APPLICATION OF RIGOROUS HIGH-ORDER METHODS AND NORMAL FORMS TO NONLINEAR SYSTEMS

By

Adrian Weisskopf

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Physics – Doctor of Philosophy

2021

ABSTRACT

APPLICATION OF RIGOROUS HIGH-ORDER METHODS AND NORMAL FORMS TO NONLINEAR SYSTEMS

By

Adrian Weisskopf

The nonlinearities of dynamical systems often display the most interesting and fascinating behavior. At the same time, those nonlinearities complicate finding closed form analytic solutions, especially for complex systems, to the point where it is often impossible. Differential algebra (DA) based methods allow us to analyze those systems with all their nonlinearities up to arbitrary order in an automated, computer based framework that operates with floating point accuracy.

This thesis will investigate repetitive dynamical systems from seemingly unrelated fields of study using DA methods such as DA based transfer and Poincaré maps, the DA normal form algorithm, normal form defect studies, and verified methods based on Taylor Models. The common mathematical underpinnings of those dynamical systems allow us to analyze them with different techniques that have the same methods at their core.

Specifically, we will analyze resonances, associated fixed point structures, and oscillation periods of particles in the accelerator storage ring of the muon g-2 experiment at Fermilab to gain a detailed understanding of the stability of the system and the potential loss mechanism of particles. If successful, the muon g-2 experiment raises existential questions about the completeness of the Standard Model of particle physics, which makes our contributions to understanding of the system's stability highly relevant.

The same methods used for the analysis of the accelerator storage ring will also be used to generate far reaching sets of satellite orbits for formation flying missions under the Earth's gravitational zonal perturbations. Our approach is particularly elegant and precise, and its theoretical limits are far beyond the range of practical applications.

One central method in both of those analyses is the DA normal form algorithm. Using the well-known example of the centrifugal governor for illustration, the special properties of the resulting

normal form, the sensitivities and limitations of the algorithm, and its resulting quantities are explained in detail.

In the last chapter, we will provide first results and an outlook for future work of the presented methods in the realm of verified methods, and illustrate the current possibilities as well as future opportunities and challenges. In particular, Taylor Model based verified global optimization is introduced and used to calculate rigorous stability estimates for different configurations of the muon g-2 storage ring.

To my parents and grandparents.

ACKNOWLEDGEMENTS

First of all, I would like to thank my academic advisor Professor Martin Berz for his continuous support, patience, and helpful guidance not only in my research, but also during my Ph.D. time in general. I really enjoyed diving into complex problems with him and developing a solid understanding of the key mechanisms at play. Martin always encouraged me to make sure that I understood the fundamental components of a problem before building sophisticated solutions, which has strongly influenced my way of structuring problems and approaching their solution.

Furthermore, I particularly appreciated the collaborative work with Roberto Armellin, David Tarazona, Kyoko Makino, and Eremey Valetov and would like to thank all of them for the insightful discussions and welcoming atmosphere on and off work. I am proud of our joined contributions to the scientific community and really enjoyed the process of getting there.

I also shall not forget to thank Kyoko Makino, Scott Pratt, Mark Dykman and Vladimir Zelevinsky for being so kind and agreeing to be on my thesis committee.

This thesis and Ph.D. was only possible due to the generous scholarship by the Studienstiftung des deutschen Volkes and support from the DOE.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Henri Poincaré was a pioneer – his three volumes on 'New Methods of Celestial Mechanics' [65] were one of the greatest methodological contributions not only to the field of celestial mechanics, but for the mathematical theory of dynamical systems in general. Numerous methods to describe and analyze dynamical systems in various research areas have been established and developed based on his work.

Poincaré's ideas and concepts were groundbreaking, but strongly limited in their application. Performing his perturbation theory approaches by hand requires a certain simplicity or algebraic structure of the considered system. Many complex systems do not exhibit this simplicity by definition and are impossible to solve in a purely analytic closed form. Consequently, those systems are often reduced in their complexity to ideal cases or simplified versions to solve them analytically.

Computer based numerical methods have been developed to solve complex systems for very specific initial conditions with floating point accuracy. However, to develop sophisticated solutions of complex systems, which are more general than just for a specific set of initial conditions, it is critical to capture as much of the algebraic structure of the problem as possible. The differential algebra (DA) framework developed by Berz *et al.* [14, 10, 13, 9] (Sec. 2.1) constitutes a hybrid structure that manages both of these aspects. It captures the algebraic structure of a system up to arbitrary order to carry out the perturbation part going back to Poincaré's theory, while its implementation in COSY INFINITY [21, 18, 53] allows for an automated calculation of algebraic solutions in a computer environment based on floating point arithmetic.

This thesis will use this powerful hybrid and its associated methods to dive into the fascinating world of nonlinear dynamical systems. The common mathematical underpinnings of many of those systems make it possible to apply the highly developed DA methods to seemingly unrelated fields of study using suitable transformations and projections. To emphasize this versatility of the methods, we analyze one problem from the field of accelerator physics in Chapter 5, and one problem from

the field of astrodynamics in Chapter 4. Additionally, we introduce a key technique – the DA normal form algorithm [14] – in Chapter 3, where we analyze the well known system of the centrifugal governor not in its usual linearized version, but with its high order nonlinearities.

The analysis in the field of accelerator physics in Chapter 5 is concerned with the stability and the oscillation frequencies of particles in the storage ring of the muon g-2 experiment at Fermilab (E989). We investigate the dependence of these frequencies on offsets in the momentum of the particles and on the amplitudes of oscillation. Nonlinear effects of the various electric field and magnetic field components of the storage rings that are used to confine the particles and bend their trajectory cause these shifts in the frequencies, which potentially influences the beam's susceptibility to resonances. In fact, for the specific ring configurations considered in this thesis, the resonance behavior and their associated fixed point structures make this analysis particularly interesting from a dynamical systems point of view.

In contrast, the analysis in the field of astrodynamics in Chapter 4 is concerned with the trajectories of satellites in low and medium Earth orbits under zonal gravitational perturbation. The perturbation significantly distorts the orbits from their Keplerian form, causing them to rotate within their orbital plane and precess around the Earth at different frequencies. We present a method that elegantly solves one of the key challenges in astrodynamics, namely the bounded motion problem under zonal perturbation. Our method generates far reaching continuous sets of orbits, which remain in close proximity to each other over decades despite the perturbation.

An essential tool in all of those applications are DA transfer maps and Poincaré maps [14, 34] (see. Sec. 2.2). Instead of continuously working with the equations of motions in the form of ordinary differential equations (ODE) as Poincaré did, we work with maps generated from those ODEs. They yield an arbitrarily high order description of a system's behavior between two discrete instances of time or location. Maps are particularly useful for the analysis of repetitive systems in the form of Poincaré return maps, where the maps represent the system's behavior in a chosen cross section of the motion for each turn. A repetitive application of the map to a state in that cross section corresponds to the propagation of the state in the system. Accordingly, the repetitive application

allows for a stroboscopic study of the repetitive motion with all the implications regarding its stability.

Origin preserving Poincaré return maps, which are expanded around a linearly stable fixed point, are the starting point of the DA normal form algorithm [14, 12, 11] (see Sec. 2.3). The linearly stable fixed point corresponds to a stable equilibrium state in the Poincaré projection of the system. With the DA normal form algorithm, the phase space behavior around the fixed point of the map is transformed to normalized coordinates, which are closely related to action-angle coordinates. In those normal form coordinates, the phase space behavior is rotationally invariant with only amplitude dependent angle advancements up to the order of calculation. Accordingly, the angle advancements and the amplitude describe the dynamics in a nutshell (see Sec. 2.3.1).

This generalized nonlinear normalization method up to arbitrary order is very powerful and has many applications making it the main component of many techniques used in this thesis. As already mentioned above, the entire Chapter 3 focuses on a detailed walk through of the DA normal form algorithm using the centrifugal governor as an example. While the principal structure of the process is rather straightforward, the implications of individual steps are not always obvious. This chapter allows discussing those intricacies in full detail.

One critical aspect of the normal form transformation is its sensitivity to resonances (see Sec. 2.3.2). Resonances can affect the normalization process such that the rotationally invariant structure of the resulting normal form is perturbed depending on the strength of the resonances.

Accordingly, those resonances constituting one of the driving factors of the normal form defect (see Sec. 2.4), which is a measure of the variance of the (pseudo-)invariants produced by the normal form. This variance yields a local rate of divergence and can therefore be used as a stability estimate. Phase space regions with large normal form defects can trigger diverging phase space behavior and indicate less stable motion.

As an outlook for future developments, Chapter 6 discusses the first steps of enhancing the methods for these specific applications by making them completely verified. We will see that fully transferring these methods to a verified version is everything but trivial and still to be further

investigated. As a starting point for the verified analysis, we introduce verified global optimization [7, 61, 22, 55, 50, 37] and its application for a verified stability estimate of the muon g-2 storage ring.

The basis of this discussion and the global optimization method (see Sec. 2.6) are Taylor Models [46, 51, 47, 48, 15, 66] (see Sec. 2.5), which yield a structure for verified computations by enhancing the DA framework with rigorous remainder bounds.

CHAPTER 2

METHODS

The methods used for this thesis are hybrids of numerical and analytical techniques based on a differential algebra (DA) framework, which was first developed to its current extent by Berz *et al.* [14, 9, 10]. The following summary and introduction to the DA framework (Sec. 2.1), DA maps (Sec. 2.2), and the DA normal form algorithm (Sec. 2.3) are based on [14] and have been given in similar form in my previous publications [88, 89, 86, 87].

In Sec. 2.3.1, the resulting quantities of the normal form, namely the tune, tune shifts, and normal form radii, are discussed in more detail. The influence of resonances on the normal form are described in Sec. 2.3.2. Sec. 2.4 yields an introduction to the normal form defect, a measure for the non-invariance of the normal form radii, based on [22].

The introduction to Taylor Models (Sec. 2.5) for verified computations and their applications including verified global optimization (Sec. 2.6) are based on the work of Makino and Berz *et al.* [46, 51, 47, 48, 22, 54].

2.1 The Differential Algebra (DA) Framework

The fundamental purpose of the DA framework [14] is to provide a mathematical backbone for computer based storage and manipulation of analytic functions. In principle, this is done by representing an analytic function f in terms of its Taylor polynomial expansion $\mathcal{T}_f$ up to order m, similar to how real numbers are represented by an approximation up to a certain arbitrary number of significant digits. In order to discuss the mathematical construction of the differential algebra framework in more detail, we require the notation '$=_m$' instead of just '$\approx$' to clarify that both sides of such an equation are equivalent up to order m.

A Taylor polynomial expansion $\mathcal{T}_f$ of order m represents multiple analytic functions which are equivalent up to order m. This gives rise to the definition of equivalence classes following [14, p. 91]. The equivalence class $[f]_m$ represents all elements f of the vector space of infinitely differentiable

functions $\mathcal{C}^\infty(\mathbb{R}^n)$ with n real variables that have identical derivatives at the origin up to order m. The origin is chosen out of convenience and without loss of generality – any other point may be selected. In the DA framework, the equivalence class $[f]_m$ is represented by a DA vector, which stores all the coefficients of the Taylor expansion of f and the corresponding order of the terms in an orderly fashion. Operations are defined on the vector space ${}_mD_n$ of all the equivalence classes $[]_m$.

There are three operations: addition, vector multiplication, and scalar multiplication, which yield results equivalent to the result up to order m of adding two polynomials, multiplying two polynomials, and multiplying them with a scalar. The first two operations on the equivalence classes (DA vectors) form a ring. The scalar multiplication makes the three operations on the real (or complex) DA vectors an algebra, where not every element has a multiplicative inverse. An example of such elements without a multiplicative inverse is functions without a constant part like $f(x) = x$, since $1/f(x) = 1/x$ is not defined at the origin and can therefore not be expanded around it.

To make the algebra a differential algebra, the derivation D satisfying Leibniz's law ($D(fg) = fD(g) + gD(f)$) is introduced, which is almost trivial in the picture of differentiating polynomial expansions. The derivation opens the door to algebraic treatment of ordinary and partial differential equations as it is common in the study of differential algebras [68, 67, 39].

Implemented in COSY INFINITY [21, 18, 53], the DA framework allows preserving the algebraic structure up to arbitrary order while manipulating the coefficients of the DA vectors with floating point accuracy. Detailed examples of the operations on ${}_1D_1$ and ${}_2D_1$ are given in [14] and [86], respectively. An example of a DA vector in the application of DA transfer maps and Poincaré maps is given in Sec. 2.2.

2.2 DA Transfer Maps and Poincaré Maps

The dynamics of a system are often described by a set of ordinary differential equations (ODE) $\dot{\vec{z}} = f(\vec{z}, t)$, which describe the incremental change of a state $\vec{z}$ over an independent variable t like time. For practical purposes, it is often advantageous to generally describe the long term propagation of a state $\vec{z}$.

In the terminology of dynamical system theory, a so-called flow operator $\mathcal{M}_T$ is used to describe the action of the system on a state $\vec{z}$ after a fixed time T. Since it is often impossible to determine the flow in a closed form, numerical integration of the ODE is required. The DA framework allows for a hybrid integration that conserves the algebraic structure up to arbitrary order during the integration. Integrating a local expansion $\delta\vec{z}_i$ around an initial state $\vec{z}_0$ yields the final state $\vec{z}_f$ in form of a m order flow map $\mathcal{M}_T$, which depends on the expansion in $(\delta\vec{z}, \delta\vec{\eta})$, where $\delta\vec{\eta}$ is the expansion around a reference set of parameters $\vec{\eta}_0$.

More generally speaking, a transfer map $\mathcal{M}$ algebraically expresses how a final state $\vec{z}_f$ is dependent on an initial state $\vec{z}_i$ and system parameters $\vec{\eta}$, as

$$\vec{z}_f = \mathcal{M}\left(\vec{z}_i, \vec{\eta}\right). \tag{2.1}$$

Transfer maps are also called propagators or simply maps. The expansion point of the map belongs to a chosen reference orbit/state of the system, e.g. a (pseudo-)closed orbit for a fixed point map and/or the ideal orbit of the unperturbed system.

There are special transfer maps called Poincaré maps [65] that constrain the initial and final state to Poincaré surfaces $\mathbb{S}_i$ and $\mathbb{S}_f$, respectively. For the simulation of storage rings and their particle optical elements, this concept is used to represent how the state directly after a storage ring element depends on system parameters and the state directly before the element. A setup of multiple consecutive storage ring elements is described by the composition of their Poincaré maps.

Poincaré return maps represent the case where $\mathbb{S}_i$ is equal to $\mathbb{S}_f$. They are particularly useful for the representation of dynamics in repetitive systems like the ones considered in this thesis. Multiple applications of a Poincaré return map correspond to the propagation of the system. The Poincaré return maps are particularly advantageous when they are origin preserving, i.e., the expansion point is a fixed point of the map, because system dynamics represented by origin preserving Poincaré return maps can be further analyzed by normal form methods and for the asymptotic stability of the system.

Constraining the map to the Poincaré surface $\mathbb{S}$ is often done by calculating the flow of an ODE and projecting it onto the surface $\mathbb{S}$. This reduces the dimension of the original map and

generates the Poincaré map. An implementation of a timewise projection onto a surface $\mathbb{S}$ defined by $\sigma(\vec{z}, \vec{\eta}) = 0$ is outlined in [34].

The projection uses DA inversion methods that compute the inverse $\mathcal{A}^{-1}$ to the auxiliary map $\mathcal{A}$, which contains the constraining conditions of the Poincaré surface $\mathbb{S}$. Given that $\mathcal{A}$ has no constant part, the auxiliary map and its inverse satisfy $\mathcal{A}^{-1} \circ \mathcal{A} =_m \mathcal{A} \circ \mathcal{A}^{-1} =_m \mathcal{I}$. The basic idea of the projection of a transfer map $\mathcal{M}$ onto a surface defined by $\sigma(\vec{z}, \vec{\eta}) = 0$ is to replace one of the variables or parameters of $\mathcal{M}$ by an expression in terms of all the other variables and parameters such that the constraint $\sigma(\mathcal{M}) = 0$ is satisfied. This eliminates the corresponding component of the map and thereby reduces its dimensionality. In [34], the timewise projection is prepared by calculating an expansion of the map $\mathcal{M}$ in time t. The DA inversion methods are then used to find the intersection time $t^\star(\vec{z}, \vec{\eta})$ dependent on the state variables $\vec{z}$ and system parameters $\vec{\eta}$ such that $\sigma(\mathcal{M}(\vec{z}, \vec{\eta}, t^\star(\vec{z}, \vec{\eta}))) = 0$.

2.3 The DA Normal Form Algorithm

The DA normal form (DANF) algorithm [14] is an advancement from the DA-Lie based version, the first arbitrary order algorithm by Forest, Berz, and Irwin [31]. Given an origin preserving map $\mathcal{M}$ of a repetitive Hamiltonian system, where the components of the map are in phase space coordinates, the DA normal form algorithm provides a nonlinear change of phase space variables by an order-by-order transformation to rotationally invariant normal form coordinates.

Implemented in COSY INFINITY [53, 19], this is a fully automated process, which can be performed up to arbitrary order. It is only limited by floating point accuracy and the capability of the computer system to handle DA vectors of the chosen computation order. In the standard configuration, order ten calculations of a six dimensional system are easily manageable.

In Chapter 3, the normal form algorithm is explained in great detail for the one dimensional system of a centrifugal governor. Here we want to explain the more general form for a $2n$ dimensional symplectic system with an optional parameter dependence on n_η parameters summarized in $\vec{\eta}$. The explanations are largely based on [14].

For parameter dependent maps, the algorithm starts by expanding the origin preserving map $\mathcal{M}(\vec{z}, \vec{\eta})$ around its parameter dependent fixed point $\vec{z}_{\mathrm{FP}}(\vec{\eta})$, which satisfies

$$\mathcal{M}\left(\vec{z}_{\mathrm{FP}}(\vec{\eta}), \vec{\eta}\right) = \vec{z}_{\mathrm{FP}}(\vec{\eta}). \tag{2.2}$$

Defining the extended map $\mathcal{N} = (\mathcal{M} - \mathcal{I}_{\vec{z}}, \vec{\eta})$, the parameter dependent fixed point $\vec{z}_{\mathrm{FP}}$ is determined by evaluating the inverse of $\mathcal{N}$ at the expansion point $\vec{z} = \vec{0}$:

$$\left(\vec{z}_{\mathrm{FP}}\left(\vec{\eta}\right), \vec{\eta}\right) = \mathcal{N}^{-1}\left(\vec{0}, \vec{\eta}\right). \tag{2.3}$$

The map $\mathcal{M}$ is then expanded around its parameter dependent fixed point $\vec{z}_{\mathrm{FP}}$.

The resulting map $\mathcal{M}_0 = \mathcal{L} + \sum_m \mathcal{U}_m$ consists of a linear part $\mathcal{L}$ and the nonlinear parts $\mathcal{U}_m$ of order m. Due to the transformation to the parameter dependent fixed point, the map has no terms, only depending on a parameter. Accordingly, the entire linear part is independent of parameters.

The variables of the map are the canonical phase space coordinates $\vec{z} = (\vec{q}_0, \vec{p}_0)$ and, if applicable, parameters $\vec{\eta}$. The normal form algorithm transforms this map order by order up to the full order of the map. For each transformation step, the transformation $\mathcal{A}_m$ and its inverse are determined and applied to the result $\mathcal{M}_{m-1}$ from the previous transformation step as follows

$$\mathcal{M}_m = \mathcal{A}_m \circ \mathcal{M}_{m-1} \circ \mathcal{A}_m^{-1}. \tag{2.4}$$

The first step of the algorithm is to linearly decouple the map into n two dimensional subspaces. The linear transformation diagonalizes the system, transforming the (parameter dependent) fixed point map into the complex conjugate eigenvector space of its linear part. We assume linearly stable behavior around the (parameter dependent) fixed point of the map with distinct complex conjugate eigenvalue pairs of magnitude one since this property is shared among all systems considered in this thesis (see [14] for other cases). If any of the eigenvalues $\lambda_\star$ had an absolute value larger than 1, the motion would be unstable since the state on the corresponding eigenvector $\vec{v}_\star$ would grow in magnitude by a factor of $\lambda_\star > 1$ with each iteration. Additionally, eigenvalues of symplectic maps come in reciprocal pairs such that eigenvalues with a magnitude smaller than 1 have a reciprocal partner eigenvalue $\lambda_\star > 1$, which are again linearly unstable.

The complex conjugate eigenvalue pairs $e^{\pm i\mu_j}$ of the diagonalized linear part are grouped together such that the matrix $\hat{R}$ of the diagonalized linear part $\mathcal{R}$ of the resulting map $\mathcal{M}_1 = \mathcal{R} + \sum_m \mathcal{S}_m$ has the following decoupled form

$$\hat{R} = \begin{pmatrix} \hat{R}_1 & & & & \\ & \ddots & & & \\ & & \hat{R}_l & & \\ & & & \ddots & \\ & & & & \hat{R}_n \end{pmatrix} \quad \text{where} \quad \hat{R}_j = \begin{pmatrix} e^{+i\mu_j} & 0 \\ 0 & e^{-i\mu_j} \end{pmatrix}. \tag{2.5}$$

The new nonlinear terms of order m that resulted from the linear transformation are denoted by $\mathcal{S}_m$. The complex phase $\pm\mu_j$ of the eigenvalue pairs will be of critical importance in the nonlinear transformations of the algorithm.

In summary, the first transformation step performed the following operation

$$\mathcal{M}_1 = \mathcal{A}_1 \circ \mathcal{M} \circ \mathcal{A}_1^{-1} = \mathcal{A}_1 \circ \mathcal{L} \circ \mathcal{A}_1^{-1} + \sum_m \mathcal{A}_1 \circ \mathcal{U}_m \circ \mathcal{A}_1^{-1} = \mathcal{R} + \sum_m S_m, \tag{2.6}$$

where $\mathcal{A}_1$ is the linear transformation from the original coordinate space $(\vec{q}_0, \vec{p}_0)$ to the complex conjugate coordinate space $(\vec{q}_1, \vec{p}_1)$ and $\mathcal{A}_1^{-1}$ is its inverse for the transformation in the opposite direction.

With the linearly decoupled map, the following steps of the normal form algorithm can be performed for each of these linearly decoupled subspaces separately. The jth subspace of the linearly decoupled map $\mathcal{M}_1$ can be explicitly written as

$$\mathcal{M}_{1,j}\left(\vec{q}_1, \vec{p}_1, \vec{\eta}\right) = \mathcal{R}_j + \sum_m \mathcal{S}_{m,j} = \begin{pmatrix} e^{+i\mu_j} & 0 \\ 0 & e^{-i\mu_j} \end{pmatrix} \begin{pmatrix} q_{1,j} \\ p_{1,j} \end{pmatrix} \tag{2.7}$$

$$+ \sum_{m=||\vec{k}^+ + \vec{k}^-||_1 + ||\vec{k}^\eta||_1} \begin{pmatrix} \mathcal{S}^+_{m(\vec{k}^+, \vec{k}^-, \vec{k}^\eta), j} \\ \mathcal{S}^-_{m(\vec{k}^+, \vec{k}^-, \vec{k}^\eta), j} \end{pmatrix} \prod_{l=1}^{n} \left(q_{1,l}\right)^{k_l^+} \left(p_{1,l}\right)^{k_l^-} \prod_{u=1}^{n_\eta} \left(\eta_u\right)^{k_u^\eta} \tag{2.8}$$

where k_l^+ represents the positive integer exponent of $q_{1,l}$, k_l^- represents the positive integer exponent of $p_{1,l}$, and k_u^η represents the positive integer exponent of η_u. The positive integer exponents are summarized in the vectors $\vec{k}^+$, $\vec{k}^-$, and $\vec{k}^\eta$, respectively. The L^1-Norm $||\cdot||_1$ of the sum of these vectors is used to ensure that only polynomial terms of order m are considered.

To get a better feeling of the expression in Eq. (2.8), we present some terms of the $\mathcal{M}^-_{1,j}$ component

$$\begin{aligned}\mathcal{M}^-_{1,j}(\vec{q}_1,\vec{p}_1,\vec{\eta}) = e^{-i\mu_j}\cdot p_{1,j} &+ \mathcal{S}^-_{2\left((2,0,\ldots,0)^T,(0,\ldots,0)^T,(0,\ldots,0)^T\right),j}\cdot q_{1,1}^2 + \ldots \\ &+ \mathcal{S}^-_{2\left((0,\ldots,0,k_j^+=1,0,\ldots,0)^T,(0,\ldots,0,k_l^-=1,0,\ldots,0)^T,(0,\ldots,0)^T\right),j}\cdot q_{1,j}p_{1,l} + \ldots \\ &+ \mathcal{S}^-_{2\left((0,\ldots,0)^T,(0,\ldots,0,1)^T,(1,0,\ldots,0)^T\right),j}\cdot p_{1,n}\eta_1 + \ldots\end{aligned} \tag{2.9}$$

Due to the linear transformation into the complex conjugate eigenvector space of the purely real linear part, the two components of each subspace form a complex conjugate pair. Accordingly, the '+' and '-' notation is used, where the sign corresponds to the sign of the complex eigenvalue phase of the map component of that subspace. Specifically, this means that $\mathcal{M}^+_{1,j} = \overline{\mathcal{M}}^-_{1,j}$, with $q_{1,j} = \overline{p}_{1,j}$ and $\mathcal{S}^+_{m(\vec{k}^+,\vec{k}^-,\vec{k}\eta),j} = \overline{\mathcal{S}}^-_{m(\vec{k}^+,\vec{k}^-,\vec{k}\eta),j}$.

This property is maintained throughout all the following nonlinear transformation steps, which are performed order by order starting with order two. The general form of the nonlinear transformation is $\mathcal{A}_m =_m \mathcal{I} + \mathcal{T}_m$, where $\mathcal{T}_m$ is a polynomial containing only terms of order m. Hence, the transformation $\mathcal{A}_m$ is a near-identity transformation and a full identity up to order $m-1$. The transformation $\mathcal{A}_m$ is determined by finding $\mathcal{T}_m$ such that the mth order of the map $\mathcal{M}_{m-1}$ is simplified or even eliminated when the transformation $\mathcal{A}_m$ and its inverse $\mathcal{A}_m^{-1} =_m \mathcal{I} - \mathcal{T}_m$ are applied to it in the mth order nonlinear transformation step (see Eq. (2.4)).

The higher order terms of the transformation $\mathcal{A}_m$ do not influence the mth order terms of the map. Accordingly, they are irrelevant for the mth order transformation step and can be chosen freely, e.g. to make the transformation symplectic with $\mathcal{A}_m = \exp(L_{\mathcal{T}_m})$ which we will do (see [14]). However, the higher orders of the resulting map $\mathcal{M}_m$ are strongly dependent on $\mathcal{A}_m$, its higher order terms, and its corresponding inverse. In Chapter 3, the influences of the second order transformation on the third order terms of the resulting map are analyzed in great detail. While these influences are not to be dismissed, the key element of this m order transformation step is the elimination of as many mth order terms of the map $\mathcal{M}_{m-1}$ as possible by a smart choice of $\mathcal{T}_m$.

Given the map $\mathcal{M}_{m-1}$, representing $\mathcal{M}$ simplified up to order $m-1$ and applying $\mathcal{A}_m$ and its inverse to it, yields [14, Eq. (7.60)]:

$$\begin{aligned}\mathcal{A}_m \circ \mathcal{M}_{m-1} \circ \mathcal{A}_m^{-1} &=_m (\mathcal{I} + \mathcal{T}_m) \circ (\mathcal{R} + \mathcal{S}_m) \circ (\mathcal{I} - \mathcal{T}_m) \\ &=_m (\mathcal{I} + \mathcal{T}_m) \circ (\mathcal{R} - \mathcal{R} \circ \mathcal{T}_m + \mathcal{S}_m) \\ &=_m \mathcal{R} + \mathcal{S}_m + [\mathcal{T}_m, \mathcal{R}]\,, \end{aligned} \tag{2.10}$$

where $\mathcal{R}$ is the diagonalized linear part and $\mathcal{S}_m$ represents only the mth order terms of the map $\mathcal{M}_{m-1}$ (the leading order of terms that have not been simplified yet).

The equations above only consider terms up to order m, since terms of order $m+1$ and larger are irrelevant for determining $\mathcal{T}_m$. The maximum simplification would be achieved by finding $\mathcal{T}_m$ such that the commutator $\mathcal{C}_m = \mathcal{T}_m \circ \mathcal{R} - \mathcal{R} \circ \mathcal{T}_m = [\mathcal{T}_m, \mathcal{R}] = -\mathcal{S}_m$, which would eliminate all nonlinear terms $\mathcal{S}_m$ of order m.

Since the commutator only involves $\mathcal{T}_m$ and $\mathcal{R}$ we can investigate this transformation separately in the n individual subspaces. The components of the jth subspace of the commutator $\mathcal{C}_m = [\mathcal{T}_m, \mathcal{R}]$ are

$$\mathcal{C}_{m,j} = \sum_{m=||\vec{k}^+ + \vec{k}^-||_1 + ||\vec{k}^\eta||_1} \begin{pmatrix} \mathcal{C}^+_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j} \\ \mathcal{C}^-_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j} \end{pmatrix} \prod_{l=1}^{n} (q_l)^{k_l^+} (p_l)^{k_l^-} \prod_{u=1}^{n_\eta} (\eta_u)^{k_u^\eta}, \tag{2.11}$$

where

$$\mathcal{C}^\pm_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j} = \mathcal{T}^\pm_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j} \left(e^{i\vec{\mu}\left(\vec{k}^+ - \vec{k}^-\right)} - e^{\pm i\mu_j} \right). \tag{2.12}$$

Accordingly, the commutator terms $\mathcal{C}^\pm_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j}$ can eliminate their corresponding nonlinear terms of the map $\mathcal{S}^\pm_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j}$ by choosing

$$\mathcal{T}^\pm_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j} = \frac{-\mathcal{S}^\pm_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j}}{e^{i\vec{\mu}\left(\vec{k}^+ - \vec{k}^-\right)} - e^{\pm i\mu_j}}, \tag{2.13}$$

if

$$e^{i\vec{\mu}\left(\vec{k}^+ - \vec{k}^-\right)} - e^{\pm i\mu_j} \neq 0. \tag{2.14}$$

In other words, only the $\mathcal{S}^{\pm}_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j}$ terms corresponding to $\mathcal{C}^{\pm}_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j}$ for which the condition [14, Eq. (7.65)]

$$\mathrm{mod}_{2\pi}\left(\mu_j(k_j^+ - k_j^- \mp 1) + \sum_{l\neq j}\mu_l\left(\vec{k}^+ - \vec{k}^-\right)\right) = 0, \tag{2.15}$$

is satisfied, survive.

A straightforward solution of the condition in Eq. (2.15) is

$$k_j^+ - k_j^- = \pm 1 \quad \wedge \quad k_l^+ = k_l^- \; \forall l \neq j, \tag{2.16}$$

where the first condition concerns the jth subspace and the second condition is regarding all the other subspaces l with $l \neq j$.

The surviving terms of the mth order transformation step in the jth subspace can be generally written as

$$\mathcal{S}^{+}_{m(\vec{k}+\vec{e}_j,\vec{k},\vec{k}^\eta),j} \quad \text{and} \quad \mathcal{S}^{-}_{m(\vec{k},\vec{k}+\vec{e}_j,\vec{k}^\eta),j} \quad \text{with} \quad 2||\vec{k}||_1 + 1 + ||\vec{k}^\eta||_1 = m, \tag{2.17}$$

where the unit vector $\vec{e}_j$ consists only of zeros except for a 1 at the jth entry.

From Eq. (2.17) it becomes clear that only certain terms of uneven order in the phase space coordinates $(\vec{q}, \vec{p})$ survive. These terms have the special property that each complex conjugate phase space variable pair is raised to the same exponent except for the phase space variable pair of the respective subspace. Accordingly, all even order terms in phase space coordinates can be eliminated by the nonlinear normal form transformations.

The remaining terms of $\mathcal{S}_m$ (from Eq. (2.17)) describe the entire dynamics of the systems in a nutshell and are the key elements of the normal form and therefore essential for further dynamic analysis.

Resonances between the complex phases $\vec{\mu}$ of the different subspaces in the denominator of Eq. (2.13) can break this special structure and therefore the rotational invariance of the normal form as will be discussed in Sec. 2.3.2. For now, we will continue only with the terms that are supposed to survive, namely the terms specified in Eq. (2.17).

Once the nonlinear transformation steps transformed the map up to its full order, the map has been significantly simplified to

$$\begin{pmatrix} \mathcal{M}^{+}_{m,j} \\ \mathcal{M}^{-}_{m,j} \end{pmatrix} = \begin{pmatrix} q_{m,j}\, f^{+}_{j}\left(q_{m,1}p_{m,1}, q_{m,2}p_{m,2}, ..., q_{m,n}p_{m,n}, \vec{\eta}\right) \\ p_{m,j}\, f^{-}_{j}\left(q_{m,1}p_{m,1}, q_{m,2}p_{m,2}, ..., q_{m,n}p_{m,n}, \vec{\eta}\right) \end{pmatrix}, \tag{2.18}$$

where

$$f^{+}_{j} = e^{+i\mu} + \sum_{m=2||\vec{k}||_1+1+||\vec{k}\eta||_1} \mathcal{S}^{+}_{m(\vec{k}+\vec{e}_j,\vec{k},\vec{k}^{\eta}),j} \prod_{l=1}^{n} \left(q_{m,l}p_{m,l}\right)^{k_l} \prod_{u=1}^{n_\eta} (\eta_u)^{k^{\eta}_u} \tag{2.19}$$

Since the original map is real, the last step of the algorithm is transforming the resulting map to the real normal form basis $(\vec{q}_{\mathrm{NF}}, \vec{p}_{\mathrm{NF}})$, which is composed of the real and imaginary parts of the current complex conjugate basis $(\vec{q}_m, \vec{p}_m)$. The relation between the bases is

$$q_{\mathrm{NF},j} = \frac{q_{m,j} + p_{m,j}}{2} \qquad p_{\mathrm{NF},j} = \frac{q_{m,j} - p_{m,j}}{2i} \tag{2.20}$$

$$q_{m,j} = q_{\mathrm{NF},j} + i p_{\mathrm{NF},j} \qquad p_{m,j} = q_{\mathrm{NF},j} - i p_{\mathrm{NF},j}. \tag{2.21}$$

The squared normal form radius $r^2_{\mathrm{NF},j}$ is given by the product of $q_{m,j}p_{m,j}$, with

$$q_{m,j}p_{m,j} = q^2_{\mathrm{NF},j} + p^2_{\mathrm{NF},j} = r^2_{\mathrm{NF},j}. \tag{2.22}$$

Applying the basis transformation to the map components of $\mathcal{M}_m$ in each subspace yields

$$\begin{aligned}
\mathcal{M}_{\mathrm{NF},j} &= \mathcal{A}_{\mathrm{real},j} \circ \mathcal{M}_{m,j} \circ \mathcal{A}^{-1}_{\mathrm{real},j} \\
&= \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix} \cdot \begin{pmatrix} f^{+}_{j}\left(r^2_{\mathrm{NF},1}, ..., r^2_{\mathrm{NF},n}, \vec{\eta}\right)\left(q_{\mathrm{NF},j} + i\, p_{\mathrm{NF},j}\right) \\ f^{-}_{j}\left(r^2_{\mathrm{NF},1}, ..., r^2_{\mathrm{NF},n}, \vec{\eta}\right)\left(q_{\mathrm{NF},j} - i\, p_{\mathrm{NF},j}\right) \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{2}\left(f^{+}_{j} + \bar{f}^{+}_{j}\right) q_{\mathrm{NF},j} + \frac{i}{2}\left(f^{+}_{j} - \bar{f}^{+}_{j}\right) p_{\mathrm{NF},j} \\ \frac{-i}{2}\left(f^{+}_{j} - \bar{f}^{+}_{j}\right) q_{\mathrm{NF},j} + \frac{1}{2}\left(f^{+}_{j} + \bar{f}^{+}_{j}\right) p_{\mathrm{NF},j} \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{Re}\left(f^{+}_{j}\left(r^2_{\mathrm{NF},1}, ..., r^2_{\mathrm{NF},n}, \vec{\eta}\right)\right) & -\mathrm{Im}\left(f^{+}_{j}\left(r^2_{\mathrm{NF},1}, ..., r^2_{\mathrm{NF},n}, \vec{\eta}\right)\right) \\ \mathrm{Im}\left(f^{+}_{j}\left(r^2_{\mathrm{NF},1}, ..., r^2_{\mathrm{NF},n}, \vec{\eta}\right)\right) & \mathrm{Re}\left(f^{+}_{j}\left(r^2_{\mathrm{NF},1}, ..., r^2_{\mathrm{NF},n}, \vec{\eta}\right)\right) \end{pmatrix} \cdot \begin{pmatrix} q_{\mathrm{NF},j} \\ p_{\mathrm{NF},j} \end{pmatrix}.
\end{aligned} \tag{2.23}$$

Writing f^{+}_{j} and its complex conjugate counterpart f^{-}_{j} in terms of complex phases with

$$f^{\pm}_{j}\left(r^2_{\mathrm{NF},1}, ..., r^2_{\mathrm{NF},n}, \vec{\eta}\right) = e^{\pm i\Lambda_j\left(r^2_{\mathrm{NF},1}, ..., r^2_{\mathrm{NF},n}, \vec{\eta}\right)} \tag{2.24}$$

yields the following normal form

$$\mathcal{M}_{\mathrm{NF},j} = \begin{pmatrix} \cos\left(\Lambda_j\left(r^2_{\mathrm{NF},1},...,r^2_{\mathrm{NF},n},\vec{\eta}\right)\right) & -\sin\left(\Lambda_j\left(r^2_{\mathrm{NF},1},...,r^2_{\mathrm{NF},n},\vec{\eta}\right)\right) \\ \sin\left(\Lambda_j\left(r^2_{\mathrm{NF},1},...,r^2_{\mathrm{NF},n},\vec{\eta}\right)\right) & \cos\left(\Lambda_j\left(r^2_{\mathrm{NF},1},...,r^2_{\mathrm{NF},n},\vec{\eta}\right)\right) \end{pmatrix} . \begin{pmatrix} q_{\mathrm{NF},j} \\ p_{\mathrm{NF},j} \end{pmatrix}, \quad (2.25)$$

which clearly shows the circular phase space behavior in normal form subspaces with only amplitude $\vec{r}_{\mathrm{NF}}$ and parameter $\vec{\eta}$ depended angle advancements $\vec{\Lambda}$.

The radii of the circular motion – the normal form radii – are constants of motion up to the calculation order. The entire dynamics in the normal form are given by the constant angle advancements $\vec{\Lambda}$ along the circular phase space curves. The rotational invariance implies an interpretation of the normal form as an averaged representation of the original Poincaré return map $\mathcal{M}$, in the limit where the map application is repeated infinitely many times.

Normalizing the angle advancements $\vec{\Lambda}$ to $[0, 1]$ yields the tunes $\vec{\nu}$ and amplitude and parameter dependent tune shifts $\delta\vec{\nu}(\vec{r}_{\mathrm{NF}}, \vec{\eta})$. Accordingly,

$$\frac{\Lambda(\delta\vec{\nu}(\vec{r}_{\mathrm{NF}}, \vec{\eta}))}{2\pi} = \nu + \delta\nu(\delta\vec{\nu}(\vec{r}_{\mathrm{NF}}, \vec{\eta})) \quad (2.26)$$

The normal form transformation $\mathcal{A}$ and its inverse $\mathcal{A}^{-1}$ are given by the composition of all the individual transformations of each transformation step with

$$\mathcal{M}_{\mathrm{NF}} = \underbrace{\mathcal{A}_{\mathrm{real}} \circ \mathcal{A}_m \circ \mathcal{A}_{m-1} \circ ... \circ \mathcal{A}_1}_{\mathcal{A}} \circ \mathcal{M} \circ \underbrace{\mathcal{A}_1^{-1} \circ ... \circ \mathcal{A}_{m-1}^{-1} \circ \mathcal{A}_m^{-1} \circ \mathcal{A}_{\mathrm{real}}^{-1}}_{\mathcal{A}^{-1}} . \quad (2.27)$$

The normal form transformation $\mathcal{A}$ yields how the normal form variables $(q_{\mathrm{NF},j}, p_{\mathrm{NF},j})$ depend on the original phase space variables $(\vec{q}_0, \vec{p}_0)$ and, if considered, system parameters $\vec{\eta}$, which suggests the following notation for $\mathcal{A}$ and its inverse

$$\mathcal{A} = \left(\vec{q}_{\mathrm{NF}}\left(\vec{q}_0, \vec{p}_0, \vec{\eta}\right), \vec{p}_{\mathrm{NF}}\left(\vec{q}_0, \vec{p}_0, \vec{\eta}\right)\right) \quad (2.28)$$

$$\mathcal{A}^{-1} = \left(\vec{q}_0\left(\vec{q}_{\mathrm{NF}}, \vec{p}_{\mathrm{NF}}, \vec{\eta}\right), \vec{q}_0\left(\vec{q}_{\mathrm{NF}}, \vec{p}_{\mathrm{NF}}, \vec{\eta}\right)\right) . \quad (2.29)$$

2.3.1 Tunes, Tune Shifts, and Normal Form Radii

DA normal form methods are used to transform the origin preserving phase space Poincaré return map to the rotationally invariant normal form up to calculation order. From the normal form, the angle advancements $\vec{\Lambda}(\vec{r}_{\mathrm{NF}}, \vec{\eta})$ as a functions of amplitude $\vec{r}_{\mathrm{NF}}$ and parameters $\vec{\eta}$ are particularly straightforward to extract. Scaling the angle advancements in each of the normal form phase spaces to $[0, 1]$ instead of $[0, 2\pi]$ provides the average number of phase space revolutions per system revolution represented by the Poincaré return map. In beam physics terminology, the frequencies of normal form phase space revolutions is known as the tunes $\vec{\nu}$ and their amplitude and parameter dependent tune shifts $\delta\vec{\nu}(\vec{r}_{\mathrm{NF}}, \vec{\eta})$.

The tune ν_j corresponds to the scaled complex phase μ_j of the complex conjugate eigenvalues $\lambda_j^{\pm}$ of the linear transformation. Hence, the tune is related to the linear motion around the expansion point, i.e., the motion 'infinitely close' to the expansion point. Interpreting the tune and its tune shifts as the phase space rotation frequency suggests that the tune – the phase space rotation frequency of the expansion point – is a rotation with no amplitude, where the frequency is determined by the linear motion around the expansion point. In particular, this means that different maps with the same expansion point can have different tunes depending on the linear motion around the expansion point. Since the tunes are calculated from the linear coefficients directly without any nonlinear transformations, performing the tune calculation with parameter dependent linear coefficients directly yields the parameter dependent tune shifts.

The tune shifts indicate the change of the phase space rotation frequency dependent on the phase space amplitudes $\vec{r}_{\mathrm{NF}}$ and variations in the system parameters $\vec{\eta}$. Since the normal form transformation is symplectic, it preserves the phase space volume, which is critical to understanding the connection between the original phase space coordinates and their normal form radii. If the system is only weakly coupled between the different phase spaces, the normal form radius $r_{\mathrm{NF},j}$ is a measure for the invariant phase space area of the jth subspace denoted by A_j. Hence, the original phase space coordinates of an invariant phase space orbit in the jth subspace enclose the area A_j, which roughly corresponds to the normal form radius of $r_{\mathrm{NF},j} = \sqrt{A_j/\pi}$.

The normal form radii are the link between the tune dependencies and the original coordinates. The dependency of the tune shifts on the normal form radii is a result of the surviving terms $\mathcal{S}_m$ of the nonlinear normal form transformations. However, the crucial terms are the $\mathcal{T}_m$ terms from Eq. (2.13) that are used to cancel all the other nonlinear terms $\mathcal{S}_m$. On the one hand, the $\mathcal{T}_m$ terms determine how the original coordinates $\vec{z} = (\vec{q}, \vec{p})$ and the system parameters $\vec{\eta}$ relate to the normal form radii $\vec{r}_{\mathrm{NF}}$, since the $\mathcal{T}_m$ are the essential part of the normal form transformation. On the other hand, they influence the higher order nonlinear terms $\mathcal{S}_l$ with $l > m$, which either survive and determine the dependency of the tune shifts on the normal form radii, or they determine the higher order terms $\mathcal{T}_l$.

2.3.2 Resonances

The denominator of $\mathcal{T}_m$ in Eq. (2.13) has a potentially large effect on the size of $\mathcal{T}_m$ the closer it is to satisfying the resonance condition in Eq. (2.15). If the condition is satisfied, the corresponding nonlinear terms in $\mathcal{S}_m$ can not be eliminated. Accordingly, terms survive which do not fit the normal form structure. They break the normal form by the size of their respective coefficient.

If the condition is almost satisfied close to a resonance, then the denominator of $\mathcal{T}_m$ becomes very small, making $\mathcal{T}_m$ very large. In this situation, there are two options. One option is to continue the procedure with the very large $\mathcal{T}_m$ coefficient, which conserves the normal form structure but yields diverging coefficients in all higher order terms. The other option is to let the corresponding term in $\mathcal{S}_m$ survive, which breaks the normal form structure but avoids a divergence of the coefficients. In practice, one chooses a cutoff value for the size of the denominator, which restricts the size of potentially diverging coefficients. If the denominator is smaller than the cutoff value, the $\mathcal{T}_m$ coefficient is set to zero, letting the corresponding $\mathcal{S}_m$ term survive.

Rewriting the resonance condition in terms of tunes yields

$$\vec{w} \cdot \vec{\nu} = g, \tag{2.30}$$

where $\vec{w}$ consists only of integer values and g is a natural number $\mathbb{N}_0$. The values in $\vec{w}$ and g are

chosen such that the greatest common divisor of all values is 1. With this definition, the order of the resonance is given by $m_{\text{res}} = ||\vec{w}||_1$.

In the normal form algorithm a tune resonance defined by $(\vec{w}, g)$ appears in all terms $\mathcal{S}^{\pm}_{m(\vec{k}^+,\vec{k}^-,\vec{k}^\eta),j}$ for which

$$w_j = k_j^+ - k_j^- \mp 1 \quad \wedge \quad w_l = k_l^+ - k_l^- \quad \forall l \neq j, \tag{2.31}$$

$$\text{and} \quad -w_j = k_j^+ - k_j^- \mp 1 \quad \wedge \quad -w_l = k_l^+ - k_l^- \quad \forall l \neq j, \tag{2.32}$$

according to Eq. (2.16). Resonances of order m_{res} appear for the first time in the normal form transformation step of order $m_{\text{NF}} = m_{\text{res}} - 1$.

Consider a four dimensional phase space system ($n = 2$) without parameter dependence, where the eigenvalue phases μ_i satisfy the following order seven resonance $2\mu_1 - 5\mu_2 = -4\pi$. This corresponds to the tune resonance condition of $-2\nu_1 + 5\nu_2 = 2$ denoted by $\left((-2, 5)^T, 2\right)$. The first terms of the normal form to encounter this resonance are the sixth order complex conjugate terms

$$\mathcal{S}^{+}_{6((0,5)^T,(1,0)^T),1} \quad \text{and} \quad \mathcal{S}^{-}_{6((1,0)^T,(0,5)^T),1} \tag{2.33}$$

$$\text{as well as} \quad \mathcal{S}^{-}_{6((0,4)^T,(2,0)^T),2} \quad \text{and} \quad \mathcal{S}^{+}_{6((2,0)^T,(0,4)^T),2}. \tag{2.34}$$

Accordingly, for each subspace, one complex conjugate pair survives due to the resonance between μ_1 and μ_2, which break the rotational symmetry structure of the resulting normal form.

2.4 The Normal Form Defect

The volume conserving property of Hamiltonian systems expressed by Liouville's theorem is maintained by the normal form transformation. Given the rotational invariants of the normal form, the size of the phase space volume is determined by the normal form radii. Accordingly, the normal form phase space radii constitute invariants of motion up to the order of the normal form transformation if no resonance conditions were encountered. However, they are usually not invariants of the full (order) motion.

While the expansion of the transfer map improves in accuracy with every additional order considered, the same is not guaranteed for the normal form transformation. It is unknown how well

or even if the normal form converges with higher orders. This is due to its sensitivity to resonances, which may initiate asymptotic behavior once the order of a close-by resonance is reached. The higher the order of the computation, the more resonances are potentially relevant. Depending on the complexity of the original transfer map, it is usually unpredictable which resonances may affect the normal form and in what way.

However, if the normal form transformation converges, its high order limit will yield the exact invariants. In the case of exact invariants, the system is integrable and can be transformed into a trivial system by introducing the invariants as variables. Those variables are known as action-angle coordinates, where the action is constant and unique for each phase space curve and each point on the phase space curve is associated with the action-angle. For complex systems such as the ones discussed in this thesis, there are no exact invariants that can be expressed in terms of finite order terms. Thus, tools to assess the error of the calculated pseudo-invariants in the form of normal form radii are useful.

The normal form defect represents the inaccuracy of the normal form radii as invariants and is locally defined for each phase space state. Given an origin preserving fixed point map $\mathcal{M}(q,p) = (Q,P)$ of a repetitive system and the corresponding normal form transformation $\mathcal{A}(q,p) = \left(q_{\mathrm{NF}}, p_{\mathrm{NF}}\right)$, the normal form defect $d_{\mathrm{NF}}(\vec{z}_0)$ of the phase space state $\vec{z}_0 = (q,p)$ is given by the difference between the normal form radius $r(\vec{z}_1 = \mathcal{M}(\vec{z}_0))$ of the mapped phase space state $\vec{z}_1 = \mathcal{M}(\vec{z}_0)$ and the normal form radius $r(\vec{z}_0)$ of the original phase space state $\vec{z}_0$. Generally, the normal form radius r of a phase space state $\vec{z}$ is the magnitude of the vector formed by the normal form phase space state $\left(q_{\mathrm{NF}}, p_{\mathrm{NF}}\right) = \mathcal{A}(\vec{z}_0)$, specifically

$$r(\vec{z}) = \sqrt{\left(q_{\mathrm{NF}}(\vec{z})\right)^2 + \left(p_{\mathrm{NF}}(\vec{z})\right)^2}. \tag{2.35}$$

Accordingly, the normal form defect is given by

$$\begin{aligned} d_{\mathrm{NF}}(\vec{z}_0) &= r_1 - r_0 = r(\vec{z}_1) - r(\vec{z}_0) = r(\mathcal{M}(\vec{z}_0)) - r(\vec{z}_0) \\ &= \sqrt{\left(q_{\mathrm{NF}}(\mathcal{M}(\vec{z}_0))\right)^2 + \left(p_{\mathrm{NF}}(\mathcal{M}(\vec{z}_0))\right)^2} - \sqrt{\left(q_{\mathrm{NF}}(\vec{z}_0)\right)^2 + \left(p_{\mathrm{NF}}(\vec{z}_0)\right)^2}. \end{aligned} \tag{2.36}$$

The application of the one turn map represents the evolution of the system by describing how each phase space state changes after one revolution of the system. The normal form defect indicates how much the normal form radii, i.e., a (pseudo-)invariants of the motion, change between two states of the motion connected by the map $\mathcal{M}$. An increasing normal form radius with time indicates diverging phase space behavior with larger amplitudes, i.e. the normal form defect measures the local rate of divergence per map application.

Analyzing the normal form defect for a whole set of states within a certain phase space domain $\mathbb{D}$ allows for stability estimations by placing an upper bound on the rate of divergence. The upper bound can be determined in various ways, including rigorous global optimization methods on the normal form defect over the given domain. The upper bound can serve as a Nekhoroshev-type stability estimate [64] that allows for the calculation of the minimum amount of revolutions of the system N, for which the motion will be guaranteed to stay within the allowed region $\mathbb{D}$:

$$N = \frac{r_{\max} - r\left(\vec{z}_{\text{ini}}\right)}{\max\left(d_{\text{NF}}\left(\vec{z}\right)\right)} \quad \text{with} \quad \vec{z} \in \mathbb{D} \tag{2.37}$$

where $r\left(\vec{z}_{\text{ini}}\right)$ is the upper bound of the normal form radius of the initial state of the system and $r_{\max}$ is the lower bound of the maximum normal form radius corresponding to motion still within the allowed region $\mathbb{D}$ (see Fig. 2.1).

The concept of the normal form defect based Nekhoroshev-type stability estimate is comparable to an augmented Lyapunov function [45]. A regular Lyapunov function L is not increasing along any phase space curve, with $L(\mathcal{M}(\vec{z})) \leq L(\vec{z})$. This works very well for systems with damping. For damped motion in a convex potential, the total energy function can serve as a Lyapunov function. For systems without damping, this is a lot less straightforward. Under the assumption that the normal form algorithm produces a normal form radius which is a true invariant of the motion, the normal form transformation to calculate the normal form radius is a regular Lyapunov function proving eternal stability. However, the errors to the limited floating point accuracy already break this hypothetical scenario. An augmented or pseudo-Lyapunov function $L_\star = L + \max\left(d_{\text{NF}}\left(\mathbb{D}\right)\right)$ is increasing in a very slow and well estimated way with a verified upper bound on the rate of increase

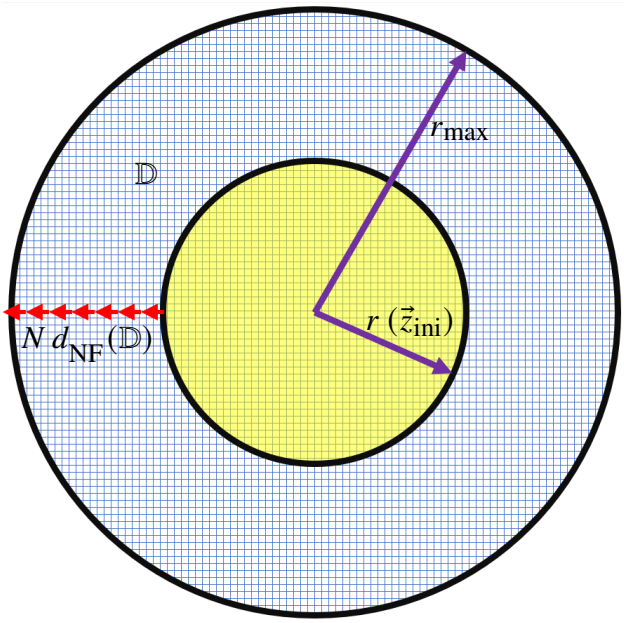

Figure 2.1: Schematic illustration of the various normal form quantities involved in the calculation of the minimum iteration number within allowed region $\mathbb{D}$.

per iteration

$$L(\mathcal{M}(\vec{z})) \leq L_\star = L(\vec{z}) + \max\left(d_{\text{NF}}\,(\mathbb{D})\right) \tag{2.38}$$

Accordingly, it can not prove eternal stability, but rigorously estimate the long term stability. See [16] and [42] for a detailed discussion.

In [22], this method was successfully used to analyze the long term stability of the Tevatron storage ring at the Fermi National Accelerator Laboratory. However, it can be generally used in dynamical systems applications to assess stability. Particularly, in complex systems where the stability in different phase space regions is not evident, the normal form defect based Nekhoroshev-type stability estimate is a great tool to capture the maximum rate of divergence.

2.5 Verified Computations Using Taylor Models (TM)

Based on DA vectors (Sec. 2.1), Taylor Models (TM) were developed by Makino and Berz [46, 51, 47, 48, 15, 66] as a structure for rigorously verified computations, which deals much better with issues known from interval arithmetic like the dependency problem [48], the wrapping effect

[54, 52, 17], and linear scaling of the overestimation with domain size. Accordingly, the following introduction to TM and their application is largely based on their work [46, 47, 48, 15, 66, 54].

To better understand the advantages of TM, we will first take a quick look at the alternative of using interval arithmetic for verified computations.

2.5.1 Interval arithmetic

Intervals are a basic concept to represent a range of numbers and are often used to capture uncertainty. The interval $I = [a, b] = \{x \mid a \leq x \leq b\}$ represents all numbers between a and b, and the values a and b themselves.

The basic interval arithmetic [59, 60, 41] for the addition, subtraction, multiplication, and division of two intervals $I_1 = [a_1, b_1]$ and $I_2 = [a_2, b_2]$ are given by the following operations. The addition yields

$$I_1 + I_2 = [a_1 + a_2, b_1 + b_2] \,. \tag{2.39}$$

The subtraction operation $I_1 - I_2$ works equivalently by performing the addition of I_1 with $-I_2 = [-b_2, -a_2]$.

The multiplication yields

$$I_1 \cdot I_2 = [\min(a_1a_2, a_1b_2, b_1a_2, b_1b_2), \max(a_1a_2, a_1b_2, b_1a_2, b_1b_2)] \,. \tag{2.40}$$

The division is only possible if the divisor interval does not contain zero. If the divisor does not contain zero, the division I_1/I_2 is equivalently defined by multiplying I_1 with

$$\frac{1}{I_2} = \left[\frac{1}{b_2}, \frac{1}{a_2}\right] \quad \text{for} \quad 0 \notin I_2. \tag{2.41}$$

This arithmetic provides the mathematically tightest bounds when the quantities represented by I_1 and I_2 are independent. But since this is rarely the case, the calculated bounds are an overestimation due to the dependency problem, which is easily illustrated by considering the difference between an interval and itself. The result of the expression $x - x$ should be zero, but from the arithmetic above the difference between two identical intervals is

$$I - I = [a, b] - [a, b] = [-(b - a), (b - a)] \,, \tag{2.42}$$

which has a width of $2\,(b-a)$ instead of zero width.

Compared to DA vectors (see Sec. 2.1), which form a ring structure, intervals do not even form a group structure, because neither for addition nor multiplication there is an inverse for intervals of nonzero width.

For the interval evaluation of functions, further rules can be established. Monotonically increasing functions $f_{\mathrm{mon}\nearrow}$ like $\exp(x)$ can be evaluated by

$$f_{\mathrm{mon}\nearrow}([a,b]) = \left[f_{\mathrm{mon}\nearrow}(a), f_{\mathrm{mon}\nearrow}(b)\right]. \tag{2.43}$$

Monotonically decreasing functions $f_{\mathrm{mon}\searrow}$ can be equivalently evaluated by

$$f_{\mathrm{mon}\searrow}([a,b]) = \left[f_{\mathrm{mon}\searrow}(b), f_{\mathrm{mon}\searrow}(a)\right]. \tag{2.44}$$

Trigonometric functions are compositions of monotonically increasing and monotonically decreasing sections, which are well known. Accordingly, the interval evaluation of a trigonometric function can be implemented based on many subcases depending on the size and position of the interval.

Considering the function $f(x) = \sin(\frac{\pi x}{2}) - \exp(x)$ and evaluating it over the domain interval $I_1 = [-1, 1]$ yields

$$f\left(I_1\right) = \sin\left(\frac{\pi I_1}{2}\right) - \exp\left(I_1\right) = I_1 - \left[\exp\left(-1\right), \exp\left(1\right)\right] \tag{2.45}$$

$$= \left[-1-e, 1-e^{-1}\right] \subset [-3.718282, 0.632121]\,. \tag{2.46}$$

We will compare this interval evaluation to the performance of different order Taylor Models in the following section.

2.5.2 Taylor Models

Taylor Models [46, 51, 47, 48, 15, 66] are remainder-enhanced DA vectors. The DA part of the TM is a mth order Taylor polynomial in form of a regular DA vector representation of a function f, which is differentiable m times, as introduced in Sec. 2.1. The remainder part complements this by

rigorously verified bounds on the error of using the truncated Taylor expansion up to order m in form of a DA vector compared to f itself.In contrast to regular DA vectors, TM need to be defined over a domain $\mathbb{D}$ to be able to rigorously bound the remainder.

This approach is based on the Taylor Remainder Theorem: Given a function $f : \vec{\mathbb{D}} = \left[\vec{a}, \vec{b}\right] \subset \mathbb{R}^n \to \mathbb{G} \subset \mathbb{R}$ being $(m+1)$ times continuously partially differentiable on the domain $\vec{\mathbb{D}}$ with $\vec{x}_0 \in \vec{\mathbb{D}}$. Then for each $\vec{x} \in \vec{\mathbb{D}}$ there is a $\eta \in (0, 1)$ such that

$$f(\vec{x}) = \underbrace{\sum_{k=0}^{m} \left. \frac{\left((\vec{x} - \vec{x}_0) \cdot \vec{\nabla}_{\vec{y}}\right)^k f(\vec{y})}{k!} \right|_{\vec{y}=\vec{x}_0}}_{\mathcal{P}_{m,f}} + \underbrace{\left. \frac{\left((\vec{x} - \vec{x}_0) \cdot \vec{\nabla}_{\vec{y}}\right)^{m+1} f(\vec{y})}{(m+1)!} \right|_{\vec{y}=\vec{x}_0+(\vec{x}-\vec{x}_0)\eta}}_{\mathcal{E}_{m,\mathbb{D},f}} \tag{2.47}$$

where $\mathcal{P}_{m,f}$ is the polynomial part and $\mathcal{E}$ is an expression for the remainder.

A Taylor Models is characterized by its order m, the function f it is representing and the domain $\mathbb{D}$ over which the representation of f is within the verified bounds of the Taylor Model. We denote a Taylor Model with

$$\mathcal{T}_{m,\mathbb{D},f} = \left(\mathcal{P}_{m,f}, \epsilon_{m,\mathbb{D},f}\right), \tag{2.48}$$

where $\mathcal{P}_{m,f}$ is the Taylor polynomial term of order m and $\epsilon_{m,\mathbb{D},f}$ is a rigorous verified estimation $\epsilon_{m,\mathbb{D},f}$ of the remainder size over the domain $\mathbb{D}$ such that for function f

$$\left|f(\vec{x}) - \mathcal{P}_{m,f}(\vec{x})\right| < \epsilon_{m,\vec{\mathbb{D}},f} \quad \forall \vec{x} \in \vec{\mathbb{D}}. \tag{2.49}$$

A Taylor Model can be visualized as a tube that wraps around the mth order DA representation with a distance ϵ such that the original expression is guaranteed to lie within the tube over the given domain $\mathbb{D}$ (see Fig. 2.2).

Except for order $m = 1$, the Taylor Model bounding of f significantly outperforms the interval bounding. The tightness of the bounding also improves drastically with higher order Taylor Models. With every additional order, the polynomial part clings closer to f, and the reminder gets smaller and smaller.

This tighter and tighter bounding with higher orders shows how the DA part of the Taylor Models avoids more and more of the dependency problem. Dependent expressions like $1 + x - x$, which

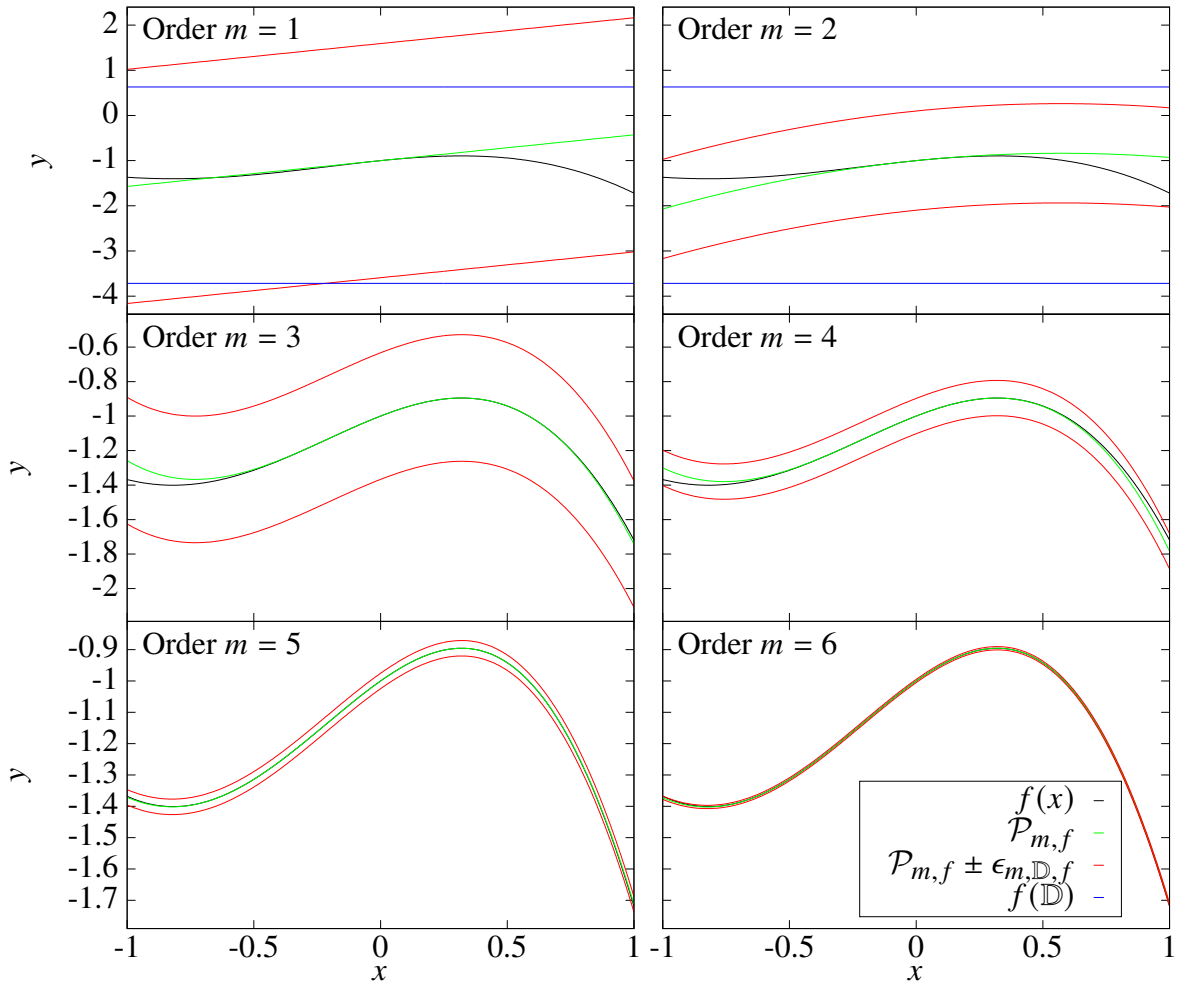

Figure 2.2: Verified representation of $f(x) = \sin(\frac{\pi x}{2}) - \exp(x)$ over the domain $\mathbb{D} = I_1 = [-1, 1]$ with interval methods using $f(\mathbb{D})$ and with Taylor Models $(\mathcal{P}_{m,f}, \epsilon_{m,\mathbb{D},f})$ of various orders m. The original function $f(x)$ is indicated by the black line, while its DA polynomial representation is shown in green. The bounds at a distance $\epsilon_{m,\mathbb{D},f}$ from the DA polynomial are red. The two straight blue lines indicate the bounds of the interval evaluation. Note that the scale of the y axis is changing to better illustrate the tightness of the Taylor Model representation with higher orders. Accordingly, the interval bounds are only shown for order $m = 1$ and order $m = 2$.

may arise as the first order part of expressions like $\exp(x) - \sin(x)$ are reduced to just $1 + 0$ in the DA part of the Taylor Model description. As we saw in Sec. 2.5.1, Interval arithmetic is not able to avoid this dependency problem.

The fourth order Taylor Model representation of the function $f(x) = \sin(\frac{\pi x}{2}) - \exp(x)$ over the domain $\mathbb{D} = I_1 = [-1, 1]$ would be

$$\mathcal{T}_{4,I_1,f}(x) = \left(-1 + \frac{(\pi - 2)x}{2} - \frac{x^2}{2} - \frac{(\pi^3 - 8)x^3}{48} - \frac{x^4}{24}, 0.102345\right) \tag{2.50}$$

2.6 Taylor Model based Verified Global Optimizers

The goal of a verified global optimizer [7, 61, 22, 55, 50, 37] is finding the optimum of a given scalar objective function $f(\vec{x})$ of n_{var} variables x_i over a predefined n_{var} dimensional global search domain box $\vec{\mathbb{B}}$. Without loss of generality, it is assumed that the optimum is a minimum. If the optimum is a maximum, consider the optimization of $-f(\vec{x})$.

Ideally, the result of global optimization yields the minimum $f^\star$ of the objective function $f(\vec{x})$ and all locations $\vec{x}^\star$, where the minimum is assumed within the global search domain box $\vec{\mathbb{B}}$. However, straightforward and exact analytic solutions of the optimization problem only exist for elementary objective functions. As soon as higher order terms and multiple variables are involved, iterative algorithms to track down the optimum are inevitable. Consequently, results are often only approximations of the actual minimum and all their locations where it is assumed. Verified global optimizers compensate for the shortcoming of being unable to pinpoint the exact minimum by yielding rigorously verified bounds on the minimum and its locations.

The fundamental idea of a global optimization algorithm is the efficient elimination of subdomains/subboxes of the initial search box $\vec{\mathbb{B}}$ by proving that those eliminated subboxes do not contain the minimum. The basic steps of the algorithm are the following:

1. Split domain box $\vec{\mathbb{B}}$ into subdomains $\vec{\mathbb{B}}_i$
2. Determine a lower bound $f_{i,\text{LB}}$ of f over $\vec{x} \in \vec{\mathbb{B}}_i$
3. Calculate/Update the cutoff value $\mathcal{C}$ – the currently lowest known upper bound of the minimum.
4. Eliminate all boxes $\mathbb{B}_i$ with a lower bound $f_{i,\text{LB}}$ larger than the cutoff value $\mathcal{C}$
5. Restart the algorithm at step 1 for each of the non-eliminated domain boxes $\vec{\mathbb{B}}_i^{\#}$

The more subdomain boxes are eliminated in step 4 in each iteration, the more effective the algorithm. Accordingly, it is essential to use methods for very tight bounding in step 2 (making $f_{i,\text{LB}}$ as large as possible), and to use heuristics to significantly improve the cutoff value $\mathcal{C}$ in step 3, making it as small as possible.

For the determination of the cutoff value $\mathcal{C}$ in step 3, any method or combination of methods that produce a tight verified upper bound on the global minimum of the search domain are useful. A typical technique is the verified evaluation of individual points within the domain box. The testing points are chosen either randomly in a Monte-Carlo based approach or by heuristics, e.g., the results of non-verified optimization over the domain. Depending on the computational effort of those methods, the improvement of the cutoff value and its benefits for the algorithm must be weighed against the computation time of the cutoff method.

For step 2, Taylor Models (see Sec. 2.5) are particularly useful, especially high order Taylor Models, since they allow for very tight bounding compared to interval methods. This property can mainly be ascribed to the avoidance of the dependency problem due to the DA vector part of the TM. For very complex objective functions like the normal form defect (see Sec. 2.4), the evaluation with very high order Taylor Models (e.g. order ten) can take considerably more time compared to evaluations with lower order Taylor Models (e.g. order three). Again, the benefits of the more precise bounding with high order Taylor Model evaluation have to be weighed against the associated computation time. A rule of thumb is that the larger the evaluation domain and the more complex the objective function, the larger the benefit of higher order Taylor Models.

For the rigorous bounding of Taylor Models, there are multiple approaches. The standard method uses order bounds, where the terms belonging to each order are bound and summed up together with the remainder bound. More sophisticated methods are discussed in great detail in [56]. They can be briefly summarized as follows. The linear dominated bounder (LDB) is very efficient for linear dominated domains. The quadratic dominated bounder (QDB) is good at determining the minimum of a multidimensional quadratic dominated function but losses its efficiency with very high dimensional problems. The quadratic fast bounder (QFB) is not as exact as the QDB but very efficient in providing a good lower bound near a local minimum, where the Hessian matrix of the objective function over the domain is positive definite.

To avoid an infinite continuation of the splitting, stop conditions are implemented, which are checked before a domain box is split. A typical stop condition sets a lower bound on the size of the

domain, either by setting a lower bound on the volume of the domain box or its side length. Another possible stop condition is a lower bound on the tightness of the bounding of the minimum of the objective function rather than the domain size. With such a stop condition in place, the algorithm would not split a non-eliminated domain box over which the bounds of the minimum are tighter than a certain given value. This is particularly useful if the exact minimum is not relevant but rather the order of magnitude of the minimum.

CHAPTER 3

AN EXAMPLE-DRIVEN WALK-THROUGH OF THE DA NORMAL FORM ALGORITHM

This chapter is based on my arXiv preprint and MSU Report MSUHEP-190617 *Introduction to the Differential Algebra Normal Form Algorithm using the Centrifugal Governor as an Example* [87].

We provide a very detailed description of the steps involved in the DA normal form algorithm (Sec. 2.3) and their implications for the normal form using the example of the centrifugal governor. We pick this example because it is one dimensional and the derivation of the equations of motion and the linearization of the motion are well known. This understanding yields the groundwork for the non-trivial analysis of the nonlinear phenomena using the steps of the DA normal form algorithm.

