{B}}_{N,1} \cup \widehat{\mathbf{B}}_{N,2}$*

$$\widehat{\mathbf{B}}_{N,1} = \{((p,q)Nt + ((p,q)N + (r_1, s_1))s) \,|s, t \in \mathbb{N}\}. \tag{3.14}$$

$$\widehat{\mathbf{B}}_{N,2} = \{((p,q)Nt + ((p,q)N + (r_0, s_0))s) \,|s, t \in \mathbb{N}\}. \tag{3.15}$$

Moreover $N = N(\varepsilon) \longrightarrow +\infty$ as $\varepsilon \longrightarrow 0$.

Proof: The existence of (r_0, s_0) are guaranteed by a theorem of Bezout in number theory (here we need that p and q have no common divisors). Remark also that $(r_1, s_1) = (p, q) - (r_0, s_0)$ satisfies $qr_1 - ps_1 = -1$. We have $\widetilde{\mathbf{B}}_{\varepsilon} = B_1 \cup B_2$, where

$$B_1 = \left\{(k_1, k_2) \in N^2 | -\frac{\lambda_1}{\lambda_2}k_1 \le k_2 \le \left(-\frac{\lambda_1}{\lambda_2} + \varepsilon\right)k_1\right\}, \tag{3.16}$$

$$B_2 = \left\{(k_1, k_2) \in N^2 | \left(-\frac{\lambda_1}{\lambda_2} - \varepsilon\right)k_1 \le k_2 \le -\frac{\lambda_1}{\lambda_2}k_1\right\}. \tag{3.17}$$

We show that $B_1 \subset \widehat{\mathbf{B}}_{N,1}$ the proof of $B_2 \subset \widehat{\mathbf{B}}_{N,2}$ is analogous. Let N be the smallest natural number for which $\frac{Nq+s_1}{Np+r_1} + \frac{\lambda_1}{\lambda_2} \le \varepsilon$. It is clear that N tends to infinity if ε tends to zero. Geometrically this means that the line with direction vector $N(p,q) + (r_1, s_1)$ has a smaller slope then the line $k_2 = \left(-\frac{\lambda_1}{\lambda_2} + \varepsilon\right)k_1$. Hence $B_1 \subset \mathbf{BB}_1$, where

$$\mathbf{BB}_1 = \left\{((p,q)Nt + ((p,q)N + (r_1, s_1))s) \,|s, t \in \mathbb{R}^+\right\} \cap \mathbb{N}^2.$$

We argue now that $\mathbf{BB}_1 = \widehat{\mathbf{B}}_{N,1}$. Indeed, suppose that the opposite is true, then there exists an $(t, s) \notin \mathbb{N}^2$ for which $v = Nt(p,q) + s(N(p,q) + (r_1, s_1)) = (s+t)N(p,q) + s(r_1, s_1) \in \mathbb{N}^2$. Such a point v is contained in the closed parallellogram determined by the integral points $\lfloor(s+t)N\rfloor(p,q) + \lfloor s\rfloor(r_1, s_1)$, $\lceil(s+t)N\rceil(p,q) + \lfloor s\rfloor(r_1, s_1)$, $\lfloor(s+t)N\rfloor(p,q) + \lceil s\rceil(r_1, s_1)$, $\lceil(s+t)N\rceil(p,q) + \lceil s\rceil(r_1, s_1)$, but is not one of its corner points, see also Figure 3.2. Such a parallellogram has surface area $|qr_1 - ps_1| = 1$. Since any parallellogram with integral corner points has a surface area that is at least 1, it follows that $v \notin \mathbb{N}^2$. □

We have now proven:

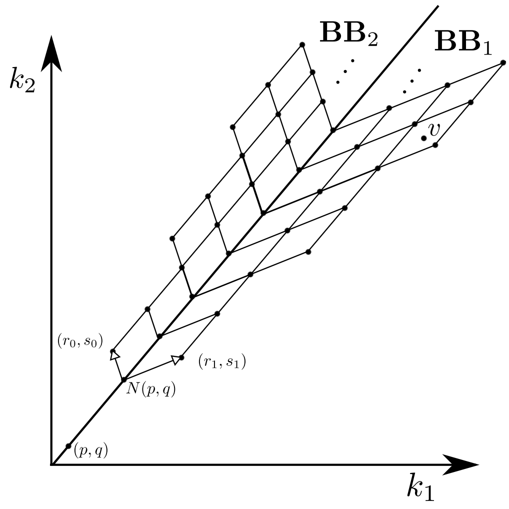


Figure 3.2: $\mathbf{BB}_1$ and $\mathbf{BB}_2$.

Theorem 3.4 *Suppose that $-\frac{\lambda_1}{\lambda_2} = \frac{q}{p}$, where q, p are natural numbers without common divisors. There exists unique $(r_0, s_0), (r_1, s_1) \in \mathbb{N}^2$ for which $qr_1 - ps_1 = -1$, $qr_0 - ps_0 = 1$, $0 \leq r_0 \leq p$, $0 \leq r_0 \leq q$ and $(r_1, s_1) = (p, q) - (r_0, s_0)$. Let $u = x^p y^q$. Suppose that $C > 0$ is small enough and suppose that μ is sufficiently close to μ_0, then the normal form determined by equation (3.5) can be written as*

$$Y_\mu : \begin{cases} \dot{x} = \lambda_1(\mu) x \left(1 + u b_{l,1}(u, \mu) + \sum\limits_{l \geq 1} u^{Nl} \left(x^{lr_0} y^{ls_0} g_{l,1}(u, \mu) + x^{lr_1} y^{ls_1} h_{l,1}(u, \mu) \right) \right) \\ \dot{y} = \lambda_2(\mu) y \left(1 + u b_{l,2}(u, \mu) + \sum\limits_{l \geq 1} u^{Nl} \left(x^{lr_0} y^{ls_0} g_{l,2}(u, \mu) + x^{lr_1} y^{ls_1} h_{l,2}(u, \mu) \right) \right) . \end{cases}$$

In this formula the coefficients $g_{l,1}(u, \mu)$, $g_{l,2}(u, \mu)$, $h_{l,1}(u, \mu)$, $h_{l,2}(u, \mu)$, $b_{l,2}(u, \mu)$, $b_{l,1}(u, \mu)$ are analytic functions in u with coefficients depending on a parameter.

Proof: This follows from $\mathbf{B}_{C,1} \subset \widetilde{\mathbf{B}}_{\varepsilon,1}$, $\mathbf{B}_{C,2} \subset \widetilde{\mathbf{B}}_{\varepsilon,2}$, $\widetilde{\mathbf{B}}_{\varepsilon,1} = (1, 0) + \widetilde{\mathbf{B}}_\varepsilon$, $\widetilde{\mathbf{B}}_{\varepsilon,2} = (0, 1) + \widetilde{\mathbf{B}}_\varepsilon$, $\widetilde{\mathbf{B}}_\varepsilon \subset \widehat{\mathbf{B}}_{N,1} \cup \widehat{\mathbf{B}}_{N,2}$ and the description of $\widehat{\mathbf{B}}_{N,1}$ and $\widehat{\mathbf{B}}_{N,2}$ in Lemma 3.3. If C is arbitrary small and μ is sufficiently close to μ_0, then ε is arbitrary small and N arbitrary large. □

We proceed to the case where the ratio $-\frac{\lambda_1}{\lambda_2} \in \mathbb{R} \setminus \mathbb{Q}$. Since we will use the continued fraction expansion of $-\frac{\lambda_1}{\lambda_2}$, it is useful to recall some well known facts concerning continued fractions. This material is standard and can e.g. be found in [43, 64, 47]. The continued fraction expansion of $\xi_0 = -\frac{\lambda_1}{\lambda_2} = [a_0; a_1, a_2, \ldots]$ is defined as follows:

$$-\frac{\lambda_1}{\lambda_2} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ldots}}$$

More formally we have $a_0 := \lfloor \xi_0 \rfloor$, $b_0 := \xi_0 - \lfloor \xi_0 \rfloor$, $a_1 := \lfloor \frac{1}{b_0} \rfloor$, $b_1 := \frac{1}{b_0} - \lfloor \frac{1}{b_0} \rfloor$, ..., $a_n := \lfloor \frac{1}{b_{n-1}} \rfloor$, $b_n := \frac{1}{b_{n-1}} - \lfloor \frac{1}{b_{n-1}} \rfloor$. We define the ratio's $\frac{q_n}{p_n} = [a_0; a_1, \ldots, a_n, 0, 0, \ldots] \in \mathbb{Q}$. For example we have that $\frac{q_0}{p_0} = \frac{a_0}{1}$, $\frac{q_1}{p_1} = \frac{a_0 a_1 + 1}{a_1}$, $\frac{q_2}{p_2} = \frac{a_0 a_1 a_2 + a_0 + a_2}{a_1 a_2 + 1}$ and so on. We have recursively

$$(q_{n+1}, p_{n+1}) = (a_{n+1} q_n + q_{n-1}, a_{n+1} p_n + p_{n-1}). \tag{3.18}$$

For any $n \in \mathbb{N}$ and any $x \in \mathbb{R}$ we define

$$[a_0, a_1, \ldots, a_{n-1}, x] = -\frac{\lambda_1}{\lambda_2} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots \cfrac{}{a_{n-1} + \cfrac{1}{x}}}}$$

Lemma 3.5 *Let $x > 0$, $n \in \mathbb{N} \setminus \{0\}$, then*

$$[a_0, a_1, \ldots, a_{n-1}, x] = \frac{x q_{n-1} + q_{n-2}}{x p_{n-1} + p_{n-2}}. \tag{3.19}$$

Proof: We make a proof by induction. The formula is trivially true for $n = 1$. Suppose that formula (3.19) is valid for n, we show that it is also true for $n + 1$. Indeed,

$$\begin{aligned} [a_0, \ldots, a_n, x] = [a_0, \ldots, a_n + \frac{1}{x}] &= \frac{\left(a_n + \frac{1}{x}\right) q_{n-1} + q_{n-2}}{\left(a_n + \frac{1}{x}\right) p_{n-1} + p_{n-2}} \\ &= \frac{x\left(a_n q_{n-1} + q_{n-2}\right) + q_{n-1}}{x\left(a_n p_{n-1} + p_{n-2}\right) + p_{n-1}} \\ &= \frac{x q_n + q_{n-1}}{x p_n + p_{n-1}}. \end{aligned}$$

□

We say that the eigenvalues (λ_1, λ_2) satisfy a τ-diophantine condition if the inequality

$$|\lambda_1 k_1 + \lambda_2 k_2| \geq \frac{C_0}{(k_1 + k_2)^\tau}$$

holds for all non-zero $(k_1, k_2) \in \mathbb{Z}^2$. We say that the eigenvalues (λ_1, λ_2) satisfy the Brjuno condition iff

$$\sum_{n\geq 2} \frac{\ln(p_{n+1})}{p_n} < \infty. \tag{3.20}$$

We proof some useful lemmas concerning continued fractions. These can for instance be found in [51, 64, 47], we repeat them for the sake of completeness.

Lemma 3.6 *The set of ratio's $-\frac{\lambda_1}{\lambda_2}$ that do not satisfy the Brjuno condition have zero Lebesgue measure in $\{(\lambda_1, \lambda_2) \in \mathbb{R}^2 |\ \lambda_1 > 0, \lambda_2 < 0\}$.*

Proof: This follows from the fact that almost every ratio satisfies a τ-diophantine condition for some τ; and every τ-diophantine number satisfies the Brjuno condition. □

Lemma 3.7 *Let $\left(\frac{q_n}{p_n}\right)_{n\in\mathbb{N}}$ be the continued fraction expansion of $-\frac{\lambda_1}{\lambda_2}$ then the sequences $(p_n)_{n\in\mathbb{N}}$ and $(q_n)_{n\in\mathbb{N}}$ increase at least as fast as the Fibonacci sequence.*

Proof: This follows from the fact that $p_{n+1} = a_{n+1}p_n + p_{n-1}$ for $a_{n+1} \geq 1$, where, for the Fibonnaci sequence $(f_n)_{n\in\mathbb{N}}$ one has $f_{n+1} = f_n + f_{n-1}$. □

Lemma 3.8 *The continued fraction expansion $\left(\frac{q_n}{p_n}\right)_{n\in\mathbb{N}}$ of $-\frac{\lambda_1}{\lambda_2}$ satisfies the recurrence relations*

$$q_n p_{n-1} - p_n q_{n-1} = (-1)^n. \tag{3.21}$$

Proof: The proof is done by induction on n. Suppose that we have already proven that $q_n p_{n-1} - p_n q_{n-1} = (-1)^n$. We have $q_{n+1} = a_{n+1}q_n + q_{n-1}$ and $p_{n+1} = a_{n+1}p_n + p_{n-1}$. Hence

$$\begin{aligned} q_{n+1}p_n - p_{n+1}q_n &= (a_{n+1}q_n + q_{n-1})p_n - (a_{n+1}p_n + p_{n-1})q_n \\ &= q_{n-1}p_n - p_{n-1}q_n = (-1)^{n+1}. \end{aligned}$$

□

Remark 3.9 *Geometrically this means that the parallellogram spanned by the vectors (q_n, p_n) and (q_{n-1}, p_{n-1}) has surface area 1. Hence there are, apart from the corner points, no other integer couples in the closure of the parallellogram spanned by these two vectors.*

Lemma 3.10 *The following sequence of inequalities are true:*

$$\frac{q_0}{p_0} \leq \ldots \leq \frac{q_{2n-2}}{p_{2n-2}} \leq \frac{q_{2n}}{p_{2n}} \leq \ldots \leq -\frac{\lambda_1}{\lambda_2} \leq \ldots \leq \frac{q_{2n+1}}{p_{2n+1}} \leq \frac{q_{2n-1}}{p_{2n-1}} \ldots \leq \frac{q_1}{p_1}.$$

Geometrically, this means that the slope of the resonant line is approached alternately from above by the continued fractions with an odd index, and from below by the continued fractions with an even index.

Proof: We give a proof by induction. Suppose that it was already shown that

$$\frac{q_0}{p_0} \leq \dots \leq \frac{q_{2n-2}}{p_{2n-2}} \leq \frac{q_{2n}}{p_{2n}} \leq \frac{q_{2n-1}}{p_{2n-1}} \leq \frac{q_{2n-3}}{p_{2n-3}} \dots \leq \frac{q_1}{p_1},$$

we prove that this implies

$$\frac{q_{2n}}{p_{2n}} \leq \frac{q_{2n+2}}{p_{2n+2}} \leq \frac{q_{2n+1}}{p_{2n+1}} \leq \frac{q_{2n-1}}{p_{2n-1}}.$$

We have $(q_{2n+1}, p_{2n+1}) = a_{2n+1}(q_{2n}, a_{2n+1}p_{2n}) + (q_{2n-1}, p_{2n-1})$. Since $\frac{q_{2n}}{p_{2n}} \leq \frac{q_{2n-1}}{p_{2n-1}}$ and a_{2n+1} is positive it follows that $\frac{q_{2n}}{p_{2n}} \leq \frac{q_{2n+1}}{p_{2n+1}} \leq \frac{q_{2n-1}}{p_{2n-1}}$. Analogous it follows from $(q_{2n+2}, p_{2n+2}) = a_{2n+2}(q_{2n+1}, p_{2n+1}) + (q_{2n}, p_{2n})$, $a_{2n+2} > 0$ and $\frac{q_{2n}}{p_{2n}} \leq \frac{q_{2n+1}}{p_{2n+1}}$ that $\frac{q_{2n}}{p_{2n}} \leq \frac{q_{2n+2}}{p_{2n+2}} \leq \frac{q_{2n+1}}{p_{2n+1}}$.

We finish the proof by explaining why $-\frac{\lambda_1}{\lambda_2}$ lies between continued fractions with an even and an odd index. Suppose therefore that $\frac{q_{2n}}{p_{2n}} \leq -\frac{\lambda_1}{\lambda_2} \leq \frac{q_{2n+1}}{p_{2n+1}}$ were false for a certain $n \in \mathbb{N}$. Then, since $\left(\frac{q_{2n}}{p_{2n}}\right)_{n\in\mathbb{N}}$ is a strictly increasing sequence and $\left(\frac{q_{2n+1}}{p_{2n+1}}\right)_{n\in\mathbb{N}}$ is a strictly decreasing sequence, the limit of $\frac{q_n}{p_n}$ would be different from $-\frac{\lambda_1}{\lambda_2}$, a contradiction. □

Lemma 3.11 *Let $a, b, c, d > 0$, then $\frac{a}{b} < \frac{c}{d}$ implies $\frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}$.*

Proof: We show the first inequality, the second being analogous. We have

$$\begin{aligned} \frac{a}{b} < \frac{c}{d} &\Leftrightarrow ad < bc \\ &\Leftrightarrow ad + ab < ab + bc \\ &\Leftrightarrow \frac{a}{b} < \frac{a+c}{b+d}. \end{aligned}$$

□

Lemma 3.12 *Let $\left(\frac{q_n}{p_n}\right)_{n\in\mathbb{N}}$ be the continued fraction expansion of $\xi_0 \in \mathbb{R}^+ \setminus \mathbb{Q}$. Then*

$$\frac{1}{p_n + p_{n+1}} < |p_n\xi_0 - q_n| < \frac{1}{p_{n+1}}. \tag{3.22}$$

Proof: We show first for even $n = 2m$ that $\frac{1}{p_{2m}+p_{2m+1}} < |p_{2m}\xi_0 - q_{2m}|$. The proof for the odd values of n is analogous and is left to the reader. Using Lemma 3.10, we observe that

$$\frac{q_{2m}}{p_{2m}} < \frac{q_{2m+2}}{p_{2m+2}} = \frac{a_{2m+1}q_{2m+1} + q_{2m}}{a_{2m+1}p_{2m+1} + p_{2m}} < \xi_0.$$

Using this inequality and Lemma 3.11, it follows that

$$\begin{aligned}\frac{q_{2m}}{p_{2m}} < \frac{q_{2m+1}+q_{2m}}{p_{2m+1}+p_{2m}} < \ldots < &\frac{(a_{2m+1}-1)q_{2m+1}+q_{2m}}{(a_{2m+1}-1)p_{2m+1}+p_{2m}} \\ &< \frac{q_{2m+2}}{p_{2m+2}} = \frac{a_{2m+1}q_{2m+1}+q_{2m}}{a_{2m+1}p_{2m+1}+p_{2m}} < \xi_0.\end{aligned}$$

As a consequence it follows that

$$\left|\frac{q_{2m}}{p_{2m}} - \xi_0\right| > \left|\frac{q_{2m+1}+q_{2m}}{p_{2m+1}+p_{2m}} - \frac{q_{2m}}{p_{2m}}\right| = \frac{1}{(p_{2m+1}+p_{2m})\,p_{2m}},$$

which shows, after multiplication by p_{2m}, that $\frac{1}{p_{2m}+p_{2m+1}} < |p_{2m}\xi_0 - q_{2m}|$.

We show now for even $n = 2m$ that $|p_{2m}\xi_0 - q_{2m}| < \frac{1}{p_{2m+1}}$, the odd case being analogous. We use Lemma 3.10 to observe that

$$\frac{q_{2m}}{p_{2m}} < \xi_0 < \frac{q_{2m+1}}{p_{2m+1}}.$$

Hence

$$\left|\frac{q_{2m}}{p_{2m}} - \xi_0\right| < \left|\frac{q_{2m}}{p_{2m}} - \frac{q_{2m+1}}{p_{2m+1}}\right| = \frac{1}{p_{2m}p_{2m+1}},$$

from which it follows after multiplication by p_{2m} that

$$|p_{2m}\xi_0 - q_{2m}| < \frac{1}{p_{2m+1}}.$$

□

We prove the following lemma that explains why $\frac{q_{n+1}}{p_{n+1}}$ is a better approximation of $-\frac{\lambda_1}{\lambda_2}$ than $\frac{q_n}{p_n}$.

Lemma 3.13 $\left(\left|-\frac{\lambda_1}{\lambda_2} - \frac{q_n}{p_n}\right|\right)_{n\in\mathbb{N}}$ *is a strictly decreasing sequence.*

Proof: Because $p_{n+1} = a_n p_n + p_{n-1}$, it follows that $p_{n+1} \geq p_n + p_{n-1}$ and hence

$$\frac{1}{p_{n+1}} \leq \frac{1}{p_n + p_{n+1}}.$$

Using Lemma 3.12, we observe that

$$|p_n\xi_0 - q_n| < \frac{1}{p_{n+1}} \leq \frac{1}{p_n + p_{n-1}} < |p_{n-1}\xi_0 - q_{n-1}|.$$

As a consequence it follows that

$$\left|\xi_0 - \frac{q_n}{p_n}\right| < \left|\xi_0 - \frac{q_{n-1}}{p_{n-1}}\right|,$$

which is what we needed to show. □

Let $\xi_0 \in \mathbb{R} \setminus \mathbb{Q}$, $\xi_0 \geq 0$. We say that $(a,b) \in \mathbb{N}^2$, $b \neq 0$ is a best rational approximation to ξ_0 if $|b\xi_0 - a| < |p\xi_0 - q|$ for all $q, p \in \mathbb{N}$ for which $0 < p \leq b$.

Lemma 3.14 *Let $\xi_0 \in \mathbb{R} \setminus \mathbb{Q}$, $\xi_0 \geq 0$. Then every best rational approximation (a,b) has a ratio $\frac{a}{b}$ that is a continued fraction approximant of ξ_0 (i.e. there exists a $k \in \mathbb{N}$ such that $\frac{q_k}{p_k} = \frac{a}{b}$).*

Proof: We give a proof by contradiction. Therefore, suppose that (a,b) is a best rational approximation and $\frac{a}{b} \neq \frac{q_n}{p_n}$, for all $n \in \mathbb{N}$. We distinguish four cases.
Case 1: $\frac{a}{b} < \frac{q_0}{p_0}$
In this case

$$|b\xi_0 - a| = b\left|\xi_0 - \frac{a}{b}\right| \geq \left|\xi_0 - \frac{a}{b}\right| > \left|\xi_0 - \frac{q_0}{p_0}\right|.$$

This is a contradiction since $p_0 = 1$.
Case 2: $\frac{a}{b} > \frac{q_1}{p_1}$
We have

$$|b\xi_0 - a| = b\left|\xi_0 - \frac{a}{b}\right| \geq b\left|\frac{q_1}{p_1} - \frac{a}{b}\right| \geq b\frac{1}{p_1 b} = \frac{1}{p_1} = \frac{1}{a_1}.$$

Since, using Lemma 3.12, $|\xi_0 - a_0| = |\xi_0 - \frac{q_0}{p_0}| < \frac{1}{p_1}$, we have a contradiction.
Case 3: $\frac{q_0}{p_0} < \frac{a}{b} < \xi_0$
In this case there is some even $n = 2m$ for which

$$\frac{q_n}{p_n} < \frac{a}{b} < \xi_0 < \frac{q_{n+1}}{p_{n+1}}.$$

It follows that

$$\frac{1}{p_{n+1}p_n} = \left|\frac{q_{n+1}}{p_{n+1}} - \frac{q_n}{p_n}\right| > \left|\frac{a}{b} - \frac{q_n}{p_n}\right| \geq \frac{1}{p_n b}.$$

This implies that $b > p_{n+1}$. We have on one hand that

$$|b\xi_0 - a| = b\left|\xi_0 - \frac{a}{b}\right| > b\left|\frac{p_{n+2}}{q_{n+2}} - \frac{a}{b}\right| \geq \frac{b}{p_{n+2}b} = \frac{1}{p_{n+2}},$$

and on the other hand, using Lemma 3.12,

$$|p_{n+1}\xi_0 - q_{n+1}| < \frac{1}{p_{n+2}},$$

which is clearly a contradiction, because $|p_{n+1}\xi_0 - q_{n+1}| < |b\xi_0 - a|$ and $p_{n+1} < b$, and the supposition that (a,b) is a best rational approximation of ξ_0.

Case 4: $\xi_0 < \frac{a}{b} < \frac{q_1}{p_1}$
This case in analogous to Case 3. □

Suppose now that $\xi_0 = [a_0, a_1, \ldots]$. We define $\xi_\nu = [a_\nu, a_{\nu+1}, \ldots]$. One can verify that

$$\begin{aligned} \xi_0 &= \frac{q_{\nu-1}\xi_\nu + q_{\nu-2}}{p_{\nu-1}\xi_\nu + p_{\nu-2}}, \\ \xi_\nu &= \frac{p_{\nu-1}\xi_0 - q_{\nu-2}}{p_{\nu-1}\xi_0 - q_{\nu-2}}. \end{aligned}$$

Let $p_{-1} = 0$, $q_{-1} = 1$. We define the following two approximation series of ξ_0, here is $\left(\frac{q_n}{p_n}\right)_{n\in\mathbb{N}}$ the continued fraction expansion of ξ_0:

$$\frac{q_0}{p_0}, \frac{q_1+q_0}{p_1+p_0}, \frac{2q_1+q_0}{2p_1+p_0}, \ldots, \frac{a_2q_1+q_0}{a_2p_1+p_0} = \frac{q_2}{p_2}, \frac{q_3+q_2}{p_3+p_2}, \ldots, \frac{a_4q_3+q_2}{a_4p_3+p_2}, \ldots \tag{3.23}$$

$$\frac{q_0+q_{-1}}{p_0+p_{-1}}, \frac{2q_0+q_{-1}}{2p_0+p_{-1}}, \ldots, \frac{a_1q_0+q_{-1}}{a_1p_0+p_{-1}} = \frac{q_1}{p_1}, \frac{q_2+q_1}{p_2+p_1}, \ldots, \frac{a_3q_2+q_1}{a_3p_2+p_1}, \ldots \tag{3.24}$$

Each member of (3.23) is smaller then ξ_0 and each member of (3.24) is larger then ξ_0. Both sequences (3.23) and (3.24) are sorted by growing denominator. We show a few approximation lemmas.

Lemma 3.15 *Let $\nu \in \mathbb{N}$, $\nu \geq 2$. Every positive rational number $\frac{A}{B}$ that is different from each of the members of (3.23) and (3.24) that is as close or closer to ξ_0 as*

$$\frac{cq_{\nu-1} + q_{\nu-2}}{cp_{\nu-1} + p_{\nu-2}},$$

for a certain $0 \leq c \leq a_\nu$, has a larger denominator (i.e. $B > cp_{\nu-1} + p_{\nu-2}$).

Proof: Let $\frac{A}{B}$ be such that

$$\left|\xi_0 - \frac{A}{B}\right| \leq \left|\xi_0 - \frac{cq_{\nu-1} + q_{\nu-2}}{cp_{\nu-1} + p_{\nu-2}}\right|,$$

and suppose that $\frac{A}{B}$ is not one of the members of (3.23) or (3.24). Then

$$\left|\frac{A}{B} - \frac{q_{\nu-1}}{p_{\nu-1}}\right| = \left|\left(\xi_0 - \frac{q_{\nu-1}}{p_{\nu-1}}\right) - \left(\xi_0 - \frac{A}{B}\right)\right| \tag{3.25}$$

$$\overset{(a)}{\leq} \left|\left(\xi_0 - \frac{q_{\nu-1}}{p_{\nu-1}}\right)\right| + \left|\left(\xi_0 - \frac{A}{B}\right)\right| \tag{3.26}$$

$$\overset{(b)}{\leq} \left|\left(\xi_0 - \frac{q_{\nu-1}}{p_{\nu-1}}\right)\right| + \left|\left(\xi_0 - \frac{cq_{\nu-1} + q_{\nu-2}}{cp_{\nu-1} + p_{\nu-2}}\right)\right| \tag{3.27}$$

$$\overset{(c)}{\leq} \left|\left(\xi_0 - \frac{q_{\nu-1}}{p_{\nu-1}}\right) - \left(\xi_0 - \frac{cq_{\nu-1} + q_{\nu-2}}{cp_{\nu-1} + p_{\nu-2}}\right)\right|. \tag{3.28}$$

Remark that both quantities between the absolute values in (3.27) have an opposite sign. Furthermore, it is impossible that we have an equality in both (a) and (b), because this would mean that $\frac{A}{B}$ lies between ξ_0 and $\frac{q_{\nu-1}}{p_{\nu-1}}$ and between ξ_0 and $\frac{cq_{\nu-1}+q_{\nu-2}}{cp_{\nu-1}+p_{\nu-2}}$, which is contradictory since $\xi_0 - \frac{q_{\nu-1}}{p_{\nu-1}}$ and $\xi_0 - \frac{cq_{\nu-1}+q_{\nu-2}}{cp_{\nu-1}+p_{\nu-2}}$ have a different sign. Consequently

$$\left|\frac{A}{B} - \frac{q_{\nu-1}}{p_{\nu-1}}\right| < \left|\frac{q_{\nu-1}}{p_{\nu-1}} - \frac{cq_{\nu-1}+q_{\nu-2}}{cp_{\nu-1}+p_{\nu-2}}\right| = \frac{1}{p_{\nu-1}\left(cp_{\nu-1}+p_{\nu-2}\right)},$$

and

$$|Ap_{\nu-1} - q_{\nu-1}B| < \frac{B}{cp_{\nu-1}+p_{\nu-2}}.$$

Since $\frac{A}{B} \neq \frac{q_{\nu-1}}{p_{\nu-1}}$ it follows that $|Ap_{\nu-1} - q_{\nu-1}B| \geq 1$, since it is a natural number. As a consequence $B > cp_{\nu-1} + p_{\nu-2}$, which finishes the proof. □

If $\frac{A}{B}$ has the property that every rational number that lies between ξ_0 and $\frac{A}{B}$ has a larger denominator, then we say that $\frac{A}{B}$ is a best rational approximation of ξ_0 of the second kind.

Lemma 3.16 *If $\frac{A}{B}$ is a best rational approximation of ξ_0 of the second kind, then $\frac{A}{B}$ is one of the members of (3.23) or (3.24).*

Proof: Suppose that $\frac{A}{B}$ is a rational that is different from all members of (3.23) and (3.24). We consider two cases.
Case 1: $\frac{A}{B} < \xi_0$.
In this case $\frac{A}{B}$ lies either between two subsequent members of 3.23 or $\frac{A}{B} < \frac{q_0}{p_0}$. If $\frac{A}{B} < \frac{q_0}{p_0} = \frac{a_0}{1} < \xi_0$, we have a contradiction, since the ratio $\frac{a_0}{1}$ has a denominator that is not strictly smaller then the denominator of $\frac{A}{B}$. If $\frac{(c-1)q_{\nu-1}+q_{\nu-2}}{(c-1)p_{\nu-1}+p_{\nu-2}} < \frac{A}{B} < \frac{cq_{\nu-1}+q_{\nu-2}}{cp_{\nu-1}+p_{\nu-2}} < \xi_0$, then also

$$\begin{aligned} 0 < \left|\frac{A}{B} - \frac{(c-1)\,q_{\nu-1}+q_{\nu-2}}{(c-1)\,p_{\nu-1}+p_{\nu-2}}\right| &< \left|\frac{(c-1)\,q_{\nu-1}+q_{\nu-2}}{(c-1)\,p_{\nu-1}+p_{\nu-2}} - \frac{cq_{\nu-1}+q_{\nu-2}}{cp_{\nu-1}+p_{\nu-2}}\right| \\ &= \frac{1}{\left((c-1)\,p_{\nu-1}+p_{\nu-2}\right)\left(cp_{\nu-1}+p_{\nu-2}\right)}. \end{aligned}$$

This implies

$$0 < \left|A\left((c-1)\,p_{\nu-1}+p_{\nu-2}\right) - B\left(cp_{\nu-1}+p_{\nu-2}\right)\right| < \frac{B}{cp_{\nu-1}+p_{\nu-2}}.$$

It follows that $B > cp_{\nu-1}+p_{\nu-2}$, because $\left|A\left((c-1)\,p_{\nu-1}+p_{\nu-2}\right) - B\left(cp_{\nu-1}+p_{\nu-2}\right)\right|$ is a nonzero natural number.
Case 2: $\frac{A}{B} > \xi_0$.
The proof is analogous and is left to the reader. □

In fact we can be more precise on which members of (3.23) and (3.24) are best rational approximation of ξ_0 of the second kind.

Lemma 3.17 *Consider the ratio*

$$\frac{cq_{\nu-1}+q_{\nu-2}}{cp_{\nu-1}+p_{\nu-2}},\ 0<c\leq a_\nu.$$

We have that if $2c>a_\nu+1$ *then this ratio is a best approximation of the second kind; and if* $2c\leq a_\nu-1$ *then this ratio is not a best approximation of the second kind.*

It is clear that $\frac{cq_{\nu-1}+q_{\nu-2}}{cp_{\nu-1}+p_{\nu-2}}$ is a best rational approximation of the second kind if

$$\left|\xi_0-\frac{cq_{\nu-1}+q_{\nu-2}}{cp_{\nu-1}+p_{\nu-2}}\right|<\left|\xi_0-\frac{q_{\nu-1}}{p_{\nu-1}}\right|,$$

which is equivalent with

$$\left|\frac{c\left(p_{\nu-1}\xi_0-q_{\nu-1}\right)+p_{\nu-2}\xi_0-q_{\nu-2}}{cp_{\nu-1}+p_{\nu-2}}\right|<\left|\frac{\xi_0p_{\nu-1}-q_{\nu-1}}{p_{\nu-1}}\right|.$$

We now use $\xi_\nu=\frac{p_{\nu-1}\xi_0-q_{\nu-2}}{p_{\nu-1}\xi_0-q_{\nu-2}}$, to obtain the equivalence with

$$\left|\frac{\left(c-\xi_\nu\right)\left(p_{\nu-1}\xi_0-q_{\nu-1}\right)}{cp_{\nu-1}+p_{\nu-2}}\right|<\frac{\xi_0p_{\nu-1}-q_{\nu-1}}{p_{\nu-1}},$$

and simplifies to

$$p_{\nu-1}\left|\left(c-\xi_\nu\right)\right|<cp_{\nu-1}+p_{\nu-2}.$$

We have that $c<a_\nu\leq\xi_\nu$ hence this is equivalent to

$$p_{\nu-1}\left(\xi_\nu-c\right)<cp_{\nu-1}+p_{\nu-2}. \tag{3.29}$$

Suppose that $2c>a_\nu+1$, then

$$2cp_{\nu-1}+p_{\nu-2}\geq\left(a_\nu+1\right)p_{\nu-1}+p_{\nu-1}=a_\nu p_{\nu-1}\leq\xi_\nu B_{\nu-1},$$

and (3.29) is fullfilled. Suppose now that $2c\leq a_\nu-1$, then

$$2cp_{\nu-1}+p_{\nu-2}\leq\left(a_\nu-1\right)p_{\nu-1}+p_{\nu-1}\leq b_\nu p_{\nu-1}\leq\xi_\nu p_{\nu-1},$$

and (3.29) is false.

Theorem 3.18 *Let* $\left(\frac{q_n}{p_n}\right)_{n\in\mathbb{N}}$ *be the continued fraction expansion of* $-\frac{\lambda_1}{\lambda_2}$*. For every* $N\in\mathbb{N}$ *there exists a* $C>0$ *small enough such that for all* μ *sufficiently close to* μ_0*, the normal form determined by equation (3.5) can be written as*

$$Y_\mu:\begin{cases}\dot{x} &= \lambda_1(\mu)x+x\displaystyle\sum_{(a,b)\in\mathbb{N}^2\setminus\{(0,0)\}}u_1^au_2^bh_{(a,b),1}(\mu)\\ \dot{y} &= \lambda_2(\mu)y+y\displaystyle\sum_{(a,b)\in\mathbb{N}^2\setminus\{(0,0)\}}u_1^au_2^bh_{(a,b),2}(\mu),\end{cases} \tag{3.30}$$

where $u_1=x^{p_{2N}}y^{q_{2N}}$, $u_2=x^{p_{2N+1}}y^{q_{2N+1}}$.