3.1 The Centrifugal Governor

The centrifugal governor (see Fig. 3.1) is a device involving gravitational and centrifugal forces with the rotation axis parallel to the direction of the gravitational force. We consider a mathematically

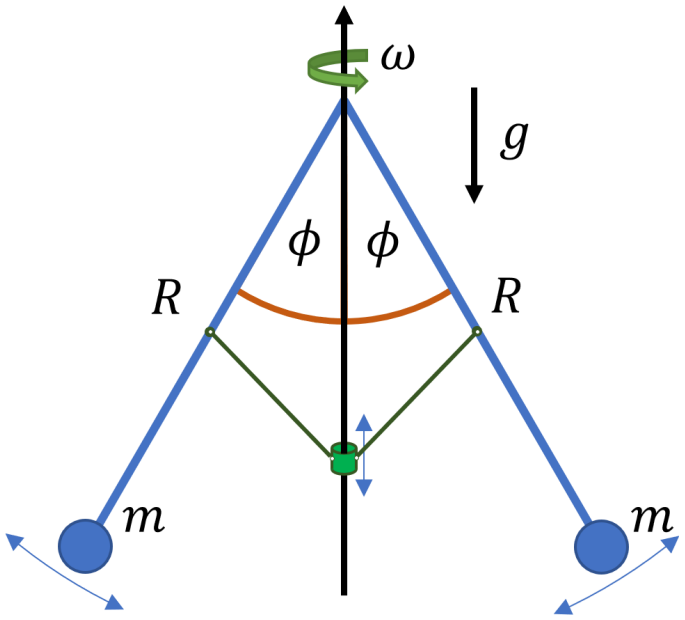

Figure 3.1: Schematic illustration of centrifugal governor.

idealized governor, which consists of two massless rods of equal length R suspended in a common

plane with the rotation axis. A point mass m is attached at the end (opposite to where the rod is mounted) of each of the rods. The angle between the rotation axis and the rod is denoted by the angle ϕ. A mechanism links the two rods and the rotation axis, which guarantees identical angles and therefore identical behavior on both sides. An external torque applied via the rotation axis ensures that the rotation frequency ω of the centrifugal governor arms is kept constant.

In the usual application of a centrifugal governor, the rotation frequency is not fixed but negatively coupled to the angle ϕ through an additional mechanism external to the governor itself. This additional mechanism makes the system self regulating by decreasing ω for an increase in ϕ. Accordingly, in those applications, e.g. the steam engine, the rotation frequency ω changes during the regulating process. However, as already mentioned above, for the introduction to the DA normal form algorithm, we consider the motion of the system for a fixed rotation frequency ω, i.e. no self-regulating coupling mechanism between ϕ and ω.

3.1.1 Units

To limit the number of parameters in the following calculations to just the rotation frequency ω, we scale time, distance, and mass in such a way that the mass m, the gravitational constant g, and the length of the rods R are all equal to one in their respective scaled units and therefore disappear from the equations. Specifically, mass is considered in units of the point mass m, distances are considered in units of the rod length R, and time is considered in units of

$$T_0[\mathrm{s}] = \sqrt{\frac{R[\mathrm{m}]}{g\left[\frac{\mathrm{m}}{\mathrm{s}^2}\right]}}, \tag{3.1}$$

such that the gravitational constant g equal one in units of distance R and time T_0.

3.1.2 The Equilibrium Point

For any given fixed rotation frequency ω, there is an angle ϕ_0 so that $\phi(t) = \phi_0$ is a solution of the motion of the centrifugal governor arms. This equilibrium angle is characterized by the alignment of the rods with the vector sum of the vertical gravitational force F_{grav} and the radial centrifugal

force F_{cent} such that there is no torque acting on the rods in the common plane of the rods and the rotation axis.

For any frequency ω, $\phi_0 = 0$ satisfies this requirement, since the centrifugal force is zero and there is only the gravitational force acting vertically downwards. However, if the rotation frequency ω is sufficiently high enough (see Eq. (3.3), a bifurcation of the equilibrium angle occurs – the angle $\phi_0 = 0$ becomes an unstable equilibrium state, while stable equilibrium angle $\phi_0(\omega) > 0$ arises, which satisfies the alignment condition with

$$\tan\phi_0 = \frac{F_{\text{cent}}}{F_{\text{grav}}} = \frac{m\omega^2 R\sin\phi_0}{mg} = \omega^2\sin\phi_0. \tag{3.2}$$

For $\phi_0 > 0$, this corresponds to

$$\cos\phi_0 = \frac{1}{\omega^2} \quad \Rightarrow \quad \phi_0 = \arccos\left(\frac{1}{\omega^2}\right) \quad \text{for} \quad \omega > 1 = \omega_{\text{min}}. \tag{3.3}$$

Fig. 3.2 visualizes the stable equilibrium angle as a function of the rotation frequency ω.

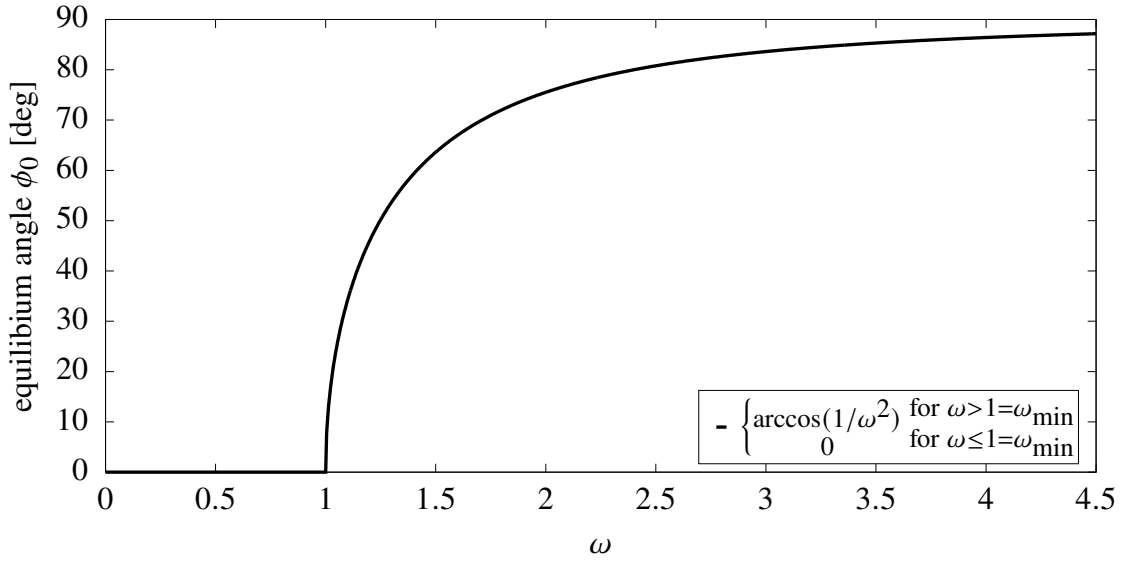

Figure 3.2: Illustration of the stable equilibrium angle ϕ_0 of the arms of the centrifugal governor as a function of the rotation frequency ω. For $\omega > \omega_{\text{min}} = 1$, $\phi_0 = 0$ is an unstable equilibrium angle.

Since the vertical contribution of the gravitational force to the vector sum is nonzero and independent of the rotation frequency, an equilibrium angle of $\phi_0 = 90°$ is only approached

asymptotically for the rotation frequency ω approaching infinity. The bifurcation of the equilibrium state at $\omega_{\min} = 1$ is also clearly visible.

Tab. 3.1 lists stable equilibrium angles for some specific rotation frequencies, especially for the fast-changing region between $\omega = 1$ and $\omega = 2$.

Table 3.1: List of stable equilibrium angles ϕ_0 of the centrifugal governor arms for some specific rotation frequencies ω.

ω	ϕ_0 [deg]	ϕ_0 [rad]
1	0°	0
$\sqrt{2}/\sqrt[4]{3}$	30°	$\frac{\pi}{6}$
$\sqrt[4]{2}$	45°	$\frac{\pi}{4}$
$\sqrt{2}$	60°	$\frac{\pi}{3}$
2	$\approx 75.52°$	≈ 1.318
20	$\approx 89.86°$	≈ 1.568
$\lim_{\omega\to\infty}$	90°	$\frac{\pi}{2}$

3.1.3 The Equations of Motion

To understand the dynamics of the centrifugal governor arms around an equilibrium state, we derive the equations of motion starting with the Lagrangian formulation of the problem. It yields

$$L = \frac{m}{2}\left(\dot{\phi}^2 R^2 + \omega^2 R^2 \sin^2\phi\right) - mgR\,(1-\cos\phi) = \frac{\dot{\phi}^2}{2} - \underbrace{\left(\frac{-\omega^2\sin^2\phi}{2} + (1-\cos\phi)\right)}_{U_{\text{eff}}}, \tag{3.4}$$

where U_{eff} is the effective or centrifugal-gravitational potential. In Fig. 3.3, we illustrate the centrifugal-gravitational potential U_{eff} for multiple rotation frequencies ω.

The minimum of the effective potential well corresponds to the stable equilibrium angle discussed in Sec. 3.1.2. The axis notations indicate that the width and the depth of the potential, in particular, increase with increasing rotation frequency ω. The higher the rotation frequency ω, the less relevant are the gravitational influences and the deeper and the more symmetric the potential well. The asymmetry of the effective potential will also be apparent in the dynamics of the system, which we

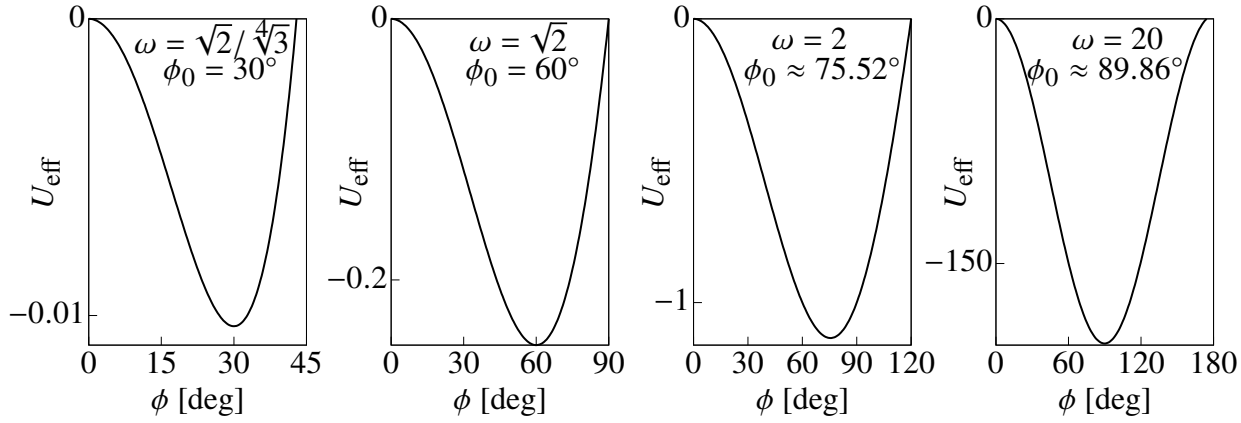

Figure 3.3: Potential well of U_{eff} for multiple oscillation frequencies ω The equilibrium angle ϕ_0 corresponds to the minimum of the potential well.

discuss in Sec. 3.1.4. For the rest of the chapter, we will focus on the case $\omega = \sqrt{2}$, which yields a clear 2:1 asymmetry left and right of its equilibrium angle.

To continue the derivation of the equations of motion, we derive the generalized canonical momentum p_ϕ to the position variable ϕ from the Lagrangian, where

$$p_\phi = \frac{dL}{d\dot{\phi}} = mR^2\dot{\phi} = \dot{\phi}. \tag{3.5}$$

Using the Legendre transformation, the Hamiltonian

$$H = \frac{p_\phi^2}{2mR^2} - \frac{m\omega^2 R^2 \sin^2\phi}{2} + mgR\,(1-\cos\phi) = \frac{p_\phi^2}{2} + U_{\text{eff}} = E \tag{3.6}$$

is obtained, which is not explicitly time dependent and therefore a constant of motion. The Hamiltonian also happens to correspond to the energy E of this system.

The equations of motions are derived from the Hamiltonian via Hamilton's equations where

$$\dot{\phi} = \frac{dH}{dp_\phi} = \frac{p_\phi}{mR^2} = p_\phi \tag{3.7}$$

$$\text{and} \quad \dot{p}_\phi = -\frac{dH}{d\phi} = -mgR\sin\phi + m\omega^2 R^2 \sin\phi\cos\phi = \sin\phi\left(\omega^2\cos\phi - 1\right). \tag{3.8}$$

In coordinates $(\delta\phi, \delta p_\phi)$ relative to the equilibrium state $(\phi_0, 0)$, the equations of motions are

$$\frac{\mathrm{d}\delta\phi}{\mathrm{d}t} = \delta p_\phi \qquad \text{and} \qquad \frac{\mathrm{d}\delta p_\phi}{\mathrm{d}t} = \sin(\phi_0+\delta\phi)\left(\omega^2\cos(\phi_0+\delta\phi) - 1\right). \tag{3.9}$$

3.1.4 Illustration of System Dynamics

With the equations of motion relative to the equilibrium state (Eq. (3.9)) and the understanding of how the shape of the effective potential well changes with the rotation frequency ω, we can now interpret the dynamics of the centrifugal governor when the angle of the rods is perturbed from the equilibrium angle $\phi_0\,(\omega)$.

In Fig. 3.4, the dynamics of the rods are shown for a rotation frequency of $\omega = \sqrt{2}$, which corresponds to an equilibrium angle of $\phi_0 = 60°$. While the oscillation is periodic, it is asymmetric

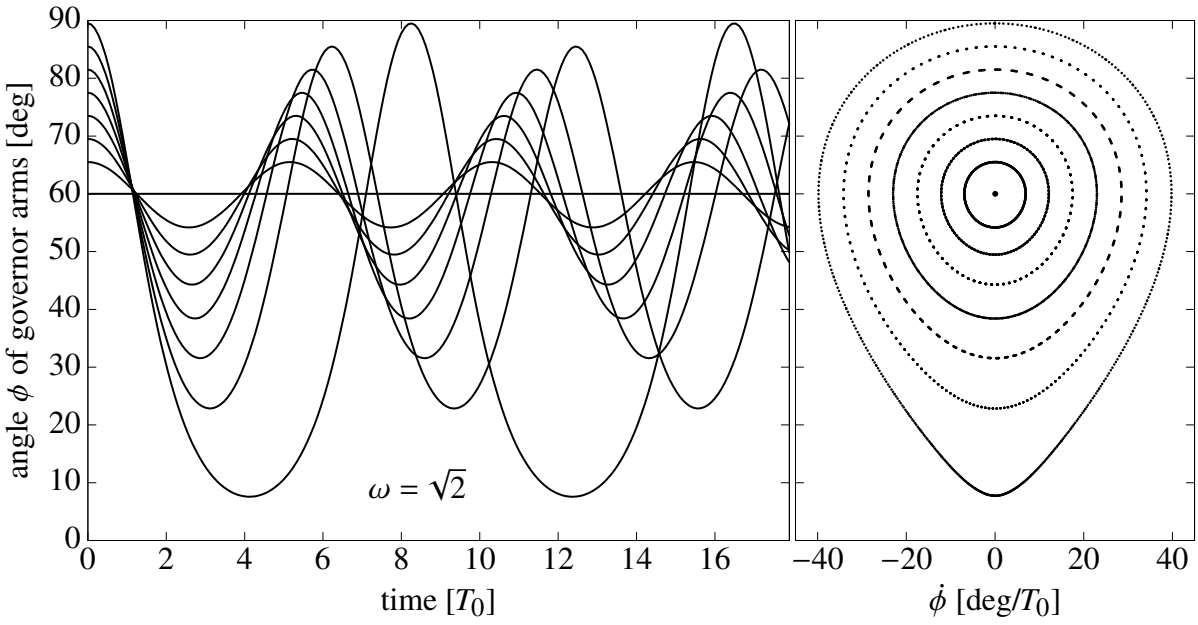

Figure 3.4: Dynamics of the centrifugal governor for a rotation frequency of $\omega = \sqrt{2}$. The centrifugal governor arms were initiated with $\dot{\phi} = p_\phi = 0$ and at the following angles: 60°, 65.5°, 69.5°, 73.5°, 77.5°, 81.5°, 85.5°, and 89.5°. The left plot shows the oscillatory behavior around the equilibrium angle at $\phi_0 = 60°$ over time. The right plot shows the stroboscopic phase space behavior from repetitive map evaluation. To related phase space behavior to the position behavior in time, the ϕ axis of both plots are aligned.

around the equilibrium point, as we would expect from the asymmetric effective potential for $\omega = \sqrt{2}$ in Fig. 3.3. The asymmetry of the oscillation is larger, the larger the angle during initiation. The maximum downward angle displacement (often more generally referred to as amplitude) and the maximum upward angle displacement of the governor's arms relative to their equilibrium angle are related through the effective potential, which corresponds to the energy of the vertical motion for

$\delta p_\phi = 0$. For both those angle displacements, the effective potential has the same maximum value or 'invariant amplitude'corresponding to the energy. The maximum amplitudes in the momentum space in the right plot of Fig. 3.4 are related the same way. In other words, the phase space motion in Fig. 3.4 corresponds to contour lines of the energy.

For future reference, it is useful to associate the term 'amplitude' not only with a physical displacement or a maximum/minimum momentum but also with an abstract quantity that relates all the different versions of phase space amplitudes like the energy in this case.

Apart from the asymmetric upward and downward position amplitudes, the left plot in Fig. 3.4 clearly shows a change in the period of oscillation depending on the angle during initiation, or more generally speaking, depending on the invariant amplitude of the motion the energy. The larger the amplitude, the longer is the period of oscillation. This is particularly prominent for the oscillation with the largest amplitude. It is also obvious, especially for the larger amplitudes that the relation between the amplitude and the period is nonlinear.

However, there is no trivial way of extracting this nonlinear relation between the amplitude and the period of oscillation from the equations of motion and/or the energy. Additionally, if we were unaware of the function for the effective potential and energy, or were considering a more complex system, it would also be very difficult to relate the different phase space amplitudes to each other. The DA normal form algorithm generates both relations in an automated process up to calculation order. In the order-by-order process, it determines an invariant amplitude up to calculation order as a function of the original phase space variables and also determines the period of oscillation as a function of that invariant amplitude.

All the normal form algorithm requires is an origin preserving transfer map (see Sec. 2.2), which represents the flow of the ODEs (see Eq. 3.9) relative to the linearly stable fixed point of the considered phase space motion. For the centrifugal governor example, the equilibrium phase space state $(\phi_0, 0)$ constitutes such a phase space fixed point, as the right plot in Fig. 3.4 already indicated. In other words, we require a functional description of how the relative phase space state $z_{\text{fin}} = (\delta\phi_{\text{fin}}, \delta p_{\phi,\text{fin}})$ after a fixed time t_0 depends on the initial relative phase space

state $z_{\text{ini}} = (\delta\phi_{\text{ini}}, \delta p_{\phi,\text{ini}})$. DA based maps (Sec. 2.2) can provide this functional description up to arbitrary order. We will use them to represent the dynamics around the equilibrium state corresponding to a rotation frequency of $\omega = \sqrt{2}$, for the later analysis with the DA normal form algorithm.

3.2 Map Calculation via Integration

As mentioned above, the following analysis of the centrifugal governor considers the system at a fixed rotation frequency of $\omega = \sqrt{2}$. We are interested in the dynamics around the corresponding equilibrium state of the centrifugal governor arms at $(60°, 0)$. The goal of this section is to generate a DA map describing the phase space dynamics relative to that equilibrium state.

For consistency with the following notation during the DA normal form algorithm introduction, we denote the phase space coordinates relative to the equilibrium point with (q_0, p_0) instead of the previously used $(\delta_\phi, \delta p_\phi)$. We will also conduct the calculations in radians rather than degrees due to their slightly easier implementation.

The map is calculated by integrating the ODEs (see Eq. (3.9)) from the initial phase space state

$$(q_{\text{ini}}, p_{\text{ini}}) = \left(\phi_0\left(\omega = \sqrt{2}\right) + \delta_\phi, \delta p_\phi\right) = \left(\frac{\pi}{3} + q_0, p_0\right) \tag{3.10}$$

from $t = 0$ until $t = t_0 = 1$. Since the flow of the ODEs in Eq. (3.9) remains expanded around the equilibrium state for any t_0, the time of the integration can be chosen freely.

The resulting map of the integration $\mathcal{M}_0 = (Q(q_0, p_0), P(q_0, p_0))^T$ has the following form: $\mathcal{M}_0 = \mathcal{C} + \mathcal{L} + \sum_m \mathcal{U}_m$, where the constant part is denoted by $\mathcal{C}$, the linear part with $\mathcal{L}$ and each of the nonlinear parts of order m with $\mathcal{U}_m$. Since the system is expanded around the equilibrium point, the constant part of the map corresponds to the equilibrium state $(\frac{\pi}{3}, 0)$. The following explicit formulation of $\mathcal{M}_0$ up to order three introduces the notation of various coefficients of the map:

$$\mathcal{M}_0(q_0,p_0)=\begin{pmatrix}\mathcal{M}_0^+(q_0,p_0)\\ \mathcal{M}_0^-(q_0,p_0)\end{pmatrix}=\begin{pmatrix}Q(q_0,p_0)\\ P(q_0,p_0)\end{pmatrix}=\underbrace{\begin{pmatrix}q_{\text{const}}\\ p_{\text{const}}\end{pmatrix}}_{\mathcal{C}}+\underbrace{\begin{pmatrix}(Q|q_0) & (Q|p_0)\\ (P|q_0) & (P|p_0)\end{pmatrix}\begin{pmatrix}q_0\\ p_0\end{pmatrix}}_{\mathcal{L}}$$
$$+\underbrace{\begin{pmatrix}\mathcal{U}^+_{2(2,0)}\\ \mathcal{U}^-_{2(2,0)}\end{pmatrix}q_0^2+\begin{pmatrix}\mathcal{U}^+_{2(1,1)}\\ \mathcal{U}^-_{2(1,1)}\end{pmatrix}q_0p_0+\begin{pmatrix}\mathcal{U}^+_{2(0,2)}\\ \mathcal{U}^-_{2(0,2)}\end{pmatrix}p_0^2}_{\mathcal{U}_2}$$
$$+\underbrace{\begin{pmatrix}\mathcal{U}^+_{3(3,0)}\\ \mathcal{U}^-_{3(3,0)}\end{pmatrix}q_0^3+\begin{pmatrix}\mathcal{U}^+_{3(2,1)}\\ \mathcal{U}^-_{3(2,1)}\end{pmatrix}q_0^2p_0+\begin{pmatrix}\mathcal{U}^+_{3(1,2)}\\ \mathcal{U}^-_{3(1,2)}\end{pmatrix}q_0p_0^2+\begin{pmatrix}\mathcal{U}^+_{3(0,3)}\\ \mathcal{U}^-_{3(0,3)}\end{pmatrix}p_0^3}_{\mathcal{U}_3}+\ldots \tag{3.11}$$

The position Q and momentum P components of the map $\mathcal{M}_0$ correspond to the upper and lower component and are denoted by '+' and '-' , respectively. The coefficients in the upper and lower component for the nonlinear $m(=a+b)$th order terms $q^a p^b$ are denoted by $\mathcal{U}^{\pm}_{m(a,b)}$. The coefficients in the linear matrix $(a|b)$ indicate the factor with which a is linearly dependent on b.

The following Tab. 3.2 lists the values of the coefficients in Eq. (3.11) above. The integration was performed with an order 20 Picard-iteration based integrator with stepsize $h = 10^{-3}$ over 1000 iterations within COSY INFINITY. Details on the implementation of the integrator under the name fixed point integrator are given in [86].

3.3 The DA Normal Form Algorithm

In Sec. 2.3, the general DA normal form algorithm [14] was introduced for a linearly stable $2n$ dimensional system. This chapter provides a detailed example-driven walk-through of the differential algebra based normal form algorithm for the symplectic one dimensional (1D) system of the centrifugal governor with a fixed rotation frequency of $\omega = \sqrt{2}$ corresponding to an equilibrium angle of $\phi_0 = 60°$.

The normal form resulting from the DA normal form algorithm constitutes circular motion with a quasi-invariant as radius and only normal form phase space amplitude (and parameter) dependent

Table 3.2: Integration result for map around equilibrium state ($\phi_0(\omega = \sqrt{2}) = \frac{\pi}{3}, 0$) integrated until $t = 1$ using an order 20 Picard-iteration based integrator with stepsize $h = 10^{-3}$ over 1000 iterations within COSY INFINITY. The component $\mathcal{M}_0^+ = Q(q_0, p_0)$ is on the left, $\mathcal{M}_0^- = P(q_0, p_0)$ on the right.

O	Coeff.	Value	Coeff.	Value
0	q_{const}	1.04719755	p_{const}	0
1	$(Q\|q_0)$	0.33918599	$(P\|q_0)$	-1.15214118
1	$(Q\|p_0)$	0.76809412	$(P\|p_0)$	0.33918599
2	$\mathcal{U}^+_{2(2,0)}$	-0.44622446	$\mathcal{U}^-_{2(2,0)}$	-0.55821731
2	$\mathcal{U}^+_{2(1,1)}$	-0.29304415	$\mathcal{U}^-_{2(1,1)}$	-0.64033440
2	$\mathcal{U}^+_{2(0,2)}$	-0.08403817	$\mathcal{U}^-_{2(0,2)}$	-0.29304415
3	$\mathcal{U}^+_{3(3,0)}$	0.31844278	$\mathcal{U}^-_{3(3,0)}$	0.50817317
3	$\mathcal{U}^+_{3(2,1)}$	0.29904862	$\mathcal{U}^-_{3(2,1)}$	0.76091921
3	$\mathcal{U}^+_{3(1,2)}$	0.13758223	$\mathcal{U}^-_{3(1,2)}$	0.46230241
3	$\mathcal{U}^+_{3(0,3)}$	0.03017663	$\mathcal{U}^-_{3(0,3)}$	0.13758223

angle advancements. Fig. 3.5 illustrates the oscillatory phase space behavior of the governor's arms around the equilibrium point (left plot already seen in different orientation in Fig. 3.4) and compares it to its associated rotationally invariant phase space behavior in the normal form representation. The orientation of the phase space in Fig 3.5 is according to the usual convention, where the position q is on the horizontal axis and the momentum p on the vertical axis. In Fig. 3.4, this convention was ignored for the sake of a better understanding when comparing the phase space behavior to the position behavior over time. Accordingly, the asymmetry with larger downwards amplitudes is shown in the horizontal (ϕ) direction in Fig 3.5a.

For the introduction of the DA normal form algorithm, we use the following notation. The starting map (see Tab. 3.2) is dependent on the 'original' variables (q_0, p_0) relative to the expansion point (see Sec. 3.1.2). The transformations of the normal form algorithm are done order by order. With each transformation step, the index of the map and the variables is going to increase by 1, i.e. as a result of the first (order) transformation we get $\mathcal{M}_1$ dependent on the variables (q_1, p_1). For each order m there is a transformation $\mathcal{A}_m$ and its inverse $\mathcal{A}_m^{-1}$, which are applied to resulting map of the previous transformation $\mathcal{M}_{m-1}$ to yield the resulting map of the mth order transformation

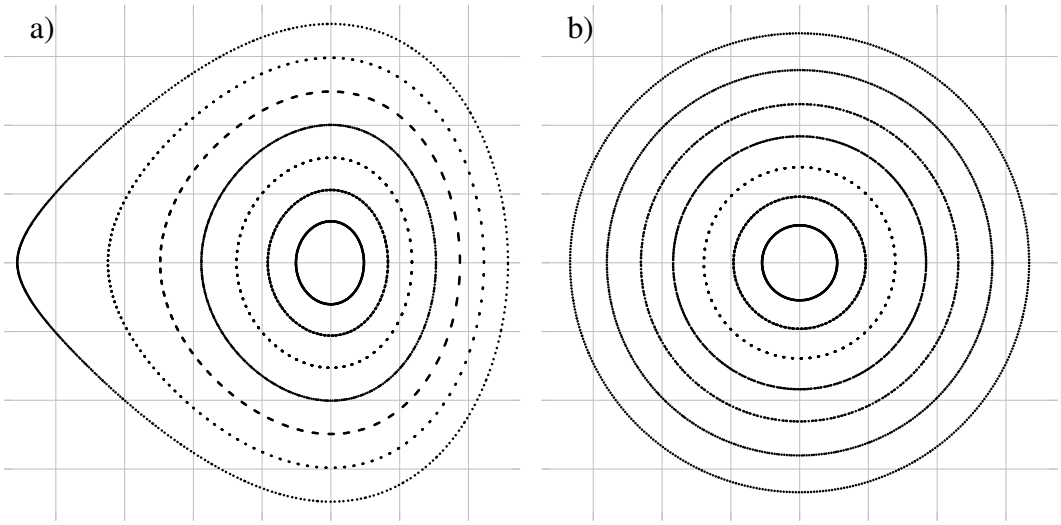

Figure 3.5: Phase space behavior of the centrifugal governor arms around their equilibrium angle of $\phi_0(\omega = \sqrt{2}) = 60°$ provided by a tenth order Poincaré map of the system. a) shows the original phase space behavior. b) shows the associated circular behavior in normal form.

$\mathcal{M}_m(q_m, p_m) = (\mathcal{A}_m \circ \mathcal{M}_{m-1} \circ \mathcal{A}_m^{-1})(q_m, p_m)$.

The transformation $\mathcal{A}_m^{-1}$ transforms (q_m, p_m) to (q_{m-1}, p_{m-1}), which are the variables of the map of the previous order $\mathcal{M}_{m-1}$. The transformation $\mathcal{A}_m$ transforms the intermediate result of $\mathcal{M}_{m-1} \circ \mathcal{A}_m^{-1}$, which is in the (q_{m-1}, p_{m-1}) phase space, back to the new phase space in (q_m, p_m). In [14], the variables (q_m, p_m) are denoted by the (s^+, s^-) notation and the normal form coordinates $(q_{\mathrm{NF}}, p_{\mathrm{NF}})$ are written as (t^+, t^-) instead.

The nonlinear normal form transformation steps below are calculated up to third order. It will become obvious during the process that transformations of higher even and odd orders follow the same pattern as the second and third order transformation, respectively.

3.3.1 The Parameter Dependent Fixed Point

The DA normal form algorithm starts with an origin preserving map. Accordingly, the result from the integration is shifted to the equilibrium/fixed point $\mathcal{M}_{\mathrm{FP}} = \mathcal{M}_0 - \mathcal{C}$, hence $\mathcal{M}_{\mathrm{FP}} = \mathcal{L} + \sum_m \mathcal{U}_m$ is an origin preserving fixed point map with $\mathcal{M}_{\mathrm{FP}}(\vec{0}) = \vec{0}$.

If the map were dependent on changes $\delta\eta$ of a system parameter η, e.g., changes in the driving

frequency $\omega = \omega_0 + \delta\omega$, the normal form algorithm would require the calculation of the parameter dependent fixed point $\vec{z}(\delta\eta) = (q_{\mathrm{FP}}(\delta\eta), p_{\mathrm{FP}}(\delta\eta))$ such that $\mathcal{M}_{\mathrm{FP}}(\vec{0}, \delta\eta) = \vec{0}$. In Eq. (3.3), the relation of the equilibrium point (fixed point) and the driving frequency was already calculated yielding the parameter dependent fixed point

$$\vec{z}\,(\delta\omega) = \left(\arccos\left(\frac{1}{(\omega_0 + \delta\omega)^2}\right), 0\right) \quad \text{for} \quad (\omega_0 + \delta\omega)^2 \geq 1.$$

For less straightforward systems, one uses the following inversion method on the extended map $(\mathcal{M}_{\mathrm{FP}} - \mathcal{I}_{\vec{z}}, \mathcal{I}_{\delta\vec{\eta}})$ to find the parameter dependent fixed point $\vec{z}(\delta\vec{\eta})$ [14, Eq. (7.47)]:

$$\left(\vec{z}\,(\delta\vec{\eta})\,, \mathcal{I}_{\delta\vec{\eta}}\right) = \left(\mathcal{M}_{\mathrm{FP}} - \mathcal{I}_{\vec{z}}, \mathcal{I}_{\delta\vec{\eta}}\right)^{-1} \left(\vec{0}, \delta\vec{\eta}\right), \tag{3.12}$$

where $\mathcal{I}_{\vec{z}}$ and $\mathcal{I}_{\delta\vec{\eta}}$ are the identity map of $\vec{z}$ and $\delta\vec{\eta}$, respectively.

Given the parameter dependent fixed point, the map is expanded around it:

$$\mathcal{M}_{\mathrm{PDFP}} = \mathcal{M}_{\mathrm{FP}}\left(\vec{z}\,(\delta\vec{\eta}) + \vec{z}, \delta\vec{\eta}\right) - \mathcal{M}_{\mathrm{FP}}\left(\vec{z}\,(\delta\vec{\eta})\,, \delta\vec{\eta}\right). \tag{3.13}$$

To limit the complexity of the introduction, we will not consider parameter dependence in the further calculations and therefore proceed with $\mathcal{M}_{\mathrm{FP}}$.

3.3.2 The Linear Transformation

The first order transformation is the diagonalization, transforming the system into the eigenvector space of the linear part $\mathcal{L}$. In order to determine the transformation $\mathcal{A}_1$ and its inverse $\mathcal{A}_1^{-1}$ for the diagonalization, we determine the eigenvalues $\lambda_\pm$ and eigenvectors $\vec{v}_\pm$ of the linear matrix $\hat{L}$ in the linear part $\mathcal{L}$. For this, we require that all eigenvalues of $\mathcal{M}_{\mathrm{FP}}$ are distinct. Furthermore, we only consider cases where $\mathcal{M}_{\mathrm{FP}}$ is linearly stable, which means that all eigenvalues have an absolute value $|\lambda| \leq 1$. This also means that $\det(\hat{L}) \leq 1$, otherwise at least one of the eigenvalues is larger than 1, making the system linearly unstable. Particularly interesting is the case $\det(\hat{L}) = 1$, which indicates that the system is symplectic and only stable in the case of complex conjugate eigenvalues $\lambda_\pm = e^{\pm i\mu}$. While there are procedures for the cases of real and degenerate eigenvalues with a

magnitude smaller than one (see [14]), this chapter only illustrates the procedures for the most relevant and common symplectic case of only complex conjugate eigenvalues and eigenvectors.

Solving the characteristic polynomial yields the eigenvalues

$$\lambda_\pm = \frac{\mathrm{tr}\left(\hat{L}\right)}{2} \pm \sqrt{\frac{\mathrm{tr}\left(\hat{L}\right)^2}{4} - \det\left(\hat{L}\right)} = re^{\pm i\mu}$$
$$\text{with} \quad r = \sqrt{\det\left(\hat{L}\right)} \quad \text{and} \quad \mu = \mathrm{sign}\left(Q|p_0\right) \arccos\left(\frac{\mathrm{tr}\left(\hat{L}\right)}{2r}\right).$$

To generalize the procedure of diagonalization, the Twiss parameters [25] are used with

$$\alpha = \frac{(Q|q_0) - (P|p_0)}{2r\sin\mu} \quad \beta = \frac{(Q|p_0)}{r\sin\mu} \quad \gamma = \frac{-(P|q_0)}{r\sin\mu}.$$

With this notation the linear matrix $\hat{L}$ can be generally written as

$$\hat{L} = \begin{pmatrix} \cos\mu + \alpha\sin\mu & \beta\sin\mu \\ -\gamma\sin\mu & \cos\mu - \alpha\sin\mu \end{pmatrix}.$$

The complex conjugate eigenvectors $\vec{v}_\pm$ associated with the complex conjugate eigenvalues $\lambda_\pm$ of $\hat{L}$ are then obtained by solving $\left(\hat{L} - \lambda_\pm \mathcal{I}\right)\vec{v}_\pm = \vec{0}$.

As a result, the following eigenvectors are calculated

$$\vec{v}_\pm = \begin{pmatrix} \beta \\ -\alpha \pm i \end{pmatrix} \quad \text{or} \quad \vec{v}_\pm = \begin{pmatrix} \alpha \pm i \\ -\gamma \end{pmatrix},$$

for the case that either $\beta = 0$ or $\gamma = 0$. The transformation $\mathcal{A}_1^{-1}$ consist of the two complex conjugate eigenvectors $\vec{v}_\pm$, guaranteeing that $\mathcal{A}_1^{-1}(q_1, p_1)$ is real just like the original variables (q_0, p_0) and the fixed point map $\mathcal{M}_{\mathrm{FP}}$. The transformation $\mathcal{A}_1$ is calculated accordingly such that the resulting map $\mathcal{M}_1 = \mathcal{A}_1 \circ \mathcal{M}_{\mathrm{FP}} \circ \mathcal{A}_1^{-1}$ is in the complex conjugate eigenvector space and has complex conjugate components $\bar{\mathcal{M}}_1^+ = \mathcal{M}_1^-$. For $\beta \neq 0$, the transformations are

$$\mathcal{A}_1^{-1} = \begin{pmatrix} (q_0|q_1) & (q_0|p_1) \\ (p_0|q_1) & (p_0|p_1) \end{pmatrix} = \frac{1}{2\sqrt{\beta}} \begin{pmatrix} \beta & \beta \\ i - \alpha & -i - \alpha \end{pmatrix} \tag{3.14}$$

$$\mathcal{A}_1 = \begin{pmatrix} (q_1|q_0) & (q_1|p_0) \\ (p_1|q_0) & (p_1|p_0) \end{pmatrix} = \frac{i}{\sqrt{\beta}} \begin{pmatrix} -i - \alpha & -\beta \\ -i + \alpha & \beta \end{pmatrix}. \tag{3.15}$$

For the centrifugal governor example with $\omega = \sqrt{2}$, the eigenvalues are $\lambda_\pm = re^{\pm i\mu}$ with $r = 1$ and $\mu = 1.22474487$. The Twiss parameters are

$$\alpha = 0 \quad \beta = 0.816496581 \approx \sqrt{\frac{2}{3}} \quad \gamma = 1.22474487 \approx \sqrt{\frac{3}{2}}.$$

The resulting diagonalized map is of the form $\mathcal{M}_1 = \mathcal{R} + \sum_m S_m$, where $\mathcal{S}_m$ are the transformed nonlinear parts of order m in the eigenvector space of $\hat{L}$ and $\mathcal{R}$ is the diagonalized linear part, where the linear matrix $\hat{R}$ of $\mathcal{R}$ only consist of the eigenvalues $e^{\pm i\mu}$ on its main diagonal:

$$\begin{aligned}\mathcal{M}_1(q_1, p_1) = &\underbrace{\begin{pmatrix} e^{i\mu} & 0 \\ 0 & e^{-i\mu} \end{pmatrix}\begin{pmatrix} q_1 \\ p_1 \end{pmatrix}}_{\mathcal{R}} + \underbrace{\begin{pmatrix} \mathcal{S}^+_{2(2,0)} \\ \mathcal{S}^-_{2(2,0)} \end{pmatrix} q_1^2 + \begin{pmatrix} \mathcal{S}^+_{2(1,1)} \\ \mathcal{S}^-_{2(1,1)} \end{pmatrix} q_1 p_1 + \begin{pmatrix} \mathcal{S}^+_{2(0,2)} \\ \mathcal{S}^-_{2(0,2)} \end{pmatrix} p_1^2}_{\mathcal{S}_2} \\ &+ \underbrace{\begin{pmatrix} \mathcal{S}^+_{3(3,0)} \\ \mathcal{S}^-_{3(3,0)} \end{pmatrix} q_1^3 + \begin{pmatrix} \mathcal{S}^+_{3(2,1)} \\ \mathcal{S}^-_{3(2,1)} \end{pmatrix} q_1^2 p_1 + \begin{pmatrix} \mathcal{S}^+_{3(1,2)} \\ \mathcal{S}^-_{3(1,2)} \end{pmatrix} q_1 p_1^2 + \begin{pmatrix} \mathcal{S}^+_{3(0,3)} \\ \mathcal{S}^-_{3(0,3)} \end{pmatrix} p_0^3}_{\mathcal{S}_3} + \ldots \end{aligned} \tag{3.16}$$

Tab. 3.3 lists the values to the coefficients above for the centrifugal governor example for a rotation frequency corresponding to an equilibrium angle of $\phi_0(\omega = \sqrt{2}) = \frac{\pi}{3} = 60°$.

3.3.3 The Nonlinear Transformations

The nonlinear transformations are the key steps of the normal form algorithm. In this first part of this subsection, we are going to look at an mth order transformation in general, before going through the nonlinear transformation for orders two and three in detail.

3.3.3.1 General mth Order Nonlinear Transformation

All the following nonlinear transformation steps are done order by order and are all of the same form: $\mathcal{M}_m = \mathcal{A}_m \circ \mathcal{M}_{m-1} \circ \mathcal{A}_m^{-1}$, where the mth transformation does **not** change any of the lower order terms of $\mathcal{M}_{m-1}$ that have already been transformed in the previous transformations. Hence, $\mathcal{M}_m$ differs from $\mathcal{M}_{m-1}$ only in the orders m and larger. The mth order transformation

Table 3.3: Coefficients of $\mathcal{M}_1$ up to order three. Note the complex conjugate property $\mathcal{S}^{\pm}_{m(k_+,k_-)} = \bar{\mathcal{S}}^{\mp}_{m(k_-,k_+)}$.

O	Coeff.	Real Part	Imaginary Part
1	$e^{i\mu}$	0.339185989	0.940719334
1	$e^{-i\mu}$	0.339185989	-0.940719334
2	$\mathcal{S}^{+}_{2(2,0)}$	-0.216977793	-0.059191831
2	$\mathcal{S}^{-}_{2(2,0)}$	0.072325931	-0.102961500
2	$\mathcal{S}^{+}_{2(1,1)}$	-0.258557455	0.368076331
2	$\mathcal{S}^{-}_{2(1,1)}$	-0.258557455	-0.368076331
2	$\mathcal{S}^{+}_{2(0,2)}$	0.072325931	0.102961500
2	$\mathcal{S}^{-}_{2(0,2)}$	-0.216977793	0.059191831
3	$\mathcal{S}^{+}_{3(3,0)}$	0.068036138	0.047162997
3	$\mathcal{S}^{-}_{3(3,0)}$	-0.045160062	-0.016282923
3	$\mathcal{S}^{+}_{3(2,1)}$	0.259415349	-0.130475661
3	$\mathcal{S}^{-}_{3(2,1)}$	-0.022283986	0.239186527
3	$\mathcal{S}^{+}_{3(1,2)}$	-0.022283986	-0.239186527
3	$\mathcal{S}^{-}_{3(1,2)}$	0.259415349	0.130475661
3	$\mathcal{S}^{+}_{3(0,3)}$	-0.045160062	-0.016282923
3	$\mathcal{S}^{-}_{3(0,3)}$	0.068036138	-0.047162997

$\mathcal{A}_m = \mathcal{I} + \mathcal{T}_m + \mathcal{O}_{\geq m+1}$, specifically the polynomial $\mathcal{T}_m$ of only mth order terms, is chosen such that the mth order terms $\mathcal{S}_m$ of the map $\mathcal{M}_{m-1}$ are simplified or even eliminated.

Effects on the higher orders of $\mathcal{M}_m$ due to the mth order transformation can only be considered by adjusting the terms of order higher than m of $\mathcal{A}_m$, namely $\mathcal{O}_{\geq m+1}$. In other words, finding $\mathcal{T}_m$ is essential to the DA normal form algorithm, while the terms $\mathcal{O}_{\geq m+1}$ can be chosen freely, e.g., to make the transformation symplectic by choosing $\mathcal{A}_m = \exp(L_{\mathcal{T}_m})$ or to avoid higher order resonances. Usually, the symplectic transformation is chosen since the calculation of the transformation $\mathcal{A}_m$ and its inverse are straightforward.

The flow operator $L_{\mathcal{T}_m} = (\mathcal{T}_m^+ \partial_q + \mathcal{T}_m^- \partial_p)$ in the exponential behaves in the following way:

$$\begin{aligned}
\exp\left(L_{\mathcal{T}_m}\right)\mathcal{I} &= \left(L^0_{\mathcal{T}_m} + L^1_{\mathcal{T}_m} + \frac{1}{2}L^2_{\mathcal{T}_m} + \mathcal{O}_{>(m+1)}\right)\mathcal{I} \\
&= \left(1 + (\mathcal{T}_m^+ \partial_q + \mathcal{T}_m^- \partial_p) + \frac{1}{2}L_{\mathcal{T}_m}(\mathcal{T}_m^+ \partial_q + \mathcal{T}_m^- \partial_p) + \mathcal{O}_{>(m+1)}\right)(q,p)^T \\
&= \mathcal{I} + \mathcal{T}_m + \frac{1}{2}L_{\mathcal{T}_m}\mathcal{T}_m + \mathcal{O}_{>(m+1)}. \qquad (3.17)
\end{aligned}$$

Accordingly, the inverse is given by

$$\mathcal{A}_m^{-1} = \exp\left(-L_{\mathcal{T}_m}\right) = \mathcal{I} - \mathcal{T}_m + \frac{1}{2}L_{\mathcal{T}_m}\mathcal{T}_m - \mathcal{O}_{>(m+1)}. \qquad (3.18)$$

In the example case of the centrifugal governor, we investigate the DA normal form algorithm up to order three, which means for $m = 3$:

$$\mathcal{A}_3 = \exp\left(L_{\mathcal{T}_3}\right)\mathcal{I} =_3 \mathcal{I} + \mathcal{T}_3 \qquad (3.19)$$

$$\mathcal{A}_3^{-1} = \exp\left(-L_{\mathcal{T}_3}\right)\mathcal{I} =_3 \mathcal{I} - \mathcal{T}_3. \qquad (3.20)$$

For the second order transformation it is necessary to consider the third order terms $\mathcal{O}_3$, since they influence the third order terms of $\mathcal{M}_2$:

$$\mathcal{A}_2 = \exp\left(L_{\mathcal{T}_2}\right)\mathcal{I} =_3 \mathcal{I} + \mathcal{T}_2 + \mathcal{O}_3 \qquad (3.21)$$

$$\mathcal{A}_2^{-1} = \exp\left(-L_{\mathcal{T}_2}\right)\mathcal{I} =_3 \mathcal{I} - \mathcal{T}_2 + \mathcal{O}_3 \qquad (3.22)$$

with

$$\mathcal{O}_3 = \frac{1}{2}L_{\mathcal{T}_m}\mathcal{T}_m = \frac{1}{2}(\mathcal{T}_2^+ \partial_q + \mathcal{T}_2^- \partial_p)\mathcal{T}_2. \qquad (3.23)$$

As introduced in Sec. 2.1, the notation '$=_m$'indicates that the quantities on both sides are equal up to expansion order m.

In order to determine $\mathcal{T}_m$, we analyze the mth order transformation and only look at terms up to order m [14, Eq. (7.62)]:

$$\begin{aligned}
\mathcal{A}_m \circ \mathcal{M}_{m-1} \circ \mathcal{A}_m^{-1} &=_m (\mathcal{I} + \mathcal{T}_m) \circ (\mathcal{R} + \mathcal{S}_m) \circ (\mathcal{I} - \mathcal{T}_m) \\
&=_m (\mathcal{I} + \mathcal{T}_m) \circ (\mathcal{R} - \mathcal{R} \circ \mathcal{T}_m + \mathcal{S}_m) \\
&=_m \mathcal{R} + \mathcal{S}_m + [\mathcal{T}_m, \mathcal{R}]\,. \qquad (3.24)
\end{aligned}$$

Various terms with orders higher than m are ignored in the equations above. The goal is to choose $\mathcal{T}_m$ such that the commutator $[\mathcal{T}_m, \mathcal{R}] = \mathcal{T}_m \circ \mathcal{R} - \mathcal{R} \circ \mathcal{T}_m = -\mathcal{S}_m$ to simplify $\mathcal{M}_m$, i.e. the result of Eq. (3.24). The polynomials in the upper and lower component of $\mathcal{T}_m$ can be express as

$$\mathcal{T}_m^{\pm}(q,p) = \sum_{\substack{m=k_+ + k_- \\ k_\pm \in \mathbb{N}_0}} \mathcal{T}^{\pm}_{m(k_+,k_-)} q^{k_+} p^{k_-}. \tag{3.25}$$

Accordingly, the commutator $\mathcal{C}_m = [\mathcal{T}_m, \mathcal{R}]$ yields

$$\mathcal{C}_m^{\pm}(q,p) = \sum_{\substack{m=k_+ + k_- \\ k_\pm \in \mathbb{N}_0}} \mathcal{T}^{\pm}_{m(k_+,k_-)} \left(e^{i\mu(k_+ - k_-)} - e^{\pm i\mu} \right) q^{k_+} p^{k_-}. \tag{3.26}$$

A term in $\mathcal{S}_m$ can only be removed if and only if the corresponding term in the commutator $\mathcal{C}_m$ is not zero. Terms of the commutator are zero, whenever the condition

$$e^{i\mu(k_+ - k_-)} - e^{\pm i\mu} = 0 \tag{3.27}$$

is satisfied, which is the case for $k_+ - k_- = \pm 1$. This (Eq. (3.27)) is the key condition of the DA normal form algorithm, since it determines the surviving nonlinear terms $\mathcal{S}_m$. All other terms that do not satisfy the condition are eliminated by choosing the coefficients of $\mathcal{T}_m$ as follows

$$\mathcal{T}^{\pm}_{m(k_+,k_-)} = \frac{-S^{\pm}_{m(k_+,k_-)}}{e^{i\mu(k_+ - k_-)} - e^{\pm i\mu}}. \tag{3.28}$$

Specifically, this means that the terms $\mathcal{S}^{+}_{m(k,k-1)}$ and $\mathcal{S}^{-}_{m(k-1,k)}$ always survive for all uneven orders m with $m = k + k - 1 = 2k - 1$.

3.3.3.2 Explicit Second Order Nonlinear Transformation

The polynomial $\mathcal{T}_m$ from Eq. (3.25) for $m = 2$ yields

$$\begin{aligned} \mathcal{T}_2(q,p) &= \left(\mathcal{T}_2^{\pm}|2,0\right) q^2 + \left(\mathcal{T}_2^{\pm}|1,1\right) qp + \left(\mathcal{T}_2^{\pm}|0,2\right) p^2 \\ &= \begin{pmatrix} \mathcal{T}^{+}_{2(2,0)} \\ \mathcal{T}^{-}_{2(2,0)} \end{pmatrix} q^2 + \begin{pmatrix} \mathcal{T}^{+}_{2(1,1)} \\ \mathcal{T}^{-}_{2(1,1)} \end{pmatrix} qp + \begin{pmatrix} \mathcal{T}^{+}_{2(0,2)} \\ \mathcal{T}^{-}_{2(0,2)} \end{pmatrix} p^2. \end{aligned} \tag{3.29}$$

The commutator $\mathcal{C}_2 = [\mathcal{T}_2, \mathcal{R}] = \mathcal{T}_2 \circ \mathcal{R} - \mathcal{R} \circ \mathcal{T}_2$ of the second order nonlinear transformation has only nonzero terms with

$$\begin{aligned}
\mathcal{C}_2(q,p) &= [\mathcal{T}_2, \mathcal{R}](q,p) = (\mathcal{T}_2 \circ \mathcal{R} - \mathcal{R} \circ \mathcal{T}_2)(q,p) \\
&= \left(\mathcal{T}_2^{\pm}|2,0\right) e^{2i\mu} q^2 + \left(\mathcal{T}_2^{\pm}|1,1\right) qp + \left(\mathcal{T}_2^{\pm}|0,2\right) e^{-2i\mu} p^2 - e^{\pm i\mu} \mathcal{T}_2^{\pm}(q,p) \\
&= \begin{pmatrix} \mathcal{T}^{+}_{2(2,0)} \left(e^{2i\mu} - e^{i\mu}\right) \\ \mathcal{T}^{-}_{2(2,0)} \left(e^{2i\mu} - e^{-i\mu}\right) \end{pmatrix} q^2 + \begin{pmatrix} -e^{i\mu} \mathcal{T}^{+}_{2(1,1)} \\ -e^{-i\mu} \mathcal{T}^{-}_{2(1,1)} \end{pmatrix} qp + \begin{pmatrix} \mathcal{T}^{+}_{2(0,2)} \left(e^{-2i\mu} - e^{i\mu}\right) \\ \mathcal{T}^{-}_{2(0,2)} \left(e^{-2i\mu} - e^{-i\mu}\right) \end{pmatrix} p^2
\end{aligned} \tag{3.30}$$

eliminating all $\mathcal{S}_2$ terms by choosing

$$\mathcal{T}^{\pm}_{2(k_+,k_-)} = \frac{-S^{\pm}_{2(k_+,k_-)}}{\left(e^{i\mu(k_+-k_-)} - e^{\pm i\mu}\right)}, \tag{3.31}$$

since the condition from Eq. (3.27) is not satisfied:

$$e^{i\mu(k_+-k_-)} - e^{\pm i\mu} \neq 0 \quad \forall k_+, k_- \in \mathbb{N}_0 \quad \text{with} \quad k_+ + k_- = 2.$$

The values of the $\mathcal{T}^{\pm}_{2(k_+,k_-)}$ for the centrifugal governor example are given in Tab. 3.4. The terms of $\mathcal{O}_3$ are calculated via Eq. (3.23) from $\mathcal{T}_2$ and are also given in Tab. 3.4 yielding all terms of the transformation $\mathcal{A}_2$ and its inverse $\mathcal{A}_2^{-1}$ from Eq. (3.21) and Eq. (3.22).

Table 3.4: The values of the $\mathcal{T}^{\pm}_{2(k_+,k_-)}$ and $\mathcal{O}^{\pm}_{3(k_+,k_-)}$. Note that $\mathcal{T}_2$ and $\mathcal{O}_3$ and therefore $\mathcal{A}_2$ and its inverse are real with $\mathcal{A}^{+}_{m(k_+,k_-)} = \mathcal{A}^{-}_{m(k_-,k_+)}$.

O	Coeff.	Value	Coeff.	Value
2	$\mathcal{T}^{+}_{2(2,0)}$	-0.195635573	$\mathcal{T}^{-}_{2(2,0)}$	0.065211858
2	$\mathcal{T}^{+}_{2(1,1)}$	0.391271145	$\mathcal{T}^{-}_{2(1,1)}$	0.391271145
2	$\mathcal{T}^{+}_{2(0,2)}$	0.065211858	$\mathcal{T}^{-}_{2(0,2)}$	-0.195635573
3	$\mathcal{O}^{+}_{3(3,0)}$	0.051031036	$\mathcal{O}^{-}_{3(3,0)}$	0
3	$\mathcal{O}^{+}_{3(2,1)}$	-0.034020691	$\mathcal{O}^{-}_{3(2,1)}$	0.051031036
3	$\mathcal{O}^{+}_{3(1,2)}$	0.051031036	$\mathcal{O}^{-}_{3(1,2)}$	-0.034020691
3	$\mathcal{O}^{+}_{3(0,3)}$	0	$\mathcal{O}^{-}_{3(0,3)}$	0.051031036

To study how the second order transformation affects the third order terms $\mathcal{S}_3$ of the map $\mathcal{M}_2$, the transformation is considered up to third order:

$$
\begin{aligned}
\mathcal{M}_2 &=_3 \mathcal{A}_2 \circ \mathcal{M}_1 \circ \mathcal{A}_2^{-1} \\
&=_3 (\mathcal{I} + \mathcal{T}_2 + \mathcal{O}_3) \circ (\mathcal{R} + \mathcal{S}_2 + \mathcal{S}_3) \circ (\mathcal{I} - \mathcal{T}_2 + \mathcal{O}_3) \\
&=_3 (\mathcal{I} + \mathcal{T}_2 + \mathcal{O}_3) \circ \Big(\mathcal{R} \circ (\mathcal{I} - \mathcal{T}_2 + \mathcal{O}_3) + \underline{\mathcal{S}_2 \circ (\mathcal{I} - \mathcal{T}_2 + \mathcal{O}_3)} + \underline{\mathcal{S}_3 \circ (\mathcal{I} - \mathcal{T}_2 + \mathcal{O}_3)} \Big) \\
&=_3 (\mathcal{I} + \mathcal{T}_2 + \mathcal{O}_3) \circ \left(\mathcal{R} - \mathcal{R} \circ \mathcal{T}_2 + \mathcal{R} \circ \mathcal{O}_3 + \overbrace{\mathcal{S}_2 + \mathcal{S}_{2\to3} + \cancel{\mathcal{O}_{\geq 4}}}^{} + \overbrace{\mathcal{S}_3 + \cancel{\mathcal{O}_{\geq 4}}}^{} \right) \\
&=_3 \mathcal{R} - \mathcal{R} \circ \mathcal{T}_2 + \mathcal{R} \circ \mathcal{O}_3 + \mathcal{S}_2 + \mathcal{S}_{2\to3} + \mathcal{S}_3 \\
&\quad + \underline{\mathcal{T}_2 \circ (\mathcal{R} - \mathcal{R} \circ \mathcal{T}_2 + \mathcal{R} \circ \mathcal{O}_3 + \mathcal{S}_2 + \mathcal{S}_{2\to3} + \mathcal{S}_3)} + \cancel{\mathcal{O}_{\geq 4}} \\
&=_3 \overbrace{\mathcal{T}_2 \circ \mathcal{R} + \mathcal{K}_{2\to3} + \cancel{\mathcal{O}_{\geq 4}}}^{} + \mathcal{R} - \mathcal{R} \circ \mathcal{T}_2 + \mathcal{R} \circ \mathcal{O}_3 + \mathcal{S}_2 + \mathcal{S}_{2\to3} + \mathcal{S}_3 \\
&=_3 \mathcal{R} + \underbrace{\mathcal{S}_2 + [\mathcal{T}_2 \circ \mathcal{R}]}_{=0} + \underbrace{\mathcal{S}_3 + \mathcal{S}_{2\to3} + \mathcal{K}_{2\to3} + \mathcal{R} \circ \mathcal{O}_3}_{\mathcal{S}_{3,\text{new}}}
\end{aligned}
\tag{3.32}
$$

All the crossed-out terms $\cancel{\mathcal{O}_{\geq 4}}$ represent terms that do not contribute to the result up to order three, since they are at least of order four. As a result of the second order transformation, the third order terms have changed and are summarized by $\mathcal{S}_{3,\text{new}}$. They are composed of the third order terms from after the linear transformation $\mathcal{S}_3$ and three new terms: $\mathcal{S}_{2\to3} =_3 \mathcal{S}_2 \circ (\mathcal{I} - \mathcal{T}_2) - \mathcal{S}_2$, $\mathcal{K}_{2\to3} =_3 \mathcal{T}_2 \circ (\mathcal{R} - \mathcal{R} \circ \mathcal{T}_2 + \mathcal{S}_2) - \mathcal{T}_2 \circ \mathcal{R}$ and $\mathcal{R} \circ \mathcal{O}_3$. While the last one is self-explanatory, the first two are not intuitively understood. In Sec. 3.3.3.4 these terms are calculated more explicitly, however, we recommend this section only for the very intrigued reader and encourage everyone else to skip it to follow the steps in the normal form algorithm.

The result of the second order transformation $\mathcal{M}_2 = \mathcal{R} + \mathcal{S}_{3,\text{new}}$ for the example case of the centrifugal governor is given in Tab. 3.5.

Table 3.5: New coefficients of third order of $\mathcal{M}_2$ after the second order transformation. Note that the first order terms remain unchanged and that the second order terms are all eliminated by the second order transformation. Interestingly, the second order transformation caused some terms of the third order to disappear in this specific case, this is not a general property of the second order transformation. The emphasized terms are surviving the third order transformation as explained in the following subsection.

O	Coeff.	Real Part	Imaginary Part
3	$\mathcal{S}^{+}_{3,\mathrm{new}(3,0)}$	0.061270641	0.073920008
3	$\mathcal{S}^{-}_{3,\mathrm{new}(3,0)}$	0	0
3	$\mathcal{S}^{+}_{3,\mathrm{new}(2,1)}$	**0.470359667**	**-0.169592994**
3	$\mathcal{S}^{-}_{3,\mathrm{new}(2,1)}$	0	0.288035295
3	$\mathcal{S}^{+}_{3,\mathrm{new}(1,2)}$	0	-0.288035295
3	$\mathcal{S}^{-}_{3,\mathrm{new}(1,2)}$	**0.470359667**	**0.169592994**
3	$\mathcal{S}^{+}_{3,\mathrm{new}(0,3)}$	0	0
3	$\mathcal{S}^{-}_{3,\mathrm{new}(0,3)}$	0.061270641	-0.073920008

3.3.3.3 Explicit Third Order Nonlinear Transformation

The third order transformation follows the same scheme as above (see Eq. (3.24)) only that the commutator $\mathcal{C}_3 = [\mathcal{T}_3 \circ \mathcal{R}]$ has terms that are zero

$$\mathcal{C}_3 = \begin{pmatrix} \mathcal{T}^{+}_{3(3,0)} \left(e^{3i\mu} - e^{i\mu}\right) \\ \mathcal{T}^{-}_{3(3,0)} \left(e^{3i\mu} - e^{-i\mu}\right) \end{pmatrix} q^3 + \begin{pmatrix} 0 \\ \mathcal{T}^{-}_{3(2,1)} \left(e^{i\mu} - e^{-i\mu}\right) \end{pmatrix} q^2 p + \begin{pmatrix} \mathcal{T}^{+}_{3(1,2)} \left(e^{-i\mu} - e^{i\mu}\right) \\ 0 \end{pmatrix} q p^2 + \begin{pmatrix} \mathcal{T}^{+}_{3(0,3)} \left(e^{-3i\mu} - e^{i\mu}\right) \\ \mathcal{T}^{-}_{3(0,3)} \left(e^{-3i\mu} - e^{-i\mu}\right) \end{pmatrix} p^3, \tag{3.33}$$

with $\mathcal{C}^{+}_{3(2,1)} = \mathcal{C}^{-}_{3(1,2)} = 0$. This means that the terms $\mathcal{S}^{+}_{3,\mathrm{new}(2,1)}$ and $\mathcal{S}^{-}_{3,\mathrm{new}(1,2)}$ cannot be eliminated. All the other terms are eliminated by choosing

$$\mathcal{T}^{\pm}_{3(k_+,k_-)} = \frac{-S^{\pm}_{3,\mathrm{new}(k_+,k_-)}}{\left(e^{i\mu(k_+-k_-)} - e^{\pm i\mu}\right)} \quad \text{for} \quad k_+ - k_- \neq \pm 1. \tag{3.34}$$

The values of $\mathcal{T}^{\pm}_{3(k_+,k_-)}$ for the centrifugal governor example are given in Tab. 3.6.

Table 3.6: The values of the $\mathcal{T}^{\pm}_{3(k_+,k_-)}$. Note that $\mathcal{T}^{+}_{3(k_+,k_-)} = \mathcal{T}^{-}_{3(k_-,k_+)}$.

O	Coeff.	Value	Coeff.	Value
3	$\mathcal{T}^{+}_{3(3,0)}$	0.051031036	$\mathcal{T}^{-}_{3(3,0)}$	0
3	$\mathcal{T}^{+}_{3(2,1)}$	0	$\mathcal{T}^{-}_{3(2,1)}$	-0.153093109
3	$\mathcal{T}^{+}_{3(1,2)}$	-0.153093109	$\mathcal{T}^{-}_{3(1,2)}$	0
3	$\mathcal{T}^{+}_{3(0,3)}$	0	$\mathcal{T}^{-}_{3(0,3)}$	0.051031036

After the third order transformation the resulting map is of the following form

$$\begin{aligned}\mathcal{M}_3 &= \underbrace{\begin{pmatrix} e^{i\mu} & 0 \\ 0 & e^{-i\mu} \end{pmatrix}\begin{pmatrix} q_3 \\ p_3 \end{pmatrix}}_{\mathcal{R}} + \underbrace{\begin{pmatrix} \mathcal{S}^{+}_{3,\mathrm{new}(2,1)} \\ 0 \end{pmatrix} q_3^2 p_3 + \begin{pmatrix} 0 \\ \mathcal{S}^{-}_{3,\mathrm{new}(1,2)} \end{pmatrix} q_3 p_3^2}_{\mathcal{S}_{3,\mathrm{transformed}}} \\ &= \begin{pmatrix} \left(e^{i\mu} + \mathcal{S}^{+}_{3,\mathrm{new}(2,1)} q_3 p_3\right) q_3 \\ \left(e^{-i\mu} + \mathcal{S}^{-}_{3,\mathrm{new}(1,2)} q_3 p_3\right) p_3 \end{pmatrix} = \begin{pmatrix} f^{+}\left(q_3 p_3\right) q_3 \\ f^{-}\left(q_3 p_3\right) p_3 \end{pmatrix}. \end{aligned} \tag{3.35}$$

The corresponding values for the coefficients can be found in Tab. 3.3 for the linear terms and in Tab. 3.5 for the third order terms. The complex conjugate property of the map $\mathcal{M}^{+}_{3} = \overline{\mathcal{M}_3}^{-}$ is maintained.

While all nonlinear transformations follow the same structure, there is a fundamental difference between even and odd order transformation steps. For even order transformations there are no regularly surviving terms as shown for the second order transformation. For uneven order transformations, there are some terms of a special structure that do survive as shown for third order transformation. Higher even and odd order transformations will behave in the same way, which is why we will stop the process of the detailed walk-through here, after the third order transformation. In principle, the calculation of the transformations can be continued up to arbitrary order. With each transformation, the higher order terms are change and in the end only the terms $\mathcal{S}^{+}_{m(k,k-1)}$ and $\mathcal{S}^{-}_{m(k-1,k)}$ of uneven orders survive. Hence, the components $\mathcal{M}^{\pm}_{m}$ can also be factorize into the $f^{\pm}\left(q_m p_m\right)$ notation (see Eq. (3.35)) for higher orders.

3.3.3.4 The Effect of the Second Order Transformation on Third Order Terms

The following calculation investigates the term $\mathcal{S}_{2\to3}$ as was previously done in [86] and was added here for sake of completeness.

$$\begin{aligned}
\mathcal{S}_{2\to3} =_3\ & \mathcal{S}_2 \circ (\mathcal{I} - \mathcal{T}_2) - \mathcal{S}_2 \\
=_3\ & \mathcal{S}_{2(2,0)} \left(q - \mathcal{T}_2^+\right)^2 + \mathcal{S}_{2(0,2)} \left(p - \mathcal{T}_2^-\right)^2 + \mathcal{S}_{2(1,1)} \left(q - \mathcal{T}_2^+\right) \left(p - \mathcal{T}_2^-\right) - \mathcal{S}_2 \\
=_3\ & \underbrace{\mathcal{S}_{2(2,0)} q^2 + \mathcal{S}_{2(1,1)} qp + \mathcal{S}_{2(0,2)} p^2 - \mathcal{S}_2}_{=0} \\
& + \underbrace{\mathcal{S}_{2(2,0)} \left(\mathcal{T}_2^+\right)^2 + \mathcal{S}_{2(1,1)} \mathcal{T}_2^+ \mathcal{T}_2^- + \mathcal{S}_{2(0,2)} \left(\mathcal{T}_2^-\right)^2}_{\geq \mathcal{O}_4} \\
& - 2\mathcal{S}_{2(2,0)} \mathcal{T}_2^+ q - \mathcal{S}_{2(1,1)} \left(\mathcal{T}_2^+ p + \mathcal{T}_2^- q\right) - 2\mathcal{S}_{2(0,2)} \mathcal{T}_2^- p
\end{aligned} \tag{3.36}$$

As derived in the beginning of Sec. 3.3.3.3, the surviving parts of $\mathcal{S}_{2\to3}$ after the third order transformation are $\mathcal{S}^+_{2\to3(2,1)}$ and its complex conjugate counterpart $\mathcal{S}^-_{2\to3(1,2)}$:

$$\begin{aligned}
\mathcal{S}^+_{2\to3(2,1)} &= -2\mathcal{S}^+_{2(2,0)} \mathcal{T}^+_{2(1,1)} - \mathcal{S}^+_{2(1,1)} \left(\mathcal{T}^+_{2(2,0)} + \mathcal{T}^-_{2(1,1)}\right) - 2\mathcal{S}^+_{2(0,2)} \mathcal{T}^-_{2(2,0)} \\
&= \frac{2\mathcal{S}^+_{2(2,0)} \mathcal{S}^+_{2(1,1)}}{1 - e^{i\mu}} + \frac{\mathcal{S}^+_{2(1,1)} \mathcal{S}^+_{2(2,0)}}{e^{2i\mu} - e^{i\mu}} + \frac{\mathcal{S}^+_{2(1,1)} \mathcal{S}^-_{2(1,1)}}{1 - e^{-i\mu}} + \frac{\mathcal{S}^+_{2(0,2)} \mathcal{S}^-_{2(2,0)}}{e^{2i\mu} - e^{-i\mu}}
\end{aligned} \tag{3.37}$$

This illustrates the complexity of these terms since every single term from $\mathcal{S}_2$ is relevant for them. Each term of $\mathcal{S}_2$ is again dependent on the terms of $\mathcal{U}_2$. The relation is given by the linear transformation $\mathcal{S}_2 = \mathcal{A}_1 \circ \mathcal{U}_2 \circ \mathcal{A}_1^{-1}$. In principle, one can extend the calculation above to express $\mathcal{S}^+_{2\to3(2,1)}$ in terms of $\mathcal{U}_2$ and the Twiss parameters as done in [86]. The main insight however is that due to the significant influence of lower order transformation on higher order terms it is almost impossible to determine a priory which terms are the relevant ones for characteristics of the normal form.

In the following calculation we are investigating the term $\mathcal{K}_{2\to3}$, which was not previously

investigated in [86].

$$
\begin{aligned}
\mathcal{K}_{2\to 3} &= \mathcal{T}_2 \circ (\mathcal{R} - \mathcal{R} \circ \mathcal{T}_2 + \mathcal{S}_2) - \mathcal{T}_2 \circ \mathcal{R} \\
&= \mathcal{T}_2 \circ (\mathcal{R} - \mathcal{K}_2) - \mathcal{T}_2 \circ \mathcal{R} \\
&=_3 \mathcal{T}_{2(2,0)} \left(e^{i\mu} q - \mathcal{K}_2^+\right)^2 + \mathcal{T}_{2(0,2)} \left(e^{-i\mu} p - \mathcal{K}_2^-\right)^2 \\
&\quad + \mathcal{T}_{2(1,1)} \left(e^{i\mu} q - \mathcal{K}_2^+\right) \left(e^{-i\mu} p - \mathcal{K}_2^-\right) - \mathcal{T}_2 \circ \mathcal{R} \\
&=_3 \underbrace{\cancel{\mathcal{T}_{2(2,0)} e^{2i\mu} q^2 + \mathcal{T}_{2(1,1)} qp + \mathcal{T}_{2(0,2)} e^{-2i\mu} p^2} - \mathcal{T}_2 \circ \mathcal{R}}_{=0} \\
&\quad + \underbrace{\cancel{\mathcal{T}_{2(2,0)} \left(\mathcal{K}_2^+\right)^2 + \mathcal{T}_{2(1,1)} \mathcal{K}_2^+ \mathcal{K}_2^- + \mathcal{T}_{2(0,2)} \left(\mathcal{K}_2^-\right)^2}}_{\geq \mathcal{O}_4} \\
&\quad - 2\mathcal{T}_{2(2,0)} \mathcal{K}_2^+ e^{i\mu} q - \mathcal{T}_{2(1,1)} \left(\mathcal{K}_2^+ e^{-i\mu} p + \mathcal{K}_2^- e^{i\mu} q\right) - 2\mathcal{T}_{2(0,2)} \mathcal{K}_2^- e^{-i\mu} p
\end{aligned} \tag{3.38}
$$

where

$$
\mathcal{K}_2 = \mathcal{R} \circ \mathcal{T}_2 - \mathcal{S}_2 \qquad \to \qquad \mathcal{K}_2^\pm = e^{\pm i\mu} \mathcal{T}_2^\pm - \mathcal{S}_2^\pm \tag{3.39}
$$

so

$$
\begin{aligned}
\mathcal{K}_{2\to 3} =_3\ & 2\mathcal{T}_{2(2,0)} \mathcal{S}_2^+ e^{i\mu} q + \mathcal{T}_{2(1,1)} \left(\mathcal{S}_2^+ e^{-i\mu} p + \mathcal{S}_2^- e^{i\mu} q\right) + 2\mathcal{T}_{2(0,2)} \mathcal{S}_2^- e^{-i\mu} p \\
& - 2\mathcal{T}_{2(2,0)} \mathcal{T}_2^+ e^{2i\mu} q - \mathcal{T}_{2(1,1)} \left(\mathcal{T}_2^+ p + \mathcal{T}_2^- q\right) - 2\mathcal{T}_{2(0,2)} \mathcal{T}_2^- e^{-2i\mu} p
\end{aligned} \tag{3.40}
$$

The surviving terms of $\mathcal{S}_{2\to 3}$ after the third order transformation are $\mathcal{K}^+_{2\to 3(2,1)}$ and its complex conjugate counterpart $\mathcal{K}^-_{2\to 3(1,2)}$

$$
\begin{aligned}
\mathcal{K}^+_{2\to 3(2,1)} &= 2\mathcal{T}^+_{2(2,0)} \mathcal{S}^+_{2(1,1)} e^{i\mu} + \mathcal{T}^+_{2(1,1)} \left(\mathcal{S}^+_{2(2,0)} e^{-i\mu} + \mathcal{S}^-_{2(1,1)} e^{i\mu}\right) \\
&\quad + 2\mathcal{T}^+_{2(0,2)} \mathcal{S}^-_{2(2,0)} e^{-i\mu} - 2\mathcal{T}^+_{2(2,0)} \mathcal{T}^+_{2(1,1)} e^{2i\mu} \\
&\quad - \mathcal{T}^+_{2(1,1)} \left(\mathcal{T}^+_{2(2,0)} + \mathcal{T}^-_{2(1,1)}\right) - 2\mathcal{T}^+_{2(0,2)} \mathcal{T}^-_{2(2,0)} e^{-2i\mu} \\
&= \frac{-2\mathcal{S}^+_{2(2,0)} \mathcal{S}^+_{2(1,1)}}{e^{i\mu} - 1} - \frac{\mathcal{S}^+_{2(1,1)}}{1 - e^{i\mu}} \left(\mathcal{S}^+_{2(2,0)} e^{-i\mu} + \mathcal{S}^-_{2(1,1)} e^{i\mu}\right) + \frac{2\mathcal{S}^+_{2(0,2)} \mathcal{S}^-_{2(2,0)}}{e^{2i\mu} - e^{-i\mu}} \\
&\quad + \frac{\mathcal{S}^+_{2(1,1)} \left(2\mathcal{S}^+_{2(2,0)} + \mathcal{S}^-_{2(1,1)}\right)}{2\left(\cos\mu - 1\right)} - \frac{\mathcal{S}^+_{2(1,1)} \mathcal{S}^+_{2(2,0)}}{2e^{2i\mu} - e^{i\mu} - e^{3i\mu}} - \frac{2\mathcal{S}^+_{2(0,2)} \mathcal{S}^-_{2(2,0)}}{2e^{2i\mu} - e^{-i\mu} - e^{5i\mu}}
\end{aligned} \tag{3.41}
$$

Also for $\mathcal{K}^{+}_{2\to3(2,1)}$ and $\mathcal{K}^{-}_{2\to3(1,2)}$ the intertwine dependency on all terms of $\mathcal{S}_2$ becomes apparent highlighting the complex relation between lower order and higher order terms.

3.3.4 Transformation back to Real Space Normal Form

Since the original map $\mathcal{M}_0$ only operates in real space, the normal form map $\mathcal{M}_{\mathrm{NF}}$ should also only operate in real space. This is why the current map $\mathcal{M}_m$, where m is the order of last transformation, is transformed to a real normal form basis $(q_{\mathrm{NF}}, p_{\mathrm{NF}})$ composed of the real and imaginary parts of the current complex conjugate basis (q_m, p_m). Based on [14, Eq. (7.58) and (7.59) and (7.67)] the bases are related as follows

$$q_{\mathrm{NF}} = \frac{q_m + p_m}{2} \quad \text{and} \quad p_{\mathrm{NF}} = \frac{q_m - p_m}{2i}, \tag{3.42}$$

and

$$q_m = q_{\mathrm{NF}} + i\, p_{\mathrm{NF}} \quad \text{and} \quad p_m = q_{\mathrm{NF}} - i\, p_{\mathrm{NF}} \quad \text{with} \tag{3.43}$$

$$q_m p_m = q_{\mathrm{NF}}^2 + p_{\mathrm{NF}}^2 = r_{\mathrm{NF}}^2. \tag{3.44}$$

The associated transfer matrix below to the real normal form basis is obtained from the equations above.

$$\mathcal{A}_{\mathrm{real}} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix} = \begin{pmatrix} (q_{\mathrm{NF}}|q_m) & (q_{\mathrm{NF}}|p_m) \\ (p_{\mathrm{NF}}|q_m) & (p_{\mathrm{NF}}|p_m) \end{pmatrix} \tag{3.45}$$

The inverse relation is given accordingly:

$$\mathcal{A}^{-1}_{\mathrm{real}} = \begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix} = \begin{pmatrix} (q_m|q_{\mathrm{NF}}) & (q_m|p_{\mathrm{NF}}) \\ (p_m|q_{\mathrm{NF}}) & (p_m|p_{\mathrm{NF}}) \end{pmatrix}. \tag{3.46}$$

The transformation back to the real space (into normal form space) yields

$$\begin{aligned}\mathcal{M}_{\mathrm{NF}} = \mathcal{A}_{\mathrm{real}} \circ \mathcal{M}_m \circ \mathcal{A}_{\mathrm{real}}^{-1} &= \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix} \cdot \begin{pmatrix} f^+\left(r_{\mathrm{NF}}^2\right)(q_{\mathrm{NF}} + i\, p_{\mathrm{NF}}) \\ f^-\left(r_{\mathrm{NF}}^2\right)(q_{\mathrm{NF}} - i\, p_{\mathrm{NF}}) \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2}\left(f^+ + \bar{f}^+\right) t_+ + \frac{i}{2}\left(f^+ - \bar{f}^+\right) t_- \\ \frac{-i}{2}\left(f^+ - \bar{f}^+\right) t_+ + \frac{1}{2}\left(f^+ + \bar{f}^+\right) t_- \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{Re}\left(f^+\left(r_{\mathrm{NF}}^2\right)\right) & -\mathrm{Im}\left(f^+\left(r_{\mathrm{NF}}^2\right)\right) \\ \mathrm{Im}\left(f^+\left(r_{\mathrm{NF}}^2\right)\right) & \mathrm{Re}\left(f^+\left(r_{\mathrm{NF}}^2\right)\right) \end{pmatrix} \cdot \begin{pmatrix} q_{\mathrm{NF}} \\ p_{\mathrm{NF}} \end{pmatrix}. \end{aligned} \tag{3.47}$$

For the example of the centrifugal governor up to order three the normal form is

$$\mathcal{M}_{\mathrm{NF}} = \begin{pmatrix} \cos\mu + \frac{1}{2}\mathrm{Re}\left(\mathcal{S}^+_{3,\mathrm{new}(2,1)}\right) r_{\mathrm{NF}}^2 & -\sin\mu - \frac{1}{2}\mathrm{Im}\left(\mathcal{S}^+_{3,\mathrm{new}(2,1)}\right) r_{\mathrm{NF}}^2 \\ \sin\mu + \frac{1}{2}\mathrm{Im}\left(\mathcal{S}^+_{3,\mathrm{new}(2,1)}\right) r_{\mathrm{NF}}^2 & \cos\mu + \frac{1}{2}\mathrm{Re}\left(\mathcal{S}^+_{3,\mathrm{new}(2,1)}\right) r_{\mathrm{NF}}^2 \end{pmatrix} \cdot \begin{pmatrix} q_{\mathrm{NF}} \\ p_{\mathrm{NF}} \end{pmatrix}. \tag{3.48}$$

The Tab. 3.7 below yields the values for the normal form map of our example case.

Table 3.7: The normal form map $\mathcal{M}_{\mathrm{NF}}$ up to order three. The component $\mathcal{M}^+_{\mathrm{NF}}$ is on the left, $\mathcal{M}^-_{\mathrm{NF}}$ on the right.

O	Coeff.	Value	Coeff.	Value
1	$\mathcal{M}^+_{\mathrm{NF}(1,0)}$	0.339185989	$\mathcal{M}^-_{\mathrm{NF}(1,0)}$	0.940719334
1	$\mathcal{M}^+_{\mathrm{NF}(0,1)}$	-0.940719334	$\mathcal{M}^-_{\mathrm{NF}(0,1)}$	0.339185989
3	$\mathcal{M}^+_{\mathrm{NF}(3,0)}$	0.470359667	$\mathcal{M}^-_{\mathrm{NF}(3,0)}$	-0.169592994
3	$\mathcal{M}^+_{\mathrm{NF}(2,1)}$	0.169592994	$\mathcal{M}^-_{\mathrm{NF}(2,1)}$	0.470359667
3	$\mathcal{M}^+_{\mathrm{NF}(1,2)}$	0.470359667	$\mathcal{M}^-_{\mathrm{NF}(1,2)}$	-0.169592994
3	$\mathcal{M}^+_{\mathrm{NF}(0,3)}$	0.169592994	$\mathcal{M}^-_{\mathrm{NF}(0,3)}$	0.470359667

The normal form transformation from $\mathcal{M}_0$ to $\mathcal{M}_{\mathrm{NF}}$ can be obtained by the combination of all the single transformations yielding

$$\begin{aligned}\mathcal{M}_{\mathrm{NF}} = &\underbrace{\mathcal{A}_{\mathrm{real}} \circ \mathcal{A}_m \circ \mathcal{A}_{m-1} \circ \ldots \circ \mathcal{A}_1 \circ \mathcal{A}_{\mathrm{FP}}}_{\mathcal{A}} \circ \mathcal{M}_0 \\ &\circ \underbrace{\mathcal{A}_{\mathrm{FP}}^{-1} \circ \mathcal{A}_1^{-1} \circ \ldots \circ \mathcal{A}_{m-1}^{-1} \circ \mathcal{A}_m^{-1} \circ \mathcal{A}_{\mathrm{real}}^{-1}}_{\mathcal{A}^{-1}}. \end{aligned} \tag{3.49}$$

Table 3.8: The normal form transformation $\mathcal{A}$ up to order three. The component $\mathcal{A}^+$ is on the left, $\mathcal{A}^-$ on the right.

O	Coeff.	Value	Coeff.	Value
1	$\mathcal{A}^+_{1(1,0)}$	1.106681920	$\mathcal{A}^-_{1(1,0)}$	0
1	$\mathcal{A}^+_{1(0,1)}$	0	$\mathcal{A}^-_{1(0,1)}$	-0.903602004
2	$\mathcal{A}^+_{2(2,0)}$	0.319471552	$\mathcal{A}^-_{2(2,0)}$	0
2	$\mathcal{A}^+_{2(1,1)}$	0	$\mathcal{A}^-_{2(1,1)}$	0.521694860
2	$\mathcal{A}^+_{2(0,2)}$	0.425962069	$\mathcal{A}^-_{2(0,2)}$	0
3	$\mathcal{A}^+_{3(3,0)}$	-0.046111747	$\mathcal{A}^-_{3(3,0)}$	0
3	$\mathcal{A}^+_{3(2,1)}$	0	$\mathcal{A}^-_{3(2,1)}$	-0.414150918
3	$\mathcal{A}^+_{3(1,2)}$	-0.399635138	$\mathcal{A}^-_{3(1,2)}$	0
3	$\mathcal{A}^+_{3(0,3)}$	0	$\mathcal{A}^-_{3(0,3)}$	0.025100056

The values of the coefficients of the full normal form transformation $\mathcal{A}$ are given in Tab. 3.8.

Writing the complex conjugate functions $f^{\pm}$ from the equations above (particularly Eq. (3.47)) in a complex notation as $f^{\pm}\left(r^2_{\mathrm{NF}}\right) = e^{\pm i\Lambda\left(r^2_{\mathrm{NF}}\right)}$ illustrates circular behavior of the normal form:

$$\mathcal{M}_{\mathrm{NF}} = \begin{pmatrix} \cos\left(\Lambda\left(r^2_{\mathrm{NF}}\right)\right) & -\sin\left(\Lambda\left(r^2_{\mathrm{NF}}\right)\right) \\ \sin\left(\Lambda\left(r^2_{\mathrm{NF}}\right)\right) & \cos\left(\Lambda\left(r^2_{\mathrm{NF}}\right)\right) \end{pmatrix} \cdot \begin{pmatrix} q_{\mathrm{NF}} \\ p_{\mathrm{NF}} \end{pmatrix}. \tag{3.50}$$

It shows that the normal form $\mathcal{M}_{\mathrm{NF}}$ consists of circular curves in phase space with only amplitude depended angle advancements $\Lambda\left(r^2_{\mathrm{NF}}\right)$.

3.3.5 Invariant Normal Form Radius

The squared normal form radius r^2_{NF} is related to the original coordinates (q_0, p_0) by the normal form transformation $\mathcal{A}$, where

$$\begin{aligned} r^2_{\mathrm{NF}}(q_0, p_0) &= \left(q^2_{\mathrm{NF}}(q_0, p_0) + p^2_{\mathrm{NF}}(q_0, p_0)\right) \\ &= \left(\mathcal{A}^2_+ + \mathcal{A}^2_-\right)(q_0, p_0). \end{aligned} \tag{3.51}$$

Explicitly calculating the squared normal form radius with the normal form transformation $\mathcal{A}$ up

to order three from Tab. 3.8 yields

$$r_{\mathrm{NF}}^2 =_3 1.224745 q_0^2 + 0.816497 p_0^2 + 0.707107 q_0^3 \tag{3.52}$$

$$\approx \sqrt{\frac{3}{2}} q_0^2 + \sqrt{\frac{2}{3}} p_0^2 + \frac{1}{\sqrt{2}} q_0^3 \tag{3.53}$$

$$=_3 \frac{2\sqrt{2}}{\sqrt{3}} \left(E\left(\frac{\pi}{3} + q_0, p_0\right) - E\left(\frac{\pi}{3}, 0\right)\right). \tag{3.54}$$

This direct relationship between the energy E, as an invariant or constant of motion, and the squared normal form radius up to order three confirms that the normal form radius constitutes a constant of motion up to calculation order.

The invariant of motion is a family of functions that remain constant for all phase space states (q, p) along their phase space motion. In particular, if $I(q, p)$ is an invariant of motion, then so is $I^2(q, p)$ or any other function $f(I)$, which is defined by the resulting values of I. Furthermore, $I(Q(q, p), P(q, p))$ is also an invariant if (Q, P) belong to the same phase space curve as (q, p). Transfer maps can yield such relations $(Q(q, p), P(q, p))$, since they can represent how a phase space final state (Q, P) depends on the phase space initial state (q, p).

Accordingly, the energy E and the normal form radius r_{NF}^2 are both functions of the same family and related by the transformations explained in the paragraph above. Up to order three, this relation includes a shift by a constant and scaling, but the relation might reveal itself to be more complex than this with higher orders.

3.3.6 Angle Advancement, Tune and Tune Shifts

In the beam physics terminology, the angle advancements $\Lambda\left(r_{\mathrm{NF}}^2\right)$ are scaled to the interval $[0, 1]$ instead of $[0, 2\pi]$ and referred to as the tune and amplitude dependent tune shifts [14]. The angle advancement can be calculated from the normal form map via

$$\Lambda\left(r_{\mathrm{NF}}^2 = q_{\mathrm{NF}}^2, p_{\mathrm{NF}} = 0\right) = \arccos\left(\frac{\mathcal{M}_{\mathrm{NF}}^+\Big|_{p_{\mathrm{NF}}=0}}{q_{\mathrm{NF}}}\right) = \arccos\left(\mathrm{Re}\left(f^+\left(r_{\mathrm{NF}}^2\right)\right)\right). \tag{3.55}$$

For the centrifugal governor example up to order three, the angle advancement is given by

$$\begin{aligned}\Lambda\left(r_{\mathrm{NF}}^2 = q_{\mathrm{NF}}^2\right) &= \arccos\left(\cos\mu + \frac{1}{2}\mathrm{Re}\left(\mathcal{S}^+_{3,\mathrm{new}(2,1)}\right) r_{\mathrm{NF}}^2\right) \\ &= \mu - \frac{\mathrm{Re}\left(\mathcal{S}^+_{3,\mathrm{new}(2,1)}\right)}{2\sin\mu} r_{\mathrm{NF}}^2.\end{aligned} \tag{3.56}$$

Note that μ is the eigenvalue phase of the original linear part. Accordingly, the tune ν is $\mu/2\pi$. For the centrifugal governor, the tune and tune shifts are

$$\frac{\Lambda\left(r_{\mathrm{NF}}^2\right)}{2\pi} = \nu\left(r_{\mathrm{NF}}^2\right) = 0.1949242 - 0.07957747 r_{\mathrm{NF}}^2 \tag{3.57}$$

With the expression of r_{NF}^2 in terms of the original coordinates (q_0, p_0) from Eq. (3.52) the tune and tune shifts are evaluated to

$$\nu\left(q_0, p_0\right) = 0.1949242 - 0.0974621 q_0^2 - 0.0649747 p_0^2 - 0.05626977 q_0^3. \tag{3.58}$$

This yields a key insight into the centrifugal governor behavior for $\omega = \sqrt{2}$. We already know that the centrifugal governor is rotating at $\frac{\sqrt{2}}{2\pi} \approx 0.225$ revolutions per T_0 for $\omega = \sqrt{2}$. The tune of about 0.195 tells us that the centrifugal governor arms oscillate at a frequency of about $0.195 + c$ oscillations per T_0 around their equilibrium position. The negative tune shifts additionally show that this frequency is decreasing for increasing amplitude of oscillation.

Since the map can only compare initial and final state of the oscillation after the integration time of 1 T_0 we only know how much the oscillation cycle has advanced over this period, but not how many additional full oscillations c have been completed in the meantime. By doing the same process as above for the centrifugal governor with $\omega = \sqrt{2}$ for a Poincaré map after time $t = \frac{2\pi}{\sqrt{2}}$, i.e. one full centrifugal governor revolution, yields

$$\nu\left(q_0, p_0\right) = 0.8660254 - 0.4330127 q_0^2 - 0.2886751 p_0^2 - 0.25000000 q_0^3, \tag{3.59}$$

which is exactly a factor of $\frac{2\pi}{\sqrt{2}}$ larger than the tunes from Eq. (3.57). This means that c must be zero and we did not miss any full oscillations during the integration up to $t = 1$.

From Eq. (3.57) we can directly calculate the period of oscillation from normal form T_{NF}, which is just $1/\nu\,(q_0, p_0)$.

To compare the calculated normal form period T_{NF} to the actual period of oscillation, we overlay the oscillatory plot from Fig. 3.4 with the calculated periods (see Fig. 3.6). The centrifugal governor

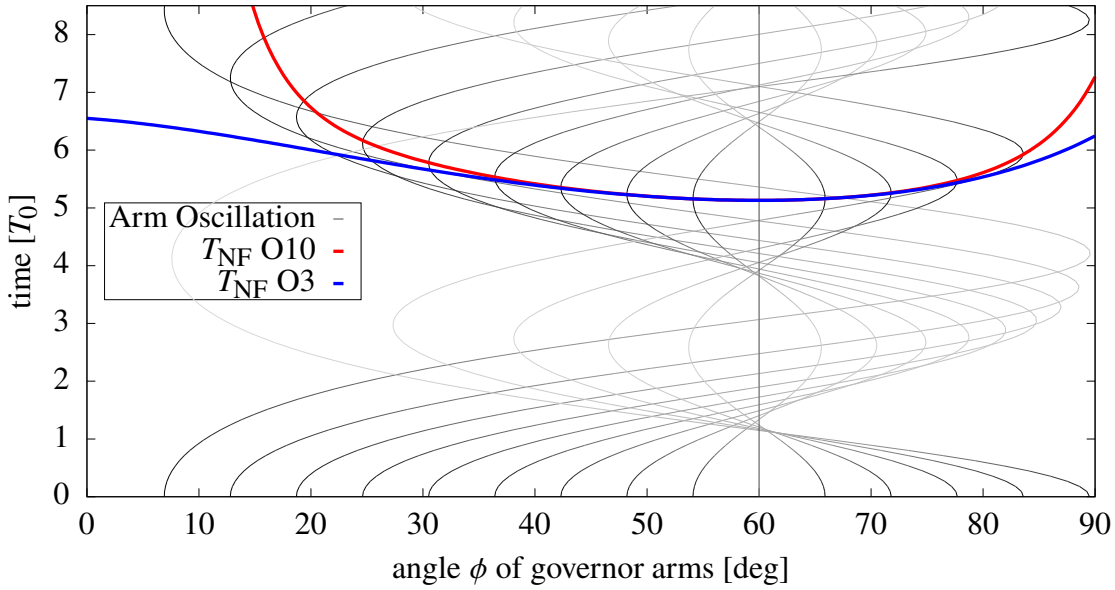

Figure 3.6: Comparison between the calculated period with normal form methods T_{NF} for calculation order ten (O10) and calculation order three (O3) to the actual period of oscillation given by the oscillatory behavior of the centrifugal governor arms for $\omega = \sqrt{2}$.

arms are initiated with multiple angle offsets with $p_\phi = 0$ relative to their equilibrium angle at $\phi_0 = 60°$. If the normal form calculation of the period is correct, the calculated period will agree with the time when the equilibrium governor arms reach their initial position amplitude after one actual period of oscillation.

The higher the amplitude of oscillation, the more relevant are higher order effects. Accordingly, the accuracy drops with larger amplitudes. The order three calculation performs well between $35°(\delta\phi = -25°)$ and $75°(\delta\phi = +15°)$, while the order ten calculation can extend an accurate description over the range from $35°(\delta\phi = -35°)$ to $85°(\delta\phi = +25°)$.

The normal form algorithm can also be performed with parameters, e.g., depending on changes to ω. In Tab. 3.9 result for the amplitude and parameter $\delta\omega$ dependent tunes shifts are listed. It

shows that the $\delta\omega$ dependent tune shifts are positive, which means that an increase in ω increases the oscillation frequency of centrifugal governor arms. This is related to the deeper potential well.

Table 3.9: Tune and coefficients of amplitude and parameter $\delta\omega$ dependent tune shifts for centrifugal governor with $\omega_0 = \sqrt{2}$.

	Exponents				Exponents		
Coefficient	q_0	p_0	$\delta\omega$	Coefficient	q_0	p_0	$\delta\omega$
0.1949242003	0	0	0	-0.0562697698	3	0	0
0.3355884937	0	0	1	-0.0307638305	2	0	1
-0.0974621002	2	0	0	0.1741334861	1	1	1
-0.0649747334	0	2	0	0.1123973696	0	2	1
0.1591549431	1	0	1	-0.0435458248	1	0	2
-0.5753522001	0	0	2	-0.0142179396	0	1	2
				0.0866936204	0	0	3

This knowledge about the dependency of the tunes on parameter shifts can help by the selection of a suitable ω, e.g., to avoid resonances between the governors revolution frequency and the oscillation frequency of the arms. While such a resonance is irrelevant in this simplified example it might be critical if the governor is part of a more complex system.

3.4 Visualization of the Different Order Normal Forms and Conclusion

In this chapter, we considered the system of a centrifugal governor with a fixed rotation frequency of $\omega = \sqrt{2}$ and analyzed it using the DA normal form algorithm.

To visualize the effect of the different steps in the DA normal form algorithm, Fig. 3.7 shows phase space tracking pictures for incomplete normal form maps. Given the tenth order Poincaré map which describes the behavior of the centrifugal governor for $\omega = \sqrt{2}$, these incomplete normal form maps stopped the normal form transformations at an order $n < 10$ such that the resulting incomplete normal form map is only normalized up to order n. There is no practical use for these incomplete normal form maps other than showing the progress of the normal form algorithm, since to make use of the normal form properties completion of the normal form transformation to the full order of the map is required. The phase space behavior in the full order normal form with its rotationally invariant property was previously shown in Fig. 3.5.

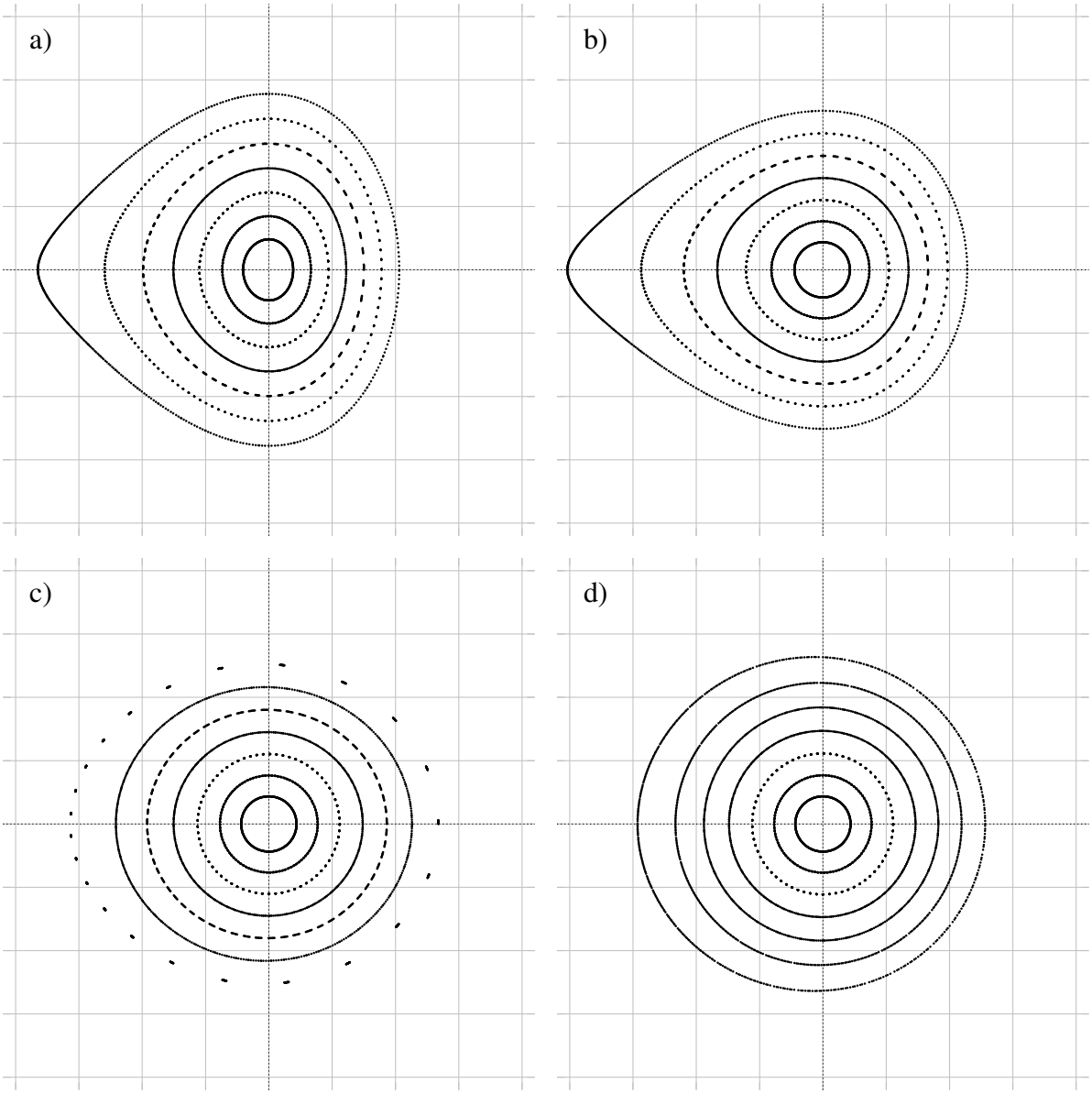

Figure 3.7: Phase space tracking of incomplete normal form maps of order ten of the centrifugal governor arms with a fixed rotation frequency of $\omega = \sqrt{2}$. The original map (a), only linear normal form transformation (b), and only normal form transformations up to order two (c) and three (d), respectively.

The difference between a) and b) in Fig. 3.7 shows the effect of the linear transformation, which scales the variables to create circles close to the expansion point. The nonlinear distortions for larger amplitudes are still present. With the second and third order transformation, these distortions are removed in the normal form.

As a result of the DA normal form algorithm, we were able to produce invariants of motion up to calculation order. Specifically, we could show how the squared normal form radius is directly related to the energy E up to calculation order, which is a constant of motion for this system.

The normal form algorithm also provided transformations from the original coordinates to the normal form coordinates, which were used to relate the phase space amplitudes to the normal form invariant.

Finally, the normal form produced the period of oscillation of the centrifugal governor arms around their equilibrium angle depending on the amplitude of oscillation. Only for very large amplitudes, the limited calculation order was not able to capture all the relevant high order effects to accurately describe the period of oscillation for these amplitudes.

CHAPTER 4

BOUNDED MOTION PROBLEM

This chapter contains large parts of my paper *Bounded motion design in the Earth zonal problem using differential algebra based normal form methods* published in *Celestial Mechanics and Dynamical Astronomy, Vol. 132, 14 (2020)* [88]. The paper was authored by Roberto Armellin, Martin Berz, and me.

Given the detailed understanding of the DA normal form algorithm from Sec. 2.3 and Chapter 3, we present its application in a new technique for the calculation of entire continuous sets of orbits, which remain in long term relative bounded motion under zonal gravitational perturbation. We will see that the application of the DA normal form algorithm in this particular case is only possible due to a well-chosen Poincaré surface for the Poincaré return map (Sec. 2.2), which captures the critical phase space behavior at the right space-time instance, which requires a combination of dimension-reducing phase space projections.

4.1 Introduction to Bounded Motion

The term 'bounded motion' is used in the field of astrodynamics to describe a special orbital flight pattern of two objects (usually man-made satellites), where the two objects remain in close proximity to each other over an extended period of time. Both objects are on orbits around a common central gravitational body like a planet, moon, asteroid, or star, and their relative distance is bounded.

In practice, bounded motion finds application in cluster flight [24] and formation flying [1] missions, which can offer many advantages compared to single spacecraft missions. From the scientific standpoint, they enable measurements of unprecedented spatial and temporal correlation, but they also have economic advantages such as allowing for redundancies within the spacecraft group, a distribution of the payload, and the adaptability of the mission by exchanging modules of the group. Missions such as PRISMA [26], GRACE [58], and TerraSAR-X and TanDEM-X [27] demonstrated the practicability of formation flying and stimulated further research in the field.

Moving from an ideal unperturbed system with elliptical Kepler orbits to the realistic mission case by considering perturbations to the dynamics makes it not trivial to find bounded motion orbits. The dominating perturbation is often due to the oblateness of the central body and the associated zonal perturbation from the second zonal harmonic coefficient J_2 of the gravitational potential. This zonal perturbation introduces a drift in the right ascension of the ascending node (RAAN) $\Delta\Omega$, the argument of periapsis, and the mean anomaly. The drift in each of the quantities is oscillating at different frequencies, which drastically increases the complexity of the bounded motion problem. Additional non-zonal gravitational perturbations break the rotational symmetry of the system and the regular oscillations in each of the quantities mentioned above, which complicates the problem even more.

To minimize the extent of formation-keeping maneuvers with control strategies during a mission, it is of great interest to the astrodynamical community to find 'naturally' bounded motion orbits for models considering as many perturbations as possible, which leave only the unmodeled perturbations to be corrected by control maneuvers. In this section of the dissertation and in [88], we present a method that allows for the design of long term relative bounded motion considering a zonal gravitational model using normal form methods. Since [88] contains an extensive literature review of previous approaches, only contributions directly linked to our technique for the zonal problem will be mentioned below.

The pioneering work by Broucke [23] on families of two dimensional quasi-periodic invariant tori around stable periodic orbits of the Ruth-reduced axially symmetric system was used by Koon *et al.* [40] in combination with Poincaré section techniques to study the J_2 problem. While this method improved first order approaches, long term bounded motion was still not achieved by placing orbits on the center manifold. Xu *et al.* [90] pointed out that long term bounded motion in the zonally perturbed system could only be achieved when the RAAN drift $\Delta\Omega$ and nodal period T_d are on average the same for each of the bounded modules (see Sec. 4.2.5). These constraints are weaker than the constraints originally derived by Martinusi and Gurfil [57].

In [5], a fully numerical technique based on stroboscopic maps was used to obtain entire families

of quasi-periodic orbits producing bounded relative motion about a periodic one. This method was then used to study both bounded motion about asteroids [4] and in low Earth, medium Earth, and geostationary orbits [6]. Numerical approaches yield bounded relative orbits with arbitrary size over very long periods of time (or infinite time in theory). However, they require complex and time-consuming algorithms.

In [35], a compromise between the analytic and numerical technique was presented based on the use of DA. DA techniques were used to expand to high order the mapping between two consecutive equatorial crossings (i.e., Poincaré maps). This enabled the study of the motion of a spacecraft for many revolutions by the fast evaluation of Taylor polynomials. The problem of designing bounded motion orbits was then reduced to the solution of two nonlinear polynomial equations, namely constraining the mean nodal period T_d and drift of the right ascension of the ascending node $\Delta\Omega$. The derived method showed an accuracy comparable with that of fully numerical methods but with a reduced complexity due to the introduced polynomial approximations. The main drawback of this technique consisted of the calculation of the mean T_d and $\Delta\Omega$ using numerical averaging over thousands of nodal crossings. This process resulted in the computationally intensive part of the algorithm and was also responsible for accuracy degradation in case of very large separations.

The advantage of our approach is that it overcomes this limitation when calculating bounded motion orbits under zonal perturbation by the introduction of DA based normal form (DANF) methods. In particular, the high-order DANF algorithm is used to determine a change of expansion variables of the Poincaré map into normal form space, in which the phase space behavior is circular and can be easily parameterized by action-angle coordinates (Fig. 4.3). The action-angle representation of the normal form coordinates is then used to parameterize the original phase space coordinates of the Poincaré return map. The original map is averaged over a full phase space revolution by a path integral along the angle parameterization, yielding the Taylor expansion of the averaged bounded motion quantities T_d and $\Delta\Omega$, for which the bounded motion conditions are straightforwardly imposed. Sets of highly accurate bounded orbits are obtained in the full zonal problem, extending over several thousand kilometers and valid for decades. This method avoids the

numerical averaging introduced in [35]. The superiority in terms of elegance, computational time, and accuracy of the new algorithm will be demonstrated using similar test cases to those presented in [35] and [6].

Before introducing our approach from [88], we start with some basics on the orbital motion under gravitational perturbation. Later we will show our results for the full zonal problem [88].

4.2 Understanding Orbital Motion Under Gravitational Perturbation

We consider the orbital motion around a single central body of mass, where the motion is only determined by the gravitational potential of the central body. Perturbations due to atmospheric drag, solar radiation pressure, or the gravitational field of other space bodies are ignored. We also ignore parabola and hyperbola orbits, which escape the gravitational potential due to their large enough kinetic energy.

4.2.1 The Perturbed Gravitational Potential

Any gravitational potential U can be expressed in terms of spherical harmonics $Y_{l,m}$ and the corresponding coefficients $k_{l,m}$:

$$U(r,\theta,\phi) = -\frac{\mu}{r}\left(1 + \sum_{l=1}^{\infty}\sum_{m=-l}^{l} k_{l,m}\left(\frac{R_0}{r}\right)^l Y_{l,m}(\theta,\phi)\right), \tag{4.1}$$

where (r,θ,ϕ) are spherical coordinates with the origin at the center of mass and where μ is the product of the gravitational constant and the mass of the central body. The coefficients of the $Y_{l,m}$ are often split into $k_{l,m}\cdot R_0^l$ to make them independent of the size R_0 of the central body.

The orientation of the coordinate system is usually chosen such that $\hat{z}$ ($\theta = 0$) aligns with the dominating symmetry axis of the central body. The plane perpendicular to $\hat{z}$, i.e. the xy plane or $\theta = \frac{\pi}{2}$ plane, is referred to as the equatorial plane or plane of reference.

The spherical harmonics can be grouped into three categories. Zonal terms ($m = 0$) are independent of the longitude ϕ creating zones in the vertical/latitudinal direction. Sectional terms ($m = l$) on the other hand are independent of the latitude θ creating sections longitudinally. Tesseral

terms ($0 < m < l$) dependent on both ϕ and θ creating a chessboard pattern on the sphere. Each of these terms is considered a gravitational perturbation to the spherically symmetric potential $U_0 = -\frac{\mu}{r}$, which only depends on the distance r.

The gravitational potentials of many rotating central bodies are dominated by their low order zonal terms, in particular, $Y_{2,0}$, since centrifugal effects of the rotation often cause a zonally dependent mass distribution with more mass at the equator and less mass at the poles compared to the sphere. Considering only the effects of zonal perturbations is also referred to as the zonal problem and is going to be the basis of our analysis. The axial symmetry conserves the angular momentum component along the symmetry axis and simplifies the potential significantly as the spherical harmonics $Y_{l,m}$ reduce to the ordinary Legendre polynomials P_l, with

$$U(r, \theta) = -\frac{\mu}{r}\left(1 + \sum_{l=1}^{\infty} J_l \left(\frac{R_0}{r}\right)^l P_l(\cos\theta)\right). \tag{4.2}$$

4.2.2 The Equations of Motion

To calculate the behavior of an object in the perturbed gravitational field, we derive the equations of motion, which describe the dynamics as a set of mathematical functions. To be consistent with previous approaches and [88], we will use cylindrical coordinates. The starting point of the derivation is the Lagrangian

$$L = \frac{1}{2}\left(\dot{\rho}^2 + \dot{z}^2 + \rho^2\dot{\phi}^2\right) - U(\rho, z, \phi) \tag{4.3}$$

of the system in cylindrical coordinates (ρ, z, ϕ), where ρ is the distance in the equatorial plane such that $r = \sqrt{\rho^2 + z^2}$ yields the total distance between the orbiting object and the center of mass. The potential takes the following form in cylindrical coordinates

$$U(\rho, z, \phi) = -\frac{\mu}{r}\left[1 + \sum_{l=1}^{\infty}\sum_{m=0}^{l}\left(\frac{R_0}{r}\right)^l P_{l,m}\left(\frac{z}{r}\right)\left(C_{l,m}\cos(m\phi) + S_{l,m}\sin(m\phi)\right)\right], \tag{4.4}$$

where $P_{l,m}$ are the associated Legendre polynomials.

For the zonal problem ($m = 0$), the $P_{l,m}$ reduce to the ordinary Legendre polynomials P_l. The coefficients $C_{l,0}$ of the zonal problem are often denoted by J_l.

The canonical momenta $\left(v_\rho, v_z, v_\phi\right)$ to the position variables (ρ, z, ϕ) are given by

$$v_\rho = \frac{\partial L}{\partial \dot{\rho}} = \dot{\rho} \quad v_z = \frac{\partial L}{\partial \dot{z}} = \dot{z} \quad v_\phi = \frac{\partial L}{\partial \dot{\phi}} = \rho^2 \dot{\phi} \doteq \mathcal{H}_z, \tag{4.5}$$

where $\mathcal{H}_z$ is the angular momentum component along the symmetry axis $\hat{z}$ and the canonical momentum to the angle ϕ. From the Lagrange-Euler equations it follows that $\dot{\mathcal{H}}_z = -\frac{\partial U}{\partial \phi}$, which is zero for the zonal problem due to the axial symmetry making $\mathcal{H}_z$ a constant of motion.

Using the Legendre transformation, the Hamiltonian

$$H = \frac{1}{2}\left(v_\rho^2 + v_z^2 + \frac{\mathcal{H}_z^2}{\rho^2}\right) + U\left(\rho, z, \phi\right) \tag{4.6}$$

is obtained. Due to the time independence of the system ($\mathrm{d}_t H = 0$), the Hamiltonian is equivalent to the energy E, which is a constant of motion.

The equations of motion are derived from the Hamiltonian via the Hamilton equations

$$\dot{\rho} = v_\rho \qquad \dot{z} = v_z \qquad \dot{\phi} = \frac{\mathcal{H}_z}{\rho^2} \tag{4.7}$$

$$\dot{v}_\rho = \frac{\mathcal{H}_z^2}{\rho^3} - \frac{dU}{d\rho} \qquad \dot{v}_z = -\frac{dU}{dz} \qquad \dot{\mathcal{H}}_z = -\frac{dU}{d\phi}. \tag{4.8}$$

The time evolution $\mathcal{X}\left(t\right)$ of the state $\mathcal{X} = (r, v_r, z, v_z, \phi, \mathcal{H}_z)^T$ of a spacecraft is determined by integrating the system of ODE's $\dot{\mathcal{X}} = f\left(\mathcal{X}\right)$ from above. The orbit $\mathcal{O}$ of the spacecraft is described by the set of all states $\mathcal{X}\left(t\right)$.

4.2.3 The Kepler Orbit

Before we investigate the orbital behavior under perturbation, it is advisable to understand the unperturbed system with the spherically symmetric gravitational potential U_0. The orbiting motion of an object in the unperturbed potential takes the Keplerian form of a closed ellipse, which makes the motion two dimensional. The plane in which the ellipse lies is called the orbital plane.

The traditional orbital elements $(a, e, i, \Omega, \omega, \nu(t))$, also called Keplerian elements, characterize the position and orbit of the object using the elliptical shape as well as the equatorial plane of the central body as a reference. The variables a and e define the size (semi-major axis) and shape

(eccentricity) of the ellipse, respectively. To describe the orientation of the orbital plane with respect to the central body, the reference direction $\hat{x}$ within the equatorial plane is defined. Except for orbits within the equatorial plane, the elliptical orbit intersects with the equatorial plane in two places. The intersection in the $\hat{z}$ direction (from south to north) is called the ascending node ☊. The angle between the equatorial plane and the orbital plane is called the inclination i. The angle between the reference direction $\hat{x}$ and the ascending node within the equatorial plane is the longitude or right ascension of the ascending node (RAAN) Ω. The argument of periapsis ω describes the orientation of the ellipse within the orbital plane as the angle between the ascending node and the periapsis (closest point of the ellipse to the origin). The true anomaly $\nu(t)$ yields the position of the object along the ellipse as the angle between the periapsis and the object. The time between two consecutive ascending nodes is called the nodal period T_d, with $T_d = t\,(☊_{n+1}) - t\,(☊_n)$.

4.2.4 Orbits Under Gravitational Perturbation

The elliptical orbits deform under gravitational perturbations such that the orbits no longer close after a revolution around the central body.

The description of perturbed orbits using Keplerian elements has to be carefully considered, since the four elements (a, e, ω, ν) are based on the assumption of an elliptical orbit in an unperturbed system. The elements i and Ω, on the other hand, only describe the orientation of the orbital plane, determined by position and velocity vectors of the orbiting object, but make no assumptions about the shape of the orbit. In practice, the Keplerian elements are calculated at each point in time assuming the orbit is an ellipse in an unperturbed system while propagating the object in the perturbed system.

This representation is particularly helpful when the gravitational potential is only slightly perturbed. It shows how the unperturbed elliptical orbit is influenced by the perturbations at each point in time. In Fig. 4.1, the orbital elements of a low Earth orbit ($\mathcal{O}_2$ from Sec. 4.4.1) under zonal perturbation are shown. As a reference, the orbit is also initiated with the same starting conditions but propagated considering only the spherically symmetrical part of the Earth gravitational field.

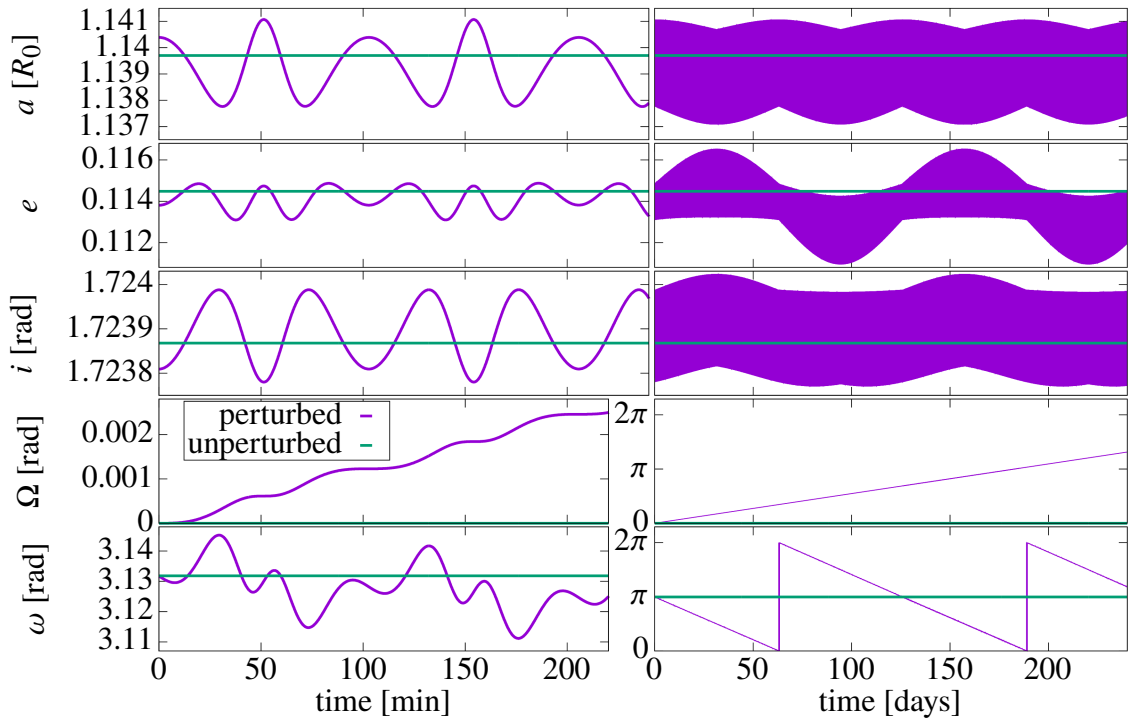

Figure 4.1: The behavior of the Keplerian elements of a low Earth orbit under zonal gravitational perturbations up to J_{15} (purple) and as a regular Kepler orbit in the unperturbed gravitational field (green) over time. Left and right plots show different time scales of the behavior.

Compared to the unperturbed motion, the behavior of the Keplerian elements under zonal perturbation is quite complex. There are multiple oscillations happening at different frequencies. On the short time scale (left plots in Fig. 4.1) there is the semi-periodic behavior associated with one orbital revolution with a nodal period of roughly 103 min. As already mentioned in the introduction, the zonal perturbation introduces a drift of the orbital plane, which is indicated by the increasing Ω in Fig. 4.1. The corresponding long term behavior suggests that the orbital plane is rotating around the symmetry axis in about 365 days. However, as we will discover in Sec. 4.4.1 and in particular in Fig. 4.4 neither the nodal period T_d nor the drift in the ascending node are constant, but they are also oscillating. The nodal period T_d, the RAAN-drift $\Delta\Omega$, and the long term behavior of a, e, and i are oscillating at the frequency of the rotation of the argument of periapsis ω, which has a period of roughly 129 days.

4.2.5 The Bounded Motion Conditions by Xu *et al.*

Considering that each orbit is individually influenced by the gravitational perturbations determining its shape and orbital period, bounded motion conditions link two orbits in space-time.

Xu *et al.* [90] showed that the conditions for bounded motion between two orbits $\mathcal{O}_1$ and $\mathcal{O}_2$ require the following conditions to be met:

$$\overline{T}_d\left(\mathcal{O}_1\right) = \overline{T}_d\left(\mathcal{O}_2\right) \tag{4.9}$$

$$\overline{\Delta\Omega}\left(\mathcal{O}_1\right) = \overline{\Delta\Omega}\left(\mathcal{O}_2\right). \tag{4.10}$$

In other words, any two orbits are in sync, if both, their average nodal period $\overline{T}_d$ and their average drift of the ascending node $\overline{\Delta\Omega}$, are the same.

The time related condition is linked to the space related condition by the space-time event at the ascending node, where the object passes through the equatorial plane from south to north. The time difference between two consecutive ascending nodes is the nodal period T_d. The angular difference between two consecutive ascending nodes is denoted by $\Delta\Omega$, also referred to as the RAAN-drift. It is defined by

$$\Delta\Omega = \phi\left(\Omega_{n+1}\right) - \phi\left(\Omega_n\right) - 2\pi\mathrm{sgn}\left(\mathcal{H}_z\right), \tag{4.11}$$

where $-2\pi\mathrm{sgn}\left(\mathcal{H}_z\right)$ ensures that $\Delta\Omega$ is the shortest angular distance between the two consecutive ascending nodes.

Under zonal perturbation, the nodal period T_d and the RAAN-drift $\Delta\Omega$ show regular oscillatory behavior (see Fig. 4.4), making their average values constants of motion. The basic goal of our approach is finding a way of cleverly calculating those average values and relating them to the constants of motion $\mathcal{H}_z$ and E. Given the relation, $\mathcal{H}_z$ and E can be chosen such that the bounded motion conditions are satisfied and the associated orbits are bound.

4.2.6 The Fixed Point Orbit

Under zonal perturbation, there are special orbits for which the nodal period T_d and the RAAN-drift $\Delta\Omega$ are constant. The associated reduced state $\mathcal{Z} = (\rho, v_\rho, z = 0, v_z)$ at the ascending nodes remains

unchanged, which is why these orbits are called fixed point orbits. The orbits are also known as quasi-circular orbits, which originates from the idea of having the elliptical reference shape of the orbit rotate within the orbital plane under zonal perturbation. Given that $r = \rho$ is constant at the ascending node for those orbits would suggest that the reference shape is a circle. The Keplerian elements of such a quasi-circular orbit (see Fig. 4.2) show however that e oscillates around a value

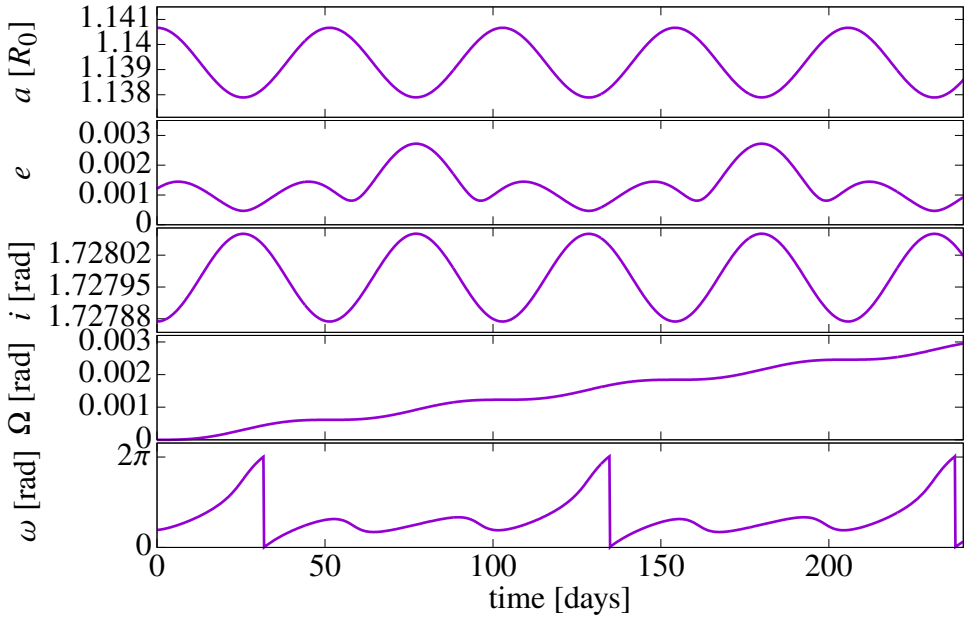

Figure 4.2: Keplerian elements of a quasi-circular low Earth orbit under Earth's zonal gravitational perturbation.

slightly greater than zero, which is the reason for the word 'quasi'. More insightful is the idea that the perturbations influence the orbit just right to yield periodic behavior after just one orbital revolution around the central body.

Compared to the Keplerian elements of non-quasi-circular orbits like the one shown in Fig. 4.1, the orbital behavior of the quasi-circular orbit is a lot more regular. Its nodal period T_d and ascending node drift $\Delta\Omega$ are constant and not oscillating as Fig. 4.4 reveals. Since the long term oscillation has no amplitude, the entire dynamics of a quasi-circular orbit are already captured by the time scale of minutes shown in Fig. 4.2.

For our approach, these fixed point orbits serve as a reference for entire families of orbits which all share the same average nodal period T_d and the same average RAAN-drift $\Delta\Omega$. Our method

calculates a manifold in $(\rho, v_\rho, z, v_z, \mathcal{H}_z, E)$ around the fixed point, where the manifold is defined such that any two points on the manifold satisfy the bounded motion condition.

In the fully gravitationally perturbed system the axial symmetry vanishes, which introduces a ϕ dependence and results in $\mathcal{H}_z$ no longer being a constant of motion. Accordingly, fixed point orbits in the fully gravitationally perturbed systems must have a fixed point property in the full state $\mathcal{X} = (\rho, v_\rho, z = 0, v_z, \phi, \mathcal{H}_z)$. We will discuss fixed point orbits in the fully perturbed system and the possibilities of creating bounded motion manifolds around them in more detail later in this chapter, but first, we will present the method and results from [88], where manifolds of bounded motion orbits for the zonal problem are calculated.

4.3 Method of Bounded Motion Design Under Zonal Perturbation [88]

The goal is to generate a Poincaré return map $\mathcal{P}$ that describes the dynamics of the system by characterizing how a state $(\mathcal{X}_{\text{ini}}, t = 0) \in \mathcal{O}$ within a Poincaré surface $\mathbb{S}$ returns to $\mathbb{S}$. Defining a suitable Poincaré surface is the first step in generating the map. Secondly, a reference orbit with fixed point properties has to be identified to ensure that the expansion point of the map returns to itself. The Poincaré return map is then calculated as an expansion around the reference orbit before being averaged using DA normal form methods. This yields the average nodal period $\overline{T}_d$ and average ascending node drift $\overline{\Delta\Omega}$ as a function of the system parameters and expansion variables around the reference orbit. Using DA inversion methods, the system parameters can be determined such that the bounded motion conditions are met.

4.3.1 The Poincaré Surface Space

The bounded motion conditions are defined regarding the ascending node of two orbits. To be able to enforce the bounded motion condition on our map, we choose the set of ascending nodes $(z = 0, v_z \geq 0)$ as the Poincaré surface. The Poincaré surface $\mathbb{S}_{\Omega}$ can be divided into subsurfaces $\mathbb{S}_{\Omega,\mathcal{H}_z,E}$ for specific angular momentum components $\mathcal{H}_z$ and energies E. These surfaces contain all states with the parameters $(\mathcal{H}_z, E)$ that lie in the equatorial plane $(z = 0)$ and satisfy $v_z > 0$. The

restriction of v_z to positive values makes the relation between E and v_z (Eq. (4.6)) bijective and therefore locally invertible in $\mathbb{S}_{\Omega,\mathcal{H}_z,E}$, so

$$\mathbb{S}_{\Omega,\mathcal{H}_z,E} = \left\{ \mathcal{X} \mid z = 0,\ v_z = \sqrt{2\left(E - U\left(r\right)\right) - v_r^2 - \left(\frac{\mathcal{H}_z}{r}\right)^2} \right\}. \tag{4.12}$$

This means that any state $\mathcal{X} \in \mathbb{S}_{\Omega,\mathcal{H}_z,E}$ is uniquely determined by (r, v_r, ϕ), since $z = 0$ and $v_z\left(r, v_r, \mathcal{H}_z, E\right)$.

4.3.2 The Fixed Point Orbit

The orbit associated with the fixed point state is called reference orbit. The reference orbit has the special property that it returns to the same reduced state $\mathcal{Z} = (r, v_r, z, v_z)^T$ after each revolution with a constant nodal period $T_d^\star$ and a constant angle advancement in ϕ, which is also referred to as the fixed point drift in the ascending node $\Delta\Omega^\star$.

For a certain set of parameters $(\mathcal{H}_z, E)$, we use DA inversion techniques iteratively to find the fixed point orbit. The iteration is initialized with the state

$$\mathcal{Z}_0 = \left(r = -1/(2E), v_r = 0, z = 0, v_z\left(r, \mathcal{H}_z, E\right)\right)^T \tag{4.13}$$

at its ascending node Ω $(v_z > 0)$ and the state is expanded in the variables (r, v_r). After a full orbit integration until the next ascending node intersection, the map $\mathcal{M}$ is timewise projected onto the Poincaré surface $\mathbb{S}_{\Omega,\mathcal{H}_z,E}$ (see Sec. 2.2). The resulting Poincaré map $\mathcal{P}$ represents the one turn map in dependence on variations $(\delta r, \delta v_r)$ in the variables (r, v_r). The difference between the constant part of the map $\mathcal{P}$ and the initial state $\mathcal{Z}_0$ in the components r and v_r is denoted by Δr and Δv_r, respectively. The Poincaré map without its constant part is indicated by $\mathcal{P}'$. The next initial state $\mathcal{Z}_1$ for the iterative process will be given by the evaluation of

$$\begin{pmatrix} \mathcal{Z}_{r,1} \\ \mathcal{Z}_{v_r,1} \end{pmatrix} = \begin{pmatrix} \mathcal{P}'_r\left(\delta r, \delta v_r\right) - \delta r \\ \mathcal{P}'_{v_r}\left(\delta r, \delta v_r\right) - \delta v_r \end{pmatrix}^{-1} \left(\delta r = -\Delta r, \delta v_r = -\Delta v_r\right). \tag{4.14}$$

The process is repeated until the offset $(\Delta r, \Delta v_r)$ is smaller than a threshold value e.g. 1E-14.

4.3.3 The Calculation of Poincaré Return Map

Given a fixed point state $\mathcal{Z}^\star$ from Sec. 4.3.2 for the parameter set $(\mathcal{H}_z, E)$, the Poincaré return map $\mathcal{P} : (\mathbb{S}_{\bigcap}, t) \to (\mathbb{S}_{\bigcap}, t)$ is calculated as a DA expansion around that reference orbit. In the first step, the flow $\mathcal{M}$ of the fixed point and its neighborhood in $\mathbb{S}_{\bigcap}$ (expansion in $(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E)$) is obtained by integrating the system of ODE's from the initial state until the reference/fixed point orbit is an element of $\mathbb{S}_{\bigcap,\mathcal{H}_z,E}$ again after $T_d^\star$. In other words, the state is integrated until the orbit of $\mathcal{X}^0$ intersects with the equatorial plane from south to north again.

While the reference orbit itself is in $\mathbb{S}_{\bigcap,\mathcal{H}_z,E} \subset \mathbb{S}_{\bigcap}$ after $T_d^\star$, the expansion around the reference orbit is not in $\mathbb{S}_{\bigcap,\mathcal{H}_z+\delta\mathcal{H}_z,E+\delta E} \subset \mathbb{S}_{\bigcap}$ due to changing nodal periods of the orbits within the expansion. In order to project the flow $\mathcal{M}$ after $T_d^\star$ onto the Poincaré surface $\mathbb{S}_{\bigcap,E+\delta\mathcal{H}_z,E+\delta\mathcal{H}_z}$, a timewise projection is calculated following Sec. 2.2 and [34]. The flow $\mathcal{M}$ is expanded in time to find the intersection time $t_{\text{intersec}}(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E)$ such that

$$\mathcal{P}_z = \mathcal{M}_z\left(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E, t_{\text{intersec}}\left(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E\right)\right) = 0 \tag{4.15}$$

and $\mathcal{P} = (\mathcal{M}(t_{\text{intersec}}), T_d^\star + t_{\text{intersec}}) \in (\mathbb{S}_{\bigcap,\mathcal{H}_z+\delta\mathcal{H}_z,E+\delta E}, t) \subset (\mathbb{S}_{\bigcap}, t)$.

The time component $\mathcal{P}_{T_d}$ of the Poincaré return map yields the dependence of the nodal period T_d on the system parameters and expansion variables.

4.3.4 The Normal Form Averaging

Given the fixed point Poincaré return map $\mathcal{P}$ with

$$\mathcal{P}\left(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E\right) = \begin{pmatrix} \mathcal{P}_r\left(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E\right) \\ \mathcal{P}_{v_r}\left(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E\right) \\ \mathcal{P}_z = 0 \\ \mathcal{P}_{v_z}\left(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E\right) \\ \mathcal{P}_\phi\left(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E\right) \\ \mathcal{P}_{T_d}\left(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E\right) \end{pmatrix} \tag{4.16}$$

we are using only the first two components (in r and v_r) of the Poincare map for the calculation of phase space transformation provided by the DA normal form algorithm, since the motion is determined by only the (r, v_r) phase space and the parameters $(\mathcal{H}_z, E)$. The reduced map is denoted by $\mathcal{K} = (\mathcal{P}_r, \mathcal{P}_{v_r})^T$.

The normal form transformation $\mathcal{A}(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E)$ (see Eq. (2.28)) and its inverse are used to transform the map $\mathcal{K}$ such that

$$\mathcal{A} \circ \mathcal{K} \circ \mathcal{A}^{-1} \left(q_{\mathrm{NF}}, p_{\mathrm{NF}}, \delta\mathcal{H}_z, \delta E\right) = \mathcal{K}_{\mathrm{NF}} \left(q_{\mathrm{NF}}, p_{\mathrm{NF}}, \delta\mathcal{H}_z, \delta E\right) \tag{4.17}$$

is rotational invariant in the normal form phase space coordinates $(q_{\mathrm{NF}}, p_{\mathrm{NF}})$ up to the order of calculation. In other words, the distorted phase space curves in original phase space coordinates $(\mathcal{P}_r(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E), \mathcal{P}_{v_r}(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E))$ are transformed to circles in the normal form coordinates $(Q_{\mathrm{NF}}(q_{\mathrm{NF}}, p_{\mathrm{NF}}, \delta\mathcal{H}_z, \delta E), P_{\mathrm{NF}}(q_{\mathrm{NF}}, p_{\mathrm{NF}}, \delta\mathcal{H}_z, \delta E))$ as Fig. 4.3 illustrates.

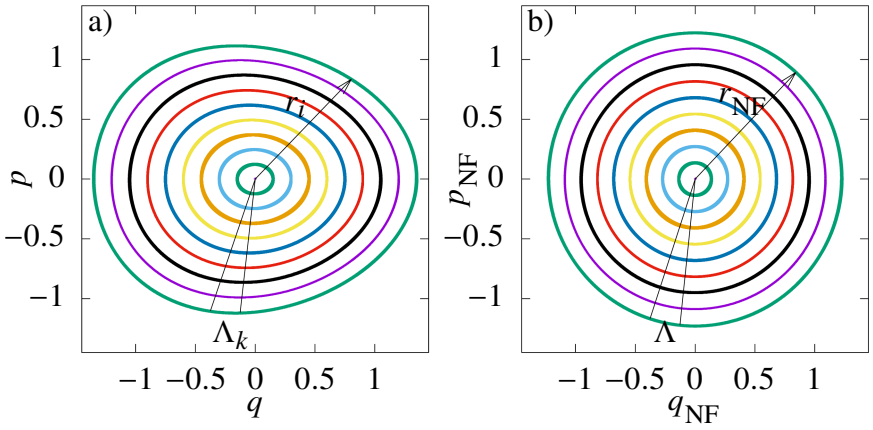

Figure 4.3: a) Distorted phase space behavior in the original phase space (q, p) and b) circular behavior in the corresponding normal form phase space $(q_{\mathrm{NF}}, p_{\mathrm{NF}})$. In a), the phase space angle advancement Λ_k and the phase space radius r_i are not constant by continuously change along each of the phase space curves. In b), the phase space behavior is rotationally invariant ('normalized') with a constant radius r_{NF} and a constant but amplitude dependent angle advancement $\Lambda(r_{\mathrm{NF}})$.

By rewriting the normal form coordinates $(q_{\mathrm{NF}}, p_{\mathrm{NF}})$ in an action-angle representation $(r_{\mathrm{NF}}, \Lambda)$ with

$$\begin{pmatrix} q_{\mathrm{NF}} \\ p_{\mathrm{NF}} \end{pmatrix} = r_{\mathrm{NF}} \begin{pmatrix} \cos\Lambda \\ \sin\Lambda \end{pmatrix}, \tag{4.18}$$

each normal form phase space curve is characterized by the normal form radius (action) r_{NF} and the path along each curve is parameterized by the angle Λ. Using the inverse normal form transformation $\mathcal{A}^{-1}$ (see Eq. (2.29)), the original phase space variables $(\delta r, \delta v_r)$ of $\mathcal{P}$ (and $\mathcal{K}$) are expressed in terms of the action-angle representation and variations in the system parameters $(\delta\mathcal{H}_z, \delta E)$:

$$(\delta r, \delta v_r) = \mathcal{A}^{-1}\left(q_{\mathrm{NF}}\left(r_{\mathrm{NF}}, \Lambda\right), p_{\mathrm{NF}}\left(r_{\mathrm{NF}}, \Lambda\right), \delta\mathcal{H}_z, \delta E\right). \tag{4.19}$$

The Poincaré map $\mathcal{P}(r_{\mathrm{NF}}, \Lambda, \delta\mathcal{H}_z, \delta E)$ is then averaged over a full phase space revolution, by integrating along the angle Λ:

$$\overline{\mathcal{P}}\left(r_{\mathrm{NF}}, \delta\mathcal{H}_z, \delta E\right) = \frac{1}{2\pi}\oint \mathcal{P}\left(r_{\mathrm{NF}}, \Lambda, \delta\mathcal{H}_z, \delta E\right) d\Lambda. \tag{4.20}$$

The numerical averaging presented in [35] is done in the time domain, which cannot incorporate the slightly different oscillation frequencies of the relevant quantities T_d and $\Delta\Omega$ for the different orbits. The key advantage of the normal form representation is that the different oscillation frequencies are captured by the amplitude dependent angle advancement in the normal form. The generalized parameterization of all normal form phase space curves makes the averaging independent of those differences in the frequency.

Splitting the integration into subsections minimizes the error of the numerical integration and considerably improves the quality and accuracy of the averaging. For n separate parameterization

$$\begin{pmatrix} q_{\mathrm{NF}} \\ q_{\mathrm{NF}} \end{pmatrix} = r_{\mathrm{NF}} \begin{pmatrix} \cos\left(\frac{2\pi(k-1)}{n}\right) & -\sin\left(\frac{2\pi(k-1)}{n}\right) \\ \sin\left(\frac{2\pi(k-1)}{n}\right) & \cos\left(\frac{2\pi(k-1)}{n}\right) \end{pmatrix} \begin{pmatrix} \cos\Lambda \\ \sin\Lambda \end{pmatrix} \quad k \in \{1, 2, ..., n\} \tag{4.21}$$

each section is integrated over the symmetric interval of $\Lambda \in \left[-\frac{\pi}{n}, \frac{\pi}{n}\right]$.

The result of the averaging yields every component of $\mathcal{P}$ averaged over a full phase space curve. In particular, it yields the averaged drift in the ascending node $\overline{\Delta\Omega}\left(r_{\mathrm{NF}}, \delta\mathcal{H}_z, \delta E\right)$ and average nodal period $\overline{T}_d\left(r_{\mathrm{NF}}, \delta\mathcal{H}_z, \delta E\right)$.

For mission design purposes the abstract quantity r_{NF} is expressed by the original coordinates $(\delta r, \delta v_r)$ and the parameters $(\delta\mathcal{H}_z, \delta E)$ with

$$r_{\mathrm{NF}}^2(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E) = \left(q_{\mathrm{NF}}^2 + p_{\mathrm{NF}}^2\right)(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E) \tag{4.22}$$

using the normal form transformation $\mathcal{A}$, which yields how $(q_{\mathrm{NF}}, p_{\mathrm{NF}})$ depend on the original coordinates $(\delta r, \delta v_r)$ and the parameters $(\delta\mathcal{H}_z, \delta E)$.

The average drift in the ascending node $\overline{\Delta\Omega}\,(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E)$ and the average nodal period $\overline{T}_d\,(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E)$ are then projected such that the bounded motion conditions are satisfied, with

$$\Delta\Omega^\star = \overline{\Delta\Omega}\,(\delta r, \delta v_r, \delta\mathcal{H}_z\,(\delta r, \delta v_r)\,, \delta E\,(\delta r, \delta v_r)) \tag{4.23}$$

$$T_d^\star = \overline{T}_d\,(\delta r, \delta v_r, \delta\mathcal{H}_z\,(\delta r, \delta v_r)\,, \delta E\,(\delta r, \delta v_r))\,. \tag{4.24}$$

In this process, DA inversion methods are used to find $\delta\mathcal{H}_z(\delta r, \delta v_r)$ and $\delta E(\delta r, \delta v_r)$. The dependence of $\mathcal{H}_z$ and E on orbital parameters for bounded motion orbits became apparent already in [84, 70].

Theoretically, one could have proceeded with the abstract invariant of motion r_{NF} to satisfy the bounded motion condition with $\delta\mathcal{H}_z(r_{\mathrm{NF}})$ and $\delta E(r_{\mathrm{NF}})$. For specific bounded orbits one would then have chosen a value for r_{NF} to calculate $(\delta\mathcal{H}_z, \delta E)$ and afterwards the initial values for (r, v_r) by using Eq. (4.19), where Λ can be chosen freely.

4.4 Bounded Motion Results from [88]

We will now apply the normal form methods for bounded motion of low Earth and medium Earth orbits. For this, we use fixed point orbits of the zonal problem that have previously been investigated by He *et al.* [35] for the low Earth orbit (LEO) and Baresi and Scheeres [6] for the medium Earth orbit (MEO).

As explained above, the fixed point Poincaré maps $\mathcal{P}$ are calculated as an expansion in the variables $(\delta r, \delta v_r, \delta\mathcal{H}_z, \delta E)$ around the respective fixed point orbit. In the calculation we consider zonal perturbations up to the J_{15}-term, since investigations in [35] indicated no considerable influence of J_k terms for $k > 15$. We are using maps of 8th order, which provide the best balance of accuracy and computation time. Additionally, the following dimensionless units are used: distances are considered in units of the average Earth radius $R_0 = 6378.137$ km and time is considered in units of $T_0 = 806.811$ s such that the gravitational constant assumes the value $\mu = 1$.

It will be shown that the DANF method provides entire sets of bounded motions that extend far beyond the realistic/practical scope. Since the approach is based on polynomial expansions, it is obvious it will have to fail at some point. After presenting the bounded motion results for the LEO and MEO case, we take a look at the limitations of the DANF method and the resulting sets for very large distances between orbits.

4.4.1 Bounded Motion in Low Earth Orbit

In a first comparison, we are investigating bounded motion around a pseudo-circular LEO that was also considered in [35]. The pseudo-circular orbit corresponds to the reduced fixed point state

$$(r^\star, v_r^\star) = (1.14016749, -1.05621369\text{E-}3) \tag{4.25}$$

for the parameters $(\mathcal{H}_z, E) = (-0.16707295, -0.43870527)$. The orbit has a fixed nodal period of $T_d^\star = 7.64916169$ ($\approx$ 103 min) and a constant ascending node drift of $\Delta\Omega^\star = 1.22871195\text{E-}3$ rad (0.0704°). The vertical position z of the Poincaré fixed point orbit are defined by the Poincaré section ($z = 0$) and Eq. (4.12) with $v_z^\star\left(r^\star, v_r^\star, \mathcal{H}_z, E\right) = 0.92518953$.

The computation of the Poincaré map took 165 seconds on a Lenovo E470 with an Intel®Core™ i5-7200U CPU 2.5GHz. The map confirms the fixed point property of the orbit, since the offset of the constant part of the map from the initial coordinates is well within the numerical error of the integration with $(\Delta r, \Delta v_r, \Delta z, \Delta v_z) = (4\text{E-}15, 5\text{E-}13, -1\text{E-}15, -4\text{E-}15)$. The normal form transformation of the reduced fixed point Poincaré map $\mathcal{K} = (\mathcal{P}_r, \mathcal{P}_{v_r})^T$ is calculated via the DA normal form algorithm (in 90 milliseconds). The circular phase space behavior in normal form space is parameterized using the action-angle notation (r_{NF}, Λ). The phase space parameterization is transformed back to the original coordinates of the Poincaré map. The Poincaré map is averaged (in 52 milliseconds) over a full phase space rotation using 8 subsections following the procedure outlined in Sec. 4.3.4. Afterwards, the variable r_{NF} is expressed in terms of δr, δv_r, $\delta\mathcal{H}_z$ and δE before the variations in the constants of motion $(\delta\mathcal{H}_z, \delta E)$ are matched dependent on $(\delta r, \delta v_r)$ such that the averaged expressions for T_d and $\Delta\Omega$ satisfy the bounded motion conditions (Eq. (4.23) and

Eq. (4.24)).

Considering bounded orbits initiated with the same v_r as the pseudo-circular orbit ($\delta v_r = 0$), the dependence of $\mathcal{H}_z(\delta r, \delta v_r = 0)$ and $E(\delta r, \delta v_r = 0)$ are provided in Tab. 4.1 below.