Proof: Remark that it is sufficient to give a proof for large numbers of N. We give a description of $\widetilde{\mathbf{B}}_\varepsilon = \left\{(k_1,k_2) \in N^2 | \left(-\frac{\lambda_1}{\lambda_2} - \varepsilon\right) k_1 \le k_2 \le \left(-\frac{\lambda_1}{\lambda_2} + \varepsilon\right) k_1 \right\}$. Therefore fix $\varepsilon > 0$ small enough and choose $N+1 \in \mathbb{N}$ the smallest integer for which one of the inequalities

$$\left|\frac{p_{2n}}{q_{2n}} + \frac{\lambda_1}{\lambda_2}\right| \le \varepsilon, \ \left|\frac{p_{2n+1}}{q_{2n+1}} + \frac{\lambda_1}{\lambda_2}\right| \le \varepsilon,$$

hold. Hence we have

$$\frac{\lambda_1}{\lambda_2} + \frac{p_{2N}}{q_{2N}} \le -\varepsilon \le 0 \le \varepsilon \le \frac{\lambda_1}{\lambda_2} + \frac{p_{2N+1}}{q_{2N+1}},$$

and

$$\frac{p_{2N}}{q_{2N}} \le -\frac{\lambda_1}{\lambda_2} - \varepsilon \le -\frac{\lambda_1}{\lambda_2} \le -\frac{\lambda_1}{\lambda_2} + \varepsilon \le \frac{p_{2N+1}}{q_{2N+1}}, \tag{3.31}$$

It is clear that N tends to infinity as ε tends to zero. Since $-\frac{\lambda_1}{\lambda_2} \in \mathbb{R} \setminus \mathbb{Q}$, there exists an $n \in \mathbb{N}$ such that

$$\max\left\{\left|\frac{p_{2n}}{q_{2n}} + \frac{\lambda_1}{\lambda_2}\right|, \left|\frac{p_{2n+1}}{q_{2n+1}} + \frac{\lambda_1}{\lambda_2}\right|\right\} \le \varepsilon.$$

Following remark 3.9, it is clear that

$$\begin{aligned} C_N : &= \left\{(k_1,k_2) \in \mathbb{N}^2 | \frac{p_{2N}}{q_{2N}} \le \frac{k_2}{k_1} \le \frac{p_{2N+1}}{q_{2N+1}}\right\} && (3.32) \\ &= \left\{(k_1,k_2) \in \mathbb{N}^2 | \, \exists a,b \in \mathbb{R}, \, (k_1,k_2) = a(p_{2N},q_{2N}) + b(p_{2N+1},q_{2N+1})\right\} \\ &= \left\{(k_1,k_2) \in \mathbb{N}^2 | \, \exists a,b \in \mathbb{N}, \, (k_1,k_2) = a(p_{2N},q_{2N}) + b(p_{2N+1},q_{2N+1})\right\}. && (3.33) \end{aligned}$$

From the definition of $\widetilde{\mathbf{B}}_\varepsilon$ and the definition of C_N in equation (3.32) and by equation (3.31) it is clear that $\widetilde{\mathbf{B}}_\varepsilon \subset C_N$. We have $\mathbf{B}_{C,1} \subset \widetilde{\mathbf{B}}_{\varepsilon,1}$, $\mathbf{B}_{C,2} \subset \widetilde{\mathbf{B}}_{\varepsilon,2}$, $\widetilde{\mathbf{B}}_{\varepsilon,1} = (1,0) + \widetilde{\mathbf{B}}_\varepsilon$, $\widetilde{\mathbf{B}}_{\varepsilon,2} = (0,1) + \widetilde{\mathbf{B}}_\varepsilon$, $\widetilde{\mathbf{B}}_\varepsilon \subset C_N$. If C is arbitrarily small and μ is close to μ_0, then ε is arbitrary small and N arbitrary large. Hence the normal $Y_\mu(x)$ can be expressed as

$$\begin{cases} \dot{x} = \lambda_1(\mu)x + x \displaystyle\sum_{k \in C_N \setminus \{(0,0)\}} x^{k_1} y^{k_2} h_{k,1}(\mu) \\ \dot{y} = \lambda_2(\mu)y + y \displaystyle\sum_{k \in C_N \setminus \{(0,0)\}} x^{k_1} y^{k_2} h_{k,2}(\mu), \end{cases}$$

where $g_{(0,0),1} = g_{(0,0),2} = 1$. Using the description of the cone C_N by formula (3.33) this can be made more explicit to obtain formula (3.30).

□

3.4 Closing the cone in a two dimensional situation

In this section we will focus on the analytic linearization of a two dimensional non-resonant saddle. This is a differential equation of the form

$$\left\{ \begin{array}{l} \dot{x} = \lambda_1 x + h(x,y) \\ \dot{y} = \lambda_2 y + k(x,y), \end{array} \right. \tag{3.34}$$

where λ_1, $-\lambda_2$ are positive real numbers, $-\frac{\lambda_1}{\lambda_2}$ is not a rational and h, k are analytic and second order in the variables.

Remark 3.19 *We do not allow parameter dependence in this section. The main reason is that the diophantine conditions that will appear later in this section cannot hold for parameters in an open set. It is not a problem to allow parameter dependence in the non-linear part, making things a little more technical, but we do not insist on it.*

In this section we want to fit, on one hand , the classical results by Siegel, Brjuno, etc. on analytic linearization (given some condition on the continued fraction of $-\frac{\lambda_1}{\lambda_2}$) and, on the other hand, the 'the formal cone'-like approach. The idea consists of 'removing terms in cones' iteratively such that the cones become more narrow each iteration, removing all terms effectively in the limit of this process. The main difficulty is to make sure that the domain of the limiting transformation does not shrink to zero. The idea of 'removing terms in cones' is the subject of [3]. However, in order to be able to iterate this idea, we need explicit estimates on the radius of convergence of the transformations corresponding to the cones. Therefore we will reprove some of the theorems from [3] with a contraction argument in order to obtain the needed sharp estimates. Let us be more precise and introduce some helpful notations.

Definition 3.20 *Let $C > 0$, we define the good set G_C and the bad set B_C*

$$\begin{aligned} G_C &= \{(k,j) \in \mathbb{N}^2 \times \{1,2\} |\ |k_1\lambda_1 + \lambda_2 k_2 - \lambda_1| > C|k_1 + k_2|, \ \textit{and}\ |k| \geq 2\} \\ B_C &= \{(k,j) \in \mathbb{N}^2 \times \{1,2\} |\ |k_1\lambda_1 + \lambda_2 k_2 - \lambda_j| \leq C|k_1 + k_2|, \ \textit{and}\ |k| \geq 2\}. \end{aligned}$$

Given $\delta > 0$, we use the following normed spaces of analytic functions:

$$\begin{aligned} \mathcal{A}_\delta &:= \{f = \sum_{|k|\geq 2} a_k x^k |\ ||f||_{\delta,\mathcal{A}} := \sum_{k \geq 2} |a_k| \delta^{|k|} < \infty\}, \\ \mathcal{B}_\delta &:= \{f = \sum_{|k|\geq 2} a_k x^k |\ ||f||_{\delta,\mathcal{B}} := \max\{||f||_{\delta,\mathcal{A}}, ||Df||_{\delta,\mathcal{A}}\} < \infty\}, \end{aligned}$$

we will drop the index $\mathcal{A}$ or $\mathcal{B}$ in the notation of the norms whenever no confusion is possible.

We want to transform equation (3.34) into

$$\begin{cases} \dot{x}_1 & = \lambda_1 x_1 + h_1(x_1, y_1) \\ \dot{y}_1 & = \lambda_2 y_1 + k_1(x_1, y_1), \end{cases} \tag{3.35}$$

where $h_1(x,y) = \sum_{(k,1)\in B_C} a_{k,j}x^k$ and $k_1(x,y) = \sum_{(k,2)\in B_C} b_{k,j}x^k$, by means of an analytic coordinate transform $x = x_1 + v_1(x_1, y_1)$, $y = y_1 + w_1(x_1, y_1)$. It is clear that at the formal level this series can be chosen such that $v_1(x,y) = \sum_{(k,1)\in G_C} c_{k,j}x^k$ and $w_1(x,y) = \sum_{(k,2)\in G_C} d_{k,j}x^k$. We explain that these series are actually convergent and we give a precise estimate on the radius of convergence of the transformation and their corresponding normal form (3.35), where $f = (h,k)$, $u = (v_1, w_1)$ and $g = (h_1, k_1)$ and $A(x,y) = (\lambda_1 x, \lambda_2 y)$. Let $\mathbf{x} = (x,y)$ and $\mathbf{x}_1 = (x_1, y_1)$. The initial equation (3.34) can be written in the form $\dot{\mathbf{x}} = A\mathbf{x} + f(\mathbf{x})$, the coordinate transform $\mathbf{x} = \mathbf{x}_1 + u(\mathbf{x}_1)$ turns this into equation $\dot{\mathbf{x}}_1 = (I + Du)^{-1}.(A+f)\circ(I+u)(\mathbf{x}_1) = A\,\mathbf{x}_1 + g(\mathbf{x}_1)$. Hence

$$\begin{aligned} & A + g = (I + Du)^{-1}.(A+f)\circ(I+u) \\ \Leftrightarrow\ & (I + Du).(A+g) = (A+f)\circ(I+u) \\ \Leftrightarrow\ & \mathcal{F}(u,g,f) = 0, \end{aligned}$$

where $\mathcal{F}$ is the functional

$$\begin{aligned} \mathcal{F} :& U \times V \times W \longrightarrow \mathcal{A}_\delta : (u,g,f) \mapsto Du.(A+g) - A.u + g - f\circ(\mathrm{id}+u), \\ X :&= \{u \in \mathcal{B}_\delta |\ u = \sum_{k\in G_C} u_{kj}x^k e_j\}; \\ U :&= \{u \in X |\ ||u||_{\delta,\mathcal{B}} < \delta_1\}; \\ V :&= \{g \in \mathcal{A}_\delta |\ g = \sum_{k\in B_C} g_{kj}x^k e_j\}; \\ W :&= \mathcal{A}_{\delta+\delta_1}; \end{aligned} \tag{3.36}$$

The idea in [3] is to find for a fixed f a solution $u(f)$, $g(f)$ of this functional equation for some $\delta > 0$ by means of an implicit function theorem. In order to do so the authors of [3] proved a somewhat different version of the following lemma.

Lemma 3.21 *The functional $\mathcal{F}$ is C^1 in a neighbourhood of the origin and the norm of $\mathcal{L} = D_{(u,g)}\mathcal{F}(0,0,0)^{-1}$ is bounded by $\frac{1}{C\delta}$, if $\delta < 1$.*

Proof: It follows from Proposition 2.11 that $(u,f,g) \mapsto f\circ(\mathrm{id}+u)$ is C^1. $Du.A - A.u$ is linear and continuous in u. $Du.g$ is bilinear and continuous. The presence of the term $Du.(A+g)$ is the main reason why there is a C^1-norm is chosen for the transformation u. The boundedness of $\mathcal{L}$ follows from:

$$\begin{aligned}
||Du.g|| &= \left|\left|\begin{pmatrix} \frac{\partial u_1}{\partial x_1} & & \cdots \frac{\partial u_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial u_n}{\partial x_1} & & \cdots \frac{\partial u_n}{\partial x_n} \end{pmatrix}.g\right|\right| \\
&= \sum_{\alpha=1}^{n}\sum_{j=1}^{n} ||\frac{\partial u_\alpha}{\partial x_j} g_j e_\alpha||_\delta \\
&\leq n^2 \max_{\alpha,j} ||\frac{\partial u_\alpha}{\partial x_j}||_\delta.||g_j||_\delta \\
&\leq n^2||Du||_\delta.||g||_\delta.
\end{aligned}$$

We calculate the Gateaux derivatives of $\mathcal{F}$ at $(0,0,0)$ and find

$$\begin{aligned}
D_u\mathcal{F}(0,0,0).u &= \lim_{t\to 0}\frac{tDu.A - tA.u}{t} = Du.A - A.u, \\
D_f\mathcal{F}(0,0,0).f &= \lim_{t\to 0}\frac{tf\circ \mathrm{id}}{t} = f, \\
D_g\mathcal{F}(0,0,0).g &= \lim_{t\to 0}\frac{-tg}{t} = -g. \qquad (3.37)
\end{aligned}$$

Hence $D_{(u,g)}\mathcal{F}(0,0,0).(\hat{u},\hat{g}) = D\hat{u}.A - A.\hat{u} - \hat{g}$. We invert this operator. Remark that $\mathcal{L} = D_{(u,g)}\mathcal{F}(0,0,0) : X \times V \to \mathcal{A}_\delta$. Take now $m = \sum_{j=1}^{n}\sum_{k\in\mathbb{N}^n} m_k x^k e_j \in \mathcal{A}_\delta$. Then $m = w + v$, where $w = \sum_{(k,j)\in\mathbf{G}_C} m_k x^k e_j$ and $v = \sum_{(k,j)\in\mathbf{B}_C} m_k x^k e_j$. Clearly $\mathcal{L}^{-1}.v = -v$, and $u = \mathcal{L}^{-1}.w = \sum_{(k,j)\in\mathbf{G}_C} \frac{m_{k,j}}{\langle\lambda,k\rangle-\lambda_j} x^k e_j$. We have:

$$\begin{aligned}
||u||_\delta &= ||\sum_{(k,j)\in\mathbf{G}_C} \frac{k_i.m_{k,j}}{\langle\lambda,k\rangle - \lambda_j} x^k e_j||_\delta \\
&\leq \frac{1}{C}||\sum_{(k,j)\in\mathbf{G}_C} m_{k,j}x^k e_j||_\delta \\
&\leq \frac{1}{C}||m||_\delta.
\end{aligned}$$

$$\begin{aligned}
||\frac{\partial u}{\partial x_i}||_\delta &= ||\sum_{(k,j)\in\mathbf{G}_C} \frac{k_i.m_{k,j}}{\langle\lambda,k\rangle-\lambda_j} x^{k-e_i} e_j||_\delta \\
&\leq \frac{1}{C}||\sum_{(k,j)\in\mathbf{G}_C} m_{k,j} x^{k-e_i} e_j||_\delta \\
&\leq \frac{1}{C}\sum_{(k,j)\in\mathbf{G}_C} |m_{k,j}|\delta^{|k|-1} \\
&\leq \frac{1}{C\delta}\sum_{(k,j)\in\mathbf{G}_C} |m_{k,j}|\delta^{|k|} \\
&\leq \frac{1}{C\delta}||m||_\delta.
\end{aligned}$$

□

We will use this lemma to construct a solution of $\mathcal{F}(u,g,f)=0$ for a fixed f by means of a fixed point construction in Section 3.4.1. In Section 3.4.2 we will then use the continued fraction expansion of $-\frac{\lambda_1}{\lambda_2}$ to show that the linearizing transformation of (3.34) is convergent in a lot of cases.

3.4.1 The fixed point construction

Suppose that we are given the functional

$$\mathcal{F}(u,g,f) = Du.(A+g) - A.u + g - f\circ(\mathrm{id}+u),$$

defined as in (3.36) and we want to solve for $u(f), g(f)$ for f small enough. We remark that the functional is well defined and C^1 in a neighbourhood of the origin if $\max(||Du||_\delta, ||u||_\delta) < \delta_1$, $||g||_\delta$ is well defined and $||f||_{\delta+\delta_1}$ is well-defined. We define $\mathcal{L} = d\mathcal{F}(0,0,0)$ and we prove that the mapping

$$T_f(u,g) = -\mathcal{L}^{-1}\left(Du.g - f\circ(\mathrm{id}+u)\right)$$

has a fixed point $(u,g) := T_f(u,g)$ whenever f is small enough by using contractive properties of $T_f(u,g)$. Let now $0<\delta_2<\delta_1$, $||u||_\delta < \frac{\delta_2}{2}, ||u'||_\delta < \frac{\delta_2}{2}$. We use the mean value theorem and Cauchy estimate and find that

$$\begin{aligned}
|f(x+u(x)) - f(x+u'(x))| &\leq \sup_{||\eta||\leq\delta+\delta_2} |Df(\eta)|.|u(x)-u'(x)| \\
&\leq \frac{||f||_{\delta+\delta_1}}{\delta_1-\delta_2}||u-u'||_\delta. \qquad (3.38)
\end{aligned}$$

Now let $||(u,g)||_\delta = \max\{||g||_\delta, ||u||_\delta, ||Du||_\delta\}$, then

$$\begin{aligned}
||T_f(u,g) - T_f(u',g')||_\delta &\leq ||\mathcal{L}^{-1}||.(||Du.g - Du'.g'||_\delta + ||f\circ(\mathrm{id}+u) - f\circ(\mathrm{id}+u')||_\delta) \\
&\leq ||\mathcal{L}^{-1}||.(||Du||_\delta||g-g'||_\delta + ||Du - Du'||_\delta||g'||_\delta \\
&\quad + ||f\circ(\mathrm{id}+u) - f\circ(\mathrm{id}+u')||_\delta) \\
&\leq ||\mathcal{L}^{-1}||.\Bigg(||Du||_\delta||g-g'||_\delta + ||Du-Du'||_\delta||g'||_\delta \\
&\quad + \frac{||f||_{\delta+\delta_1}||u-u'||_\delta}{\delta_1-\delta_2}\Bigg).
\end{aligned}$$

We now introduce a majorating scheme. In order to do so, we define the following: $M = ||\mathcal{L}^{-1}||$; $(u_0,g_0) = T_f(0,0)$; $(u_{n+1},g_{n+1}) := T_f(u_n,g_n)$; $\tilde{y}_n := ||T_f(u_n,g_n)||_\delta$ and $\tilde{x}_n := ||(u_{n+1},g_{n+1}) - (u_n,g_n)||_\delta = ||T_f(u_{n+1},g_{n+1}) - T_f(u_n,g_n)||_\delta$. Then clearly we have that

$$\begin{aligned}
\tilde{y}_{n+1} &= ||T_f(u_{n+1},g_{n+1})||_\delta = ||(u_{n+2},g_{n+2})||_\delta \\
&\leq ||(u_{n+2},g_{n+2}) - (u_{n+1},g_{n+1})||_\delta + ||(u_{n+1},g_{n+1})||_\delta \leq \tilde{x}_{n+1} + \tilde{y}_n \\
\tilde{x}_{n+1} &= ||T_f(u_{n+1},g_{n+1}) - T_f(u_n,g_n)|| \qquad (3.39)\\
&\leq M.\Bigg((||(u_{n+1},g_{n+1})||_\delta + ||(u_n,g_n)||_\delta)||(u_{n+1}-u_n, g_{n+1}-g_n)||_\delta \\
&\quad + \frac{||f||_{\delta+\delta_1}||u_{n+1}-u_n||_\delta}{\delta_1-\delta_2}\Bigg) \\
&\leq M(\tilde{y}_n\tilde{x}_n + \tilde{y}_{n-1}\tilde{x}_n) + \frac{M||f||_{\delta+\delta_1}}{\delta_1-\delta_2}\tilde{x}_n.
\end{aligned}$$

Hence it is natural to consider

$$\begin{cases} y_0 = x_0 = ||T_f(0,0)||_\delta = ||f||_\delta \\ y_{n+1} = x_{n+1} + y_n \\ x_{n+1} = \beta x_n + \alpha x_n y_n, \end{cases} \qquad (3.40)$$

where $\alpha := 2M$ and $\beta := \dfrac{M||f||_{\delta+\delta_1}}{\delta_1-\delta_2}$. Remark that whenever we find a convergent, positive solution (x_n,y_n) of the difference equation (3.40), it will be a majorant of $(\tilde{x}_n,\tilde{y}_n)$, due to the inequalities given by (3.39). Hence the corresponding (u_n,g_n) converges to some (u,g) (it is a Cauchy sequence in the Banach space $\mathcal{A}_\delta \times \mathcal{B}_\delta$) that is a solution of the fixed point problem $(u,g) = T_f(u,g)$. For sure we have to take into account that for all $n \in \mathbb{N}$ we maintain $y_n \leq \frac{\delta_2}{2}$ during the iteration process to make sure that the Cauchy estimate given by (3.38) and used in the estimate of equation (3.39) is valid. We first study the convergence of equation (3.40) by means of the following theorem. Afterwards we deal with the initial conditions and translate them in terms of the original problem.

Theorem 3.22 *Consider the recursively defined sequences*

$$\begin{cases} x_{n+1} = \beta x_n + \alpha x_n y_n, \\ y_{n+1} = y_n + \beta x_n + \alpha x_n y_n, \end{cases} \tag{3.41}$$

where $\alpha > 0$, $0 < \beta < 1$. Suppose that $x_0 = y_0 > 0$ and suppose that for certain $\beta < \beta' < 1$ we have that

$$\beta + \frac{\alpha x_0}{1-\beta'} \leq \beta',$$

which is clearly equivalent to

$$x_0 \leq \frac{(\beta' - \beta)(1-\beta')}{\alpha}. \tag{3.42}$$

Then the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are both convergent.

Proof: We show by induction that:

$$x_n \leq (\beta')^n x_0 \text{ and } y_n \leq \sum_{k=0}^{n} (\beta')^k x_0.$$

This is clearly true for $n = 0$. Suppose that this statement is true for all $0 \leq m \leq n$; then it is also true for $n+1$. Indeed:

$$y_n \leq \sum_{k=0}^{n} (\beta')^k x_0 \leq \frac{x_0}{1-\beta'}, \tag{3.43}$$

Hence

$$\begin{aligned} x_{n+1} = x_n(\beta + \alpha y_n) &\leq (\beta')^n x_0 (\beta + \frac{\alpha x_0}{1-\beta'}) \\ &\leq (\beta')^{n+1} x_0, \end{aligned}$$

and

$$\begin{aligned} y_{n+1} = y_n + x_n &\leq \sum_{k=0}^{n} (\beta')^k x_0 + (\beta')^{n+1} x_0 \\ &\leq \sum_{k=0}^{n+1} (\beta')^k x_0, \end{aligned}$$

This concludes the theorem. □

Corollary 3.23 *Consider the recursively defined sequences*

$$\begin{cases} x_{n+1} = \beta x_n + \alpha x_n y_n, \\ y_{n+1} = y_n + \beta x_n + \alpha x_n y_n, \end{cases}$$

where $\alpha > 0$, $0 < \beta < 1$. Suppose also that $x_0 = y_0 > 0$ and

$$x_0 \leq \frac{(1-\beta)^2}{4\alpha};$$

Then the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are both convergent. Moreover

$$y_n \leq \frac{2x_0}{1-\beta}. \tag{3.44}$$

Proof: The condition (3.42) is equivalent to

$$x_0 \leq \frac{(1-\beta)^2}{4\alpha};$$

because

$$\frac{(\beta'-\beta)(1-\beta')}{\alpha}$$

reaches its supremum at $\beta' = \dfrac{1+\beta}{2}$. Inequality (3.44) clearly follows from substituting $\beta' = \dfrac{1+\beta}{2}$ in inequality (3.43) in the proof of the previous lemma. □

Proposition 3.24 *$(T_f(u_n, g_n))_{n\in\mathbb{N}}$ converges if all of the following three properties hold:*

(i) $\beta < 1$

(ii) $y_n \leq \frac{\delta_2}{2}$*, for each $n \in \mathbb{N}$ (this is necessary to keep $||u_n||_\delta \leq \frac{\delta_2}{2}$ and $||g_n||_\delta \leq \frac{\delta_2}{2}$).*

(iii) $x_0 = y_0 = ||T_f(0,0)||_\delta = M||f||_\delta \leq \dfrac{(1-\beta)^2}{4\alpha}$

Proof: The above three conditions ensure that the estimates that lead to the majorating equation (3.40) are valid and that corollary 3.23 can be applied. □

Lemma 3.25 *The three properties in Proposition 3.24 are valid when*

(i) $||f||_{\delta+\delta_1} < \frac{\delta_1-\delta_2}{2M}$,

(ii) $||f||_{\delta+\delta_1} \leq \frac{\delta_2(\delta_1-\delta_2)}{4M\delta_1}$,

(iii) $||f||_{\delta+\delta_1} \leq \frac{\frac{1}{2}+M(\delta_1-\delta_2)-\sqrt{M^2(\delta_1-\delta_2)^2+M(\delta_1-\delta_2)}}{M}$.

This simplifies if $\delta_1 = 2\delta_2$ to

(i) $||f||_{\delta+\delta_1} \leq \frac{\delta_1}{8M}$,

(ii) $||f||_{\delta+\delta_1} < \frac{\delta_1}{8M}$,

(iii) $||f||_{\delta+\delta_1} \leq \frac{1+M\delta_1-\sqrt{M^2\delta_1^2+2M\delta_1}}{2M}$.

Proof: The first property in Proposition 3.24 is clearly equivalent to $||f||_{\delta+\delta_1} < \frac{\delta_1-\delta_2}{2M}$. For the second property we use (we choose $\beta' = \frac{1+\beta}{2}$, in the proof above)

$$y_n \leq x_0 \sum_{k=0}^{\infty} (\beta')^k \leq x_0 \frac{1}{1-\beta'} \leq x_0 \frac{2}{1-\beta}.$$

Hence the second property is valid when

$$\begin{aligned} M||f||_\delta = x_0 &\leq \frac{\delta_2(1-\beta)}{4} \\ \Leftrightarrow ||f||_\delta &\leq \frac{\delta_2(1-\beta)}{4M} \\ \Leftarrow ||f||_{\delta+\delta_1} &\leq \frac{\delta_2(1-\beta)}{4M} \\ \Leftrightarrow ||f||_{\delta+\delta_1} &\leq \frac{\delta_2}{4M}(1 - \frac{2M||f||_{\delta+\delta_1}}{\delta_1-\delta_2}) \\ \Leftrightarrow \frac{4M}{\delta_2}||f||_{\delta+\delta_1} + \frac{2M||f||_{\delta+\delta_1}}{\delta_1-\delta_2} &\leq 1 \\ \Leftrightarrow ||f||_{\delta+\delta_1} &\leq \frac{\delta_2(\delta_1-\delta_2)}{4M\delta_1}. \end{aligned}$$

Since $||f||_\delta \leq ||f||_{\delta+\delta_1}$, the third property is clearly valid if

$$\begin{aligned} ||f||_{\delta+\delta_1} &\leq \frac{(1-\beta)^2}{4M\alpha} \\ \Leftrightarrow 4M\alpha||f||_{\delta+\delta_1} &\leq (1 - \frac{2M||f||_{\delta+\delta_1}}{\delta_1-\delta_2})^2 \\ \Leftrightarrow 8M^2||f||_{\delta+\delta_1} &\leq 1 - \frac{4M||f||_{\delta+\delta_1}}{\delta_1-\delta_2} + \frac{4M^2||f||^2_{\delta+\delta_1}}{(\delta_1-\delta_2)^2}. \end{aligned} \tag{3.45}$$

We compute the roots of the equation

$$\begin{aligned} 4M\alpha x &= 1 - \frac{4Mx}{\delta_1-\delta_2} + \frac{4M^2x^2}{(\delta_1-\delta_2)^2} \\ \Leftrightarrow 1 - 4M\alpha x - \frac{4Mx}{\delta_1-\delta_2} &+ \frac{4M^2x^2}{(\delta_1-\delta_2)^2} = 0. \end{aligned}$$

We find two positive roots. The smallest is given by

$$\frac{\frac{1}{2} + M(\delta_1-\delta_2) - \sqrt{M^2(\delta_1-\delta_2)^2 + M(\delta_1-\delta_2)}}{M}.$$

Hence condition (3.45) is valid if

$$||f||_{\delta+\delta_1} \leq \frac{\frac{1}{2} + M(\delta_1 - \delta_2) - \sqrt{M^2(\delta_1 - \delta_2)^2 + M(\delta_1 - \delta_2)}}{M}$$

The conditions that appear in case $2\delta_2 = \delta_1$ are obvious. □

Summarizing the above, we have the following

Proposition 3.26 *$(T_f(u_n, g_n))_{n\in\mathbb{N}}$ converges if $\delta_1 = 2\delta_2$, δ_1 is smaller than $\frac{2}{5}$ and the following properties hold:*

$$||f||_{\delta+\delta_1} \leq \min\{\frac{1}{16M^2\delta_1}, \frac{\delta_1}{8M}\}. \tag{3.46}$$

Here, $M = \frac{1}{C\delta}$.

Proof: The Taylor series of $h(x) = 1 + x - \sqrt{x^2 + 2x}$ at infinity starts as $\frac{1}{8x} + \ldots$. Hence it is natural to consider the function $g(x) = 1 + x - \sqrt{x^2 + 2x} - \frac{1}{8x}$. The derivative of this function is $g'(x) = \dfrac{8x^2\sqrt{x\,(x+2)} - 8\,x^3 - 8x^2 + \sqrt{x\,(x+2)}}{8\sqrt{x\,(x+2)}x^2}$. Take now an arbitrary $x \geq 1$. Then we have the estimates

$$\begin{aligned}
8x^2\sqrt{x\,(x+2)} - 8\,x^3 - 8x^2 + \sqrt{x\,(x+2)} &\leq 8x^2(x+2) - 8\,x^3 - 8x^2 + x + 2 \\
&\leq -6x^3 - 6x^2 + x + 2 \\
&\leq -12x^2 + x + 2 \\
&\leq -11x + 2 \\
&\leq -9,
\end{aligned}$$

from which it follows that $g'(x) < 0$; and hence g is decreasing on $[1, \infty[$. Because $\frac{1}{10} < g(1) = \frac{15}{8} - \sqrt{3} < \frac{1}{5}$ and since

$$\lim_{x\to\infty} g(x) = 1 - \lim_{x\to\infty}(x - \sqrt{x^2 + 2x}) = 1 - \lim_{x\to\infty} \frac{-2x}{x + \sqrt{x^2 + 2x}} = 0,$$

it follows that $g(x) \geq 0$ for all $x \geq 1$ i.e. $h(x) \geq \frac{1}{8x}$. Since $h'(x) = 1 - \frac{x+1}{\sqrt{x^2+2x}} = 1 - \frac{\sqrt{x^2+2x+1}}{\sqrt{x^2+2x}} < 0$, h is decreasing on the interval $[0, 1]$ and hence $\frac{1}{10} \leq g(1) \leq h(1) \leq h(y)$, for each $y \in [0, 1]$. We can summarize that for each $x \geq 0$ we have that $\min(\frac{1}{10}, \frac{1}{8x}) \leq h(x)$. Hence also $\min\{\frac{1}{16M^2\delta_1}, \frac{1}{20M}\} \leq \frac{1+M\delta_1-\sqrt{M^2\delta_1^2+2M\delta_1}}{2M}$. We conclude that all three inequalities in Proposition 3.25 are valid if $||f||_{\delta+\delta_1} \leq \min\{\frac{1}{16M^2\delta_1}, \frac{\delta_1}{8M}\}$ and δ_1 is smaller than $\frac{2}{5}$. □

3.4.2 Narrowing the cone

We are now ready to prove the following theorem:

Theorem 3.27 *Let* $\left(\frac{q_n}{p_n}\right)_{n\in\mathbb{N}}$ *be the continued fraction approximants of* $-\frac{\lambda_1}{\lambda_2}$*. Suppose furthermore that*

$$\sum_{n\geq 2} \frac{|\ln(C_{n+1})|}{p_{n+1}+q_{n+1}} < \infty, \text{ where } C_{n+1} = \frac{|\lambda_1 p_{n+1} + \lambda_2 q_{n+1}|}{p_{n+1}+q_{n+1}+1}. \tag{3.47}$$

Then equation (3.34) can be linearized to

$$\begin{cases} \dot{x} &= \lambda_1 x \\ \dot{y} &= \lambda_2 y \end{cases}$$

by means of a convergent transformation.

Proof: Since the proof is rather technical we explain the idea of proof first. The idea is to give an iterative approach of the transformation

$$x + u(x) = (x + u_2(x)) \circ (x + u_3(x)) \circ \ldots$$

to the normal form by narrowing the cone a little further with each transformation. We will narrow the cones in close relation to the continued fraction expansion $\left(\frac{q_n}{p_n}\right)_{n\in\mathbb{N}}$ of the ratio $-\frac{\lambda_1}{\lambda_2}$. Indeed we will consider the subsequent bad sets (consisting of the indices (k, j) of corresponding terms $a_k x^k e_j$ that are not yet removed)

$$B_{C_n} = \{(k,j) \in \mathbb{N}^2 \times \{1,2\} | \ |k_1\lambda_1 + \lambda_2 k_2 - \lambda_j| \leq C_n|k_1 + k_2|, \text{ and } |k| \geq 2\}. \tag{3.48}$$

At each step of the procedure we have that $(x+u_2(x))\circ(x+u_3(x))\circ\ldots\circ(x+u_{n-1}(x))$ transforms the original vector field determined by $\dot{x} = Ax+f(x)$ to $\dot{x} = Ax+g_{n-1}(x)$. We then put $f_n := g_{n-1}$ and consider at this iteration step n the functional equation:

$$\mathcal{F}_{C_{n+1}}(u_n, g_n, f_n) = 0.$$

Here we have a function f_n that contains only terms in B_{C_n} and we look for a solution (u_n, g_n) such that g_n contains only terms in $B_{C_{n+1}}$. This solution is obtained by applying Proposition 3.26. Roughly speaking, this proposition establishes two things.

(i) It provides a bound to the solution

$$||g_n||_{\delta_n - \delta_1^{(n)}} \leq \frac{\delta_2^{(n)}}{2} = \frac{\delta_1^{(n)}}{4}, \ ||u_n||_{\delta_n - \delta_1^{(n)}} \leq \frac{\delta_2^{(n)}}{2} = \frac{\delta_1^{(n)}}{4}. \tag{3.49}$$

There is some freedom to choose the $\delta_1^{(n)}$, but one has to take into account a few constraints. First of all we will obtain that $\delta^{(n+1)} = \delta^{(n)} - \delta_1^{(n)}$, and since $\delta_{lim} := \lim_{n\to\infty} \delta_n$ will play the role of the radius of convergence of the transformation u_n, we want that $\delta_{lim} > 0$. Hence it makes sense to choose $\delta_1^{(n)} = \frac{\alpha}{2^n}$, where $\alpha < 2\delta$.

(ii) It provides a solution of $\mathcal{F}_{C_{n+1}}(u_n, g_n, f_n) = 0$ provided that $||f_n||_{\delta_n}$ satisfies the corresponding bound (3.46).

It is however not always true that $||f_n||_{\delta_n}$ satisfies the bound (3.46). Hence, we can not apply Proposition 3.26 directly because $||f_n||_{\delta_n}$ is possibly too large. Instead we will need to rescale the equation. We explain what we mean.

Define $R_\kappa : \mathbb{C}^n \longrightarrow \mathbb{C}^n : x \mapsto \kappa x$. We will extensively use the rescaling operator $\mathcal{R}_\kappa(f) = R_\kappa^{-1} \circ f \circ R_\kappa$. We do this because if we want to find a solution (u, g) of the equation $\mathcal{F}(u, g, f) = 0$ where $||f||_{\delta+\delta_1}$ is too large to apply Proposition 3.26, we can solve $\mathcal{F}(\tilde{u}, \tilde{g}, \mathcal{R}_\kappa(f)) = 0$ instead. It is then readily verified that $u = \mathcal{R}_\kappa{}^{-1}(\tilde{u})$ and $g = \mathcal{R}_\kappa{}^{-1}(\tilde{g})$ is a solution of $\mathcal{F}(u, g, f) = 0$. Moreover $||\mathcal{R}_\kappa(f)||_{\delta+\delta_1} \leq \kappa^{w-1}||f||_{\delta+\delta_1}$, if f is an analytic function for which $Df(0) = 0, \ldots, D^{w-1}f(0) = 0$ we call this number w the minimal order of f. This approach has the advantage that Proposition 3.26 can be applied if we choose κ small enough. Observe that if the minimal order w of f is big, the factor κ can be chosen rather large. To find a lower bound for w at each iteration step is actually the key to the solution. Such a bound is shown in a subsequent technical Lemma 3.28. We will also need the following technical remark concerning the numbers C_n: using Lemma 3.13 it follows that

$$\begin{aligned}
&\left|-\frac{\lambda_1}{\lambda_2} - \frac{q_{n+1}}{p_{n+1}}\right| < \left|-\frac{\lambda_1}{\lambda_2} - \frac{q_n}{p_n}\right| \\
\Longrightarrow &\left|\frac{\lambda_1 p_{n+1} + \lambda_2 q_{n+1}}{\lambda_2 p_{n+1}}\right| < \left|\frac{\lambda_1 p_n + \lambda_2 q_n}{\lambda_2 p_n}\right| \\
\Longrightarrow &\left|\frac{\lambda_1 p_{n+1} + \lambda_2 q_{n+1}}{p_{n+1}}\right| < \left|\frac{\lambda_1 q_n + \lambda_2 p_n}{q_n}\right| \\
\Longrightarrow & C_{n+1} = \left|\frac{\lambda_1 p_{n+1} + \lambda_2 q_{n+1}}{q_{n+1} + p_{n+1} + 1}\right| < \left|\frac{\lambda_1 p_n + \lambda_2 q_n}{q_n + p_n + 1}\right| = C_n
\end{aligned}$$

Geometrically this means that $B_{C_n} \subset B_{C_{n+1}}$ for all $n \in \mathbb{N}$.

Let us start the proof. We define for $n \geq N_0$

$$\delta_1^{(n)} = \frac{\alpha}{2^{n-1}},\ \delta_2^{(n)} = \frac{\alpha}{2^n}\ \alpha < \frac{\delta}{2}.$$

We will start the procedure from a certain N_0 that will be specified in a subsequent technical Lemma 3.28. We define

$$\delta^{(N_0)} = \delta - \delta_1^{(N_0)},\ \delta^{(n)} = \delta^{(n-1)} - \delta_1^{(n)},$$

for all $n > N_0$. It is readily verified that $\delta_{\lim} = \lim_{n\to\infty} \delta^{(n)} > 0$. Let $f_{N_0} := f \in \mathcal{A}_\delta$. We choose κ_{N_0} small enough in such a way that the norm of $\widetilde{f}_{N_0} := \mathcal{R}_{\kappa_{N_0}}(f_{N_0})$ is small enough to apply Proposition 3.26 to the equation $\mathcal{F}_{C_{N_0+1}}(\widetilde{u}_{N_0}, \widetilde{f}_{N_0+1}, \widetilde{f}_{N_0}) = 0$

with $\delta = \delta^{(N_0)}$ and $\delta_1 = \delta_1^{(N_0)}$. We obtain a solution satisfying the bounds

$$||\widetilde{u}_{N_0}||_{\delta^{(N_0)}} \leq \frac{\delta_2^{(N_0)}}{2} = \frac{\alpha}{2^{N_0+1}},$$
$$||\widetilde{f}_{N_0+1}||_{\delta^{(N_0)}} \leq \frac{\delta_2^{(N_0)}}{2} = \frac{\alpha}{2^{N_0+1}}.$$

Define $u_{N_0} = \mathcal{R}^{-1}_{\kappa_{N_0}}(\widetilde{u}_{N_0})$ and $f_{N_0+1} = \mathcal{R}^{-1}_{\kappa_{N_0}}(\widetilde{f}_{N_0+1})$.