Table 4.1: The expansion of $\mathcal{H}_z(\delta r, \delta v_r = 0)$ and $E(\delta r, \delta v_r = 0)$ for relative bounded motion orbits with an average nodal period $\overline{T_d} = 7.64916169$ (≈103 min) and an average ascending node drift of $\overline{\Delta\Omega} = 1.22871195\text{E-}3$ rad. The expansion is relative to the pseudo-circular LEO from [35].

$\mathcal{H}_z(\delta r, \delta v_r = 0) =$		$E(\delta r, \delta v_r = 0) =$	
-0.16707295		-0.43870527	
$+0.32072807$	δr^2	$-0.31602983\text{E-}3$	δr^2
$+0.25767948\text{E-}3$	δr^3	$-0.25390482\text{E-}6$	δr^3
-0.19132824	δr^4	$-0.31003174\text{E-}3$	δr^4
$+0.53296708\text{E-}4$	δr^5	$-0.85361819\text{E-}6$	δr^5
$+0.12006391\text{E-}1$	δr^6	$-0.32152252\text{E-}3$	δr^6
$+0.60713391\text{E-}3$	δr^7	$-0.24661573\text{E-}5$	δr^7
-0.19751494	δr^8	$-0.21784073\text{E-}3$	δr^8

To show that the expansion of $\delta\mathcal{H}_z$ and δE provide relative bounded motion orbits, we illustrate the long term behavior of three LEOs relative to one another. The first orbit is the fixed point/pseudo-circular orbit and is denoted by $\mathcal{O}_0$. The second orbit ($\mathcal{O}_1$) is initiated at $\delta r = 0.06$ with $\delta v_r = 0$. The third orbit ($\mathcal{O}_2$) is initiated at $\delta r = 0.13$ with $\delta v_r = 0$. The last two both have an initial longitudinal offset of $\phi = 0.5°$ relative to $\mathcal{O}_0$. The specific values of the orbits are given in Tab. 4.2.

Table 4.2: The LEOs below are all initiated at $v_{r,0} = -1.05621369\text{E-}3$ and $r_0 = 1.14016749 + \delta r$, and have an average nodal period of $\overline{T_d} = 7.64916169$ (≈103 min) and an average ascending node drift of $\overline{\Delta\Omega} = 1.22871195\text{E-}3$ rad. The pseudo-circular LEO from [35] is denoted by $\mathcal{O}_0$.

	δr	δv_r	ϕ	$\mathcal{H}_z$	E
$\mathcal{O}_0$	0.00	0	0.0°	-0.16707295	-0.43870527
$\mathcal{O}_1$	0.06 (383 km)	0	0.5°	-0.16592075	-0.43870642
$\mathcal{O}_2$	0.13 (829 km)	0	0.5°	-0.16170668	-0.43871071

In Fig. 4.4 we show that the bounded motion conditions are met: the oscillatory behavior of the nodal period T_d and the ascending node drift $\Delta\Omega$ of the two orbits $\mathcal{O}_1$ and $\mathcal{O}_2$ average out to the same value, respectively, which corresponds to the constant nodal period $T_d^\star$ and constant ascending node drift $\Delta\Omega^\star$ of the fixed point orbit $\mathcal{O}_0$.

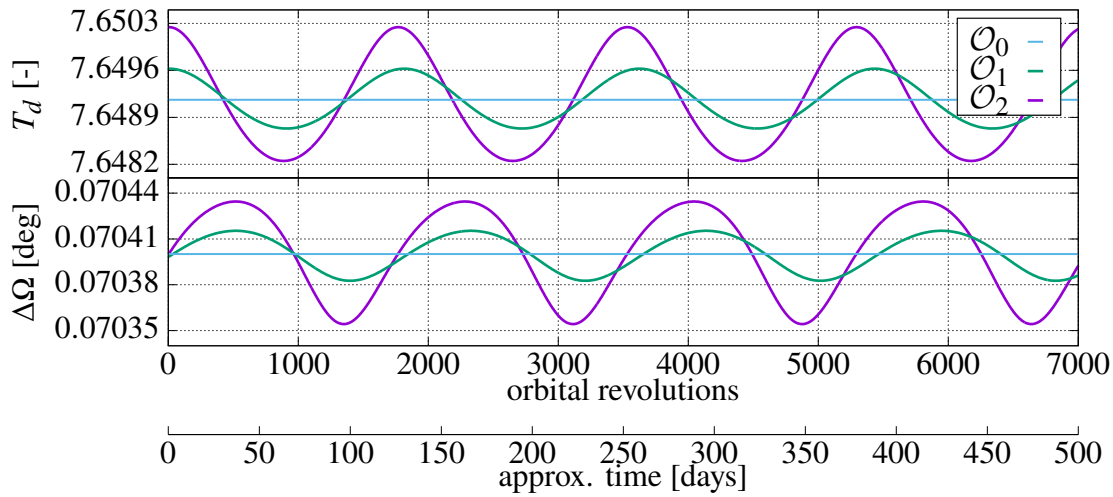

Figure 4.4: Oscillatory behavior of the bounded motion quantities T_d and $\Delta\Omega$ of the bounded LEOs $\mathcal{O}_1$ and $\mathcal{O}_2$ initiated at $\delta r = 0.06$ and $\delta r = 0.13$, respectively. Additionally, the constant nodal period $T_d^\star = 7.64916169$ and constant ascending node drift of $\Delta\Omega^\star = 0.0704°$ of the fixed point orbit $\mathcal{O}_0$ are shown. The periods of oscillation are 1763 orbital revolutions (126 days) for $\mathcal{O}_2$, 1810 orbital revolutions (129 days) for $\mathcal{O}_1$, and 1823 orbital revolutions (130 days) for $\delta r \to 0$ of $\mathcal{O}_0$. The shown results are generated by numerical integration. The time domain is based on the average orbital revolution $\approx T_d^\star$.

The bounded motion is further confirmed by Fig. 4.5, which shows the total distance between the three LEOs respectively for 14 years. Furthermore, Fig. 4.5 illustrates the relative radial and along-track distance between the orbit pairs from the perspective of one of the orbits in the pair.

Apart from yielding long term bounded motion, the normal form methods also provide the average angle advancement Λ in the (r, v_r) phase space. This angle advancement is directly linked to the rotation frequency ω_p of the orbit (and its apsides) within its orbital plane, which causes the oscillation of T_d and $\Delta\Omega$ shown in Fig. 4.4 with ω_p. One (r, v_r) phase space rotation corresponds to one revolution of the orbit (and its apsides) within its orbital plane. Accordingly, the frequency $\omega_p = \Lambda/2\pi$ is equivalent to the definition of the tune and the tune shifts $\nu + \delta\nu$, which are just the normalized angle advancement separated into its constant part (the tune ν) and its amplitude dependent part (the tune shifts $\delta\nu$). The normal form yields the average angle advancement Λ dependent on $(r_{\mathrm{NF}}, \delta\mathcal{H}_z, \delta E)$. After normalizing Λ, by division by 2π, and replacing r_{NF} by an expression of $(\delta r, \delta v_r)$ and $(\delta\mathcal{H}_z, \delta E)$ according to Eq. (4.22), and using the expressions for $(\delta\mathcal{H}_z(\delta r, \delta v_r), \delta E(\delta r, \delta v_r))$ from earlier, the frequency $\omega_p(\delta r, \delta v_r)$ is obtained for the bounded

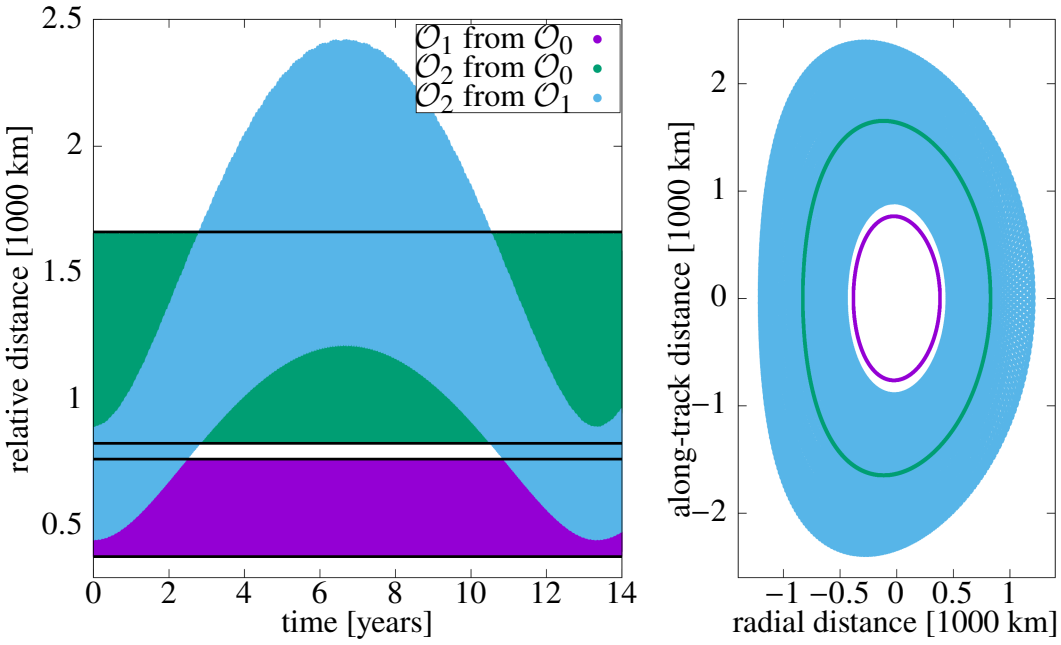

Figure 4.5: Relative bounded motion of LEOs with an average nodal period of $\overline{T_d}$ = 7.64916169 (≈103 min) and an average node drift of $\overline{\Delta\Omega}$ = 1.22871195E-3 rad for 14 years. The total relative distance between the orbits is shown in the left plot and the right plot shows the relative radial and along-track distance between orbit pairs from the perspective of one of the orbits in the pair. The oscillation in the relative distance between $\mathcal{O}_2$ and $\mathcal{O}_1$ is caused by the rotating orbital orientation of the orbits at different frequencies.

motion orbits around the fixed point LEO. The coefficients of ω_p for $\delta v_r = 0$ are given in Tab. 4.3 below.

Table 4.3: Expansion of $\omega_p(\delta r, \delta v_r = 0)$ of relative bounded motion LEOs with an average nodal period $\overline{T_d}$ = 7.64916169 (≈103 min) and an average node drift of $\overline{\Delta\Omega}$ = 1.22871195E-3 rad. The expansion is relative to the pseudo-circular LEO from [35].

$\omega_p(\delta r, \delta v_r = 0) =$	
+0.54868728E-3	
+0.10803872E-2	δr^2
+0.86800515E-6	δr^3
+0.10552068E-2	δr^4
+0.29106874E-5	δr^5
−0.76284414E-3	δr^6
+0.39324207E-5	δr^7
−0.35077526E-1	δr^8

Accordingly, the periods of the oscillations of the nodal periods T_d and the ascending node

drifts $\Delta\Omega$ in Fig. 4.4 (in units of orbital revolutions) are just the inverse of the frequencies $\omega_p(\delta r = 0.06) = 5.52590498\text{E-}4$ and $\omega_p(\delta r = 0.13) = 5.67242676\text{E-}4$. These frequencies also help explain the oscillation of the total relative distance range between $\mathcal{O}_1 \rightleftarrows \mathcal{O}_2$ over 13.3 years in Fig. 4.5.

While $\mathcal{O}_1$ shows repetitive behavior after 1809.7 orbital revolutions (129.3 days), the behavior of $\mathcal{O}_2$ is repetitive after 1762.9 orbital revolutions (125.9 days). Accordingly, the two orbits will be in and out of sync regarding their orbital orientation, while maintaining bounded due to the matching average nodal period and ascending node drift. Specifically, the two orbits will be back in sync after about 68170 orbital revolutions (4869 days/13.3 years) as Fig. 4.5 illustrates, since $\mathcal{O}_1$ will have turned 37.7 times while $\mathcal{O}_2$ will have turn exactly once less, namely, 36.7 times, bringing them both back into the same orbital orientation to one other before moving apart again.

In conclusion, our first comparison showed the superiority of the normal form methods, particularly, compared to the iterative map evaluation method in [35], where numerical adjustments to the method were required to provide long term relative bounded motion for $\delta r = 0.11$.

In Sec. 4.4.3 we will show that the DANF method even provides hypothetical long term bounded motion up to $\delta r = 0.3$, which covers all realistic cases until $\delta r = 0.14$ and further hypothetical (non-practical) cases with altitudes below the Earth's surface.

In the next comparison, we are going to investigate bounded motion much farther from the Earth's surface. Accordingly, we expect a larger theoretical and practical bounded motion range from the DANF method, due to a weaker influence of the zonal perturbations.

4.4.2 Bounded Motion in Medium Earth Orbit

In this comparison, we are considering a medium Earth orbit (MEO) from [6, p. 11] initiated at $r = 26562.58$ km, $v_r = -9.05\text{E-}4$ km/s and $v_z = 3.18$ km/s. In the units of $R_0 = 6378.137$ km and $T_0 = 806.811$ s, the zonal problem with J_2 to J_{15} yields a fixed point orbit at $(r^\star, v_r^\star) = (4.17198963, -1.14150072\text{E-}4)$ and $v_z^\star = 0.40154964$ for the parameters $(\mathcal{H}_z, E) = (1.16863390, -0.11984818)$. The fixed point orbit has a fixed nodal period

$T_d^\star = 53.5395648$ (≈12 hours) and constant drift in the ascending node of $\Delta\Omega^\star = -3.35410945\text{E-}4$ rad (-0.0192°).

The same computer system as in Sec. 4.4.1, took 131 seconds for the computation of the map. The offset of the integration with $(\Delta r, \Delta v_r, \Delta z, \Delta v_z) = (-4\text{E-}15, -2\text{E-}13, -4\text{E-}15, 2\text{E-}16)$ is well within the range of the numerical error of the integration. After the normal form transformation (in 100 milliseconds) and the averaging (in 62 milliseconds) following the same procedure as in 4.4.1, the dependencies of the constants of motion $(\mathcal{H}_z, E)$ on $(\delta r, \delta v_r)$ were calculated. Below, Tab. 4.4 yields $\mathcal{H}_z(\delta r, \delta v_r = 0)$ and $E(\delta r, \delta v_r = 0)$.

Table 4.4: The expansion of $\mathcal{H}_z(\delta r, \delta v_r = 0)$ and $E(\delta r, \delta v_r = 0)$ for relative bounded motion MEOs with an average nodal period of $\overline{T_d} = 53.5395648$ (≈ 12 h) and an average ascending node drift of $\overline{\Delta\Omega} = -3.35410945\text{E-}4$ rad. The expansion is relative to the pseudo-circular MEO from [6].

$\mathcal{H}_z(\delta r, \delta v_r = 0) =$		$E(\delta r, \delta v_r = 0) =$	
$+1.16863390$		-0.11984818	
-0.16787983	δr^2	$-0.11295792\text{E-}05$	δr^2
$-0.57819536\text{E-}5$	δr^3	$-0.38903865\text{E-}10$	δr^3
$+0.72342680\text{E-}2$	δr^4	$-0.16786161\text{E-}07$	δr^4
$+0.16208617\text{E-}6$	δr^5	$-0.34176382\text{E-}11$	δr^5
$-0.69493130\text{E-}4$	δr^6	$-0.28279909\text{E-}08$	δr^6
$+0.11561378\text{E-}6$	δr^7	$+0.27190622\text{E-}12$	δr^7
$+0.54888817\text{E-}4$	δr^8	$-0.51224108\text{E-}10$	δr^8

To illustrate that the DANF methods also provide bounded motion for this set of parameters, we consider the long term behavior of three MEOs relative to one another. The first orbit is the fixed point/pseudo-circular orbit and is denoted by $\mathcal{O}_0$. Since $r^\star$ of the fixed point MEO is about four times the $r^\star$ of the low Earth fixed point orbit from the previous section, the bounded orbits are initiated at four times the distance compared to the LEO investigation in Sec. 4.4.1. The orbit $\mathcal{O}_1$ is initiated at $\delta r = 0.24$ (1531 km) with $\delta v_r = 0$ and $\mathcal{O}_2$ is initiated at $\delta r = 0.52$ (3317 km) with $\delta v_r = 0$. These relative distances are already at the border or larger than distances that are used in practice. Again, both orbits have an initial longitudinal offset of $\phi = 0.5°$ relative to $\mathcal{O}_0$. The specific values of the orbits are given in Tab. 4.5.

Equivalent to Fig. 4.4 we show that the bounded motion conditions are met for the chosen MEOs

Table 4.5: The MEOs below are all initiated at $v_{r,0} = -1.14150072\text{E-}4$ and $r_0 = 4.17198963 + \delta r$, and have an average nodal period of $\overline{T}_d = 53.5395648$ (≈ 12 h) and an average ascending node drift of $\overline{\Delta\Omega} = -3.35410945\text{E-}4$ rad. The orbit $\mathcal{O}_0$ is the pseudo-circular MEO from [6].

	δr	δv_r	ϕ	$\mathcal{H}_z$	E
$\mathcal{O}_0$	0.0	0	0.0°	1.16863390	-0.119848175
$\mathcal{O}_1$	0.24 (1531 km)	0	0.5°	1.15898794	-0.119848240
$\mathcal{O}_2$	0.52 (3317 km)	0	0.5°	1.123766254	-0.119848482

in Fig. 4.6. The oscillatory behavior of the nodal period T_d and the ascending node drift $\Delta\Omega$ of the two orbits $\mathcal{O}_1$ and $\mathcal{O}_2$ average out to the same value, respectively, which correspond to the constant nodal period $T_d^\star$ and constant ascending node drift $\Delta\Omega^\star$ of the fixed point orbit $\mathcal{O}_0$. In contrast to the investigated LEOs, the oscillation period of the bounded motion quantities of the MEOs increases with increasing δr. The period of oscillation in the MEO cases is also about two orders of magnitude longer with periods of 47 and 53 years for $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively, compared to the LEOs.

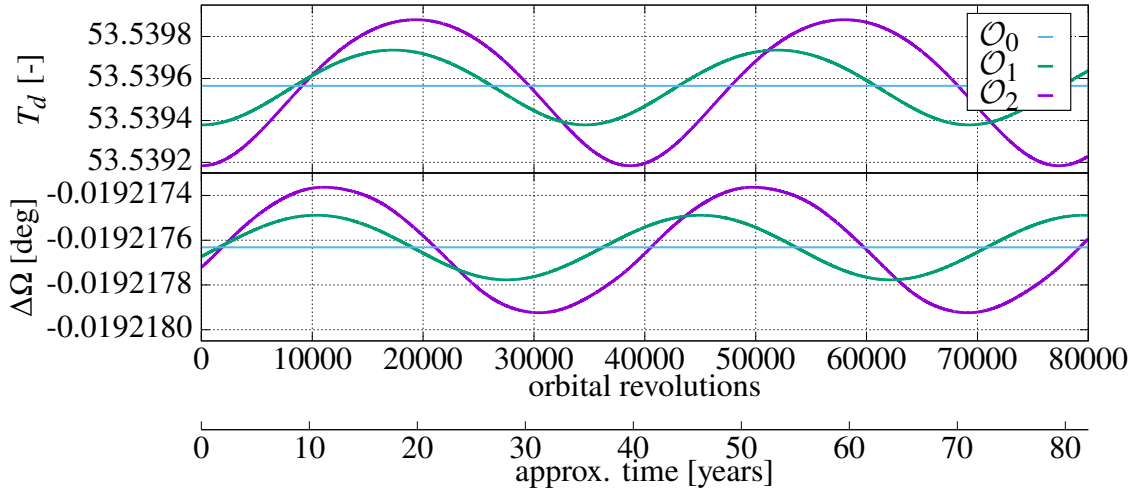

Figure 4.6: Oscillatory behavior of the bounded motion quantities T_d and $\Delta\Omega$ of the bounded MEOs $\mathcal{O}_1$ and $\mathcal{O}_2$ initiated at $\delta r = 0.24$ and $\delta r = 0.52$, respectively. Additionally, the constant nodal period $T_d^\star = 53.5395648$ and constant ascending node drift of $\Delta\Omega^\star = -0.0192176316$ deg of the fixed point orbit $\mathcal{O}_0$ are shown. The periods of oscillation are 38682 orbital revolutions (52.9 years) for $\mathcal{O}_2$, 34621 orbital revolutions (47.4 years) for $\mathcal{O}_1$, and 33671 orbital revolutions (46.1 years) for $\delta r \to 0$ of $\mathcal{O}_0$. The shown results are generated by numerical integration. The time domain was added assuming that on average one orbital revolution $\approxeq T_d^\star$.

Using the normal form methods, the rotation frequency ω_p of the orbital orientation within its

orbital plane is calculated as described in Sec. 4.4.1. The results from the expansion of ω_p confirm these periods of oscillation with $\omega_p(0.24)$ = 2.88842404E-5 and $\omega_p(0.52)$ = 2.58516089E-5. The expansion of ω_p dependent on δr is given in Tab. 4.6.

Table 4.6: Expansion of $\omega_p(\delta r, \delta v_r = 0)$ of relative bounded motion orbits with an average nodal period of $\overline{T}_d$ = 53.5395648 ($\approx$ 12 h) and an average ascending node drift of $\overline{\Delta\Omega}$ = −3.35410945E-4 rad. The expansion is relative to the pseudo-circular MEO from [6].

$\omega_p(\delta r, \delta v_r = 0) =$	
+0.29699500E-04	
−0.14137545E-04	δr^2
−0.48691156E-09	δr^3
−0.22644327E-06	δr^4
−0.43912160E-10	δr^5
−0.10717280E-05	δr^6
−0.10374073E-09	δr^7
+0.23789772E-05	δr^8

Fig. 4.7 shows the long term bounded motion behavior by illustrating the relative total distance between the orbits and their relative radial and along-track distances. Due to the long oscillation periods in the bounded motion quantities of 47 and 53 years for $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively, the oscillation in the total distance between $\mathcal{O}_1$ and $\mathcal{O}_2$ is about 456 years and can therefore only be partially shown. After 456 years the orbital orientation of $\mathcal{O}_1$ will have turned 9.6 times and align again with the orbital orientation of $\mathcal{O}_2$, which will have turned 8.6 times.

The 'breathing' of the relative distance between the orbits is particularly noticeable for the orbit pair of $\mathcal{O}_2$ and $\mathcal{O}_0$. The frequency of the 'breathing'is $2\omega_p$ which is a result of the rotation of the orbital orientation of the pseudo-elliptical $\mathcal{O}_2$ relative to the pseudo-circular $\mathcal{O}_0$. Since the orbital shape of the pseudo-elliptical $\mathcal{O}_2$ is approximately symmetric along its semi-major axis, one full rotation of the orbital orientation corresponds to two breathing cycles.

In conclusion, our methods also provided an entire set of long term relative bounded motion around the considered fixed point MEO from [6], which was validated far beyond practical relative distances. In the following section, the limitations of our method are investigated. The investigations will show that the validity of the sets presented in Sec. 4.4.1 and Sec. 4.4.2 extends over about twice

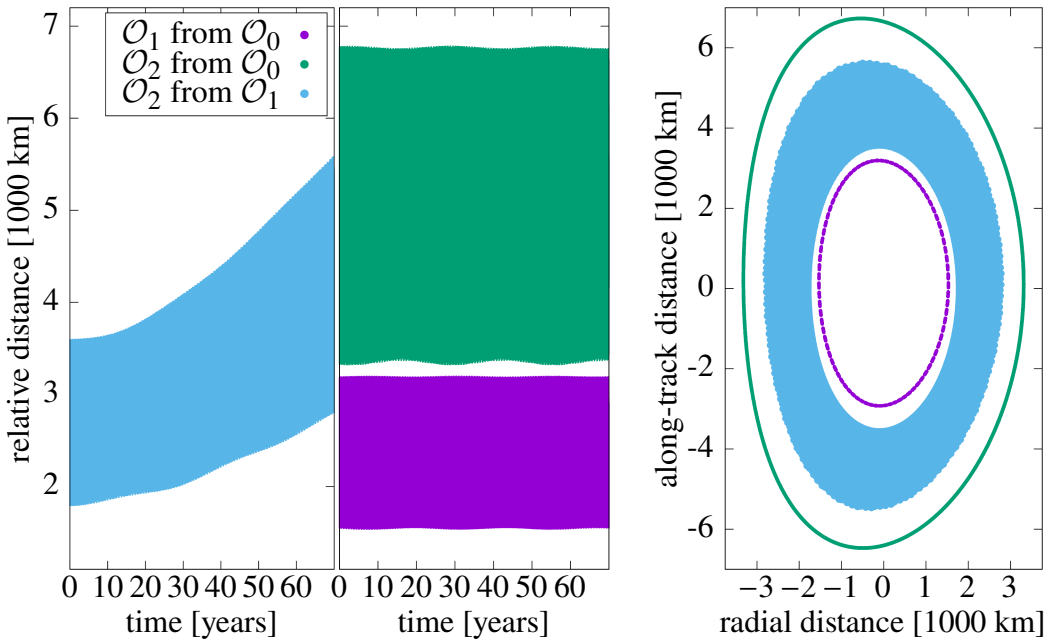

Figure 4.7: Relative bounded motion of MEOs from Tab. 4.5 with an average nodal period of $\overline{T_d}$ = 53.5395648 (≈ 12 h) and an average ascending node drift of $\overline{\Delta\Omega}$ = −3.35410945E-4 rad over 70 years. The total relative distance between the orbits is shown in the left plots and the right plot shows the relative radial and along-track distance between orbit pairs from the perspective of one of the orbits in the pair. The 'breathing' of the relative total distance between $\mathcal{O}_2$ and $\mathcal{O}_0$ originates from the rotating orbital orientation of pseudo-elliptical $\mathcal{O}_2$ relative to the pseudo-circular $\mathcal{O}_0$. Due to the very long rotation periods, only the first 70 years of the relative distance oscillation and radial/along-track behavior between $\mathcal{O}_2$ and $\mathcal{O}_1$ could be shown.

the already presented distance from their respective fixed point orbits.

4.4.3 Testing the Limitations of the DANF Method

The previous two sections illustrated the validity of the DANF method for all practical relative distances for bounded motion and beyond. In this section, we move even further away from any practical relevance of the calculated sets of bounded motion to the limitations of our method. Since it is based on polynomial expansions, it is obvious it will fail at some point and we want to show when and how this failing process takes place.

First we pick a number of test orbits from the calculated bounded motion sets (see Tab. 4.7). In contrast to previous examples, no initial longitudinal offset relative to the respective fixed point orbits are set.

Table 4.7: The following orbit parameters are obtained by evaluating $\mathcal{H}_z(\delta r, \delta v_r = 0)$ and $E(\delta r, \delta v_r = 0)$ from Tab. 4.1 and Tab. 4.4 for various δr keeping $\delta v_r = 0$.

Test LEOs				Test MEOs			
	δr	$\mathcal{H}_z$	E		δr	$\mathcal{H}_z$	E
$\mathcal{O}_0$	0.00	-0.16707295	-0.43870527	$\mathcal{O}_0$	0.0	1.1686339	-0.11984817
$\mathcal{O}_{0.15}$	0.15	-0.15995246	-0.43871254	$\mathcal{O}_{0.6}$	0.6	1.1091311	-0.11984854
$\mathcal{O}_{0.20}$	0.20	-0.15454760	-0.43871843	$\mathcal{O}_{0.7}$	0.7	1.0881027	-0.11984873
$\mathcal{O}_{0.25}$	0.25	-0.14777078	-0.43872632	$\mathcal{O}_{0.8}$	0.8	1.0641420	-0.11984890
$\mathcal{O}_{0.30}$	0.30	-0.13975416	-0.43873648	$\mathcal{O}_{0.9}$	0.9	1.0373802	-0.11984910
$\mathcal{O}_{0.35}$	0.35	-0.13066556	-0.43874929	$\mathcal{O}_{1.0}$	1.0	1.0079682	-0.11984932
$\mathcal{O}_{0.40}$	0.40	-0.12071669	-0.43876526	$\mathcal{O}_{1.1}$	1.1	0.97607833	-0.11984957
				$\mathcal{O}_{1.2}$	1.2	0.94190725	-0.11984984
				$\mathcal{O}_{1.3}$	1.3	0.90567972	-0.11985014
				$\mathcal{O}_{1.4}$	1.4	0.86765361	-0.11985047

Fig. 4.8 illustrates the behavior of the bounded motion quantities T_d and $\Delta\Omega$ for the chosen orbits of the LEO bounded motion set. Both quantities show oscillatory behavior centered at or close to $T_d^\star$ and $\Delta\Omega^\star$, respectively. With increasing distance δr, the influence of higher order oscillations becomes apparent. The frequency and amplitude of oscillation of the bounded motion quantities also increase with increasing distance δr.

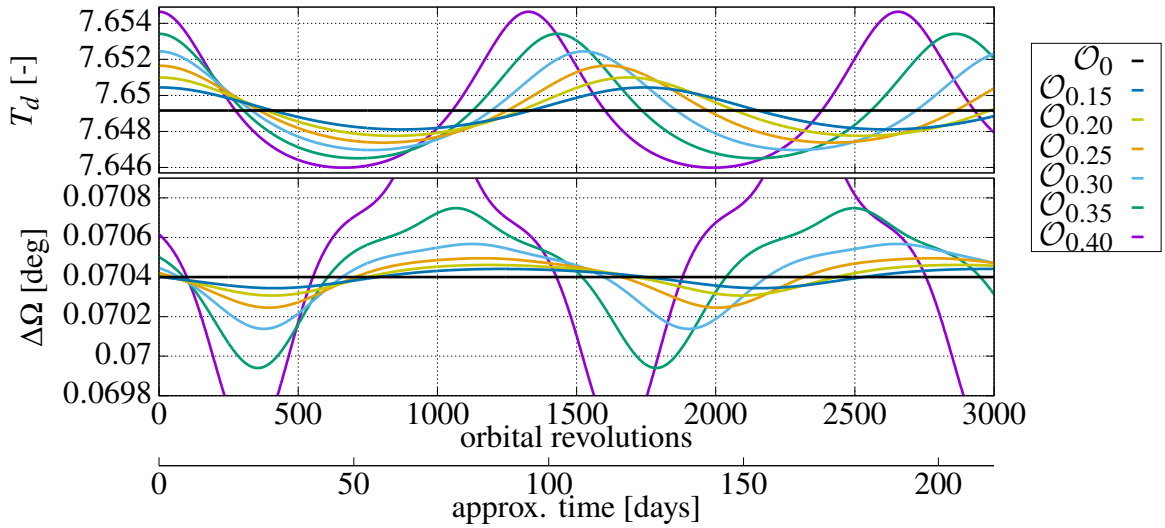

Figure 4.8: The behavior of the bounded motion quantities T_d and $\Delta\Omega$ for the test orbits from Tab. 4.7 of the calculated LEO bounded motion set generated by numerical integration. For large δr, the influences of higher order oscillations are apparent. The frequency and amplitude of oscillation increase with increasing δr. The amplitude of $\Delta\Omega$ is particularly sensitive to δr.

If the bounded motion conditions are not met or only met approximately, the orbits will start drifting apart. This effect is illustrated in Fig. 4.9, which shows very slowly diverging behavior of approximately 2.6 km/year for $\delta r = 0.3$ (1913 km) and a stronger divergence of approximately 10.6 km/year for $\delta r = 0.4$ (2551 km) in the left plot. The thickening curves in the radial/along-track representation of the relative orbit motion are a further indication of divergence. The strength of divergence in Fig. 4.9 can be directly linked to the size of the offsets in the bounded motion quantities from $T_d^\star$ and $\Delta\Omega^\star$, shown in Fig. 4.8.

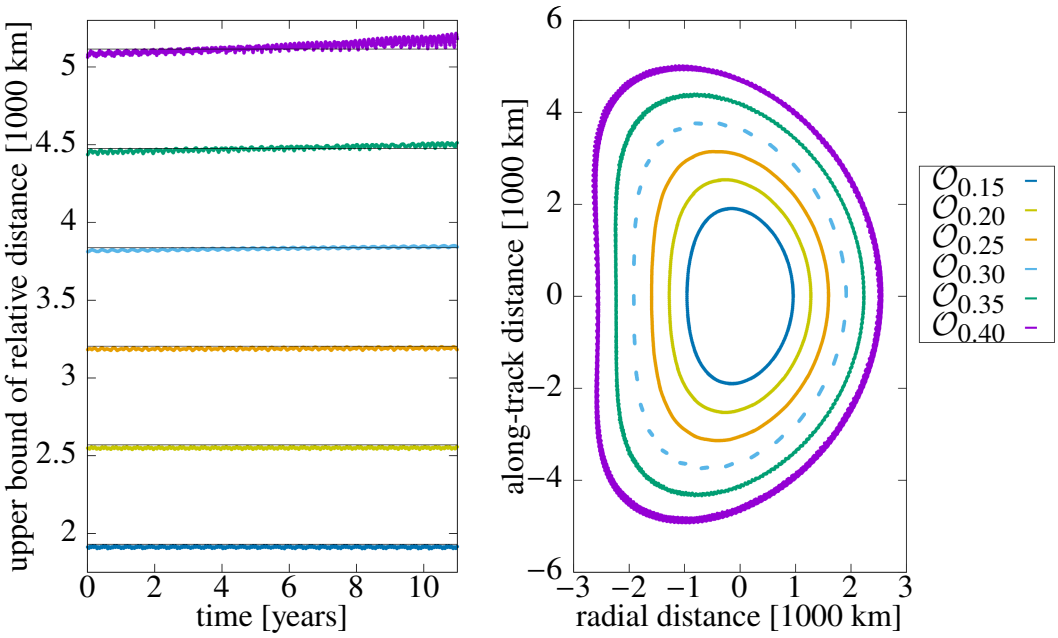

Figure 4.9: Distance between the orbits in the calculated bounded motion set and $\mathcal{O}_0$ is determined in regular time intervals with numerical integration over more than ten years. The left plot only shows the upper bound to avoid overlaps. Thin horizontal lines at the initial upper bound emphasize small changes. The dotted light blue curve (right) originates from an unintended near-resonance between the chosen time interval for distance evaluations and the orbital behavior. A measurable increase in relative distances (left) over 10 years for $\delta r \geq 0.3$ is supported by thickening curves in the radial/along-track behavior (right).

From Fig. 4.9 and Fig. 4.8 we conclude that our method and the resulting expansions in $\mathcal{H}_z$ and E for long term bounded motion of at least 10 years around the fixed point LEO from [35] start to lose their significant accuracy for $\delta r \geq 0.3$ to satisfy the bounded motion conditions with the required precision. Note that $\delta r = 0.3$ (1913 km) is already a purely theoretical orbit with altitudes

of more than 1000 km below the Earth's surface, which means that our expansions in $\mathcal{H}_z$ and E provided reliable bounded motion beyond realistic distances ($\delta r \leq 0.14$) between orbits.

The behavior of the bounded motion quantities T_d and $\Delta\Omega$ for the chosen orbits of the MEO bounded motion set (from Tab. 4.7) are shown in Fig. 4.10. In contrast to the test LEOs, the amplitude and period of oscillation of the bounded motion quantities are decreasing with increasing distance δr, which causes the almost steady behavior of $\delta r = 1.4$ over the shown timespan and generally suppresses higher order oscillations that were seen for the LEOs. While the oscillations of T_d are approximately centered around $T_d^\star$ (except for $\mathcal{O}_{1.4}$), the center of oscillation is increasingly diverging from $\Delta\Omega^\star$ to lower $\Delta\Omega$ for $\delta r \geq 0.8$. In other words, the expansions in $\delta\mathcal{H}_z$ and δE start failing in producing related orbits that satisfy the bounded motion condition.

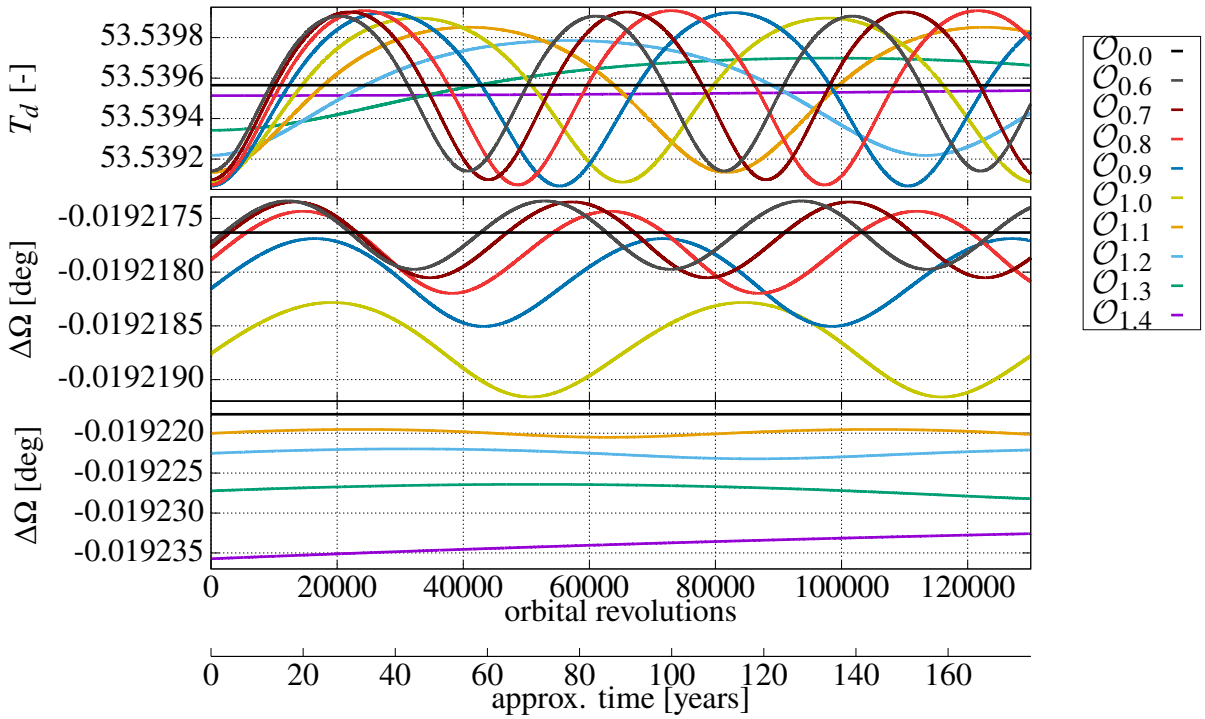

Figure 4.10: Behavior of the bounded motion quantities T_d and $\Delta\Omega$ for the test orbits from Tab. 4.7 of the calculated MEO bounded motion set generated by numerical integration. In contrast to the investigated LEOs, the frequency and amplitude of oscillation decrease with increasing δr such that $\mathcal{O}_{1.4}$ appears almost steady. For $\delta r \geq 0.8$ the center of oscillation of $\Delta\Omega$ start to drift to more negative values and away from $\Delta\Omega^\star$.

The consequence of this offset in the bounded motion condition is diverging behavior between

the orbits, which can be seen in Fig. 4.11. The upper bound of the total distance between the orbits starts diverging for those very large distances and the thickening curves in the radial/along-track representation of the distance of the orbits from the perspective of $\mathcal{O}_0$ further indicate this divergence. Additionally, Fig. 4.11 shows the 'breathing' in the total relative distance between the orbits with $2\omega_p$, which is due to the rotating orbital orientation of the orbits relative to the pseudo-circular fixed point orbit as already mentioned in the section above.

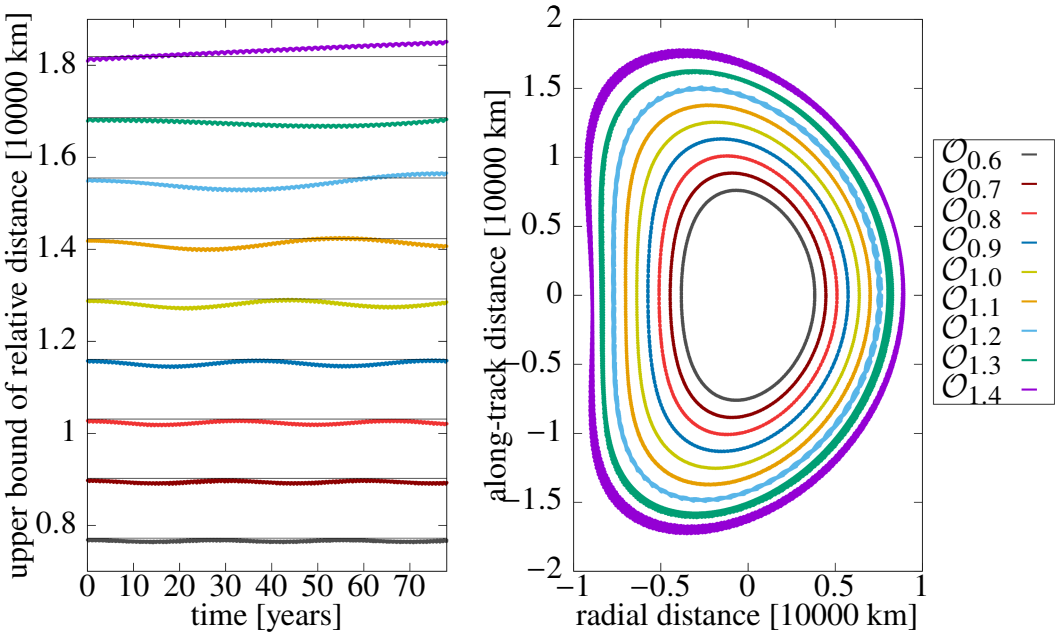

Figure 4.11: Distance between the orbits in the calculated bounded motion set and $\mathcal{O}_0$ is determined in regular time intervals by numerical integration over more than 70 years. The left plot only shows the upper bound to avoid overlaps. Thin horizontal lines at the initial upper bound emphasize small changes. The 'breathing' of the total relative distance from the orbital rotation is clearly visible. Its period increases with increasing δr until being unrecognizable due to the strong divergence for $\delta r \geq 1.4$, which is supported by thinker curves in the right plot. The weaker divergence over the 70-year timespan is already noticeable for $\delta r \geq 0.9$. The divergence is caused by the offset in respective bounded motion quantities (see. Fig. 4.10).

From Fig. 4.11 and Fig. 4.10 we conclude that our method and the resulting expansions in $\mathcal{H}_z$ and E for long term bounded motion of at least 70 years around the fixed point MEO from [6] start to lose their significant accuracy for $\delta r \geq 0.9$ to satisfy the bounded motion conditions with the required precision. Interestingly, the very long 'breathing' periods for very large distances like $\delta r = 1.3$ suggested (temporary) bounded motion for the first 70 years when looking at Fig. 4.11,

while Fig. 4.10 reveals the underlying diverging behavior due to the mismatched bounded motion conditions.

4.5 Conclusion

The normal form methods presented in this chapter yield parameterized sets of the constants of motion $(\mathcal{H}_z(\delta r), E(\delta r))$ for bounded orbits with an average nodal period and average ascending node drift corresponding to the fixed nodal period and ascending node drift of the reference (fixed point) orbit. The range of δr for which bounded motion orbits can be obtained is dependent on the closeness to the Earth. The closer to the Earth, the stronger the influence of the zonal perturbation and the stronger the dynamics of the orbit relatively depend on δr.

In comparison to the approach in [35], our method avoided the time-consuming and inaccurate numerical averaging, by using a normal form based parameterization for the averaging. As a result, the range of the bounded motion provided by our methods is more than twice as large as the range of the results in [35]. Additionally, our method does not require a separate calculation for each δr, but rather provides an expansion in $(\delta r, \delta v_r)$, which covers all orbits up to a certain maximum range that varies with the altitude of the reference trajectory.

While the method in [6] has the advantage of allowing for the calculation of bounded orbits up to arbitrary distances δr, it lacks the ability to provide parameterized sets of bounded motion just like [35].

The normal form methods are also able to provide parameterized sets of the rotation frequency of the orbits within their orbital plane. This rotation is due to the zonal perturbations in the gravitational field of the Earth since there is no rotation of the orbit for the spherically symmetric case. With increasing distance from the Earth's center ρ, the zonal perturbations J_l fall off with ρ^{-l-1}. Accordingly, it is not surprising that the rotation frequency of the MEOs is so much lower than the rotation frequency of the LEOs. Similarly, the δr dependence of the bounded motion is a lot less sensitive for the MEOs compared to the LEOs.

CHAPTER 5

STABILITY ANALYSIS OF MUON G-2 STORAGE RING

This chapter contains parts from my paper *Computation and consequences of high order amplitude- and parameter-dependent tune shifts in storage rings for high precision measurements* published in the *International Journal of Modern Physics A, Vol. 34, No. 36, 1942011 (2019)* [89]. The paper was authored by David Tarazona, Martin Berz, and me. The analysis and results from [89] are presented here and complemented by additional investigations into muon loss mechanisms, which were partly discussed in [80].

The DA map methods (Sec. 2.2) and DA normal form methods (Sec. 2.3) are used to understand the oscillations of particles in the storage ring of the muon g-2 experiment at Fermilab. In contrast to the previous examples, this case of study considers two phase space dimensions. We chose a configuration of the ring which was utilized during one of the first data-collecting stages of the muon g-2 experiment at Fermilab. For this configuration, the phase space behavior is particularly interesting due to resonances (Sec. 2.3.2) and how they affect the stability and loss rates of particles.

5.1 Introduction

Nonlinear effects of electric and magnetic field components of storage rings to confine the particles and bend their trajectory can cause substantial amplitude dependent tune shifts within the beam. Additionally, tune shifts are often sensitive to variations of system parameters, e.g., total particle momentum offsets δp relative to the reference momentum of the storage ring. Such amplitude and parameter dependent tune shifts lead to particles within the beam that oscillate at different frequencies, which potentially influences the beam's susceptibility to resonances and therefore its dynamics and stability. Thus, it is critical for high precision measurements like the muon g-2 experiment to analyze and understand these influences.

In this chapter, we investigate the dynamics within the muon g-2 storage ring, which is the fundamental component of the muon g-2 experiment, using Poincaré return maps and DA normal

form methods. A one-turn Poincaré return map yields the state of particles at a certain azimuthal location within the ring dependent on their state in the previous turn and on system parameters. The application of DA normal form methods to such maps allows for the calculations of tune shifts and quasi-invariants for motion around a (stable) fixed point of the map. Additionally, these maps can be used to track the phase space behavior stroboscopically. Before explaining the methods and the results, the following paragraphs will yield a short introduction to the muon g-2 experiment and its relevance.

The goal of the muon g-2 experiment at Fermilab (E989) [3] is the measurement of the anomalous magnetic dipole moment of the muon

$$a_\mu \equiv \frac{g_\mu - 2}{2}, \tag{5.1}$$

where the g-factor relates the spin and magnetic moment of a particle. Dirac theory predicts the factor to be two for charged leptons like the muon [28, 29], but hyperfine structure experiments in 1940 showed that $g \neq 2$ [62, 63]. The largest radiative correction was introduced by Schwinger in 1948 to explain the difference [71, 72]. Over the years more corrections were explored to gain an understanding of the deviation (g-2) [2], the name-giver of the experiment.

Today, the most successful theory in particle physics is the standard model (SM). It considers high order effects including quantum electrodynamics, electro-weak interactions, and quantum chromodynamics in the calculation of the magnetic dipole moment anomaly of the muon. The most accurate calculation of the magnetic dipole moment anomaly of the muon using the standard model a_μ^{SM} reaches a precision of 0.39 ppm [2]. The muon g-2 experiment E821 conducted at the Brookhaven National Laboratory (BNL) in 2006 yielded a result with a precision of 0.54 ppm [32], which differed from the calculation by 3.6 standard deviations. The E989 at Fermilab is the latest experiment in a series of measurements aimed at pushing the precision of the measured result even higher to reach a precision of 0.14 ppm and push the discrepancy between measurement and calculation to more than five standard deviations [33]. If successful, the result would be a very strong indication that the standard model is unable to describe this anomaly and would call for adjustments to the model or entire new theories. The first results from the latest measurement

at E989 [3, 83, 81, 82] were in agreement with the measurement from BNL had a precision of 0.46 ppm. Combined with the result from BNL, this yields an experimental precision of 0.35 ppb and a discrepancy of 4.2 standard deviations from theory predictions [3]. Expectations are that the five standard deviations are reached with the next set of results from E989.

The experimental technique can be briefly summarized as follows: A highly spin-polarized beam of muons[1] is created as the decay product of high energy pions. The muons are delivered by the Muon Campus as part of the accelerator complex at Fermilab [75, 79, 74] and injected into the muon g-2 storage ring. During the first revolutions within the storage ring, the muon beam is prepared to minimize the emittance and limit fluctuations of the total momentum acceptance to about $\pm 0.5\%$. The prepared beam is then orbiting in the storage ring with only the vertical magnetic field and the four electrostatic quadrupole systems (ESQ) acting on it. The constant magnetic field forces the beam around the ring and causes the spin of the muons to precess. The four ESQ confine and focus the muons vertically [73]. The muons are decaying while they are orbiting and their spins are precessing. Their decay products are measured by the calorimeter system [36] around the beamline in order to determine the spin precession frequency of the muons, which is then used to calculate the muon anomalous magnetic moment [32].

Understanding the behavior of the prepared muon beam in the storage ring is particularly important to identify and address problems. One issue is muon loss, which decreases the number of detected muons. The loss of muons introduces a systematic bias for the average polarization of the remaining particles, which will influence the overall result of the measurement.

In [89], we published an analysis focused on tune shifts, which is the basis of subsequent investigations into the relevance of resonances in muon loss mechanisms.

Accordingly, the following description of the methods draws from [89]. The tune shift analysis from [89] will be presented and complemented by additional investigation regarding period-3 fixed point structures and their relevance in muon loss mechanisms, which was partly presented in [80].

We are going to start with the introduction from [89] into how the Poincaré maps for the storage

[1] Anti-muons are dubbed as muons throughout as customary in the muon g-2 collaboration.

ring are generated and the concept of closed orbits. Afterward, we will be discussing results regarding the relevance of the momentum dependent fixed point of the maps, momentum and amplitude dependent tune shifts, and the presence of period-3 fixed point structures.

5.2 Storage Ring Simulation Using Poincaré Maps

A storage ring is composed of various particle optical elements, each of which can be simulated in COSY INFINITY [53, 20], mostly by a multipole expansion of the involved fields or corresponding potentials. For each particle optical element, there is a hypothetical ideal orbit for which it is calibrated, usually along the center of the element [14]. The ideal orbit is often characterized by a predetermined set of system parameters $\vec{\eta}_0$, for example, a specific total reference momentum of the particles. If the element is simulated as ideal, namely without perturbations, the actual trajectory of a particle initiated on the ideal orbit when entering the element (at $\vec{z}_0$) follows the ideal orbit throughout the element. However, with perturbations like imperfections in the associated fields of the element, a particle initiated at $\vec{z}_0$ might follow a trajectory different from the ideal orbit. Hence, the ideal orbit describes the actual trajectory of a particle initiated at $\vec{z}_0$ in the unperturbed case.

To analyze how an element influences the transverse phase space behavior around the ideal orbit, Poincaré maps (see Sec. 2.2) are used. The Poincaré surfaces correspond to the vertical storage ring cross section perpendicular to the optical axis at azimuthal locations before ($\mathbb{S}_i$) and after the element ($\mathbb{S}_f$). The Poincaré map $\mathcal{P}$ is expanded around the ideal orbit and expresses how the relative phase space state $\vec{z}_f \in \mathbb{S}_f$ after the particle optical element depends on variations in $\vec{\eta}$ and on the relative phase space state $\vec{z}_i \in \mathbb{S}_i$ before the element, with $\vec{z}_f = \mathcal{P}(\vec{z}_i, \vec{\eta})$. The phase space states relative to the ideal orbit $\vec{z}$ consist of the horizontal $(q_1, p_1) = (x, a)$ and vertical $(q_2, p_2) = (y, b)$ phase space components within the Poincaré surface $\mathbb{S}$. For unperturbed elements, the Poincaré map $\mathcal{P}$ is origin preserving, with $\mathcal{P}(\vec{0}, \vec{0}) = \vec{0}$, since the trajectory follows the ideal orbit.

The transverse phase space behavior after a full revolution in the storage ring is given by the Poincaré return map $\mathcal{M}$, which is generated by composing the individual Poincaré maps $\mathcal{P}_i$ of the individual storage ring elements according to the storage ring setup ($\mathcal{M} = \mathcal{P}_k \circ \mathcal{P}_{k-1} \circ \ldots \circ \mathcal{P}_2 \circ \mathcal{P}_1$)

such that the ideal orbits connect.

For the simulation of the muon g-2 storage ring, a detailed nonlinear model [78] of the storage ring particle optical elements has been set up using COSY INFINITY. The simulation considers the magnetic field that guides the beam around the storage ring and the four-fold symmetric electrostatic quadrupole system [73] (ESQ), which focuses the beam vertically. Additionally, perturbations due to the ESQ fringe fields and nonlinearities of the main field, or imperfections in the vertical magnetic field can be taken into account based on experimental data. Superimposing both perturbations when fringe fields are accounted for has been unsuccessful so far due to technical limitations.

The model represents the magnetic field inhomogeneities by fitting 2D magnetic multipoles up to fifth order to measurement data of the magnetic field within the muon g-2 storage ring (see [82, 78] for details). The ESQ [73] is considered by the corresponding electrostatic potential as a 2D multipole expansion up to tenth order to accurately model the nonlinearities of the system up to the significant contribution of the 20th-pole. The fringe fields of the ESQ – the fall-off of the electric field at the edges of the ESQ components – are simulated based on numerical calculations performed with the code COULOMB [85].

The generated Poincaré return maps are expanded in the horizontal (x, a) and vertical (y, b) phase space coordinates relative to the ideal orbit. Additionally, the maps are expanded in relative offset $\delta p = \Delta p / p_0$ with respect to the initial reference momentum p_0 to represent particles within the momentum acceptance range of about ±0.5% of the E989 storage ring. The relative change δp corresponds to the change of the system parameter $\vec{\eta}$.

To distinguish the influences of various elements of the storage ring and their perturbations, we simulated different configurations of the components as shown in [89]. Specifically, the influence of perturbations due to ESQ fringe fields and influences from imperfections in the vertical magnetic field are treated separately. We also considered the system for two ESQ voltages, namely 18.3 kV and 20.4 kV. In this chapter of the thesis, however, we will only consider an ESQ voltage of 18.3 kV, since it offers the most interesting nonlinear dynamics and is a set-point used during the first data collection of the muon g-2 experiment. We are also only considering the map with the magnetic

field imperfections since investigations in [89] indicated that it is the dominating perturbation and therefore yields the most realistic results. The main insights from [89] regarding the other cases will still be mentioned at the appropriate places in the text below.

5.3 The Closed Orbit

Closed orbits return to themselves after each storage ring revolution, which makes them fixed points of the Poincaré return maps. There are also low period closed orbits that return to themselves after a few turns n. These orbits correspond to low period fixed point structures in the n-turn Poincaré return map. While there are also unstable fixed points, which are discussed later, we will first focus on the properties of the stables ones.

The closed orbit is a reference for the associated particles since they oscillate around it with the closed orbit representing an equilibrium state. Accordingly, the closed orbit is sometimes also referred to as the reference orbit. In the stroboscopic view of the Poincaré return maps, the fixed point mimics an equilibrium point of the oscillatory phase space behavior around it. Using the DA normal form algorithm (see Sec. 2.3) on an origin preserving Poincaré return map, the transverse oscillation frequencies around the fixed point can be calculated. In the rest of this section, we will focus on how these closed orbits and their associated fixed points in the Poincaré return maps are determined.

5.3.1 The Closed Orbit Under Perturbation

If all components are simulated to be unperturbed, then the Poincaré return map is a composition of origin preserving Poincaré maps and hence also origin preserving. However, if the simulation considers perturbations, the actual trajectory of the expansion point may be distorted from the ideal orbit and hence not a closed orbit. Accordingly, the expansion point of the associated Poincaré return map may not be a fixed point and the map may not be origin preserving.

However, if the perturbation is sufficiently small, a fixed point $\vec{z}_{\mathrm{FP}}$ will continue to exist. Parameterizing the strength of the perturbation with $\vec{\eta}$, the origin preserving fixed point map

of the unperturbed system is given by $\mathcal{M}(\vec{z}, \vec{\eta} = 0)$. To analyze the preservation of the parameter dependent fixed point, an extended map $\mathcal{N}(\vec{z}, \vec{\eta}) = (\mathcal{M}(\vec{z}, \vec{\eta}) - \vec{z}, \vec{\eta})$ is defined [14]. If $\det(\mathrm{Jac}(\mathcal{N}(\vec{z}, \vec{\eta})))|_{(\vec{z},\vec{\eta})=(\vec{0},\vec{0})} \neq 0$ then, according to the inverse function theorem, an inverse of the map $\mathcal{N}$ exists for a neighborhood $\mathbb{D}$ around the evaluation point $(\vec{0}, \vec{0})$ of the Jacobian. The parameter dependent fixed point $\vec{z}_{\mathrm{FP}}(\vec{\eta})$ of $\mathcal{M}$ and hence the closed orbit of the system exists as long as $(0, \vec{\eta})$ is within the neighborhood for which invertibility has been asserted. If this is the case and the inverse $\mathcal{N}^{-1}$ around $(\vec{0}, \vec{0})$ is given, then the parameter dependent fixed point can be calculated via

$$(\vec{z}_{\mathrm{FP}}(\vec{\eta}), \vec{\eta}) = \mathcal{N}^{-1}\left(\vec{0}, \vec{\eta}\right). \tag{5.2}$$

Expanding the map around the parameter dependent fixed point yields the origin preserving Poincaré return map under perturbations in the system parameters.

The perturbation due to imperfections in the magnetic field distorts particles from the ideal orbit of the E989 storage ring. Accordingly, the Poincaré return map from the composition of the individual particle optical elements is not origin preserving. Using the method above, the fixed point of the map – the phase space coordinates of the closed orbit at the azimuthal location of the map – is calculated and the map is expanded around it. The result is an origin preserving fixed point map.

Calculating the fixed point for Poincaré return maps at multiple azimuthal locations of the ring indicates the form of the closed orbit (see Fig. 5.1).

The collimator locations are highlighted because they are of particular relevance for muon losses. They constitute the narrowest part around the storage region restricting the muons to amplitudes of $r = \sqrt{x^2 + y^2} < 45\ \mathrm{mm} = r_0$ relative to the center of the ring, i.e. the ideal orbit. Muons hitting a collimator during data taking for the measurement are known as *lost muons*.

While the radial motion of the closed orbit along the storage ring is close to sinusoidal, the vertical phase space motion is disturbed into more complex behavior. In the xy projection, distorted elliptical motion around the ideal orbit along the center of the ring is indicated. All these deviations from the ideal orbit are triggered by the weak coupling of radial and vertical motion due to ppm-level imperfections of the skew quadrupole magnetic field. The form of the closed orbit is determined by

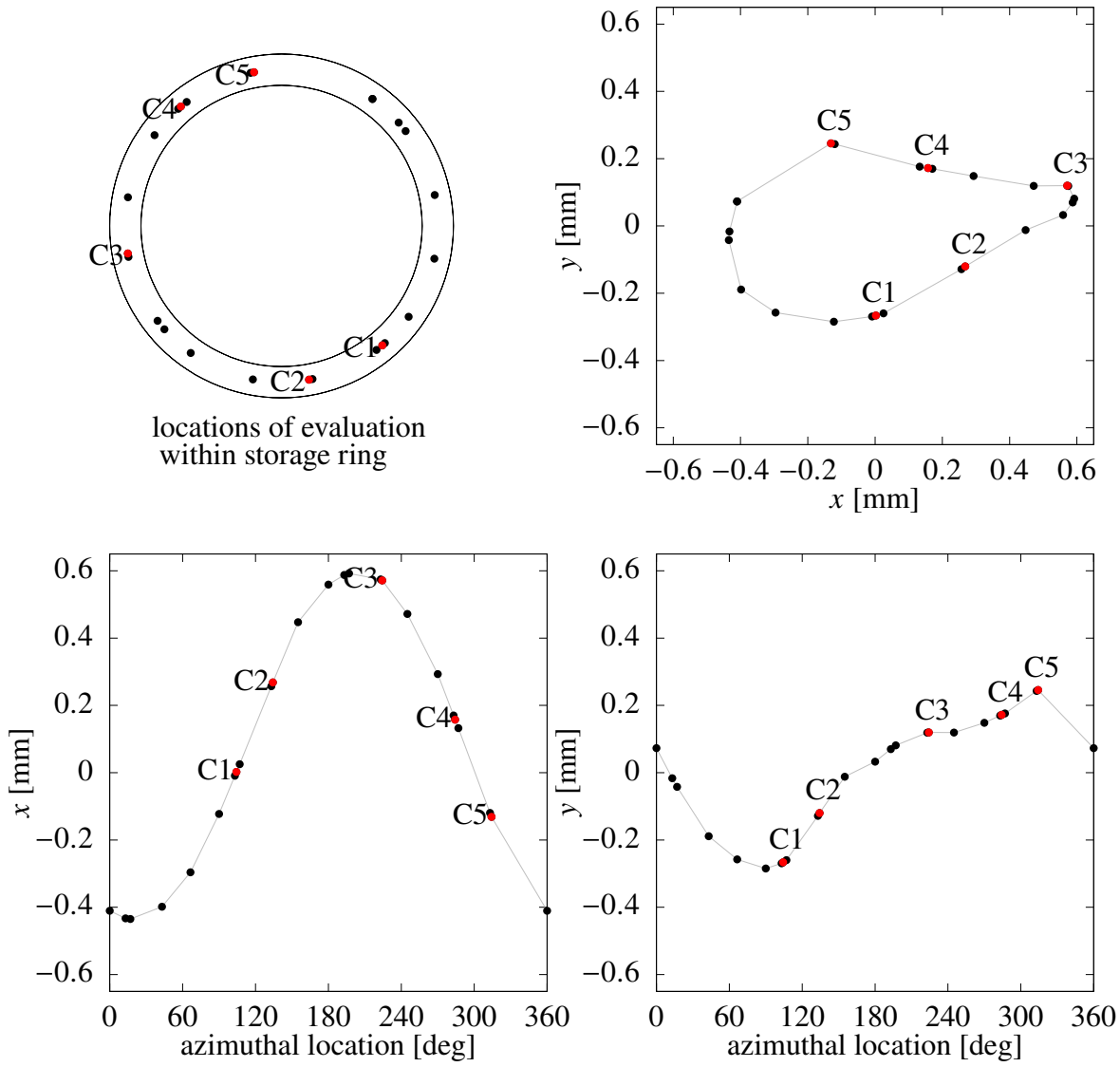

Figure 5.1: The fixed points of Poincaré return maps from various azimuthal locations around the ring indicate the behavior of the closed orbit (for $\delta p = 0$). The projections of the four dimensional fixed points into subspaces illustrate the influence of the magnetic field perturbations on the closed orbit around the ring. The results from the five collimator locations (C1-C5) are highlighted with red color.

the distribution of such magnetic field imperfections as well as the fields and voltage of the ESQ.

The closed orbit we found here and showed in Fig. 5.1 is considering a particle with no momentum offset ($\delta p = 0$). Following the argumentation above the closed orbit continues to exist with perturbations in δp as will be investigated in the next section.

5.3.2 The Momentum Dependence of the Closed Orbit

The closed orbit additionally depends on system parameters like the momentum offset of the particles. Just like for the magnetic field perturbation, Eq. (5.2) is used to calculate the parameter dependent fixed point of the origin preserving Poincaré return map, where the parameter is the momentum offset δp. The phase space coordinates of the momentum dependent fixed point at the collimator locations in the ring are shown in Fig. 5.2.

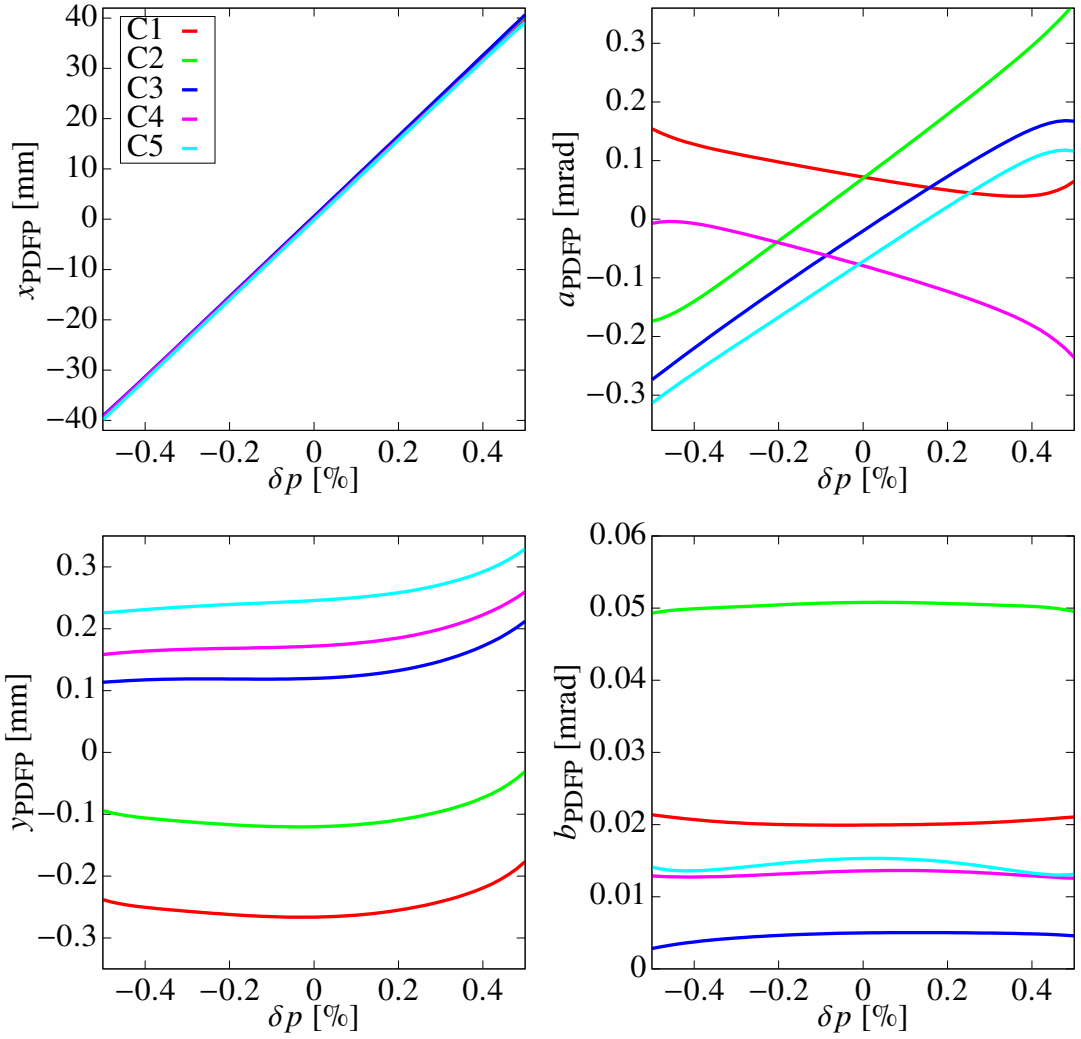

Figure 5.2: Changes of the closed orbits due to relative changes δp in the total initial momentum. The plots illustrate absolute coordinates with respect to the ideal orbit at the center of the ring for the five collimator locations (C1-C5).

The primary effect from the momentum offset comes from the interaction of the charged particles

with the unperturbed part of the vertical magnetic field. The Lorentz force, which determines the orbit radius, is directly proportional to the momentum of the particle. This behavior is clearly visible in the horizontal components of Fig. 5.2. The radial position of the parameter dependent fixed point x_{PDFP} changes linearly at about 79 mm/% with the momentum offset at all collimator locations. The associated dependence of the horizontal momentum a_{PDFP} incorporates the changing radial orientation of the momentum dependent closed orbit with respect to the Poincaré surface and the different orientations at the various collimator locations.

The vertical components y_{PDFP} and b_{PDFP} of the closed orbit are mostly dependent on the azimuthal location of the map and change only slightly with a momentum offset.

5.3.3 The Relevance of Closed Orbits

The momentum dependent closed orbits correspond to fixed points in the Poincaré return maps. Particles that are not on a closed orbit oscillate around the momentum dependent closed orbit corresponding to their specific momentum offset. In the stroboscopic view of the Poincaré return maps, this corresponds to stroboscopic oscillatory behavior around the fixed point in both phase spaces as Fig. 5.3 indicates. The amplitudes of these transverse oscillations are determined by the phase space position of the particle and the momentum dependent fixed point.

Particles with the same oscillation amplitudes but different momentum offsets will follow roughly the same motion, but at different locations in phase space. On the other hand, particles at the same phase space location may follow entirely different orbital motion depending on their corresponding momentum dependent fixed point. In summary, the phase space motion of a particle is characterized by its momentum dependent fixed point, its amplitudes of oscillation, and its oscillation frequencies, which are addressed in detail in Sec. 5.4.

The collimators restrict the maximum amplitudes of oscillation around the associated momentum dependent fixed points. The viable phase space region for particles decreases with increasing momentum offset (see Fig. 5.4). The closeness of the reference closed orbit to the collimators increases the risk of muon loss. While particles with low momentum offset are only at risk of getting

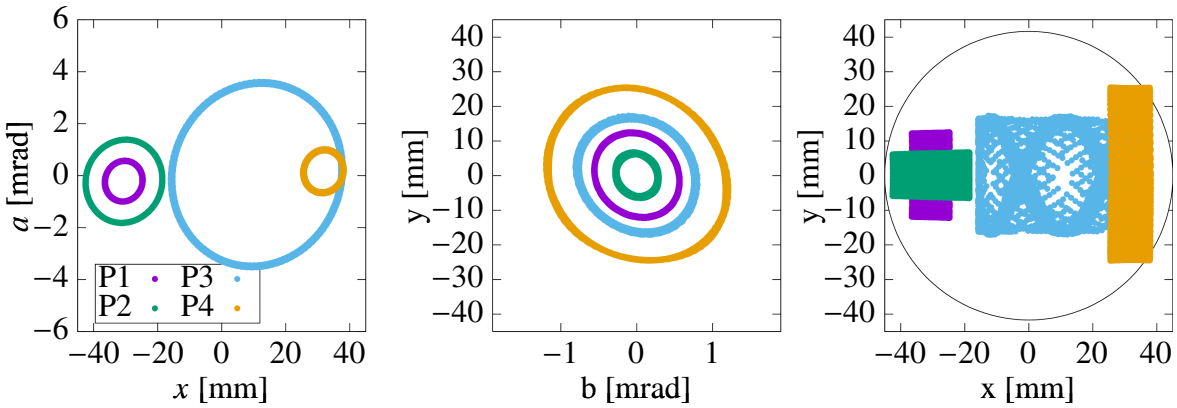

Figure 5.3: Phase space behavior of four particles in different phase space regions with various amplitudes and momentum offsets. Particle 4 (yellow) hits the collimator and is lost. The momentum dependent radial position x of the particles is particularly prominent. The individual particles are characterized by the parameter set $(x_{\text{amp}}, y_{\text{amp}}, \delta p)$ with (6 mm, 12 mm, −0.39%) for particle 1 (P1), (12 mm, 6 mm, −0.39%) for particle 2 (P2), (27 mm, 16 mm, +0.13%) for particle 3 (P3), and (6 mm, 25 mm, +0.39%) for particle 4 (P4).

lost when they have relatively large oscillation amplitudes, particles with a large momentum offset may already be lost with seemingly small amplitudes of oscillation.

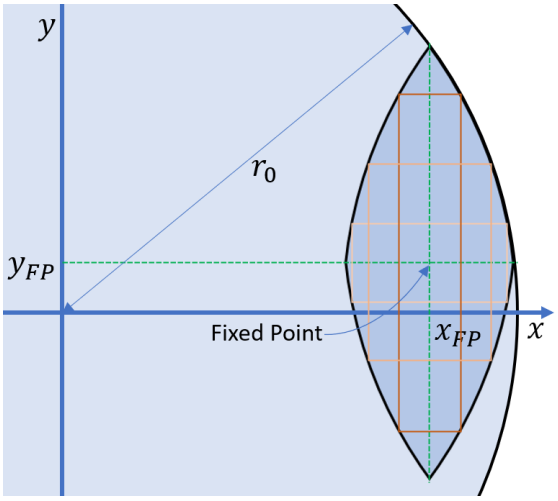

Figure 5.4: Schematic illustration of viable xy region around a momentum dependent fixed point.

Since the semi-major and semi-minor axis of the distorted elliptical phase space behavior are not necessarily aligned with the position and momentum axis and vary for each particle, there is no straightforward definition of the amplitude of oscillation. The DA normal form algorithm takes care of this by transforming the distorted ellipses in phase space to circles such that the amplitudes of

oscillation are just the radii of the circles – the normal form radii. We will investigate the relationship between the original phase space coordinates and the normal form radii more closely later on and also use its advantages, but for now, we want to focus on practically relatable quantities in the original phase space, rather than abstract quantities like the normal form radius.

Given a particle with distorted elliptical phase space behavior and its corresponding momentum dependent fixed point $\vec{z}_{\mathrm{PDFP}}(\delta p)$, we define the oscillation amplitudes x_{amp} and y_{amp} independently from each other. In the radial phase space $x_{\mathrm{amp}} = |x_0 - x_{\mathrm{PDFP}}(\delta p)|$ for $a_0 = a_{\mathrm{PDFP}}(\delta p)$ and $y_{\mathrm{amp}} = |y_0 - y_{\mathrm{PDFP}}(\delta p)|$ for $b_0 = b_{\mathrm{PDFP}}(\delta p)$ in the vertical phase space.

5.4 Tune analysis

The following tune analysis investigates the oscillation frequency around the reference closed orbits depending on the momentum offset and the amplitude of oscillation. The tunes shall shed light on average loss times and the involvement of resonances.

5.4.1 Tunes of the Momentum Dependent Closed Orbit

Given the parameter dependent fixed point map representing the phase space behavior around the momentum dependent closed orbit of the muon g-2 storage ring model, the diagonalization in the DA normal form algorithm is used to determine the tunes of the momentum dependent closed orbit.

The calculated tunes of the closed orbit (for $\delta p = 0$) differ only very slightly depending on the azimuthal location of the Poincaré return map yielding $\nu_x = 0.944462633(8 \pm 3)$ and $\nu_y = 0.330814444(7 \pm 6)$, which is expected since they all describe the linear motion around same closed orbit. The proximity of the vertical tune ν_y to the low $1/3$-resonance will be investigated more closely later. The radial tune ν_x is even closer to a higher order resonance namely the $17/18$-resonance. Without loss of generality, we will use the Poincaré return map at collimator C3 for our further map investigations.

The Fig. 5.5 illustrates the momentum dependence of the tunes over the momentum offset range of $\delta p \in [-0.5\%, 0.5\%]$ and indicates the linear dependence (chromaticities) ξ_i as a reference.

For $|\delta p| < 0.25\%$ the momentum dependence of both tunes is predominantly linear with $\xi_x =$

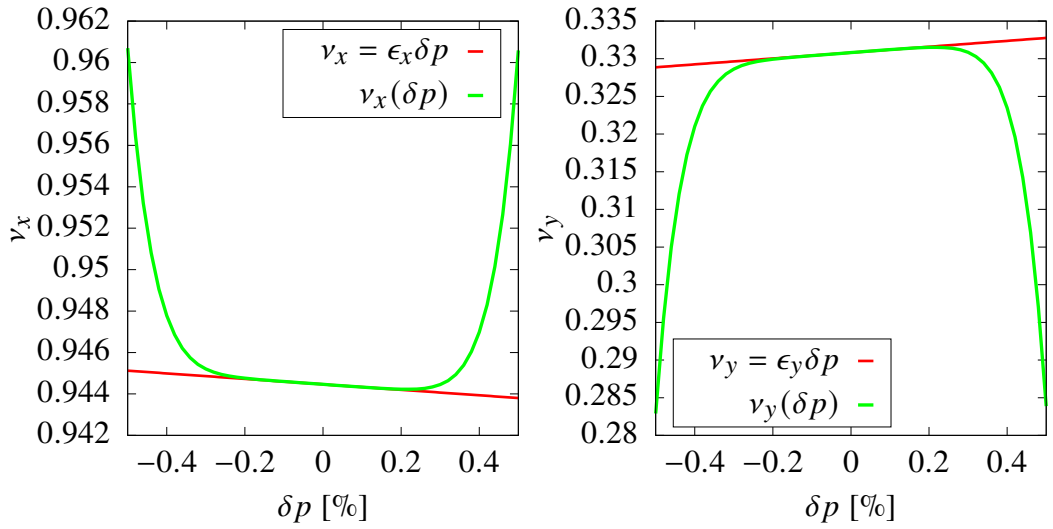

Figure 5.5: Vertical and horizontal tune dependence in the model of the muon g-2 storage ring of E989 on relative offsets δp from the reference momentum p_0.

-0.131999346 and $\xi_y = 0.389753993$. For $|\delta p| > 0.33\%$ however, the tunes are dominated by an order eight dependence on relative momentum offsets δp. This eighth order dependence results from the strong ninth order terms in the original map, which are linear in the phase space components and of order eight in the momentum dependence, representing the earlier mentioned significant influence of the 20th-pole of the ESQ potential.

Interestingly, the linear coefficient and the eighth order coefficient of the vertical momentum dependent tune shifts are both larger by a factor of three and opposite in sign compared to their radial counterparts. Additionally, the momentum dependent vertical tune shifts away from the $1/3$-resonance.

The investigation in [89] indicated a strong influence of the ESQ voltage on the linear motion around the respective expansion points and therefore the tunes. The momentum dependence of the tunes – the momentum dependent tune shifts – however is only slightly changed by the ESQ voltage (see [89] for more details).

5.4.2 The Amplitude Dependent Tune Shifts

The DA normal form algorithm provides the transformation $\mathcal{A}_{\mathrm{NF}}$ from the original phase space coordinates (x, a) and (y, b) to rotationally invariant normal form coordinates $(q_{\mathrm{NF},1}, p_{\mathrm{NF},1})$ and $(q_{\mathrm{NF},2}, p_{\mathrm{NF},2})$. The amplitude and parameter dependent tune shifts $\nu_i(r_{\mathrm{NF},1}, r_{\mathrm{NF},2}, \delta p)$ can be extracted from the normal form map, where the amplitudes are given by the normal form radii $r_{\mathrm{NF},i} = \sqrt{q^2_{\mathrm{NF,i}} + p^2_{\mathrm{NF,i}}}$.

This full description of the tunes and their dependence on phase space amplitudes and momentum offsets is extremely powerful. However, the abstract normal form radii are not as practically useful as the previously defined oscillation amplitudes x_{amp} and y_{amp} in original phase space coordinates. To address this, Fig. 5.6 illustrates the dependence of the tunes on the radial phase space amplitude x_{amp} and the dependence on the vertical phase space amplitude y_{amp}, separately. This is done by calculating the corresponding normal form coordinates and normal form radii and using those for the tune evaluation.

The amplitude dependence is never linear but always appears as even orders. Investigations in [89] indicated that amplitude dependent tune shifts, just like momentum dependent tune shifts, are only weakly influenced by the ESQ voltages and the field perturbations. Similar to the purely momentum dependent tune shifts, the sign of the momentum offset seems to only play a minor role compared to the magnitude of the offset.

The radial amplitude dependence of the tunes is relatively well behaved. Again, there is the dominating eighth order dependence related to the strong ninth order nonlinear terms resulting from the 20th-pole of the ESQ potential, which shifts the tunes of the radial phase space up and tunes of the vertical phase space down with increasing radial amplitude and magnitude of the momentum offset.

The vertical amplitude dependence however is more complex as it varies strongly with the magnitude of the momentum offset. Regarding the vertical tune, this is particularly critical due to the crossing of the $1/3$-resonance tune for some vertical amplitude and momentum offset combinations. Such low resonances can have a major influence on the dynamics of particles which is why we will

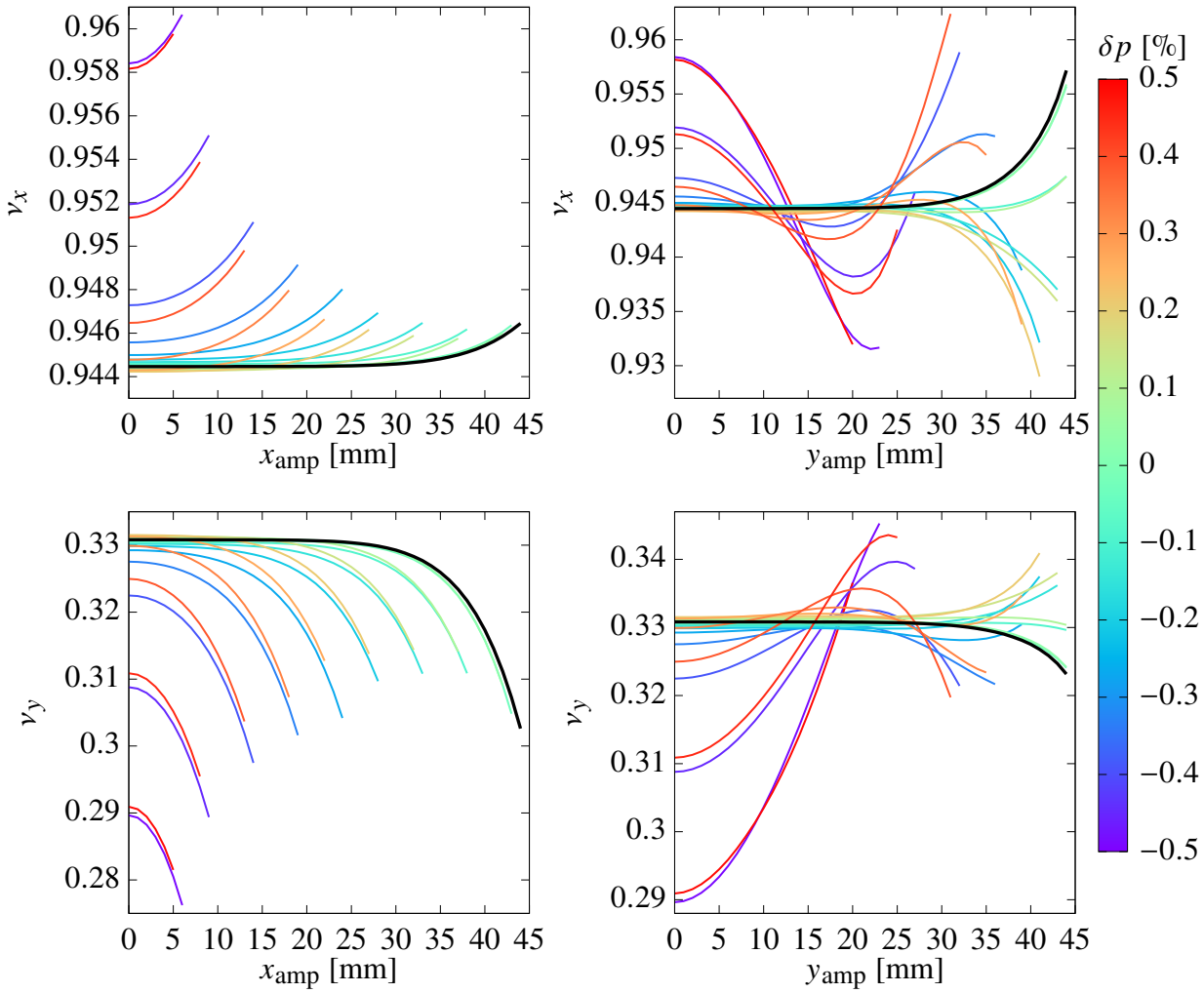

Figure 5.6: Amplitude dependent tune shifts in the model of the muon g-2 storage ring of E989. The black line indicates the amplitude dependent tune shifts for $\delta p = 0$, while the other lines have a momentum offset specified by their color. For the left plots regarding the radial amplitude dependence, the vertical amplitude relative to the momentum dependent fixed point is set to zero and vice versa for the plots regarding the vertical amplitude dependence on the right. The lines end when the total xy amplitude of the particle relative to the ideal orbit reaches the collimator at $r_0 = 45$ mm.

closely investigate these cases later.

Even though the purely momentum dependent tune shifts ($x_{\mathrm{amp}} = 0, y_{\mathrm{amp}} = 0$) and the tune shifts purely dependent on the vertical amplitude ($x_{\mathrm{amp}} = 0, \delta p = 0$) shift in the same direction – up for radial tunes and down for vertical tunes – there are opposing cross-terms, which depend both on the vertical amplitude and the momentum offset that trigger this nontrivial tune shift behavior.

In Fig. 5.7 to Fig. 5.9 the combined effects of simultaneous radial and vertical amplitudes on the tune shifts are illustrated for selected momentum offsets. The behavior for the intermediate momentum offsets may be interpolated from the given plots. Again, the sign of the momentum offset has only a minor influence on the form of the tune shifts compared to its magnitude.

Note that Fig. 5.7 to Fig. 5.9 only illustrates tunes for phase space states within the viable phase space around the corresponding momentum dependent fixed point. Accordingly, not all lines extend over the full 45 mm range of y_{amp} and some lines for large x_{amp} are not shown, since their total xy amplitude of the particle relative to the ideal orbit reaches the collimator at $r_0 = 45$ mm.

The combined effects in Fig. 5.7 to Fig. 5.9 emphasize the strong nonlinear character of the tune dependencies, which was already indicated in Fig. 5.6. The wave-like structure illustrates how different order terms dominate at different vertical amplitudes y_{amp} depending on both, the radial amplitude x_{amp} and the momentum offset δp. Additionally, for almost every momentum offset there are combinations of oscillation amplitudes for which the vertical $1/3$-resonance tune is crossed. Investigations in [89] did not show this strong nonlinear behavior of the combined effects on the tune shifts in such clarity.

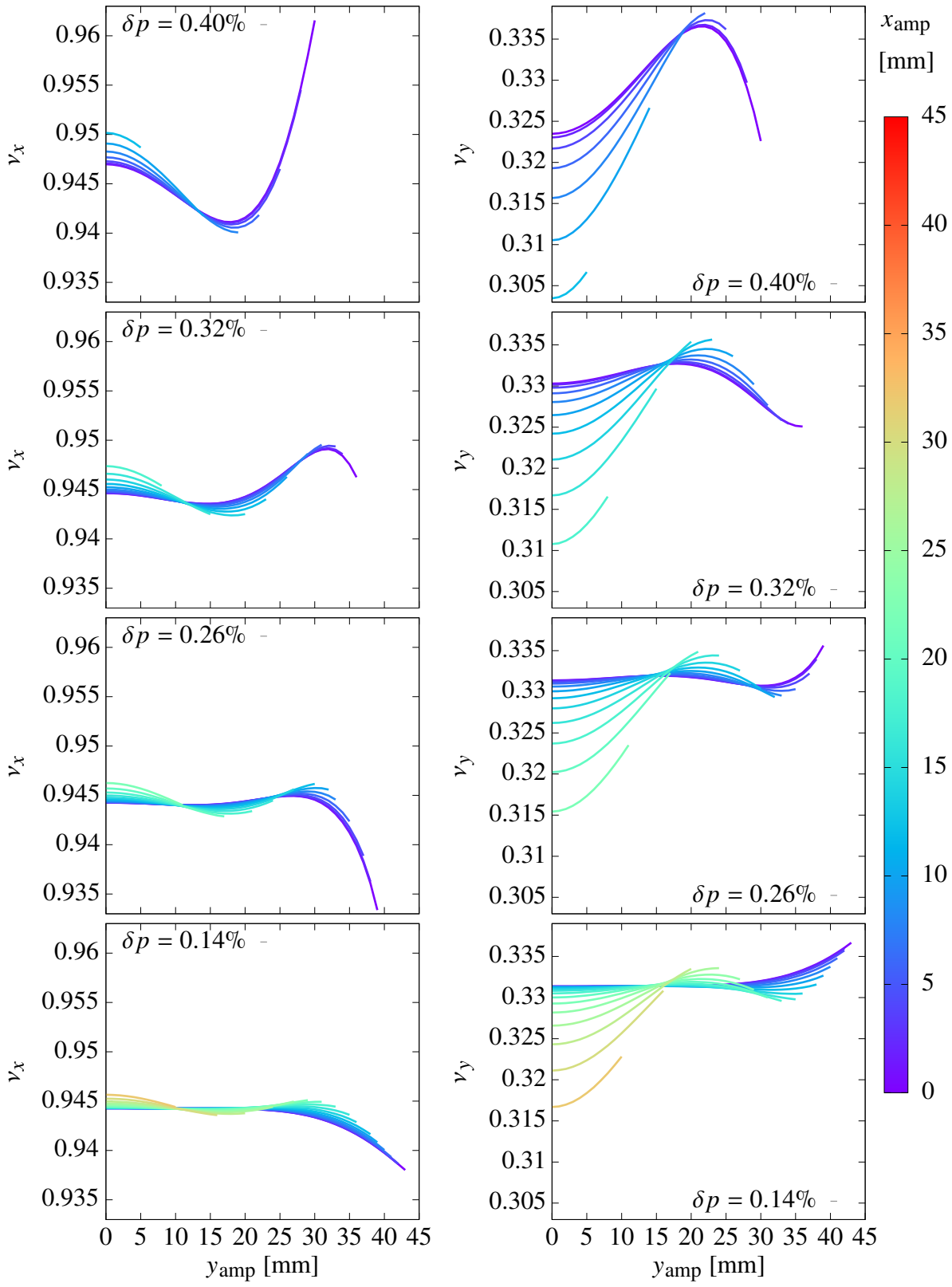

Figure 5.7: Behavior of combined amplitude dependent tune shifts at multiple momentum offsets.

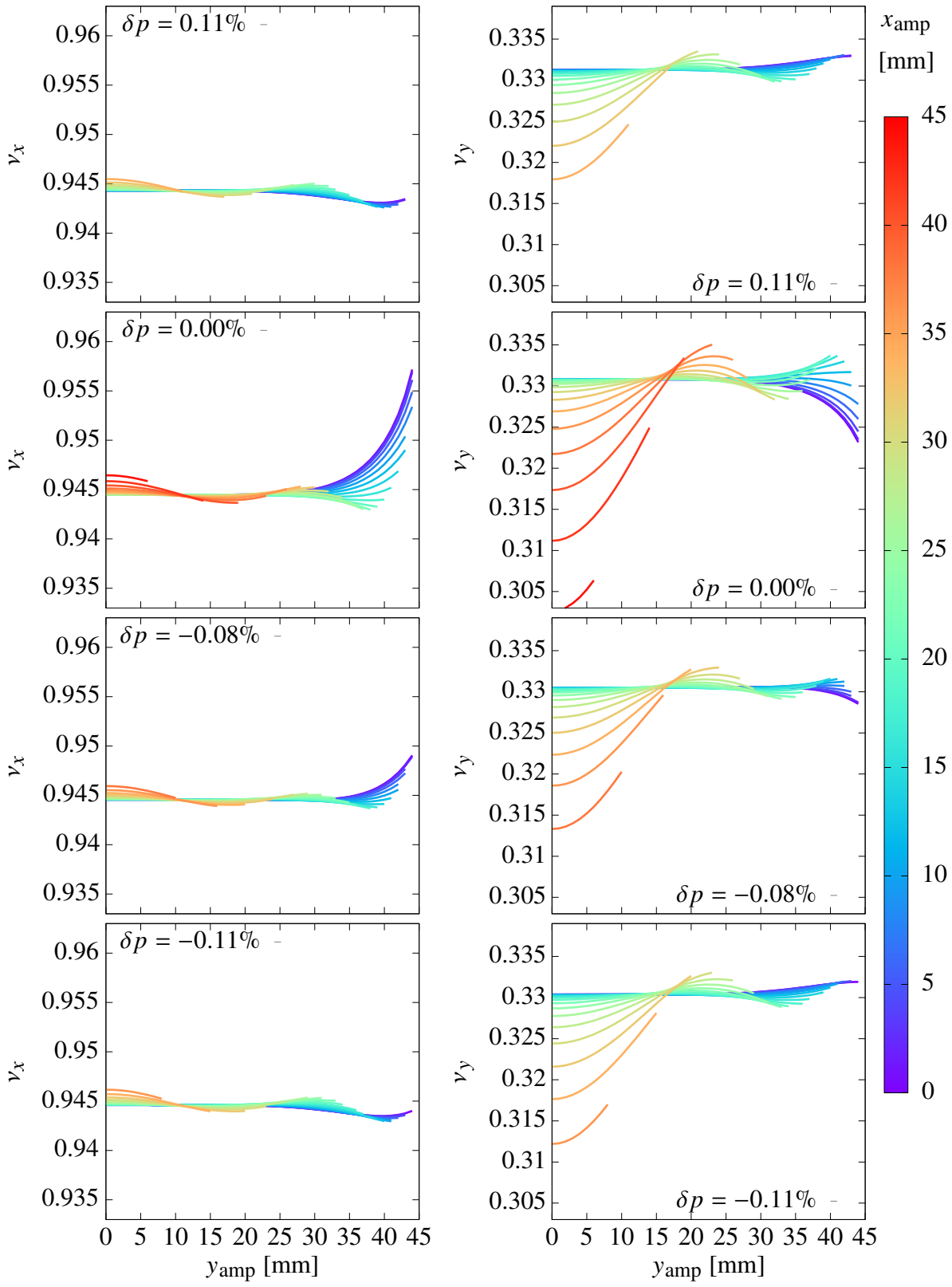

Figure 5.8: Behavior of combined amplitude dependent tune shifts at multiple momentum offsets.

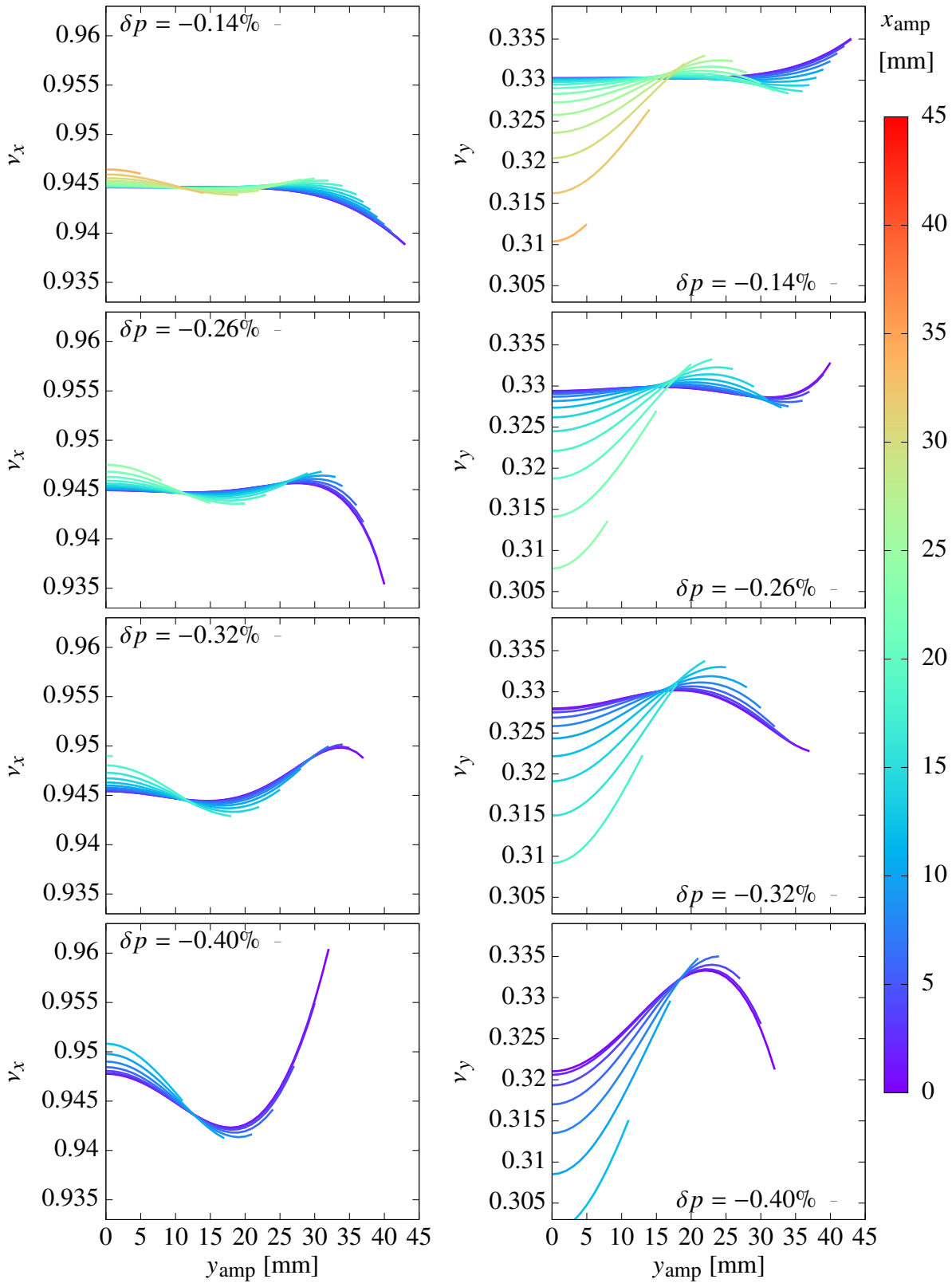

Figure 5.9: Behavior of combined amplitude dependent tune shifts at multiple momentum offsets.

5.4.3 The Tune Footprint

The tune footprint visualizes the projection of a beam distribution into tune space. The COSY INFINITY based model [78] of the muon g-2 storage ring is used to generate the realistic beam distribution from orbit tracking of the muon beam until it is circulating in the storage ring, prepared for data analysis. In particular, the model considers the imperfect injection process, which attempts to align the injected beam with the ideal orbit of the storage ring as well as possible. The model also considers the mispowered ESQ components to imitate the preparation mechanism during the first turns of the beam in the storage ring at E989. Further details of the tracking model and on how a distribution is obtained are elaborated in [77, 78].

The variables $(x, a, y, b, \delta p)$ relative to the ideal orbit are illustrated in Fig. 5.10 as projections into the (x, a), (y, b), and (x, y) subspaces.

The beam distribution tends towards higher total momenta in the range of $\delta p \in [-0.2\%, 0.4\%]$ while overall staying well within the momentum acceptance range of $\pm 0.5\%$. The spread of the vertical momentum component b is about a factor two to three smaller than its horizontal counterpart a. The position space (xy) is filled up to the limitations due to the collimators.

The distributions of the horizontal and vertical tunes are illustrated by the tune footprint in Fig. 5.11, where the vertical tunes of the particle distribution are plotted against their horizontal tunes as previously done in [49]. The tune footprint of the tenth order calculation is overlaid by the result of an eighth order calculation to emphasize the influence of the strong ninth order nonlinearities of the map caused by the 20th-pole of the ESQ potential. The tune footprint of the tenth order calculation is five to 6 times larger in each dimension than its eighth order counterpart.

Additionally, particles in different momentum offset ranges are highlighted to illustrate the behavior of this specific group. The tune footprint can be segmented into three groups characterized by their momentum offset which generates a tune footprint in the shape of a 'T'.

The tune footprint for the other ESQ voltages in [89] has a similar distribution for the order eight and order ten calculations, respectively. While the reference tunes are mainly determined by the ESQ voltage, the relative tune shifts behave very similarly. If the ESQ voltage were to place the

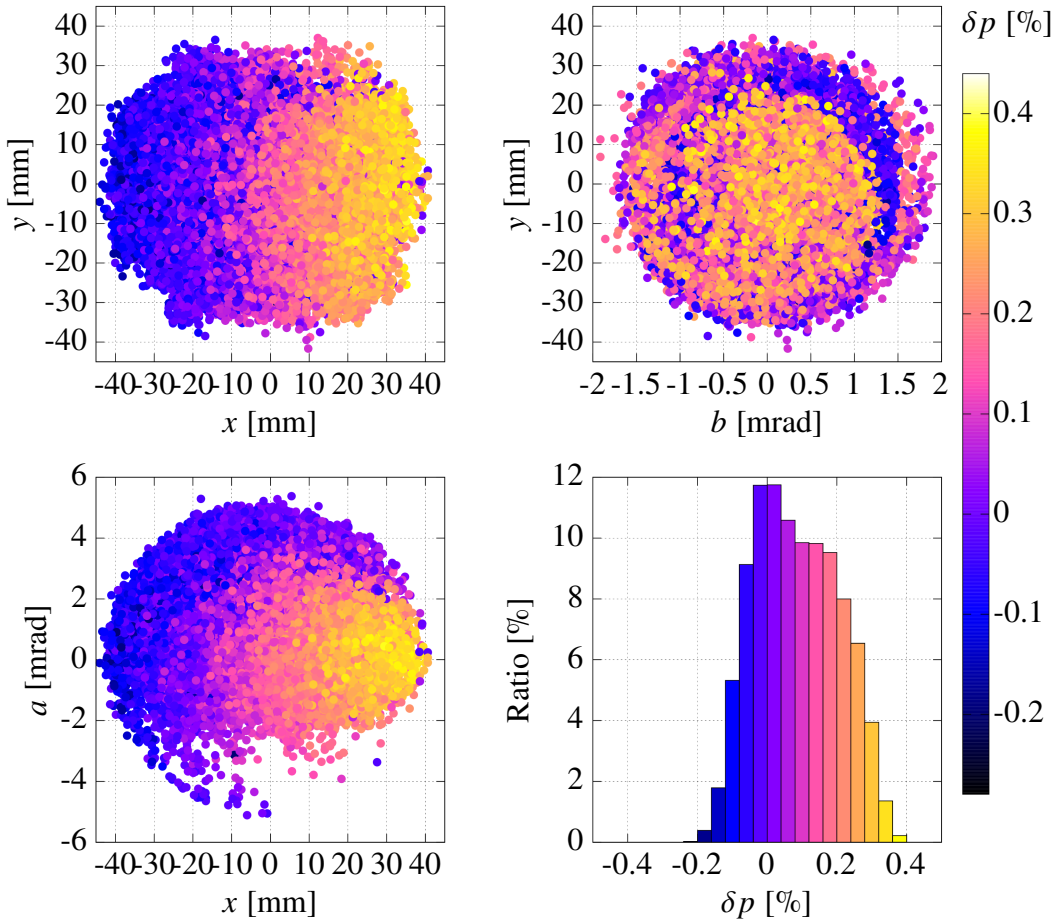

Figure 5.10: Projections of the distribution of the variables $(x, a, y, b, \delta p)$ in the realistic beam simulation at the azimuthal ring location of the central kicker.

reference tunes very close to a resonance line, we expect the tune distribution and tune shifts to behave differently.

Fig. 5.11 shows that the vertical $1/3$-resonance tune can not only be reached hypothetically for the apparent case of a nominal set-point away from resonances. A substantial part of particles is close to or on this low order resonance. The overlaid eighth order calculation shows that this is triggered by the strong ninth order nonlinearities of the map caused by the 20th-pole of the ESQ potential. The segmentation with regard to the momentum offset of the particles into subgroups additionally shows that the vertical $1/3$-resonance tune is crossed in each of those groups. The resonance point $(17/18, 1/3)$ is also covered and surrounded by many particles and might have a

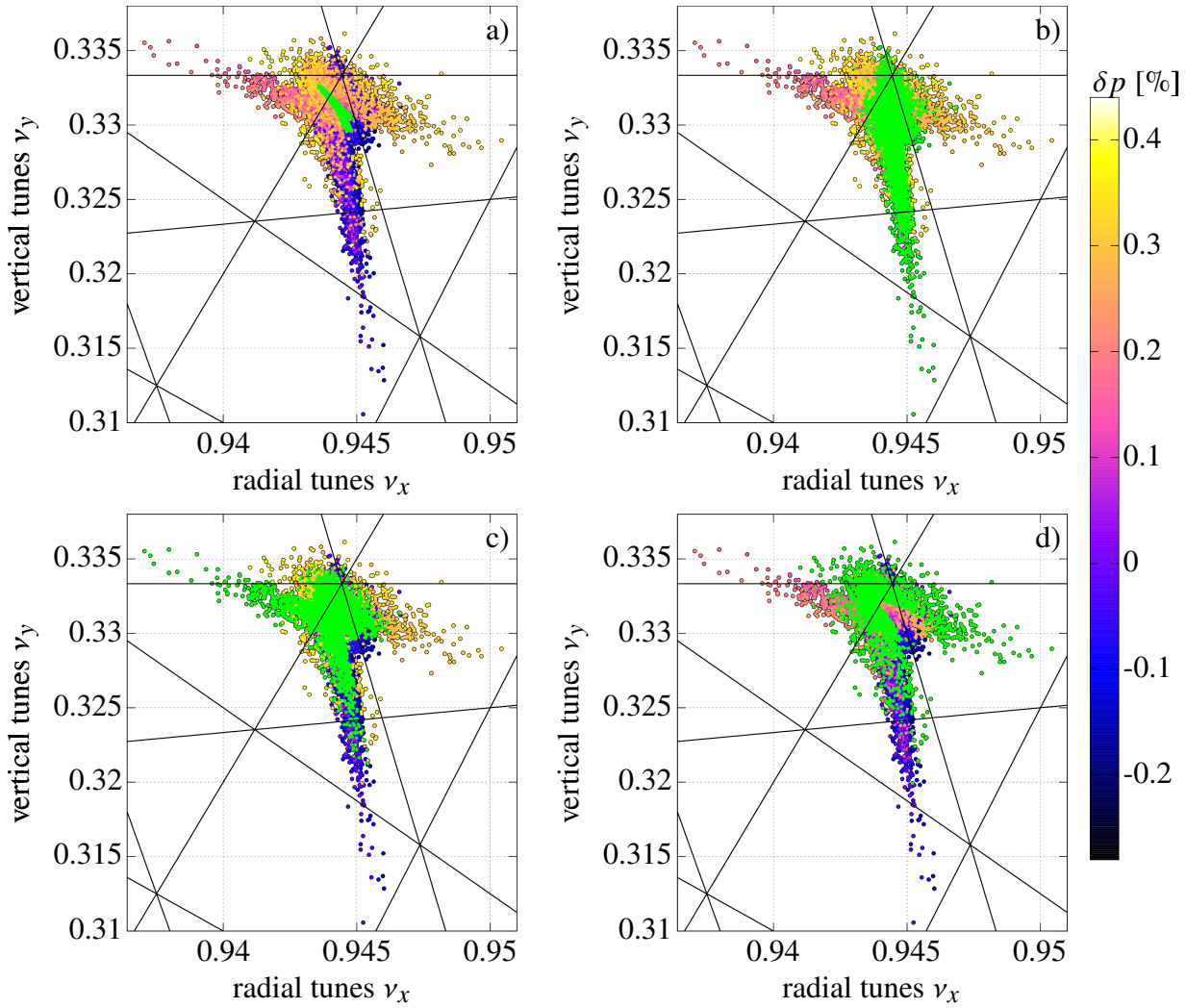

Figure 5.11: The tune footprint of a realistic beam distribution at the azimuthal ring location of the central kicker. The tune footprint from the 10th order calculation is colored according to the momentum offset of the individual particles. The black lines correspond to resonance conditions. In a) the 8th order calculation (green) is overlaid to illustrate the drastic influence of the strong ninth order nonlinearities of the map caused by the 20th-pole of the ESQ potential. In b) the particles with a momentum offset $-0.3\% < \delta p < 0.1\%$ are overlaid in green. In c) the particles with a momentum offset $0.1\% < \delta p < 0.28\%$ are overlaid in green. In d) the particles with a momentum offset $0.28\% < \delta p < 0.5\%$ are overlaid in green.

particularly strong impact.

5.5 Stability and Loss Mechanisms

Muons are lost when they hit structural parts of the storage ring and lose the energy necessary to remain within the storage region during data taking. Collimators, which are inserted at various azimuthal locations in the ring (see Fig. 5.1), constitute the narrowest part around the storage region. They restrict the muons to amplitudes of $r = \sqrt{x^2 + y^2} < 45$ mm $= r_0$ relative to the ideal orbit.

Our previous analysis is very helpful for gaining a general understanding of certain properties of the system, e.g. the momentum dependence of the reference orbit, and the momentum and amplitude dependent shifts in the oscillation frequency of orbits around their repetitive reference orbit. This analysis showed that the vertical $1/3$-resonance tune is relevant for various combinations of amplitudes and momentum offsets. However, only tracking analysis can yield the actual phase space behavior of lost particles and particles involved with the vertical $1/3$-resonance tune. Additionally, we saw that the radial tune is very close to the high order $17/18$-resonance, which will also look at more closely.

For the tracking analysis, we use both one-turn maps as well as sectional maps. The one-turn Poincaré return map yields the state of a muon at the azimuthal location of the central kicker (K2) dependent on its state in the previous turn. Sectional maps transfer the state of a muon to the azimuthal location of the collimators. Accordingly, the muons are not tracked continuously, but stroboscopically at specific azimuthal locations e.g. at the respective collimator locations.

There are two common approaches for tracking analysis. For a general understanding of the phase space dynamics of the storage ring, one could track a particle distribution, which is evenly distributed in all phase space dimensions and over momentum offset range. However, the implication from such an analysis for the actual muon beam might be limited, since the actual muon beam is not evenly distributed. Accordingly, we track the realistic particle distribution of 37738 particles from Sec. 5.4.3.

The particle distribution is given after turn 200 at K2, which is about 30 μs after injection when data taking begins. During this initial 30 μs after injection, the quadrupole system is still ramping up and scraping techniques are deployed for the final preparation of the beam [81]. The beam is tracked

for additional 4500 turns (670 μs), while determining and documenting various orbit parameters that shall be analyzed in detail below.

5.5.1 The Normal Form Defect of Tracked Particles

As explained in Sec. 2.4, the normal form defect yields the inaccuracies in the normal form, i.e., how much the pseudo-invariants (the normal form radii) vary per turn. Using tracking simulations, one can evaluate a related quantity that we will call the long term normal form defect. It yields the difference between the maximum and the minimum normal form radius of a single particle orbit over many turns during long term tracking. It is therefore able to detect instabilities on a large time scale.

In Fig. 5.12 the particles are grouped by the maximum per turn normal form defect they encountered during the 4500 turns of tracking. The rate of particles getting lost is strongly correlated

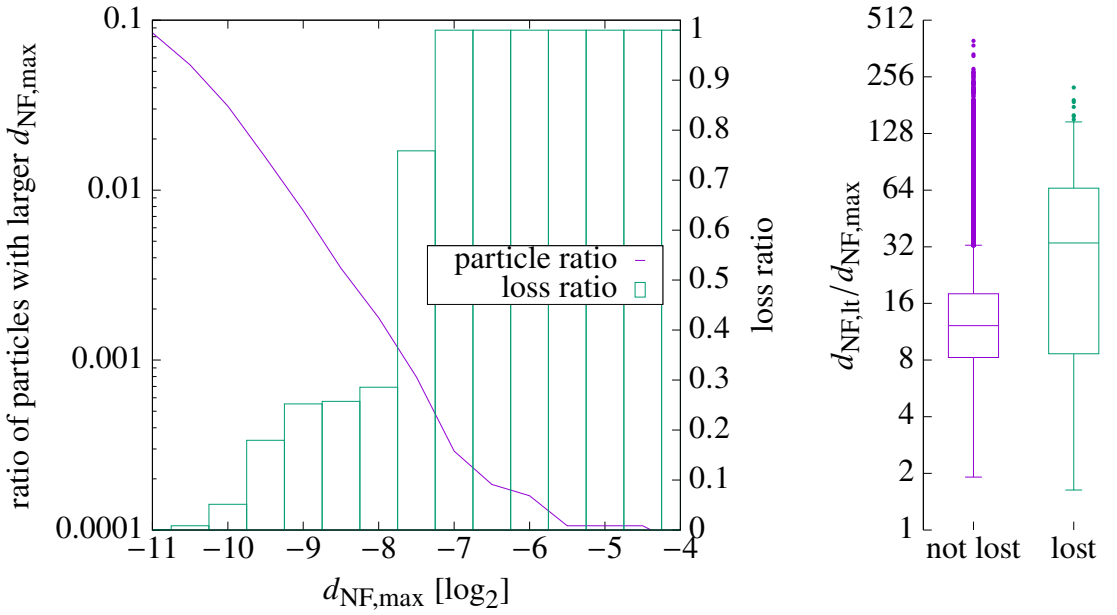

Figure 5.12: Relation between losses and the normal form defect.

with the size of the maximum normal form defect they encountered. There are no losses for particles with a maximum normal form defect smaller than 2^{-11}, which are more than 91% of particles. For larger normal form defects, the loss rate increases significantly with 100% of particles getting lost that encounter a maximum normal form defect larger than 2^{-7}. This confirms that the size of the

per turn normal form defect is a good indication for losses, in the sense that the larger the per turn normal form defect the more likely the particle gets lost.

The right side of Fig. 5.12 illustrates the ratio of the long term normal form defect to the maximum per turn normal form defect of a particle. Considering that the long term normal form defect over 4500 turns is only a factor of eight to 16 larger than the maximum per turn normal form defect for particles that are not lost illustrates the overestimating implications for Nekhoroshev-type stability estimates (see Sec. 2.4) based on the per turn normal form defect for this particular map. On the other hand, the ratio is much more shifted to higher factors for lost particles, indicating loss overestimation for lost particles with Nekhoroshev-type stability estimates.

In Fig. 5.13 the relevance of the resonances – especially low order resonances like the vertical $1/3$-resonance tune – on the long term normal form defect becomes obvious. Since the tunes are dependent on the normal form radii, a larger long term normal form defect automatically corresponds to a larger tune range of a particle.

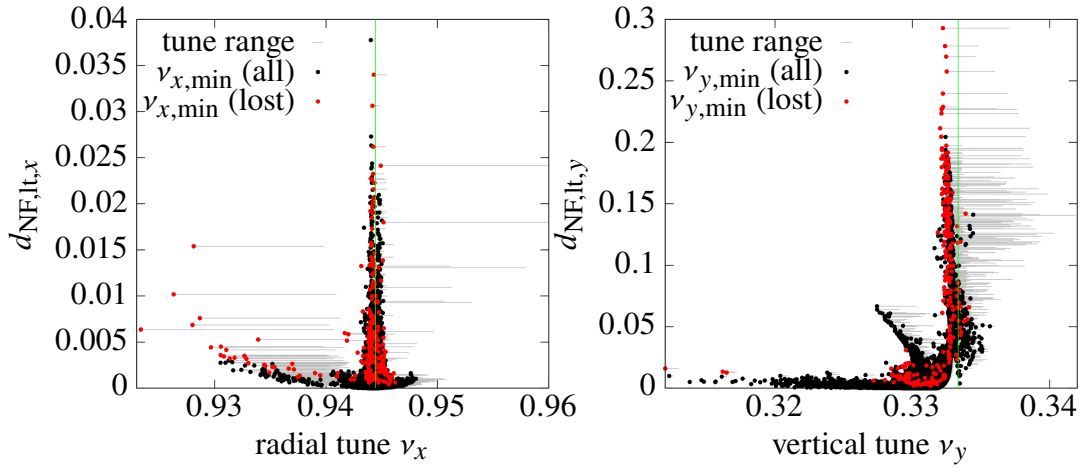

Figure 5.13: The plots show the long term normal form defect dependent on the calculated tune range of each particle. The dots are the minimum calculated tune of each particle while tracking. Red dots indicate that the respective particle is lost over the 4500 tracking turns. The gray lines show the calculated tune range of each particle. The left plot illustrates the radial long term normal form defect with respect to the radial tune and the $17/18$ resonance (green line). The right plot shows the vertical long normal form defect with respect to the vertical tune and the $1/3$ resonance (green line).

In the plot of the vertical tune against the vertical long term normal form defect, there is a ‘spike ’ facing roughly 45° away from the resonance line. In Fig. 5.14, the tune range of these ‘spike ’

particles is analyzed to determine a resonance as a potential trigger of the increasing normal form defect. The analysis indicates that the 10th order $6\nu_x + 4\nu_y = 7$ resonance might be the cause of

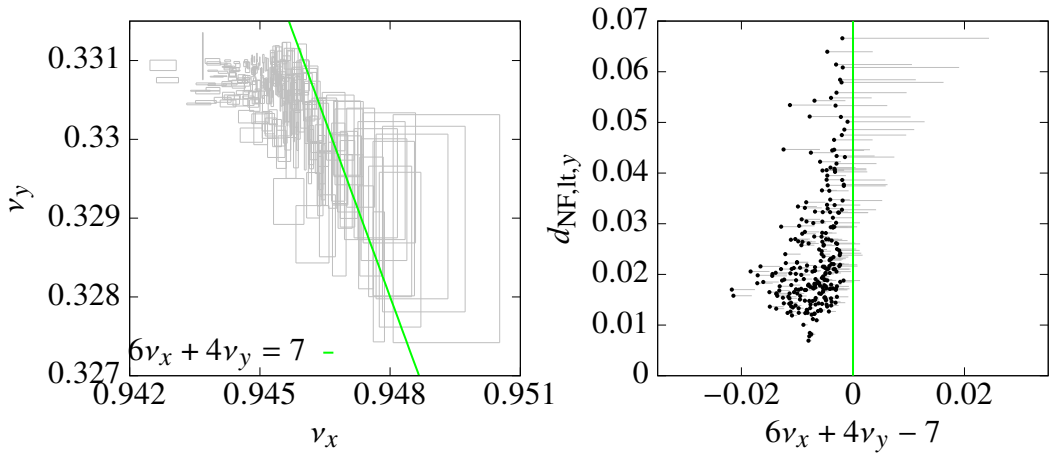

Figure 5.14: The tune range of the particles forming the spike in Fig. 5.13 are shown on the left. The right plot shows the normal form defect of the particles depends on their closeness to the $6\nu_x + 4\nu_y = 7$ resonance (green line).

this spike, but it remains unknown why the normal form defect increases along this resonance with increasing distance from the $1/3$-resonance.

The normal form radii are the oscillation amplitudes in the high order normalized, linearly decoupled phase space. They are closely related to the oscillation amplitudes in the respective phase spaces relative to the momentum dependent closed orbit. The strong variation in the normal form radii (the large long term normal form defect) of some orbits indicates that the corresponding oscillation amplitude of those orbits around their respective reference orbits is also not constant. To investigate this more closely, the following section investigates the orbits of all lost particles.

5.5.2 Lost Muon Studies

In this section, we track and investigate all muons of the distribution from Sec. 5.4.3 that are lost at collimator C3 and/or C4 over the 4500 turns. In Fig. 5.15 to Fig. 5.29, 15 out of the 259 lost particles are picked to illustrate the different phase space behaviors observed for lost particles.

One striking property that many of the lost muons share is the appearance of threefold-symmetry patterns in the vertical phase space projections. The calculated tunes of these lost particles are all

crossing or proceed very close to the vertical $1/3$-resonance. These threefold-symmetry patterns often include significant modulations in the vertical oscillation amplitude, which is additionally shown by the changing overall normal form radius $r_{\mathrm{NF}} = \sqrt{r_{\mathrm{NF},1}^2 + r_{\mathrm{NF},2}^2}$ and the variations in the calculated tunes. While there are many patterns, there are two that stick out, namely, the island pattern (see for example Fig. 5.17) and the shuriken pattern (see for example Fig. 5.23). In Sec. 5.5.3, we will understand how all these patterns are related to period-3 fixed point structures.

The patterns come in stable, semi-stable, and unstable forms. This tendency to unstable behavior is often associated with a large radial amplitude and/or a closeness to the $(\nu_x, \nu_y) = (17/18, 1/3)$ resonance point. The 'fuzzyness' of the vertical phase space pattern in (y, b) compared to the pattern in the corresponding normal form phase space $(q_{\mathrm{NF},2}, p_{\mathrm{NF},2})$ is related to the radial phase space motion. Due to the weak coupling between the radial and vertical phase space from the imperfections in the magnetic field, large amplitudes in (x, a) notably affect the motion in (y, b), which does not happen in the decoupled normal form phase space. This 'fuzzyness' might also trigger the jumping between different patterns for orbits, which are close to the border between two patterns (see Fig. 5.31). More thoughts on this also in Sec. 5.5.3.

Another property that many lost particles share is a significant momentum offset, which radially shifts their respective reference orbit closer to the boundaries of the collimator. The dependence of the radial position of the reference orbit on the momentum offset decreases the maximum survivable size of those rectangular shapes in xy space significantly as previously discussed in Sec. 5.3.3 and illustrated in Fig. 5.4.

Last but not least, there are also particles like the one shown in Fig. 5.15, which get lost simply because of their constant but large oscillation amplitudes in the radial and/or vertical direction. However, it is not always obvious to distinguish them from particles that are under the influence of a period-3 fixed point structure like the particle in Fig. 5.16.

Fig. 5.15 to Fig. 5.29 also indicate that the xy pattern of lost particles often only barely touches the collimator boundary. For these cases, it may take many revolutions for both oscillations, in the radial and vertical direction, to reach their maximum simultaneously [76].

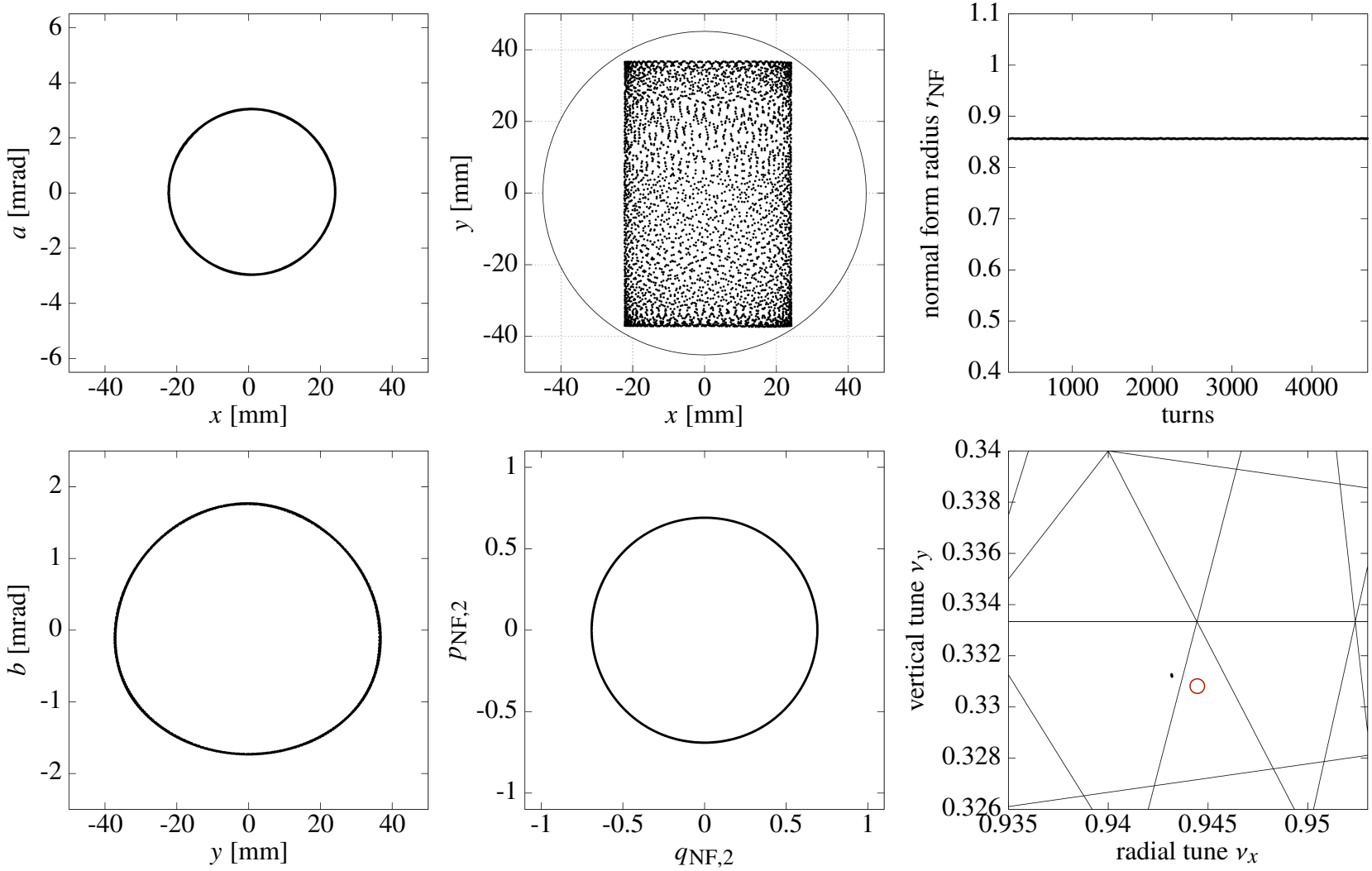

Figure 5.15: The radial and vertical phase space behavior indicates that this particle ($\delta p = 0.015\%$) oscillates at constant amplitudes around its momentum dependent reference orbit. The overall normal form radius is constant and confirms this. Accordingly, the tune footprint of the particle is a single dot. This is a trivial large amplitude loss.

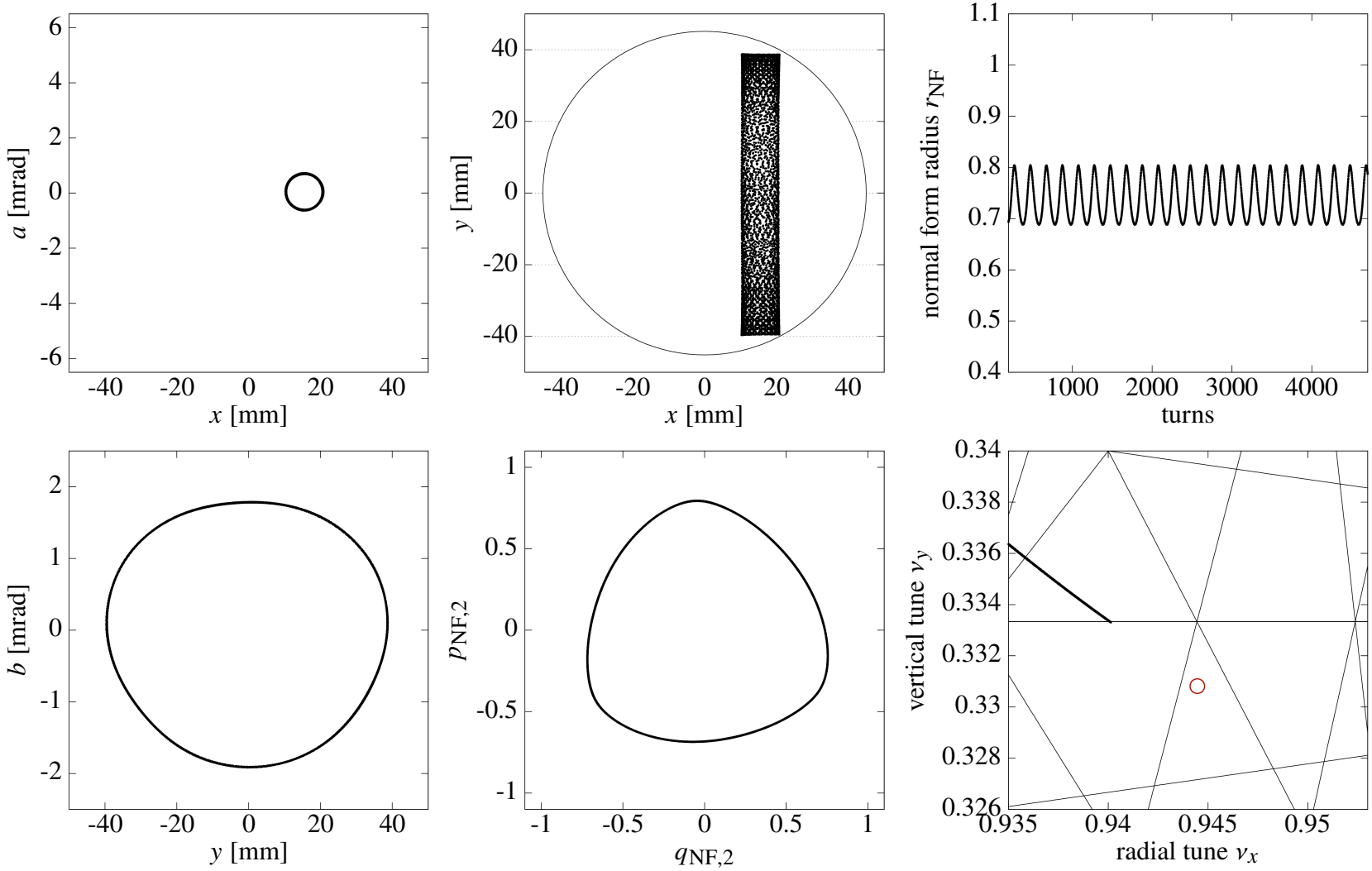

Figure 5.16: The vertical phase space behavior of this particle ($\delta p = 0.196\%$) has a slight triangular deformation. The overall normal form radius indicates a modulated amplitude and the spread out tune footprint starts right after the vertical $1/3$-resonance line. Despite slight influence of the resonance, the rather elliptical phase space behavior makes this a trivial large amplitude loss.

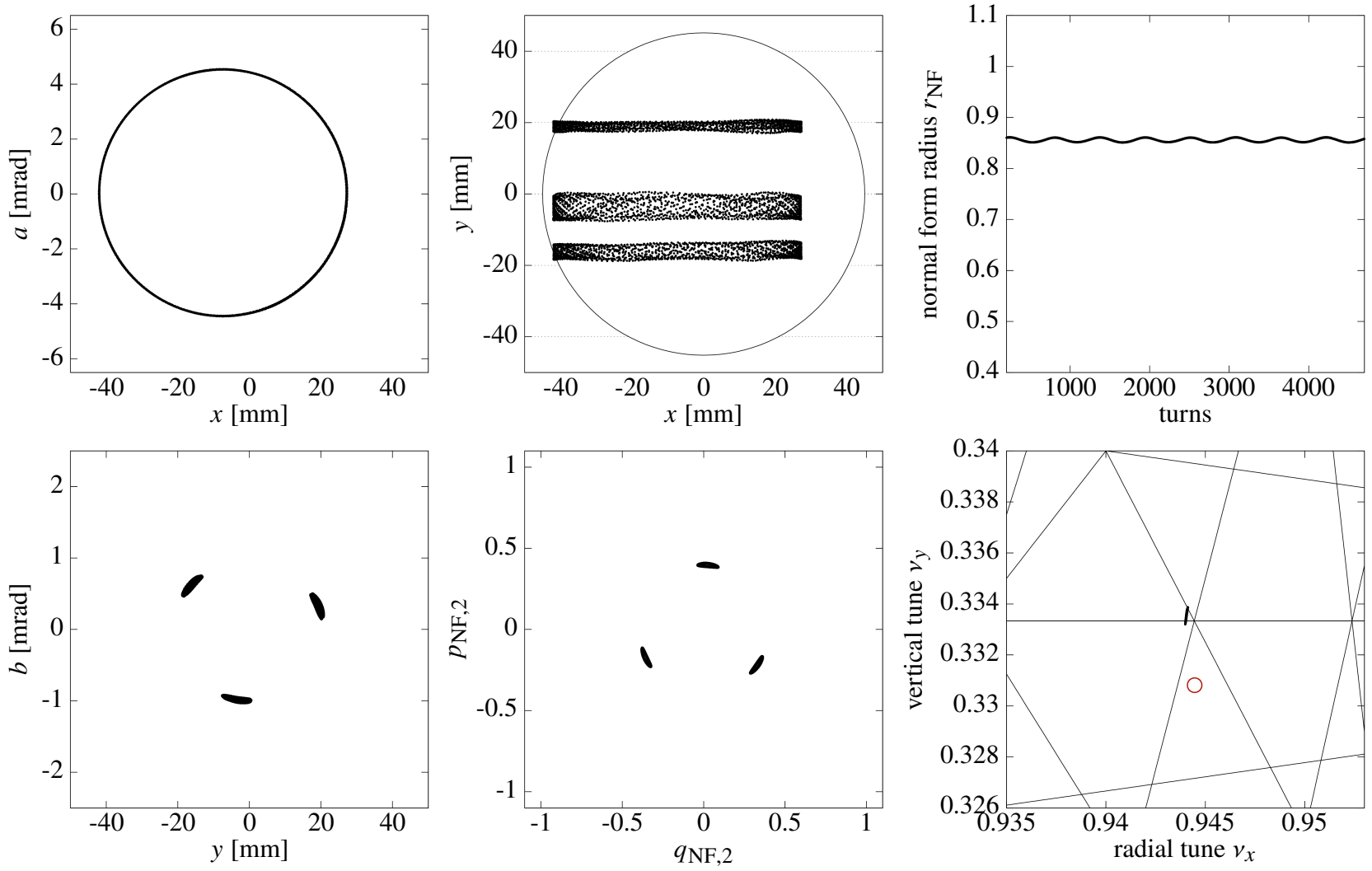

Figure 5.17: This particle ($\delta p = -0.088\%$) is caught around a period-3 fixed point structure in the vertical phase space, which is related to the vertical $1/3$-resonance. We refer to these structures as islands and the loss mechanisms is called island related loss.

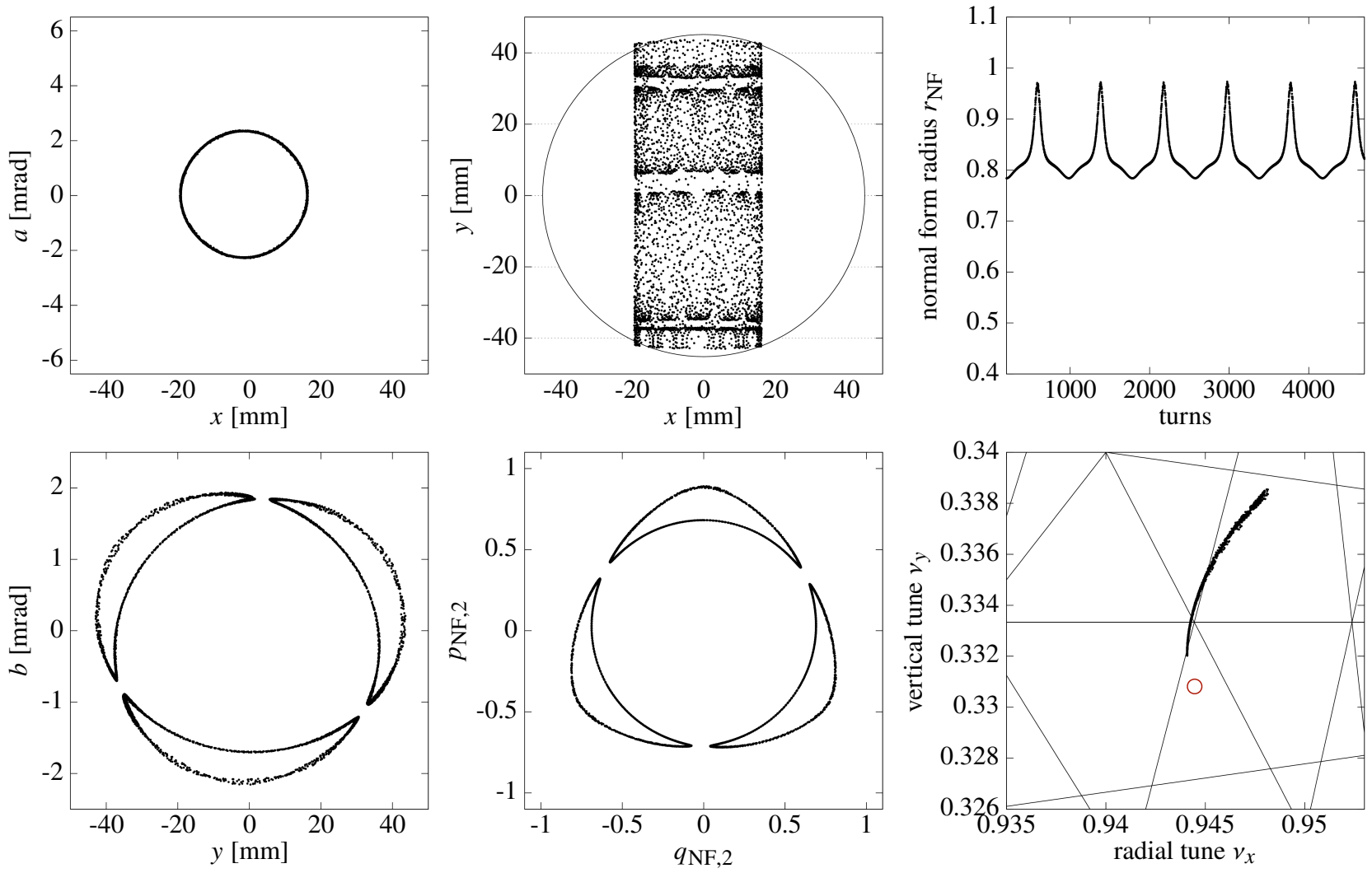

Figure 5.18: This particle ($\delta p = -0.015\%$) forms large islands around a period-3 fixed point structure in the vertical phase space, which is associated with a major modulation of the oscillation amplitude.

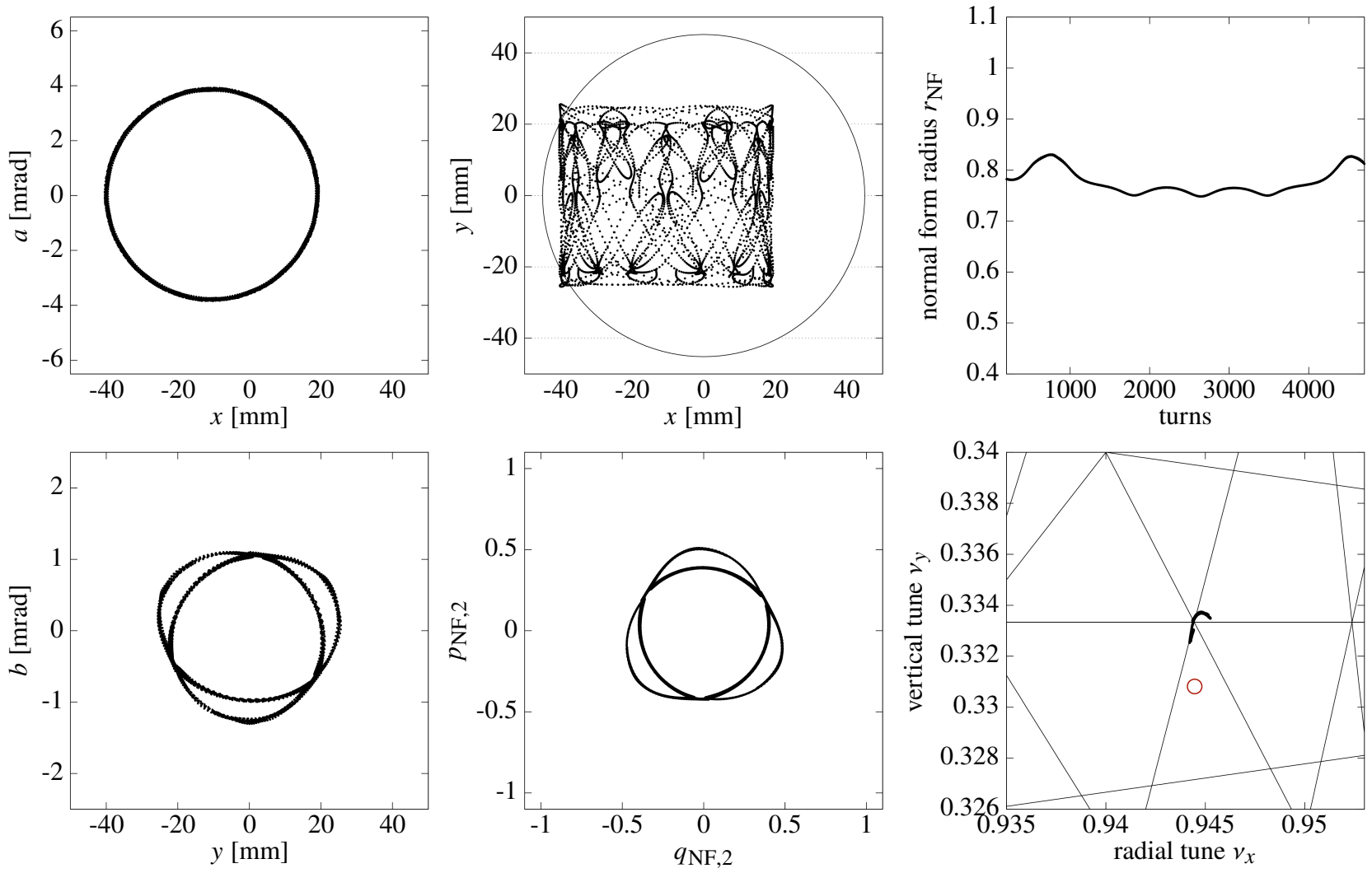

Figure 5.19: This particle ($\delta p = -0.127\%$) jumps between the islands. The large radial amplitude and/or the closeness to the $(17/18, 1/3)$ resonance point might have triggered the jump. This is an example of moderate unstable behavior around a period-3 fixed point structure.

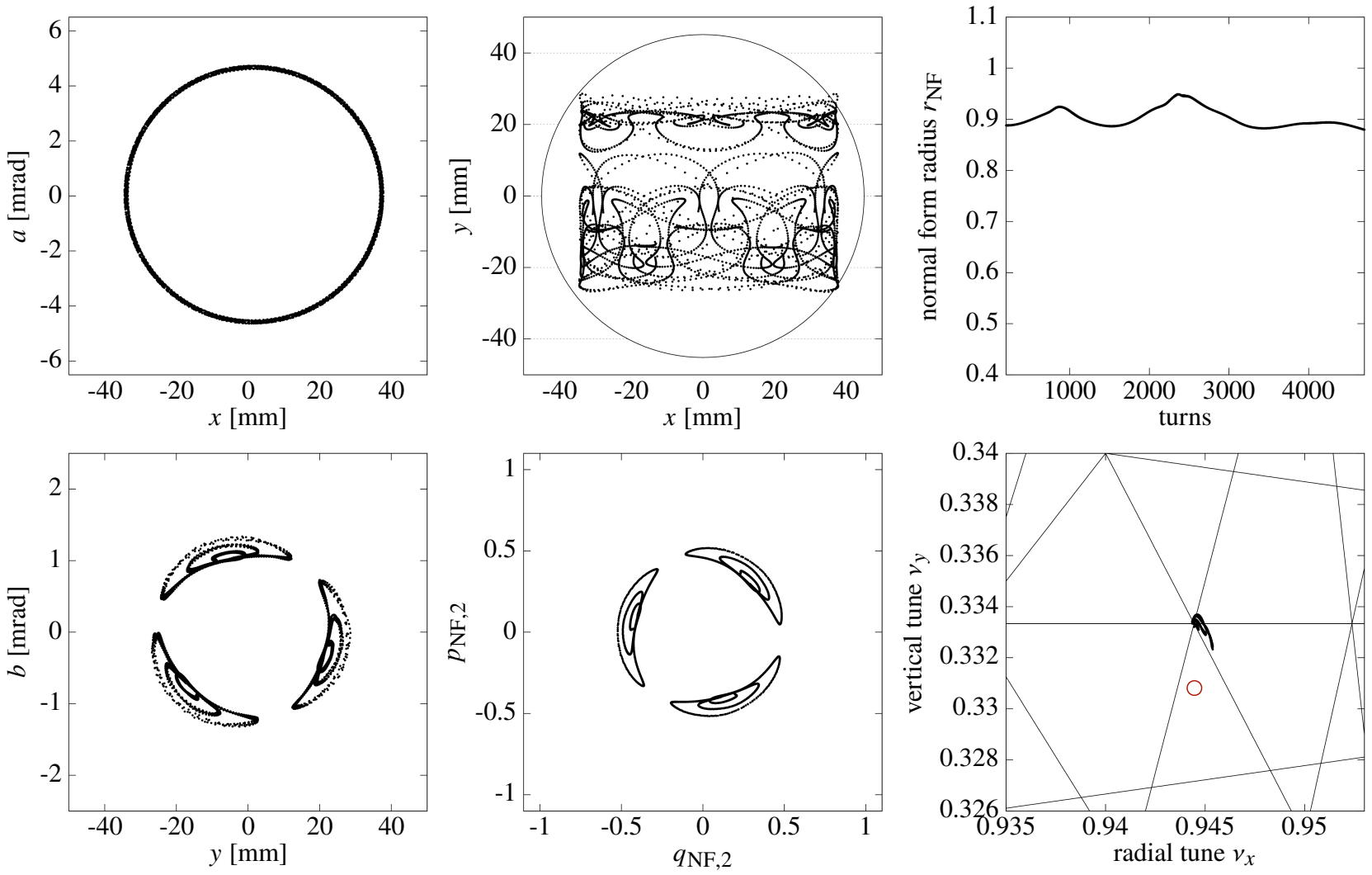

Figure 5.20: This particle ($\delta p = 0.024\%$) shows a different kind of moderate unstable behavior around a period-3 fixed point structure, where the island size varies. The particle has both, a large radial amplitude and the closeness to the $(17/18, 1/3)$ resonance point.