Suppose now that we have defined $(u_n, \kappa_n, \widetilde{u}_n, \widetilde{f}_{n+1}, \gamma_n)$ that satisfy

$$\begin{cases} \widetilde{u}_n = \mathcal{R}_{\gamma_n}(u_n) \\ \widetilde{f}_n = \mathcal{R}_{\gamma_n}(f_n) \\ \gamma_n = \prod_{i=N_0}^{n} \kappa_i = \gamma_{n-1}\kappa_n \\ \delta^{(n)} = \delta - \sum_{i=N_0}^{n} \delta_1^{(i)} \\ ||\widetilde{u}_n||_{\delta^{(n)}} \leq \delta_1^{(n)} = \frac{\alpha}{2^{n-1}} \\ ||\widetilde{f}_{n+1}||_{\delta^{(n)}} \leq \delta_1^{(n)} = \frac{\alpha}{2^{n-1}} \\ \mathcal{F}_{C_{n+1}}(\widetilde{u}_n, \widetilde{f}_{n+1}, \widetilde{f}_n) = 0, \end{cases} \tag{3.50}$$

We define $(u_{n+1}, \kappa_{n+1}, \widetilde{u}_{n+1}, \widetilde{f}_{n+2}, \gamma_{n+1})$, and show that they satisfy the same equations where n is replaced by $n+1$. It is sufficient to determine κ_{n+1}, $\widetilde{u}_{n+1}$, $\widetilde{f}_{n+2}$, to check the two inequalities and the functional equation. One can define $\gamma_{n+1} = \gamma_n\kappa_{n+1}$, $\mathcal{R}^{-1}_{\gamma_{n+1}}(\widetilde{u}_{n+1}) = u_{n+1}$ and $\mathcal{R}^{-1}_{\gamma_{n+1}}(\widetilde{f}_{n+2}) = f_{n+2}$ afterwards.

We rescale $\widetilde{f}_{n+1}$ to $\mathcal{R}_{\kappa_{n+1}}(\widetilde{f}_{n+1})$ in order to be able to apply Proposition 3.26, with $\delta = \delta^{(n)} = \delta^{(n+1)} + \delta_1^{(n+1)}$, $\delta_1 = \delta_1^{(n+1)} = \frac{\alpha}{2^n}$, $C = C_{n+2}$, $f = \mathcal{R}_{\kappa_{n+1}}(\widetilde{f}_{n+1})$. Let w_{n+1} the minimal order of $\widetilde{f}_{n+1}$. We want κ_{n+1} to be chosen such that

$$\begin{aligned} ||\mathcal{R}_{\kappa_{n+1}}(\widetilde{f}_{n+1})||_{\delta^{(n+1)}} \leq ||\mathcal{R}_{\kappa_{n+1}}(\widetilde{f}_{n+1})||_{\delta^{(n)}} &\leq \kappa_{n+1}^{w_{n+1}-1}\frac{\alpha}{2^{n+2}} \\ &\leq \min\left\{\frac{C_{n+2}^2\left(\delta^{(n)}\right)^2}{16\delta_1^{(n+1)}}, \frac{C_{n+2}\delta^{(n)}\delta_1^{(n)}}{8}\right\}. \end{aligned}$$

A sufficient choice is

$$\kappa_{n+1} := \left(\min\left\{\frac{2^n C_{n+2}^2 \delta_{lim}^2}{4\alpha}, \frac{2^n C_{n+2}\delta_{lim}}{2\alpha}\right\}\right)^{\frac{1}{w_{n+1}-1}}.$$

We apply Proposition 3.26 with this choice and find solutions $u = \widetilde{u}_{n+1}$, $g = \widetilde{f}_{n+2}$ of $\mathcal{F}_{C_{n+2}}(\widetilde{u}_{n+1}, \widetilde{f}_{n+2}, \mathcal{R}_{\kappa_{n+1}}(f_{n+1})) = 0$, that are bounded by

$$||\widetilde{u}_{n+1}||_{\delta^{(n)}} \leq \frac{\delta_2^{(n+1)}}{2} = \frac{\delta_1^{(n+1)}}{4} = \frac{\alpha}{2^{n+1}} \leq \frac{\alpha}{2^n}$$
$$||\widetilde{f}_{n+2}||_{\delta^{(n)}} \leq \frac{\delta_2^{(n+1)}}{2} = \frac{\delta_1^{(n+1)}}{4} = \frac{\alpha}{2^{n+1}} \leq \frac{\alpha}{2^n}.$$

We suppose that $\lim_{n\to\infty}\gamma_n = \gamma$ exists and $\gamma > 0$ in order to finish the proof. We will check afterwards that this condition is equivalent to condition (3.47). We explain why the formal power series

$$x + u(x) := \lim_{n\to\infty} v_n(x) := \lim_{n\to\infty} (x + u_{N_0}(x)) \circ (x + u_{N_0+1}(x)) \circ \ldots \circ (x + u_{N_0+n}(x))$$

converges in a neighbourhood of the origin. Remark that $x + u(x)$ is well-defined as a power series because the minimal order of u_n tends to infinity whenever n tends to infinity. We show that $||\mathcal{R}_\gamma(v_n)||_{\delta_{lim}} \leq D$ for each n and a constant D independent of n. Remark first that

$$\mathcal{R}_\gamma(v_n)(x) = (x + \mathcal{R}_\gamma(u_{N_0})(x)) \circ (x + \mathcal{R}_\gamma(u_{N_0+1})(x)) \circ \ldots \circ (x + \mathcal{R}_\gamma(u_{N_0+n})(x)).$$

We use the estimate

$$\begin{aligned} ||(x + \mathcal{R}_\gamma(u_n)(x))||_{\delta_{lim}} &\leq ||(x + \mathcal{R}_{\gamma_n}(u_{N_0+n})(x))||_{\delta^{(N_0+n)}} \\ &\leq \delta^{(N_0+n)} + ||\mathcal{R}_{\gamma_n}(u_{N_0+n})(x)||_{\delta^{(N_0+n)}} \\ &\leq \delta^{(N_0+n)} + ||\widetilde{u}_{N_0+n}||_{\delta^{(N_0+n)}} \\ &= \delta^{(N_0+n)} + \frac{\alpha}{2^{N_0+n-1}} \leq \delta^{(N_0+n-1)}, \end{aligned}$$

and observe that

$$\begin{aligned} &||(x + \mathcal{R}_\gamma(u_{N_0})(x)) \circ (x + \mathcal{R}_\gamma(u_{N_0+1})(x)) \circ \ldots \circ (x + \mathcal{R}_\gamma(u_{N_0+n})(x))||_{\delta^{(N_0+n)}} \\ &\qquad \leq ||(x + \mathcal{R}_\gamma(u_{N_0})(x))||_{\delta^{(N_0)}} \leq \delta_{N_0} + \frac{\alpha}{2^{N_0-1}} =: D. \end{aligned}$$

It follows that $||\mathcal{R}_\gamma(v_n)||_{\delta_{lim}} \leq ||\mathcal{R}_\gamma(v_n)||_{\delta^{(N_0+n)}} \leq D$ for each $n \geq 0$. Hence also $||\mathcal{R}_\gamma(v)||_{\delta_{lim}} \leq D$ showing that v is analytic on $\bar{B}(0;\delta_{lim})$.

We still need to show that $\lim_{n\to\infty}\gamma_n = \gamma$ exists and $\gamma > 0$ if condition (3.47) holds. We will need a sharp bound on w_n first, proven in the subsequent lemma. The existence of the limit is then proven afterwards.

Lemma 3.28 *Let $l > 1$ and consider the formal power series $f_l = \sum_{(k,j)\in\mathbf{B}_{C_l}} a_k x^k e_j$. Define w_l to be the smallest natural number for which $D^{w_l} f_l \neq 0$; then there exists an $N_0 \in \mathbb{N}$ such that $w_l \geq \frac{p_{l+1}+q_{l+1}+1}{5}$ for each $l \geq N_0$.*

Proof: Throughout the proof we will suppose that $\lambda_1 > 0$ and $\lambda_2 < 0$; the case where $\lambda_1 < 0$ and $\lambda_2 > 0$ is analogous. We give a proof for l odd, i.e. $l = 2n+1$ for some $n \in \mathbb{N}$. The even case is similar. We define the following lines:

$$\begin{cases} L_1: & \lambda_1(k_1 - 1) + \lambda_2 k_2 = C_{2n+1}\,(k_1 + k_2) \\ L_2: & \lambda_1(k_1 - 1) + \lambda_2 k_2 = -C_{2n+1}\,(k_1 + k_2)\,, \end{cases}$$

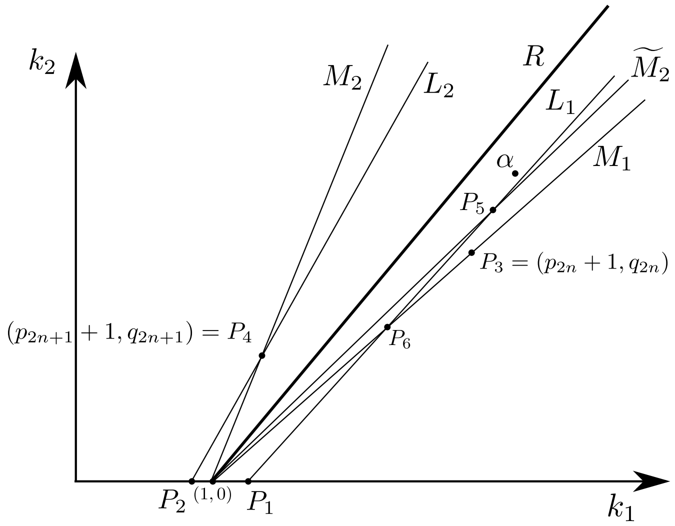


Figure 3.3: The cones

or, equivalently,

$$\begin{cases} L_1: & k_2 = \frac{C_{2n+1}-\lambda_1}{\lambda_2 - C_{2n+1}} k_1 + \frac{\lambda_1}{\lambda_2 - C_{2n+1}} \\ L_2: & k_2 = \frac{C_{2n+1}+\lambda_1}{-\lambda_2 - C_{2n+1}} k_1 - \frac{\lambda_1}{\lambda_2 - C_{2n+1}}. \end{cases}$$

L_1 has slope $\frac{C_{2n+1}-\lambda_1}{\lambda_2 - C_{2n+1}} < -\frac{\lambda_1}{\lambda_2}$ and intersection with the line $k_2 = 0$ at the point $P_1 = \left(\frac{-\lambda_1}{C_{2n+1}-\lambda_1}, 0\right)$. Remark that $\frac{-\lambda_1}{C_{2n+1}-\lambda_1} = \frac{\lambda_1}{-C_{2n+1}+\lambda_1} > 1$. L_2 has slope $\frac{C_{2n+1}+\lambda_1}{-\lambda_2 - C_{2n+1}} > -\frac{\lambda_1}{\lambda_2}$ and intersection with the line $k_2 = 0$ at the point $P_2 = \left(\frac{\lambda_1}{C_{2n+1}-\lambda_1}, 0\right)$. Remark that $\frac{\lambda_1}{C_{2n+1}-\lambda_1} < 1$. The lines L_1 and L_2 determine the bounds of the set

$$\begin{aligned} \mathbf{B}_{C_{2n+1},1} &= \{k \in \mathbb{N}^2 |\ |(k_1-1)\lambda_1 + \lambda_2 k_2| < C_{2n+1}|k_1 + k_2|, \text{ and } |k| \geq 2\}, \\ C_{2n+1} &= \frac{|\lambda_1 p_{2n+1} + \lambda_2 q_{2n+1}|}{p_{2n+1} + q_{2n+1} + 1}. \end{aligned}$$

We define the lines

$$\begin{cases} M_2: & k_2 = \frac{q_{2n+1}}{p_{2n+1}}(k_1 - 1) \\ M_1: & k_2 = \frac{q_{2n}}{p_{2n}}(k_1 - 1). \end{cases}$$

M_1 is the line passing through $(1,0)$ and $P_3 = (p_{2n}+1, q_{2n})$ and M_2 is the line passing through $(1,0)$ and $P_4 = (p_{2n+1}+1, q_{2n+1})$, see also Figure 3.3. If $n \geq N_0$, then the interior of the triangle with corner points $(1,0)$, P_2, P_4 does not contain any point of $\mathbb{Z}^2$: indeed, following remark 3.9 there is no positive integral point (this is a point $P=(x,y)$, where both x,y are positive integers) in the interior of this triangle if the determinant

$$\det\begin{vmatrix} p_{2n+1} & \frac{\lambda_1}{C_{2n+1}+\lambda_1}-1 \\ q_{2n+1} & 0 \end{vmatrix} = \frac{q_{2n+1}C_{2n+1}}{C_{2n+1}+\lambda_1} < 1.$$

Because $q_{2n+1}C_{2n+1} = q_{2n+1}\frac{|\lambda_1 p_{2n+1}+\lambda_2 q_{2n+1}|}{p_{2n+1}+q_{2n+1}+1} = |\lambda_1 p_{2n+1}+\lambda_2 q_{2n+1}|\frac{1}{1-\frac{\lambda_1}{\lambda_2}+\frac{1}{q_{2n+1}}} \to 0$, if n tends to ∞, this is true if $n \geq N_0$ for a certain N_0 large enough.

Using the same argumentation one shows that there exists no positive integral point in the interior of the triangle determined by $(1,0)$, P_1, P_3, and hence a fortiori there exists no integral point in the interior of the triangle determined by $(1,0)$, P_1 and P_6. Consequently, any positive integral point in $B_{C_{2n+1},1}$ is can be written as

$$(1,0)+a\,(p_{2n},q_{2n})+b\,(p_{2n+1},q_{2n+1})\,,\ \text{where } a,b\in\mathbb{N}\setminus\{0\}.$$

We define $\widetilde{M_2}$ to be the reflection of M_2 with respect to the line R; and $P_5 = \widetilde{M_2}\cap R$. Hence each positive integral point that lies in the set $B_{C_{2n+1},1}$ that is determined by the lines L_1 and L_2 lies either in the triangle Δ with corner points $(1,0), P_5, P_1$ or in the cone K_2 determined by the lines M_2 and $\widetilde{M_2}$. Let now $P_5 = (a,b)$. Consider now any point $P=(a_1,b_1)$ in Δ; then, because the slopes of $\widetilde{M_2}$ anc L_1 are positive, we clearly have that $a_1+b_1 \leq a+b$. We compute $\widetilde{M_2}$ and P_5. Since $\widetilde{M_2}$ is the reflection of M_2 with respect to R, it is sufficient to compute the reflexion of $(p_{2n+1}+1,q_{2n+1})\in M_2$. Define $\xi_0=-\frac{\lambda_1}{\lambda_2}$, then $R: k_2=\xi_0(k_1-1)$ and $n=(-\xi_0,1)/\sqrt{1+\xi_0^2}$ is a unit vector that is perpendicular to R. Hence the reflexion of the point $(p_{2n+1}+1,q_{2n+1})$ is given by

$$\left(p_{2n+1}+1+\frac{2\xi_0}{1+\xi_0^2}\left(-p_{2n+1}\xi_0+q_{2n+1}\right), q_{2n+1}-\frac{2}{1+\xi_0^2}\left(-p_{2n+1}\xi_0+q_{2n+1}\right)\right).$$

It follows that $\widetilde{M_2}:\ k_2=\beta(k_1-1)$, where

$$\beta=\frac{q_{2n+1}\left(1+\xi_0^2\right)-2\left(-p_{2n+1}\xi_0+q_{2n+1}\right)}{p_{2n+1}\left(1+\xi_0^2\right)+2\xi_0\left(-p_{2n+1}\xi_0+q_{2n+1}\right)}.$$

We find now

$$P_5=(a,b)=\left(\frac{-\lambda_2\beta-\lambda_1+C_{2n+1}\beta}{-\lambda_2\beta+C_{2n+1}-\lambda_1+C_{2n+1}\beta},\frac{-C_{2n+1}\beta}{-\lambda_2\beta+C_{2n+1}-\lambda_1+C_{2n+1}\beta}\right).$$

A straightforward computation results in

$$a+b=\frac{\left(p_{2n+1}+q_{2n+1}+1\right)\left(\lambda_1^2+\lambda_2^2\right)}{\lambda_1^2+\lambda_2^2+2\lambda_1\left(p_{2n+1}\lambda_1+q_{2n+1}\lambda_2\right)+2\lambda_2\left(p_{2n+1}\lambda_1+q_{2n+1}\lambda_2\right)}.$$

It follows that

$$\lim_{n\to\infty} \frac{a+b}{(p_{2n+1}+q_{2n+1}+1)} = 1.$$

As a consequence of this it follows that there exists an $N_0 \in \mathbb{N}$ such that for all $n \geq N_0$ and for all $P = (a_1, a_2) \in \Delta$ we have that

$$a_1 + a_2 \leq a + b < 2p_{2n+1} + 2q_{2n+1} + 1. \tag{3.51}$$

We consider two cases:
Case 1: There exist positive integral points in Δ.
Consider such an arbitrary point. We already observed that such a point equals

$$(1 + ap_{2n+1} + bp_{2n}, aq_{2n+1} + bq_{2n}),$$

for some $a, b \in \mathbb{N} \setminus \{0\}$. The weight(=the sum of the coordinates) of this point is given by $1 + ap_{2n+1} + bp_{2n} + aq_{2n+1} + bq_{2n}$. Since this point lies in Δ, (3.51) is valid, and hence $a = 1$. Hence the point with lowest weight in Δ is given by $(1 + p_{2n+1} + p_{2n}, q_{2n+1} + q_{2n})$. Clearly $Q = (1 + 2p_{2n+1} + p_{2n}, 2q_{2n+1} + q_{2n}) \notin \Delta$. Because the slope of the line determined by $(p_{2n} + 1, q_{2n})$ and $\mathbb{Q}$ is strictly larger than the slope of $\widetilde{M}_2$, every point $(\widetilde{c}p_{2n+1} + p_{2n}, \widetilde{c}q_{2n+1} + q_{2n})$ with $\widetilde{c} \geq 1$ is either in Δ or in the cone K_1 determined by the lines M_2 and $\widetilde{M}_2$. As a consequence Q must lie in the cone K_1 (since it is not in Δ). Using Lemma 3.17, we know that if $c \in \mathbb{N}$ and $2c > a_{2n+2} + 1$, then $\left|\frac{cq_{2n+1}+q_{2n}}{cp_{2n+1}+p_{2n}} - \xi_0\right| \leq \left|\frac{q_{2n+1}}{p_{2n+1}} - \xi_0\right|$ and hence $(cp_{2n+1} + p_{2n} + 1, cq_{2n+1} + q_{2n}) \in K_1$. Again from Lemma 3.17 it follows that if $c \in \mathbb{N}$ and $2c \leq a_{2n+2} - 1$, then $(cp_{2n+1} + p_{2n} + 1, cq_{2n+1} + q_{2n}) \notin K_1$. Because $Q = (1 + 2p_{2n+1} + p_{2n}, 2q_{2n+1} + q_{2n}) \in K_1$, we must have that $2.2 > a_{2n+1} - 1$, equivalent with $a_{2n+1} < 5$. The point with lowest weight in $B_{C_{2n+1},1}$ is in this case $\alpha = (p_{2n+1} + p_{2n} + 1, q_{2n+1} + q_{2n})$ and has weight $p_{2n+1} + p_{2n} + q_{2n+1} + q_{2n} + 1$. This weight satisfies

$$\begin{aligned} p_{2n+1} + p_{2n} + q_{2n+1} + q_{2n} + 1 &\geq \frac{1}{5}\left(5p_{2n+1} + p_{2n} + 5q_{2n+1} + q_{2n}\right) + 1 \\ &\geq \frac{1}{5}\left(p_{2n+2} + q_{2n+2}\right) + 1. \end{aligned}$$

Case 2: There are no positive integral points in Δ.
All positive integral points in $\mathbf{B}_{C_{2n+1},1}$ lie in the open cone K_1 determined by M_2 and $\widetilde{M}_2$. Since clearly $K_1 = \left\{(a+1, b) \in \mathbb{N}^2 |\ \left|\xi_0 - \frac{b}{a}\right| < \left|\xi_0 - \frac{q_{2n+1}}{p_{2n+1}}\right|\right\}$. The cone K_1 is contained in the cone $K_2 = \{(1,0) + a(p_{2n}, q_{2n}) + b(p_{2n+1}, q_{2n+1}) |\ a, b \geq 1\}$ determined by the lines M_2 and M_1. Using Lemma 3.17 we see that each point $(1,0) + (cp_{2n+1} + p_{2n}, cq_{2n+1} + q_{2n})$ is in K_1 if $2c > a_{2n+2} + 1$ and is not in K_1 if $2c \leq a_{2n+2} - 1$. It follows that the point with lowest weight in K_1 is given

by $(1,0)+(cp_{2n+1}+p_{2n},cq_{2n+1}+q_{2n})$, for a certain $c\in\mathbb{N}$ that is larger than $\max\{1,\frac{a_{2n+2}-1}{2}\}\geq\frac{a_{2n+2}}{3}$. Hence

$$\begin{aligned}w_{2n+1}&\geq 1+\frac{a_{2n+2}}{3}p_{2n+1}+p_{2n}+\frac{a_{2n+2}}{3}q_{2n+1}+q_{2n}\\&\geq\frac{p_{2n+2}+q_{2n+2}}{3}+1\geq\frac{p_{2n+2}+q_{2n+2}}{5}+1.\end{aligned}$$

Hence the set $\mathbf{B}_{C_{2n+1},1}$ has minimal weight larger than $\frac{1+p_{2n+2}+q_{2n+2}}{5}$ Completely analogous one shows that the same is true for the set $\mathbf{B}_{C_{2n+1},2}=\{(k,2)\in\mathbf{B}_{C_{2n+1}}\}$, finishing the proof. □

We finish the theorem by showing that $\lim_{n\to\infty}\gamma_n=\gamma<\infty$. Note that $\gamma_n=\prod_{k=N_0}^n\kappa_k$, and $\ln(\gamma_n)=\sum_{k=N_0}^n\ln(\kappa_k)$. κ_n is defined by

$$\kappa_n=\left(\min\left\{\frac{2^{n-1}C_{n+1}^2(\delta_{lim})^2}{4\alpha^2},\frac{2^{n-1}C_{n+1}\delta_{lim}}{2\alpha}\right\}\right)^{\frac{1}{w_n-1}}.$$

We have that for n large enough

$$\kappa_n\leq\left(\frac{2^{n-1}C_{n+1}^2(\delta_{lim})^2}{4\alpha^2}.\frac{2^{n-1}C_{n+1}\delta_{lim}}{2\alpha}\right)^{\frac{1}{w_n-1}}.$$

It is hence sufficient to show that the two sums in the right hand side of

$$\sum_{k=N_0}^{\infty}\ln(\kappa_k)=\sum_{n\geq N_0}\frac{1}{w_n-1}\ln\left(\frac{4^n\delta_{lim}^3}{2\alpha^3}\right)+\sum_{n\geq N_0}\frac{1}{w_n-1}\ln\left(C_{n+1}^3\right)\tag{3.52}$$

converge if condition 3.47 holds. We use the previous lemma to see that $w_n-1\geq p_n+q_n$ in both sums. We consider the first sum. Since $w_n-1\geq\frac{p_{n+1}+q_{n+1}}{5}$ and p_{n+1},q_{n+1} go to infinity at least as fast as the Fibonacci sequence, the first sum is obviously convergent. For the second sum we have

$$\left|\sum_{n\geq N_0}\frac{1}{w_n-1}\ln\left(C_{n+1}^3\right)\right|\leq 5\sum_{n\geq N_0}\frac{|\ln(C_{n+1})|}{p_{n+1}+q_{n+1}},$$

which converges by supposition. □

We finish this chapter by showing that the condition on $-\frac{\lambda_1}{\lambda_2}$ appearing in the heading of Theorem 3.27 holds if the Brjuno condition holds.

Proposition 3.29 *The Brjuno condition (3.20) implies*

$$\sum_{n\geq 1}\frac{|\ln(C_{n+1})|}{p_{n+1}+q_{n+1}}<\infty,\ \textit{where}\ C_{n+1}=\frac{|\lambda_1p_{n+1}+\lambda_2q_{n+1}|}{p_{n+1}+q_{n+1}+1}.$$

Proof: We have

$$-\sum_{n\geq 0}\frac{\ln(C_{n+1})}{p_{n+1}+q_{n+1}} = -\sum_{n\geq 0}\frac{\ln\left(\frac{|\lambda_1 p_{n+1}+\lambda_2 q_{n+1}|}{p_{n+1}+q_{n+1}+1}\right)}{p_{n+1}+q_{n+1}}$$

$$= -\sum_{n\geq 0}\frac{\ln\left(|\lambda_1 p_{n+1}+\lambda_2 q_{n+1}|\right)}{p_{n+1}+q_{n+1}} + \sum_{n\geq 0}\frac{\ln\left(p_{n+1}+q_{n+1}+1\right)}{p_{n+1}+q_{n+1}}.$$

We investigate the two sums on the right hand side of this equation separately. The last sum converges because $(p_n)_{n\in\mathbb{N}}$ and $(q_n)_{n\in\mathbb{N}}$ increase at least as fast as the Fibonacci series. For the first sum we use Lemma 3.12 which implies that

$$|\lambda_1 p_{n+1}+\lambda_2 q_{n+1}| \leq \frac{|\lambda_2|}{p_{n+2}}.$$

We also use Lemma 3.10 and obtain

$$\frac{1}{p_{n+1}+q_{n+1}} = \frac{1}{p_{n+1}\left(1+\frac{q_{n+1}}{p_{n+1}}\right)} \leq \frac{1}{p_{n+1}\left(1+\frac{q_0}{p_0}\right)}.$$

Hence

$$-\sum_{n\geq 1}\frac{\ln(|C_{n+1}|)}{p_{n+1}+q_{n+1}} \leq \left(\frac{1}{1+\frac{q_0}{p_0}}\right)\sum_{n\geq 1}\frac{\ln(p_{n+2})}{p_{n+1}} + \ln(|\lambda_2|)\sum_{n\geq 1}\frac{1}{p_{n+1}}.$$

It follows that if the Brjuno condition $\sum_{n\geq 1}\frac{\ln(p_{n+2})}{p_{n+1}} < \infty$ holds, then

$$\sum_{n\geq 1}\frac{|\ln(C_{n+1})|}{p_{n+1}+q_{n+1}} < \infty.$$

□

3.5 An invariant manifold problem

In this section we make a thorough investigation of the recursion process that we have encountered in Theorem 3.22:

$$\begin{cases} x_{n+1} = \beta x_n + \alpha x_n y_n, \\ y_{n+1} = y_n + \beta x_n + \alpha x_n y_n. \end{cases} \tag{3.53}$$

Remark that this kind of recursion is encountered in many conjugacy questions, not only the one in Section 3.4.1.

In a first stage we will show that the problem has a converging solution whenever $x_0 = y_0 > 0$ is small enough. A first result concerning this problem was proven in Theorem 3.22. In this paragraph, we sharpen the results, and investigate what happens if β tends to 1. We start with sharpening the result of Theorem 3.22.

Theorem 3.30 *Consider the recursively defined sequences*

$$\begin{cases} x_{n+1} = x_n(\beta + \alpha y_n), \\ y_{n+1} = y_n + x_n(\beta + \alpha y_n), \end{cases}$$

where $\alpha > 0$, $0 < \beta < 1$. Suppose also that $x_0 = y_0 > 0$. Now suppose that

$$f : \mathbb{R} \to \mathbb{R}^+ : x \longmapsto f(x) \tag{3.54}$$

is a non-increasing strictly positive function for which

$$\frac{f(n+1)}{f(n)} \geq \left(\beta + \alpha \left(\int_0^n f(x)\, dx + f(0)\right)\right) \tag{3.55}$$

holds for each $n \in \mathbb{N}$. Suppose now that $x_0 = y_0 = f(0)$, then the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are both convergent whenever the integral $\int_0^{+\infty} f(x)\, dx$ converges.

Proof: We show by induction on n that

$$\begin{cases} x_n \leq \beta_0\beta_1 \ldots \beta_n, \\ y_n = y_n + \sum_{k=0}^{n}(\beta_0\beta_1 \ldots \beta_k), \end{cases} \tag{3.56}$$

where $\beta_n = f(n)/f(n-1)$ $(n \neq 0)$ and $\beta_0 = f(0)$. Suppose that this is already proven for a certain n, we now prove this fact for $n+1$. Because f is a non-increasing function, it is clear that

$$\sum_{k=0}^{n} f(k) \leq \int_0^n f(x)\, dx + f(0).$$

Hence it follows from (3.55) that

$$\beta_{n+1} = \frac{f(n+1)}{f(n)} \geq \beta + \alpha \sum_{k=0}^{n} f(k) = \beta + \alpha \sum_{k=0}^{n}(\beta_0\beta_1 \ldots \beta_k)$$

Since

$$x_{n+1} = x_n(\beta + \alpha y_n) \leq \beta_0 \ldots \beta_n(\beta + \alpha \sum_{k=0}^{n}(\beta_0\beta_1 \ldots \beta_k)) \leq \beta_0 \ldots \beta_n\beta_{n+1}$$

$$y_{n+1} = y_n + x_{n+1} \leq \sum_{k=0}^{n}(\beta_0\beta_1 \ldots \beta_k) + \beta_0 \ldots \beta_n\beta_{n+1} \leq \sum_{k=0}^{n+1}(\beta_0\beta_1 \ldots \beta_k)$$

The statement of the theorem now follows since $\sum_{k=0}^{+\infty} f(k)$ converges whenever the corresponding integral $\int_0^{+\infty} f(x)\, dx$ converges. □

Remark 3.31 *For $f(x) = \dfrac{1}{(x+p)^{1+\epsilon}}$, $\epsilon > 0$ we find (numerically) that for $0 < \beta < 1$ and α small, p large the condition (3.55) is satisfied. It seems however not easy to find α and p explicitly in terms of β (hence hoping for a large starting value of x_0).*

3.5.1 Finding a trapping region

Let's reconsider the equations (3.53). So far we showed that for $0 \leq x_0 = y_0 \leq \dfrac{(1-\beta)}{4\alpha}$ the sequences x_n and y_n are convergent. In this paragraph we show that we can do better. First remark that we can restrict to the case $\alpha = 1$. Indeed, a simple scaling $\tilde{x}_n = \alpha x_n$, $\tilde{y}_n = \alpha y_n$ gives the following equation:

$$\begin{cases} x_{n+1} = \beta x_n + x_n y_n, \\ y_{n+1} = y_n + \beta x_n + x_n y_n, \end{cases} \tag{3.57}$$

where we renamed $\tilde{x}_n$, $\tilde{y}_n$ back to x_n and y_n. We introduce the corresponding diffeomorphism

$$F(x, y) = (x(\beta + y), y + x(\beta + y))$$

Remark that $F(0, y) = (0, y)$ for every y, so we have a line of fix-points.

Furthermore, it is easily seen that the sequence $(y_n)_{n\in\mathbb{N}}$ diverges whenever it exceeds $1 - \beta$. Indeed from that point on $x_{n+1} = x_n(\beta + y_n) > x_n$. Hence one can ask the following question: given $0 \leq y_0 \leq 1 - \beta$ fixed; what is the biggest value of x_0 such that $\lim_{n\to\infty} F^{\circ n}(x_0, y_0)$ converges. In order to study this question, we translate the point $(0, 1 - \beta)$ to the origin. Hence we define

$$G(x, y) = F(x, y + 1 - \beta) - (0, 1 - \beta) = (x + xy, y + x + xy). \tag{3.58}$$

This equation has the extra advantage that it does not depend on β.

Since whenever a starting point (x_0, y_0) has the property that if $G^{\circ n}(x_0, y_0)$ has a y-coordinate bigger than zero, then the corresponding sequence $\{G^{\circ n}(x_0, y_0)\}_{n\in\mathbb{N}}$ diverges, it seems that an invariant manifold $x = h(y)$ must exist, such that every point (x_0, y_0) on that manifold with starting coordinate $y_0 < 0$ will converge to $(0, 0)$ under forward iteration of G. Any point on that manifold with starting coordinate $y_0 > 0$ will converge to $(0, 0)$ under backward iteration of G.

If such an invariant manifold exists (we will show later on that this is the case; and it is C^∞, summable in the real direction, but not analytic) then it satisfies:

$$h(y) + yh(y) = h(h(y) + y + yh(y)).$$

The first few coefficients of the Taylor expansion of the formal solution can be computed. One finds that

$$\begin{aligned}
h(y) = {}& \frac{1}{2}y^2 - \frac{5}{12}y^3 + 0.3958333333\,y^4 - 0.3937500000\,y^5 + 0.3971064815\,y^6 \\
& - 0.4003720238\,y^7 + 0.4014663938\,y^8 - 0.4004878220\,y^9 + 0.3988441484\,y^{10} \\
& - 0.3981564720\,y^{11} + 0.3990716391\,y^{12} - 0.4006322184\,y^{13} + 0.4009436547\,y^{14} \\
& - 0.3991413019\,y^{15} + 0.3971210333\,y^{16} - 0.3983918860\,y^{17} + 0.4031488438\,y^{18} \\
& - 0.4043911924\,y^{19} + 0.3946329054\,y^{20} - 0.3839675917\,y^{21} + 0.4026071848\,y^{22} \\
& - 0.4488182650\,y^{23} + 0.4220720384\,y^{24} - 0.2365457604\,y^{25} + 0.2034570814\,y^{26} \\
& - 0.9596286694\,y^{27} + 1.663211180\,y^{28} + 1.513264784\,y^{29} - 7.502382125\,y^{30} \\
& - 5.954470260\,y^{31} + 51.48174429\,y^{32} + 1.536214833\,y^{33} - 348.1644985\,y^{34} \\
& + 214.6279884\,y^{35} + 2524.258166\,y^{36} - 3397.040393\,y^{37} - 19332.92241\,y^{38} \\
& + 43302.76372\,y^{39} + 155095.1356\,y^{40} - 534241.9666\,y^{41} - 1276536.102\,y^{42} \\
& + 6720479.161\,y^{43} + 10321819.55\,y^{44} - 88003202.03\,y^{45}. \qquad (3.59)
\end{aligned}$$

Remark 3.32 *This is a series with coefficients that have a Gevrey-behaviour. Notice how one can be tempted to conjecture analyticity on the series by only looking at the first 27 coefficients in the Taylor-series. Such behaviour is well known to exist for Gevrey-type power series. Indeed the series* $\sum_{n=1}^{\infty} n! R^n x^n$ *has coefficients that are smaller than 1 until order* N*, if* $R = \frac{1}{N}$*. This is also the reason why one should avoid rescaling transformations in the numerical investigation of a problem where one expects a Gevrey-type solution: a sufficiently large rescaling will hide the Gevrey-growth at the beginning of the series.*

We make a picture (Figure 3.4) of the subsequent Taylor series cut-off of degree N of $h(y)$. Notice how the first few approximations lie close to the invariant manifold for a relative large distance.

We look for a trapping region, i.e. a set of points D for which $G(D) \subset D$. The idea is that the curve $y^2/2$ is close to the invariant manifold, we try to use this curve in order to find a trapping region. We now show that the region D bounded by the curves $L = \{(x, y) | x = y^2/2,\ x \geq 0\}$, $y = c$ for a certain $-1 \leq c < 0$ and the negative y-axis is a trapping region. It is sufficient to check that the image of L under G remains in this region. We have the following

Lemma 3.33 *Suppose that* $(x, y) \in L$*,* $-1 \leq y < 0$ *then* $G(x, y)$ *lies in* D*.*

Proof: Suppose (x, y) satisfies the above properties i.e. $(x, y) = (y^2/2, y)$ We need to

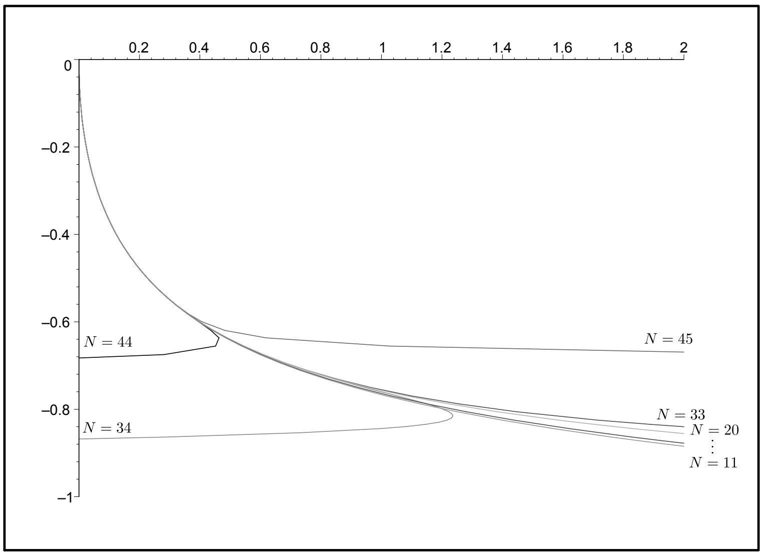


Figure 3.4: The Taylor approximations of $h(y)$

check that $G(x, y) = (x + xy, x + y + xy)$ satisfies

$$\begin{aligned} x + xy < \frac{(x + y + xy)^2}{2} &\Leftrightarrow \\ \frac{y^2}{2} + \frac{y^3}{2} < \frac{(y + \frac{y^2}{2} + \frac{y^3}{2})^2}{2} &\Leftrightarrow \\ y^4(-\frac{5}{8} - \frac{1}{4}y - \frac{1}{8}y^2) < 0 & \end{aligned}$$

which is clearly ok. □

We give an alternative version of proof, which can be used for higher order approximations.

Proof: The image of the curve C_1 determined by $x = h(y)$ is the curve C_2, $(\frac{y^2}{2} + \frac{y^3}{2}, \frac{(y + \frac{y^2}{2} + \frac{y^3}{2})^2}{2})$ parameterized by $y < 0$. We can parameterize this curve by putting $\alpha(y) = y + y^2/2 + y^3/2 = t$. It is clear that $\alpha'(y) = 1 + y + 3y^2/2 > 0$. Hence it is increasing. It follows easily that α is a bijection between $\{y < 0\}$ and $\{t < 0\}$. Hence a valid parameterization of C_2 is given by $(t - \alpha^{-1}(t), t)$. If we can show that the 'horizontal difference' between C_1 an C_2 is positive, then we are done.

This difference is given by

$$v(t) = \frac{t^2}{2} - (t - \alpha^{-1}(t))$$

As α is a bijection between $y < 0$ and $t < 0$, this is equivalent with showing that $v(\alpha(y)) > 0$. Now $v(\alpha(y)) = \frac{y^4}{8}(y^2 + 2y + 5)$; which is clearly positive. □

Remark 3.34 *The same type of proof can be adapted for higher order approximations of the invariant manifold. This was done by hand for the approximation* $x = \frac{1}{2}y^2 - \frac{5}{12}y^3$*, for higher order approximations, there is numerical evidence.*

Corollary 3.35 *Let D be the region as described in Lemma 3.33. For every point (x, y) in this region the sequence $(G^{\circ n}(x, y))_{n\in\mathbb{N}}$ converges. The translation of this property to our original problem defined by equation (3.57) implies that whenever* $x_0 \leq \frac{(1-\beta)^2}{2}$, $y_0 = 0$ *this system converges. This is an improvement by a factor two of the result stated in Theorem 3.22.*

Proof: The corresponding sequence of $(y_n)_{n\in\mathbb{N}}$ is increasing and bounded. The corresponding sequence of $(x_n)_{n\in\mathbb{N}}$ approaches 0. □

3.5.2 The existence of an invariant manifold

In a first subsection we explain the existence of a C^∞-manifold of the form

$$x = y^2/2 + O(y^3).$$

In a second subsection we explain why the manifold is of Gevrey-type. And in a third subsection we explain why the manifold is 1-summable in the real direction. In the last section we explain how the summable character of the Taylor series can be used to give good estimates concerning our manifold. One could argue that the first two subsections are overhead: they contain only weaker results. This is true, but we want to stress the existence of some gaps in the literature (at least we think, as we do not know any references) concerning Gevrey-type/summable embeddings of diffeomorphisms in a flow.

The manifold exists and is C^∞

We use a result of [23] which states that any diffeomorphism with nilpotent linear part of class C^∞ can be embedded in a C^∞-flow. Moreover, the first few terms of the Taylor series of the corresponding vector field X can be calculated and are given by:

$$\left(-\frac{x^2}{2} + xy - \frac{y^2x}{2} + \frac{yx^2}{3} + O(4)\right)\frac{\partial}{\partial x} + \left(x - \frac{x^2}{6} + \frac{xy}{2} + \frac{x^3}{20} - \frac{y^2x}{6} + O(4)\right)\frac{\partial}{\partial y}$$

Using a blow-up transformation $x = v^2(u+1)/2$, $y = v$ we find the vector field

$$\left(-vu - \frac{5v^2}{4} - vu^2 - 2v^2u + \frac{3v^3}{4} + vO(3)\right)\frac{\partial}{\partial u} + \left(\frac{v^2}{2} + \frac{v^2u}{2} + \frac{v^3}{4} + vO(3)\right)\frac{\partial}{\partial v}. \tag{3.60}$$

The idea of the transformation is that it maps the parabola $x = cy^2/2$ onto the line $u = c - 1$. (Remember that the formal expansion of our invariant manifold starts with $x = y^2/2$.) The above vector field is equivalent with

$$\left(-u - \frac{5}{4}v - u^2 - 2vu + \frac{3}{4}v^2 + O(3)\right)\frac{\partial}{\partial u} + \left(\frac{1}{2}v + \frac{1}{2}vu + \frac{1}{4}v^2 + O(3)\right)\frac{\partial}{\partial v}. \tag{3.61}$$

This vector field is a hyperbolic saddle: its linear part has an eigenvector -1 with corresponding eigenvector $(1,0)$ and an eigenvector $1/2$ with corresponding eigenvector $(-6/5, 1)$. Since the vector field is C^∞ it has a C^∞ unstable manifold $v = g(u)$; in (x,y)-coordinates this manifold is expressed as $x = y^2(1+g(y))/2$. In (u,v)-coordinates the first order approximation of our manifold is $v = -6u/5$, which gives, translated in (x,y)-coordinates $x = y^2/2 - 5y^3/12$ in correspondence with the formal invariant manifold (3.59).

Further literature study revealed that the same type of results where also obtained in [14].

Remark 3.36 *Using this method it is quite easy to show the existence of a C^∞ invariant manifold. But because we are not familiar with a general result showing that the embedding of a diffeomorphism with a nilpotent linear part into a vector field X is of Gevrey-type; the technique of embedding in a vector field cannot directly be used to conclude that the invariant manifold is also of Gevrey-type. We show in the next paragraph how one can obtain this result.*

The manifold is Gevrey

We again use a blow-up method. This time we first make a blow-up of the diffeomorphism and embed it in a vector field afterwards. The idea is to use a transformation that straightens the invariant manifold; and then shift it back to the origin. So we consider the transformation

$$g(u,v) = \left(\frac{v^2(1+u)}{2}, v\right).$$

This transforms the diffeomorphism into $h(u,v)$ given by:

$$\left(u - uv - \frac{5v^2}{4} - vu^2 - \frac{7v^2u}{4} + \frac{3v^3}{4} + O(4), v + \frac{v^2}{2} + \frac{v^2u}{2} + \frac{v^3}{2} + \frac{v^3u}{2} + O(4)\right).$$

We are now in position to use a result of [8], which shows that there exists a Gevrey-1 vector field $X(u,v)$ for which the time one map is exactly $h(u,v)$. Since we used the same blow-up transformation as in the previous section, this vector field is exactly the same as (3.60). Hence this time we can conclude the equivalent vector field (3.61) is also Gevrey-1. Now it is a direct consequence of Theorem 1 in [3], (using $s = 1$, and the bad set $S = \{1\} \times \{(0, k_2)|k_2 \in \mathbb{N}\}$) that the formal unstable manifold is Gevrey-1. Suppose now that this manifold is parameterized by $u = \alpha(v)$, then, in original coordinates this manifold can be written as $x = \frac{y^2(\alpha(y)+1)}{2}$. Hence the manifold is also Gevrey-1 in the original coordinates.

Remark 3.37 *Using this method we conclude that the manifold is Gevrey-1. We can however not draw any conclusion on the summability using this technique.*

The manifold is summable

Because the associated vector field (3.60) has the origin as point that is formally of Briot and Bouquet-type, we can apply directly Corollary 6.2 in the thesis of L. Lopez Hernanz [39] to conclude that the manifold $u = \alpha(v)$ is summable in every direction, except the two imaginary ones.

In fact we have tried to prove this result independently using a somewhat different idea. However the techniques that appear in [39] are more natural and more generally applicable then the unfinished result we had in mind.

Chapter 4

Normal forms for vector fields with nilpotent linear part

4.1 Introduction and statement of the results

We consider vector fields where the linear part at a singularity is nilpotent, with no restriction on the dimension. This, for example, includes the case of a coupled Takens-Bogdanov system, see e.g. [41]. See [46] for an introduction to the subject.

We briefly give some history (non-exhaustive) of the subject. In [61] the planar case $y\frac{\partial}{\partial x} + \ldots$ was considered and a formal normal form $(y + a(x))\frac{\partial}{\partial x} + b(x)\frac{\partial}{\partial y}$ was derived. It was shown in [59], also in the planar case, that an analytic vector field with nilpotent linear part $y\frac{\partial}{\partial x}$ can be analytically transformed to a normal form. Other results related to the planar case are in [11].

More recently it was shown in [37] that the analytic vector fields with linear part $y\frac{\partial}{\partial x} + z\frac{\partial}{\partial y}$ can be Gevrey-1 reduced to a normal form using a specific normal form procedure that is also described later on in this chapter. This framework was extended in [37] to the case of quasihomogeneous vector fields. In [37] it is explained what the generalization of the so called small denominators are for non-diagonal linear vector fields (and more general quasihomogeneous vector fields); and some results of convergence and Gevrey-1 normalization are explained (See also Theorems 4.4 and 4.5).

Somewhat before that, we have the results of [17] on the formal structure of the normal forms with a nilpotent linear part, using representation theory of $\mathrm{sl}(2, \mathbb{C})$. More recently [60] and [41] have also made contributions to this subject in the multidimensional case, on the formal level.

The purpose of this chapter is to combine both ideas : we will show how to use representation theory of $\mathrm{sl}(2, \mathbb{C})$ in nilpotent cases in order to calculate the small denominators in the framework of [37] and hence obtain qualitative information on the growth of coefficients appearing inside the normal form procedure.

In Section 4.2 of this chapter we repeat some results of the framework created in [37], in order for this text to be self-contained. In Section 4.3 we state some results on the representation theory of $\mathrm{sl}(2,\mathbb{C})$. In Section 4.4 we prove some propositions that lead to the main result, stated as Theorem 4.1 below and proven in Section 4.5.

Let us state the main result. We say that the linear part N of a vector field X is nilpotent at 0 if it acts as a nilpotent linear operator on the space of polynomials of degree δ, for each $\delta \in \mathbb{N} \setminus \{0\}$. Note that this means, up to a linear change of the coordinates, that the linear part of the vector field can be written as $N = \sum_{i=1}^{n-1} a_i x_{i+1} \frac{\partial}{\partial x_i}$, for certain $a_1, \ldots, a_{n-1} \in \mathbb{R}$.

Theorem 4.1 *Let $\alpha \geq 0$. Every formal Gevrey-α vector field $X = N + R$, where N is a nilpotent linear part and R is a part of higher order, admits a formal Gevrey-$(1+\alpha)$ transformation to Gevrey-$(1+\alpha)$ normal form. If the Gevrey-α vector field is formally linearizable, then the transformation and corresponding normal form are Gevrey-α.*

Remark that the cases where $N = y\frac{\partial}{\partial x}$ and $N = (y\frac{\partial}{\partial x} + z\frac{\partial}{\partial y})$ have already been treated in [34]. This theorem provides a generalization and a geometric explanation using representation theory of $\mathrm{sl}(2,\mathbb{C})$ of these examples. Considering the results in [59] and [40] one could wonder whether or not the given normal form actually converges, when X is analytic (i.e. $\alpha = 0$). We think however that, in general, this is not the case.

4.2 Background and notation

We recall some standard preliminaries about the used normal form procedure. We follow the outline of [37].

4.2.1 Setting

Let $X = N + R$ be a formal vector field in the neighbourhood of the origin, N its linear part and R the part or order ≥ 2. We will look for a coordinate transform $\Phi^{-1} = I + U$, U of order ≥ 2, such that the pullback $\Phi_*(X) = X' = N + R'$. A minor calculation shows that

$$\begin{aligned} &\Phi_*(X) = N + R' \\ \Leftrightarrow &X \circ \Phi^{-1} = D\Phi^{-1}.X' \\ \Leftrightarrow &(S + R) \circ (I + U) = D(I + U).(N + R') \\ \Leftrightarrow &R' + [U, N] = R(I + U) - DU.R'. \end{aligned} \tag{4.1}$$

Now we are going to determine the terms of order δ for the formal series $U = U_2 + U_3 + U_4 + \ldots$ and $R' = R'_2 + R'_3 + R'_4 + \ldots$ by induction. Therefore suppose that we already know $U_2, \ldots, U_{\delta-1}$ and $R'_2, \ldots, R'_{\delta-1}$. We take the projection of the terms of order δ in (4.1) and obtain:

$$R'_\delta + [U_\delta, N] = RHS_\delta.$$

where RHS_δ denotes the projection of order δ of the right hand side of (4.1) and depends only on U_l, R'_l with index strictly smaller than δ. Therefore it is natural to introduce the Lie-operator

$$d_{0,\delta} : \mathcal{V}_\delta \longrightarrow \mathcal{V}_\delta : U_\delta \mapsto [U_\delta, N];$$

where

$$\mathcal{V}_\delta = \left\{ \sum_{i=1}^{n} P_i \frac{\partial}{\partial x_i} | P_i \in \mathcal{P}_{\delta+1} \right\}, \tag{4.2}$$

$$\mathcal{P}_\delta = \{P |\, P \text{ is a homogeneous polynomial of degree } \delta\}, \tag{4.3}$$

and decompose the space $\mathcal{V}_\delta$ of vector fields of degree δ as $\mathcal{V}_\delta = \mathrm{Im}(d_{0,\delta}) \oplus \mathcal{W}_\delta$, where $\mathcal{W}_\delta$ is a particular choice of a complementary space that is induced by an inner product. This is explained in detail in the next section. Remark that we will sometimes drop the δ in the notation whenever no confusion is possible. Note also that the polynomials in the definition of $\mathcal{V}_\delta$ have degree $\delta + 1$, this is mainly because we look at $\frac{\partial}{\partial x_i}$ as an object of degree -1.

4.2.2 The choice of the complementary subspaces $\mathcal{W}_\delta$

In order to define suitable complementary spaces $\mathcal{W}_\delta$ we need the adjoint action of d_0 with respect to an inner product. Therefore we introduce the following

Definition 4.2 *We define an inner product on $\mathcal{P}_\delta$, the space of polynomials of degree δ as*

$$\left\langle \sum_{|\alpha|=\delta} a_\alpha x^\alpha, \sum_{|\beta|=\delta} b_\beta x^\beta \right\rangle = \sum_{|\alpha|=\delta} a_\alpha \bar{b}_\alpha \frac{\alpha!}{|\alpha|!}.$$

This induces an inner product on the space $\mathcal{V}_{\delta-1}$ of vector fields of degree $\delta - 1$ as follows:

$$\left\langle \sum_{k=1}^{n} V_k \frac{\partial}{\partial x_k}, \sum_{k=1}^{n} W_k \frac{\partial}{\partial x_k} \right\rangle = \sum_{k=1}^{n} \langle V_k, W_k \rangle_\delta, \tag{4.4}$$

Where the V_k, W_k are elements of $\mathcal{P}_\delta$.

Now we define d_0^* as the adjoint action of d_0 with respect to the above inner product. We repeat that d_0^* is defined as the unique linear map satisfying

$\langle d_0^*(V), W\rangle = \langle V, d_0(W)\rangle$, for all V, $W \in \mathcal{V}_\delta$. We define the operators $\Box_\delta = d_0 d_0^*$. In a similar way we define N^* to be the adjoint of a given linear mapping N defined on the inner product space $\mathcal{V}_\delta$ or on the inner product space $\mathcal{P}_\delta$. From linear algebra we know that:

1. These operators are self-adjoint.
2. These operators are diagonizable.
3. The operators have real positive eigenvalues.
4. $\mathcal{V}_\delta = \mathrm{Ker}(\Box_\delta) \oplus \mathrm{Im}(\Box_\delta) = \mathrm{Ker}(d_0^*) \oplus \mathrm{Im}(d_0)$

We will from now on choose complementary subspace W_δ as $\mathrm{Ker}(d_0^*) = \mathrm{Ker}(\Box_\delta)$.

We recall from [37] a nice way to calculate the adjoint operator d_0^*. Let us first define the isomorphism:

$$\phi : \mathcal{V}_\delta \longrightarrow \mathcal{P}_{\delta+1}^n : \sum_{k=1}^n V_k \frac{\partial}{\partial x_k} \mapsto (V_1, V_2, \ldots, V_n)$$

Lemma 4.3 ([37], p.691) *Suppose that* $V = \sum_{k=1}^n V_k \frac{\partial}{\partial x_k} \in \mathcal{V}_\delta$. *Then we have*

$$\phi(d_0^*(V)) = \begin{pmatrix} N^* - \left(\frac{\partial N_1}{\partial x_1}\right)^* & -\left(\frac{\partial N_2}{\partial x_1}\right)^* & \cdots & -\left(\frac{\partial N_n}{\partial x_1}\right)^* \\ -\left(\frac{\partial N_1}{\partial x_2}\right)^* & N^* - \left(\frac{\partial N_2}{\partial x_2}\right)^* & \cdots & -\left(\frac{\partial N_n}{\partial x_2}\right)^* \\ \vdots & & & \vdots \\ -\left(\frac{\partial N_1}{\partial x_n}\right)^* & \cdots & -\left(\frac{\partial N_{n-1}}{\partial x_n}\right)^* & N^* - \left(\frac{\partial N_n}{\partial x_n}\right)^* \end{pmatrix} \begin{pmatrix} V_1 \\ \vdots \\ \vdots \\ V_n \end{pmatrix}.$$

We give an example in the case that $n = 2$ and $N = x_2 \frac{\partial}{\partial x_1}$. In this case we have

$$\phi(d_0^*(V_1 \frac{\partial}{\partial x_1} + V_2 \frac{\partial}{\partial x_2})) = \begin{pmatrix} N^* & 0 \\ -1 & N^* \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}.$$

Hence $d_0^*(V_1 \frac{\partial}{\partial x_1} + V_2 \frac{\partial}{\partial x_2}) = N^* V_1 \frac{\partial}{\partial x_1} + (-V_1 + N^* V_2) \frac{\partial}{\partial x_2}$.

4.2.3 Resonant terms and small denominators

When solving equation (4.1), decompose $RHS_\delta = Q_\delta \oplus T_\delta$, where $Q_\delta \in \mathrm{Ker}(\Box_\delta)$ and $T_\delta \in \mathrm{Im}(\Box_\delta) = \mathrm{Im}(d_0)$. Now let Λ_δ be the list of eigenvalues (counted with multiplicity) of the operator $\Box_\delta$ and Λ_δ^* be the list of nonzero eigenvalues. Since $\Box_\delta$ is diagonizable, it is possible to decompose T_δ in a base of eigenvectors of $\Box_\delta$. More precisely:

$$T_\delta = \Box_\delta(V_\delta) = \sum_{\lambda \in \Lambda_\delta^*} \Box_\delta(V_{\delta,\lambda}) = \sum_{\lambda \in \Lambda_\delta^*} \lambda V_{\delta,\lambda}.$$

If we define $W_{\delta,\lambda} = d_0^*(V_{\delta,\lambda})$ and $W_\delta = \sum_{\lambda\in\Lambda_\delta^*} W_{\delta,\lambda}$, then

$$d_0(W_\delta) = d_0(\sum_{\lambda\in\Lambda_\delta^*} W_{\delta,\lambda}) = \sum_{\lambda\in\Lambda_\delta^*} \Box_\delta(V_{\delta,\lambda}) = \sum_{\lambda\in\Lambda_\delta^*} \lambda V_{\delta,\lambda} = T_\delta.$$

Moreover since

$$\langle W_{\delta,\lambda}, W_{\delta,\lambda}\rangle = \langle d_0^*(V_{\delta,\lambda}), d_0^*(V_{\delta,\lambda})\rangle = \langle V_{\delta,\lambda}, \Box_\delta(V_{\delta,\lambda})\rangle = \lambda \langle V_{\delta,\lambda}, V_{\delta,\lambda}\rangle\,,$$

it follows that we have the estimates:

$$||T_\delta||^2 = ||d_0(W_\delta)||^2 = \sum_{\lambda\in\Lambda_\delta^*} \lambda^2 ||V_{\delta,\lambda}||^2 = \sum_{\lambda\in\Lambda_\delta^*} \lambda ||W_{\delta,\lambda}||^2 \geq (\min_{\lambda\in\Lambda_\delta^*}(\sqrt{\lambda})||W_\delta||)^2.$$

This estimate makes clear that the λ's will play the role of the small denominators. We explain now what we mean by 'S satisfies a diophantine condition'. Therefore, following [37], p .675, we introduce the numbers $a_\delta = \min_{\lambda\in\Lambda_\delta^*}(\sqrt{\lambda})$ and define the numbers η_δ, for $\delta \geq 0$, recursively by (let $\eta_0 = 1$)

$$a_\delta \eta_\delta = \max_{\delta_1+\ldots+\delta_r=\delta} \eta_{\delta_1}\ldots\eta_{\delta_r},$$

where the maximum is taken over the set where at least two of the δ_i's are strictly positive. We say that S satisfies a diophantine condition if the $\eta_\delta \leq cM^\delta$ for certain positive constants c, M.

We will say that 'S satisfies a Siegel condition of order τ' whenever we have the estimates:

$$\frac{1}{\delta^\tau} \leq Ca_\delta$$

for a certain $C \geq 1$ and $\tau \geq 0$. These conditions are important because of the following two theorems, proven in [37].

Theorem 4.4 ([37], Theorem 5.6 p.676+Remark 6.7 on p.686) *Suppose that $X = N + R$ is a formal Gevrey-α vector field that is formally linearizable to its linear part N. Suppose that N satisfies a diophantine condition, then X is Gevrey-α linearizable.*

Theorem 4.5 ([37], Theorem 6.2 p.683+Remark 6.7 on p.686) *Suppose that $X = N + R$ is a Gevrey-α vector field that has a formal normal form $X' = N + R'$ by means of the procedure explained in this section, and suppose that the linear part N of X satisfies a Siegel condition of order τ, then X' and Φ are formal power series of type Gevrey-$(1 + \tau + \alpha)$.*

If we want to show the main theorem, we want $\tau = 0$, i.e. the non-zero eigenvalues of $\Box_\delta$ do not accumulate to 0 as $\delta \longrightarrow \infty$. This is the topic of Section 4.3.

4.3 Representations of $\mathrm{sl}(2,\mathbb{C})$

We briefly recall the definition of a Lie algebra, a representation of a Lie algebra and some related algebraic concepts.

Definition 4.6 *A Lie algebra* $(\mathfrak{g},[\,,])$ *is a vector space* $\mathfrak{g}$ *provided with a multiplication* $[\,,]:\mathfrak{g}\times\mathfrak{g}\mapsto\mathfrak{g}:(x,y)\mapsto[x,y]$ *that satisfies the relations*

- $[g_1,g_2]=-[g_2,g_1]$,
- $[g_1,[g_2,g_3]]+[g_2,[g_3,g_1]]+[g_3,[g_1,g_2]]=0$.

We list the following concepts:

(a) A Lie algebra $\mathfrak{g}$ *is called simple iff* $[\mathfrak{g},\mathfrak{g}]=\mathfrak{g}$.

(b) A lie algebra homomorphism is a linear map $L:\mathfrak{g}\longrightarrow\mathfrak{h}$ *between two Lie algebra's preserving the product structure :* $L([g_1,g_2])=[L(g_1),L(g_2)]$.

(c) $gl(V)$ *is a Lie algebra when considering the product* $[A,B]=AB-BA$.

(d) A Lie algebra representation of $\mathfrak{g}$ *is a Lie algebra homomorphism* $L:\mathfrak{g}\longrightarrow gl(V)$, *where* V *is a vector space and* $gl(V)$ *is the group of linear transformations from* V *to* V.

(e) A Lie algebra representation $L:\mathfrak{g}\longrightarrow gl(V)$ *is irreducible, iff there exist no subspace* W *different from* V *or* $\{0\}$ *such that* $L(g)(w)\in W$, *for every* $w\in W$ *and every* $g\in\mathfrak{g}$. *A subspace* W *with this property defines a subrepresentation.*

We will need one of the key results of representations of simple Lie algebra's. A proof can be found e.g. in [29].

Theorem 4.7 *Every finite dimensional representation of a simple Lie algebra* $\mathfrak{g}$ *can be written as a direct sum of irreducible representations of* $\mathfrak{g}$.

We now recall some basic facts of the representations of the simple Lie algebra $\mathrm{sl}(2,\mathbb{C})$. Let us first recall the definition.

Definition 4.8 *We define the Lie algebra* $\mathrm{sl}(2,\mathbb{C})$ *as the subalgebra of* $gl_2(\mathbb{C})$ *of matrices with trace* 0. *It is generated by the matrices*

$$N=\begin{pmatrix}0&1\\0&0\end{pmatrix},M=\begin{pmatrix}0&0\\1&0\end{pmatrix},H=\begin{pmatrix}1&0\\0&-1\end{pmatrix}.$$

Remark 4.9 *Any Lie algebra generated by three elements* N,M,H *and subject to the relations*

- $[H,N]=2N$,
- $[H,M]=-2M$,

- $[N, M] = H$.

is isomorphic to $\mathrm{sl}(2, \mathbb{C})$. *Moreover it is now clear that* $\mathrm{sl}(2, \mathbb{C})$ *is a simple Lie algebra.*

The following theorem is well-known: a proof can be found e.g. in [29].

Theorem 4.10 *For every n the representation of* $\mathrm{sl}(2, \mathbb{C})$ *defined by*

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \mapsto \widetilde{N}_n = \begin{pmatrix} 0 & n & 0 & 0 & \dots & 0 \\ 0 & 0 & n-1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 2 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \mapsto \widetilde{M}_n = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 2 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & n-1 & 0 & 0 \\ 0 & 0 & \dots & 0 & n & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \mapsto \widetilde{H}_n = \begin{pmatrix} n & 0 & 0 & 0 & \dots & 0 \\ 0 & n-2 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & -(n-2) & 0 \\ 0 & 0 & \dots & 0 & 0 & -n \end{pmatrix}$$

and acting on $\mathbb{C}^n$ is irreducible. Moreover, any other irreducible representation of $\mathrm{sl}(2, \mathbb{C})$ *is isomorphic to one of these representations.*

4.4 Construction of some particular $\mathrm{sl}(2, \mathbb{C})$ representations

In this section we focus on the construction of some particular $\mathrm{sl}(2, \mathbb{C})$ representations. In order to make the computations a bit more transparent, we use the correspondence between matrices and linear vector fields by a bijection $\phi : \sum_i \sum_j a_{ij} x_j \frac{\partial}{\partial x_i} \mapsto A = (a_{ij})$. Now suppose that we have two vector fields $A_v = \sum_{j=1}^n \sum_{i=1}^n a_{ij} x_i \frac{\partial}{\partial x_j}$ and $B_v = \sum_{j=1}^n \sum_{i=1}^n b_{ij} x_i \frac{\partial}{\partial x_j}$ with corresponding matrices A and B, then the Lie bracket transforms as $\phi([A_v, B_v]) = AB - BA$.

We start now with the construction. Therefore we define the following vector fields:

$$\begin{aligned}
N_n &:= \alpha_1 x_2 \frac{\partial}{\partial x_1} + \alpha_2 x_3 \frac{\partial}{\partial x_2} + \ldots + \alpha_n x_{n+1} \frac{\partial}{\partial x_n} \\
M_n &= \alpha_1 x_1 \frac{\partial}{\partial x_2} + \alpha_2 x_2 \frac{\partial}{\partial x_3} + \ldots + \alpha_n x_n \frac{\partial}{\partial x_{n+1}} \\
H_n &:= [N_n, M_n].
\end{aligned}$$

It is important to note that M_n is the adjoint of N_n with respect to the inner product (4.4). We will use the same notation N_n, M_n and H_n for the associated matrices and drop the index n where no confusion is possible. We want to choose the coefficients $\alpha_1, \ldots, \alpha_n$ in such a way that they are non-zero and that the triple N, M, H is isomorphic to the Lie algebra $\mathrm{sl}(2, \mathbb{C})$. Therefore it is sufficient to ensure that the relations described in remark 4.9 are satisfied. The third relation is automatic from the construction. We focus on the first relation. In matrix notation this relation becomes $HN - NH - 2N = 0$. Now remark that

$$N = \begin{pmatrix} 0 & \alpha_1 & 0 & \ldots & 0 \\ 0 & 0 & \alpha_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & \alpha_n \\ 0 & 0 & \ldots & 0 & 0 \end{pmatrix}, M = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ \alpha_1 & 0 & 0 & \ldots & 0 \\ 0 & \alpha_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \alpha_n & 0 \end{pmatrix},$$

$$H = \begin{pmatrix} \alpha_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \alpha_2^2 - \alpha_1^2 & 0 & \ldots & 0 \\ 0 & 0 & \alpha_3^2 - \alpha_2^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \alpha_n^2 - \alpha_{n-1}^2 & 0 \\ 0 & 0 & \ldots & 0 & -\alpha_n^2 \end{pmatrix}.$$

Hence this relation becomes

$$\begin{pmatrix} 0 & b_1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & b_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & b_n \\ 0 & 0 & \ldots & 0 & 0 & 0 \end{pmatrix} = 0,$$

where $b_i = \alpha_i(-\alpha_{i-1}^2 + 2\alpha_i^2 - \alpha_{i+1}^2) - 2\alpha_i$ for $2 \le i \le n-1$, $b_1 = \alpha_1(2\alpha_1^2 - \alpha_2^2) - 2\alpha_1$ and $b_n = \alpha_n(-\alpha_{n-1}^2 + 2\alpha_n^2) - 2\alpha_n$; and we need to solve the equations

$$\begin{cases} \alpha_1(2\alpha_1^2 - \alpha_2^2) = 2\alpha_1 \\ \alpha_2(-\alpha_1^2 + 2\alpha_2^2 - \alpha_3^2) = 2\alpha_2 \\ \vdots \\ \alpha_n(-\alpha_{n-1}^2 + 2\alpha_n^2) = 2\alpha_n. \end{cases}$$

Since we suppose that none of the α_i vanishes, this simplifies to

$$\begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} \alpha_1^2 \\ \alpha_2^2 \\ \vdots \\ \alpha_{n-1}^2 \\ \alpha_n^2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ \vdots \\ 2 \end{pmatrix}.$$

One can verify that a solution is given by $\alpha_i^2 = i(n+1-i)$ for $1 \le i \le n$. We choose the positive solutions and put $\alpha_i = \sqrt{i(n+1-i)}$. Then it is readily verified (repeat the above calculations) that also the second relation $[H, M] = -2M$ from remark 4.9 is satisfied. We have now proven the

Lemma 4.11 *Let $n \in \mathbb{N}$ and define $N_n = \sum_{i=1}^{n} \sqrt{i(n+1-i)} x_{i+1} \frac{\partial}{\partial x_i}$, then the triple N_n, $M_n := N_n^*$, $H = [N_n, M_n]$ defines a Lie-algebra isomorphic to* $\mathrm{sl}(2, \mathbb{C})$.

We are now in a position to show that

Lemma 4.12 *Let $\delta \in \mathbb{N} \setminus \{0\}$. For a given $N_n = \sum_{i=1}^{n} \sqrt{i(n+1-i)} x_{i+1} \frac{\partial}{\partial x_i}$ the associated triple d_0, d_0^*, $D = [d_0, d_0^*]$ defines an $\mathrm{sl}(2, \mathbb{C})$ representation. Here, d_0 is the Lie operator $U \mapsto [N_n, U]$ acting on $\mathcal{V}_\delta$ and d_0^* its adjoint.*

Proof: Put $\alpha_i = \sqrt{i(n+1-i)}$ and I the identity operator. The Lie operator acting on vector fields $d_0(\sum_{i=1}^{n+1} V_i \frac{\partial}{\partial x_i})$ can be expressed using matrix notation as

$$\begin{pmatrix} N & -\alpha_1 I & 0 & 0\dots & 0 & \\ 0 & N & -\alpha_2 I & 0\dots & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & N & -\alpha_n I \\ 0 & \dots & 0 & 0 & 0 & N \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \\ V_{n+1} \end{pmatrix};$$

and its adjoint $d_0^*(\sum_{i=1}^{n+1} V_i \frac{\partial}{\partial x_i})$ as (see also Lemma 4.3 and remember that $M = N^*$)

$$\begin{pmatrix} M & 0 & 0 & 0\dots & 0 & \\ -\alpha_1 I & M & 0 & 0\dots & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & -\alpha_{n-1} I & M & 0 \\ 0 & \dots & 0 & 0 & -\alpha_n I & M \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \\ V_{n+1} \end{pmatrix}.$$

Hence it is readily verified that the commutator $D = [d_0, d_0^*] = d_0 d_0^* - d_0^* d_0$ can be expressed as

$$\begin{pmatrix} H+\alpha_1^2 I & 0 & 0 & \dots & 0 \\ 0 & H+(\alpha_2^2-\alpha_1^2)I & 0 & \dots & 0 \\ 0 & 0 & H+(\alpha_3^2-\alpha_2^2)I & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & H+(\alpha_n^2-\alpha_{n-1}^2)I & 0 \\ 0 & 0 & \dots & 0 & H-\alpha_n^2 I \end{pmatrix},$$

where $H=[N,M]$. Now the commutator $[D,d_0]$ simplifies as

$$\begin{pmatrix} a_1 & b_1 & 0 & 0\dots & 0 & \\ 0 & a_2 & b_2 & 0 & 0\dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0\dots & a_n & b_n \\ 0 & 0 & 0 & 0\dots & 0 & a_{n+1} \end{pmatrix},$$

with

$$\begin{aligned} a_i &= HN-NH+(\alpha_i^2-\alpha_{i-1}^2)(NI-IN)=[H,N]=2N, \\ b_i &= -\alpha_i(H+\alpha_i^2 I-\alpha_{i-1}^2 I)+\alpha_i(H+\alpha_{i+1}^2 I-\alpha_i^2 I)= \\ &= -\alpha_i(-\alpha_{i+1}^2+2\alpha_i^2-\alpha_{i-1}^2)I \\ &= -2\alpha_i I; \end{aligned}$$

where we have put $\alpha_0=0$ and $\alpha_{n+1}=0$ in the above calculation. We also used the fact that the triple N, M, H defines a Lie algebra isomorphic to $\mathrm{sl}(2,\mathbb{C})$. Hence $[D,d_0]=2d_0$.
Making analogous calculations, one verifies that also $[D,d_0^*]=-2d_0^*$. □

As a corollary of this lemma we can consider the case of multiple nilpotent blocks as follows. Remark that we allow zero blocks (i.e. $k_i=0$ for some i).

Proposition 4.13 *Let k_1, k_2, …, k_n be natural numbers and let x^i be a k_i+1-dimensional variable $(x_1^i,\dots,x_{k_i+1}^i)$, for $1\le i\le n$. Let $N=N_{k_1}(x^1)+\dots+N_{k_n}(x^n)$, where*

$$N_{k_j}(x^j)=\sum_{i=1}^{k_j}\sqrt{i(n-i+1)}x_{i+1}^j\frac{\partial}{\partial x_i^j},\ N_0=0.$$

Then the triple N, $M:=N^*$, $H=[N,M]$ defines a Lie algebra isomorphic to $\mathrm{sl}(2,\mathbb{C})$. *Moreover let d_0 be the associated Lie operator, then also the triple d_0, d_0^*, $D=[d_0,d_0^*]$ defines a Lie algebra isomorphic to* $\mathrm{sl}(2,\mathbb{C})$.

Proof: Use the concept of a direct sum, Lemma 4.11 and Lemma 4.12. □

4.5 Proof of Theorem 4.1

This is rather a summary of all the foregoing. From linear algebra we know that, up to a linear change of variables, it is no restriction to start with a vector field $X = N + R$ where N is as in Proposition 4.13. Let now d_0 be the associated Lie operator. Let $\delta \in \mathbb{N} \setminus \{0\}$. We are interested in the calculation of eigenvalues of the associated operator $\Box_\delta = d_0 d_0^*$ acting on $\mathcal{V}_\delta$. According to Proposition 4.13, we know that the triple d_0, d_0^* and $D = [d_0, d_0^*]$ defines a Lie algebra isomorphic to $\mathrm{sl}(2, \mathbb{C})$. It follows, using Theorem 4.7, that the associated representation acting on $\mathcal{V}_\delta$ can be split in a direct sum of irreducible representations. Hence, up to a linear coordinate transform φ (acting on the space $\mathcal{V}_\delta$), we can suppose that we are dealing with a representation of the form

$$\begin{aligned} N &= \widetilde{N}_1 \oplus \widetilde{N}_2 \oplus \ldots \oplus \widetilde{N}_l, \\ M &= \widetilde{M}_1 \oplus \widetilde{M}_2 \oplus \ldots \oplus \widetilde{M}_l, \\ H &= \widetilde{H}_1 \oplus \widetilde{H}_2 \oplus \ldots \oplus \widetilde{H}_l; \end{aligned}$$

where $\widetilde{N}_i$, $\widetilde{M}_i$ and $\widetilde{H}_i$ are as in Theorem 4.10. Hence φ transforms the operator $\Box_\delta = d_0 d_0^*$ into NM. The nonzero eigenvalues of the operator $NM = \widetilde{N}_1 \widetilde{M}_1 \oplus \ldots \oplus \widetilde{N}_l \widetilde{M}_l$ are positive integers because each $\widetilde{N}_i \widetilde{M}_i$ is a diagonal matrix containing integers on the diagonal. Hence the same is true for the operator $\Box_\delta$. Now using Theorem 4.5 with $\tau = 0$ (or Theorem 4.4 in case the vector field is formally linearizable) finishes the proof.

Chapter 5

Perturbations of diffeomorphisms of quasi-homogeneous degree 0 and Gevrey normalization of diffeomorphisms with diagonal linear part

5.1 A quasi-homogeneous framework for analytic diffeomorphisms and linearization

5.1.1 Introduction and motivation

In this chapter we explore a generalization of the following problem : let V be a neighbourhood of the origin in $\mathbb{C}^n$ and $F : V \subset \mathbb{C}^n \longrightarrow \mathbb{C}^n$ be a local analytic diffeomorphism fixing the origin. Hence $F = A + f$, where A is its linear part; and $f(x)$ is of the type $O(||x||^2)$. We look for a local analytic near-identity transformation $U = \mathrm{id} + u$, where $u(x)$ is of the type $O(||x^2||)$, in such a way that the conjugation is easier on a neighbourhood $\hat{V} \subset V$ of the origin. More precisely we want to obtain

$$U^{-1} \circ F \circ U = A + g, \tag{5.1}$$

where $g(x)$ is analytic and $O(||x||^2)$. The intention is to make g as simple as possible, the ideal case being $g = 0$. This is however not possible in general due to the presence of resonances or quasi-resonances, see e.g. Chapter 1, Chapter 2 and below.

Let λ_i, $1 \leq i \leq n$, be the eigenvalues of the linear part A of F. We shall say in this chapter that A is non-resonant if for all $k \in \mathbb{N}^n \setminus \{0\}$ we have that $\lambda^k \neq 1$. It is known that the existence of resonances are in general a formal obstruction to solve problem (5.1) with $g = 0$.