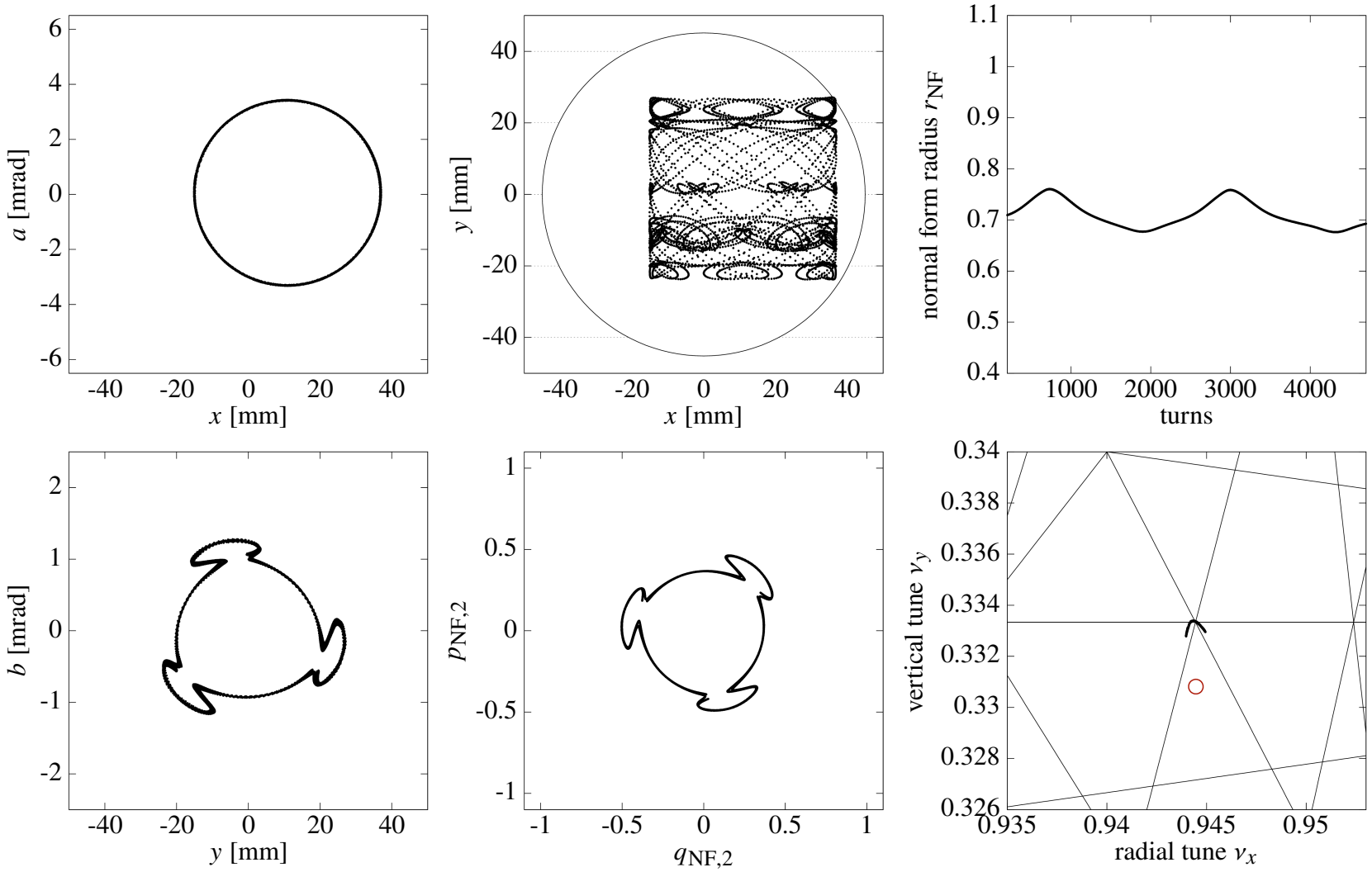

Figure 5.21: This particle ($\delta p = 0.140\%$) forms a shuriken like shape in the vertical phase space. In this pattern there are two period-3 fixed point structures involved indicated by the double crossing of the vertical $1/3$ resonance line.

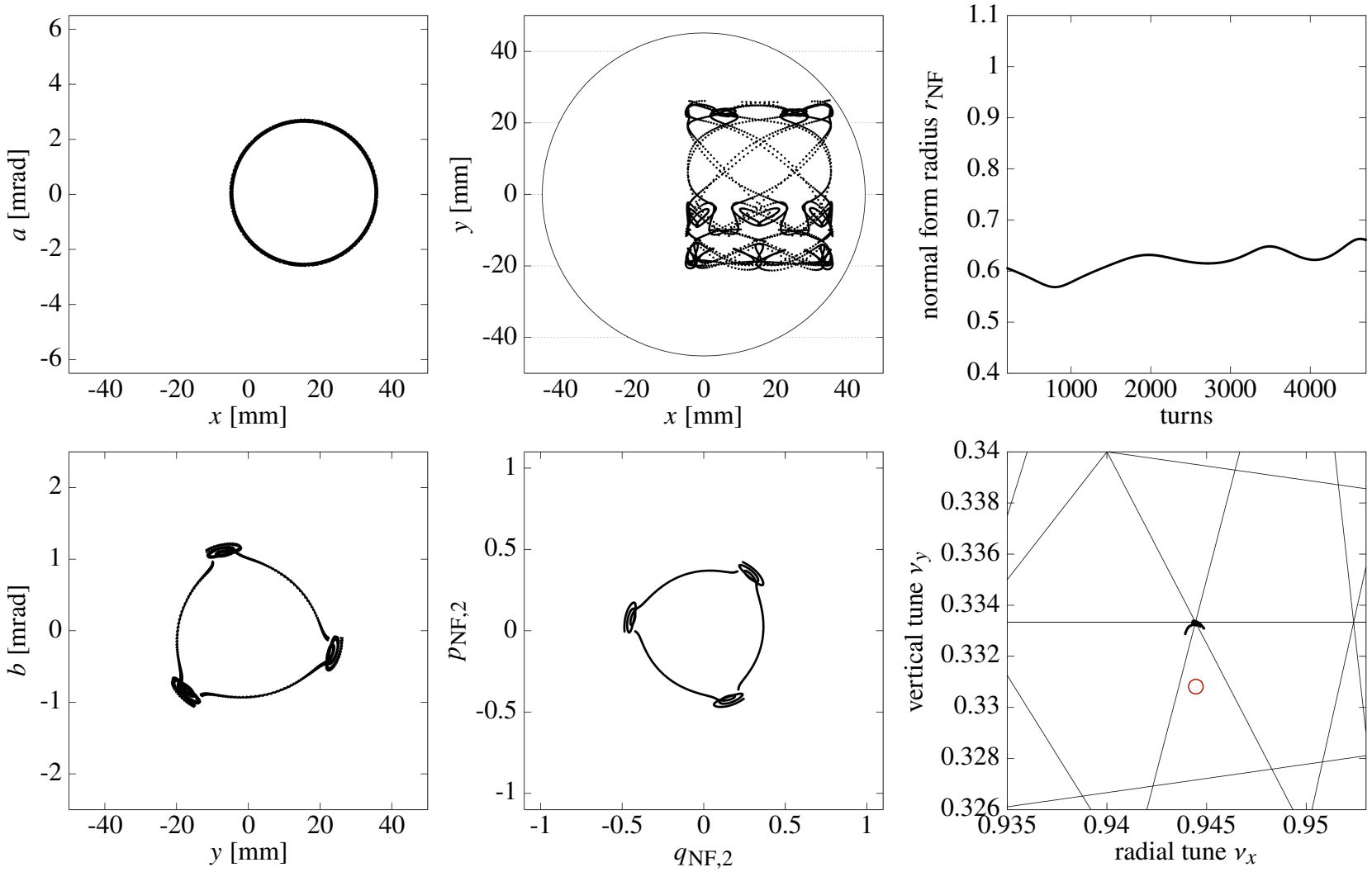

Figure 5.22: This particle ($\delta p = 0.196\%$) illustrates moderate unstable behavior in a shuriken pattern. The radial amplitude is not particularly large, but the resonance point $(17/18, 1/3)$ is very close, which might be the trigger of the unsuitability.

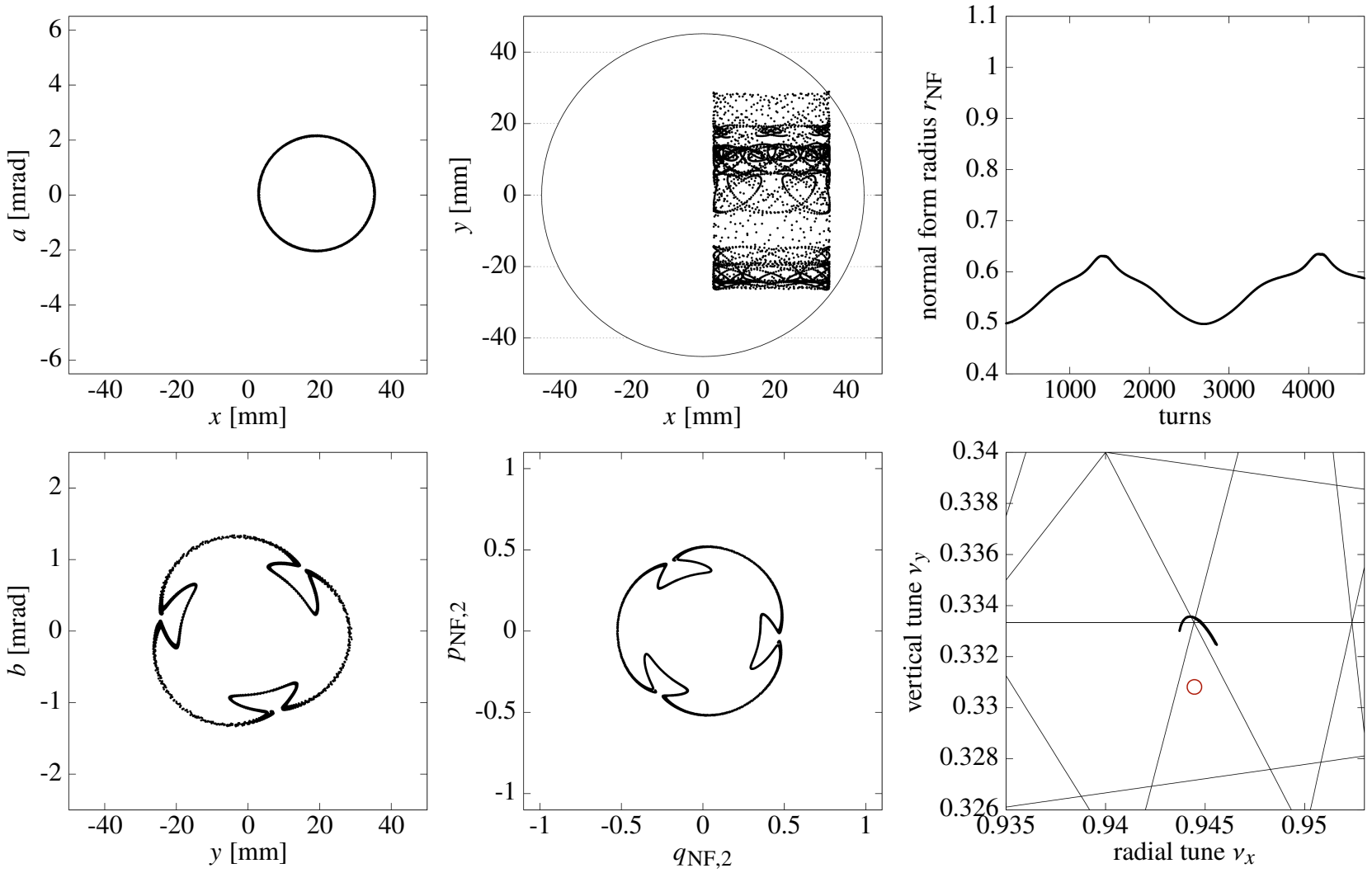

Figure 5.23: This particle ($\delta p = 0.242\%$) illustrates a shuriken pattern, where the two period-3 fixed point structures are more obvious. The muon experiences a major modulation in the vertical oscillation amplitude and performs a double crossing of the vertical $1/3$ resonance line.

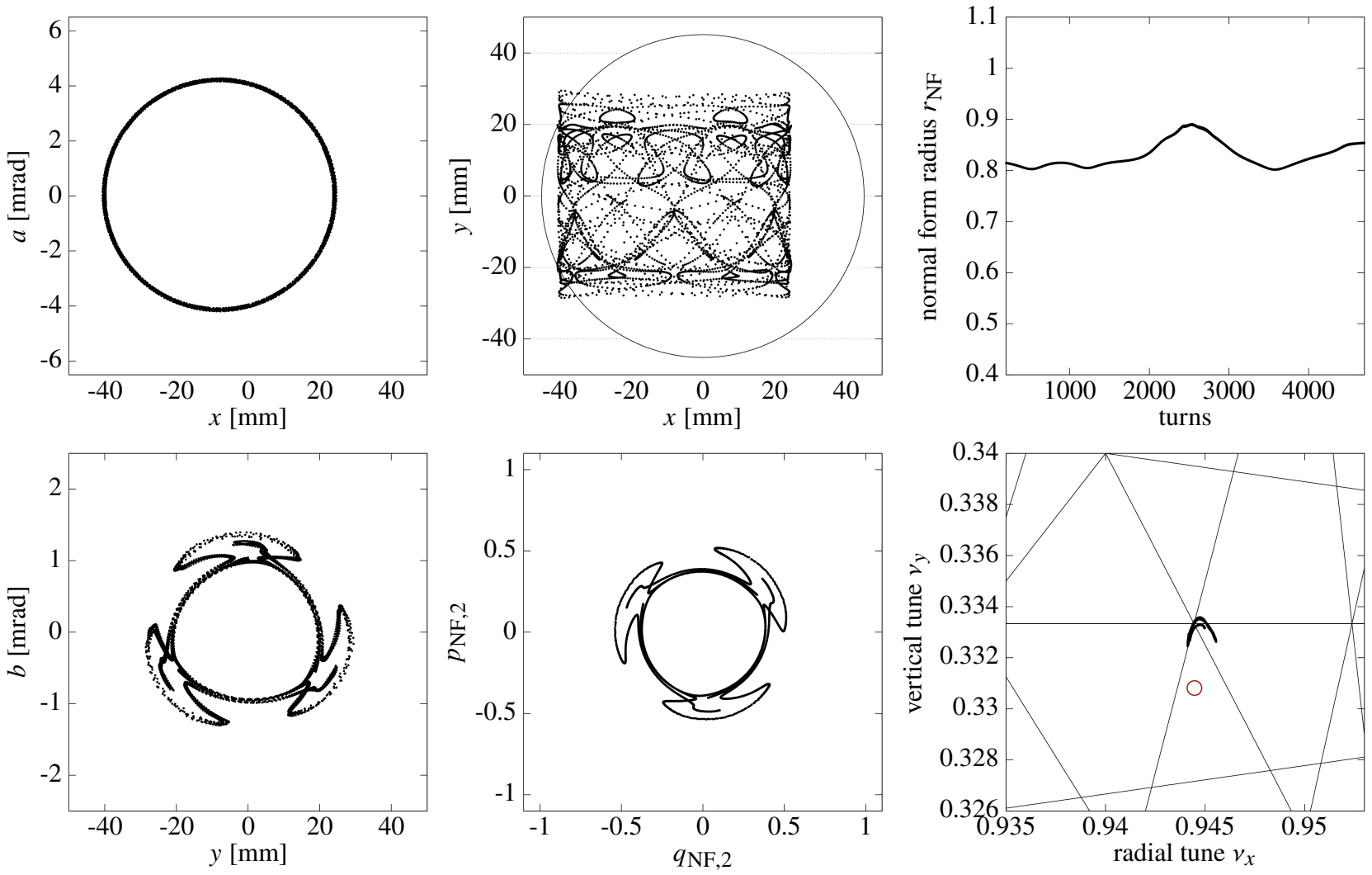

Figure 5.24: This particle ($\delta p = -0.096\%$) shows a shuriken pattern with unstable tendencies. The large radial amplitude and/or the closeness to the radial $17/18$ resonance line might be the trigger for the instability.

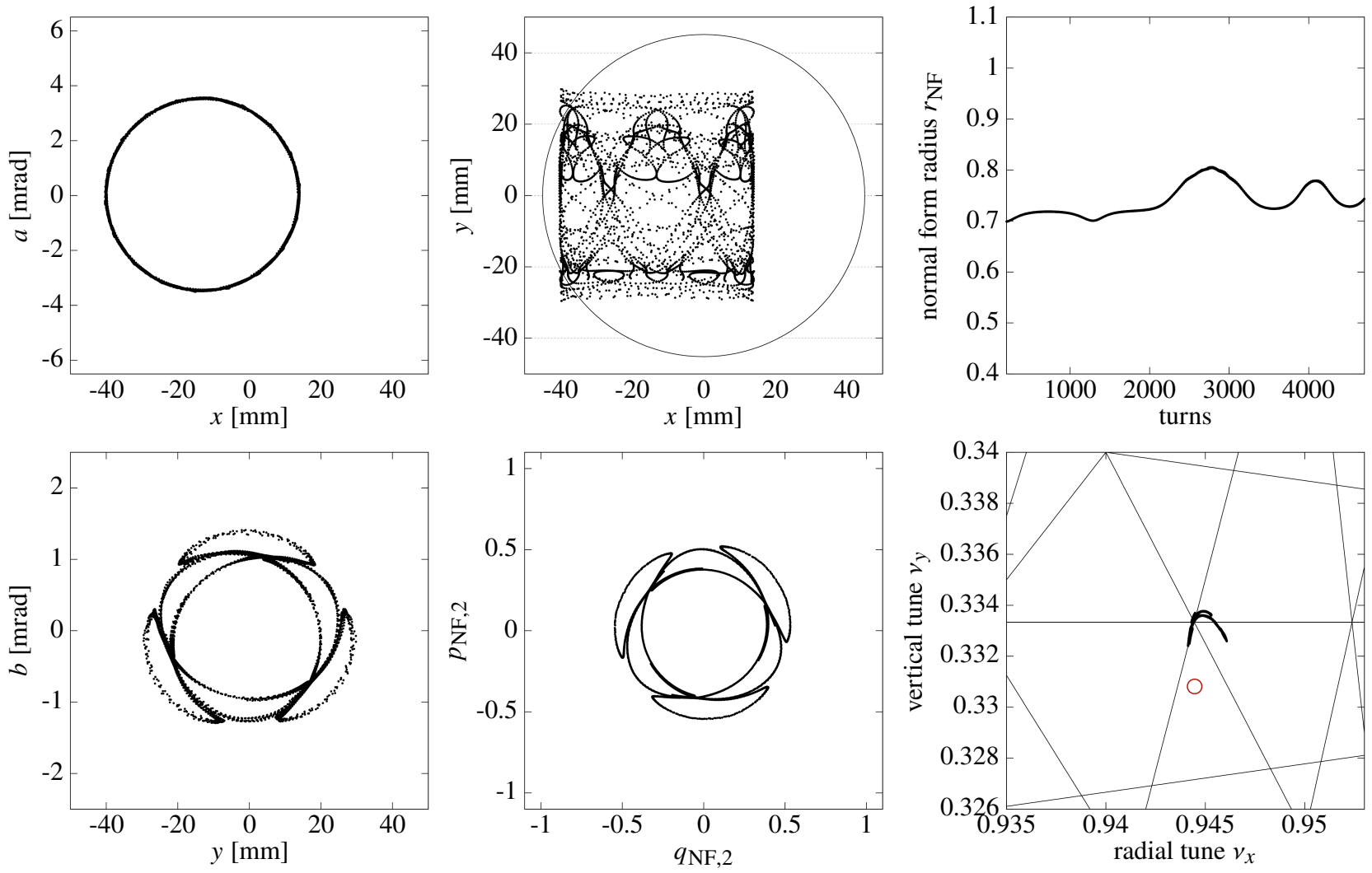

Figure 5.25: This particle ($\delta p = -0.159\%$) shows a shuriken pattern with a moderate instability. The two period-3 fixed point structures are so close together that the particle gets temporarily caught around the inner one of them in an island pattern.

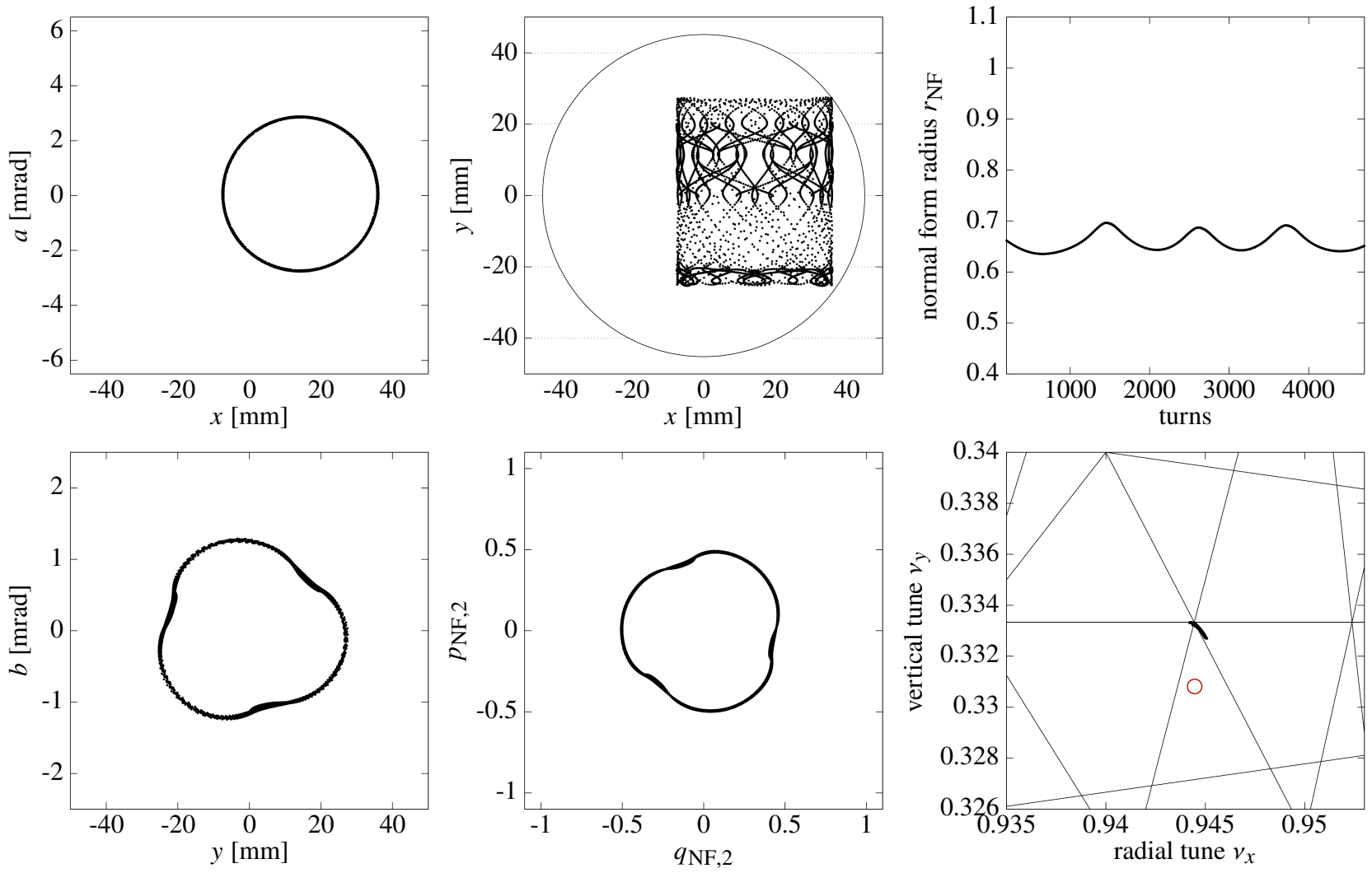

Figure 5.26: This particle ($\delta p = 0.181\%$) shows the pattern of a very blunt shuriken. The vertical amplitude oscillation is only moderate and illustrates there can be almost regular behavior between two period-3 fixed point structures.

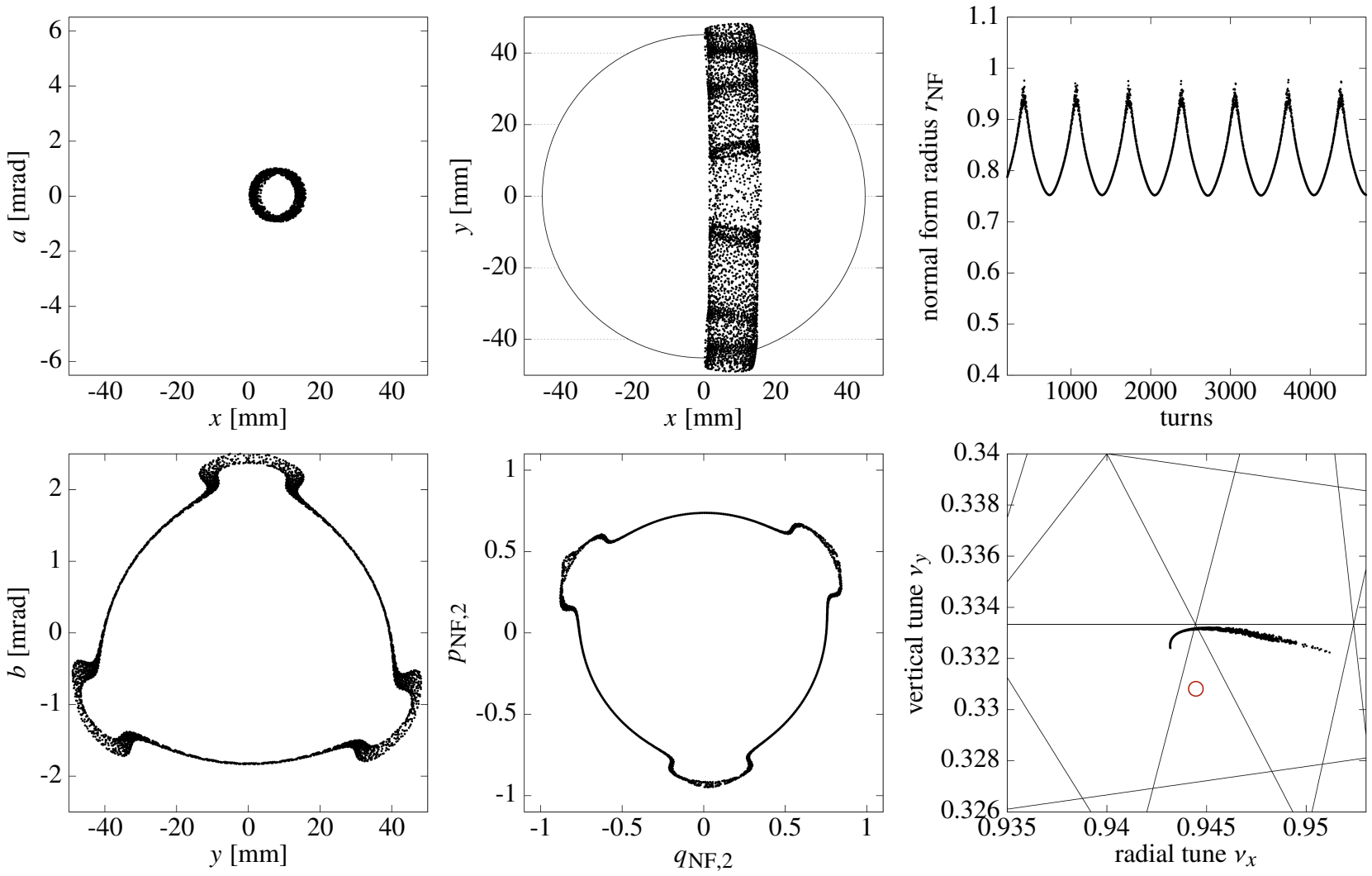

Figure 5.27: This particle ($\delta p = 0.106\%$) is characterized by a very large vertical amplitude, which is additionally modulated by the shuriken pattern. Its one of the very few particles for which the orbit considerably overlaps with the collimator boundary.

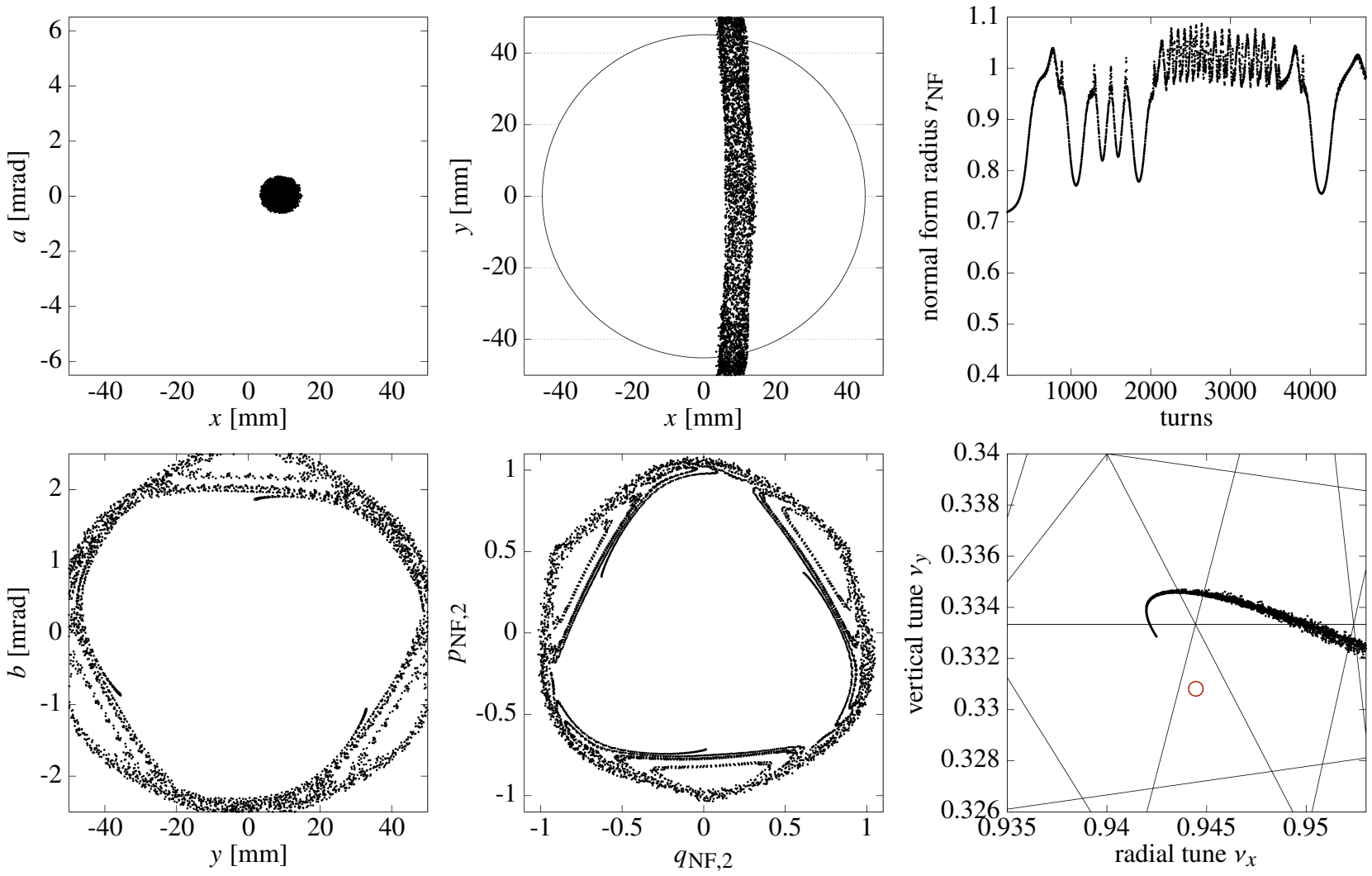

Figure 5.28: This particle ($\delta p = 0.118\%$) shows strong instabilities caused by a combination of a very large vertical amplitude in combination with a period-3 fixed point structure, which occasionally captures the orbit in an island pattern.

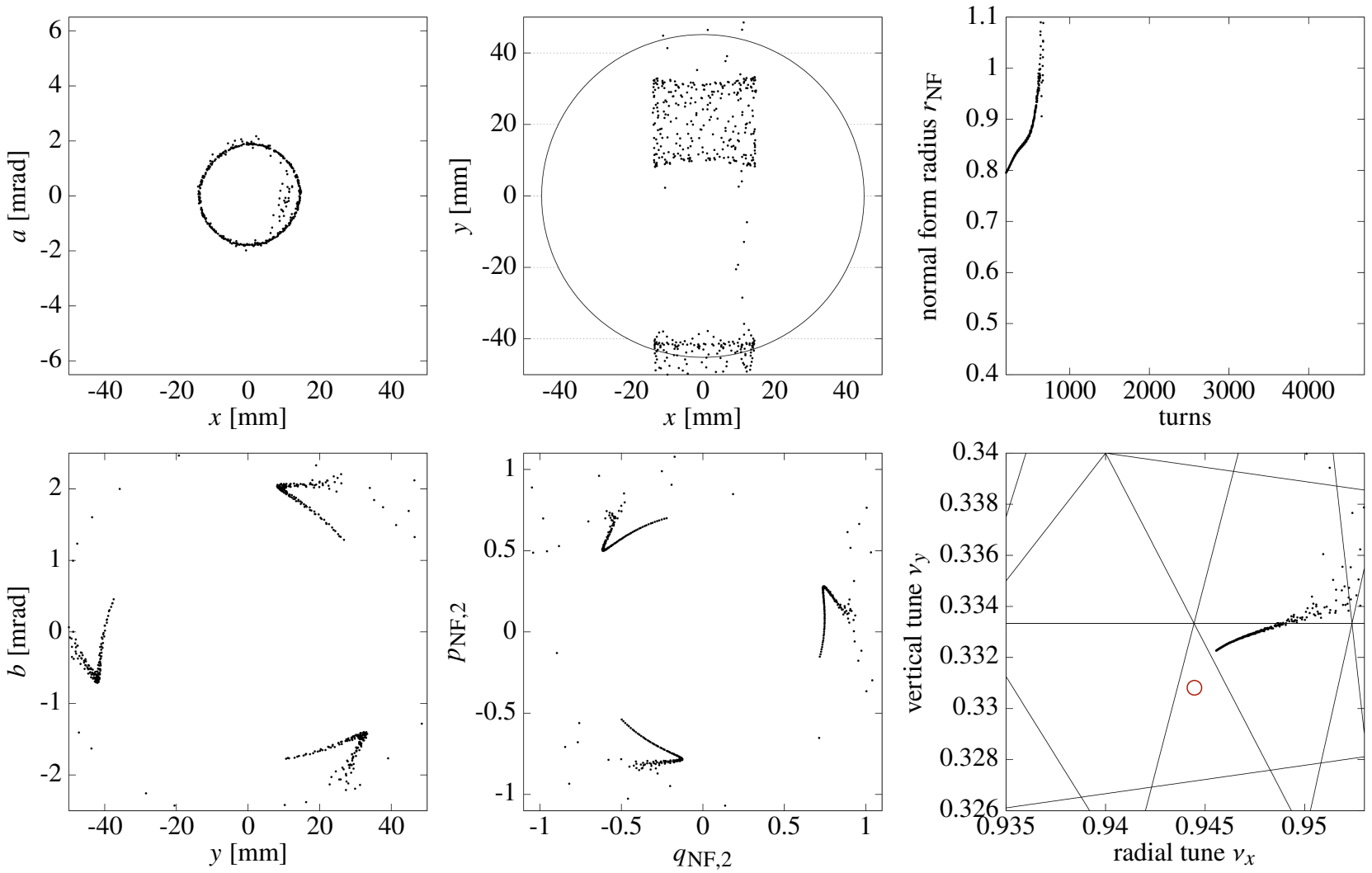

Figure 5.29: This particle ($\delta p = 0.010\%$) diverges due to its unstable orbit. The approach of the unstable fixed point with such a with the large vertical amplitude are likely the trigger of the divergence.

5.5.3 Period-3 Fixed Point Structures

There are period-3 fixed point structures in the vertical phase space as the particle in Fig. 5.17 suggests. The period-3 fixed points are a property of the vertical projection of the stroboscopic muon tracking. They are associated with the vertical $1/3$-resonance, which is particularly relevant due to the strong eight order nonlinear tune shifts from the strong ninth order nonlinear field contributions of the 20th order multipole of the potential from the electrostatic quadrupole system [73].

The period-3 fixed point structure corresponds to an orbit, which vertically oscillates around its momentum dependent reference orbit with a period of exactly three turns, i.e. a vertical betatron tune of $1/3$. However, such an orbit is not necessarily a closed orbit, which closes after three turns, because while the vertical behavior might be exactly resonant after three turns, the radial behavior is not.

There are attractive fixed points and repulsive fixed points within the period-3 fixed point structures. Accordingly, the term 'period-3 fixed points' describes a set of 6 fixed points at the same amplitude in yb, where every other fixed point is attractive. The positions of the period-3 fixed points in the vertical phase space depend on the momentum offset δp and the radial phase space (due to coupling). Attractive fixed points 'capture' particles in their reach, creating island patterns (see Fig. 5.30). The unstable fixed points push the particles away. They are at those blank spaces between any two islands.

The inner red orbit and the adjacent blue island orbit in Fig. 5.30 illustrate how abruptly the vertical phase space behavior changes around these period-3 fixed point structures. The muon initiated on the inner red orbit exhibits an oscillation with constant amplitude. The muon initiated at a slightly larger amplitude on the blue orbit initially seems to follow a similar elliptical orbit with constant amplitude as the red particle before it gets pushed back and outwards by the unstable fixed point, which drastically increases the vertical amplitude of the particle. In the case shown in Fig. 5.30, the attraction of the stable fixed point is strong enough to keep the particle in an island orbit. In Fig. 5.29, on the other hand, the particle can not remain on a bounded orbit and diverges.

It is also not uncommon for two period-3 fixed point structures to be present simultaneously

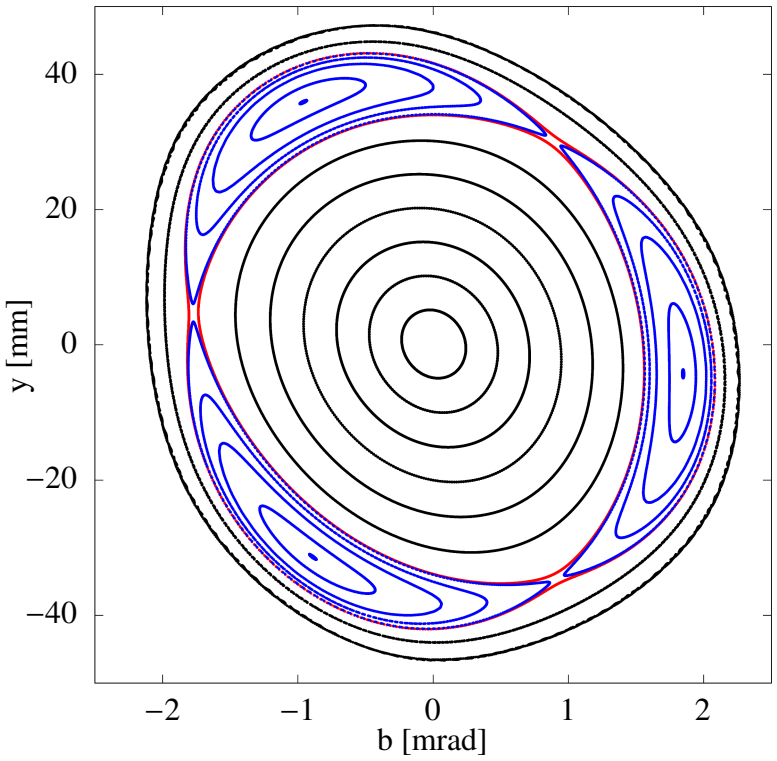

Figure 5.30: Stroboscopic tracking in the vertical phase space illustrating orbit behavior with a single period-3 fixed point structure present. The orbits only differ in their vertical phase space behavior – they all have the same momentum offset of $\delta p = 0.126$ % and are at the momentum dependent equilibrium point in radial phase space ($x = 10.64$ mm, $a = 0.045$ mrad) and therefore have no radial oscillation amplitude. The blue orbits indicate the island patterns around the attractive fixed points in the middle of the islands. The red orbits are right at the edge before being caught around the fixed points. The three repulsive fixed points are in the space between the two red orbits, where the islands almost touch.

in the vertical phase space. The structures are oriented similarly, only having their attractive and repulsive fixed points switched, such that an attractive fixed point of structure with the larger amplitude is 'above' a repulsive fixed point of the structure with the lower amplitude. In Fig. 5.31 a phase space region with two period-3 fixed point structures for orbits with $\delta p = 0.339$ % are shown. The different plots illustrate how the relative position and interaction of the two period-3 fixed point structures change with different oscillation amplitudes in the radial phase space. The two period-3 fixed point structures can be well separated, yielding the known island patterns with 'regular ' orbits in between. However, the structures can also move into each other such that some orbits are caught between the two period-3 fixed point structures and follow the shape of a threefold

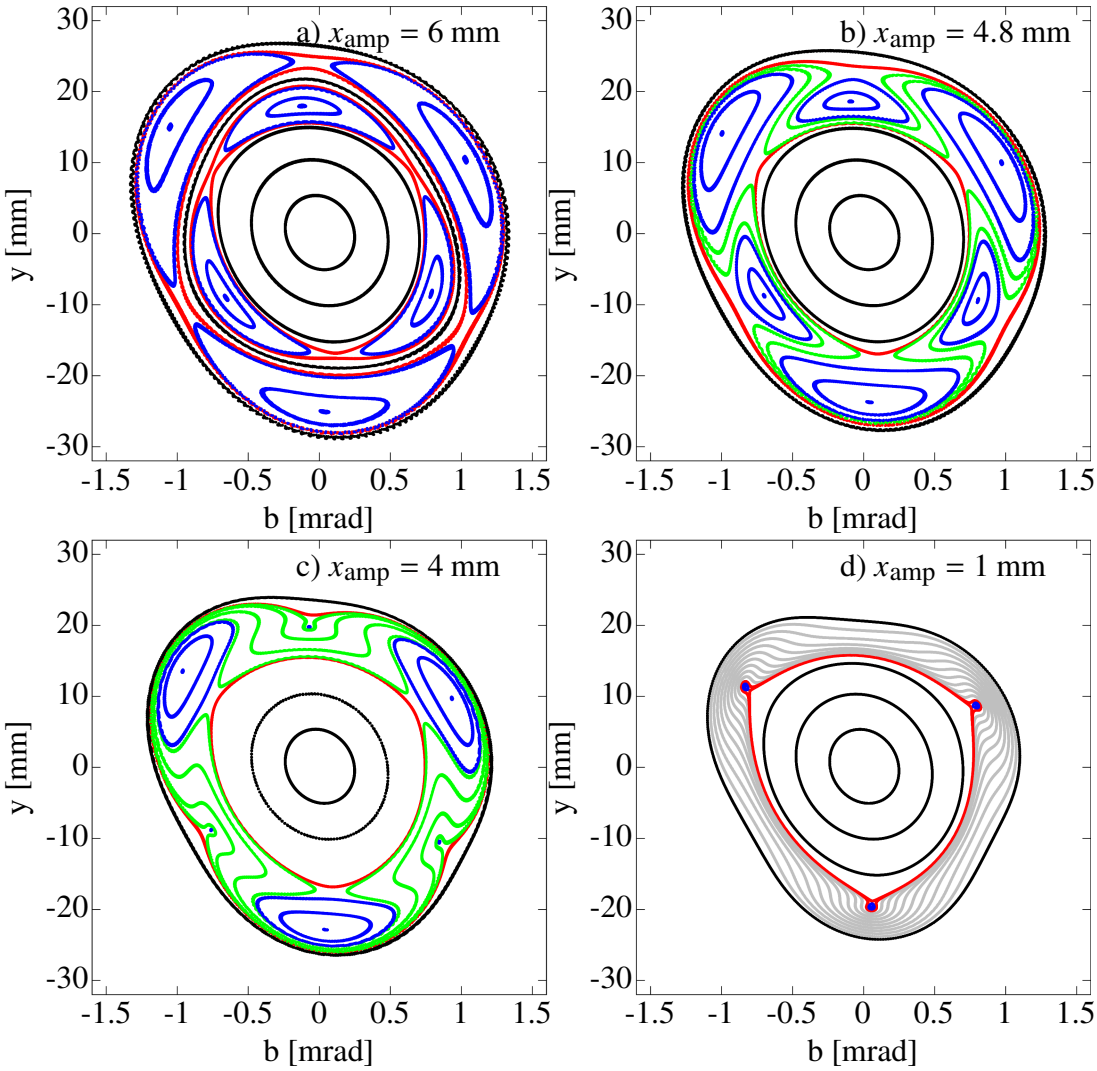

Figure 5.31: Stroboscopic tracking in the vertical phase space illustrating orbit behavior with two period-3 fixed point structures present. The orbits in each plot only differ in their vertical phase space behavior. All orbits have the same momentum offset of $\delta p = 0.339$ %. The four plots differ by their radial amplitude around the momentum dependent equilibrium point in radial phase space at ($x = 27.7$ mm, $a = 0.144$ mrad). The radial amplitudes are: a) $x_{amp} = 6$ mm, b) $x_{amp} = 4.8$ mm, c) $x_{amp} = 4$ mm, d) $x_{amp} = 1$ mm. The blue orbits indicate the island patterns around the attractive fixed points. The red orbits are right at the edge before being caught around the period-3 fixed points. The green orbits are caught around both period-3 fixed point structures. The gray orbits in d) emphasize that half of the fixed points from c) have indeed been annihilated.

shuriken around the two island patterns. When the two period-3 fixed point structures come even closer, the opposite fixed points of the two period-3 fixed point structures can annihilate each other, resulting in triangular patterns with rounded corners.

While the period-3 fixed point structures often lead to a significant vertical amplitude modulation,

many of them are well within the boundary of the collimators like the examples shown in Fig. 5.30 and Fig. 5.31. So, the involvement of a particle in a period-3 fixed point structure or two does not necessarily mean that it is lost, but the additional modulation of the vertical amplitude surely increases the general risk of getting lost.

All orbit patterns shown in Fig. 5.15 to Fig. 5.29 can be found in a similar form either in Fig. 5.30 or Fig. 5.31. In other words, we fully understand what is causing the different types of patterns. The major difference for some particles is the stability of their pattern. The phase space regions chosen in Fig. 5.30 and Fig. 5.31 are stable and do not share the characteristics of unstable orbits which are large radial amplitudes and/or closeness to the $17/18$ resonance point.

5.5.4 Muon Loss Rates from Simulation

We have seen what different phase space tracking patterns can arise due to period-3 fixed point structures. We also saw that these structures can be responsible for losses due to the modulation of the oscillation amplitude in the vertical phase space. To get a more general understanding of how prominent these patterns are among the entire distribution and how common they are among lost particles, we need a mechanism to characterize these patterns in a way that can be automatically detected.

The various degrees of instabilities, especially among particles involved with period-3 fixed point structures make a generalized categorization difficult. There is no obvious distinction between certain unstable islands and certain shuriken patterns, and also no clear distinction between very blunt shuriken patterns and very triangularly deformed elliptical patterns. Accordingly, we only make two distinctions. First, we distinguish between particles involved with the vertical $1/3$-resonance and particles that are not. Among the particles that are involved with the vertical $1/3$-resonance, we make a further distinction between pure island patterns and everything else. A pure island pattern is a (non-across-jumping) island pattern. Fig. 5.19 shows an across-jumping island structure, where the orbit jumps from one fixed point island to another. In comparison, Fig. 5.20 shows an island pattern that is also unstable but remains on the island around the fixed points.

For reference, we will call particles involved with the vertical $1/3$-resonance 'period-3 particles'and all the others 'regular particles'. Of the period-3 particles, we will only give a special name to the 'island particles', because the period-3 non-island particles are a very diverse group, which is not easily described by a single word without mischaracterizing at least some of its elements.

Since the transition between patterns is continuous as the gray orbits in Fig 5.31d illustrates, the category of period-3 particles and the category non-period-3 particles might have elements that are almost identical.

To make those distinctions, we start by explaining how period-3 particles are identified. We consider the vertical phase space in polar coordinates and look at the phase space behavior in steps of three. The first three vertical phase space angles during tracking are denoted by $\phi_{1,0}$, $\phi_{2,0}$ and $\phi_{3,0}$, and the next three angles are denoted by $\phi_{1,1}$, $\phi_{2,1}$ and $\phi_{3,1}$, and so forth. Additionally, we define the angle advances $\Delta\phi_{i,n} = \phi_{i,n} - \phi_{i,n-1}$. To avoid ambiguity in the value for the angles, we require that value for $\phi_{i,n}$ is chosen such that $\phi_{i,n} \in \left[\phi_{i,n-1} - \pi, \phi_{i,n-1} + \pi\right]$. If there is a sign change from $\Delta\phi_{i,n-1}$ to $\Delta\phi_{i,n-1}$ for all three angle advances, then the $1/3$-resonance tune was crossed and we categorize the particle as period-3 particle.

To identify island particles, we use the definitions from above and additionally introduce the range $\mathbb{D}_{i,0} = \left[\phi_{i,0}, \phi_{i,0}\right]$ of the angles for each of the three potential island locations. With every iteration step the ranges of the angles are updated to

$$\mathbb{D}_{i,n} = \left[\mathbb{D}_{i,n-1,\mathrm{LB}}, \mathbb{D}_{i,n-1,\mathrm{UB}}\right] = \left[\min\left(\phi_{i,n-1}, \mathbb{D}_{i,n-1,\mathrm{LB}}\right), \max\left(\phi_{i,n-1}, \mathbb{D}_{i,n-1,\mathrm{UB}}\right)\right]. \tag{5.3}$$

The abbreviations 'LB' and 'UB' denote the lower and upper bound of the domain respectively. Note that the rule to avoid ambiguity in the value for the angles from above also applies here. All particles for which the total range over the three potential island domains after the 4500 tracking turns is less than a full revolution (2π) are considered island particles. In other words, island particles satisfy

$$\sum_{i=1}^{3} |\mathbb{D}_{i,1500}| < 2\pi. \tag{5.4}$$

With these recognition mechanisms implemented, we were able characterize all particles and determine their proportion as presented in Tab. 5.1. Period-3 particles are over-represented among

Table 5.1: Percentages of different characterization groups. Read as follows: x % of *Base* particles have the property *Property*. All particles that hit a collimator during the 4500 turns of tracking are considered lost.

Property \ Base	All	Lost	Period-3	Island
Lost	0.686%	100%	7.44%	22.2%
Period-3	7.06%	76.4%	100%	100%
Island	1.00%	32.4%	14.2%	100%

lost particles by a factor of almost eleven compared to their appearance in the entire distribution. For island particles, this discrepancy is even more drastic with a factor of 32. Accordingly, period-3 particles and island particles, in particular, are more prone to be lost. But by far not every period-3 particle or island particle is lost. More than 87% of island particles and more than 92% of period-3 particles survive the 4500 turns. As Fig. 5.30 illustrates, sometimes the amplitude of these period-3 structures is so low that the additional modulation of the amplitude is not enough to be critical.

While island particles make up only $1/7$ of period-3 particles, they are responsible for almost half the losses associated with period-3 particles. This is particularly surprising because the island particle category excludes most unstable patterns by definition (exceptions are moderate instabilities that do not contravene the recognition criteria like the particle shown in Fig. 5.20). On the other hand, period-3 particles cover a wide range of patterns some of which barely show a modulation of the vertical amplitude as the example of the gray orbits in Fig. 5.31d shows.

To understand how the losses occur over time, we plot the accumulative loss ratio over the 4500 turns in Fig. 5.32. Island loss is the fastest growing loss over the first 1000 turns before settling almost asymptotically. This is explained by different modulation frequencies around the period-3 fixed point structures. The closer to the stable fixed point, the faster the modulation, and the closer to the unstable fixed points the slower the modulation. Accordingly, the island modulation is on average faster than the shuriken modulation.

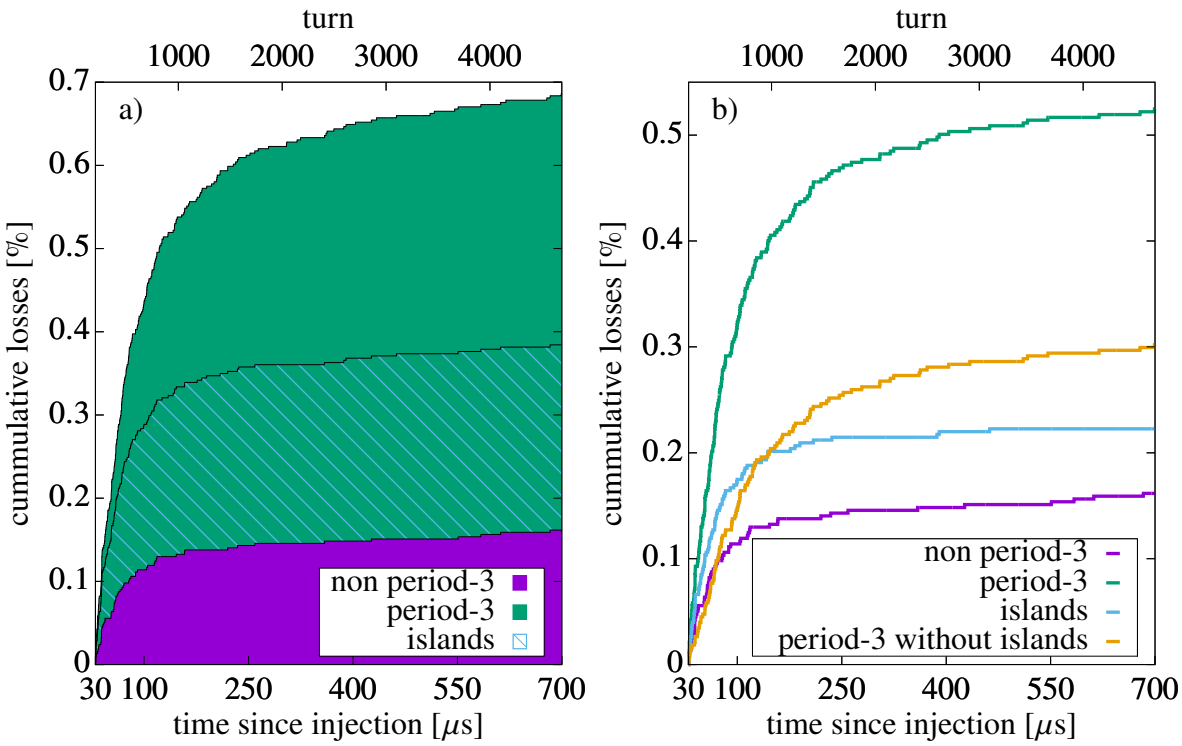

Figure 5.32: a) Shows how the muon loss ratio is composed of particles with constant oscillation amplitudes (purple) and particles involved with resonances (green). Of the particles involved with resonances (green), the fraction caught in islands structures is indicated by the blue stripe pattern. In b) the loss ratio over time is shown for each subgroup of lost particles to better understand which losses drive to overall loss from plot a). The tracking starts after the initial 30 μs of scraping when data taking is initiated.

5.6 Conclusion

The Poincaré return map description of the storage ring model of the muon g-2 experiment [77] and its analysis with DA normal form methods yielded many insightful characteristics of the system. We gained an understanding of the form of the closed orbit within the storage ring as well as details on how it changes with an offset in the momentum δp. Considering that particles oscillate around their corresponding reference orbit, which is the closed orbit of their momentum offset, the radial shift of the closed orbit with momentum offset is particularly critical. This shift brings the equilibrium state of the radial oscillation closer to the collimator boundary, which increases the risk of getting lost.

The tune analysis provided a detailed understanding of how the oscillation frequencies of

particles dependent on their momentum offset and their amplitudes relative to their respective reference orbit. This analysis showed that particles over the entire momentum offset range could cross the vertical $1/3$-resonance frequency for certain vertical and radial amplitude combinations.

The strong ninth order nonlinearities of the map caused by the 20th-pole of the ESQ potential have a significant effect on amplitude and parameter dependent tune shifts. This property manifests itself in the dominating eighth order dependencies in the amplitude and momentum dependent tune shifts and the drastic change in the tune footprint for calculations of order $m > 8$, which include the ninth order terms of the original map.

Further tracking analysis revealed period-3 fixed point structures in the vertical phase space. They are associated with the vertical $1/3$-resonance tune and cause significant vertical amplitude modulations to the particles that are caught around them. We were able to connect all vertical phase space patterns of lost particles with patterns that arise around one or two of these period-3 fixed point structures. Additionally, instabilities caused by large radial amplitudes and/or closeness to the $(17/18)$ resonance point significantly mixed multiple of the known orbit patterns. This only allowed for a limited automatic recognition of patterns, which in turn revealed valuable insights about the effect of these period-3 fixed point structures on the loss rates of muons in the storage ring. Particles associated with period-3 fixed point structures are at an eleven-fold to 32-fold risk of getting lost, compared to particles not crossing the vertical $1/3$-resonance frequency.

CHAPTER 6

VERIFYING CALCULATIONS USING TAYLOR MODELS

In this chapter, we take steps towards making the methods presented above fully self-verified. Since many aspects that have to be carefully considered for a rigorous transfer to the verified world lay beyond the scope of this thesis, this chapter will only yield a discussion of the basic principles behind some of them. However, the aspect of verified global optimization and its application to the normal form defect for verified stability estimates will be analyzed in greater detail.

To introduce the concept of verified global optimization using Taylor Models, we apply it to two example optimization problems. First, in Sec. 6.1, we run a Taylor Model based global optimization in different operating modes on the generalized Rosenbrock function, as it is one of the most commonly used examples to test global optimization algorithms. In Sec. 6.2, we discuss the optimization problem of finding minimum energy configurations of particles that have their pairwise interaction energy modeled by the Lennard-Jones potential. It is one of the simplest examples to explain, yet arbitrarily complex to solve depending on the number of particles in the configuration and the dimensionality of the configuration.

In Sec. 6.3, we will discuss the intricacies of verifying the methods from the applications from Chapter 4 and Chapter 5 for a verified stability analysis of those dynamical systems. In particular, we will take a detailed look at the normal form defect and use the gained understanding of verified global optimization from Sec. 6.1 and Sec. 6.2 to analyze its application to the normal form defect of the simulated phase space behavior in the muon g-2 storage ring.

6.1 The Rosenbrock Optimization Problem

6.1.1 The Rosenbrock Function

The Rosenbrock function

$$f(x, y) = (a - x)^2 + b\left(y - x^2\right)^2 \tag{6.1}$$

was introduced by Howard H. Rosenbrock in 1960 [69]. It is a non-convex function that is commonly used as a test problem for global optimization algorithms. The parameters are usually set to $(a, b) = (1, 100)$, and so we will use those parameters here as well. Fig. 6.1 illustrates the Rosenbrock function for those parameters.

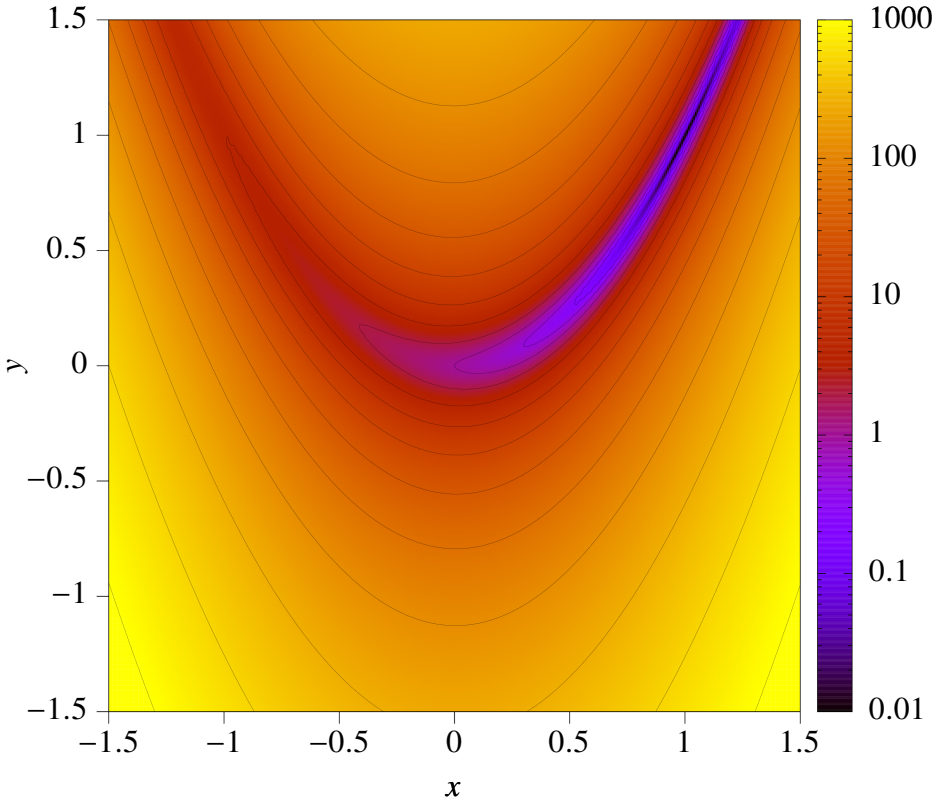

Figure 6.1: The Rosenbrock function with $(a, b) = (1, 100)$.

It is also referred to as Rosenbrock's valley function or Rosenbrock's banana function for obvious reasons. The Rosenbrock function is characterized by a very deep valley, the floor of which constitutes a very shallow valley. This shallowness is one of the aspects that challenges global optimizers.

There are various multidimensional generalizations of the Rosenbrock function to compare more advance global optimization algorithms. In this work, we will use the following generalized form

$$f_{n\mathrm{D}}\left(\vec{x}\right) = \sum_{i=1}^{n-1} \left[100 \left(x_{i+1} - x_i^2 \right)^2 + (1 - x_i)^2 \right], \tag{6.2}$$

where $n \geq 2$ is the dimension and x_i are the optimization variables. Note that this generalized definition is consistent for the definition of the 2D Rosenbrock function from above and also retains the difficulties of the original problem of steep valley with a shallow valley floor, but with a complexity that increases with n. Unless specified otherwise, we will refer to the generalized Rosenbrock function as the Rosenbrock function or the objective function of the optimization.

The Rosenbrock function is a composition of quadratic expressions. None of the individual terms in the sum can be negative. Accordingly, a global minimum would be reached if all individual terms of the sum are zero. The $(1 - x_i)^2$ terms are only zero for $x_i = 1$, which also yields zero for the remaining terms. Accordingly, $\vec{x}^\star = (1, 1, ..., 1)$ is the single global minimum of the Rosenbrock function for which every term is zero and therefore the overall objective function is zero.

In Fig. 6.2, the Rosenbrock function is illustrated in multiple 2D projections around its minimum at $\vec{x}^\star$. In other words, all x_i are set to one except for the variables shown in the projection.

The dependency problem of the Rosenbrock function is rather mild. For the first variable x_1, the following dependent terms appear

$$100 \left(x_2 - x_1^2 \right)^2 + (1 - x_1)^2 \tag{6.3}$$

For any of the variables x_i with $1 < i < n$, there is one additional dependent term with

$$100 \left(x_{i+1} - x_i^2 \right)^2 + (1 - x_i)^2 + 100 \left(x_i - x_{i-1}^2 \right)^2 \tag{6.4}$$

However, there is no dependency problem with regard to the last variable x_n as it only appears in the term

$$100 \left(x_n - x_{n-1}^2 \right)^2 \tag{6.5}$$

Because of the double squares, the Rosenbrock function is always a fourth order polynomial.

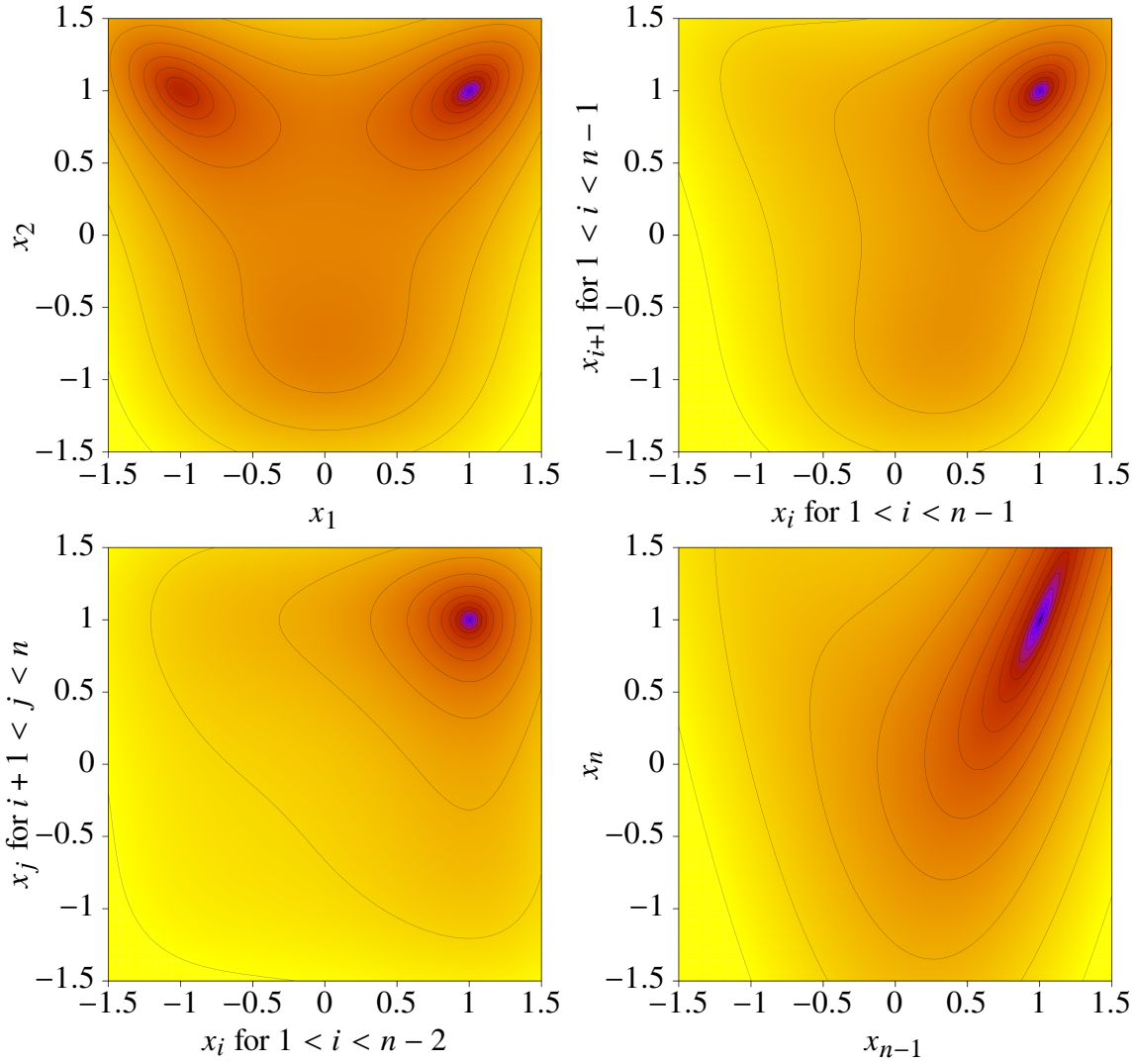

Figure 6.2: Projections of the multidimensional generalizations of the Rosenbrock function (Eq. (6.2)) into 2D-subspaces around minimum at $\vec{x} = (1, 1, ..., 1)$, i.e., all variables are equal one except for the ones shown in the respective plot.

6.1.2 Global Optimization Using COSY-GO

The global optimization is performed using COSY-GO [55, 56]. In the most advanced setting (QFB/LDB), the algorithm uses both of the advanced Taylor Model based bounding methods, namely, the quadratic fast bounder (QFB) and the linear dominated bounder (LDB), which were mentioned in Sec. 2.6 and were introduced in [56]. Additionally, COSY-GO also uses naive Taylor Model bounding and interval evaluations (IN). For comparisons, COSY-GO offers to run an optimization with some of the advanced methods disabled. By ranking the bounding methods in the order: QFB,

LDB, naive TM, and IN, the operating mode is denoted by its highest ranking bounding method, e.g., the running mode LDB indicates that LDB, naive TM, and IN are used but not QFB.

Because the global minimum is already known, we are just interested in the algorithm's performance to narrow down the domain of the minimum and its value. Accordingly, we can choose an arbitrary search domain for the optimization that includes $\vec{x}^{\star}$. We will investigate the Rosenbrock function over the domain $[-1.5, 1.5]^n$.

For the optimization, we evaluate the objective function the way it is written in Eq. (6.2) and not expanded out in a single second, third, and fourth order polynomials. In particular, the optimization is performed with no additional knowledge about the derivatives of the objective function.

In Fig. 6.3, the performance of the algorithm on the 2D Rosenbrock function is visualized in the form of its splitting pattern. It shows the individual boxes analyzed by in the various operation modes. All calculations are performed with fourth order Taylor Models except for the interval evaluation, which does not use TM.

The significant differences in the splitting patterns are the number of splits, and the way boxes are split. For the operating mode in naive TM and IN, boxes are always split in half. With LDB and QFB, the boxes are split as the respective method sees fit. Especially close to the minimum this avoids the cluster effect [38, 30]. In Fig. 6.4, the boxing close to the minimum is illustrated, which clearly shows the cluster effect.

Another advantage of the Taylor Model based approach is the avoidance of the dependency problem. However, due to the simplicity of the 2D Rosenbrock function and its weak dependency problem, the advantages of the Taylor Model based methods are not so prominent relative to the IN evaluations. For more complex higher dimensional comparisons, a visualization of the boxing is not easily possible. However, to still visually emphasize the advantages of the TM operations over intervals, we artificially increase the dependency problem in the objective function by modifying it to $f = f_{2\mathrm{D}} - f_{2\mathrm{D}} + f_{2\mathrm{D}}$. In Fig. 6.5, the QFB/LDB methods using fourth order Taylor Models are compared to the interval method for the modified objective function.