Some classical cases where (5.1) is known to have a solution for polynomial g are:

i) If all eigenvalues λ_i of A satisfy $|\lambda_i| < 1$. In this case the normal form $A + g$ can be chosen a polynomial containing only resonant terms. Remark that from the dynamical point of view the diffeomorphism is a hyperbolic contraction.

ii) When all eigenvalues λ_i of A satisfy $|\lambda_i| > 1$. In this case the normal form $A+g$ can again be chosen a polynomial containing only resonant terms. Remark that from the dynamical point of view the diffeomorphism is a hyperbolic repulsion.

iii) When F contains no resonances and $g = 0$ is chosen, A is semi-simple, and the Brjuno condition

$$-\sum_{k\geq 1} \frac{\ln \omega_k}{2^k} < +\infty, \text{ where } \omega_k := \inf\{| \langle \lambda, k \rangle - \lambda_j | \,|\, k \in \mathbb{N}^n, \, 1 \leq j \leq n\}, \tag{5.2}$$

is satisfied.

We will reprove all these results later on by considering a more general setting. We remark that, when the diffeomorphism $A + f$ is not a hyperbolic contraction nor a hyperbolic repulsion, the presence of resonances can (and usually will) destroy the analytic character of the normal form $A + g$ and of the transformation id + u. However, if the eigenvalues satisfy a so called Siegel condition (see later), the divergence generated by the presence of such terms is not too bad: the Taylor series expansion of the non-linear normal form $A+g$ as well as the transformation U will have at most a Gevrey-growth rate that is determined by the type of the Siegel condition.

In the general setting we consider we have a slightly different point of view in mind. We will not necessarily linearize the system as is commonly done. Instead we start with a diffeomorphism $F = S + f$, but this time we use a quasi-homogeneous grading. We want to reduce the diffeomorphism F to its non-perturbed part S that has a quasi-homogeneous degree 0 if possible. The perturbation terms f are of a higher quasi-homogeneous degree then S. Remark that we will use the notation S when working with a quasi-homogeneous non-perturbed part and the notation A when working with a linear non-perturbed part. The idea of this approach is based on the same approach for vector fields in [37], therefore we need to give a lot of credit to the authors of this work, because we can recycle a lot of their lemmas there. We will therefore give a reference in the heading of the corresponding lemmas and proofs to their corresponding lemma each time we recycle a proof.

In this context it is also worth mentioning the results of [19, 26]. They generalize a counterexample of [64] stating ‘if there exists eigenvalues of A that satisfy $|\lambda_\alpha| > 1$ and there exists eigenvalues $|\lambda_\beta| < 1$, then the non-diagonalizability of A is an obstruction for analytic linearization’, see also Chapter 6 of this thesis. This remark is important with respect to the first part of this text, since one could wonder whether or not there exist non-trivial examples for the diophantine conditions we demand. We will comment on these results later on in Chapter 6.

We proceed as follows: to start we introduce an adapted grading, and use the concept of quasi-homogeneous polynomials. We also prove some lemmas and propositions that will mainly carry the estimates that will be needed to build the general theory in three main sections :

- **The formal section**: We build the formal normal form by considering a special choice of the complementary subspaces of the kernel of the Lie operator that we construct with the help of the inner product structure that is present on the space of quasi-homogeneous polynomials.

- **Analytic section**: We show that, under a certain type of diophantine condition comparable to the Brjuno condition (5.2), the generalization of the problem (5.1) with $g = 0$ or g contained in a specific chosen ideal has an analytic solution. We will then proceed to show that in case A is not semi-simple, this condition is never satisfied. We use the same techniques as [64, 26, 19]. This is an important remark, since one could wonder whether problem (5.1) can be solved for more general S that are either not linear or not semi-simple.

- **Gevrey section**: In this section we will deliver only partial results. We will restrict to the case where $S = A$ is linear and diagonal. We show that if the eigenvalues of the box-operator satisfy a Siegel bound, then the formal normal form $S + g$ as well as the formal normal form transformation U has a Taylor series with coefficients that have at most a Gevrey growth. Remark that if $S = A$ is linear and semi-simple the Siegel condition is almost always satisfied (in Lebesgue sense). We do not know whether or not the same holds for non-linear or non-semi-simple S.

We will then continue with a numerical study for some examples, where we mainly focus on the numerical calculation of the growth of the eigenvalues of the box-operators, that according to the theorems proven at this point control the Gevrey behaviour of the normal form and the normal form transformation. We have mainly two cases in mind: the non-semi-simple saddle case with eigenvalues satisfying a Siegel condition and example 5.4. The numerical calculation of the eigenvalues is rather delicate, not only because the matrices in casu have a high number of entries that need to be calculated separately, but especially because they have some eigenvalues that tend to zero rather fast as the order increases.

5.1.2 Quasi-homogeneous polynomials

Let $p = (p_1, \ldots, p_n)$ be an n-tuple of non-zero natural numbers with greatest common divisor 1. We introduce the following notations:

$$\begin{aligned} \langle p, \alpha \rangle &:= \sum_{i=1}^{n} p_i \alpha_i, \\ x^\alpha &:= x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}, \text{ where } \alpha \in \mathbb{N}^n,\ x = (x_1, \ldots, x_n) \in \mathbb{C}^n. \end{aligned}$$

Then, for each $\delta \in \mathbb{N}$ we call a polynomial that can be expressed as $\sum\limits_{\langle p,\alpha\rangle=\delta} a_\alpha x^\alpha$ a polynomial of p-quasi-degree δ. Furthermore we define for each $\delta \in \mathbb{N}$ the vector space $\mathcal{P}_{p,\delta}$ consisting of polynomials of p-quasi-degree δ. Remark that this space can be trivial if there exist no n-tuple α of natural numbers for which $\langle p, \alpha \rangle = \delta$ (for example if $p = (2, 7)$, then the equation $2\alpha_1 + 7\alpha_2 = 5$ has no solutions); but there exists a $\delta_0 \in \mathbb{N}$ depending on p such that if $\delta > \delta_0$ we may assume that $\mathcal{P}_{p,\delta}$ is not empty (See [37], Lemma 3.3 p.665). We introduce a norm on $\mathcal{P}_{p,\delta}$. Therefore, let $f = \sum_{\langle p,\alpha\rangle=\delta} f_\alpha x^\alpha$, $g = \sum_{\langle p,\alpha\rangle=\delta} g_\alpha x^\alpha \in \mathcal{P}_{p,\delta}$ and define their inner product

$$\langle f, g \rangle_{p,\delta} := \sum_{\langle p,\alpha\rangle=\delta} f_\alpha \overline{g_\alpha} \frac{(\alpha_1!)^{p_1} \ldots (\alpha_n!)^{p_n}}{\langle p, \alpha \rangle!} = \sum_{\langle p,\alpha\rangle=\delta} f_\alpha \overline{g_\alpha} \frac{(\alpha!)^p}{\langle p, \alpha \rangle!} \tag{5.3}$$

and associated norm $|f|_{p,\delta} := \sqrt{\langle f, f \rangle_{p,\delta}}$. Remark that there are multiple choices possible for suitable norms on this space, we suggest reading appendix A.2 in [37] for more information on the choice of the norm and more examples of norms. For technical reasons we will sometimes need to split a polynomial $f_\delta \in \mathcal{P}_{p,\delta}$ into pieces of usual degree. Hence we have that $f_\delta = \sum_{r \geq 0} f_{\delta,r}$, where $f_{\delta,r}$ is a polynomial of quasi-degree δ and regular degree r. It is well known that to each polynomial of degree r one can associate a unique r-linear symmetric form. We will denote this associated r-linear form with a tilde e.g. $\widetilde{f_{\delta,r}}$. Since we are mainly interested in local analytic functions f, it is interesting to decompose these functions with respect to new quasi-grading. More explicitly we have at the formal level that $f(x) = \sum\limits_{\delta \geq 0} f_\delta(x) = \sum\limits_{\delta \geq 0} \big(\sum\limits_{\langle p,\alpha\rangle=\delta} a_\alpha x^\alpha \big)$. In the classical situation where $p = (1, 1, \ldots, 1)$ such a formal power series is local analytic if $|f_{p,\delta}| \leq \frac{C}{R^\delta}$ for certain positive constants C and R. We have an analogous characterization for local analyticity with respect to a general quasi-homogeneous degree p:

Proposition 5.1 ([37], Proposition A.8, p. 701–702) *Suppose that* $f = \sum\limits_{\delta \geq 0} f_\delta$ *is a formal power series; then the following properties are equivalent:*

1. *The power series converges uniformly in a neighbourhood of the origin.*

2. *There exist positive constants C, R such that $|f_\delta| \leq \frac{C}{R^\delta}$ for every $\delta \in \mathbb{N}$.*

3. *There exist positive constants C, R such that $|f_{\delta,r}| \leq \frac{C}{R^r}$ for every $\delta, r \in \mathbb{N}$.*

We define the set $\mathcal{O}_{p,\delta}$ of vector-valued p-quasi-homogeneous polynomials of degree δ. This is an n-tuple of polynomials $P(x) = (P_1(x), P_2(x), \dots, P_n(x))$ where each $P_i(x)$ is a polynomial of quasi-homogeneous degree $p_i + \delta$. The reason for this shift in the degree is to be sure that the corresponding Lie operator that we will encounter in the following sections

$$d_{0,p,\delta} : \mathcal{O}_{p,\delta} \to \mathcal{O}_{p,\delta} : U \mapsto DS.U - U \circ S, \tag{5.4}$$

is a well-defined linear operator whenever $S \in \mathcal{O}_{p,0}$. Indeed, the i-th component of $d_0(U)$ is given by $\sum_{j=1}^n \frac{\partial S_i(x)}{\partial x_j} U_j(x) - U_i(S(x))$. Here $\frac{\partial S_i}{\partial x_j}$ is a polynomial of quasi-homogeneous degree $p_i - p_j$ and U_j of degree $\delta + p_j$ such that the sum $\sum_{j=1}^n \frac{\partial S_i(x)}{\partial x_j} U_j(x)$ has the grading $\delta + p_i$. The substitution $U \circ S$ is also well-defined, because we substitute the polynomial $S_i(x)$ of degree p_i for the polynomial x_i of the same degree. Remark that the inner product (5.3) induces an inner product on $\mathcal{O}_{p,\delta}$. Indeed, let $U, V \in \mathcal{O}_{p,\delta}$ their inner product is given by

$$\langle U, V \rangle_{p,\delta} := \sum_{i=1}^n \langle U_i, V_i \rangle_{p,\delta+p_i} \tag{5.5}$$

For later use we also introduce the polynomial vector fields of p-quasi-homogeneous degree δ. This space is defined as

$$\mathcal{V}_{p,\delta} = \{ V_1 \frac{\partial}{\partial x_1} + \ldots + V_n \frac{\partial}{\partial x_n} | \, (V_1, V_2, \dots, V_n) \in \mathcal{O}_{p,\delta} \}$$

and inherits the inner product of $\mathcal{O}_{p,\delta}$ through the natural bijection between $\mathcal{O}_{p,\delta}$ and $\mathcal{V}_{p,\delta}$ and we will use the same notation for this inner product. Remark that, with respect to vector fields, $\frac{\partial}{\partial x_i}$ has weight $-p_i$.

5.1.3 Analytic quasi-homogeneous normal forms

Formal framework

Suppose that $(p_1, p_2, \dots, p_n)$ is an n-tuple of natural numbers with greatest common divisor equal to 1. Let $F = S + f$, where $S \in \mathcal{O}_{p,0}$ and f can be decomposed as

$$f = \sum_{\delta \geq 1} f_\delta, \; f_\delta \in \mathcal{O}_{p,\delta}.$$

We will refer to such an f as 'the terms of higher order' without further notice. Remark that we do not require S nor F to be locally invertible i.e. F is not necessarily a diffeomorphism. We consider the following problem: try to find a formal (resp: analytic, Gevrey) diffeomorphism $U = \mathrm{id} + u$, where $u = \sum_{\delta \geq 1} u_\delta$ and $u_\delta \in \mathcal{O}_{p,\delta}$; such that the pullback of F by U

$$U^{-1} \circ F \circ U = S + g \tag{5.6}$$

becomes 'as simple as possible'. By this we mean that the Taylor series of g is reduced as much as possible, the optimal case being $g = 0$. It is generally well known that this is impossible due to the presence of resonances. We explain this now in more detail. First we rewrite the equation (5.6) and obtain

$$\begin{aligned} d_0(u) = & S \circ (\mathrm{id} + u) - S - DS.u + f \circ (\mathrm{id} + u) - g \\ & - u \circ (S + g) + u \circ S \end{aligned} \tag{5.7}$$

Let us first describe the formal procedure. We will define $u_\delta, g_\delta \in \mathcal{O}_{p,\delta}$ for each $\delta \geq 1$ by induction on δ. Suppose that we have already defined u_1, u_2, ..., $u_{\delta-1}$, we define u_δ. In order to do so, we consider the projection π_δ on the terms of degree δ of equation (5.7). On the left hand side we get $\pi_\delta(d_0(u)) = d_0(\pi_\delta(u)) = d_0(u_\delta)$. While on the right hand side we obtain

$$\begin{aligned} & \pi_\delta(S \circ (\mathrm{id} + u) - S - DS.u + f \circ (\mathrm{id} + u) - g - u \circ (S + g) + u \circ S) \\ & = f_\delta - g_\delta + EXPR_\delta, \end{aligned}$$

where the last term is a polynomial depending on the coefficients $u_\alpha, f_\alpha, g_\alpha$ with indices $1 \leq \alpha < \delta$. It is immediately clear that whenever we have that $d_{0,p,\delta}$ is surjective – and hence bijective – we may choose $g_\delta = 0$ and $u_\delta = d_{0,p,\delta}^{-1}(f_\delta + EXPR_\delta)$. However, if $d_{0,p,\delta}$ is not surjective, we need to choose a complementary subspace $\mathcal{C}_{p,\delta}$ to $\mathrm{Im}(d_{0,p,\delta})$ in $\mathcal{O}_{p,\delta}$. Remark that the choice of such a complementary subspace is not unique. In the next sections we will make a choice for $\mathcal{C}_{p,\delta}$ that is based on the inner product defined in (5.3). Since we then have a direct sum decomposition

$$\mathcal{O}_{p,\delta} = \mathrm{Im}(d_{0,p,\delta}) \oplus \mathcal{C}_{p,\delta},$$

with a corresponding projection operator $\pi_{\mathcal{C}_{p,\delta}} : \mathcal{O}_{p,\delta} \longrightarrow \mathcal{C}_{p,\delta}$, we can choose $g_\delta = \pi_{\mathcal{C}_{p,\delta}}(f_\delta + EXPR_\delta)$ and u_δ as the solution of $d_{0,p,\delta}(u_\delta) = f_\delta - g_\delta + EXPR_\delta$. To conclude we have proven the following theorem:

Theorem 5.2 *Let $F = S + f$ be a formal transformation, where S is a polynomial of p-quasi-homogeneous degree 0, and f contains terms of higher order, then there exists*

- *A polynomial transformation $U = id + \sum_{\delta=1}^{k} u_\delta$ such that the conjugation of U and F is simplified in the following sense:*

$$\pi_\delta(g) \in \mathcal{C}_{p,\delta}, \ \textit{for } 1 \leq \delta \leq k.$$

- *A formal transformation $U = id + \sum_{\delta=1}^{+\infty} u_\delta$ such that the conjugation of U and F consists is simplified in the following sense:*

$$\pi_\delta(g) \in \mathcal{C}_{p,\delta}, \text{ for } 1 \leq \delta \leq \infty,$$

we will refer to $S + g$ as the formal normal form (with respect to the chosen $\mathcal{C}_{p,\delta}$, that are usually clear from the context).

Remark 5.3 *It is in general not true that the formal transformation converges. We will see in the next section some sufficient conditions to obtain convergence of this formal transformation.*

Example 5.4 *A typical example that we have in mind is the following: $S(x, y) = (\alpha x + \beta y^2, \gamma y)$; $F(x, y) = S(x, y) + f(x, y)$, where $\alpha, \beta, \gamma \in \mathbb{C}$. The weights are chosen $p = (2, 1)$; and the perturbation $f(x, y)$ contains only non-zero terms that are of strict positive quasi-degree. Even for such a seemingly easy example it is tough to compute the eigenvalues of the box-operator. We perform a numerical study at the end of this chapter.*

Analytic results

Most of the notations, ideas and proofs in this section are borrowed from [37], where an analogous statement for vector fields is proven.

We use the same notations as in the previous subsection and obtain convergence when certain formal and number-theoretic conditions are satisfied. In order to do so we will make a specific choice of the complementary subspaces $\mathcal{C}_{p,\delta}$ that is based upon the inner products defined by equation (5.3). It allows us to define the adjoint operator $d^*_{0,p,\delta}$ of $d_{0,p,\delta}$ with respect to this inner product. Remember that this operator is completely determined by the relations

$$\left\langle d^*_{0,p,\delta}(U), V \right\rangle = \left\langle U, d_{0,p,\delta}(V) \right\rangle, \text{ for any } U, V \in \mathcal{O}_{p,\delta}. \tag{5.8}$$

We have the following lemma inspired from linear algebra:

Lemma 5.5 *The box-operator $\square_{p,\delta} := d_{0,p,\delta} d^*_{0,p,\delta}$ is a self-adjoint operator. It is diagonalizable and has only real, positive eigenvalues (zero eigenvalues are allowed). Moreover we have the decomposition*

$$\begin{aligned} \mathcal{O}_{p,\delta} &= \operatorname{Im}(d_{0,p,\delta}) \oplus \operatorname{Ker}(d^*_{0,p,\delta}) \\ &= \operatorname{Im}(\square_{p,\delta}) \oplus \operatorname{Ker}(\square_{p,\delta}). \end{aligned}$$

Proof: It follows from the relations (5.8) that the box-operator is self-adjoint. The fact that it is diagonalizable and has positive real eigenvalues follows from the self-adjointness property and

$$\left\langle \square_{p,\delta} v, v \right\rangle = \left\langle d_{0,p,\delta} d^*_{0,p,\delta} v, v \right\rangle = \left\langle d^*_{0,p,\delta} v, d^*_{0,p,\delta} v \right\rangle \geq 0, \text{ for all } v$$

The decomposition of $\mathcal{O}_{p,\delta} = \mathrm{Im}(\Box_{p,\delta}) \oplus \mathrm{Ker}(\Box_{p,\delta})$ is an immediate consequence of the diagonalizability of the box-operator. Indeed define Λ to be the set of eigenvalues of the box-operator with multiplicity and Λ_0 the set of non-zero eigenvalues with multiplicity. Choose now for each $\lambda \in \Lambda$ an eigenvector e_λ for the operator $\Box_{p,\delta}$. Then we can decompose each vector v as

$$v = \sum_{\lambda\in\Lambda} v_\lambda e_\lambda = \sum_{\lambda\in\Lambda_0} v_\lambda e_\lambda + \sum_{\lambda\in\Lambda\setminus\Lambda_0} v_\lambda e_\lambda .$$

It is clear that the second sum belongs to the kernel and the first has a zero projection onto the kernel. We finish the proof by observing that

- $\mathrm{Ker}(\Box_{p,\delta}) = \mathrm{Ker}(d^*_{0,p,\delta})$:
  Suppose that v is an element of $\mathrm{Ker}(\Box_{p,\delta})$ then $\Box_{p,\delta}(v) = d_{0,p,\delta}d^*_{0,p,\delta}(v) = 0$, from which it follows that $\langle \Box_{p,\delta}(v), v\rangle = 0$ and hence $\left\langle d^*_{0,p,\delta}(v), d^*_{0,p,\delta}(v)\right\rangle = 0$. Proving that $d^*_{0,p,\delta}(v) = 0$. The other inclusion is obvious.
- $\mathrm{Im}(d_{0,p,\delta}) = \mathrm{Im}(\Box_{p,\delta})$:
  Suppose that $u \in \mathrm{Im}(d_{0,p,\delta})$, then $u = d_{0,p,\delta}(w)$ for some w. Hence it follows that $\langle u, v\rangle = \langle d_{0,p,\delta}(w), v\rangle = \left\langle w, d^*_{0,p,\delta}(v)\right\rangle = 0$ for all $v \in \mathrm{Ker}(d^*_{0,p,\delta}) = \mathrm{Ker}(\Box_{p,\delta})$. It follows that u is orthogonal to $\mathrm{Ker}(\Box_{p,\delta})$ and hence belongs to $\mathrm{Im}(\Box_{p,\delta})$. The other inclusion is again obvious.

□

Suppose that $\mathcal{I}$ is the ideal of $\mathbb{C}\{x\}$ generated by the polynomials $h_1, h_2, \ldots, h_s$ of resp. p-quasi-homogeneous degree $\alpha_1, \alpha_2, \ldots, \alpha_s$. We define the multiplication operators $\mathcal{M}_i : \mathcal{O}_{p,\delta} \longrightarrow \mathcal{O}_{p,\delta+e_i} : f \mapsto h_i.f$ and the submodule

$$\mathcal{M} = \mathcal{M}_1\mathcal{O} + \ldots + \mathcal{M}_s\mathcal{O}.$$

We also define $\mathcal{M}_{p,\delta} = \mathcal{M} \cap \mathcal{O}_{p,\delta}$. It then follows that we have a decomposition

$$\mathcal{O}_{p,\delta} = \mathcal{M}_{p,\delta} \overset{\perp}{\oplus} \mathcal{V}_{p,\delta},$$

with corresponding projection $\pi_{\mathcal{I}^\perp}$ on $\mathcal{V}_{p,\delta}$. We define $\mathcal{V} = \oplus_{\delta=1}^{\infty}\mathcal{V}_{p,\delta}$ and

$$\mathcal{W} = \left\{u \in \mathcal{O} \,|\, u_\delta \in \mathrm{Im}(d^*_{0,p,\delta}) = \mathrm{Ker}(d_0)^\perp, \forall\delta \text{ and } [U, S] \in \mathcal{V}\right\}. \qquad (5.9)$$

It follows that $\mathcal{O} = \mathcal{M} \overset{\perp}{\oplus} \mathcal{V}$. In the same way we can construct the formal ideals $\widetilde{\mathcal{M}}$, $\widetilde{\mathcal{V}}$ and the corresponding projection $\pi_{\widetilde{\mathcal{I}}^\perp}$. As in [37] we have the following lemma:

Lemma 5.6 ([37], Lemma 5.1, p.675) *Using the above notation we obtain that* $\mathcal{V}_{p,\delta} = \cap_{i=1}^{s}\mathrm{Ker}(\mathcal{M}_i^*|_{\mathcal{O}_{p,\delta}})$*, where* $\mathcal{M}_i^*|_{\mathcal{O}_{p,\delta}}$ *is the adjoint of the mapping*

$$\mathcal{M}_i|_{\mathcal{O}_{p,\delta}} : \mathcal{O}_{p,\delta} \longrightarrow \mathcal{O}_{p,\delta+\alpha_i}.$$

We define now $\sigma_{\delta,\backslash\mathcal{I}}$ as the set of non-zero eigenvalues of $\Box_{p,\delta}$ for which there exists an associated eigenvector that is orthogonal to $\mathcal{I}_\delta$. We therefore introduce the following constants:

$$a_\delta = \min_{\lambda\in\sigma_{\delta,\backslash\mathcal{I}}} \sqrt{\lambda},$$

and define recursively the associated η_δ. We put $\eta_0 = 1$ and

$$a_\delta \eta_\delta = \overset{*}{\max_{\delta_1+\delta_2+\ldots+\delta_r=\delta}} \eta_{\delta_1}\eta_{\delta_2}\ldots\eta_{\delta_r}, \tag{5.10}$$

where we use a * above the maximum to indicate that it is taken over all possible $(\delta_1,\ldots,\delta_r)$ that have at least two non-zero entries. This leads to the following definition.

Definition 5.7 *We say that S is α-diophantine with respect to $\mathcal{I}$, if the formal power series $\sum_{\delta\geq 0}\frac{\eta_\delta}{\delta!^\alpha}x^\delta$ converges. We say that S is diophantine if it is 0-diophantine.*

We are now ready to formulate the following theorem that is the 'diffeo version' of Theorem 5.6, p.676 from [37].

Theorem 5.8 *Let $\mathcal{I}$ be the ideal in $\mathbb{C}\{x\}$ that has the properties as described above. Suppose that $S\in\mathcal{O}_{p,0}$ and $f\in\mathbb{C}\{x\}$ that contains only higher order terms. I.e. $f=\sum_{\delta\geq 1} f_\delta$, where $f_\delta\in\mathcal{O}_{p,\delta}$, and f is analytic. Assume that there exists a formal transformation $U\in\mathcal{W}$ that transforms $S+f$ to $S+g$ such that g belongs to the module $\mathcal{M}$. Then we have the following consequences:*

(i) If S is diophantine with respect to the ideal $\mathcal{I}$, then there is an analytic transformation $U = id + u$ to $S+g$.

(ii) If S is α-diophantine with respect to the ideal $\mathcal{I}$, then there exists a Gevrey-α transformation to its Gevrey-α normal form $S+g$.

Proof: Remark that if $\mathcal{I}$ is empty and if we use the classical grading then this theorem is a linearization result. We rewrite the formal conjugation formula $U^{-1}\circ F\circ U = S+g$ as in (5.7) and consider the projection $\pi_{\mathcal{I}^\perp}$ of this formula. We obtain:

$$\pi_{\mathcal{I}^\perp}\left(S\circ(\mathrm{id}+u) - S - DS.u + f\circ(\mathrm{id}+u) - g - u\circ(S+g) + u\circ S\right) = \pi_{\mathcal{I}^\perp}\left(d_0(u)\right) \tag{5.11}$$

Now, because $u\in\left\{v\in(\mathrm{Ker}(d_0))^\perp \,|\, d_0(v)\in\mathcal{V}\right\}$, it follows that $\pi_{\mathcal{I}^\perp}(d_0(u)) = d_0(u)$. Furthermore, since $g\in\mathcal{M}$, we also have that $\pi_{\mathcal{I}^\perp}(g)=0$ and from the observation that

$$u\circ(S+g) - u\circ S = \sum_{n\geq 1}\frac{(Du)^{(n)}(S)(g,g,\ldots,g)}{n!},$$

and hence belongs to $\mathcal{I}$, we can conclude that $\pi_{\mathcal{I}^\perp}(u \circ (S+g) - u \circ S) = 0$. As a consequence equation (5.11) reduces to:

$$\pi_{\mathcal{I}^\perp}\left(S \circ (\mathrm{id}+u) - S - DS.u + f \circ (\mathrm{id}+u)\right) = d_0(u).$$

We now determine the coefficients of u using induction on the p-quasi-homogeneous degree. Suppose therefore that we already defined u_β, for $0 \leq \beta \leq \delta - 1$. We obtain

$$d_0(u_\delta) = \pi_\delta(\pi_{\mathcal{I}^\perp}\left(S \circ (\mathrm{id}+u) - S - DS.u + f \circ (\mathrm{id}+u)\right))$$

Now, since $u_\delta \in \mathrm{Ker}(d_{0,p,\delta})^\perp$ by supposition and because $\mathcal{O}_{p,\delta} = \mathrm{Ker}(d_{0,p,\delta})^\perp$ we know that there exists a $v_\delta \in \mathrm{Im}(d^*_{0,p,\delta}) =$ such that $u_\delta = d^*_{0,p,\delta}(v_\delta)$. We can suppose that the projection of v_δ on $\mathrm{Ker}(d^*_{0,p,\delta})$ is zero. [Because if the projection were $w_\delta \neq 0$, then $v_\delta - w_\delta$ would do the job.] Remember that $\mathrm{Ker}(d_{0,p,\delta}) = \mathrm{Ker}(\Box_\delta)$ as proven in Lemma 5.5. Let now $\pi_{p,\delta,\setminus\mathcal{I}_{p,\delta}}$ be the projection onto $\mathcal{O}_{p,\delta}$ of the eigenvectors of $\Box_{p,\delta}$ that are orthogonal to $\mathcal{I}_{p,\delta}$. Because $\Box_{p,\delta}(v_\delta) = d_{0,p,\delta}(u_\delta)$; we conclude that

$$\begin{aligned}
\pi_{p,\delta,\setminus\mathcal{I}_{p,\delta}} \circ \pi_{\mathcal{I}^\perp}(\Box_{p,\delta}(v_\delta)) &= \pi_{p,\delta,\setminus\mathcal{I}_{p,\delta}} \circ \pi_{\mathcal{I}^\perp}(\Box_{p,\delta}(\sum_{\lambda \in \sigma_{p,\delta,\setminus\mathcal{I}_{p,\delta}}} v_\lambda)) \\
&= \sum_{\lambda \in \sigma_{p,\delta,\setminus\mathcal{I}_{p,\delta}}} \lambda v_\lambda
\end{aligned}$$

Now put $u_\lambda = d^*_{0,p,\delta} v_\lambda$, then $u_\delta = \sum_{\lambda \in \sigma_{p,\delta,\setminus\mathcal{I}_{p,\delta}}} u_\lambda$, and

$$\begin{aligned}
\left(\min_{\lambda \in \sigma_{p,\delta,\setminus\mathcal{I}_{p,\delta}}} \sqrt{\lambda}\right) ||u||_{p,\delta} &\leq \left|\left| \sum_{\lambda \in \sigma_{p,\delta,\setminus\mathcal{I}_{p,\delta}}} \lambda v_\lambda \right|\right| \\
&\leq \left|\left| \sum_{\lambda \in \sigma_{p,\delta,\setminus\mathcal{I}_{p,\delta}}} \Box_{p,\delta} v_\lambda \right|\right| \\
&\leq ||d_{0,p,\delta} u_\delta|| \\
&\leq ||\pi_\delta \pi_{\mathcal{I}^\perp}\left(S \circ (\mathrm{id}+u) - S - DS.u + f \circ (\mathrm{id}+u)\right)|| \\
&\leq ||S \circ (\mathrm{id}+u) - S - DS.u + f \circ (\mathrm{id}+u)|| \,.
\end{aligned} \tag{5.12}$$

We now make a further estimate of the right hand side of the last inequality. Remember that we have a natural decomposition of $f = \sum_{\mu \geq 1} f_\mu$, where $f_\mu \in \mathcal{O}_{p,\mu}$. We can decompose f_μ in terms of the regular degree. We have that $f_\mu = \sum_{r=\mu_*}^{\mu^*} f_{\mu,r}$, where $\mu_* = \left\lfloor \dfrac{\min_{1\leq i \leq n}(\mu + p_i)}{\overline{p}} \right\rfloor$, $\mu^* = \left\lfloor \dfrac{\max_{1\leq i \leq n}(\mu + p_i)}{\underline{p}} \right\rfloor$, $\overline{p} = \max_{1 \leq i \leq n} p_i$ and $\underline{p} = \min_{1 \leq i \leq n} p_i$. This decomposition allows us to give an estimate on the composition (we use $u_0 = \mathrm{id}$):

$$\begin{aligned}
\pi_\delta f(\mathrm{id}+u) = \pi_\delta\left(f_\mu(\mathrm{id}+u)\right) &= \pi_\delta\left(\sum_{\mu\geq 1}\sum_{r=\mu_*}^{\mu^*} f_{\mu,r}(\mathrm{id}+u)\right) \\
&= \pi_\delta\left(\sum_{\mu\geq 1}\sum_{r=\mu_*}^{\mu^*} \widetilde{f}_{\mu,r}(\mathrm{id}+u,\ldots,\mathrm{id}+u)\right) \\
&= \sum_{\mu\geq 1}\sum_{r=\mu_*}^{\mu^*}\sum_{\delta_1+\ldots+\delta_r=\delta} \widetilde{f}_{\mu,r}(u_{\delta_1},\ldots,u_{\delta_r}),
\end{aligned} \tag{5.13}$$

in this formula $\widetilde{f}_{\mu,r}$ is the unique r-linear symmetric form associated to $f_{\mu,r}$. We will use the estimate

$$\left|\left|\widetilde{f}_{\mu,r}(u_{\delta_1},\ldots,u_{\delta_r})\right|\right|_{p,\delta} \leq \frac{M}{\rho^r}\,||u_{\delta_1}||_{p,\delta_1}\cdots||u_{\delta_r}||_{p,\delta_r}\,, \tag{5.14}$$

following from the fact that f is analytic and Proposition 5.1. We also need the estimate

$$\begin{aligned}
||\pi_\delta(S\circ(\mathrm{id}+u) - DS.u - S)||_{p,\delta} &= \left|\left|\sum_{r=0}^{0^*}\sum_{\delta_1+\ldots+\delta_r=\delta} S_{0,r}(u_{\delta_1},\ldots,u_{\delta_r})\right|\right|_{p,\delta} \\
&\leq M'\sum_{r=0}^{0^*}\sum_{\delta_1+\ldots+\delta_r=\delta} ||u_{\delta_1}||_{p,\delta_1}\cdots||u_{\delta_r}||_{p,\delta_r}\,.
\end{aligned} \tag{5.15}$$

We define

$$\sigma_\delta := \sum_{\mu\geq 1}\sum_{r=\mu_*}^{\mu^*}\sum_{\delta_1+\ldots+\delta_r=\delta}\frac{M}{\rho^r}\,||\sigma_{\delta_1}||_{p,\delta_1}\cdots||\sigma_{\delta_r}||_{p,\delta_r} + \sum_{r=0}^{0^*}\sum_{\delta_1+\ldots+\delta_r=\delta}||\sigma_{\delta_1}||_{p,\delta_1}\cdots||\sigma_{\delta_r}||_{p,\delta_r}\,. \tag{5.16}$$

For the reader who is not familiar with the notation we repeat that $0^* = \lfloor \overline{p}/\underline{p}\rfloor$. We make the following observation:

Lemma 5.9 *The estimate $||u_\delta||_{p,\delta} \leq \eta_\delta\sigma_\delta$ holds recursively, where η_δ is defined in (5.10).*

Proof: The proof is done by induction on δ and follows from the estimates:

$$\begin{aligned}
a_\delta \, ||u||_{p,\delta} &= \left(\min_{\lambda \in \sigma_{p,\delta,} \setminus \mathcal{I}_{p,\delta}} \sqrt{\lambda} \right) ||u||_{p,\delta} \\
&\overset{(*)}{\leq} \sum_{\mu \geq 1} \sum_{r=\mu_*}^{\mu^*} \sum_{\delta_1+\ldots+\delta_r=\delta} \frac{M}{\rho^r} \, ||u_{\delta_1}||_{p,\delta_1} \ldots ||u_{\delta_r}||_{p,\delta_r} \\
&\quad + \sum_{r=0}^{0^*} \sum_{\delta_1+\ldots+\delta_r=\delta} ||u_{\delta_1}||_{p,\delta_1} \ldots ||u_{\delta_r}||_{p,\delta_r} \\
&\leq \sum_{\mu \geq 1} \sum_{r=\mu_*}^{\mu^*} \sum_{\delta_1+\ldots+\delta_r=\delta} \frac{M}{\rho^r} \sigma_{\delta_1} \eta_{\delta_1} \ldots \sigma_{\delta_r} \eta_{\delta_r} \\
&\quad + M' \sum_{r=0}^{0^*} \sum_{\delta_1+\ldots+\delta_r=\delta} \sigma_{\delta_1} \eta_{\delta_1} \ldots \sigma_{\delta_r} \eta_{\delta_r} \\
&\leq \overset{*}{\max_{\delta_1+\delta_2+\ldots+\delta_r=\delta}} (\eta_{\delta_1} \eta_{\delta_2} \ldots \eta_{\delta_r}) \sigma_\delta .
\end{aligned}$$

Combine (5.12), (5.12), (5.12), (5.12) to obtain $(*)$. Use the definition of η_δ given by (5.10) to finish the proof. □

We show in a separate Lemma 5.10 that the series $\sigma(t) = \sum_{\delta \geq 1} \sigma_\delta t^\delta$ is an analytic power series. Consequently we are able to conclude that $\sigma_\delta \leq R_1^\delta$ for a certain positive R_1. Observing that the diophantine condition given by Definition 5.7 is equivalent with $\eta_\delta \leq R_2^\delta (\delta!)^\alpha$, for a certain α; we conclude that $||u_\delta||_{p,\delta} \leq R_1 R_2 (\delta!)^\alpha$ and hence the transformation $U = \mathrm{id} + u$ is Gevrey-α. □

Lemma 5.10 *Let σ_δ be defined as in (5.16), then the corresponding power series $\sigma(t) = \sum_{\delta \geq 1} \sigma_\delta t^\delta$ is convergent.*

Proof: We will use the implicit function theorem. We rewrite the first sum appearing in the definition of σ_δ using the Newton formula. We see that

$$\sum_{\mu=1}^{\delta} \sum_{r=\mu_*}^{\mu^*} \sum_{\delta_1+\ldots+\delta_r+\mu=\delta} ||\sigma_{\delta_1}||_{p,\delta_1} \ldots ||\sigma_{\delta_r}||_{p,\delta_r} \frac{M}{\rho^r} = M \sum_{\mu=1}^{\delta} \pi_{\delta-\mu} \left(\left(\frac{\sigma(t)}{\rho} \right)^{\mu_*} + \ldots + \left(\frac{\sigma(t)}{\rho} \right)^{\mu^*} \right).$$

Hence we observe that it is the coefficient of t^δ in the Taylor series of $F(\sigma(t), t)$, where

$$F(z,t) := M \sum_{\mu=1}^{\delta} \sum_{r=\mu_*}^{\mu^*} \left(\frac{z}{\rho} \right)^r t^\mu .$$

We define

$$P(z) = \sum_{r=0}^{0^*} \left(z^r - \sigma_0^r - r\sigma_0^r \left(z - \sigma_0\right)\right),$$

and remark that $P(\sigma_0) = 0$, $DP(\sigma_0) = 0$ and

$$\pi_\delta \left(P(\sigma(t))\right) = M \sum_{r=0}^{0^*} \sum_{\delta_1+\ldots+\delta_r=\delta} \sigma_{\delta_1} \ldots \sigma_{\delta_r},$$

where the last sum is taken over the set of tuples $(\delta_1, \ldots, \delta_r)$ such that $\delta_1+\ldots+\delta_r = \delta$ and at least two of the $\delta_i \geq 1$. As a consequence, the power series $\sigma(t)$ is a formal solution of $G(\sigma(t), t) = \sigma(t) - \sigma_0$, where $G(z, t) = F(z, t) + P(z)$ and $\sigma(0) = \sigma_0$. Since the equation is analytic, using the implicit function theorem we have an analytic solution. It follows that the power series $\sigma(t)$ converges for $|t|$ small enough. □

Gevrey results

The results in this section are of a different nature then the results in Section 5.1.3: instead of linearizing a diffeomorphism, we will normalize (=remove all non-resonant terms) for a given diffeomorphism F. We will need some simplifications in order to be able to perform the proofs for these Gevrey-type results. These simplifications may not be necessary, but, until now, we have not yet found a method that can be used without these simplifications. Therefore, in this section we will restrict ourselves to the classical case where $p = (1, \ldots, 1)$, and $S = A$ is an invertible, semi-simple linear part. We will also suppose that $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is in diagonal form. Remark that this is not a further restriction. We will also use a somewhat different notation. We put $F = A \circ (\mathrm{id} + f)$ as the original diffeomorphism, where f contains terms of degree at least 2, $U = \mathrm{id} + u$ for the coordinate transformation and $G = A \circ (\mathrm{id} + g)$ for the ultimate normal form. The conjugation equation becomes

$$\begin{aligned}
& U^{-1} \circ F \circ U = G \\
\Leftrightarrow\ & A \circ (\mathrm{id} + f) \circ (\mathrm{id} + u) = (\mathrm{id} + u) \circ A \circ (\mathrm{id} + g) \\
\Leftrightarrow\ & \mathrm{id} + u + f \circ (\mathrm{id} + u) = \mathrm{id} + g + A^{-1} \circ u \circ A \circ (\mathrm{id} + g).
\end{aligned}$$

Put now $v = A^{-1} \circ u \circ A$, then $u = A \circ v \circ A^{-1}$ and

$$f \circ (\mathrm{id} + A \circ v \circ A^{-1}) + v - v \circ (\mathrm{id} + g) = g + v - A \circ v \circ A^{-1}. \tag{5.17}$$

Remark that we used the linearity of the operators A and A^{-1} at this point, but not the fact that they are semi-simple. For this section, it is useful to introduce the following Lie operator:

$$d_{0,\delta} : \mathcal{O}_\delta \longrightarrow \mathcal{O}_\delta : v \mapsto v - A \circ v \circ A^{-1}.$$

As before, we also define the adjoint operator $d^*_{0,\delta}$ and the corresponding box operator $\Box_\delta = d_{0,\delta}d^*_{0,\delta}$. In this case the calculation of the eigenvectors and eigenvalues is not difficult. This follows from the fact that A is diagonal. Let us explain this first. Let $k \in \mathbb{N}^n$ and $j \in \{1,\ldots,n\}$. We use the short hand notation $x^ke_j = (0,\ldots,0,x_1^{k_1}x_2^{k_2}\ldots x_n^{k_n},0,\ldots,0)$, where the non-zero position is at position j. Now $d_{0,\delta}(x^ke_j) = (1-\frac{\lambda_j}{\lambda^k})x^ke_j$ and it follows that $\Box_\delta(x^ke_j) = \left|1-\frac{\lambda_j}{\lambda^k}\right|^2 x^ke_j$. We introduce a Siegel type condition. This condition is stronger than the diophantine condition defined in definition 5.7, and will hence also be sufficient to show convergence in the case that the equation happens to be formally linearizable. It is an open question whether the results we prove in this section will also hold under the assumptions of the weaker diophantine conditions.

Definition 5.11 *Let $C, \tau > 0$, we say that A satisfies a Siegel condition of type τ if*

$$\frac{1}{\left|1-\frac{\lambda_j}{\lambda^k}\right|} \leq C|k|^\tau$$

holds for any $k \in \mathbb{N}^n$, $|k| \geq 2$, and all $j \in \{1,\ldots,n\}$.

At this point we can start the normal form iteration procedure. It should be noted to the reader that in a first stage of the proof of the main theorem we will construct the majorating series

$$\frac{z_\delta}{3C\delta^\tau} = \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} z_l z_{\delta_1}\ldots z_{\delta_l},$$

see formula (5.25), of both the normal form and the normal form transformation, separately. The Gevrey-character of the solution of this recursively defined series is studied afterwards, and is of independent interest. We will again encounter the same complication as in Section 2.2.5: the removal of terms cannot be done in one step by the current method. Therefore we introduce the good set

$$\mathbf{G} = \left\{(k,j) \in \mathbb{N}^n \times \{1,\ldots,n\} \,|\, |k| \geq 2 \text{ and } \left(\left|\frac{\lambda_j}{\lambda^k}\right| \leq 1 \text{ and } \frac{\lambda_j}{\lambda^k} \neq 1\right)\right\}$$

$\mathbf{G}$, this is the set of terms that will be removed from the normal form as an initial step. We also define the complementary bad set

$$\mathbf{B} = \{(k,j) \in \mathbb{N}^n \times \left\{1,\ldots,n\} \,|\, |k| \geq 2 \text{ and } \left(\left|\frac{\lambda_j}{\lambda^k}\right| > 1 \text{ or } \frac{\lambda_j}{\lambda^k} = 1\right)\right\}. \tag{5.18}$$

Remark that $\mathbf{G}$ (resp. $\mathbf{B}$) depends on the linear part A. We will use the notation $\mathbf{G}_A$ (resp. $\mathbf{B}_A$) to note this dependence whenever necessary. For future use, we also define

$$\widetilde{\mathbf{B}} := \left\{(k,j) \in \mathbb{N}^n \times \{1,\ldots,j\} \,|\, |k| \geq 2, \text{ and } \left(\left|\frac{\lambda_j}{\lambda^k}\right| < 1 \text{ or } \frac{\lambda_j}{\lambda^k} = 1\right)\right\}. \tag{5.19}$$

We define, for formal power series $f = \sum_{k \in \mathbb{N}^n} f_k x^k e_j$, the corresponding projections $\pi_{\mathbf{G}}(f) = \sum_{(k,j) \in \mathbf{G}} f_k x^k e_j$ and $\pi_{\mathbf{B}}(f) = \sum_{(k,j) \in \mathbf{B}} f_k x^k e_j$. We state our main result concerning Gevrey-normal forms.

Theorem 5.12 *Let $F = A + f$ be an analytic diffeomorphism, where A is its semi-simple linear part and $f = \sum_{\delta \geq 2} f_\delta$. Suppose that A satisfies the Siegel condition of order τ. There exists a transformation $U = id + u$ with $\pi_{\mathbf{B}}(u) = 0$ and a normal form $G = id + g$ with $\pi_{\mathbf{G}}(G) = 0$ such that the conjugation $U^{-1} \circ F \circ U = G$ holds, moreover U and G are Gevrey-$(1 + \tau)$.*

Proof: During the proof we will repeatedly use the following lemma with trivial proof without explicitly quoting it.

Lemma 5.13 *Let δ_0 be a fixed natural number and let $z_{\delta_0} \geq 0$. Suppose that for every $\delta > \delta_0$ in $\mathbb{N}$ we have that z_δ is recursively defined by polynomials with positive coefficients H_δ that depend on all z_i, $\delta_0 \leq i \leq \delta - 1$. Hence $(z_\delta)_{\delta \geq \delta_0}$ is a solution to the recursively defined equations*

$$z_\delta = H_\delta.$$

If there exists polynomials J_δ with positive coefficients depending on all z_i, $\delta_0 \leq i \leq \delta - 1$ and if $H_\delta \leq J_\delta$, then the unique solution w_δ, $w_{\delta_0} = z_{\delta_0}$, of the recursively defined equations

$$w_\delta = J_\delta,$$

majorates z_δ. This is $z_\delta \leq w_\delta$, for all $\delta \geq \delta_0$.

In order to prove Theorem 5.12 we will build an inductively defined one-dimensional formal power series $z(y) = \sum_{\delta \geq 2} z_\delta y^\delta$ with positive coefficients z_δ that is a common majorant of u and g. By this we mean that if $g = \sum_{\delta \geq 2} g_\delta$ and $u = \sum_{\delta \geq 2} u_\delta$, then $\max(||g_\delta||_\delta, ||u_\delta||_\delta) \leq z_\delta$ for all $\delta \geq 2$. In order to do so, we need to write down the compositions appearing in equation (5.17) explicitly. We recall the so called Faà di Bruno formula for the formal composition of two formal power series $h = \sum_{k \geq 1} h_k$ and $m = \sum_{k \geq 1} m_k$.

$$h \circ m = \sum_{\delta \geq 1} \sum_{l=1}^{\delta} \sum_{\delta_1 + \delta_2 + \ldots + \delta_l = \delta} \widetilde{h}_l(m_{\delta_1}, \ldots, m_{\delta_l}). \tag{5.20}$$

We apply $\pi_{\mathbf{G}}$ and $\pi_{\mathbf{B}}$ to equation (5.17). We obtain

$$\begin{aligned} \pi_{\mathbf{G}}(v - A \circ v \circ A^{-1} + g) &= v - A \circ v \circ A^{-1} \\ &= \pi_{\mathbf{G}}(f \circ (\mathrm{id} + A \circ v \circ A^{-1}) + v - v \circ (\mathrm{id} + g)) \qquad (5.21) \\ \pi_{\mathbf{B}}(v - A \circ v \circ A^{-1} + g) &= g = \pi_{\mathbf{B}}(f \circ (\mathrm{id} + A \circ v \circ A^{-1}) + v - v \circ (\mathrm{id} + g)). \end{aligned}$$

We use the norm $||.||_\delta$ associated to the inner product defined by (5.5) and observe that

$$\begin{aligned}\frac{1}{C|\delta|^\tau}||v_\delta||_\delta &\leq \min_{|k|=\delta}\left(\left|1-\frac{\lambda_j}{\lambda^k}\right|\right)||v_\delta||_\delta \leq \min_{|k|=\delta}\left(\left|1-\frac{\lambda_j}{\lambda^k}\right|\right)\sum_{|k|=\delta}|v_k|\frac{k!}{\delta!}\\ &\leq \sum_{|k|=\delta}\left|1-\frac{\lambda_j}{\lambda^k}\right|.|v_k|\frac{k!}{\delta!} = ||d_0(v_\delta)||_\delta\end{aligned}$$

Hence it follows that

$$\begin{aligned}\frac{1}{C|\delta|^\tau}\,||v_\delta||_\delta &\leq ||d_0(v_\delta)||_\delta = \left|\left|\pi_\delta(v - A\circ v\circ A^{-1})\right|\right|_\delta\\ &= \left|\left|\pi_\delta\left(\pi_{\mathbf{G}}\left(f\circ(\mathrm{id}+A\circ v\circ A^{-1})+v-v\circ(\mathrm{id}+g)\right)\right)\right|\right|_\delta\\ &\leq \left|\left|\pi_\delta\left(f\circ(\mathrm{id}+A\circ v\circ A^{-1})+v-v\circ(\mathrm{id}+g)\right)\right|\right|_\delta\\ &\leq \left|\left|\pi_\delta\left(f\circ(\mathrm{id}+A\circ v\circ A^{-1})\right)\right|\right|_\delta + ||\pi_\delta\left(v-v\circ(\mathrm{id}+g)\right)||_\delta \qquad (5.22)\\ ||g_\delta||_\delta &= \left|\left|\pi_\delta\left(\pi_{\mathbf{B}}\left(f\circ(\mathrm{id}+A\circ v\circ A^{-1})+v-v\circ(\mathrm{id}+g)\right)\right)\right|\right|_\delta\\ &\leq \left|\left|\pi_\delta\left(f\circ(\mathrm{id}+A\circ v\circ A^{-1})\right)\right|\right|_\delta + ||\pi_\delta\left(v-v\circ(\mathrm{id}+g)\right)||_\delta\,. \qquad (5.23)\end{aligned}$$

We now make an estimate on the common right hand side of (5.23) and (5.22), using the composition formula (5.20). We start with the first term. We put $u_1 = \mathrm{id}$ for convenience. Let $\widetilde{f_l}$ be the l-multi-linear mapping associated to the l-homogeneous polynomial f_l in the Taylor series expansion of f. According to Proposition A.9 from [37] we have that f is analytic if and only if for all $l \geq 2$, $\max(||\widetilde{f_l}||_l, ||f_l||_l) \leq MR^l$ for a certain $M, R > 0$. We may suppose that $M = 1$ and $R = E$ is as small as we want after applying a rescaling operator. Indeed if we define the rescaling of a power series as $\mathcal{R}_\kappa(h)(x) = \frac{1}{\kappa}h(\kappa x)$, then $U^{-1}\circ F\circ U = G$ if and only if $\mathcal{R}_\kappa(U^{-1})\circ\mathcal{R}_\kappa(F)\circ\mathcal{R}_\kappa(U) = \mathcal{R}_\kappa(G)$, and we can solve the latter problem instead. This problem admits the estimates $\max(||\mathcal{R}_\kappa(\widetilde{f_l})||, ||\mathcal{R}_\kappa(f)_l||) \leq M\kappa^{l-1}R^l$, and it suffices to choose κ small enough in order to obtain $M\kappa^{l-1}R^l \leq E^l$. We rescale a little more to make sure that $\max(||\widetilde{v}_2||_2, ||g_2||_2) \leq \sqrt{E}$. We will also make sure that $E \leq \min\{1, (C.2.3^\tau)^2\}$. We then proceed the proof with the rescaled series of U, F, G, but stick to the original notation and obtain

$$\begin{aligned}||f\circ(\mathrm{id}+A\circ v\circ A^{-1})||_\delta &= ||\pi_\delta(f\circ(\mathrm{id}+u))||_\delta\\ &\leq \sum_{l=2}^{\delta}\sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta}||\widetilde{f}_l(u_{\delta_1},\ldots,u_{\delta_l})||_\delta\\ &\leq \sum_{l=2}^{\delta}\sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta}||\widetilde{f}_l||_l.||u_{\delta_1}||_{\delta_1}\ldots||u_{\delta_l}||_{\delta_l}\\ &\leq \sum_{l=2}^{\delta}\sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta}E^l||u_{\delta_1}||_{\delta_1}\ldots||u_{\delta_l}||_{\delta_l}\\ &\overset{(*)}{\leq} \sum_{l=2}^{\delta}\sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta}E^l||v_{\delta_1}||_{\delta_1}\ldots||v_{\delta_l}||_{\delta_l}.\end{aligned}$$

In the last inequality follows $(*)$ from the fact that v contains only good terms according to equation (5.21) and $u(x)=(A\circ v\circ A^{-1})(x)=\sum_{(k,j)\in\mathbf{G}}v_k\frac{\lambda_j}{\lambda^k}x^ke_j$, such that

$$u_\delta(x)=\sum_{(k,j)\in\mathbf{G},|k|=\delta}v_k\frac{\lambda_j}{\lambda^k}x^ke_j.$$

Because $(k,j)\in\mathbf{G}$ implies $|\frac{\lambda_j}{\lambda^k}|\leq 1$, we have that $||u_\delta||_\delta\leq||v_\delta||_\delta$. We estimate the second term using the same technique. Put $g_1=\mathrm{id}$.

$$\begin{aligned}||\pi_\delta(v-v\circ(\mathrm{id}+g))||_\delta &\leq ||\sum_{l=2}^{\delta-1}\sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta}\widetilde{v}_l(g_{\delta_1},\ldots,g_{\delta_l})||_\delta\\ &\leq \sum_{l=2}^{\delta-1}\sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta}||\widetilde{v}_l||.||g_{\delta_1}||_{\delta_1}\ldots||g_{\delta_l}||_{\delta_l}.\end{aligned}$$

Define

$$\begin{aligned}z_1&=1\\ z_2&=\sqrt{E}\\ \frac{z_\delta}{C\delta^\tau}&=\sum_{l=2}^{\delta}\sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta}E^lz_{\delta_1}\ldots z_{\delta_l}+\sum_{l=2}^{\delta-1}\sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta}z_lz_{\delta_1}\ldots z_{\delta_l},\end{aligned}\tag{5.24}$$

for $\delta\geq 3$. Then the series $z_\delta\geq\max(||g_\delta||_\delta,||v_\delta||_\delta,||\widetilde{v}_\delta||_\delta)\geq\max(||g_\delta||_\delta,||u_\delta||_\delta)$. (Remember that the rescaling is chosen such that $\max(||\widetilde{v}_2||_2,||g_2||_2)\leq\sqrt{E}$.) Hence we have that if the series $\sum_{\delta\geq 2}z_\delta x^\delta$ is a Gevrey power series, then the same is true for u and g.

We show by induction that $z_l \geq E^l$, for $l \geq 2$. For $l = 2$, we have $z_2 = \sqrt{E} \geq E^2$, since $E \leq \min\{1, (C.2.3^\tau)^2\}$. Let now $\delta \geq 3$ and suppose that we have shown that $z_l \geq E^l$ for $2 \leq l \leq \delta - 1$. It then follows that

$$z_\delta \overset{(a)}{\geq} C.\delta^\tau \left((\delta-1)z_{\delta-1}z_2 + \ldots\right) \geq C(\delta-1)\delta^\tau z_{\delta-1}z_2 \overset{(b)}{\geq} C.2.3^\tau E^{\delta-1}.\sqrt{E} \overset{(c)}{\geq} E^\delta.$$

In (a), the term that is separated is the term corresponding to $l = \delta - 1$ in the sum $\sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} z_l z_{\delta_1} \ldots z_{\delta_l}$, in (b) we used the induction hypothesis $z^{l-1} \geq E^{l-1}$ and $z_2 = \sqrt{E}$, and in (c) we used $E \leq \min\{1, (C.2.3^\tau)^2\}$. It follows that for all $\delta \geq 3$, we have that $z_\delta \geq E^\delta$ and hence:

$$\frac{z_\delta}{C\delta^\tau} \leq E^2 z_{\delta-1}(\delta-1) + E^\delta + \sum_{l=3}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} z_l z_{\delta_1} \ldots z_{\delta_l} + \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} z_l z_{\delta_1} \ldots z_{\delta_l}.$$

Now, because

$$E^2 z_{\delta-1}(\delta-1) = E^{\frac{3}{2}} z_2(\delta-1) \leq E^{\frac{3}{2}} \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} z_l z_{\delta_1} \ldots z_{\delta_l};$$

and

$$E^\delta \leq E^{\delta-1}.E \leq z_{\delta-1}E = \sqrt{E}(\delta-1)z_2 z_{\delta-1} \leq \sqrt{E} \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} z_l z_{\delta_1} \ldots z_{\delta_l},$$

it follows that

$$\begin{aligned} \frac{z_\delta}{C\delta^\tau} &\leq \left(1 + \sqrt{E} + E^{\frac{3}{2}}\right) \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} z_l z_{\delta_1} \ldots z_{\delta_l} \\ &\leq 3 \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} z_l z_{\delta_1} \ldots z_{\delta_l}. \end{aligned}$$

Hence the solution of (5.24) is majorated by the solution of the recursively defined numbers

$$\begin{aligned} z_1 &= 1 \\ z_2 &= \sqrt{E} \\ \frac{z_\delta}{3C\delta^\tau} &= \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} z_l z_{\delta_1} \ldots z_{\delta_l}. \end{aligned} \tag{5.25}$$

We put $M = 3C$ for convenience.

Remark 5.14 *The above equation is very useful for numerical experimentation. We implement the recursion process, and we make a picture of* $\frac{\ln(z_\delta)}{\ln(\delta!)}$. *If* $z_\delta \leq R^\delta \delta!^s$, *for a certain* s, *then we expect that*

$$\frac{\ln(z_\delta)}{\ln(\delta!)} \leq s + \frac{\ln(R)\delta}{\ln(\delta!)} \approx s,$$

for large values of δ. *For numerical computations it seems beneficial to plot* $\frac{\ln(z_\delta)}{\ln(\delta^\delta)}$. *One can argue, using the Stirling formula, that they are asymptotically the same. I include two figures for the case* $\tau = 0$, $M = 2$; *where we expect that the result is Gevrey of order* 1. *In Figure 5.1(a) we plot* $\frac{\ln(z_\delta)}{\ln(\delta!)}$ *as a function of* δ, *in Figure 5.1(b) we plot* $\frac{\ln(z_\delta)}{\ln(\delta^\delta)}$.

One can see that especially Figure 5.1(b) is convincing.

We will proceed now with the proof of Theorem 5.12. We define the formal power series $z(x) = \sum_{\delta \geq 2} z_\delta x^\delta$. It follows from equations (5.25) that

$$\sum_{\delta \geq 2} \frac{z_\delta x^\delta}{M \delta^\tau} = z(x + z(x)) - z(x). \tag{5.26}$$

After rescaling the conjugation equation, we may suppose that the first N coefficients satisfy $z_2 < \alpha, \ldots, z_N < \alpha$ (this follows e.g. recursively since z_2 can be chosen initially as small as needed). The subsequent proposition shows that the power series $z(x)$ is Gevrey-$(1 + \tau)$ and finishes the proof. □

Proposition 5.15 *For each* $M > 0$ *there exists an* α *and an* N *such that if* $z_2 < \alpha, \ldots, z_N < \alpha$, *then the solution to equation (5.26) is Gevrey-*$1 + \tau$.

Proof: We start the proof with a refined estimate of the right hand side of equation (5.26) using the Faà di Bruno formula. Therefore let $v(x) = \sum_{\delta \geq 2} v_\delta x^\delta$ be an arbitrary power series. We put $v_1 = 1$ for convenience. We obtain:

$$v(x + v(x)) - v(x) = \sum_{\delta \geq 2} \sum_{l=2}^{\delta - 1} v_l \sum_{\delta_1 + \ldots + \delta_l = \delta} v_{\delta_1} \ldots v_{\delta_l}.$$

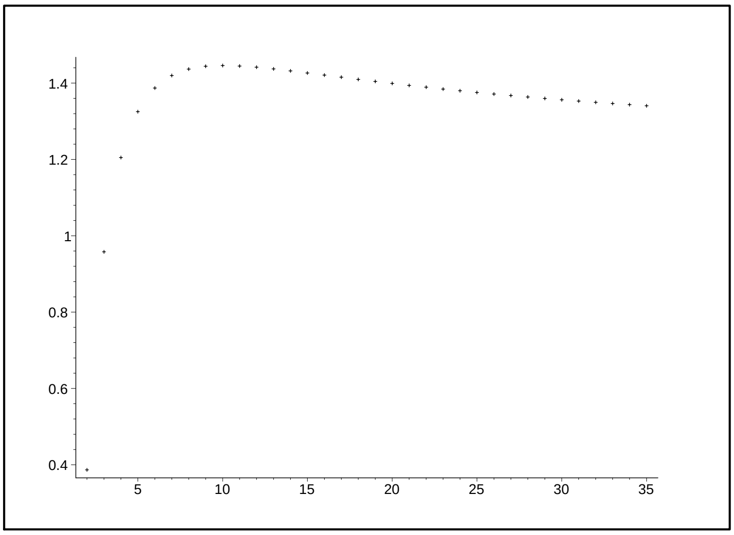
Define now for $1 \leq l \leq \delta$ and $\delta \geq 2$

$$H_{l,\delta} = v_l \sum_{\delta_1 + \ldots + \delta_l = \delta} v_{\delta_1} \ldots v_{\delta_l},$$

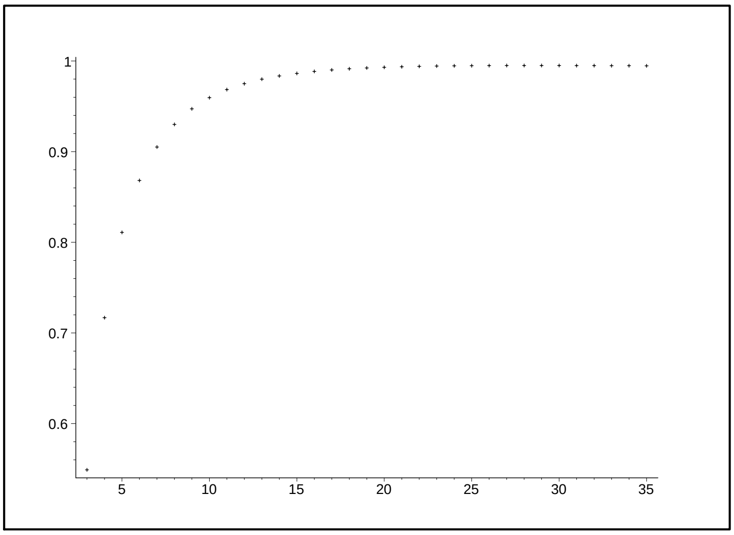
$$H_\delta = \sum_{l=2}^{\delta - 1} H_{l,\delta}.$$

For future use we remark that

$$H_{1,\delta} = v_\delta,\ H_{\delta,\delta} = v_\delta. \tag{5.27}$$



(a) A plot of $\frac{\ln(z_\delta)}{\ln(\delta!)}$



(b) A plot of $\frac{\ln(z_\delta)}{\ln(\delta^\delta)}$

Figure 5.1: Estimating Gevrey order numerically

We observe that

$$
\begin{aligned}
H_{l,\delta} &= v_l \left(v_1 \sum_{\delta_1+\ldots+\delta_{l-1}=\delta-1} v_{\delta_1}\ldots v_{\delta_{l-1}} + \ldots + v_{\delta-l+1} \sum_{\delta_1+\ldots+\delta_{l-1}=l-1} v_{\delta_1}\ldots v_{\delta_{l-1}} \right) \\
&= \frac{v_l}{v_{l-1}} \sum_{m=l-1}^{\delta-1} v_{\delta-m} H_{l-1,m}.
\end{aligned}
$$

It follows that (I explain in Remark 5.16 why I want to split H_δ in this way!)

$$
\begin{aligned}
H_\delta &= \sum_{l=2}^{\delta-1} H_{l,\delta} \\
&= \sum_{l=2}^{\delta-1} \frac{v_l}{v_{l-1}} \sum_{m=l-1}^{\delta-1} v_{\delta-m} H_{l-1,m} \\
&= \sum_{m=1}^{\delta-1} v_{\delta-m} \sum_{l=2}^{m+1} \frac{v_l}{v_{l-1}} H_{l-1,m} - \frac{v_\delta}{v_{\delta-1}} H_{\delta-1,\delta-1} \\
&= \sum_{m=1}^{\delta-1} v_{\delta-m} \sum_{l=1}^{m} \frac{v_{l+1}}{v_l} H_{l,m} - \frac{v_\delta}{v_{\delta-1}} H_{\delta-1,\delta-1} \\
&= \sum_{m=3}^{\delta-1} v_{\delta-m} \sum_{l=1}^{m} \frac{v_{l+1}}{v_l} H_{l,m} - \frac{v_\delta}{v_{\delta-1}} H_{\delta-1,\delta-1} + v_{\delta-1} v_2 H_{1,1} + v_{\delta-2} \frac{v_2}{v_1} H_{1,2} + v_{\delta-2} \frac{v_3}{v_2} H_{2,2} \\
&= \sum_{m=3}^{\delta-1} v_{\delta-m} \sum_{l=2}^{m-1} \frac{v_{l+1}}{v_l} H_{l,m} + \sum_{m=3}^{\delta-1} v_{\delta-m} \frac{v_2}{v_1} H_{1,m} + \sum_{m=3}^{\delta-1} v_{\delta-m} \frac{v_{m+1}}{v_m} H_{m,m} \\
&\quad - \frac{v_\delta}{v_{\delta-1}} H_{\delta-1,\delta-1} + \left(v_{\delta-1} v_2 H_{1,1} + v_{\delta-2} \frac{v_2}{v_1} H_{1,2} + v_{\delta-2} \frac{v_3}{v_2} H_{2,2} \right) \\
&= \underbrace{\sum_{m=3}^{\delta-1} v_{\delta-m} \sum_{l=2}^{m-1} \frac{v_{l+1}}{v_l} H_{l,m}}_{T_1} + \overbrace{\sum_{m=3}^{\delta-1} v_{\delta-m} \frac{v_2}{v_1} H_{1,m}}^{T_2} + \underbrace{\sum_{m=3}^{\delta-2} v_{\delta-m} \frac{v_{m+1}}{v_m} H_{m,m}}_{T_3} \\
&\quad + \overbrace{\left(v_{\delta-1} v_2 H_{1,1} + v_{\delta-2} \frac{v_2}{v_1} H_{1,2} + v_{\delta-2} \frac{v_3}{v_2} H_{2,2} \right)}^{T_4}.
\end{aligned}
$$

Notice how the fraction $\frac{v_{l+1}}{v_l}$ plays an important role. Let $v(x) = \sum_{\delta\geq 2} v_\delta x^\delta = \sum_{\delta=2}^{N} \alpha x^\delta + \sum_{\delta\geq N+1} (\delta!)^{1+\tau} x^\delta$ and put $\beta := \frac{1}{M}$ for convenience. We show by induction on δ, for this choice of the v_δ and corresponding H_δ, that:

1. $z_\delta \leq v_\delta$,
2. $H_\delta \leq \frac{\beta}{\delta^\tau} (\delta!)^{1+\tau}$,

if α is sufficiently small and N is sufficiently large (we will choose α and N later). This finishes the proof since $z(x)$ is then majorated by the Gevrey-$(1+\tau)$ formal power series $v(x)$.

Clearly we have

$$\frac{z_\delta}{M\delta^\tau} \leq H_\delta.$$

It is hence sufficient to show $H_\delta \leq \frac{\beta}{\delta^\tau} (\delta!)^{1+\tau}$, since this implies that $z_\delta \leq (\delta!)^{1+\tau} v_\delta$. We estimate T_1, T_2, T_3, T_4 in order to obtain this, using the induction hypothesis for the estimates. We have

$$\begin{aligned} T_1 &= \sum_{m=3}^{\delta-1} v_{\delta-m} \sum_{l=2}^{m-1} \frac{v_{l+1}}{v_l} H_{l,m} \\ &\leq \sum_{m=3}^{\delta-1} v_{\delta-m} \frac{v_m}{v_{m-1}} \sum_{l=2}^{m-1} H_{l,m} \quad \text{(notice that } \frac{v_{m+1}}{v_m} \text{ is increasing with } m\text{)} \\ &= \sum_{m=3}^{\delta-1} v_{\delta-m} \frac{v_m}{v_{m-1}} H_m \qquad (5.28) \\ &= \frac{v_{\delta-1}}{v_{\delta-2}} H_{\delta-1} + v_2 \frac{v_{\delta-2}}{v_{\delta-3}} H_{\delta-2} + v_{\delta-3}\frac{v_3}{v_2} H_3 + \sum_{m=4}^{\delta-4} v_{\delta-m}\frac{v_m}{v_{m-1}} H_m \\ &\leq \frac{v_{\delta-1}}{v_{\delta-2}} H_{\delta-1} + v_2 \frac{v_{\delta-2}}{v_{\delta-3}} H_{\delta-2} + v_{\delta-3}\frac{v_3}{v_2} H_3 + v_4 \frac{v_{\delta-4}}{v_{\delta-5}} H_{\delta-4}(\delta-7) \end{aligned}$$

Remark 5.16 *We use expansion (5.28) to give a heuristic explanation of the entire estimate of H_δ. By looking at the terms closely, we observe that those for m small ($m \approx 3$) and m large ($m \approx \delta-1$) are biggest. To see this, observe that $H_m \approx (m!)^{1+\tau} m^\tau$, $v_{\delta-m} \approx ((\delta-m)!)^{1+\tau}$, and that the symmetric $(m!)(\delta-m)!$ decreases for $0 \leq m \leq \frac{\delta\pm1}{2}$ and increases for $\frac{\delta\pm1}{2} \leq m \leq \delta$. Using this idea to see the big terms, we split them off and estimate the remaining terms at once obtaining*

$$\begin{aligned} T_1 &\approx \frac{\beta}{\delta^\tau} (\delta!)^{1+\tau} (1 - \frac{1}{\delta} + \frac{C}{\delta^2}), \\ T_2 &\approx \frac{\beta}{\delta^\tau} (\delta!)^{1+\tau} (\frac{\alpha}{\delta} + \frac{C\alpha^2}{\delta^2}). \end{aligned}$$

A similar estimate holds for T_3 and T_4. α is now chosen such that the coefficients $H_\delta \approx (T_1+T_2+T_3+T_4) \leq \frac{\beta}{\delta^\tau} (\delta!)^{1+\tau}$. We need that the first few coefficients, say $v_2, \ldots, v_N$, are relatively small to obtain this bound.

Using the induction hypothesis we obtain that

$$\begin{aligned}
\frac{T_1\delta^\tau}{\beta v_\delta} \leq \frac{\delta^\tau}{\beta(\delta!)^{1+\tau}} &\Bigg[\frac{(\delta-1)!^{1+\tau}}{(\delta-2)!^{1+\tau}}\frac{\beta(\delta-1)!^{1+\tau}}{(\delta-1)^\tau} + \alpha\beta\frac{(\delta-2)!^{1+\tau}}{(\delta-3)!^{1+\tau}}\frac{(\delta-2)!^{1+\tau}}{(\delta-2)^\tau} + \\
&(\delta-3)!^{1+\tau}\alpha\beta + \alpha\beta\frac{(\delta-4)!^{1+\tau}}{(\delta-4)!^{1+\tau}}\frac{(\delta-4)!^{1+\tau}}{(\delta-4)^\tau}(\delta-7)\Bigg] \\
&\leq \frac{\delta-1}{\delta} + \frac{3^\tau\alpha}{\delta}\left(\frac{\delta-2}{\delta-1}\right)^{1+\tau} + \frac{\alpha}{\delta[(\delta-1)(\delta-2)]^{1+\tau}} + \frac{5^\tau\alpha(\delta-7)}{\delta[(\delta-1)(\delta-2)(\delta-3)]^{1+\tau}} \\
&\leq 1 - \frac{1}{\delta} + \frac{3^\tau\alpha}{\delta} + \frac{\alpha + 10.5^\tau\alpha}{\delta^2}. \qquad (5.29)
\end{aligned}$$

Continuing with the same technique and plugging in formulas (5.27) we have

$$\begin{aligned}
T_2 &= \sum_{m=3}^{\delta-1} v_{\delta-m}\frac{v_2}{v_1}H_{1,m} \\
&= v_2v_{\delta-1} + v_2v_2v_{\delta-2} + 2v_2v_3v_{\delta-3} + \sum_{m=4}^{\delta-4} v_2v_{\delta-m}v_m \\
&= v_2v_{\delta-1} + v_2v_2v_{\delta-2} + 2v_2v_3v_{\delta-3} + v_2v_4v_{\delta-4}(\delta-7),
\end{aligned}$$

and

$$\begin{aligned}
\frac{T_2\delta^\tau}{\beta v_\delta} &\leq \frac{\delta^\tau}{\beta(\delta!)^{1+\tau}}\left[\alpha(\delta-1)!^{1+\tau} + \alpha^2(\delta-2)^{1+\tau} + \alpha^2(\delta-4)!^{1+\tau}(\delta-7)\right] \\
&\leq \frac{\alpha}{\beta\delta} + \frac{\alpha^2 C_1}{\beta\delta^2}; \qquad (5.30) \\
T_3 &= \sum_{m=3}^{\delta-2} v_{\delta-m}\frac{v_{m+1}}{v_m}H_{m,m} = v_2v_{\delta-1} + v_3v_{\delta-2} + \sum_{m=3}^{\delta-4} v_{\delta-m}v_{m+1} \\
&\leq v_2v_{\delta-1} + v_3v_{\delta-2} + (\delta-6)v_4v_{\delta-3}; \\
\frac{T_3\delta^\tau}{\beta v_\delta} &\leq \frac{\delta^\tau}{\beta(\delta!)^{1+\tau}}\left[\alpha(\delta-1)!^{1+\tau} + \alpha(\delta-2)!^{1+\tau} + (\delta-4)\alpha(\delta-3)!^{1+\tau}\right] \\
&\leq \frac{\alpha}{\beta\delta} + \frac{\alpha C_2}{\delta^2}; \qquad (5.31) \\
T_4 &= v_{\delta-1}v_2H_{1,1} + v_{\delta-2}\frac{v_2}{v_1}H_{1,2} + v_{\delta-2}\frac{v_3}{v_2}H_{2,2} \\
&= v_{\delta-1}v_2 + v_{\delta-2}v_2 + v_{\delta-2}v_2; \\
\frac{T_4\delta^\tau}{\beta v_\delta} &\leq \frac{\delta^\tau}{\beta(\delta!)^{1+\tau}}\left[\alpha(\delta-1)!^{1+\tau} + \alpha(\delta-2)!^{1+\tau} + \alpha(\delta-2)!^{1+\tau}\right] \\
&\leq \frac{\alpha}{\delta} + \frac{\alpha C_3}{\delta^2}; \qquad (5.32)
\end{aligned}$$

Therefore using (5.29), (5.30), (5.31), (5.32), we see that

$$\frac{(T_1+T_2+T_3+T_4)\delta^\tau}{\beta v_\delta} \leq 1$$

if α is small enough and $\delta \geq N$. This determines the choice of α and N. □

We have the following extension of Theorem 5.12 to Gevrey-classes:

Theorem 5.17 *Let $F = A + f$ be a Gevrey-α diffeomorphism, where A is its semi-simple linear part and $f = \sum_{\delta\geq 2} f_\delta$. Suppose that A satisfies a Siegel condition of type τ. There exists a transformation $U = id + u$ with $\pi_{\mathbf{B}}(u) = 0$ and a normal form $G = id+g$ with $\pi_{\mathbf{G}}(G) = 0$ such that the conjugation $U^{-1} \circ F \circ U = G$ holds, moreover U and G are Gevrey-$(1+\tau+\alpha)$.*

Proof: The proof follows the same structure as the proof of Theorem 5.12. We explain the differences. From the start of the proof, the only different estimate is $||\widetilde{f_l}|| \leq E^l l!^\alpha$. Using this estimate, one has instead of equation (5.24), the following equation

$$\frac{z_\delta}{C\delta^\tau} = \sum_{l=2}^{\delta} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} E^l l!^\alpha z_{\delta_1}\ldots z_{\delta_l} + \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} z_l z_{\delta_1}\ldots z_{\delta_l}.$$

Define $w_\delta := z_\delta/\delta!^\alpha$, then

$$\begin{aligned}\frac{w_\delta}{C\delta^\tau} =& \sum_{l=2}^{\delta} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} E^l \left(\frac{l!\delta_1!\ldots\delta_l!}{\delta!}\right)^\alpha w_{\delta_1}\ldots w_{\delta_l} \\ &+ \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} \left(\frac{\delta_1!\ldots\delta_l!}{\delta!}\right)^\alpha w_l w_{\delta_1}\ldots w_{\delta_l} \\ \leq& \sum_{l=2}^{\delta} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} E^l w_{\delta_1}\ldots w_{\delta_l} + \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} w_l w_{\delta_1}\ldots w_{\delta_l}.\end{aligned}$$

Hence w_δ is majorated by the solution of

$$\frac{u_\delta}{C\delta^\tau} = \sum_{l=2}^{\delta} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} E^l u_{\delta_1}\ldots u_{\delta_l} + \sum_{l=2}^{\delta-1} \sum_{\delta_1+\delta_2+\ldots+\delta_l=\delta} u_l u_{\delta_1}\ldots u_{\delta_l}.$$

That solution satisfies $u_\delta \leq R^\delta(\delta!)^{1+\tau}$: this follows from the proof of Theorem 5.12 when starting from equation (5.24). Hence $z_\delta \leq R^\delta(\delta!)^{1+\tau+\alpha}$, finishing the theorem. □

In fact, at this point we have removed only 'half of the terms' that we intended to remove. We proceed with a similar idea as in Section 2.2.5 to remove the remaining non-resonant terms. Afterwards we give two interesting corollaries.