Even though the fourth order TM representation of the modified objective function only differs

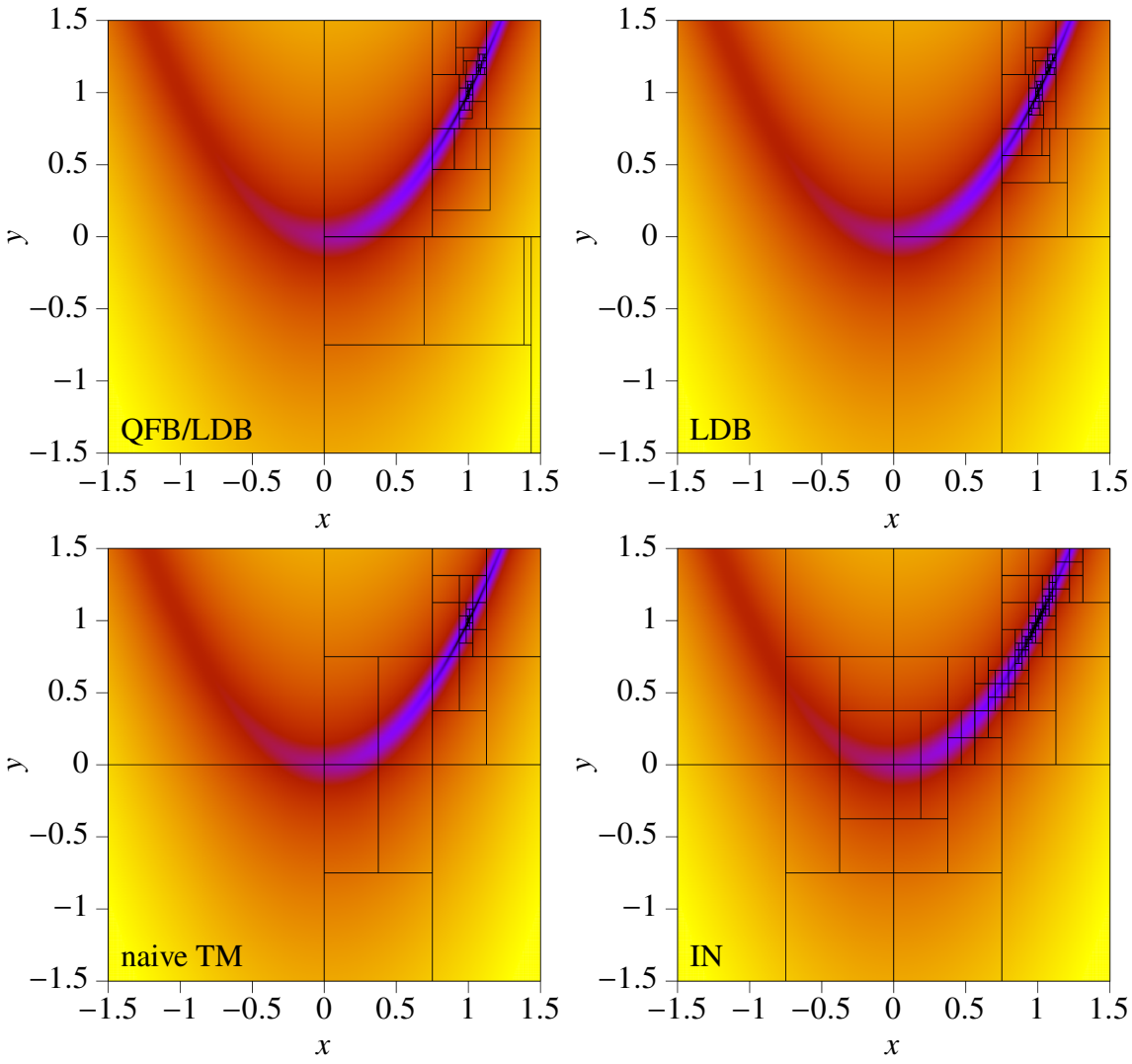

Figure 6.3: Global optimization of the 2D Rosenbrock function using COSY-GO in different operation modes with fourth order Taylor Models for all modes except interval evaluations (IN).

from the TM representation of the unmodified objective function by a slightly larger remainder bound, the behavior and efficiency of the algorithm changes more for the modified objective function than one would initially expect. This is due to the fact that the algorithm also performs intermediate steps with lower order Taylor Models, which are quicker to evaluate but less accurate. Those lower order evaluations are more sensitive to the dependency problem, which explains the affect of those intermediate steps on the splitting decisions.

Next, we analyze the performance of COSY-GO for the optimization of the higher dimensional Rosenbrock functions without the artificially added dependency problem. The search domain of the

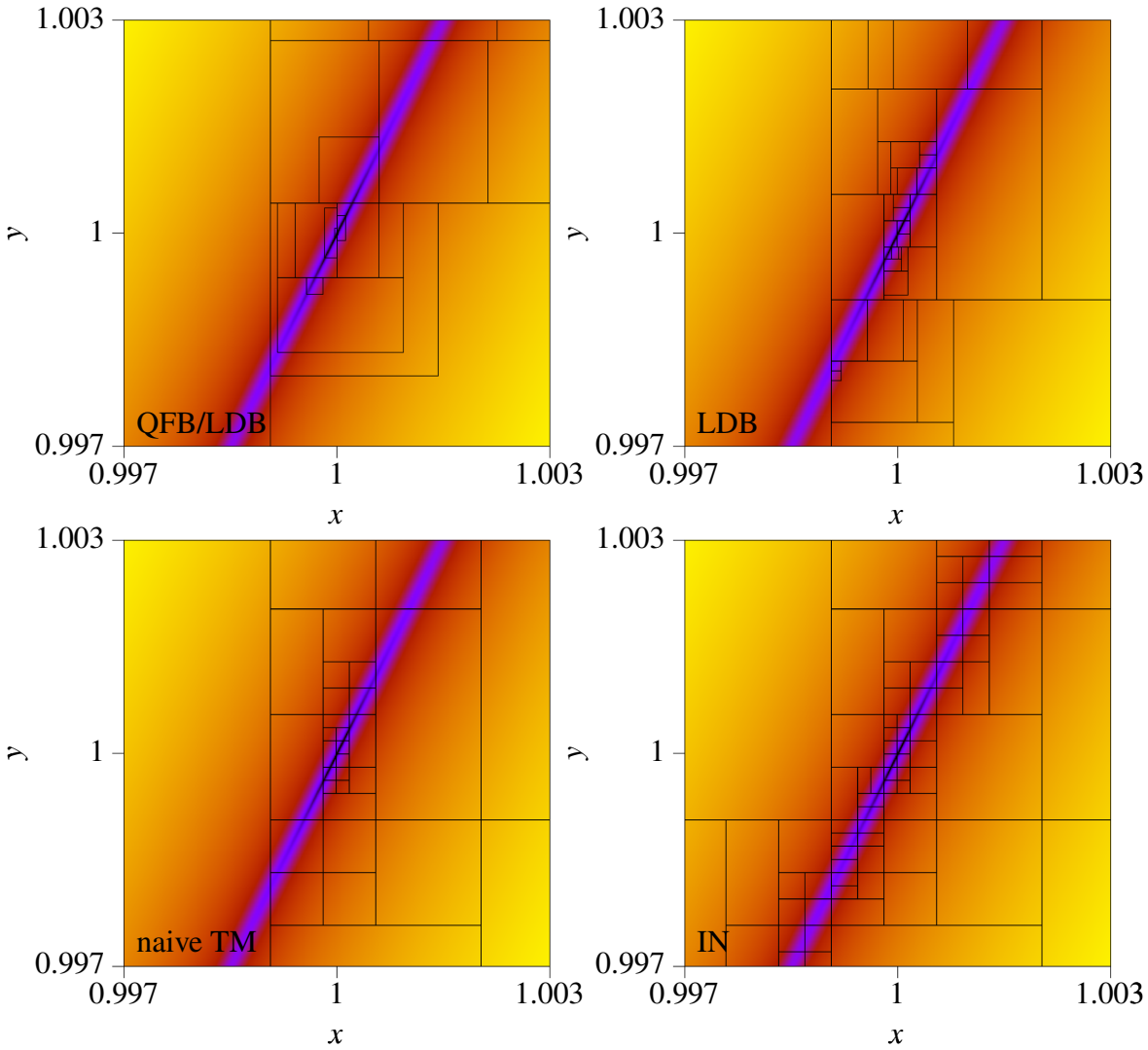

Figure 6.4: No cluster effect for the COSY-GO operating mode QFB/LDB, but a significant cluster effect for the IN evaluation.

optimization is always set to $[-1.5, 1.5]^n$. Accordingly, the search volume increases exponentially with the dimension of the objective function.

We require that boxes with a side length $s <$ 1E-6 are not split as a stopping condition of the algorithm. Ideally, the optimizer reduces the search volume by at least a factor of $3{,}000{,}000^n$. In the most advanced setting (QFB/LDB), which requires a minimum Taylor Model order of two, COSY-GO manages to reduce the search domain to a single box with a side length $s <$ 1E-6 for every dimension n that we tested.

Fig. 6.6 illustrates how the performances of COSY-GO in the evaluation of the generalized

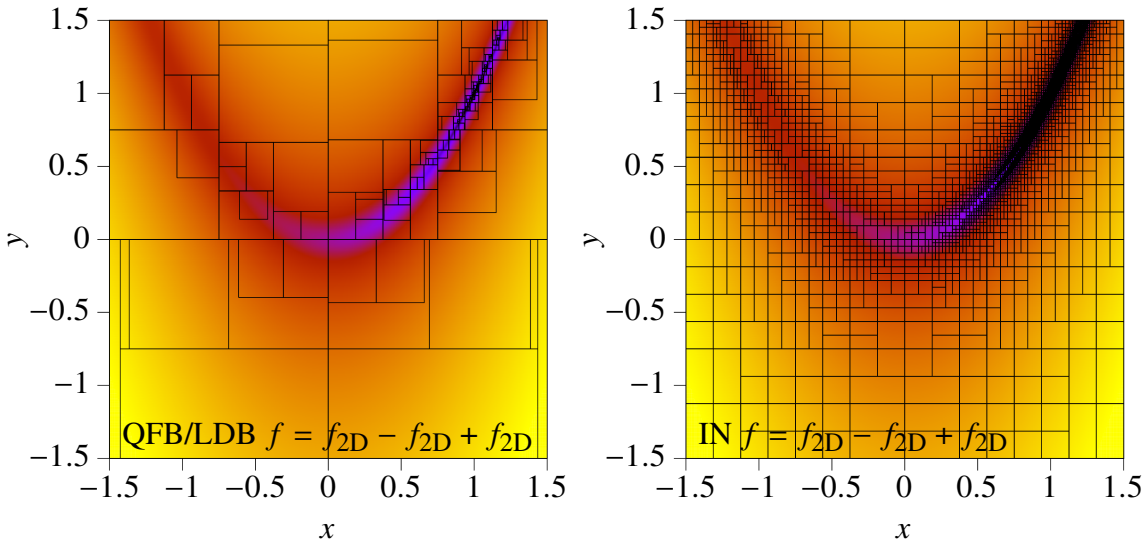

Figure 6.5: Splitting comparison between fourth order Taylor Model approach with QFB/LDB enabled and interval evaluation using the example of the modified 2D Rosenbrock function.

Rosenbrock function from Eq. (6.2) varies for different Taylor Model orders.

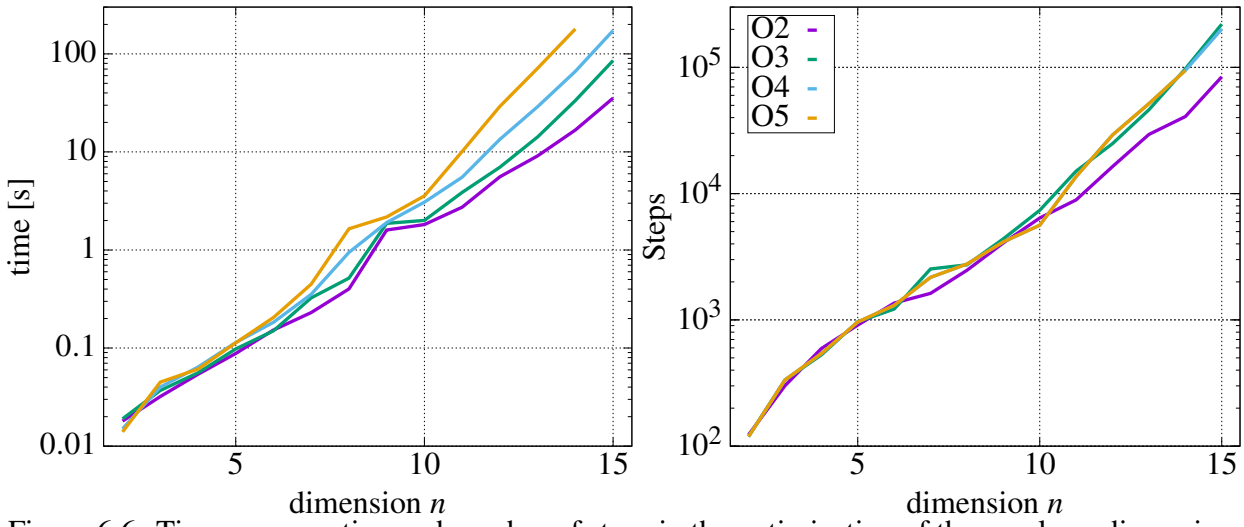

Figure 6.6: Time consumption and number of steps in the optimization of the regular n dimensional Rosenbrock function from Eq. (6.2) at various orders with COSY-GO and QFB/LDB enabled.

The second order calculation outperforms the higher order calculations in both aspects, regarding the speed and the number of required steps. This is rather unusual because even though the time per step increases with higher orders, the number of required steps usually shirks due to the tighter bounding of higher order calculations. However, in special cases like this one, the higher order

Taylor Models do not bound tighter than the lower order ones.

If we analyze the Rosenbrock function with the artificially increased dependency problem, the second order calculation behaves as one would expected, namely requiring more steps than its higher order counter parts (see Fig. 6.7).

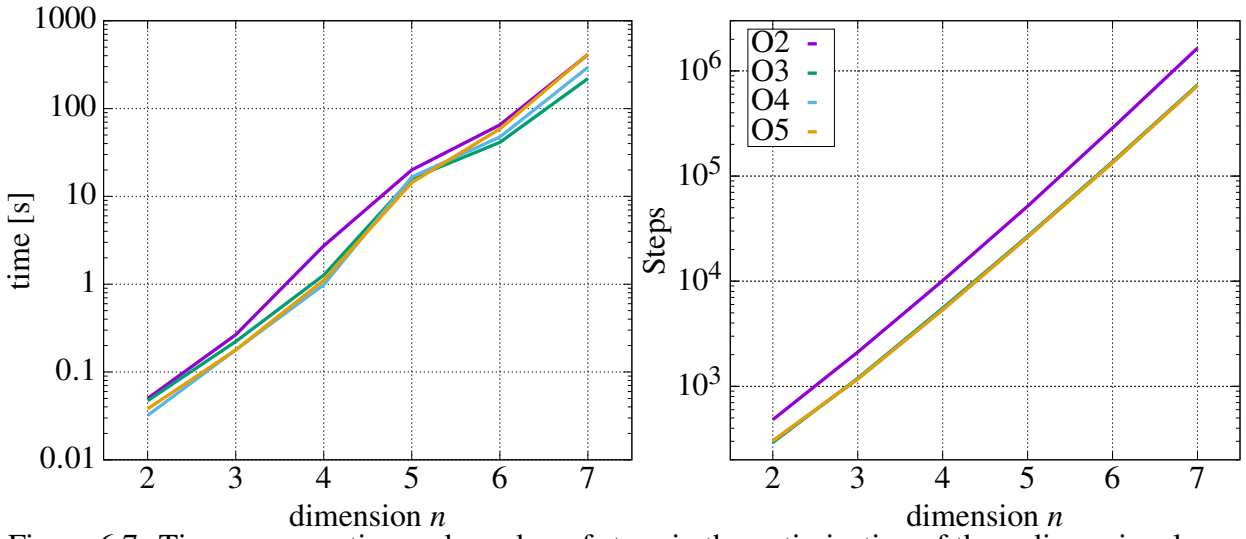

Figure 6.7: Time consumption and number of steps in the optimization of the n dimensional Rosenbrock function with an additional artificial dependency problem $f = f_{n\mathrm{D}} - f_{n\mathrm{D}} + f_{n\mathrm{D}}$ at various orders with COSY-GO and QFB/LDB enabled.

For all calculations, the global minimum of the generalized Rosenbrock function could be bound to $[-1\mathrm{E}306, 2\mathrm{E}\text{-}27]$. The optimization variables of all calculations are contained in $[0.999999998, 1.000000002]^n$, which is a box of side length 4E-9 and hence almost three orders of magnitude smaller than the minimum split size. This is because QFB and LDB are not bound to splitting boxes in half, but they can decrease their size as far as their rigorous methods allow them to.

In summary, the example cases of the Rosenbrock functions illustrated that Taylor Model based global optimizers, and COSY-GO in particular, can handle high dimensional objective functions very efficiently. The QFB and LDB avoid the cluster effect, while the Taylor Model evaluation significantly decreases the dependency problem. For the $n = 15$ dimensional Rosenbrock function, a reduction of the search volume by a factor of more than 4E157 was accomplished in 84017 steps and less than 36 seconds (see Fig. 6.6) on an Intel®Core$^{\mathrm{TM}}$ i5-7200U CPU 2.5GHz.

6.2 The Lennard-Jones Potential Problem

In this section, the capabilities of a Taylor Model based verified global optimizer (see Sec. 2.6) are demonstrated on the example of finding minimum energy configurations of particles when the well-known Lennard-Jones potential models their pairwise interactions.

This problem is particularly interesting and challenging for global optimization because the objective function is non-convex, highly nonlinear, and potentially high dimensional depending on the number of particles k considered. Accordingly, the dimensionality and hence the complexity of the optimization problem can be increased as desired by simply increasing the number of particles.

A further prominent aspect of the system is the enormous dependency problem that comes from the fact that every particle interacts with every other particle – changing the position of a single particle of a k-particle configuration changes $k - 1$ interactions and their contributions to the objective function. Furthermore, the function values become exceedingly large when two particles get too close to each other, while the actual resulting local minima are often very shallow, a situation that is reminiscent of the Rosenbrock function and its shallow valley with rapidly rising function values outside the valley.

The complexity of the objective function makes not only the optimization process itself challenging, but also finding appropriate variables and a rigorous initial search domain that is guaranteed to contain all global solutions. For a fully rigorous global optimization, the infinite search space must be analyzed unless one can prove that certain regions can be excluded because they cannot contain the minimum. Accordingly, mathematical arguments are required to reduce the infinite search space to a finite search domain box for the verified global optimizer. We will present arguments and methods that define a rigorous but sufficiently tight and finite initial search domain box while being very transparent. Additionally, we will propose ideas about even more involved methods that might yield an even tighter initial search domain.

6.2.1 The Lennard-Jones Potential

The 12-6 Lennard-Jones potential

$$U_{\mathrm{LJ}}(r) = 4U_0 \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{6.6}$$

was proposed by Lennard-Jones in 1931 [44] as a specific version of the more general r^{-a}-r^{-b} type potentials he proposed in 1925 [43]. The potential is used as a simplified model to describe the interaction between two electrically neutral atoms or molecules with a distance $r > 0$ between them.

The r^{-12} term models the strong repulsion of particles at very small distances. The attraction for moderate distances, which quickly decreases with larger distances, is modeled by the r^{-6} term. The parameter U_0 scales the depth of the potential well, which is related to the strength of the interaction between the two particles. The Van-der-Waals radius σ is also referred to as the particle size and indicates where the sign of the potential changes. It represents the distance at which the potential assumes the same value as for the configuration where the two particles are infinitely far away from each other.

The potential assumes its single minimum at the equilibrium distance of $r^\star = \sqrt[6]{2}\sigma$. For distances smaller than the equilibrium distance, the potential is strictly monotonically decreasing, and for distances larger than the equilibrium distance, the potential strictly monotonically increasing.

The values σ and U_0 depend on the particles involved in the modeled pairwise interaction. For our analysis, we will only consider one sort of particle corresponding to only one set of values for σ and U_0. To simplify the potential, we consider distances r and σ in units of the equilibrium distance $\sqrt[6]{2}\sigma$, and energy in units of U_0. Additionally, we are will offset the Lennard-Jones potential by one so that its single minimum $U^\star_{\mathrm{LJ}}$ has an energy of zero at the equilibrium distance $r^\star$ equal one. As a result, we are defining the pairwise interaction energy of two identical particles with a distance $r > 0$ between them as

$$U_{\mathrm{LJ}}(r) = 1 + r^{-12} - 2r^{-6}. \tag{6.7}$$

In Fig. 6.8, the single pairwise interaction potential between two identical particles from Eq. (6.7) is visualized. Note the shallowness of the potential and the large range of function values.

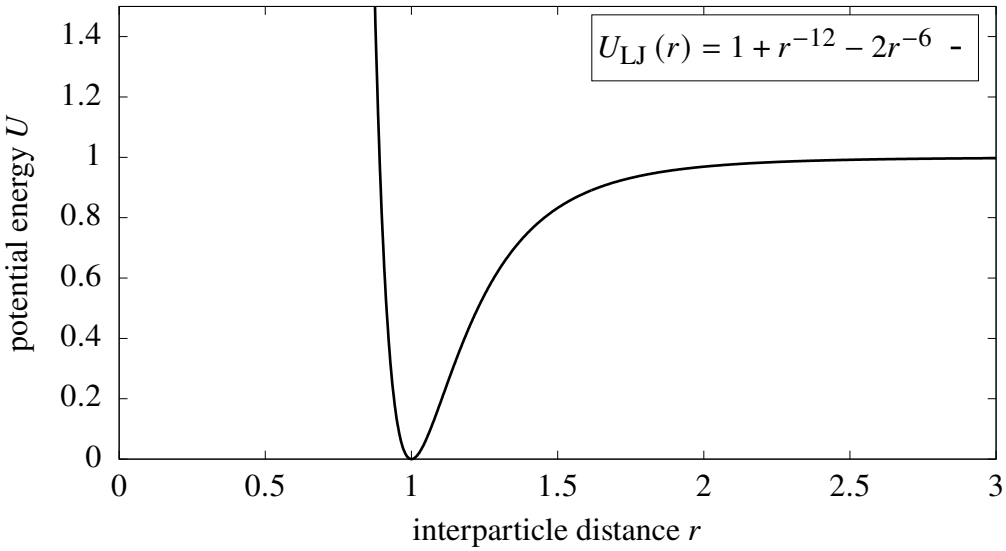

Figure 6.8: The Lennard-Jones potential for a pairwise interaction between two particles. For distances larger than the equilibrium distance $r^\star$ equal one, the potential quickly approaches its asymptotic value of one.

6.2.2 Configurations of Particles

Consider a configuration $\mathcal{S}_k$ of k identical particles that have their pairwise interaction modeled by the Lennard-Jones potential from Eq. (6.7). The overall interaction potential U_k of that configuration is given by

$$U_k = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{\mathrm{LJ}}\left(r_{ij}\right), \tag{6.8}$$

the sum of all pairwise interaction potentials U_{LJ}, where $r_{ij} = r_{ji}$ is the distance between the particles p_i and p_j.

The number of pairwise interactions $n_{\mathrm{pairs}} = \frac{k(k-1)}{2}$ increases with the square of the number of particles. The global minimum of the overall interaction potential $U_k^\star$ corresponds to a lowest energy state of the configuration. Those minimum energy states are of practical importance for the formation of molecules, assuming nature is sufficiently described by this model and finds the lowest energy when assembling molecules instead of just a local minimum.

Additionally, it is interesting to analyze the lowest energy states under an external constraint that limits the spatial dimensionality n_{dim} of the configuration. To indicate this constraint, we denote the

overall interaction potential with $U_{k,n_{\text{dim}}}$ and its global minimum with $U^{\star}_{k,n_{\text{dim}}}$. The corresponding configurations are denoted by $\mathcal{S}_{k,n_{\text{dim}}}$ and $\mathcal{S}^{\star}_{k,n_{\text{dim}}}$, respectively.

6.2.3 The Lennard-Jones Optimization Problem and its Challanges

The goal of the Lennard-Jones optimization problem is the following: Given k identical particles with their pairwise interaction modeled by the Lennard-Jones potential from Eq. (6.7) and the dimension of the configuration space n_{dim} (one, two, or three spatial dimensions), find the global minimum of the overall interaction energy (Eq. (6.8)) and the corresponding optimal configurations in the n_{dim} configuration space.

This optimization problem is trivial for two particles ($k = 2$) because its only a single Lennard-Jones interaction for which the minimum is known and discussed in Sec. 6.2.1. Furthermore, the optimization problem is also trivial when $k \leq n_{\text{dim}} + 1$, since obvious configurations exist where every single pairwise interaction potential of the n_{pairs} pairwise Lennard-Jones interactions is at its minimum $U^{\star}_{\text{LJ}} = 0$. In other words, all distances between all particles are optimal with $r^{\star}_{ij} = 1$. In particular, the configuration $S^{\star}_{k,n_{\text{dim}} \geq k-1}$ is an equilateral triangle for $k = 3$ and a tetrahedron for $k = 4$ with $U^{\star}_{k,n_{\text{dim}} \geq k-1} = 0$. However, the complexity of this optimization problem increases rapidly with the number of particles k due to the strong dependency problem of the $\sim k^2$ pairwise interactions.

To find the minimum energy configurations for such nontrivial cases using a verified global optimizer, it is critical to gain an understanding of the solution space. This understanding is needed to describe the potential minimum energy configurations with suitable optimization variables and limit the associated search domains to a finite solution space that tightly captures all the possible minimum energy configurations.

In the following sections, various conditions are presented to transparently exclude regions from the global search domain that cannot include minimum energy configurations. In Sec. 6.2.3.1 we develop a rigorous upper bound r_{UB} on the maximum inter-particle distance within the optimal configuration. This upper bound limits the maximum size of any possible minimum energy

configuration and hence limits the initial search space. Additionally, Sec. 6.2.3.1 provides an upper bound $r_{\vec{a},\mathrm{UB}}$ on the distance between projections of the particle positions onto an arbitrary axis $\vec{a}$. This property of minimum energy configurations is used for the definition of optimization variables and their associated domain in Sec. 6.2.3.7.

To help the global optimizer identify discardable boxes from the beginning, we present methods for calculating good initial upper bounds on the global minimum, i.e., the initial cutoff value $\mathcal{C}$. In Sec. 6.2.3.2, such upper bound configurations are discussed to calculate an initial upper bound on the global minimum $U_{k,\mathrm{UB}}$ for a configuration of k particles. The methods to characterize those upper bound configurations are based on the optimal configurations of $k-1$ particles, which we assume to know from a previous optimization run. In such a fashion, configurations are iteratively developed with increasing particle number, beginning with the obvious arrangements of $k+1$ particles in k dimensional search space. Some of the methods also use the upper bound on the inter-particle distance from Sec. 6.2.3.1 for the calculation.

In Sec. 6.2.3.3, a lower bound r_{LB} on the inter-particle distance in minimum energy configurations is determined using the upper bound on the minimum energy from Sec. 6.2.3.2. This lower bound is essential to formally exclude configurations from the search space for which the Lennard-Jones potential is not defined, namely, configurations for which at least one inter-particle distance is zero.

Sec. 6.2.3.4 describes a way to represent any minimum energy particle configuration in a coordinate system. However, rotated or mirrored versions of a configuration might have distinct coordinate representations, which yield multiple equivalent solutions. In Sec. 6.2.3.5, we list the different versions of equivalent coordinate representations that can occur. Sec. 6.2.3.6 discusses suppression mechanism to limit the representation of equivalent minimum energy configurations to ideally only one representative in the search space in order to reduce computational effort.

In Sec. 6.2.3.7, the optimization variables are defined based on the general coordinate system description of minimum energy configurations. Additionally, bounds are placed on those optimization variables based on the bounds on the inter-particle distances from Sec. 6.2.3.1 and Sec. 6.2.3.3.

6.2.3.1 The Rigorous Upper Bound on the Maximum Distance

Considering any configuration $\mathcal{S}_k$ of k particles and the x axis in an arbitrary orientation to it, we number the particles from 1 to k by their x coordinate from low to high. The numbering of particles with identical x coordinates is irrelevant for the further argumentation. The x distance between particle p_l and particle p_{l+1} is denoted by $v_{x,l}$, which by definition of the arrangement of the particles satisfies $v_{x,l} \geq 0$.

We are considering the distances $v_{x,l}$ as independent variables instead of the absolute x conditions of the particles. Accordingly, changing $v_{x,l}$ moves the entire subconfiguration composed of the particles p_j with $j \geq l + 1$ along the x axis, leaving all $v_{x,i\neq l}$ unchanged.

If $v_{x,l} > 1$ for any l, all inter-particle distances $r_{v_{x,l}}$ that are dependent on $v_{x,l}$ are at least of length $r_{v_{x,l}} \geq v_{x,l} > 1$. By setting $v_{x,l} = 1$, all distances involving $v_{x,l}$ are shortened to distances $r_{v_{x,l}} \geq 1$, monotonically improving the pairwise interaction potentials of every involved potential while leaving the uninvolved interaction energies unchanged, which overall monotonically improves the overall interaction potential (see Fig. 6.9).

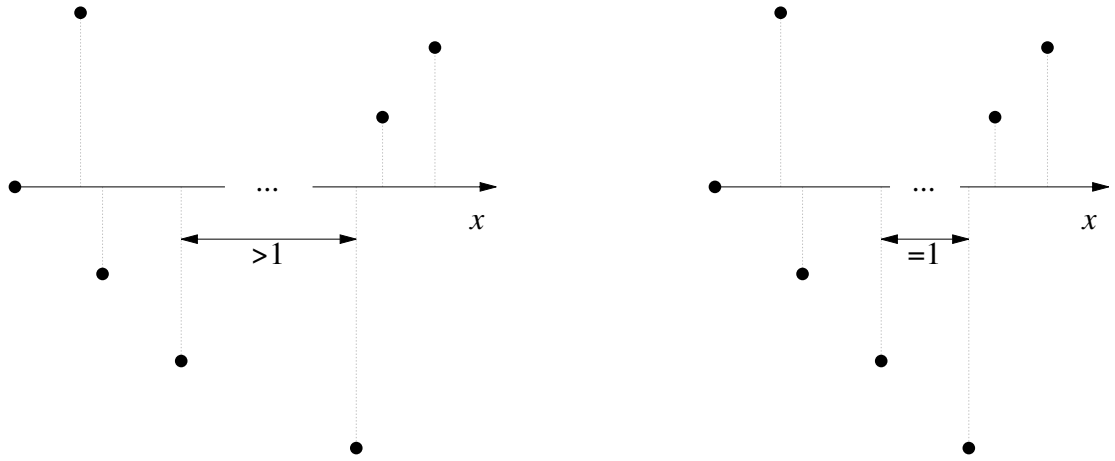

Figure 6.9: Monotonically improving the overall potential of a configuration for which the projected distance of two adjacent particles larger than one is.

Accordingly, the overall interaction potential of any configuration $\mathcal{S}_k$ of k particles can be improved if there is a $v_{x,l} > 1$ for any orientation of the x axis by setting $v_{x,l} = 1$. Thus, the optimal

configuration must satisfy $v_{x,l} \leq 1 = r_{x,\mathrm{UB}}$ with regard to any orientation of the x axis.

By placing the x axis along the maximum inter-particle distance $r_{\max}$ of all optimal configuration, we can conclude that $r_{\max} = \sum_l v_{x,l} \leq \sum_l 1 = k - 1 = r_{\mathrm{UB}}$.

This upper bound on the maximum inter-particle distance of the configuration holds for all n_{dim} but is only a tight upper bound for $n_{\mathrm{dim}} = 1$. For higher dimensional configurations, finding such an upper bound on the maximum inter-particle distance is a lot less trivial. However, using the maximum inter-particle distance of all minimum energy configuration of k particles in 2D can serve as an upper bound on the maximum inter-particle distance of the minimum energy configuration of k particles in 3D.

6.2.3.2 The Rigorous Upper Bound on the Minimum Energy

Any configuration $\mathcal{S}_{k,n_{\mathrm{dim}}}$ can serve as an upper bound configuration $\mathcal{S}_{k,n_{\mathrm{dim}},\mathrm{UB}}$, with the corresponding potential $U_{k,n_{\mathrm{dim}}} = U_{k,n_{\mathrm{dim}},\mathrm{UB}} \geq U^{\star}_{k,n_{\mathrm{dim}}}$ providing an upper bound on the minimum energy of a k particle configuration. The upper bound is used as an initial cutoff value $\mathcal{C}$ for the optimizer. The following approaches use the optimal configuration of $k-1$ particles to put tight upper bounds on $U^{\star}_{k,n_{\mathrm{dim}}}$. The first approach is specific for 1D configuration, while the second approach is suitable also for higher dimensional configurations.

1. Given a minimum energy configuration $\mathcal{S}^{\star}_{k-1,1\mathrm{D}}$ of $k-1$ particles, mirroring plane the first half of the configuration onto the second half by placing the mirror either on particle $p_{(k+1)/2}$ when k is odd, or in the middle between particle $p_{k/2}$ and $p_{k/2+1}$ when k is even. This upper bound configuration $\mathcal{S}_{k,1\mathrm{D},\mathrm{UB}}$ of k particles will be symmetric, satisfying $r_{i,i+1} = r_{k-i,k+1-i}$.

2. Consider the minimum energy configuration of $k-1$ particles fixed in place in a coordinate system. We now add a kth particle in a small and simple global optimization on its own, where only the coordinates of the kth particle are the optimization variables. From Sec. 6.2.3.1, we know that when we place an axis in any orientation in the minimum energy configuration, the

distances between projections onto that axis are less or equal to one. Accordingly, the search domain for the optimization of the position of the kth particle is determined by the maximum and minimum coordinates of the configuration of $k-1$ particles along each axis, plus a band of width one around it (see Fig. 6.10).

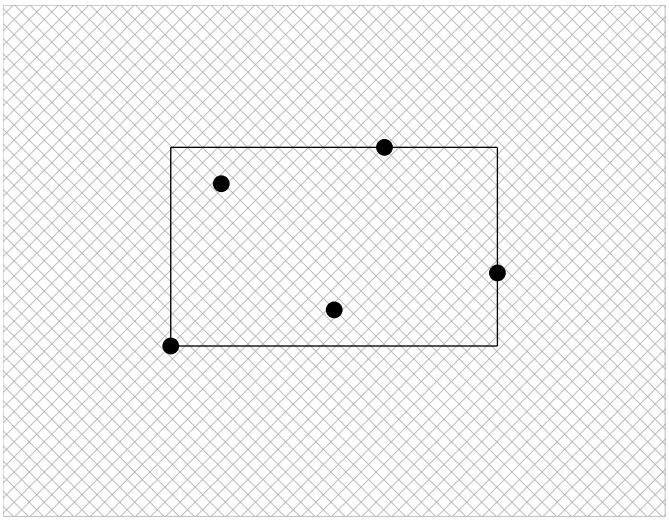

Figure 6.10: Search domain for the optimization of placing the sixth particle optimally relative to the fixed optimal configuration of five particles in 2D.

The last approach is particularly powerful for 2D and 3D configurations. It mimics the process of adding a particle to a given minimum energy configuration similarly to how molecules are sometimes formed one new element at a time. The first method is very effective for 1D configurations.

6.2.3.3 The Rigorous Lower Bound on the Minimum Distance

Given the optimal k-particle configuration $\mathcal{S}_k^\star$, we denote the pairwise interaction with the largest (worst) contribution to $U_k^\star$ by $U_{\alpha\beta} = U_{\mathrm{LJ}}\left(r_{\alpha\beta}\right)$, where α and β are the two particles involved in the interaction. We define the interaction energy of all particles with particle α as $U_\alpha = \sum_{i\neq\alpha} U_{\mathrm{LJ}}\left(r_{i\alpha}\right)$. For any possible $(k-1)$-particle subconfiguration $\mathcal{S}_{k-1}^{\subset\mathcal{S}_k^\star}$ of $\mathcal{S}_k^\star$, the interaction energy $U\left(\mathcal{S}_{k-1}^{\subset\mathcal{S}_k^\star}\right)$ will never be better than $U\left(\mathcal{S}_{k-1}^\star\right) = U_{k-1}^\star$, since by definition $\mathcal{S}_{k-1}^\star$ is the optimal configuration of

all $(k-1)$-particle configurations. Considering the potential of the $(k-1)$-particle subconfiguration of $\mathcal{S}_k^\star$ that excludes the particle α yields

$$U_{k-1}^\star \leq U_k^\star - U_\alpha \leq U_{k,\mathrm{UB}} - U_\alpha \leq U_{k,\mathrm{UB}} - U_{\alpha\beta}. \tag{6.9}$$

The largest (worst) contribution $U_{\alpha\beta}$ is overestimated by $U_{k,\mathrm{UB}} - U_{k-1}^\star \geq U_{\alpha\beta}$. Using the quadratic equation hidden in the Lennard-Jones potential from Eq. (6.7) yields the inverse

$$U_{\mathrm{LJ}}^{-1} = \begin{cases} r_{\min} = \left(1 + \sqrt{U}\right)^{-\frac{1}{6}} & \text{for } 0 \leq U \\ r_{\max} = \left(1 - \sqrt{U}\right)^{-\frac{1}{6}} & \text{for } 0 \leq U \leq 1 \end{cases}, \tag{6.10}$$

where $r_{\min}$ is strictly monotonically decreasing with increasing U, and $r_{\max}$ is strictly monotonically increasing with increasing U. Accordingly,

$$r_{\mathrm{LB}} = r_{\min}\left(U_{k,\mathrm{UB}} - U_{k-1}^\star\right) \leq r_{\min}\left(U_{\alpha\beta}\right) \tag{6.11}$$

provides a lower bound on the minimum inter-particle distance.

In practice, the exact value of the minimum energy $U_{k-1}^\star$ is most likely unknown, and we only have bounds on the minimum energy from the global optimization of the configuration of $k-1$ particles. For a rigorous calculation of the lower bound on the minimum inter-particle distance, a lower bound on $U_{k-1}^\star$ has to be used in Eq. (6.11) to remain verified.

This approach could be further improved by reducing the overestimation of $U_{\alpha\beta}$, e.g., by considering the minimum size of the potential contribution of all the other interactions $U_{\neq\alpha\beta}$ involved in U_α.

6.2.3.4 The Coordinate System

We will use a right-handed coordinate system to define the positions of the particles in the minimum energy configurations of maximal size. The definition of the placement of the coordinate system must be general enough to capture all possible minimum energy configurations.

Before we describe the positioning of the coordinate system relative to the configuration, we define a special subgroup of particles of the configuration that we call ‘outer’ particles. Given

an axis $\vec{a}$ relative to the configuration, we consider the perpendicular projections of the particles positions onto that axis. All particles that have the smallest projected a value are called 'lower outer' particles of the configuration with respect to $\vec{a}$, and all particles with the largest projected a value are called 'higher outer' particles of the configuration with respect to $\vec{a}$. Lower outer particles are higher outer particles with respect to $-\vec{a}$ and vice versa.

We begin by picking two outer particles, denoted p_1 and p_k, out of the k particle configuration and place the x' axis through them. We choose the two particles such that p_1 is a lower outer particle with respect to the x' axis and p_k is a higher outer particle with respect to the x' axis. The origin of the coordinate system is placed at the position of particle p_1 such that p_k is at the position $\vec{p}_k = \left(x'_k \geq 0, 0, 0\right)$.

The other particles are numbered from 2 to $k-1$ according to their x' coordinate yielding $x'_i \leq x'_j$ for $i < j$. The numbering scheme might be ambiguous for certain configurations that have two or more particles with the same x' position and is addressed in Sec. 6.2.3.6.

Without loss of generality, the y' axis is oriented such that an arbitrary particle p_i with $1 < i < k$ lies in the $x'y'$ plane and has a non-negative y' coordinate. To avoid ambiguity, we chose $i = 2$. The orientation of the z' axis follows from the right hand rule.

Fig. 6.11 illustrates the coordinate system for a configuration of six particles in 2D.

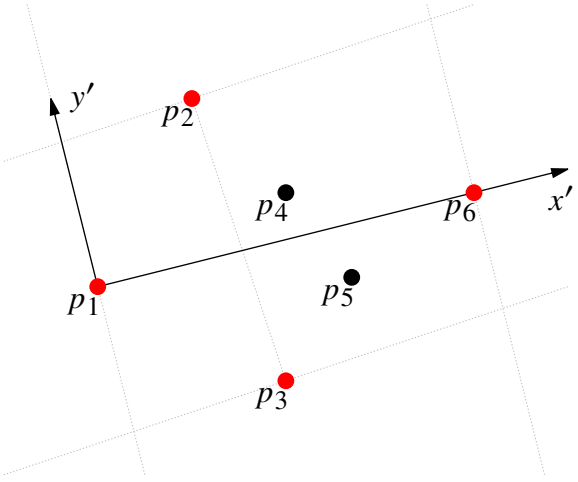

Figure 6.11: One possible placement of the coordinate system for a six particle configuration in 2D. The outer particles are shown in red together with their corresponding axis.

6.2.3.5 Equivalent Representations of Minimum Energy Configurations

Every minimum energy configuration can be characterized using the definition of the coordinate system introduced above. However, rotated and mirrored versions of the same configuration might have multiple distinct coordinate representations. For the optimization problem, those rotated and mirrored versions of a configuration are equivalent, because the objective function only depends on the inter-particle distances and not the position of the particles or the orientation of the configuration. As a consequence, the optimization algorithm will chase down every single one of those equivalent configurations and their representations. Ideally, we want to limit the search space such that it only includes one representative of a group of equivalent configurations.

Before we discuss mechanisms that exclude such redundant representations of equivalent configurations from the search domain in Sec. 6.2.3.6, we have to clarify what kind of equivalent configurations arise in our current coordinate representation.

First of all, there is multiple coordinate representations of a configuration that allows for multiple definitions of the x' axis defined by p_1 and p_k, i.e., any configuration with more than two outer particles. In other words, these configurations are equivalent by rotation of the configuration relative to the coordinate system. The six particle configuration in 2D shown in Fig. 6.11 has two outer particle pairs, which allows for two definitions of p_1 and p_k for each pair (see Fig. 6.12).

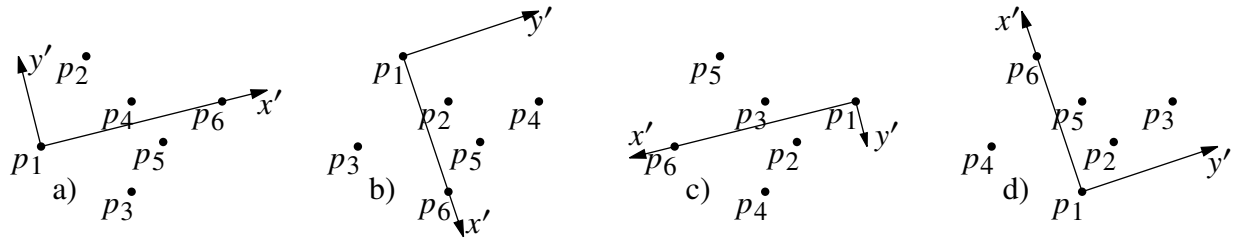

Figure 6.12: All possible placement of the coordinate system for a six particle configuration in 2D for different choices of p_1 and p_k.

A second aspect is the potential ambiguity in the numbering scheme when multiple particles have the same x' coordinate. Each numbering scheme of a configuration yields an additional coordinate representation.

Last but not least, each of those multiple representations also exist for configurations that are equivalent by mirroring. There are three mirror planes to consider, the $x'y'$ plane, the $x'z'$ plane and the $y'z'$ plane placed at $x' = x'_k/2$. The definition of the y' axis such that p_2 lies in the $x'y'$ plane and has a non-negative y' coordinate avoids all mirror configurations with respect to the $x'z'$ plane if $y_2 \neq 0$ and $\vec{p}_{i\notin\{1,2,k\}} \neq (x'_2, -y'_2, z'_2)$. But, equivalent configurations which are related though mirroring by the $y'z'$ plane placed at $x' = x'_k/2$ or the $x'y'$ plane are not avoided.

6.2.3.6 Suppression Schemes of Equivalent Configurations

To restrict the number of equivalent representation of the same configuration due to different choices of p_1 and p_k, one can limit the search domain to representations where the distance r_{1k} is greater or equal to any other distance r_{ij} of the configuration. We call $\overrightarrow{p_1 p_k}$ the major axis of the configuration. With this requirement, the coordinate representations shown in Fig. 6.12b) and Fig. 6.12d) are excluded from the search space.

To avoid mirror configurations with regard to mirror planes defined by the major axis, namely, the $x'y'$ plane, the $x'z'$ plane and the $y'z'$ plane placed at $x' = x'_k/2$, we developed two approaches.

The first approach **(1)** extends the $y_2 \geq 0$ requirement. We only consider configurations with $z'_j \geq 0$ where $j \notin \{1, 2, k\}$, without loss of generality. Because p_1, p_2, and p_k are by definition within the $x'y'$ plane, there is a mirror configuration with $z'_j > 0$ for every of the excluded configuration that has $z'_j < 0$. To avoid ambiguity, we choose $j = 3$. To limiting the search domain to only one configuration of mirror configurations with regard to the $y'z'$ plane placed at $x' = x'_k/2$ are using the property that p_1 and p_k are mirrored onto each other through that mirror plane. Accordingly, one can restrict the search domain to configurations for which $\sum_{i=2}^{k-1} U_{\mathrm{LJ}}(r_{1i}) \geq \sum_{i=2}^{k-1} U_{\mathrm{LJ}}(r_{ik})$ without loss of generality.

The second approach **(2)** uses the center of mass of the configuration [8]. Instead of requiring that $y'_i \geq 0$ and $z'_j \geq 0$, we require that the center of mass of the configuration

$$\vec{p}_{\mathrm{CM}} = \frac{1}{k} \sum_{i=1}^{k} \vec{p}_i \tag{6.12}$$

satisfies $y'_{\mathrm{CM}} \geq 0$, $z'_{\mathrm{CM}} \geq 0$, and $x'_{\mathrm{CM}} \geq x'_k/2$. In other words, without loss of generality, we only consider coordinate representations of configurations for which the center of mass of the configuration is within the specified octant out of the eight octants formed by the mirror planes.

The center of mass approach is more appealing because it captures the essence of those mirror configurations in a single concept, while the first approach is formulated in terms of three seemingly unrelated conditions. However, it is not clear which of the approaches performs better.

Avoiding an ambiguous numbering scheme is by far the most difficult and least precise because it requires knowledge about the structure of the minimum energy configuration, which we are trying to determine in the first place using the optimization. A reasonable assumption is that the minimum energy configurations tends to comprise some sort of symmetry between particles with regard to the major axis $\overrightarrow{p_1 p_k}$. Below we define a new coordinate system with respect to the current one that addresses the ambiguity in the numbering scheme of those configuration. The goal is to tilt the major axis with respect to the x axis of the new coordinate system such that symmetries in the optimal configurations with regard to the major axis will not exist with respect to the x axis. Accordingly, the numbering of the particles in the configuration determined by the x axis instead of the x' axis.

However, depending on the number of particles and the angle of the tilt, another ambiguity in the numbering scheme might possibly be introduced. In a way, this approach just makes it less likely to have an ambiguous numbering scheme for an optimal configuration under the assumption that optimal configurations comprise some symmetries with regard to the major axis.

The new coordinate system is defined as follows. The origin is at p_1. The tilt of the major axis with respect to the x axis is implemented by placing the new coordinate system such that p_k is at $\vec{p}_k = \left(x_k \geq 0, \epsilon_y \geq 0, \epsilon_z \geq 0\right)$, where ϵ_y and ϵ_z are small. The new y axis is within the $x'y'$ plane and the new z axis within the $x'z'$ plane. In general, the new coordinate system satisfies that for $\epsilon_y = 0$ and $\epsilon_z = 0$, the old and the new coordinate system are identical.

The $x'z'$ plane in the new coordinates is defined by a normal vector in the y' direction with

$$\vec{n}_{y'} = \vec{e}_z \times \overrightarrow{p_1 p_k} = \vec{e}_z \times \vec{p}_k = -\epsilon_y \vec{e}_x + x_k \vec{e}_y, \tag{6.13}$$

which points in the y direction for $\epsilon_y = 0$.

The $x'y'$ plane in the new coordinates is defined by a normal vector in the z' direction with

$$\vec{n}_{z'} = \overrightarrow{p_1 p_k} \times \vec{e}_y = \vec{p}_k \times \vec{e}_y = -\epsilon_z \vec{e}_x + x_k \vec{e}_z, \tag{6.14}$$

which points in the z direction for $\epsilon_z = 0$.

Accordingly, the particle p_2 (of the new numbering scheme) must satisfy the Hesse normal form

$$0 = \vec{n} \cdot \vec{p}_2 = -\epsilon_z x_2 + x_k z_2 \quad \Rightarrow \quad z_2 = \epsilon_z \frac{x_2}{x_k}, \tag{6.15}$$

to be in the $x'y'$ plane.

The definition of having a coordinate $y'_i \geq 0$, where $i \in \{2, \mathrm{CM}\}$, is equivalent to requiring that the dot product of the corresponding position $\vec{p}_i$ and the normal vector $\vec{n}_{y'}$ is greater or equal to zero. Accordingly,

$$0 \leq \vec{p}_i \cdot \vec{n}_{y'} = -\epsilon_y x_i + x_k y_i \quad \Rightarrow \quad y_i \geq \epsilon_y \frac{x_i}{x_k}. \tag{6.16}$$

As we would expect, this requirement breaks down to $y_i \geq 0$ for ϵ_y equal zero.

The definition of having a coordinate $z'_j \geq 0$, where $j \in \{3, \mathrm{CM}\}$, is equivalent to requiring that the dot product of the corresponding position $\vec{p}_j$ and the normal vector $\vec{n}_{z'}$ is greater or equal to zero. Accordingly,

$$0 \leq \vec{p}_j \cdot \vec{n}_{z'} = -\epsilon_z x_j + x_k z_j \quad \Rightarrow \quad z_j \geq \epsilon_z \frac{x_j}{x_k}. \tag{6.17}$$

As we would expect, this requirement breaks down to $z_j \geq 0$ for ϵ_z equal zero.

For the requirement of having $x'_{\mathrm{CM}} \geq x'_k/2$ is satisfied in the new coordinates if

$$0 \leq \left(\vec{p}_{\mathrm{CM}} - \frac{\vec{p}_k}{2} \right) \cdot \vec{p}_k. \tag{6.18}$$

The requirement $\sum_{i=2}^{k-1} U_{\mathrm{LJ}}(r_{1i}) \geq \sum_{i=2}^{k-1} U_{\mathrm{LJ}}(r_{ik})$ is independent of the coordinate system and hence does not have to be adjusted.

Unfortunately, the tilt to avoid ambiguous numbering of symmetry configurations comes at a cost. For some configurations, the tilt of the major axis will yield x coordinates of particles p_i with $x_i < 0$ and/or $x_i > x_k$, which breaks the useful ($x_i \leq x_j$ for $i < j$)-relation for p_1 and/or p_k (see Fig. 6.13). However, those particles p_i that potentially break the relation have to be very close to

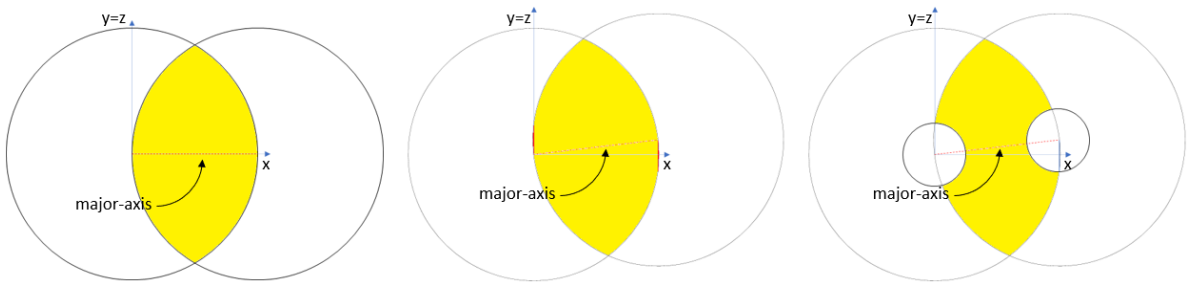

Figure 6.13: Given $\epsilon_y = \epsilon_z$, we consider the projection into the plane spanned by the x axis and the y axis. The red dotted line illustrates the major axis. The upper bound on the maximum inter-particle distance r_{UB} is the distance between the center of the two circles and their radius. Hence all particles of the configuration must lie both in the left and right circle simultaneously (the yellow area). In the left picture, the solution space for a particle contains only x coordinates between the two major axis particles. By tilting the major axis relative to the x axis, some areas of the solution space for the particle now have x coordinates outside the range defined by the two major axis particles (red), as shown in the middle picture. The right picture shows how the lower bound on the minimum inter-particle distance r_{LB} eliminates those critical (red) areas from the solution space, leaving a solution space that is again only associated with x coordinates between the two major axis particles (yellow).

either p_1 or p_k for this to happen. Specifically, only particles with a distance $r < 2\sqrt{\epsilon_y^2 + \epsilon_z^2}$ to p_1 and/or p_k are affected.

Given a lower bound r_{LB} on the minimum particle distance from Sec. 6.2.3.3, ϵ_y and ϵ_z can be chosen such that all those problematic configurations are excluded from the solution space. This way $x_i \leq x_j$ for $i < j$ holds true for all i, j for the remaining configurations in the solution space.

In summary, any configuration in the new coordinate system is defined as follows

I The major axis of the configuration is determined by $\overrightarrow{p_1 p_k}$, where p_1 is at the origin and p_k is at $(x_k > 0, \epsilon_y, \epsilon_z)$ with any choice of ϵ_y and ϵ_z that satisfies $\sqrt{\epsilon_y^2 + \epsilon_z^2} \leq \frac{r_{\mathrm{LB}}}{2}$.

II The x coordinates of any two particles p_i and p_j satisfy $x_i \leq x_j$ for $i < j$.

III p_2 is at $\left(x_2, y_2, \frac{\epsilon_z x_2}{x_k}\right)$

To limit the number of equivalent representations of configurations and their rotated and mirrored versions we require

$$r_{1k} \geq r_{ij} \quad \forall (i,j) \neq (1,k) \tag{6.19}$$

to have the maximum inter-particle distance of the configuration between particle p_1 and p_k. Additionally, we enforce either approach **(1)** with

$$\sum_{i=2}^{k-1} U_{\mathrm{LJ}}\left(r_{1i}\right) \geq \sum_{i=2}^{k-1} U_{\mathrm{LJ}}\left(r_{ik}\right), \tag{6.20}$$

$$y_2 \geq \epsilon_y \frac{x_2}{x_k} \quad \text{and} \tag{6.21}$$

$$z_3 \geq \epsilon_z \frac{x_2}{x_k}, \tag{6.22}$$

or approach **(2)** with

$$\left(\vec{p}_{\mathrm{CM}} - \frac{\vec{p}_k}{2}\right) \cdot \vec{p}_k \geq 0, \tag{6.23}$$

$$y_{\mathrm{CM}} \geq \epsilon_y \frac{x_{\mathrm{CM}}}{x_k} \quad \text{and} \tag{6.24}$$

$$z_{\mathrm{CM}} \geq \epsilon_z \frac{x_{\mathrm{CM}}}{x_k} \tag{6.25}$$

to avoid mirror versions of a configuration and their representation.

Note that for $n_{\mathrm{dim}} = 2$, only the xy-plane of the coordinate system is relevant and all z related variables and parameters are zero. For $n_{\mathrm{dim}} = 1$, the old and new coordinate system are identical because all z and y related variables and parameters are zero.

6.2.3.7 Definition and Bounding of the Optimization Variables

To incorporate the ordering of the particles of **II** in the optimization variables, we define $k-1$ variables as the x distance between any two consecutive numbered particles with $v_{x,i} = x_{i+1} - x_i \geq 0$ for $i \leq k-1$. The variables in y and z correspond to the y and z positions of the particles with $v_{y,i} = y_i$ for $2 \leq i \leq k-1$ and $v_{z,i} = z_i$ for $3 \leq i \leq k-1$ such that the distance r_{ij} between any two

particles p_i and p_j can be reconstructed from the variables as follows:

$$r_{ij}^2 = \left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2 + \left(z_j - z_i\right)^2 \quad \text{where} \tag{6.26}$$

$$x_j - x_i = \sum_{n=i}^{j-1} v_{x,n} \quad \text{for} \quad j > i; \tag{6.27}$$

$$y_1 = 0,\ y_k = \epsilon_y, \quad \text{and} \quad y_i = v_{y,i} \quad \text{for} \quad 2 \leq i \leq k-1; \tag{6.28}$$

$$z_1 = 0,\ z_2 = \frac{\epsilon_z x_2}{x_k},\ z_k = \epsilon_z, \quad \text{and} \quad z_i = v_{z,i} \quad \text{for} \quad 3 \leq i \leq k-1. \tag{6.29}$$

Based on Sec. 6.2.3.4 and the bounds on the minimum and maximum inter-particle distances from Sec. 6.2.3.3 and Sec. 6.2.3.1, respectively, the initial search domain is rigorously defined below.

In Sec. 6.2.3.1, we showed that $v_{x,i} \in \left[0, r_{x,\mathrm{UB}} = 1\right]$. For 1D, this lower bound on the inter-particle distance can be directly applied to the variable domain yielding $v_{x,i} \in \left[r_{\mathrm{LB}}, r_{x,\mathrm{UB}} = 1\right]$. For $n_{\mathrm{dim}} > 1$ however, this incorporation of the lower bound on the inter-particle distance into the initial search domain is not possible. The associated problems with configurations that have for inter-particle distances of length zero, for which the Lennard-Jones potentail is not defined, are addressed in Sec. 6.2.4.

The variables $v_{y,i}$ and $v_{z,i}$ are only bound by the maximum inter-particle distance r_{UB}. Accordingly, the variables can be bound to $v_{\dagger,i} \in [-1, 1]\ \frac{\sqrt{3}}{2} r_{\mathrm{UB}} + \frac{\epsilon_\dagger}{2}$ with $\dagger \in \{y, z\}$. This corresponds to half the $\dagger$-offset of the last particle plus and minus the height of an equilateral triangle of side length r_{UB}.

Assuming we follow the suppression approach **(1)**, the requirement of $y_2 \geq \epsilon_y \frac{x_2}{x_k}$ and $z_3 \geq \epsilon_z \frac{x_2}{x_k}$ can not fully be represented by a fixed initial search domain, because the condition is changing depending on the variables x_2 and x_k. But, the search domain box for those variables can still be decreased to $v_{y,2} \in \left[0, \frac{\sqrt{3}}{2} r_{\mathrm{UB}} + \frac{\epsilon_y}{2}\right]$ and $v_{z,3} \in \left[0, \frac{\sqrt{3}}{2} r_{\mathrm{UB}} + \frac{\epsilon_z}{2}\right]$.

In summary, the initial search domain box $\mathbb{B}$ is composed of the domains of the variables, which

are defined as follows

$$v_{x,i} \in [r_{\mathrm{LB}}, 1] \quad \text{for 1D} \quad \text{and} \quad v_{x,i} \in [0, 1] \quad \text{for 2D and 3D} \tag{6.30}$$

$$v_{y,2} \in \left[0, \frac{\sqrt{3}}{2} r_{\mathrm{UB}} + \frac{\epsilon_y}{2}\right] \quad \text{and} \quad v_{z,3} \in \left[0, \frac{\sqrt{3}}{2} r_{\mathrm{UB}} + \frac{\epsilon_z}{2}\right] \tag{6.31}$$

$$v_{\dagger,i} \in [-1, 1] \frac{\sqrt{3}}{2} r_{\mathrm{UB}} + \frac{\epsilon_\dagger}{2} \quad \text{for} \quad i > 2 \quad \text{and} \quad \dagger \in \{y, z\}\,. \tag{6.32}$$

The parameters ϵ_y and ϵ_z must satisfy $\sqrt{\epsilon_y^2 + \epsilon_z^2} \leq \frac{r_{\mathrm{LB}}}{2}$.

In Fig. 6.14, the initial search domains of individual particles in 2D are visualized.

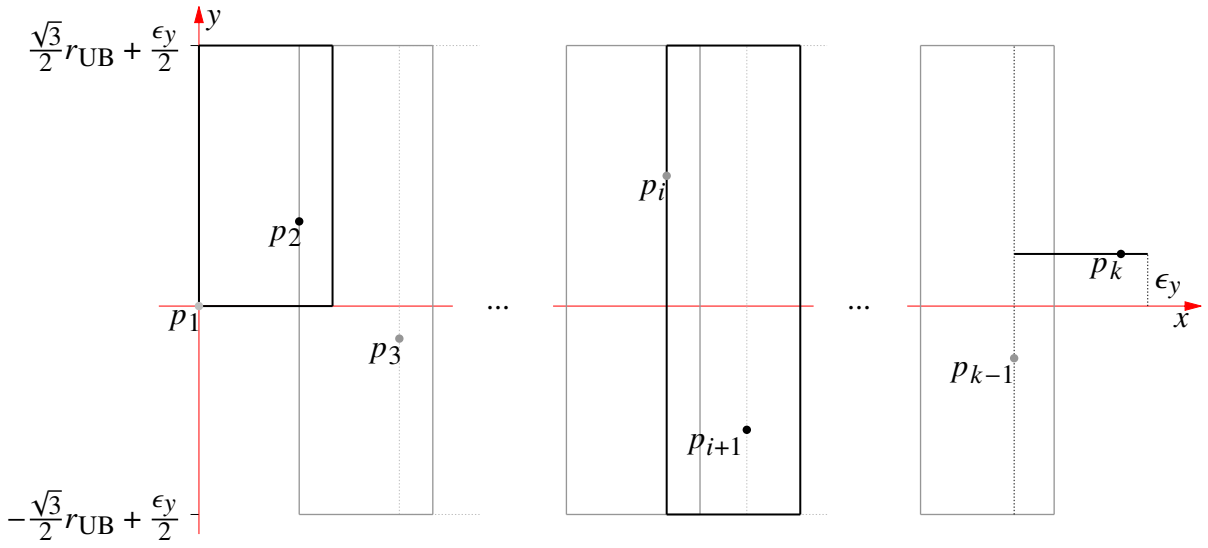

Figure 6.14: Initial search domain of global optimization problem for configuration of k particles in 2D. Note that the box width in x direction is always one and that the x position of particle p_i determines the starting position in x of the domain box of particle p_{i+1}. Particle p_1 is fixed to the origin. Particle p_k has a fixed y value of ϵ_y. Accordingly, its domain is just a line and not a box.

To enforce the suppression schemes from Sec. 6.2.3.6 (Eq. (6.19) to Eq. (6.25)), it is possible to devise penalty functions, or use methods of constrained optimization. Since all the requirements are of the form $a \geq b$, a general penalty function f_{pen} is defined, which is zero for $a \geq b$ and monotonically increasing with increasing $b - a$ for $a < b$.

Note that for $n_{\mathrm{dim}} > 1$, the domain can not exclude configurations with inter-particle distances $r_{ij} < r_{\mathrm{LB}}$. The Lennard-Jones interaction potential is already a penalty function for those

configurations, but the problem is that the Lennard-Jones potential is not defined for $r_{ij} = 0$. To address this we define a modified Lennard-Jones potential for those configurations in Sec. 6.2.4.

6.2.4 The Evaluation of the Objective Function

To be fully transparent, we elaborate on the evaluation of the objective function.

The objective function

$$U_k = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{\mathrm{LJ}}\left(r_{ij}\right), \tag{6.33}$$

is composed of $n_{\mathrm{pairs}} = \frac{k(k-1)}{2}$ individual Lennard-Jones interactions with

$$U_{\mathrm{LJ}}\left(r_{ij}\right) = 1 + r_{ij}^{-12} - 2r_{ij}^{-6}. \tag{6.34}$$

Since Eq. (6.26) yields only squared distances r_{ij}^2, we implement a Lennard-Jones potential that takes the squared distance $r^2 = r_{\mathrm{sqr}}$ as its argument with

$$U_{\mathrm{LJ,sqr}}\left(r_{\mathrm{sqr}}\right) = 1 + r_{\mathrm{sqr}}^{-3}\left(r_{\mathrm{sqr}}^{-3} - 2\right) \quad \text{where} \quad r_{\mathrm{sqr}}^{-3} = \frac{1}{r_{\mathrm{sqr}} \cdot r_{\mathrm{sqr}}^2}. \tag{6.35}$$

The squared distance r_{ij}^2 is evaluated using Eq. (6.26), where $x_j - x_i$, y_i, and z_i are calculated according to Eq. (6.27), Eq. (6.28), and Eq. (6.29), respectively.

For $n_{\mathrm{dim}} \geq 2$, the initial search domain includes configurations, for which the distance between particles is zero. Fig. 6.14 visualizes that any two particles p_i and p_{i+l} in the search domain box are at the same position if $v_{x,m} = 0$ for all $m \in [i, i + l]$ and if the vertical variables $v_{y,i}$ and $v_{y,i+l}$ are identical. Any domain box that contains such a configuration cannot be evaluated, because the Lennard-Jones potential is not defined for an argument of zero. Accordingly, the global optimizer can not process such boxes.

We were not able to exclude those configurations from the search domain because they form high dimensional manifolds within it. However, Sec. 6.2.3.3 showed that all configurations with a single inter-particle distance below r_{LB} cannot be a minimum energy configuration. In other words, any configuration with at least one inter-particle interaction energy of $U_{\mathrm{LJ}}(r_{\mathrm{LB}})$ or larger is not a minimum energy configuration.

This means that we can modify the objective function for inter-particle distances smaller than r_{LB} without changing the optimization problem if this modified Lennard-Jones potential $U^{\ddagger}_{\mathrm{LJ,sqr}}$ satisfies [8]

$$U_{\mathrm{LJ,sqr,mod}}\left(r_{\mathrm{sqr}}\right) \geq U_{\mathrm{LJ,sqr}}\left(r_{\mathrm{LB}}^2\right) \quad \forall \quad r_{\mathrm{sqr}} < r_{\mathrm{LB}}^2. \tag{6.36}$$

Accordingly, we define the modified Lennard-Jones potential piecewise and compose it of a the regular Lennard-Jones potential for $r_{\mathrm{sqr}} \geq r_{\mathrm{LB}}^2$ and the tangential extension at r_{LB}^2 for $r_{\mathrm{sqr}} < r_{\mathrm{LB}}^2$. The modified Lennard-Jones potential [8] is then given by

$$U^{\ddagger}_{\mathrm{LJ,sqr}}\left(r_{\mathrm{sqr}}, r_{\mathrm{LB}}\right) =) \begin{cases} U_{\mathrm{LJ,sqr}}\left(r_{\mathrm{sqr}}\right) & \text{for } r_{\mathrm{sqr}} \geq r_{\mathrm{LB}}^2 \\ U'_{\mathrm{LJ,sqr}}\left(r_{\mathrm{LB}}^2\right) \cdot \left(r_{\mathrm{sqr}} - r_{\mathrm{LB}}^2\right) + U_{\mathrm{LJ,sqr}}\left(r_{\mathrm{LB}}^2\right) & \text{for } r_{\mathrm{sqr}} < r_{\mathrm{LB}}^2 \end{cases} \tag{6.37}$$

where $U'_{\mathrm{LJ,sqr}}$ is the first derivative of $U_{\mathrm{LJ,sqr}}$ with

$$U'_{\mathrm{LJ,sqr}}\left(r_{\mathrm{sqr}}\right) = 6\left(r_{\mathrm{sqr}}^2\right)^2\left(1 - r_{\mathrm{sqr}} \cdot r_{\mathrm{sqr}}^2\right). \tag{6.38}$$

The modified Lennard-Jones potential is visualized in Fig. 6.15.

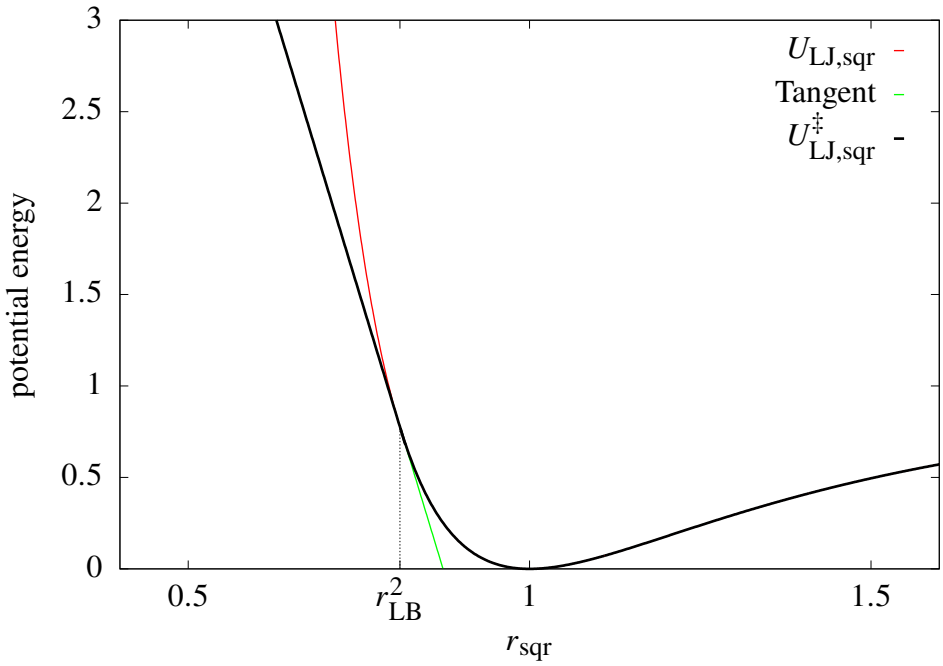

Figure 6.15: Piecewise defined modified Lennard-Jones potential.

6.2.5 Taylor Model Evaluation of Piecewise Defined Functions

Given the continues piece wise defined function

$$f(x) = \begin{cases} f_{\leq}(x) & \text{for} \quad x \leq x_0 \\ f_{\geq}(x) & \text{for} \quad x \geq x_0 \end{cases}, \tag{6.39}$$

where each of the pieces is m times differentiable in its domain, we want to find a Taylor Model that tightly captures $f(x)$ over the domain $\mathbb{D}_f = [a, b]$ with $x_0 \in \mathbb{D}_f$. In a first step, we calculate the Taylor Models for each of the subdomains $\mathbb{D}_{\leq} = [a, x_0]$ and $\mathbb{D}_{\geq} = [x_0, b]$. We denote those two Taylor Models by $f_{\leq,\mathrm{TM},[a,x_0]}$ and $f_{\geq,\mathrm{TM},[x_0,b]}$, respectively. Generally, any Taylor Model $f_{\mathrm{TM},[a,b]} = \left(\mathcal{P}_f, \epsilon_f\right)$ is a verified description of f over $\mathcal{D}_f$, if the Taylor Model $f_{\mathrm{TM},[a,b]}$ contains $f_{\leq,\mathrm{TM},[a,x_0]}$ over $\mathbb{D}_{\leq}$ and $f_{\geq,\mathrm{TM},[x_0,b]}$ over $D_{\geq}$ (see Fig. 6.16). In particular, for any $\mathcal{P}_f$ the

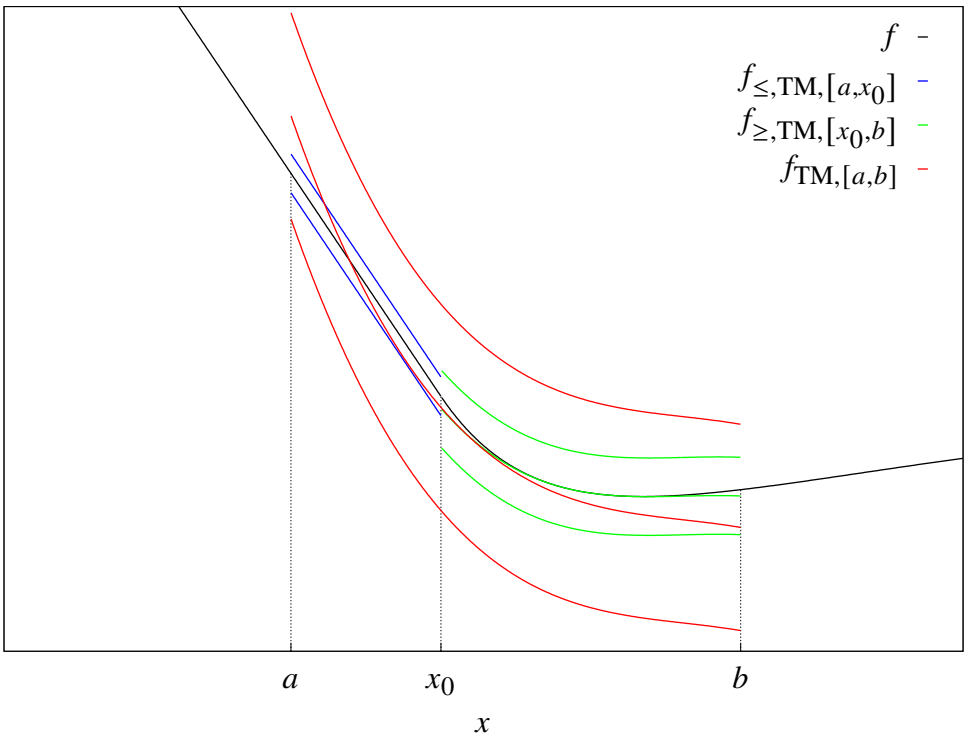

Figure 6.16: Taylor Model description of piecewise defined function.

remainder ϵ_f can be adjusted/increased such that this requirement is satisfied. The remainder

interval is determined by the union of $I_{\mathbb{D}_\leq} \cup I_{\mathbb{D}_\geq}$, where

$$I_{\mathbb{D}_\leq} = \left| \mathcal{P}_f - f_{\leq,\mathrm{TM},[a,x_0]} \right|_{\mathbb{D}_\leq} \quad \text{and} \tag{6.40}$$

$$I_{\mathbb{D}_\geq} = \left| \mathcal{P}_f - f_{\geq,\mathrm{TM},[x_0,b]} \right|_{\mathbb{D}_\geq} . \tag{6.41}$$

The notation $|g|_{\mathbb{D}}$ indicates the bounds of g over $\mathbb{D}$.

The challenge in determining $f_{\mathrm{TM},[a,b]}$ is finding a polynomial $\mathcal{P}_f$ for which the remainder interval is small. Good candidates for $\mathcal{P}_f$ can be generated by weighted compositions of the polynomial parts of the two Taylor Models over the two subdomains. However, there are some intricacies to consider regarding the implementation of this composition, which we will discuss below.