Proposition 5.18 *Let* $\mathbf{B}$ *be defined by (5.18) and* $\widetilde{\mathbf{B}}$ *by (5.19) and suppose that we have a diffeomorphism* $F(x) = x + f(x)$*, where* $f(x) = \sum_{(k,j)\in\mathbf{B}} f_{k,j}x^k e_j$ *(resp.* $f(x) = \sum_{(k,j)\in\widetilde{\mathbf{B}}} f_{k,j}x^k e_j$*). This means*

$$\left(\textit{resp. } f_{k,j}x^k e_j \neq 0 \implies \left(\left|\frac{\lambda_j}{\lambda^k}\right| > 1 \textit{ or } \frac{\lambda_j}{\lambda^k} = 1 \right) \right),$$

$$\left(\textit{resp. } f_{k,j}x^k e_j \neq 0 \implies \left(\left|\frac{\lambda_j}{\lambda^k}\right| < 1 \textit{ or } \frac{\lambda_j}{\lambda^k} = 1 \right) \right).$$

Then the inverse function admits a similar expression: $F^{-1}(x) = x + \sum_{(k,j)\in\mathbf{B}} g_{k,j}x^k e_j$ *(resp.* $F^{-1}(x) = x + \sum_{(k,j)\in\widetilde{\mathbf{B}}} g_{k,j}x^k e_j$*).*

Proof: We only prove this theorem for $\mathbf{B}$, the proof for $\widetilde{\mathbf{B}}$ is analogous. We use induction. Suppose that we have already verified that

$$g_{k,j}x^k e_j \neq 0 \text{ and } |k| \leq l-1 \implies \left(\left|\frac{\lambda_j}{\lambda^k}\right| > 1 \text{ or } \frac{\lambda_j}{\lambda^k} = 1 \right),$$

we show that the same statement for l holds. We use $F^{-1} \circ F(x) = x$. It follows from the induction hypothesis that

$$F^{-1}(x) = x + \sum_{\substack{(k,j)\in\mathbf{B} \\ |k|\leq l-1}} g_{k,j}x^k e_j + \sum_{|k|\geq l} g_{k,j}x^k e_j.$$

Take now an arbitrary $m \in N^n$ for which $|m| = l$. The projection $\pi_{m,j}$ on the terms with index (m,j) of $F^{-1}(F(x))$ is given by

$$\begin{aligned} \pi_{(m,j)}(F^{-1}(F(x))) &= \pi_{(m,j)} \left(\sum_{(k,j)\in\mathbf{B}} f_{k,j}x^k e_j \right) \circ \left(\sum_{(k,j)\in\mathbb{N}^n} g_{k,j}x^k \right) \\ &= \sum f_{k,j} g_{r^1_{1},1} x^{r^1_1} \ldots g_{r^1_{k_1},1} x^{r^1_{k_1}} \ldots g_{r^n_{1},n} x^{r^n_1} \ldots g_{r^n_{k_n},n} x^{r^n_{k_n}}. \end{aligned} \tag{5.33}$$

Here the last sum is expanded over all decompositions of $m = (m_1, \ldots, m_n)$:

$$r^1_1 + \ldots + r^1_{k_1} + \ldots + r^n_1 + \ldots + r^n_{k_n} = m.$$

A term in sum 5.33 is non-zero if all of the $g^{\beta}_{r^{\beta}_{\alpha}} \neq 0$ and $f_{k,j} \neq 0$. Consider now a term with index (m,j) and suppose that $(m,j) \notin \mathbf{B}$. There are two possibilities:

1. At least one of the $\left(r^{\beta}_{\alpha}, \beta\right) \notin \mathbf{B}$ implying $g_{r^{\beta}_{\alpha},\beta}$=0. The corresponding term vanishes.

2. All the $(r_\alpha^\beta, \beta) \in \mathbf{B}$; hence

$$\left|\frac{\lambda_\alpha}{\lambda^{r_\alpha^\beta}}\right| > 1 \text{ or } \frac{\lambda_\alpha}{\lambda^{r_\alpha^\beta}} = 1.$$

Since $(m, j) \notin \mathbf{B}$, we have $(\left|\frac{\lambda_j}{\lambda^m}\right| \geq 1$ and $\frac{\lambda_j}{\lambda^m} \neq 1)$. On the other hand we have

$$\frac{\lambda_j}{\lambda^m} = \frac{\lambda_j}{\lambda^{r_1^1} \ldots \lambda^{r_{k_1}^1} \ldots \lambda^{r_1^n} \ldots \lambda^{r_{k_n}^n}} = \frac{\lambda_1}{\lambda^{r_1^1}} \cdots \frac{\lambda_1}{\lambda^{r_{k_1}^1}} \cdots \frac{\lambda_n}{\lambda^{r_1^n}} \cdots \frac{\lambda_n}{\lambda^{r_{k_n}^n}} \frac{\lambda_j}{\lambda^k}. \quad (5.34)$$

If for all α, β we have that $\frac{\lambda_\alpha}{\lambda^{r_\alpha^\beta}} = 1$, then $\frac{\lambda_j}{\lambda^m} = \frac{\lambda_j}{\lambda^k}$, which is again a contradiction since $(k, j) \in \mathbf{B}$, and hence also $(m, j) \in \mathbf{B}$. If for one of the α, β we have that $\left|\frac{\lambda_\alpha}{\lambda^{r_\alpha^\beta}}\right| > 1$, then $\left|\frac{\lambda_j}{\lambda^m}\right| > \left|\frac{\lambda_j}{\lambda^k}\right| \geq 1$. Hence $(m, j) \in \mathbf{B}$, which is a contradiction.

It follows that $g_{m,j} = 0$ since every terms with this index vanishes. □

We have two corollaries:

Corollary 5.19 *Let $F(x) = A \circ (x + f(x))$ be a Gevrey-α diffeomorphism with diagonal linear part $A = \operatorname{diag}(\lambda_1, \ldots \lambda_n)$ and $f(x)$ containing terms of degree ≥ 2. Suppose that the eigenvalues of A satisfy a Siegel condition of order τ_1 and that the eigenvalues of A^{-1} satisfy a Siegel condition of order τ_2 i.e. there exist $C_1, C_2 > 0$ such that for all $k \in \mathbb{N}^n$, $j \in \{1, \ldots, n\}$,*

$$\left|\frac{\lambda_j}{\lambda^k} - 1\right| \geq \frac{C_1}{|k|^{\tau_1}}, \quad \left|\frac{\lambda^k}{\lambda_j} - 1\right| \geq \frac{C_2}{|k|^{\tau_2}},$$

then there exists a Gevrey-$(2 + \alpha + \tau_1 + \tau_2)$ coordinate transform $U(x)$ such that

$$U^{-1} \circ F \circ U = A \circ G(x) = A \circ (x + \sum_{(k,j) \in \mathbf{R}} g_{k,j} x^k e_j),$$

where

$$\mathbf{R} = \left\{(k, j) \in \mathbb{N}^n \times \{1, \ldots, j\} \mid \frac{\lambda^k}{\lambda_j} = 1\right\}.$$

The normal form $G(x)$ is also Gevrey-$(2 + \alpha + \tau_1 + \tau_2)$.

Proof: Let $\mu_j := \frac{1}{\lambda_j}$. Since $F^{-1}(x) = A^{-1} \circ \left(x + \widetilde{f}(x)\right) = A^{-1} \circ \left(\sum_{(k,j) \in \mathbb{N}} \widetilde{f}_{k,j} x^k e_j\right)$, where we have

$$\widetilde{\mathbf{B}} = \left\{(k, j) \in \mathbb{N}^n \times \{1, \ldots, j\} | \left|\frac{\mu_j}{\mu^k}\right| \geq 1\right\} = \left\{(k, j) \in \mathbb{N}^n \times \{1, \ldots, j\} | \left|\frac{\lambda_j}{\lambda^k}\right| \leq 1\right\}.$$

We apply Theorem 5.17 and find a Gevrey-$(1+\alpha+\tau_2)$ transformation U_1 to a normal form

$$\hat{G}(x) = A^{-1} \circ (x + g_1(x)) = A^{-1} \circ (x + \sum_{(k,j)} g^1_{k,j} x^k e_j)$$

that is Gevrey-$1+\alpha+\tau_2$. Hence, according to Proposition 5.18, it follows that

$$\hat{G}^{-1}(x) = \hat{F}(x) = A \circ (x + \sum_{(k,j)\in\widetilde{\mathbf{B}}} \hat{f}^1_{k,j} x^k e_j).$$

Moreover $\hat{F}$ is also Gevrey-$(1+\alpha+\tau_2)$. Hence we can apply Theorem 5.17 a second time to find a coordinate transform U_2 that is Gevrey-$(2+\alpha+\tau_1+\tau_2)$ to a normal form

$$G = A \circ (x + \sum_{(k,j)\in\widetilde{\mathbf{B}}\cap\mathbf{B}} \hat{f}^1_{k,j} x^k e_j)$$

in the same Gevrey-class. It suffices to remark that $\mathbf{R} = \widetilde{\mathbf{B}} \cap \mathbf{B}$.

□

Remark 5.20 *In fact one can always choose $\tau_1 = \tau_2 = \tau$. We explain that if the condition for $\tau_1 = \tau$ holds, then the same condition (with another value of C) for $\tau_2 = \tau$ is valid. Indeed, suppose that $\left|\frac{\lambda^k}{\lambda_j} - 1\right| \geq \frac{C}{|k|^\tau}$. We consider two cases.*

Case 1: $\left|\frac{\lambda_j}{\lambda^k}\right| < \frac{1}{2}$.

In this case clearly $\left|\frac{\lambda_j}{\lambda^k} - 1\right| \geq \frac{1}{2} \geq \frac{\widetilde{C}}{|k|^\tau}$ if $\widetilde{C}$ is small enough.

Case 2: $\left|\frac{\lambda_j}{\lambda^k}\right| \geq \frac{1}{2}$.

In this case

$$\left|\frac{\lambda_j}{\lambda^k} - 1\right| = \left|\frac{1 - \frac{\lambda^k}{\lambda_j}}{\frac{\lambda^k}{\lambda_j}}\right| \geq \frac{1}{2}\frac{C}{|k|^\tau}.$$

Hence, $\left|\frac{\lambda_j}{\lambda^k} - 1\right| \geq \frac{\min\{C/2, 2^{\tau-1}\}}{|k|^\tau}$. The proof for the other direction is analogous.

Corollary 5.21 *Define*

$$\mathbf{R} = \left\{(k,j) \in \mathbb{N}^n \times \{1,\ldots,j\} \mid \frac{\lambda^k}{\lambda_j} = 1\right\}.$$

Let $F(x) = A \circ (x + f(x)) = A \circ (x + \sum_{(k,j)\in\mathbf{R}} f_{k,j} x^k e_j)$ be a formal normal form. Then the inverse is again a normal form.

Proof: This follows immediately by applying Proposition 5.17 simultaneously for $\mathbf{B}$ and $\widetilde{\mathbf{B}}$ and remarking that $\mathbf{R} = \mathbf{B} \cap \widehat{\mathbf{B}}$. □

Example 5.22 *It is interesting to notice that if the diffeomorphism is attracting (resp. repelling), then its normal form is always a polynomial, since there are only a finite number of resonances. Because the inverse of a normal form is also a normal form, and the inverse is repelling (resp. attracting), we can conclude that it is also a polynomial. For example* $F(x,y,z) = (16\,x + y^2 + y^2z^2 + z^4, 4y + 2z^2, 2z)$ *has inverse*

$$F^{-1}(x,y,z) = \left(\frac{x}{16} - \frac{y^2}{256} + \frac{yz^2}{256} - \frac{y^2z^2}{1024} - \frac{5z^4}{1024} + \frac{yz^4}{1024} - \frac{z^6}{4096}, \frac{y}{4} - \frac{z^2}{8}, \frac{z}{2}\right).$$

5.2 A few numerical simulations

We compute numerically the smallest non-zero eigenvalue $\lambda_{\min,\delta}$ of the boxoperator $\Box_\delta^1 = d_{0,\delta}^1 d_{0,\delta}^{1^*}$ –we temporarily add an upper index 1 to make distinction with another operator that is also called $d_{0,\delta}$– associated to the quasi-homogeneous part S of degree 0, where $d_{0,\delta}(U) = d_{0,\delta}^1(U) = -(U \circ S - DS \circ U)$ is given by (5.4). We have

$$\sqrt{\lambda_{\min,\delta}} \geq \frac{C}{\delta^\tau} \Leftrightarrow -\frac{\ln(\sqrt{\lambda_{\min,\delta}})}{\ln(\delta)} + \frac{\ln(C)}{\ln(\delta)} \leq \tau.$$

Hence it is useful to make a plot of $-\frac{\ln(\sqrt{\lambda_{\min,\delta}})}{\ln(\delta)}$ as a function of δ. If this quantity remains bounded, then Theorem 5.8 can be applied with $\alpha = 0$. If this quantity is asymptotically smaller than τ, then corollary 5.19 can be applied with $\tau_1 = \tau_2 = \tau$. It should be noted that $d_{0,\delta}$ is defined differently with respect to corollary 5.19. Indeed, here $d_{0,\delta}(U) = d_{0,\delta}^2(U) = U - S \circ U \circ S^{-1}$, and should only be used if S is linear, diagonal and invertible. Let $\Box_\delta^2 = d_{0,\delta}^2 d_{0,\delta}^{2^*}$ be the box operator associated to such S. Note that the eigenvalues of the box-operator $\Box_\delta^1$ are given by $\left|\lambda^k - \lambda_j\right|$, while the eigenvalues of the box-operator $\Box_\delta^2$ are given by $\left|1 - \frac{\lambda_j}{\lambda^k}\right|$.

We explain why we can savely use $\Box_\delta^1$ to make conclusions concerning $\Box_\delta^2$ if S is linear and diagonal. Suppose that, for a fixed $k \in \mathbb{N}^n$, we have that

$$\left|\lambda^k - \lambda_j\right| \geq \frac{C}{k^\tau}.$$

We distinguish three cases:
Case 1: $\left|\lambda^k\right| \geq 2\left|\lambda_j\right|$
It follows that $\left|\frac{\lambda_j}{\lambda^k}\right| \leq \frac{1}{2}$. Hence,

$$\left|1 - \frac{\lambda_j}{\lambda^k}\right| \geq \frac{1}{2} \geq \frac{C}{|k|^\tau},$$

for values of $|k|$ that are large enough.
Case 2: $\left|\lambda^k\right| \leq \frac{|\lambda_j|}{2}$
It follows that $2 < \left|\frac{\lambda_j}{\lambda^k}\right|$. Hence

$$\left|1 - \frac{\lambda_j}{\lambda^k}\right| \geq 1 \geq \frac{C}{|k|^\tau},$$

for values of $|k|$ that are large enough.
Case 3: $\frac{|\lambda_j|}{2} \leq \left|\lambda^k\right| \leq 2\,|\lambda_j|$.
In this case it follows that

$$\left|1 - \frac{\lambda_j}{\lambda^k}\right| \geq \frac{C}{k^\tau\,|\lambda^k|} \geq \frac{C}{2\,|\lambda_j|\,k^\tau} \geq \frac{\widetilde{C}}{k^\tau},$$

where $\widetilde{C} = \frac{2C}{\lambda_{\max}}$, $\lambda_{\max} = \max\{|\lambda_j|\,|\,j \in \mathbb{N}\,;\, j \leq n\}$.

We conclude that in each of these cases,

$$\left|1 - \frac{\lambda_j}{\lambda^k}\right| \geq \frac{\widetilde{C}}{k^\tau},$$

for a certain $\widetilde{C} > 0$.

Conversely, one can check, using the same reasoning that if

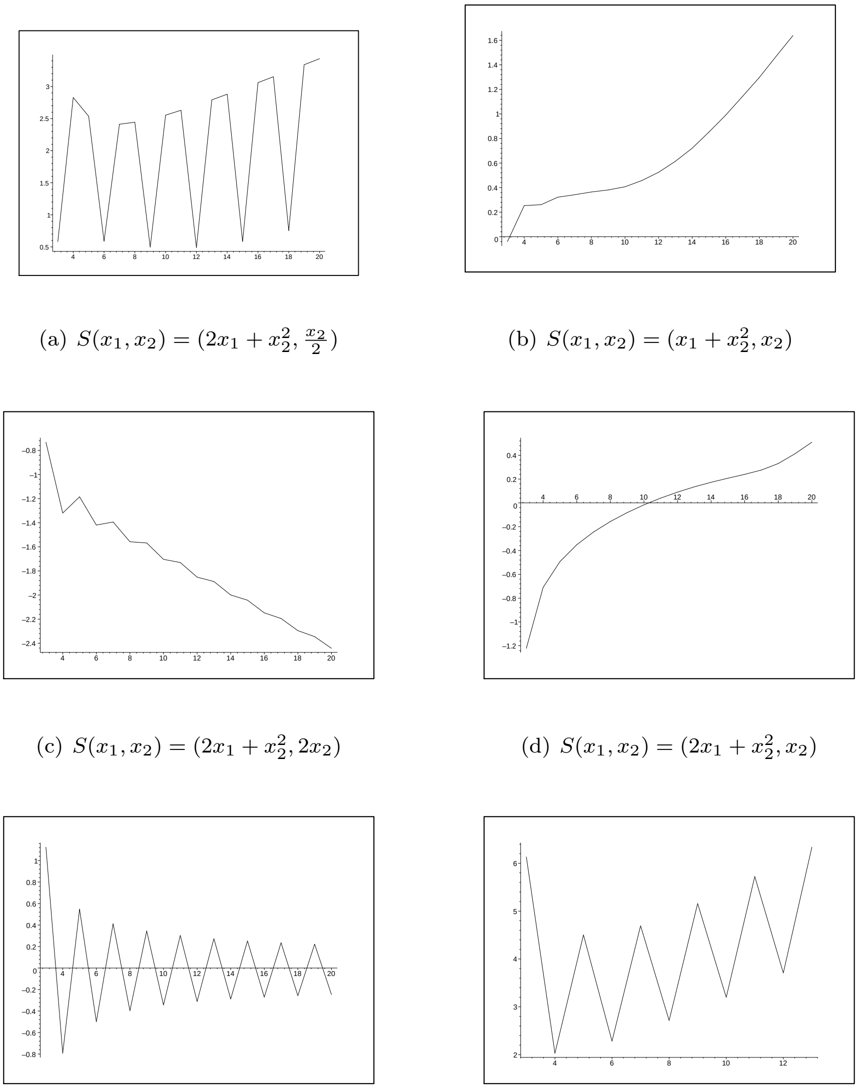
$$\left|1 - \frac{\lambda_j}{\lambda^k}\right| \geq \frac{C}{k^\tau},$$

it follows that

$$\left|\lambda^k - \lambda_j\right| \geq \frac{\widetilde{C}}{k^\tau},$$

for a certain $\widetilde{C} > 0$.

Hence, both conditions agree up to a constant for these systems. We have chosen to run the numerical algorithm using $\Box^1_\delta$. I discuss the results briefly. In Figures 5.2(a), 5.2(b), 5.2(d) the quantity $-\frac{\ln(\sqrt{\lambda_{\min,\delta}})}{\ln(\delta)}$ seems to be unbounded. In Figure 5.2(c) we can safely conjecture that $\tau = 0$, this is also expected since we are in a hyperbolic repelling situation. We have added one three-dimensional linear S, see Figure 5.2(f). This S is the main subject of Chapter 6.



(a) $S(x_1, x_2) = (2x_1 + x_2^2, \frac{x_2}{2})$

(b) $S(x_1, x_2) = (x_1 + x_2^2, x_2)$

(c) $S(x_1, x_2) = (2x_1 + x_2^2, 2x_2)$

(d) $S(x_1, x_2) = (2x_1 + x_2^2, x_2)$

(e) $S(x_1, x_2) = (x_1 + x_2^2, 2x_2)$

(f) $S(x_1, x_2, x_3) = (2(x_1 + x_2), 2x_2, \frac{x_3}{2})$

Figure 5.2: $-\frac{\ln(\sqrt{\lambda_{\min\,,\delta}})}{\ln(\delta)}$ as function of δ for the indicated S

Chapter 6

Non semi-simple saddles and divergence

6.1 Introduction and motivation

The main motivation to start this section is to give a better understanding of the normal form of analytic vector fields with a non-diagonalizable linear part. We will focus on the three dimensional situation. We start with a vector field $X = S + R$ with linear part $S = \lambda(x + ty)\frac{\partial}{\partial x} + \lambda y\frac{\partial}{\partial y} - \mu z\frac{\partial}{\partial z}$, where $\lambda, \mu > 0$ and $t \in \mathbb{R}$. We will identify S with the matrix

$$S = \begin{pmatrix} \lambda & t\lambda & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \mu \end{pmatrix}. \tag{6.1}$$

It is well known by a theorem of Siegel that if we assume that the eigenvalues satisfy the Siegel condition

$$|\lambda_1 k_1 + \lambda_2 k_2 - \mu k_3| \geq \frac{C}{|k_1 + k_2 + k_3|^\tau},$$

for each k_1, k_2, $k_3 \in \mathbb{Z}$ and a certain $\tau > 0$ and $C > 0$, that, if $t = 0$ (the semi-simple case) it follows that the vector field $X = S + R$ can be linearized to S by a unique coordinate transform $I + U$. We show that when $t \neq 0$ this result fails. In order to do so, we will study the Lie operator d_0. We use the same idea as the counterexample in [64]. This was already explained in [26]. Remark that the analogue for diffeomorphisms has been treated by [19]. All these results use a 'Divergence dichotomy'-type theorem that was first used by Il'yashenko in [30]. This results states roughly that if there exists one perturbation $X = S + R$ that has a linearizing series to normal form S that is divergent, then almost all perturbations of the form

$X = S + \alpha R$, $\alpha \in \mathbb{C}$ are divergent. As a corollary one can conclude that most solutions are divergent if the linearized problem $d_0(G) = R$ has a divergent solution. Here $d_0 : U \mapsto [U, S]$ is defined as before.

In this chapter we proof a 'Divergence dichotomy' theorem that holds in Gevrey-classes, showing that if there exists a divergent solution to the linearized equation $d_0(G) = R$ in a certain Gevrey class, then for almost all $\alpha \in \mathbb{C}$ the corresponding perturbation $X = S + \alpha R$ cannot be linearized in that Gevrey-class. We extend the result to show that the same holds for all Gevrey-classes at once. We proceed then with the study of the linearized equation $d_0(G) = R$ and obtain an explicit bound on the growth of the coefficients of G. We then explain how this bound can be used to give an upper bound on the coefficients of the normal form transformation. The bound we obtain is rather weak, and it would be interesting to see whether it would be beneficial to use the approach of [37] to compute the small divisors. So far we have not been able to find an estimate. However, when using the theory of quasi-homogeneous normal forms with the appropriate quasi-homogeneous weights, one can reinterpret the 'Nilpotent linear term' as a 'Resonant perturbation term'. Hence we can use the Gevrey-type theorems in this setting, that are known to be valid under the assumptions of the Siegel condition.

6.2 The capacity of sets

In this section we explain very briefly the notion of the capacity of a set, this section is inspired by [32] and [50] and contains no new results. The concept 'capacity of a set' will be needed only in order to use a generalization of the Bernstein inequality and has a long history. Historically the notion capacity comes from electrostatics, where the potential of a point mass due to a unit charge and placed in the point $x \in \mathbb{C}$ is given by $U(z) = -\ln(|z - x|)$. Remark that this function is harmonic, except at the point x. This harmonicity plays a key role in the study of potential theory for more advanced distributions.

The definition of a potential is extended to a charge distribution $d\mu$, Borel measure on a compact set K as

$$U(d\mu, K) = \int_K -\ln(|z - x|) d\mu(x).$$

This potential is again harmonic outside K. The corresponding electrostatic energy is defined as

$$E(d\mu, K) = \iint_{K \times K} -\ln(|z - x|) d\mu(x) d\mu(z).$$

If for some non-negative Borel measure the energy is bounded, then we define the infimum of the normalized energy as

$$E^*(K) = \inf\left\{ E(d\mu, K) | d\mu \text{ is a non-negative, normalized Borel measure: } \int_K d\mu = 1 \right\}.$$

If no such measure exists, we define $E^*(K) = 0$. In fact one can show that the infimum is actually a minimum more precisely one has

Theorem 6.1 *Let K be a compact set. If $E^*(K) > 0$, then there exists a unique non-negative Borel measure $d\mu_K$ on K for which $E^*(K) = E(d\mu_K, K)$.*

Proof: See e.g. [62]. It should be noted that the proof is not trivial. □

For compact sets K, the minimal electrostatic energy is used to define the capacity $\mathcal{C}(K) = \exp(-E^*(K))$. This quantity is either zero if $E^*(K) = \infty$ or equals $\mathcal{C}(K) = \exp(-E(d\mu_K, K))$. Analogous as in measure theory one defines for a general set B its inner capacity $\mathcal{C}_i$ and outer capacity $\mathcal{C}_o$:

$$\begin{aligned}\mathcal{C}_i(B) &:= \sup\left\{\mathcal{C}(K)|\, K \subset B \text{ and } K \text{ is compact}\right\},\\ \mathcal{C}_o(B) &:= \inf\left\{\mathcal{C}_i(O)|\, O \supset B \text{ and } O \text{ is open}\right\},\end{aligned}$$

and a set B is called capacitable if its inner capacity equals its outer capacity, for capacitable sets one defines the capacity as its inner or outer capacity. Using this definition it is clear that the inner capacity is always smaller than the outer capacity and that the open sets are capacitable. It can also be shown that compact sets are capacitable. We only need the following propositions.

Proposition 6.2 *For capacitable sets we have the following properties:*

(i) A countable union of sets with zero capacity has also zero capacity.

(ii) Any compact set with zero capacity has zero Lebesgue measure.

Proof: See e.g. [62]. □

Proposition 6.3 (Bernstein's inequality) *Suppose that $K \subset \mathbb{C}$ is a set of positive capacity, then for any polynomial $p \in \mathbb{C}[z]$ of degree $r \geq 0$,*

$$|p(z)| \leq ||p||_K \exp(rG_K(z)),$$

where $||p||_K = \max_{z\in K} |p(z)|$ is the supremum norm of p and $G_K(z) = \ln(\frac{1}{C(K)}) - d\mu_K(z))$.

Proof: See e.g. [33], [62]. □

Remark 6.4 *It can be shown that $G_K(z)$ is the non-negative Green's function of the complement $\mathbb{C} \setminus K$ with source at infinity.*

6.3 Divergence Dichotomy for Gevrey-classes

In this section we extend a result of [30] to more general Gevrey-classes. We follow the same ideas as in [33] in order to prove the following divergence dichotomy.

Theorem 6.5 (Divergence dichotomy in a specific Gevrey class) *Let $\eta \geq 0$. Consider a vector field $X_\alpha = S + \alpha R$ that is Gevrey-η, where S is a linear part that contains only non-resonant eigenvalues (we allow equal eigenvalues), R is a part of order ≥ 2 and α a complex parameter. We have the following dichotomy:*

1. *The linearizing series $I + U_\alpha$ is of class Gevrey-η for all values of $\alpha \in \mathbb{C}$.*

2. *The linearizing series $I + U_\alpha$ is not of class Gevrey-η for all values of α, except for a set $K \subset \mathbb{C}$ of capacity zero.*

Proof: Assume that $X_\alpha = S + \alpha R$ is a vector field of class Gevrey-η and assume that the formal series $I + U_\alpha$ linearizing X is Gevrey-η for each α in a certain set $K^* \subset \mathbb{C}$ of positive capacity. Consider now the subsets

$$K_{c,\rho} = \left\{\alpha \in \mathbb{C} | \; |\pi_m(U_\alpha)| \leq c(m!)^\eta \rho^{-m}, \, \forall m \in \mathbb{N}\right\},$$

where π_m is the projection on the terms of degree m. Then it is clear that $K^* = \cup_{c>0,\rho>0} K_{c,\rho}$ since each Gevrey-η series admits a bound of the form $c(m!)^\eta \rho^{-m}$. Moreover it is easy to see that the union can be replaced by a countable union, $K^* = \cup_n K_{c_n,\rho_n}$ for a certain well-chosen sequence (c_n, ρ_n). (Use the natural nesting of the sets $K_{c,\rho}$.) Because the countable union of sets of capacity zero is still a set of capacity zero, at least one of the sets $K := K_{c_{n_0},\rho_{n_0}}$ has positive capacity. For this set we have that

$$|\pi_m(U_\alpha)| \leq c_{n_0}(m!)^\eta \rho_{n_0}^{-m}, \; \forall \alpha \in K, \, \forall m \in \mathbb{N}.$$

Now, since K is a set of positive capacity and $\pi_m(U_\alpha)$ is a polynomial in α of degree at most $m - 1$, we can use Bernstein's inequality and conclude that

$$\begin{aligned} |\pi_m(U_\alpha)| &\leq c_{n_0}(m!)^\eta \rho_{n_0}^{-m} \exp[(m-1)G_K(\alpha)] \\ &\leq c_{n_0}(m!)^\eta \left(\frac{\rho_{n_0}}{\exp(G_K(\alpha))}\right)^{-m}, \; \forall \alpha \in \mathbb{C}, \, \forall m \in \mathbb{N}. \end{aligned}$$

This means that the corresponding series U_α is Gevrey-η for each $\alpha \in \mathbb{C}$. □

Corollary 6.6 *Let $X = S + \alpha R$ be a vector field of class Gevrey-η, S the non-resonant linear part and R the part of order ≥ 2 and let G be the formal solution of equation $d_0(G) = R$. Furthermore suppose that this solution G is not in the class Gevrey-η. Then the series linearizing the vector field $X = S + \alpha R$, is not in the class Gevrey-η for most values of the parameter α; the set of parameters α for which this vector field X is Gevrey-η has zero capacity.*

Proof: We give a proof by contradiction. Therefore suppose that the series $U_\alpha = I + u_\alpha$ linearizing X is Gevrey-η for all values of α in a set K of positive capacity. Then according to Theorem 6.5, it it is Gevrey-η for any value of $\alpha \in \mathbb{C}$. Now we differentiate the linearization equation

$$DU_\alpha(S + \alpha R) = SU_\alpha$$

to α and observe that $G = \frac{\partial u_\alpha}{\partial \alpha}\big|_{\alpha=0}$ is a Gevrey-η solution of the equation

$$DG.S - S.G = R,$$

a contradiction. □

Corollary 6.7 *Assume that* $S = \lambda(x+ty)\frac{\partial}{\partial x} + \lambda y \frac{\partial}{\partial y} - \mu z \frac{\partial}{\partial z}$, *where* $\lambda > 0$, $\mu > 0$; *in such a way that* λ, μ *are non resonant. I.e.* $\frac{\lambda}{\mu} \notin \mathbb{Q}$. *Then, for each* $\eta \geq 0$, *there exists a Gevrey-*η *vector field* $X = S + R$ *that is formally linearizable (this is not surprising since there are no resonances), but not Gevrey-*η *linearizable to* S.

Proof: The idea is do construct a divergent solution of $d_0(G) = R$ and to apply Corollary 6.6. It will be sufficient to study this operator on subspaces. Therefore we start with the following computations :

$$\begin{aligned}[S, x^{k_1} y^{k_2} z^{k_3} \frac{\partial}{\partial x}] =& (\lambda k_1 + \lambda k_2 - \mu k_3 - \lambda) x^{k_1} y^{k_2} z^{k_3} \frac{\partial}{\partial x} \\ &+ \lambda k_1 t x^{k_1 - 1} y^{k_2+1} z^{k_3} \frac{\partial}{\partial x}.\end{aligned}$$

We fix two integers r, s and consider the vector space $B_{r,s}$ with basis $x^\alpha y^{r-\alpha} z^s \frac{\partial}{\partial x}$ for $0 \leq \alpha \leq r$. This space is clearly invariant under the operator d_0. Moreover, using the above computating, it follows that the operator $d_0|_{B_{r,s}}$ in matrix notation is given by:

$$M_{r,s} = \begin{pmatrix} \beta_{r,s} & \lambda t & 0 & 0 & \dots & 0 \\ 0 & \beta_{r,s} & 2\lambda t & 0 & \dots & 0 \\ 0 & 0 & \beta_{r,s} & 3\lambda t & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \beta_{r,s} & r\lambda t \\ 0 & 0 & 0 & 0 & \dots & \beta_{r,s} \end{pmatrix}, \text{ with } \beta_{r,s} = \lambda r - \mu s - \lambda. \tag{6.2}$$

We calculate its inverse $M_{r,s}^{-1}$. Remark that $M_{r,s} = \beta_{r,s}(I + N_{r,s})$ where $N_{r,s}$ is the nilpotent matrix

$$N_{r,s} = \begin{pmatrix} 0 & \gamma_{r,s} & 0 & 0 & \dots & 0 \\ 0 & 0 & 2\gamma_{r,s} & 0 & \dots & 0 \\ 0 & 0 & 0 & 3\gamma_{r,s} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & r\gamma_{r,s} \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \text{ where } \gamma_{r,s} = \frac{\lambda t}{\beta_{r,s}}.$$

It follows that the inverse of $M_{r,s}$ can be computed as

$$M_{r,s}^{-1} = \frac{1}{\beta_{r,s}} \sum_{k=0}^{r} (-1)^k N_{r,s}^k,$$

because $N_{r,s}^{r+1} = 0$. Hence

$$M_{r,s}^{-1} = \frac{1}{\beta_{r,s}} \begin{pmatrix} 1 & -\gamma_{r,s} & 2!\gamma_{r,s}^2 & -3!\gamma_{r,s}^3 & \ldots & (-1)^r r!\gamma_{r,s}^r \\ 0 & 1 & * & * & \ldots & * \\ 0 & 0 & 1 & * & \ldots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & * \\ 0 & 0 & 0 & 0 & \ldots & 1 \end{pmatrix},$$

and

$$\begin{aligned} &M_{r,s}^{-1}(((r+s)!)^\eta, -((r+s)!)^\eta, \ldots, (-1)^r((r+s)!)^\eta))^T \\ &\qquad = (\sum_{k=0}^{r} ((r+s)!)^\eta \gamma_{r,s}^k k!, m_1, \ldots, m_r)^T. \end{aligned}$$

It follows that

$$\begin{aligned} &d_0^{-1}(\sum_{\alpha=0}^{r} (-1)^\alpha ((r+s)!)^\eta x^\alpha y^{r-\alpha} z^s \frac{\partial}{\partial x}) = \\ &\qquad \left(\sum_{l=0}^{r} ((r+s)!)^\eta \gamma_{r,s}^l l! \right) y^r z^s \frac{\partial}{\partial x} + \sum_{\alpha=1}^{r} m_{\alpha,r,s} x^\alpha y^{r-\alpha} z^s \frac{\partial}{\partial x}. \end{aligned}$$

Let us choose two increasing series of integers $(r_k)_{k\in\mathbb{N}}$,$(s_k)_{k\in\mathbb{N}}$ that have the property $\lim_{k\to\infty} \beta_{r_k,s_k} = 0$. It is then not difficult to see that $\frac{r_k}{r_k+s_k} \geq C_0 > 0$ for a certain constant C_0 that depends on λ and μ (this follows from the fact that $\lim_{k\to\infty} \frac{r_k}{s_k} = \frac{\lambda}{\mu}$). Let

$$R = \sum_{k=0}^{\infty} \left(\sum_{\alpha=0}^{r_k} ((r_k+s_k)!)^\eta (-1)^\alpha x^\alpha y^{r_k-\alpha} z^{s_k} \frac{\partial}{\partial x} \right). \tag{6.3}$$

Hence

$$\begin{aligned} G &= d_0^{-1}(R) \\ &= \sum_{k=0}^{\infty} \left(\left(((r_k+s_k)!)^\eta \sum_{l=0}^{r_k} \gamma_{r_k,s_k}^l l! \right) y^{r_k} z^{s_k} \frac{\partial}{\partial x} + \sum_{\alpha=1}^{r_k} m_{\alpha,r_k,s_k} x^\alpha y^{r_k-\alpha} z^s \frac{\partial}{\partial x} \right). \end{aligned} \tag{6.4}$$

Now, for k large enough, we have that $\gamma_{r_k,s_k} = \frac{\lambda t}{\beta_{r_k,s_k}} \geq 1$ because β_{r_k,s_k} tends to zero. Thus $\sum_{l=0}^{r_k} \gamma_{r_k,s_k}^l l! \geq r_k! \geq \Gamma(C_0(r_k+s_k))$ and it follows that the series (6.4) is not Gevrey-η. It is now sufficient to apply corollary 6.6 to finish the proof using R defined in (6.3). Indeed, $d_0(G) = R$ and G has clearly a zero radius of convergence. □

We give another flavour of the same type of theorem:

Theorem 6.8 (Divergence dichotomy for all Gevrey classes) *Consider a vector field $X_\alpha = S + \alpha R$, where S is a linear part that contains only non-resonant eigenvalues (we allow equal eigenvalues), R is a part of order ≥ 2 and α a complex parameter. Suppose also that for each value of α the vector field X is Gevrey of a certain order η. We have the following dichotomy:*

1. *There exists an $l \in \mathbb{N}$ such that for each value of $\alpha \in \mathbb{C}$ we have that the linearizing series $I + U_\alpha$ is of class Gevrey-l.*
2. *The linearizing series $I + U_\alpha$ is not of class Gevrey-l for all $l \in \mathbb{N}$ and for all values of α, except for a set $K_f \subset \mathbb{C}$ of capacity zero.*

Proof: Assume that $X_\alpha = S + \alpha R$ is a vector field of class Gevrey-η for each value of α. Assume that for each α in a certain set $K^* \subset \mathbb{C}$ of positive capacity the formal series $I+U_\alpha$ linearizing X is Gevrey-l (the value of l may well depend on α). Consider now the subsets

$$K_{c,\rho,l} = \left\{\alpha \in \mathbb{C} | \, |\pi_m(U_\alpha)| \leq c(m!)^l \rho^{-m}, \, \forall m \in \mathbb{N}\right\},$$

where π_m is the projection on the terms of degree m. Then it is clear that $K^* = \cup_{c,\rho,l} K_{c,\rho,l}$ since each Gevrey-l series admits a bound of the form $c(m!)^l \rho^{-m}$. Moreover it is easy to see that the union can be replaced by a countable union, $K^* = \cup_n K_{c_n,\rho_n,l_n}$ for a certain well-chosen sequence (c_n, ρ_n, l_n). (Use the natural nesting of the sets $K_{c,\rho,l}$.) Now, because the countable union of sets of capacity zero is still a set of capacity zero, at least one of the sets in this countable union, say $K := K_{c_{n_0},\rho_{n_0},l_{n_0}}$, has positive capacity. For this set we obtain

$$|\pi_m(U_\alpha)| \leq c_{n_0}(m!)^{l_{n_0}} \rho_{n_0}^{-m}, \ \forall \alpha \in K, \forall m \in \mathbb{N}.$$

Now, since K is a set of positive capacity and $\pi_m(U_\alpha)$ is a polynomial in α of degree at most $m-1$, we can use the Bernstein inequality and conclude that

$$\begin{aligned} |\pi_m(U_\alpha)| &\leq c_{n_0}(m!)^{l_{n_0}} \rho_{n_0}^{-m} \exp[(m-1)G_K(\alpha)] \\ &\leq c_{n_0}(m!)^{l_{n_0}} \left(\frac{\rho_{n_0}}{\exp(G_K(\alpha))}\right)^{-m}, \ \forall \alpha \in \mathbb{C}, \forall m \in \mathbb{N}. \end{aligned}$$

This means that the corresponding series U_α is Gevrey-l_{n_0} for each $\alpha \in \mathbb{C}$. □

Remark 6.9 *The existence of non-Brjuno numbers can be used to give examples of non-Gevrey normalizable analytic vector fields.*

6.4 An upper bound on the growth of the coefficients

Suppose that we start with an analytic vector field $X = S + R$ with linear part $S = \lambda(x+ty)\frac{\partial}{\partial x} + \lambda y\frac{\partial}{\partial y} - \mu z\frac{\partial}{\partial z}$, where $\lambda, \mu > 0$. We have shown that the normal form transformation in this case can be divergent, however we have not yet gained control of the degree of divergence. In order to do so, we will study the Lie operator d_0 more closely.

Remark 6.10 *In order to bound the degree of divergence one has to ask at least some bound on the small denominators. Common conditions that are encountered are the Siegel condition, Rüssman condition and Brjuno condition. These conditions are natural: in case* $t = 0$ *it is known that the corresponding normal form transformation is convergent. Moreover the eigenvalues that satisfy such a condition have full Lebesgue measure. We will in the first place concentrate on the Siegel condition. Mainly because in the resonant semi-simple case it is known to produce normal forms of Gevrey type. See e.g. [37].*

As before $\mathcal{V}_\delta$ is the space of homogeneous vector fields of regular degree δ and

$$d_{0,\delta} : \mathcal{V}_\delta \longrightarrow \mathcal{V}_\delta : U \mapsto [S, U].$$

In this section we start with a more thorough study of the operator $d_{0,\delta}^{-1}$. We will assume that t is small, $|t| \leq \frac{1}{2|\lambda|}$, which is not a restriction, since the linear part can always brought in such a form by a linear coordinate transformation. A bound on the operator norm of $d_{0,\delta}^{-1}$ will result in an upper bound of the coefficients of the normal form transformation as we will explain later on. We will calculate a bound of the operator norm of $d_{0,\delta}^{-1}$ by studying invariant subspaces V for this operator. For such an invariant subspace we will sometimes abuse the notation $d_{0,\delta}|_V$ for the operator d_0 as a map from V to V (instead of from V to the whole space), so that it makes sense to consider $(d_{0,\delta}|_V)^{-1}$. We will use the following calculation to start our digression:

$$\begin{aligned}
[S, x^{k_1}y^{k_2}z^{k_3}\frac{\partial}{\partial x}] &= (\lambda k_1 + \lambda k_2 - \mu k_3 - \lambda)x^{k_1}y^{k_2}z^{k_3}\frac{\partial}{\partial x} \\
&\quad + \lambda k_1 t x^{k_1-1}y^{k_2+1}z^{k_3}\frac{\partial}{\partial x}. \\
[S, x^{k_1}y^{k_2}z^{k_3}\frac{\partial}{\partial y}] &= (\lambda k_1 + \lambda k_2 - \mu k_3 - \lambda)x^{k_1}y^{k_2}z^{k_3}\frac{\partial}{\partial y} \\
&\quad + \lambda k_1 t x^{k_1-1}y^{k_2+1}z^{k_3}\frac{\partial}{\partial y} - \lambda x^{k_1}y^{k_2}z^{k_3}\frac{\partial}{\partial x}, \\
[S, x^{k_1}y^{k_2}z^{k_3}\frac{\partial}{\partial z}] &= (\lambda k_1 + \lambda k_2 - \mu k_3 - \mu)x^{k_1}y^{k_2}z^{k_3}\frac{\partial}{\partial z} \\
&\quad + \lambda k_1 t x^{k_1-1}y^{k_2+1}z^{k_3}\frac{\partial}{\partial z}.
\end{aligned}$$

Hence it follows that the spaces

$$\begin{aligned} V_{r,s} &= \operatorname{span}\left\{ x^{\alpha}y^{r-\alpha}z^{s}\frac{\partial}{\partial x}, x^{\alpha}y^{r-\alpha}z^{s}\frac{\partial}{\partial y} \,|\, 0 \leq \alpha \leq r \right\} \\ W_{r,s} &= \operatorname{span}\left\{ x^{\alpha}y^{r-\alpha}z^{s}\frac{\partial}{\partial z} \,|\, 0 \leq \alpha \leq r \right\} \end{aligned}$$

are invariant spaces for each $r, s \in \mathbb{N}$. The corresponding matrix of $d_{0,\delta}|_{V_{r,s}}$ is

$$N_{r,s} = \begin{pmatrix} M_{r,s} & -\lambda I \\ 0 & M_{r,s} \end{pmatrix},$$

where $M_{r,s}$ is the same matrix as in equation (6.2) and I is the identity matrix. We calculate the inverse and find that

$$N_{r,s}^{-1} = \begin{pmatrix} M_{r,s}^{-1} & \lambda(M_{r,s}^{-1})^{2} \\ 0 & M_{r,s}^{-1} \end{pmatrix}.$$

Moreover, one can verify the formula

$$M_{r,s}^{-1} = \frac{1}{\beta_{r,s}} \begin{pmatrix} 1 & -\frac{1!\gamma_{r,s}}{0!} & \frac{2!\gamma_{r,s}^{2}}{0!} & -\frac{3!\gamma_{r,s}^{3}}{0!} & \dots & \frac{(-1)^{r}r!\gamma_{r,s}^{r}}{0!} \\ 0 & 1 & -\frac{2!\gamma_{r,s}}{1!} & \frac{3!\gamma_{r,s}^{2}}{1!} & \dots & \frac{(-1)^{r-1}r!\gamma_{r,s}^{r-1}}{1!} \\ 0 & 0 & 1 & -\frac{3!\gamma_{r,s}}{2!} & \dots & \frac{(-1)^{r-2}r!\gamma_{r,s}^{r-2}}{2!} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 1 & -\frac{r!\gamma_{r,s}}{(r-1)!} \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

We calculate an estimate on the coefficients of $L_{r,s} = -\lambda(M_{r,s}^{-1})^{2}$. It is clear that it is upper triangular and that the diagonal is $-\lambda\frac{1}{\beta_{r,s}^{2}}$. For the other entries we take an arbitrary row R_k and column C_l of $M_{r,s}^{-1}$ and multiply them. Here, $1 \leq k < l \leq r+1$.

$$\begin{aligned} R_k &= \frac{1}{\beta_{r,s}}\left(0, \dots, 0, 1, -\frac{k!\gamma_{r,s}}{(k-1)!}, \dots, \frac{(-1)^{r-2}r!\gamma_{r,s}^{r-k}}{(k-1)!}\right) \\ C_l &= \frac{1}{\beta_{r,s}}\left(\frac{(-1)^{(l-1)}(l-1)!\gamma_{r,s}^{(l-1)}}{0!}, \frac{(-1)^{l-2}l!\gamma_{r,s}^{l-2}}{1!}, \dots, -\frac{(l-1)!\gamma_{r,s}}{(l-2)!}, 1, 0, \dots, 0\right)^{T} \end{aligned}$$

We will use [.] to indicate the position in a matrix/vector in order to avoid confusion with the subindices r, s already present. Hence

$$\begin{aligned}
R_k[\alpha] &= 0, \text{ for } 1 \leq \alpha \leq k-1 \\
R_k[\alpha] &= \frac{1}{\beta_{r,s}} \frac{(\alpha-1)!(-1)^{\alpha-k}\gamma_{r,s}^{\alpha-k}}{(k-1)!}, \text{ for } k \leq \alpha \leq r+1 \\
C_l[\alpha] &= \frac{1}{\beta_{r,s}} \frac{(-1)^{l-\alpha}(l-\alpha)!\gamma_{r,s}^{l-\alpha}}{(\alpha-1)!}, \text{ for } 1 \leq \alpha \leq l \\
C_l[\alpha] &= 0, \text{ for } l+1 \leq \alpha \leq r+1.
\end{aligned}$$

Suppose that $k < l$; then

$$\begin{aligned}
R_k.C_l &= \sum_{\alpha=0}^{r+1} R_k[\alpha]C_l[\alpha] = \sum_{\alpha=k}^{l} R_k[\alpha]C_l[\alpha] \\
&= \frac{1}{\beta_{r,s}^2} \sum_{\alpha=k}^{l} \left(\frac{(\alpha-1)!(-1)^{\alpha-k}\gamma_{r,s}^{\alpha-k}}{(k-1)!} \right) \left(\frac{(-1)^{l-\alpha}(l-\alpha)!\gamma_{r,s}^{l-\alpha}}{(\alpha-1)!} \right) \\
&= \frac{1}{\beta_{r,s}^2} \frac{\gamma_{r,s}^{l-k}}{(k-1)!} \sum_{\alpha=k}^{l} \left((-1)^{l-k}(l-\alpha)! \right).
\end{aligned}$$

Suppose that $r+s=\delta$. We define $\beta_\delta := \min\{|\beta_{r,s}| \, | \, r+s=\delta\}$ and remember that $|t| \leq \frac{1}{2|\lambda|}$. Then

$$\begin{aligned}
|L_{r,s}[k,l]| = |R_k.C_l| &\leq \frac{\lambda}{|\beta_{r,s}|^2} \frac{|\gamma_{r,s}|^{l-k} r!(r+1)}{(k-1)!} \\
&\leq \frac{|\lambda|}{|\beta_{r,s}|^2} |\gamma_{r,s}^{l-k}|(r+1)! \\
&\leq \frac{|\lambda|}{|\beta_{r,s}|^2} (r+1)! \left(\max\{|\gamma_{r,s}|^p \, | \, 0 \leq p \leq r\} \right) \\
&\leq \frac{|\lambda|}{|\beta_{r,s}|^2} (r+1)! \left(\max\left\{ \frac{(|\lambda|t)^p}{|\beta_{r,s}|^p} \, | \, 0 \leq p \leq r \right\} \right) \\
&\leq \frac{|\lambda|}{|\beta_{r,s}|^2} (r+1)! \left(\max\left\{ \frac{1}{(2|\beta_{r,s}|)^p} \, | \, 0 \leq p \leq r \right\} \right), \text{ for } k \leq l
\end{aligned}$$

We distinguish two cases to make a further estimate.
Case 1: $|\beta_{r,s}| = |(r-1)\lambda - s\mu| \geq 1$

In this case $\frac{1}{|\beta_{r,s}|} \le 1$ and

$$\begin{aligned}
|L_{r,s}[k,l]| &\le \frac{|\lambda|}{|\beta_{r,s}|^2}(r+1)!\left(\max\left\{\frac{1}{(2|\beta_{r,s}|)^p} \,|\, 0 \le p \le r\right\}\right) \\
&\le \frac{|\lambda|(r+1)!}{|\beta_{r,s}|^2} \le |\lambda|(r+1)! \\
&\le |\lambda|(\delta+1)!\frac{\beta_\delta^\delta}{\beta_\delta^\delta}
\end{aligned}$$

Because $\lim_{\delta\to\infty} \beta_\delta = 0$ it follows that $|\beta_\delta| \le \frac{1}{2}$ if $\delta > \delta_0$. And because

$$0 \le 2|\lambda|(\delta+1)^2\beta_\delta^\delta \le 2|\lambda|(\delta+1)^2\frac{1}{2^\delta},$$

also $\lim_{\delta\to\infty} 2|\lambda|(\delta+1)^2\beta_\delta^\delta = 0$. Hence it follows that $2|\lambda|(\delta+1)^2\beta_\delta^\delta \le 1$ if $\delta > \delta_1$ and

$$|L_{r,s}[k,l]| \le \frac{(\delta)!}{2(\delta+1)\beta_\delta^\delta}, \text{ if } \delta > \max\{\delta_0, \delta_1\}.$$

Case 2: $|\beta_{r,s}| = |(r-1)\lambda - s\mu| < 1$
Remark first that if $r = 0$, then $-\lambda < s\mu + 1$ and $\delta = s < \frac{-\lambda+1}{\mu}$. Hence if we suppose that $\delta \ge \delta_2 = \frac{-\lambda+1}{\mu}$, then $r \ge 1$. Analogous we can argue that if $\delta > \delta_3$, then $s \ge 2$ and hence $r+2 \le r+s \le \delta$. Now $|(r-1)\lambda - s\mu| < 1$ which implies that $s\mu - 1 < (r-1)\lambda$ equivalent with $s < \frac{r\lambda-\lambda+1}{\mu}$. Now if $\delta \ge \delta_2$, and hence $r \ge 1$, it follows that $s < \frac{r\lambda+r}{\mu}$. Consequently $\delta = r + s < (1 + \frac{\lambda+1}{\mu})r$ which implies $r \ge \dfrac{\delta}{1+\frac{\lambda+1}{\mu}} =: \kappa\delta$; and it follows $\frac{1}{2^r} \le \frac{1}{2^{\kappa\delta}}$. Thus, in this case, if we suppose that $\delta > \delta_3$ which implies $r + 2 \le \delta$,

$$\begin{aligned}
|L_{r,s}[k,l]| &\le \frac{|\lambda|}{|\beta_{r,s}|^2}(r+1)!(\max\{\frac{1}{(2|\beta_{r,s}|)^p} \,|\, 0 \le p \le r\}) \\
&\le \frac{|\lambda|(r+1)!}{2^r\beta_\delta^{r+2}} \\
&\le \frac{|\lambda|(\delta+1)!}{2^{\kappa\delta}\beta_\delta^\delta}
\end{aligned}$$

Now $\lim_{\delta\to\infty} \frac{|\lambda|(\delta+1)^2}{2^{\kappa\delta}} = 0$ and hence $|2\frac{|\lambda|(\delta+1)^2}{2^{\kappa\delta}}| < 1$ for $\delta > \delta_4$. It follows that

$$|L_{r,s}[k,l]| \le \frac{\delta!}{2(\delta+1)\beta_\delta^\delta}, \text{ if } \delta > \max\{\delta_2, \delta_3, \delta_4\}.$$

Hence we can conclude that in both cases

$$|L_{r,s}[k,l]| \le \frac{\delta!}{2(\delta+1)\beta_\delta^\delta}, \text{ if } \delta > \delta_5, \text{ where } \delta_5 = \max\{\delta_0, \delta_1, \delta_2, \delta_3, \delta_4\}. \tag{6.5}$$

It is clear that

$$|M_{r,s}[k,l]| \leq \frac{r!}{|\beta_{r,s}|}(\max\{|\gamma_{r,s}|^p \,|\, 0 \leq p \leq r\}),$$

and using the same argumentation as above one can prove that there exists a $\delta_6 > 0$ such that

$$|M_{r,s}[k,l]| \leq \frac{\delta!}{2(\delta+1)\beta_\delta^\delta}, \text{ if } \delta > \delta_6 \tag{6.6}$$

From inequality (6.5) and (6.6), we can conclude that $N_{r,s}^{-1}$ has only entries that are bounded by

$$\frac{\delta!}{2(\delta+1)\beta_\delta^\delta}, \text{ for } \delta > \delta_7 := \max\{\delta_5, \delta_6\}.$$

Now suppose that $v = (a_1, a_2, \ldots, a_{r+1}, b_1, b_2, \ldots, b_{r+1})$ is a vector with $|v| \leq 1$. Then, if $\delta > \delta_7$,

$$\begin{aligned} |N_{r,s}^{-1}.v| &\leq \frac{2(r+1)\delta!}{2(\delta+1)\beta_\delta^\delta}|v| \leq \frac{\delta!}{\beta_\delta^\delta}|v|, \text{ and} \\ ||N_{r,s}^{-1}|| &\leq \frac{\delta!}{\beta_\delta^\delta}, \end{aligned}$$

and we can conclude that $||d_0|_{V_{r,s}}^{-1}|| \leq \frac{\delta!}{\beta_\delta^\delta}$.

Define $\beta'_{r,s} := |\lambda r - \mu s - \mu|$ and $\beta'_\delta = \max\{\beta'_{r,s} \,|\, r+s = \delta\}$. Completely analogous, but somewhat easier, one can argue that if $\delta > \delta_8$ then $||d_0|_{W_{r,s}}^{-1}|| \leq \frac{\delta!}{\beta'^\delta_\delta}$. Let $\gamma_\delta := \max\{\beta'_\delta, \beta_\delta\}$. Then it follows that

$$||d_{0,\delta}^{-1}|| \leq \frac{\delta!}{\gamma_\delta^\delta}, \text{ for } \delta > \delta_9 := \max\{\delta_7, \delta_8\}.$$

We define the growth constants η_δ as follows.

$$\begin{aligned} \eta_0 &= 1 \\ \eta_\delta &= \frac{\delta!}{\gamma_\delta^\delta}\left(\max_{0 \leq \mu \leq \delta,\, \delta_1 + \ldots + \delta_r + \mu = \delta} \eta_{\delta_1} \cdots \eta_{\delta_r}\right) \end{aligned}$$

Theorem 6.11 *Suppose that $X = S + R$ is an analytic vector field with linear part $S = \lambda(x+ty)\frac{\partial}{\partial x} + \lambda y \frac{\partial}{\partial y} - \mu z \frac{\partial}{\partial z}$ where $\lambda, \mu > 0$, $\frac{\lambda}{\mu} \notin \mathbb{Q}$ and R of order $\geq \delta_9$. Then there exists a $C > 0$ and a $\rho > 0$ such that the formal coordinate transform $\Phi^{-1} = I + U$ to normal form, satisfies $||\pi_\delta(U)||_\delta \leq C\rho^\delta \eta_\delta$.*

Remark 6.12 *The fact that we choose R of order $\geq \delta_9$ is not a severe restriction, since we can always bring a vector field with non-resonant linear part X to this form by considering a polynomial transformation.*

Proof: The structure of the proof is borrowed inspired by the work of [37]. We know that Φ transforms X into the linear vector field S. Hence

$$\begin{aligned}
&\Phi_*(X) = S\\
\Leftrightarrow &X \circ \Phi^{-1} = D\Phi^{-1}.X'\\
\Leftrightarrow &(S+R) \circ (I+U) = D(I+U).(S)\\
\Leftrightarrow &[S,U] = R(I+U).
\end{aligned}$$

We will now define the formal transformation $U = \sum_{\delta=\delta_9}^{+\infty} U_\delta$, where U_δ is a homogeneous polynomial of degree δ inductively. We define $U_0 = I$, $U_1 = 0, \dots, U_{\delta_9 - 1} = 0$. Now suppose that we have already defined U_l for all $l \leq \delta - 1$. We determine now U_δ. In order to do so we will solve $[S, U_\delta] = \pi_\delta([S,U]) = \pi_\delta(R(I+U))$. Now define $V_\delta := \pi_\delta(R(I+U))$, $U_\delta = d_0^{-1}(V_\delta)$ and remark that it does only depend on $U_{\delta'}$ for $\delta' < \delta$:

$$\begin{aligned}
\pi_\delta(R(I+U)) &= \sum_{k=2}^{\delta} R_k(I+U, \dots, I+U)\\
&= \sum_{k=2}^{\delta} \sum_{\delta_1+\ldots+\delta_r=\delta} R_k(U_{\delta_1}, \dots, U_{\delta_r}).
\end{aligned}$$

Moreover since

$$|R_k(U_{\delta_1}, \dots, U_{\delta_r})| \leq \frac{M}{\rho^k} |U_{\delta_1}| \dots |U_{\delta_r}|,$$

following from the analyticity of R, it follows that

$$|\pi_\delta(R(I+U))| \leq \sum_{k=2}^{\delta} \sum_{\delta_1+\ldots+\delta_r=\delta} \frac{M}{\rho^k} |U_{\delta_1}| \dots |U_{\delta_r}|.$$

Now we introduce the following constants

$$\begin{aligned}
\sigma_0 &= |I|,\\
\sigma_\delta &= \sum_{k=2}^{\delta} \sum_{\delta_1+\ldots+\delta_r=\delta} \frac{M}{\rho^k} \sigma_{\delta_1} \dots \sigma_{\delta_r}.
\end{aligned}$$

We prove now by induction that $|U_\delta| \leq \eta_\delta \sigma_\delta$. Indeed:

$$\begin{aligned}
|U_\delta| = |d_{0,\delta}^{-1}(V_\delta)| &\leq \frac{\delta!}{\gamma_\delta^\delta}|V_\delta| \\
&\leq \frac{\delta!}{\gamma_\delta^\delta}|\pi_\delta(R(I+U))| \\
&\leq \frac{\delta!}{\gamma_\delta^\delta}\left|\sum_{k=2}^{\delta}\sum_{\delta_1+\ldots+\delta_r=\delta}\frac{M}{\rho^k}|U_{\delta_1}|\ldots|U_{\delta_r}|\right| \\
&\leq \frac{\delta!}{\gamma_\delta^\delta}\sum_{k=2}^{\delta}\sum_{\delta_1+\ldots+\delta_r=\delta}\frac{M}{\rho^k}\eta_{\delta_1}\sigma_{\delta_1}\ldots\eta_{\delta_r}\sigma_{\delta_r} \\
&\leq \eta_\delta\sigma_\delta.
\end{aligned}$$

Because the series $\sum_{\delta\geq 0}\sigma_\delta t^\delta$ has positive radius of convergence (see Lemma 5.10) we have finished the proof. □

Remark 6.13 *The constants η_δ depend on a product of terms of the form $\frac{\delta_k!}{\gamma_{\delta_k}^{\delta_k}}$, for $\delta_k < \delta$ and can grow arbitrary large if no assumption is made on the eigenvalues λ, μ. If we assume a so called Siegel condition holds i.e. if*

$$\gamma_\delta \geq \frac{1}{\delta^\tau},$$

for a certain τ positive then

$$\frac{\delta!}{\gamma_\delta^\delta} \leq \delta^{\tau\delta}\delta! \leq \delta^{(1+\tau)\delta}.$$

Bibliography

[1] V. I. Arnol'd. *Geometrical methods in the theory of ordinary differential equations*, volume 250 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, second edition, 1988.

[2] J. Basto-Gonçalves and I. Cruz. Analytic linearizability of some resonant vector fields. *Proc. Amer. Math. Soc.*, 129(8):2473–2481, 2001.

[3] P. Bonckaert and P. De Maesschalck. Gevrey and analytic local models for families of vector fields. *Discrete Contin. Dyn. Syst. Ser. B*, 10(2-3):377–400, 2008.

[4] P. Bonckaert, I. Hoveijn, and F. Verstringe. Local analytic reduction of families of diffeomorphisms. *J. Math. Anal. Appl.*, 367(1):317–328, 2010.

[5] P. Bonckaert, V. Naudot, and J. Yang. Linearization of hyperbolic resonant germs. *Dyn. Syst.*, 18(1):69–88, 2003.

[6] P. Bonckaert and F. Verstringe. On the flat remainder in normal forms of families of analytic planar saddles. *C. R. Math. Acad. Sci. Paris*, 346(9-10):553–558, 2008.

[7] A. D. Brjuno. Analytic form of differential equations. I, II. *Trudy Moskov. Mat. Obšč.*, 25:119–262; ibid. 26 (1972), 199–239, 1971.

[8] F. E. Brochero Martínez and L. López-Hernanz. Gevrey class of the infinitesimal generator of a diffeomorphism. *Astérisque*, (323):33–40, 2009.

[9] H. W. Broer, G. B. Huitema, and M. B. Sevryuk. *Quasi-periodic motions in families of dynamical systems*, volume 1645 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1996.

[10] N. N. Brušlinskaja. A finiteness theorem for families of vector fields in the neighborhood of a singular point of Poincaré type. *Funkcional. Anal. i Priložen.*, 5(3):10–15, 1971.

[11] M. Canalis-Durand and R. Schäfke. Divergence and summability of normal forms of systems of differential equations with nilpotent linear part. *Ann. Fac. Sci. Toulouse Math. (6)*, 13(4):493–513, 2004.

[12] T. Carletti. Exponentially long time stability near an equilibrium point for non-linearizable analytic vector fields. *Z. Angew. Math. Phys.*, 56(4):559–571, 2005.

[13] T. Carletti and S. Marmi. Linearization of analytic and non-analytic germs of diffeomorphisms of $(\mathbb{C}, 0)$. *Bull. Soc. Math. France*, 128(1):69–85, 2000.

[14] J. Casasayas, E. Fontich, and A. Nunes. Invariant manifolds for a class of parabolic points. *Nonlinearity*, 5(5):1193–1210, 1992.

[15] S.-N. Chow, C. Z. Li, and D. Wang. *Normal forms and bifurcation of planar vector fields.* Cambridge University Press, Cambridge, 1994.

[16] C. Christopher, P. Mardešić, and C. Rousseau. Normalizability, synchronicity, and relative exactness for vector fields in $\mathbb{C}^2$. *J. Dynam. Control Systems*, 10(4):501–525, 2004.

[17] R. Cushman and J. A. Sanders. Nilpotent normal forms and representation theory of sl$(2, \mathbf{R})$. In *Multiparameter bifurcation theory (Arcata, Calif., 1985)*, volume 56 of *Contemp. Math.*, pages 31–51. Amer. Math. Soc., Providence, RI, 1986.

[18] D. DeLatte. On normal forms in Hamiltonian dynamics, a new approach to some convergence questions. *Ergodic Theory Dynam. Systems*, 15(1):49–66, 1995.

[19] D. DeLatte and T. Gramchev. Biholomorphic maps with linear parts having Jordan blocks: linearization and resonance type phenomena. *Math. Phys. Electron. J.*, 8:Paper 2, 27 pp., 2002.

[20] A. Delshams and J. T. Lázaro. Pseudo-normal form near saddle-center or saddle-focus equilibria. *J. Differential Equations*, 208(2):312–343, 2005.

[21] M. di Bernardo, C. J. Budd, A. R. Champneys, and P. Kowalczyk. *Piecewise-smooth dynamical systems*, volume 163 of *Applied Mathematical Sciences.* Springer-Verlag London Ltd., London, 2008.

[22] H. Dulac. Sur les cycles limites. *Bull. Soc. Math. France*, 51:45–188, 1923.

[23] F. Dumortier, P. R. Rodrigues, and R. Roussarie. *Germs of diffeomorphisms in the plane.* Lecture Notes in Mathematics. 902. Berlin-Heidelberg-New York: Springer-Verlag., 1981.

[24] J. Écalle. Compensation of small denominators and ramified linearisation of local objects. *Astérisque*, (222):4, 135–199, 1994.

[25] I. A. Gorbovitskis. Normal forms of families of mappings in the Poincaré domain. *Tr. Mat. Inst. Steklova*, 254(Nelinein. Anal. Differ. Uravn.):101–110, 2006.

[26] T. Gramchev and M. Yoshino. Normal forms for commuting vector fields near a common fixed point. In *SPT 2007—Symmetry and perturbation theory*, pages 81–91. World Sci. Publ., Hackensack, NJ, 2008.

[27] D. M. Grobman. Homeomorphy of dynamical systems. *Differencial'nye Uravnenija*, 5:1351–1359, 1969.

[28] P. Hartman. A lemma in the theory of structural stability of differential equations. *Proc. Amer. Math. Soc.*, 11:610–620, 1960.

[29] J. E. Humphreys. *Introduction to Lie algebras and representation theory.* Springer-Verlag, New York, 1972.

[30] Y. S. Il'yashenko. Divergence of series that reduce an analytic differential equation to linear normal form at a singular point. *Funktsional. Anal. i Prilozhen.*, 13(3):87–88, 1979.

[31] Y. S. Il'yashenko and S. Yakovenko. Finitely smooth normal forms of local families of diffeomorphisms and vector fields. *Uspekhi Mat. Nauk*, 46(1(277)):3–39, 240, 1991.

[32] Y. S. Il'yashenko and S. Yakovenko. Nonlinear Stokes phenomena in smooth classification problems. In *Nonlinear Stokes phenomena*, volume 14 of *Adv. Soviet Math.*, pages 235–287. Amer. Math. Soc., Providence, RI, 1993.

[33] Y. S. Il'yashenko and S. Yakovenko. *Lectures on analytic differential equations*, volume 86 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, 2008.

[34] G. Iooss and E. Lombardi. Polynomial normal forms with exponentially small remainder for analytic vector fields. *J. Differential Equations*, 212(1):1–61, 2005.

[35] K. Khanin, J. Lopes Dias, and J. Marklof. Multidimensional continued fractions, dynamical renormalization and KAM theory. *Comm. Math. Phys.*, 270(1):197–231, 2007.

[36] H. Koch and J. Lopes Dias. Renormalization of Diophantine skew flows, with applications to the reducibility problem. *Discrete Contin. Dyn. Syst.*, 21(2):477–500, 2008.

[37] E. Lombardi and L. Stolovitch. Normal forms of analytic perturbations of quasihomogeneous vector fields: Rigidity, invariant analytic sets and exponentially small approximation. *Ann. Sci. Éc. Norm. Supér. (4)*, 43(4):659–718, 2010.

[38] J. Lopes Dias. Renormalization scheme for vector fields on $\mathbb{T}^2$ with a Diophantine frequency. *Nonlinearity*, 15(3):665–679, 2002.

[39] L. Lopez-Hernanz. Existence and summability of invariant curves in complex dynamics in dimension two, 2010.

[40] F. Loray. A preparation theorem for codimension-one foliations. *Ann. of Math. (2)*, 163(2):709–722, 2006.

[41] D. M. Malonza. Stanley decomposition for coupled Takens-Bogdanov systems. *J. Nonlinear Math. Phys.*, 17(1):69–85, 2010.

[42] P. Mardešić, D. Marín, and J. Villadelprat. Unfolding of resonant saddles and the Dulac time. *Discrete Contin. Dyn. Syst.*, 21(4):1221–1244, 2008.

[43] S. Marmi, P. Moussa, and J.-C. Yoccoz. Complex Brjuno functions. *J. Amer. Math. Soc.*, 14(4):783–841, 2001.

[44] K. R. Meyer. The implicit function theorem and analytic differential equations. In *Dynamical systems—Warwick 1974 (Proc. Sympos. Appl. Topology and Dynamical Systems, Univ. Warwick, Coventry, 1973/1974; presented to E. C. Zeeman on his fiftieth birthday)*, pages 191–208. Lecture Notes in Math., Vol. 468. Springer, Berlin, 1975.

[45] J. Moser. On the generalization of a theorem of A. Liapounoff. *Comm. Pure Appl. Math.*, 11:257–271, 1958.

[46] J. Murdock. *Normal forms and unfoldings for local dynamical systems.* Springer Monographs in Mathematics. Springer-Verlag, New York, 2003.

[47] I. Niven. *Irrational numbers.* The Carus Mathematical Monographs, No. 11. The Mathematical Association of America. Distributed by John Wiley and Sons, Inc., New York, N.Y., 1956.

[48] R. Pérez-Marco. Linearization of holomorphic germs with resonant linear part. 2000.

[49] R. Pérez-Marco. Total convergence or general divergence in small divisors. *Comm. Math. Phys.*, 223(3):451–464, 2001.

[50] R. Pérez-Marco. Convergence or generic divergence of the Birkhoff normal form. *Ann. of Math. (2)*, 157(2):557–574, 2003.

[51] O. Perron. *Die Lehre von den Kettenbrüchen.* Chelsea Publishing Co., New York, N. Y., 1950.

[52] J. Raissy. Linearization of holomorphic germs with quasi-Brjuno fixed points. *Math. Z.*, 264(4):881–900, 2010.

[53] C. Robinson. *Dynamical systems. Stability, symbolic dynamics, and chaos.* Studies in Advanced Mathematics., 1995.

[54] H. Rüssmann. Stability of elliptic fixed points of analytic area-preserving mappings under the Bruno condition. *Ergodic Theory Dynam. Systems*, 22(5):1551–1573, 2002.

[55] D. Schlomiuk and N. Vulpe. Global classification of the planar Lotka-Volterra differential systems according to their configurations of invariant straight lines. *J. Fixed Point Theory Appl.*, 8(1):177–245, 2010.

[56] C. L. Siegel. Über die Existenz einer Normalform analytischer Hamiltonscher Differentialgleichungen in der Nähe einer Gleichgewichtslösung. *Math. Ann.*, 128:144–170, 1954.

[57] L. Stolovitch. Sur un théorème de Dulac. *Ann. Inst. Fourier (Grenoble)*, 44(5):1397–1433, 1994.

[58] L. Stolovitch. Family of intersecting totally real manifolds of $(\mathbb{C}^n, 0)$ and CR-singularities. *arXiv:math/0506052*, 2005.

[59] E. Stróżyna and H. Żoładek. The analytic and formal normal form for the nilpotent singularity. *J. Differential Equations*, 179(2):479–537, 2002.

[60] E. Stróżyna and H. Żoładek. Multidimensional formal Takens normal form. *Bull. Belg. Math. Soc. Simon Stevin*, 15(5, Dynamics in perturbations):927–934, 2008.

[61] F. Takens. Singularities of vector fields. *Inst. Hautes Études Sci. Publ. Math.*, (43):47–100, 1974.

[62] M. Tsuji. *Potential theory in modern function theory.* Maruzen Co. Ltd., Tokyo, 1959.

[63] W. Tucker. Robust normal forms for saddles of analytic vector fields. *Nonlinearity*, 17(5):1965–1983, 2004.

[64] J.-C. Yoccoz. Théorème de Siegel, nombres de Bruno et polynômes quadratiques. *Astérisque*, (231):3–88, 1995.

[65] E. Zehnder. A simple proof of a generalization of a theorem by C. L. Siegel. In *Geometry and topology (Proc. III Latin Amer. School of Math., Inst. Mat. Pura Aplicada CNPq, Rio de Janeiro, 1976)*, pages 855–866. Lecture Notes in Math., Vol. 597. Springer, Berlin, 1977.

[66] X. Zhang. Planar analytic systems having locally analytic first integrals at an isolated singular point. *Nonlinearity*, 17(3):791–801, 2004.