Without loss of generality, Taylor Models are always expanded around zero over the domain $[-1, 1]$. Accordingly, a linear transformation is required to map the Taylor Model domain $[-1, 1]$ to the domain of interest and vice versa. We denote the variable in the original function domain by $x \in \mathbb{D}_f$ and the local variable in the Taylor Model domain by $x' \in [-1, 1]$. The midpoint and width of a domain $\mathbb{D}$ are denoted by $\dagger(\mathbb{D})$ and $w(\mathbb{D})$, respectively, which yields the mapping relation

$$x = \dagger(\mathbb{D}) + \frac{w(\mathbb{D})}{2} x' \quad \text{for} \quad x \in \mathbb{D}, x' \in [-1, 1]. \tag{6.42}$$

With the Taylor Model of the identity denoted by $\mathcal{I}_{\mathrm{TM}} = (x', \epsilon)$, we calculate the Taylor Models for each of the subdomains with

$$f_{\leq,\mathrm{TM},[a,x_0]}(x') = f_\leq \left(\dagger(\mathbb{D}_\leq) + \frac{w(\mathbb{D}_\leq)}{2} \mathcal{I}_{\mathrm{TM}} \right) \quad \text{and} \tag{6.43}$$

$$f_{\geq,\mathrm{TM},[x_0,b]}(x') = f_\geq \left(\dagger(\mathbb{D}_\geq) + \frac{w(\mathbb{D}_\geq)}{2} \mathcal{I}_{\mathrm{TM}} \right). \tag{6.44}$$

Note that each of the Taylor Models is expanded around the midpoint of the respective domain and scaled according to the width of the domain. Thus, also the polynomial parts of the TMs are expanded around $\dagger(\mathbb{D}_\leq)$ and $\dagger(\mathbb{D}_\geq)$, respectively, and the variables are scaled by the receptive widths of the domains. For the composition of two polynomials, we need to make sure that the function domains of the two polynomials are identical, i.e., both expanded around the same point

with the same scaling. In other words, the two polynomials, denoted $\mathcal{P}_{\leq}(x')$ and $\mathcal{P}_{\geq}(x')$, are expanded around $\dagger\left(\mathbb{D}_f\right)$ and scaled by $w\left(\mathbb{D}_f\right)$, specifically

$$x' = 2\frac{x - \dagger\left(\mathbb{D}_f\right)}{w\left(\mathbb{D}_f\right)}. \tag{6.45}$$

Since all steps are equivalent for the left and the right side, we will calculate $\mathcal{P}_{\leq}(x')$ with the note that the calculation of $\mathcal{P}_{\geq}(x')$ can be followed by replacing $\leq$ with $\geq$. The critical aspect is the initial expansion point, where $f_{\leq}$ is expressed in terms of a local expansion. The later translation in the polynomial description and the scaling of the variables can be done in any order. The polynomial expansion of $f_{\leq}$ around $\dagger(\mathbb{D}_{\leq})$ is given by

$$\mathcal{P}_{\geq,\dagger}(x) = f_{\leq}\left(\dagger(\mathbb{D}_{\leq}) + x\right). \tag{6.46}$$

With a linear transformation of $\mathcal{P}_{\geq,\dagger}(x)$ that transfers the expansion point to $\dagger\left(\mathbb{D}_f\right)$ and scales x to the width of $\mathbb{D}_f$, so,

$$\mathcal{P}_{\geq}\left(x'\right) = \mathcal{P}_{\geq,\dagger}\left(\dagger\left(\mathbb{D}_f\right) - \dagger\left(\mathbb{D}_{\geq}\right) + \frac{2x}{w\left(\mathbb{D}_f\right)}\right). \tag{6.47}$$

6.2.6 The Infinite 1D Equidistant Configuration

Before we investigate minimum energy configurations of k particles in 1D, we want to study the minimum energy state of an infinite one dimensional equidistant configurations. This one dimensional optimization problem can be solved analytically and potentially yields more insights into the minimum energy state of finite one dimensional configurations.

Consider infinitely many particles on a line with pairwise Lennard-Jones interaction, where the distance between any two adjacent particles is a constant value r. In order to calculate the distance $r^{\star}$ for which this configuration reaches its minimum energy state, we solve for

$$\left.\frac{dU(r)}{dr}\right|_{r=r^{\star}\in\mathbb{R}^{+}} = 0. \tag{6.48}$$

The overall potential $U(r)$ is the infinite sum of all the individual pairwise Lennard-Jones interaction potentials, such that the expression above yields

$$\frac{dU(r)}{dr} = \frac{d}{dr} \lim_{N\to\infty} \sum_{k=1}^{N} (N-k)\, U_{\mathrm{LJ}}(kr) = 0 \tag{6.49}$$

$$= \lim_{N\to\infty} \sum_{k=1}^{N} \frac{N}{N} \frac{dU_{\mathrm{LJ}}(kr)}{dr} - \cancel{\frac{k}{N} \frac{dU_{\mathrm{LJ}}(kr)}{dr}} = \frac{0}{N} \tag{6.50}$$

Each unique Lennard-Jones potential has a weight of 1 in the overall sum when considering an infinite configuration. The fact that there is one more distance of length l than there is of length $l + r$ becomes irrelevant when approaching infinity. Accordingly,

$$0 = \sum_{k=1}^{\infty} \frac{dU_{\mathrm{LJ}}(kr)}{dr} = \sum_{k=1}^{\infty} \frac{-12}{k^{12} r^{13}} + \frac{12}{k^6 r^7} = \frac{-12}{r^7} \left(\frac{1}{r^6} \sum_{k=1}^{\infty} \frac{1}{k^{12}} - \sum_{k=1}^{\infty} \frac{1}{k^6} \right) \tag{6.51}$$

$$= \frac{-12}{r^7} \left(\frac{\zeta(12)}{r^6} - \zeta(6) \right) = \frac{-4\pi^6}{315 r^7} \left(\frac{691\pi^6}{675675 r^6} - 1 \right), \tag{6.52}$$

where $\zeta(s)$ is the Riemann zeta function. Solving the expression above for r yields

$$r^\star = \pi \cdot \sqrt[6]{\frac{691}{675675}} \in [0.9971792638858069273, 0.9971792638858069274]\,. \tag{6.53}$$

This $r^\star$ would be a very good lower bound r_{LB} on the minimum inter-particle distance of any one dimensional configuration if we additionally prove that the infinite minimum energy configuration is indeed equidistant. While this seems to be the case from the results below and also intuitive from a symmetry point of view, it is not trivial to show this. Similar to the problem with the Kepler conjecture, there is no trivial way of excluding all irregular patterns.

6.2.7 The Verified Global Optimization Results for Configurations of k Particles in 1D

The configuration of k particles is placed on the positive x axis with the particle p_1 at $x_1 = 0$. The $k-1$ variables $v_{x,i} = x_{i+1} - x_i \geq 0$ describe the distances between two adjacent particles p_i and p_{i+1}. Accordingly, the position $x_i = \sum_{j=1}^{i-1} v_{x,j}$ of particle p_i is determined by the sum of all the inter-particle distances $v_{x,j}$ to the left of particle p_i.

The lower bound r_{LB} on the minimal inter-particle distance from Sec. 6.2.3.3 provides the lower bound on the variables $\vec{v}$, with $v_i \in [r_{\mathrm{LB}}, 1]$. Accordingly, the upper and lower bounds on the distances between particles, r_{LB} and $r_{\mathrm{UB}} = k - 1$, are directly incorporated into the domain restrictions of the variables.

Mirror configurations can still occur if there are asymmetric minimum energy configurations. Even if there are no asymmetric minimum energy configurations, the existence of asymmetric configurations that yield a local minimum in the overall interaction potential could potentially slow down the global optimizer. Accordingly, the mirror suppression from Eq. (6.20) is implemented.

The technique 1 from Sec. 6.2.3.2 creates symmetric upper bound configurations $\mathcal{S}_{k,\mathrm{UB}}$ using the configuration $\mathcal{S}^{\star}_{k-1}$, yielding a good initial cutoff value $\mathcal{C} = U_{\mathrm{UB},k}$.

The verified global optimization is performed with the Taylor Model based verified optimizer COSY-GO [55, 56], which is implemented in COSY INFINITY [21, 18, 53]. The performance of the optimizer varies with the order of the Taylor Models.

As a stopping condition, we set the minimum side length s_{min} for a box to be split to 1E-6.

Given $\mathcal{S}^{\star}_2$ with $v^{\star}_1 = 1$ and the associated minimum energy $U^{\star}_2 = 0$, we iteratively increase the number of particles k, beginning with $k = 3$. In Tab. 6.1, the performance of COSY-GO with LDB/QFB enabled and stopping condition $s_{\mathrm{min}} = 1\text{E-}6$ is shown for different Taylor Model orders.

Since the QFB requires a minimum order of two, the order one calculations are particularly bad and are not fully comparable to the higher order calculations.

The computation time is the product of the number of steps times the average computation time per step. The higher the order of computation, the tighter the bounding and the lower the required number of steps. At the same time, higher order computations are more time demanding. However,

Table 6.1: Performance of verified global optimization using COSY-GO with LDB/QFB enabled and stopping condition $s_{\min}$ = 1E-6 on minimum energy search of a one dimensional configuration of k particles, where the particle interaction energy is modeled by the Lennard-Jones potential. The Taylor Model orders are denoted by 'O'. The QFB requires a minimum order of two, which is why the order one (O1) calculation underperforms so significantly. The number of optimization variables n_{var} equals $k - 1$. All computation (except for O1) were able to reduce the search space to a single final box ($n_{\text{fin,boxes}} = 1$).

		Time [s]					Steps				
k	n_{pairs}	O1	O2	O3	O4	O5	O1	O2	O3	O4	O5
3	3	0.046	0.015	0.015	0.015	0.016	265	9	6	6	6
4	6	0.265	0.015	0.015	0.015	0.016	1445	16	11	9	9
5	10	3.398	0.016	0.015	0.015	0.031	7871	22	15	12	12
6	15	22.62	0.032	0.031	0.031	0.047	44185	28	18	16	16
7	21	145.4	0.032	0.032	0.047	0.078	241111	40	22	20	20
8	28	993.9	0.062	0.031	0.047	0.172	1265531	55	25	25	23
9	36		0.094	0.062	0.094	0.265		81	29	28	28
10	45		0.156	0.094	0.140	0.469		101	35	30	31
11	55		0.265	0.156	0.265	1.312		141	49	39	37
12	66		0.765	0.296	0.625	2.486		249	75	51	49
13	78		2.484	0.719	1.328	3.859		565	122	75	69
14	91		4.470	1.469	2.471	10.51		1158	194	103	91
15	105		11.55	2.286	7.180	20.45		2189	278	173	131

these two factors do not scale the same way with higher orders.

For this particular example, calculations of order three are the most time efficient. Compared to the second order calculation, the longer computation times of the third order calculations are overcompensated by the tighter bounding and the associated reduction in the number of steps required for the optimization. With higher order calculations, the number of steps can be reduced even further, but the computation time per step increases significantly, such that O4 is the second most time efficient an O5 is the least time efficient (not considering O1) despite the significant reduction of calculation steps.

Calculations of all orders were able to narrow the minimum energy state down to a single box (except for the O1 calculation). In Fig. 6.17, the position and the side lengths of this final domain box, corresponding to the optimized inter-particle distances $v_l^\star$ of the minimum energy configuration, are illustrated.

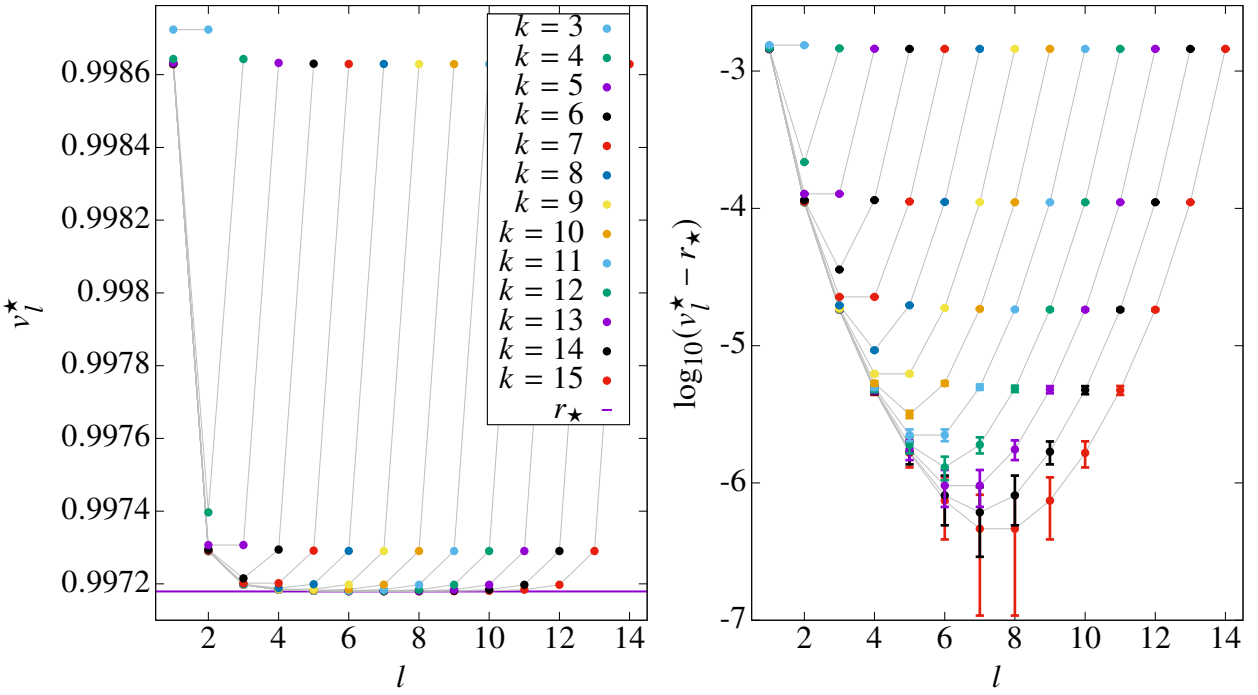

Figure 6.17: The plots show the values for the distances $v_l^\star$ of the minimum energy configuration of k particles that resulted from the global optimization. The minimum energy configuration seems to be symmetric with the middlemost distances asymptotically approaching a value that could very well be $r^\star$ from Sec. 6.2.6. The right plot emphasizes this hypothesis by plotting the logarithm of the difference between the calculated distances from the optimization and $r^\star$. The error bars indicate the side length of the resulting box.

The results are very symmetric and seem to asymptotically approach a minimum inter-particle distance that corresponds to $r^\star$ from Sec. 6.2.6. The specific values of the $v_l^\star$ are documented in Tab. A.1 and Tab. A.2, and the values for r_{LB} are listed in Tab. A.3. The side length of the box is illustrated by the error bars. The left plot emphasizes the short range of the potential because the distances between particles are almost the same except for the outermost particles on each side of the configuration. Only the logarithmic plot on the right can show that the distances get shorter in the middle of the configuration and seem to approach $r^\star$ from Sec. 6.2.6.

The resulting bounds on the minimum energy of the k-particle configurations are listed in Tab. 6.2.

Table 6.2: Verified global optimization results on the minimum energy $U^\star$ of a one dimensional configuration of k particles, where the particle interaction energy is modeled by the Lennard-Jones potential. The initial upper bound U_{UB} on the minimum energy was calculated using method 1 from Sec. 6.2.3.2. Optimizer: COSY-GO with LDB/QFB enabled and stopping condition s_{min} = 1E-6. The initial search volume is denoted by V_0, and the volume of the remaining $n_{\mathrm{fin,boxes}}$ boxes is represented by V_{fin}.

k	n_{var}	n_{pairs}	$n_{\mathrm{fin,boxes}}$	V_0/V_{fin}	U_{UB}	$U^\star$
3	2	3	1	6.4E11	0.968994140625000	0.968875869644^{86}_{68}
4	3	6	1	5.3E17	2.934929941521004	$2.93486371189^{826}_{799}$
5	4	10	1	3.6E23	5.900342654544756	$5.90034204308^{608}_{571}$
6	5	15	1	2.0E29	9.865688399486586	9.865688070463^{57}_{06}
7	6	21	1	8.2E34	14.83099005904018	$14.830990045365^{79}_{10}$
8	7	28	1	3.5E40	20.79627461693900	$20.7962746094^{684}_{598}$
9	8	36	1	8.6E45	27.76155137645425	$27.761551375^{70033}_{69916}$
10	9	45	1	2.3E51	35.72682430087387	35.726824300^{979}_{832}
11	10	55	1	4.9E56	44.69209518551273	$44.6920951854^{906}_{721}$
12	11	66	1	6.0E61	54.65736491976315	$54.657364919^{2057}_{1810}$
13	12	78	1	7.8E66	65.62263397159556	$65.6226339715^{563}_{250}$
14	13	91	1	9.1E71	77.58790260142806	$77.587902601^{2260}_{1870}$
15	14	105	1	9.5E76	90.55317096078301	$90.553170960^{8222}_{7742}$

6.2.8 The Verified Global Optimization Results for Symmetric Configurations of k Particles in 1D

Assuming that the minimum energy configurations are indeed symmetric, this section analyses the associated optimization problem. Considering symmetric 1D configurations roughly cuts the number of variables in half, since $v_i = v_{k-i}$. Additionally, the symmetry avoids distinct mirror configurations of local minimum energy configurations in the optimization process. All other parameters of the optimization like the bounds on the variables and the technique for the cutoff remain unchanged.

Given $\mathcal{S}_2^\star$ with $v_1^\star = 1$, we iteratively increase the number of particles k, beginning with $k = 3$.

In Tab. 6.3, the performance of COSY-GO with LDB/QFB enabled and stopping condition s_{min} = 1E-6 is shown for different Taylor Model orders.

In Fig. 6.18, the results from Tab. 6.3 on the time efficiency and the number of steps required

Table 6.3: Performance of verified global optimization using COSY-GO with LDB/QFB enabled and stopping condition $s_{\min}$ = 1E-6 on minimum energy search of a one dimensional symmetric configuration of k particles, where the particle interaction energy is modeled by the Lennard-Jones potential. The Taylor Model orders are denoted by 'O'.

				Time [s]			Steps		
k	n_{var}	n_{pairs}	$n_{\text{fin,boxes}}$	O2	O3	O4	O2	O3	O4
3	1	1	1	0.016	0.015	0.016	6	5	5
4	2	2	1	0.015	0.016	0.016	11	7	6
5	2	2	1	0.038	0.015	0.016	12	8	8
6	3	3	1	0.015	0.016	0.016	17	12	11
7	3	3	1	0.016	0.032	0.031	18	13	12
8	4	4	1	0.031	0.031	0.022	24	18	15
9	4	4	1	0.031	0.032	0.047	24	17	16
10	5	5	1	0.069	0.032	0.053	33	20	20
11	5	5	1	0.091	0.047	0.069	31	20	20
12	6	6	1	0.139	0.063	0.101	45	23	23
13	6	6	1	0.131	0.063	0.116	43	24	23
14	7	7	1	0.187	0.125	0.200	59	30	26
15	7	7	1	0.390	0.125	0.239	57	30	26
16	8	8	1	0.419	0.234	0.486	80	41	29
17	8	8	1	0.433	0.328	0.448	80	41	32
18	9	9	1	0.586	0.469	1.225	114	52	39
19	9	9	1	1.215	0.484	1.663	119	56	39
20	10	10	1	2.545	1.062	3.070	180	71	50
21	10	10	1	2.745	0.799	2.344	209	73	53
22	11	11	2	4.230	1.843	4.882	365	105	74
23	11	11	2048	57.269	84.550	243.494	4494	4195	4167
24	12	12	4096	111.483	193.292	672.107	8885	8315	8302
25	12	12	4096	122.401	209.074	734.148	8932	8327	8292
26	13	13	8192	290.751	514.414	1771.820	17549	16558	16531
27	13	13	8192	286.457	537.120	1855.865	17651	16568	16456

are visualized together with the corresponding results from the previous section (Sec. 6.2.7). The reduction of the optimization variables by assuming symmetric configurations significantly reduces the computation time and the number of steps as the left plot indicates. With the symmetric assumption, a configuration of k particles can be represented by $(k-1)/2$ variables instead of $k-1$, when k is odd. For $k < 23$, the calculations of order three are the most time efficient, while only requiring slightly more steps compared to higher order computations.

For $k \geq 23$, the number of final boxes increases drastically. Due to the high dimensionality, the

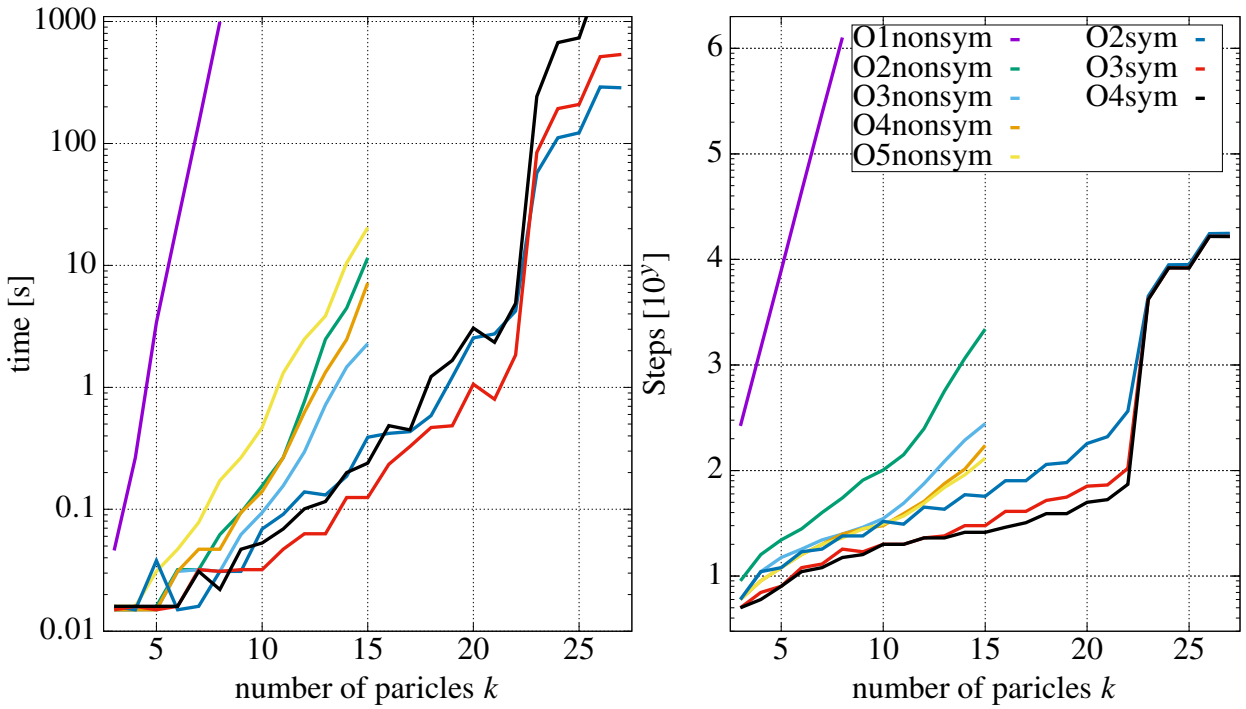

Figure 6.18: Performance of minimum energy search of a one dimensional configuration of k particles, where the particle interaction energy is modeled by the Lennard-Jones potential using COSY-GO at different Taylor Model orders with LDB/QFB enabled and stopping condition $s_{\min}$ = 1E-6. The order of the Taylor Models of the optimization is denoted by 'O'. The results from Sec. 6.2.7 are denoted by 'nonsym', because they assume that the minimum energy configuration is symmetric. Accordingly, the results from Sec. 6.2.8 are labeled with 'sym'.

overall interaction potential gets so shallow over the n_{var} dimensional domain that the floating-point accuracy restrictions prohibit narrowing down the minimum to a single final box of side length $s_{\min}$ = 1E-6. This behavior is the only aspect of the cluster effect [38, 30] that is not solved by higher order Taylor Models. It can only be improved by increasing the internal floating-point precision of the calculation.

Accordingly, the order of the Taylor Models is no longer an advantage when evaluating boxes in this plateau region but rather a disadvantage, since the number of those boxes are the same for every order calculation, while the higher order evaluation takes longer as the plots indicate.

In Tab. 6.4, the resulting bounds on the minimum energy of the symmetric configurations are listed. As expected, the results for symmetric 1D configurations agree with the previous results, where this symmetry was not assumed.

Table 6.4: Verified global optimization results on the minimum energy $U^\star$ of a symmetric one dimensional configuration of k particles, where the particle interaction energy is modeled by the Lennard-Jones potential. The initial upper bound U_{UB} on the minimum energy was calculated using method 1 from Sec. 6.2.3.2. Optimizer: COSY-GO with LDB/QFB enabled and stopping condition s_{min} = 1E-6.

k	n_{var}	n_{pairs}	$n_{\mathrm{fin,boxes}}$	V_0/V_{fin}	U_{UB}	$U^\star$
3	1	3	1	1.1E6	0.968994140625000	$0.96887586964{4}^{86}_{68}$
4	2	6	1	9.3E11	2.934929941521004	$2.9348637118{9}^{826}_{799}$
5	2	10	1	1.2E12	5.900342654544756	$5.9003420430{8}^{608}_{571}$
6	3	15	1	7.6E17	9.865688399486586	$9.86568807046{3}^{57}_{07}$
7	3	21	1	8.3E17	14.83099005904018	$14.830990045{365}^{79}_{11}$
8	4	28	1	4.2E23	20.79627461693900	$20.7962746094{7}^{684}_{599}$
9	4	36	1	4.0E23	27.76155137645425	$27.761551{375}^{70031}_{69919}$
10	5	45	1	1.4E29	35.72682430087387	$35.7268243004{4}^{979}_{835}$
11	5	55	1	1.3E29	44.69209518551273	$44.6920951854{3}^{906}_{725}$
12	6	66	1	3.1E34	54.65736491976315	$54.657364919{7}^{2056}_{1818}$
13	6	78	1	2.4E34	65.62263397159556	$65.6226339715{8}^{563}_{258}$
14	7	91	1	4.8E39	77.58790260142806	$77.587902601{4}^{2260}_{1879}$
15	7	105	1	3.7E39	90.55317096078301	$90.5531709607^{8222}_{7752}$
16	8	120	1	7.1E44	104.5184391413806	$104.518439141{3}^{8064}_{7511}$
17	8	136	1	9.5E44	119.4837072006288	$119.4837072006{2}^{876}_{298}$
18	9	153	1	1.8E50	135.4489751755410	$135.448975175{5}^{4100}_{3442}$
19	9	171	1	1.3E50	152.4142430906076	$152.414243090^{60765}_{59982}$
20	10	190	1	1.5E55	170.3795109624117	$170.379510962{4}^{1168}_{0223}$
21	10	210	1	1.1E55	189.3447788024156	$189.344778802{4}^{1559}_{0447}$
22	11	231	2	1.0E60	209.3100466186911	$209.310046618{6}^{9111}_{7777}$
23	11	253	2048	1.0E59	230.2753144170244	$230.275314417{0}^{2442}_{0216}$
24	12	276	4096	5.6E63	252.2405822016606	$252.240582201^{60768}_{58036}$
25	12	300	4096	3.6E63	275.2058499756172	$275.205849975{5}^{4204}_{1037}$
26	13	325	8192	2.1E68	299.1711177412260	$299.171117741{1}^{4227}_{0526}$
27	13	351	8192	1.4E68	324.1363855002667	$324.136385500{1}^{5532}_{1313}$

Note that each final box corresponds to a minimum energy configuration up to the precision of the calculation. By finding the smallest box that contains all those final box solutions, one can summarize them. The side lengths of this summarizing box will be larger than the side lengths of

the individual final boxes. In Fig. 6.19, the results for the distances v_l are given. For $k > 21$, the

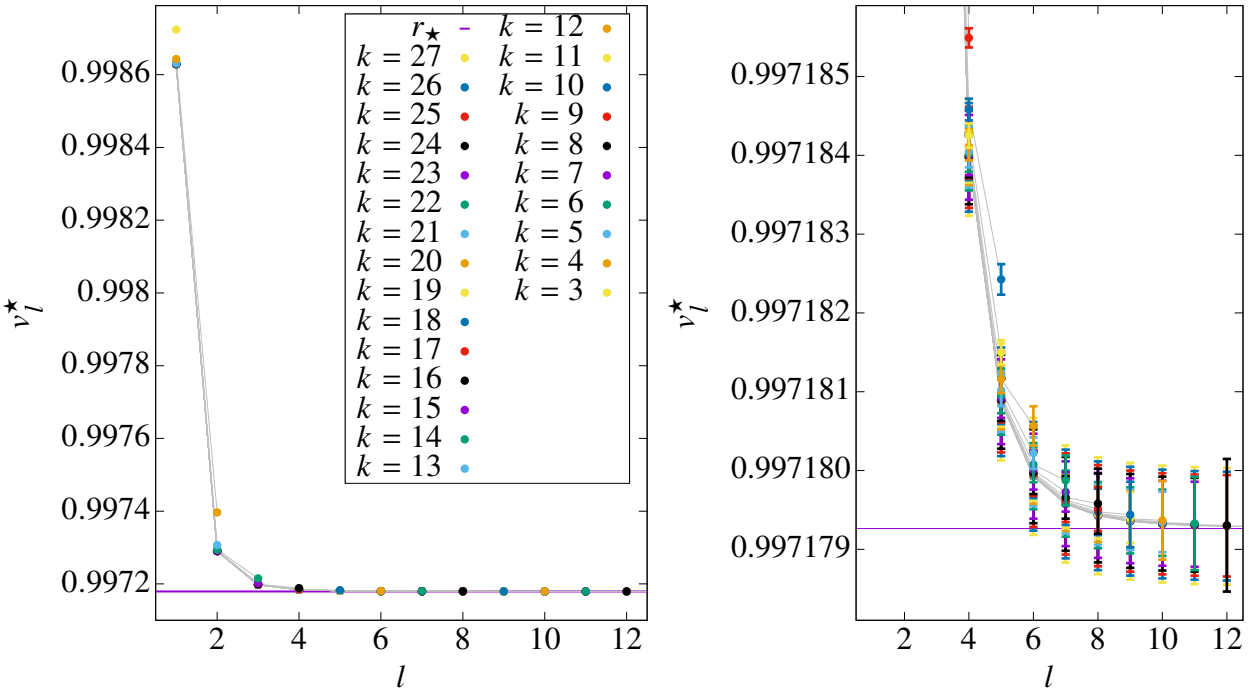

Figure 6.19: The plots show the values for the distances $v_l^\star$ of the minimum energy configuration of k particles that resulted from the global optimization. Again, the middlemost distances asymptotically approaching a value that could very well be $r^\star$ from Sec. 6.2.6. However, the right plot shows that the increasing error bars with a higher dimensionality of the optimization problem do not allow for clear conclusions.

final boxes are summarized, which explains the larger error bars. The specific values of the $v_l^\star$ of the final (summarized) box are documented in Tab. A.4, Tab. A.5, and Tab. A.6, and the values for r_{LB} are listed in Tab. A.7.

6.2.9 The Verified Global Optimization Results for Configurations of k Particles in 2D

The variables for a configuration of k particles in 2D are defined as explained in Sec. 6.2.3.7. The particles are numbered according to their x coordinate, such that $x_l \leq x_j$ for $l < j$. The first particle p_1 is set at the origin, so $x_1 = 0$. The $k - 1$ variables $v_{x,l}$ denote the x distance between particle p_l and particle p_{l+1}. Thus, the x position of particle p_l is given by $\sum_{i=1}^{l-1} v_{x,i}$.

The $k - 2$ variables $v_{y,l}$ denote the y position of the particles p_2 to p_{k-1}. The y position of p_k

is fixed to the positive value

$$\epsilon_y = 0.99 \cdot \frac{r_{\mathrm{LB}}}{2}, \tag{6.54}$$

where r_{LB} is determined by Eq. (6.11).

The initial search domain is defined according to Sec. 6.2.3.7 with $\epsilon_z = 0$ (see Fig. 6.14). As already mentioned in Sec. 6.2.4, the search domain includes entire manifolds where at least one inter-particle distance is zero, which makes the evaluation of the classical Lennard-Jones potential on those manifolds impossible for the global optimizer. As introduced in Sec. 6.2.4, the modified Lennard-Jones potential [8] from Eq. (6.37) is used to avoid this.

To avoid multiple equivalent representations of a minimum energy configuration, the suppression mechanisms from Eq. (6.19), Eq. (6.20), and Eq. (6.21) are implemented.

The verified global optimization is performed with the Taylor Model based verified optimizer COSY-GO [55, 56], which is implemented in COSY INFINITY. As a stopping condition, we use a minimum side length s_{min} = 1E-6 for boxes that are split.

Given $\mathcal{S}_3^\star$ with the associated minimum energy $U_3^\star = 0$, we iteratively increase the number of particles k, beginning with $k = 4$.

As before, the first step is determining the most time efficient calculation order since the performance of the optimizer varies with the order of the Taylor Models. Due to the drastically increasing complexity and computation time of the calculations with an increasing number of particles, we will only use the four and five particle configurations for this evaluation (see Tab. 6.5). Again, order three turns out to be the most time efficient calculation order. The increased computation

Table 6.5: Performance of verified global optimization using COSY-GO with LDB/QFB enabled and stopping condition s_{min} = 1E-6 on minimum energy search of a two dimensional configuration of k particles, where the particle interaction energy is modeled by the Lennard-Jones potential. The Taylor Model orders are denoted by 'O'.

		Time [s]				Steps			
k	n_{var}	O2	O3	O4	O5	O2	O3	O4	O5
4	5	2.506	1.442	1.975	3.392	8916	6053	5722	5687
5	7	142.137	112.273	164.691	259.418	434776	312476	302708	300832

time per step is more than compensated by the reduction in the number of required steps.

The minimum energy configuration for four particles in 2D is shown in Fig. 6.20. From Fig. 6.20,

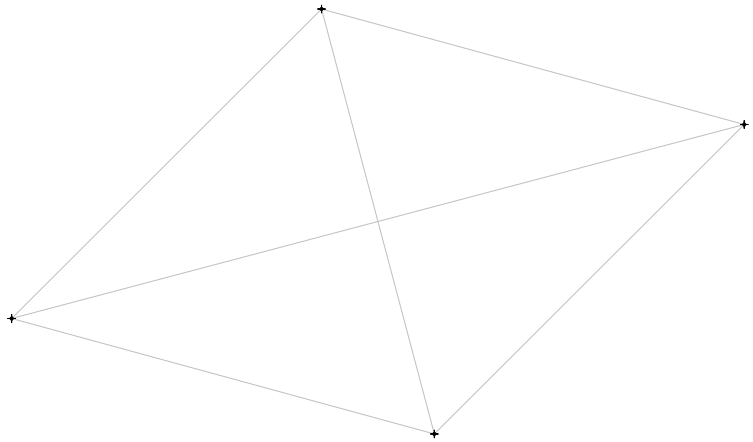

Figure 6.20: Minimum energy configuration of four particles in 2D, where the pairwise particle interaction is modeled by the Lennard-Jones potential. Note the tilt of the major axis. It avoids that the middle two particles have the same x position, which would otherwise yield two ambiguous numbering schemes. Interestingly, the minimum energy configuration is not a square but a rhombus.

the configuration is indistinguishable from two connected equilateral triangles. Tab. 6.6 reveals the distances between the individual particles and the difference between the minimum energy Lennard-Jones configuration and two connected equilateral triangles. It also includes the resulting

Table 6.6: Verified global optimization results for the minimum energy configurations of four particles in 2D, where the pairwise particle interaction is modeled by the Lennard-Jones potential (see Sec. 6.2.9). The distance r_{ij} between particle p_i and p_j emphasize the difference to the closest packing structure. The variable $v_{x,l}$ is the x distance between particles p_l and p_{l+1}. The variable $v_{y,l}$ is the y position of particle p_l. Note that $v_{y,k}$ is not a variable but the constant ϵ_y and that $v_{y,1}$ is not defined because $y_1 = 0$.

k	i	j	r_{ij}
4	1	2	0.998012^{419}_{220}
4	1	3	0.998012^{478}_{162}
4	1	4	1.726251^{685}_{264}
4	2	3	1.00208^{3083}_{2785}
4	2	4	0.998012^{478}_{162}
4	3	4	0.998012^{419}_{220}

k	†	l	$v^{\star}_{\dagger,l}$
4	x	1	0.705906^{70}_{42}
4	x	2	0.256794^{88}_{60}
4	x	3	0.705906^{70}_{42}
4	y	2	0.70549^{607}_{581}
4	y	3	-0.26312^{521}_{495}
4	y	4	0.44237086

variables of the optimization problem as well as the vertical offset of the last particle ϵ_y.

Compared to two equilateral triangles, the minimum energy Lennard-Jones configuration brings the outermost particles closer together, which drives the two particles in the middle (p_2 and p_3) slightly apart. This 'squishing' of an equidistance structure to yield the minimum energy Lennard-Jones configuration could already be observed in the 1D minimum energy configurations. This is possible because incremental changes of the distance between particles from the optimal distance $r^\star = 1$ come at a lower cost to the overall potential than the benefit of reducing the distance between to particles with $r_{ij} > 1$.

For five particles in 2D, this is 'squishing' can also be observed. Fig. 6.21, which visualizes the minimum energy configuration is again barely distinguishable from the three equilateral triangles. The distances between the individual particles, provided by Tab. 6.7, emphasize the 'squishing'.

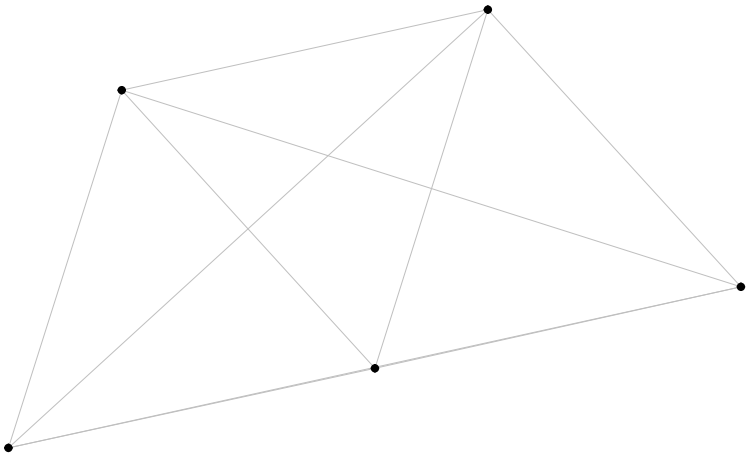

Figure 6.21: Minimum energy configuration of five particles in 2D, where the pairwise particle interaction is modeled by the Lennard-Jones potential.

Tab. 6.7 also provides the values of the optimized variables and the vertical offset of the last particle ϵ_y. The first and last particles move closer together to a distance below two. Particles two and four are pulled downwards, reducing their distance to each other and their distance to the first and last particle. Particle three, which is the one in the middle of the configuration, does something special. It preserves the ideal distance of 1 as well as possible with the upper two particles p_2 and p_4.

Table 6.7: Verified global optimization results for the minimum energy configurations of five particles in 2D, where the pairwise particle interaction is modeled by the Lennard-Jones potential (see Sec. 6.2.9). The distance r_{ij} between particle p_i and p_j emphasize the difference to the closest packing structure. The variable $v_{x,l}$ is the x distance between particles p_l and p_{l+1}. The variable $v_{y,l}$ is the y position of particle p_l. Note that $v_{y,k}$ is not a variable but the constant ϵ_y and that $v_{y,1}$ is not defined because $y_1 = 0$.

k	i	j	r_{ij}
5	1	2	0.99800^{7224}_{6996}
5	1	3	0.996784^{635}_{187}
5	1	4	1.72679^{5218}_{4649}
5	1	5	1.993561^{902}_{129}
5	2	3	1.000010^{583}_{190}
5	2	4	0.99610^{8205}_{7773}
5	2	5	1.72679^{5224}_{4643}
5	3	4	1.000010^{601}_{171}
5	3	5	0.996784^{619}_{203}
5	4	5	0.99800^{7242}_{6977}

k	†	l	$v^{\star}_{\dagger,l}$
5	x	1	0.301274^{57}_{13}
5	x	2	0.67279^{60}_{27}
5	x	3	0.300037^{44}_{09}
5	x	4	0.672868^{63}_{27}
5	y	2	0.95144^{745}_{14}
5	y	3	0.21160^{125}_{087}
5	y	4	1.165539^{60}_{26}
5	y	5	0.42847346

For six particles, the computation times on a single machine are very long. Accordingly, parallel computations are very helpful. COSY-GO is implemented in a way that easily allows for parallel computations using MPI. A critical parameter of the parallel computations is the time between processor communication and the associated load balancing. After this time, the processors exchange their remaining domain boxes and redistribute them. They also share their most recent cutoff value. If the time between communication is chosen too long, some processors will run out of work while others still have a lot of boxes to evaluate. If the time is chosen too short, too much of the computation time is wasted on communication.

The time for communication depends on multiple factors. Each processor runs the same repetitive code with different content. At some point in the repetitive process, the code checks if it is time to communicate. If it is time to communicate, the processor gathers all the data for communication and waits for all the other processors to do the same. The exchange of data only happens when all processors are ready for it. The more processors there are, the longer the potential wait time. The wait time becomes additionally problematic if the computation steps in the repetitive process take a

lot of time since it reduced the frequency of checking whether it is time to communicate. High-order Taylor Model evaluation can increase the time for individual calculation steps in the process.

To evaluate a good time between communications t_{com} for order three calculations, we investigate the optimization for six particles in 2D for multiple t_{com} on 64 cores on NERSC (see Tab. 6.8).

Table 6.8: Performance of verified global optimization using COSY-GO with LDB/QFB enabled and stopping condition $s_{\min}$ = 1E-6 on minimum energy search of a two dimensional configuration of six particles, where the particle interaction energy is modeled by the Lennard-Jones potential. The Taylor Model orders are denoted by 'O'. The computation is run in parallel on 64 cores on NERSC using different times between communication t_{com}.

t_{com} [s]	computation time [s]	steps
0.25	1497.9	152156157
0.5	1283.1	153261678
1	1267.0	152479475
2	1262.1	152242640
3	1254.0	152327673
4	1300.1	152867144

A good time between communications lies somewhere in the range from one to three seconds. This analysis is used for the optimization of the configuration with more than six particles. Specifically, we chose t_{com} = 3 s for parallel computations on 64 cores on NERSC.

In Fig. 6.22, the minimum energy Lennard-Jones configuration of six particles in 2D is illustrated. This structure is composed of four almost equilateral triangles. Connecting the points as done in Fig. 6.22 makes the shape look like an envelope.

Tab. 6.9 yields the distances between the particles in the minimum energy configuration and the associated variables of the optimization problem. In the corresponding triangle structure, there are nine unit distances, four distances of $\sqrt{3}$ from the height of two stacked triangles, and two distances of value two. As we already saw previously, distances larger than the optimal unit distance are shorter in the minimum energy configuration at the cost of the optimal unit distances slightly diverting from one to either smaller or larger values. The symmetry of the configuration is captured in the symmetry of the values in the left table of Tab. 6.9. Note that the only two connections that do not have a symmetric partner are the distances between p_1 and p_5, and p_2 and p_6.

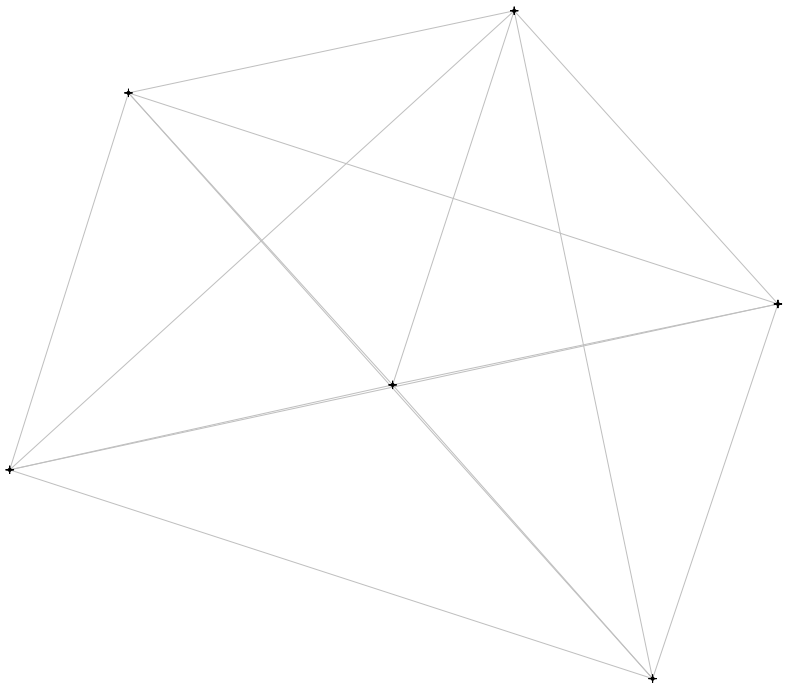

Figure 6.22: Minimum energy configuration of six particles in 2D, where the pairwise particle interaction is modeled by the Lennard-Jones potential.

The minimum energy Lennard-Jones configuration for seven particles in 2D is highly symmetrical, as Fig. 6.23 illustrates. It is a regular hexagon with a particle at each of the edges and one particle right in the middle.

This symmetry is further supported by the values for the distances between the individual particles in Tab. 6.10. The table also shows the optimized variables.

The configuration is equivalent to an equilateral hexagon with a side length of roughly 0.996434.

The results of the bounds on the minimum energy of the minimum energy configurations in 2D are listed in Tab. 6.11.

As for the symmetric 1D optimization in Sec. 6.2.8, the cluster effect due to the floating-point accuracy occurs. Luckily, the two boxes of the calculation of seven particles could easily be summarized as they are right next to each other in $v_{y,4}$.

To calculate configurations with more particles, it is reasonable to increase the minimum side

Table 6.9: Verified global optimization results for the minimum energy configurations of six particles in 2D, where the pairwise particle interaction is modeled by the Lennard-Jones potential (see Sec. 6.2.9). The distance r_{ij} between particle p_i and p_j emphasize the difference to the closest packing structure. The variable $v_{x,l}$ is the x distance between particles p_l and p_{l+1}. The variable $v_{y,l}$ is the y position of particle p_l. Note that $v_{y,k}$ is not a variable but the constant ϵ_y and that $v_{y,1}$ is not defined because $y_1 = 0$.

k	i	j	r_{ij}	k	i	j	r_{ij}
6	1	2	1.000179^{827}_{337}	6	2	6	1.72842^{2485}_{0325}
6	1	3	0.99288^{3667}_{2721}	6	3	4	0.99590^{8620}_{7690}
6	1	4	1.726^{601011}_{599805}	6	3	5	0.99288^{3862}_{2527}
6	1	5	1.7111^{31111}_{29095}	6	3	6	0.99659^{7556}_{5718}
6	1	6	1.9894^{51746}_{49154}	6	4	5	1.726^{600834}_{599983}
6	2	3	0.99659^{7061}_{6212}	6	4	6	0.99821^{7628}_{6406}
6	2	4	0.99821^{7474}_{6561}	6	5	6	1.000179^{832}_{332}
6	2	5	1.9894^{51208}_{49692}				

k	†	l	$v^{\star}_{\dagger,l}$
6	x	1	0.300727^{59}_{09}
6	x	2	0.668624^{70}_{31}
6	x	3	0.307642^{43}_{02}
6	x	4	0.35044^{932}_{856}
6	x	5	0.31717^{329}_{265}
6	y	2	0.953898^{64}_{27}
6	y	3	0.214881^{68}_{22}
6	y	4	1.162082^{44}_{03}
6	y	5	-0.528578^{90}_{57}
6	y	6	0.41997833

length $s_{\min}$ of boxes that are split.

The values for the lower bound on the inter-particle distance r_{LB} are listed in Tab. 6.12.

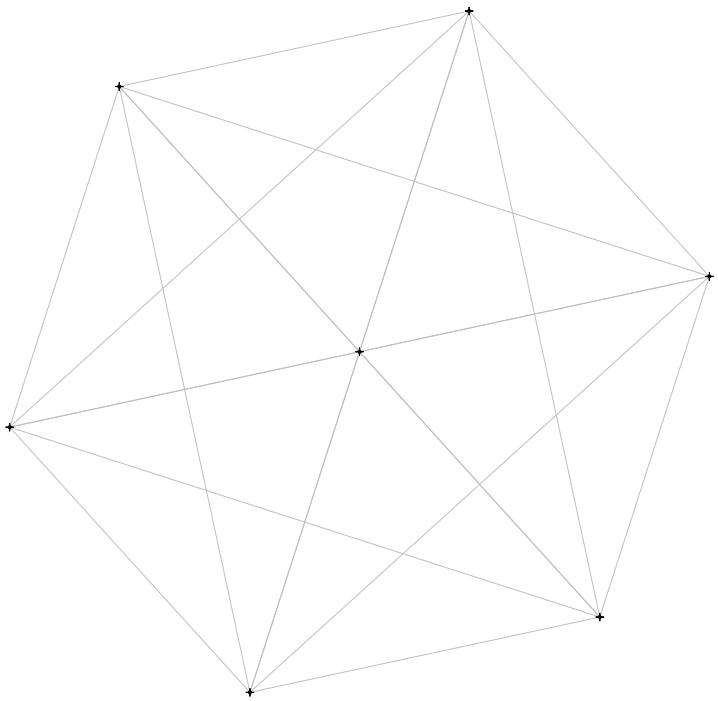

Figure 6.23: Minimum energy configuration of seven particles in 2D, where the pairwise particle interaction is modeled by the Lennard-Jones potential.

Table 6.10: Verified global optimization results for the minimum energy configurations of seven particles in 2D, where the pairwise particle interaction is modeled by the Lennard-Jones potential (see Sec. 6.2.9). The distance r_{ij} between particle p_i and p_j emphasize the difference to the closest packing structure. The variable $v_{x,l}$ is the x distance between particles p_l and p_{l+1}. The variable $v_{y,l}$ is the y position of particle p_l. Note that $v_{y,k}$ is not a variable but the constant ϵ_y and that $v_{y,1}$ is not defined because $y_1 = 0$.

k	i	j	r_{ij}	k	i	j	r_{ij}
7	1	2	0.996434^{890}_{384}	7	3	4	0.996434^{894}_{217}
7	1	3	0.99643^{5167}_{4108}	7	3	5	1.99286^{9787}_{8762}
7	1	4	0.99643^{5421}_{3817}	7	3	6	0.99643^{5464}_{3811}
7	1	5	1.72587^{6314}_{4522}	7	3	7	1.72587^{6314}_{4522}
7	1	6	1.72587^{6746}_{4090}	7	4	5	0.99643^{5058}_{4381}
7	1	7	1.9928^{70842}_{67707}	7	4	6	0.99643^{5130}_{4017}
7	2	3	1.72587^{5849}_{4987}	7	4	7	0.99643^{5458}_{3853}
7	2	4	0.99643^{5257}_{4145}	7	5	6	1.72587^{5849}_{4987}
7	2	5	0.99643^{5464}_{3811}	7	5	7	0.99643^{5167}_{4108}
7	2	6	1.9928^{70259}_{68290}	7	6	7	0.996434^{890}_{384}
7	2	7	1.72587^{6746}_{4090}				

k	†	l	$v^{\star}_{\dagger,l}$
7	x	1	0.30518^{721}_{666}
7	x	2	0.36368^{640}_{578}
7	x	3	0.30518^{717}_{670}
7	x	4	0.30518^{717}_{670}
7	x	5	0.36368^{640}_{578}
7	x	6	0.30518^{721}_{666}
7	y	2	0.948547^{98}_{61}
7	y	3	-0.738573^{74}_{34}
7	y	4	0.209974^{44}_{08}
7	y	5	1.15852^{226}_{186}
7	y	6	-0.528599^{46}_{09}
7	y	7	0.41994852

Table 6.11: Verified global optimization results on the minimum energy $U^{\star}$ of a two dimensional configuration of k particles, where the particle interaction energy is modeled by the Lennard-Jones potential. The optimization was performed using COSY-GO with LDB/QFB enabled and the stopping condition $s_{\min}$ = 1E-6.

k	n_{var}	n_{pairs}	$n_{\text{fin,boxes}}$	V_0/V_{fin}	U_{UB}	$U^{\star}$
4	5	6	1	1.2E34	0.927297668038409	$0.92657914153^{731}_{682}$
5	7	10	1	2.8E47	2.823589476701818	$2.82197624549^{239}_{164}$
6	9	15	1	4.2E61	5.647178953403635	$5.64172565099^{515}_{415}$
7	11	21	2	4.1E74	8.470768430105453	$8.46513348231^{335}_{195}$

Table 6.12: Results for the calculated lower bounds r_{LB} on the minimum distance between particles in a 2D configuration of k particles (see Eq. (6.11) and Sec. 6.2.9).

k	r_{LB}
4	0.893678512
5	0.865602947
6	0.848441067
7	0.848380848

6.3 Verified Stability Analysis of Dynamical Systems

Verified calculations are particularly important for the stability analysis of dynamical systems. With a verified upper bound on the rate of divergence, a system's long term stability can be rigorously estimated. Both of the previously discussed applications in Chapter 4 and Chapter 5 will benefit to different degrees from such verified stability estimates.

6.3.1 The Potential Implications for the Bounded Motion Problem

For the bounded motion orbits under zonal perturbation in the Earth's gravitational field (see Chapter 4), a stability estimate is the maximum rate at which two bounded orbits drift apart. Below we want to list aspects to consider for the calculation of such a verified upper bound on the rate of divergence.

The bounded motion conditions from Sec. 4.2.5 require that the average nodal period $\overline{T}_d$ and the average drift of the ascending node $\overline{\Delta\Omega}$ of two bounded orbits are the same. In other words, two orbits drift apart if those two averaged quantities are not the same for the two orbits. Additionally, each of the orbits might be diverging on its own by slowly increasing or decreasing its distance from the Earth. A verified upper bound on each of those diverging factors must be determined to combine them to an overall verified upper bound on the rate at which the two bounded orbits drift apart.

An upper bound on the radial drift rate of the bounded orbits moving apart is determined by the maximum difference between the individual radial drifts of each of the bounded orbits. The normal form defect of the radial phase space can be used as a measure for this radial drift. However, both the maximum and the minimum normal form defect of each orbit are relevant to determine the worst-case scenario of one of the orbits decreasing its amplitude and one of the orbits increasing its amplitude.

The longitudinal drift rate of the bounded orbits moving apart is determined by the difference in the average revolution frequency of the orbital planes around the symmetry axis. The revolution frequency is proportional to the drift of the ascending node $\overline{\Delta\Omega}$ per nodal period $\overline{T}_d$. Since both of

these quantities are oscillating at the same rate, the average revolution frequency can be calculated as the ratio of the average drift of the ascending node $\overline{\Delta\Omega}$ and the average nodal period $\overline{T}_d$.

Even if the orbital planes of the two bounded orbits are not radially or longitudinally drifting apart, the satellites on those orbits might still be drifting apart due to different average nodal periods, which constitutes the third drift factor.

These three factors have to be taken into account and rigorously estimated to calculate an overall maximum drift rate. The combination of the individual factors is not trivial since they are not independent of each other, e.g., the individual radial drifts of the orbits have nonlinear influences on the bounded motion quantities $\overline{\Delta\Omega}$ and $\overline{T}_d$. Global verified optimization of the overall drift rate is required to determine the maximum rate of divergence for any possible combination of the individual radial drift rates.

Given that the overall maximum drift rate is formally defined, we need to determine verified versions of the involved quantities. Accordingly, the starting point of the rigorous calculation of the maximum drift rate is a rigorous map of the system.

The map is based on the equations of motion of the system, which include the zonal coefficients of the Earth's gravitational potential based on measurements. To be rigorous is has to be decided if these coefficients are assumed to be exact or if the uncertainty about these coefficients is considered in the calculation. Given that the approach from Chapter 4 considers the zonal problem, ignoring sectional and tesseral terms, it seems reasonable to consider an idealized system where these coefficients are assumed to be exact.

In the next step, the verified integration of the equations of motion is required to calculate a verified map representation of the system. In our approach (see Chapter 4), we express the vertical momentum component v_z in terms of the other variables and system parameters. This operation includes the calculation of an inverse, which requires special methods to be performed rigorously. For the projection of the transfer map onto the Poincaré surface representing a generalized ascending node state, another rigorous computation of an inverse is required. Additionally, every step of the normal form based averaging procedure from Sec. 4.3.4 for the determination of the averaged

quantities $\overline{\Delta\Omega}$ and $\overline{T}_d$ has to be performed rigorously. The approach then calls for another inversion to calculate the constants of motion $\mathcal{H}_z$ and E as a function of the phase space variables such that the averaged bounded motion quantities match between any two orbits in the phase space.

If all those procedures are performed rigorously, one can calculate rigorous bounds on the normal form defect of the system, which can then be used together with the rigorous estimations of the averaged quantities $\overline{\Delta\Omega}$ and $\overline{T}_d$ to calculate the rigorous overall rate of divergences.

In summary, much effort is required to establish a verified upper bound on the maximum rate at which bounded orbits of the zonal problem drift apart. However, the practical implications of such an estimate are limited since the approach does not consider the fully perturbed system. Accordingly, we want to focus our attention on the application of a rigorous stability analysis for the system discussed in Chapter 5.

6.3.2 The Implications for the Stability Analysis of the Muon g-2 Storage Ring

A verified stability estimate of the muon g-2 storage ring is the verified maximum rate at which particles escape the storage region of the storage ring. The number of such escaped particles is very important for this high precision experiment, because for reasons not to be discussed here, they will introduce a systematic bias for the average polarization of the remaining particles, which will influence the overall result of the measurement. Below we want to discuss the aspects to consider for the calculation of such a verified upper bound on the rate of divergence and rigorously compute such a verified stability estimate in form of the normal form defect using Taylor Model based verified global optimization.

The normal form defect analysis will be conducted using a nonverified map as more research is required to calculate a fully verified phase space map of the storage ring. As already mentioned above, there are many intricacies to consider for a fully rigorous map calculation. A major challenge regarding the verified calculation of the storage ring map is the verified representation of every storage ring component, including all its perturbations, e.g., perturbations from ESQ fringe fields and imperfection in the magnetic field. To assess whether our computation order is high enough –

the main numerical error which is not based on measurement errors – we estimate inaccuracies in the map by computing maps of various orders and show that these are sufficiently small and will not affect the motion.

Accordingly, we will use the nonverified map from Chapter 5 and assume that it is exact. For comparison, we will additionally analyze a storage ring map that considers an ESQ voltage of 17.5 kV instead of 18.3 kV. The tunes of particles under the influence of an ESQ voltage of 17.5 kV are further away from the vertical $1/3$ resonance tune. Accordingly, we expect less diverging behavior for this map compared to the 18.3 kV map.

The goal is to rigorously analyze the stability of the entire five dimensional storage phase space $(x, a, y, b, \delta p)$ of the storage ring maps using verified global optimization of the normal form defect. In Sec. 6.3.3, we specify the normal form defect function as the objective function of the optimization problem. To be able to distinguish the diverging behavior in different areas of the storage region, we divide the five dimensional space into partitions. Each of those partitions is then used as the search domain for the verified global optimizer to find the maximum normal form defect in it. In Sec. 6.3.4, we present the onion layer approach [22, 8], which partitions the storage region according to the dynamics in the phase space. Next, we illustrate the complexity and strong nonlinearity of the normal form defect in multiple such onion layers and how it changes for different phase space regions and ESQ voltages (see Sec. 6.3.5). In Sec. 6.3.6, the results of the verified global optimization for the two maps are presented and compared to each other and the results of a nonverified analysis.

To understand the differences between the map for 17.5 kV and 18.3 kV, we present a tune shift analysis of the map for 17.5 kV. In Sec. 6.3.8, we show how different order of the normal form transformation affect the normal form defect.

6.3.3 The Normal Form Defect as the Objective Function for the Optimization

In Sec. 2.4, the normal form defect for the propagation of a state $\vec{z}$ with a map $\mathcal{M}$ was introduced as the difference between the normal form radius of the mapped state $\mathcal{M}(\vec{z})$ and the normal form

radius of the original state $\vec{z}$. If the motion occurs in multiple phase space dimensions, there is some ambiguity to the term 'normal form radius' and the associated normal form defect.

From the definition and algorithms of normal form transformations discussed in Sec. 2.3 and Sec. 2.4, it follows that there is a normal form radius for each normal form phase space. Each of these radii, yields the radius of the circular motion in this particular normal form phase space with

$$r_{\mathrm{NF},i}\left(\vec{z}_0\right) = \sqrt{\left(q_{\mathrm{NF,i}}\left(\vec{z}_0\right)\right)^2 + \left(p_{\mathrm{NF,i}}\left(\vec{z}_0\right)\right)^2}. \tag{6.55}$$

Accordingly, as defined in Sec. 2.4, there is a normal form defect defined for each of those normal form radii, with

$$d_{\mathrm{NF},i}\left(\vec{z}_0\right) = r_{\mathrm{NF},i}\left(\mathcal{M}\left(\vec{z}_0\right)\right) - r_{\mathrm{NF},i}\left(\vec{z}_0\right). \tag{6.56}$$

Additionally, we define the (overall) normal form radius of the motion as the euclidean composition of the individual normal form radii, with

$$r_{\mathrm{NF}}\left(\vec{z}_0\right) = \sqrt{\sum_i r^2_{\mathrm{NF},i}\left(\vec{z}_0\right)}. \tag{6.57}$$

This definition of the (overall) normal form radius corresponds to the following definition for the (overall) normal form defect

$$d_{\mathrm{NF}}\left(\vec{z}_0\right) = r_{\mathrm{NF}}\left(\mathcal{M}\left(\vec{z}_0\right)\right) - r_{\mathrm{NF}}\left(\vec{z}_0\right). \tag{6.58}$$

Unless stated otherwise, we will be using and referring to the (overall) normal form radius and the (overall) normal form defect.

6.3.4 The Search Domain in the Form of Onion Layers

The onion layer approach describes a way to partition the phase space regions and determine the associated variables for the global optimization. For the partitioning, it is important to consider the dynamics of the system. In Chapter 5, we saw that the main characteristics of the phase space motion in the storage ring are the oscillation amplitudes and the momentum offset δp. Accordingly,

we want to calculate the verified stability estimates on the rate of divergence based on partitions categorized by those criteria.

While the partitioning according to the momentum offset δp is straightforward, defining the partitions of different phase space amplitudes is not, because the phase space curve of a particle with a certain amplitude forms a nonlinearly distorted elliptical shape in the original phase space. The onion layer approach (see Fig. 6.24) partitions the phase space along those nonlinearly distorted elliptical phase space curves using the normal form transformation.

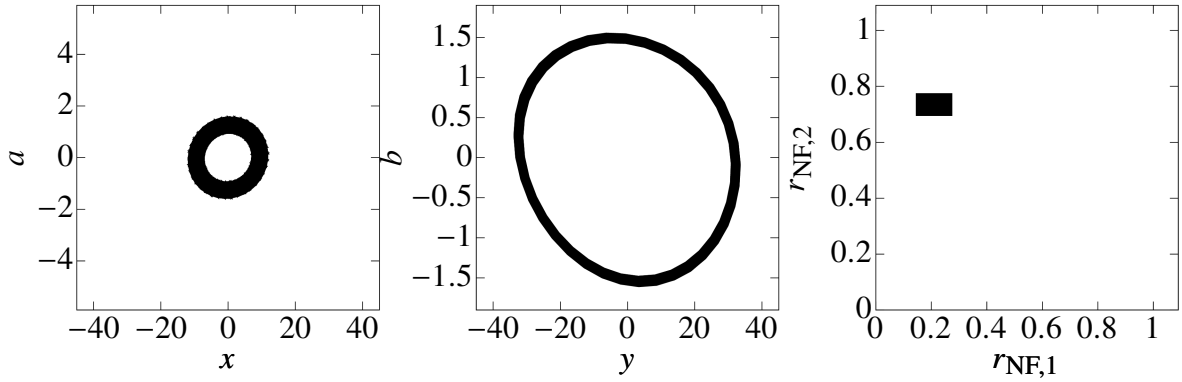

Figure 6.24: The left and the middle plot show the representation of an onion layer (black region) in regular phase space coordinates. The thickness of the onion layer is determined by the range in $r_{NF,1}$ and $r_{NF,2}$ as well as the range in δp. For this particular example, we set δp to a fixed value of $\delta p = 0\%$ instead of a range. The range in the normal form radii is given by $r_{NF,1} \in [0.15, 0.25]$ and $r_{NF,2} \in [0.7, 0.75]$. Note that the thickness in $r_{NF,1}$ is twice the thickness in $r_{NF,2}$. Accordingly, the projection of the onion layer into the radial phase space (x, a) appears roughly twice as thick as the projection into the vertical phase space (y, b).

As illustrated in Fig. 6.24, the normal form coordinates allow us to partition by amplitude. They are our best approximation of mapping the orbital phase space behavior onto circles. Accordingly, we can use the normal form description of the motion to define the onion layers for the global optimization. Specifically, we chose the normal form radii $r_{NF,1}$ and $r_{NF,2}$ as well as the corresponding normal form phase space angles $\phi_{NF,1}$ and $\phi_{NF,2}$ as the optimization variables. Additionally, the momentum offset δp is also considered an optimization variable.

The normal form phase space variables $(q_{NF,1}, p_{NF,1})$ and $(q_{NF,2}, p_{NF,2})$ are expressed in terms

of the polar optimization variables with

$$\begin{pmatrix} q_{\mathrm{NF},1} \\ p_{\mathrm{NF},1} \end{pmatrix} = r_{\mathrm{NF},1} \begin{pmatrix} \cos\left(\phi_{\mathrm{NF},1}\right) \\ \sin\left(\phi_{\mathrm{NF},1}\right) \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} q_{\mathrm{NF},2} \\ p_{\mathrm{NF},2} \end{pmatrix} = r_{\mathrm{NF},2} \begin{pmatrix} \cos\left(\phi_{\mathrm{NF},2}\right) \\ \sin\left(\phi_{\mathrm{NF},2}\right) \end{pmatrix}. \tag{6.59}$$

The inverse normal form transformation $\mathcal{A}^{-1}$ is then used as a vehicle to express the relevant phase space regions in original phase space (x, a, y, b) in terms of the optimization variables $(r_{\mathrm{NF},1}, \phi_{\mathrm{NF},1}, r_{\mathrm{NF},2}, \phi_{\mathrm{NF},2}, \delta p)$.

Moving along the angles $\phi_{\mathrm{NF},1}$ and $\phi_{\mathrm{NF},2}$ will approximately move along the phase space curve in the original coordinates. Accordingly, the search domain in those optimization variables is always $[-\pi, \pi]$. The domain on the normal form radii and the momentum offset determines the thickness of the onion layer, as illustrated in Fig. 6.24, and is set to 0.04 for normal form radii and to 0.04% in the momentum offset space.

6.3.5 The Complexity and Nonlinearity of the Normal Form Defect Function

In Chapter 5, we analyzed the normal form defect that individual particles encounter during stroboscopic tracking. In other words, we only probed individual phase space points of a particle's orbit for its normal form defect. We found that muons that encounter phase space regions with larger normal form defects are more likely to get lost (see Fig. 5.12). However, the probing only yields an incomplete picture of the normal form defect that a particle can potentially encounter. Fig. 6.25 illustrates how much the normal form defect can vary for fixed normal form amplitudes that approximately represent the normal form defect landscape along the phase space curve of a single particle.

Fig. 6.25 illustrates the radial normal form defect of an onion layer of zero thickness, which is given by a single point in the 3D onion layer thickness space of $r_{\mathrm{NF},1}$, $r_{\mathrm{NF},2}$, and δp. The landscape is characterized by highly nonlinear behavior with many local minima and maxima, which are extreme points of very steep valleys and hills. Accordingly, the stroboscopic normal form defect probing while tracking can significantly underestimate the maximum normal form defect of an orbit

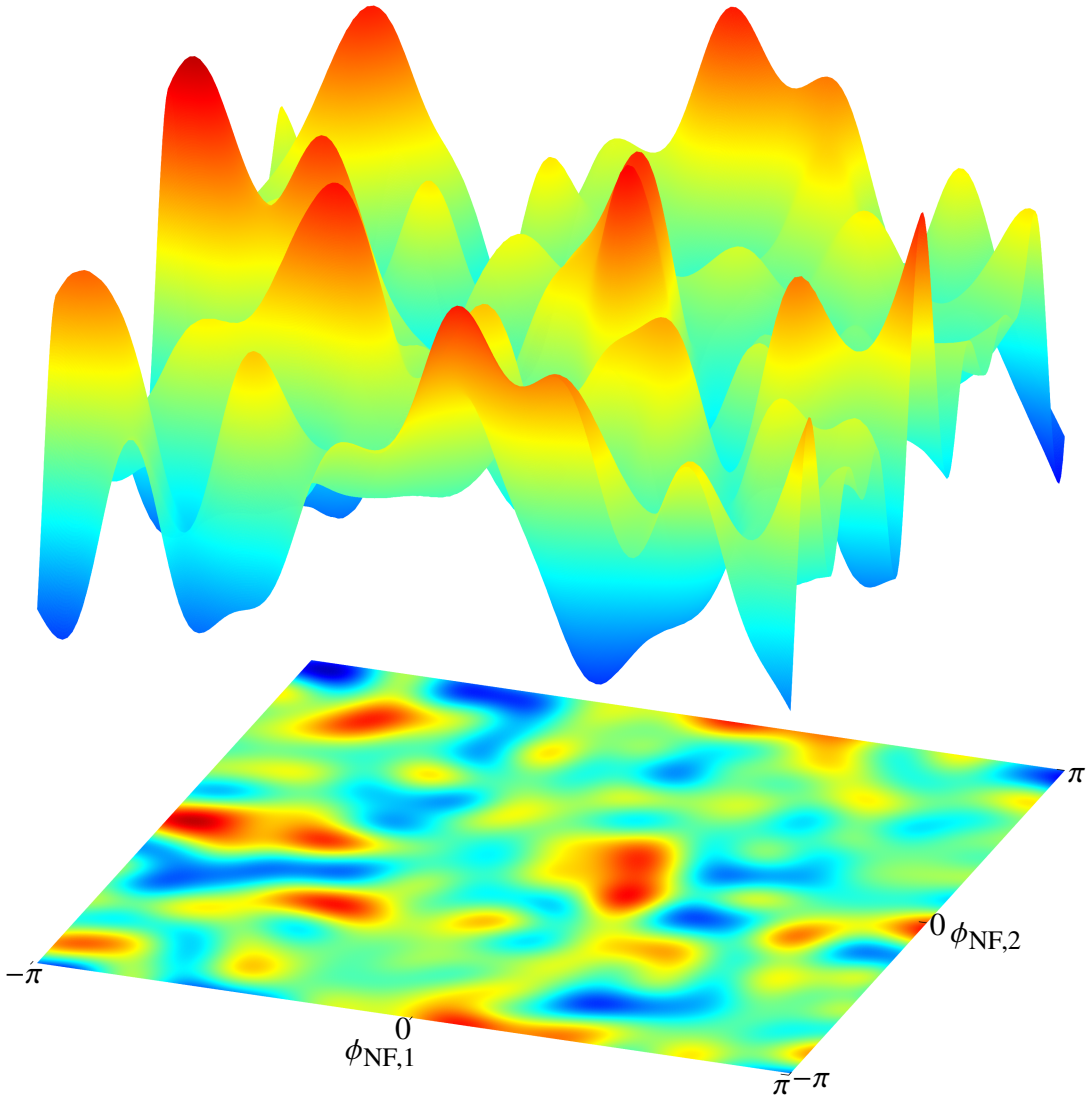

Figure 6.25: Normal form defect landscape of the radial phase space in $\phi_{\mathrm{NF},1}$ and $\phi_{\mathrm{NF},2}$ for fixed normal form amplitudes of $r_{\mathrm{NF},1} = 0.4$ and $r_{\mathrm{NF},2} = 0.4$, and with $\delta p = 0\%$. The underlying map considers an ESQ voltage of 18.3 kV.

in a certain phase space region, which motivates a rigorous analysis of the normal form defect for those phase space regions.

Before we discuss the optimization process and its results, we look at different normal form defect landscapes to emphasize how much the landscapes change in shape and magnitude for different normal form phase space points.

In Fig. 6.26 and Fig. 6.27, the normal form defect landscapes in the vertical and radial direction

are shown for maps considering an ESQ voltage of 18.3 kV and 17.5 kV, respectively.

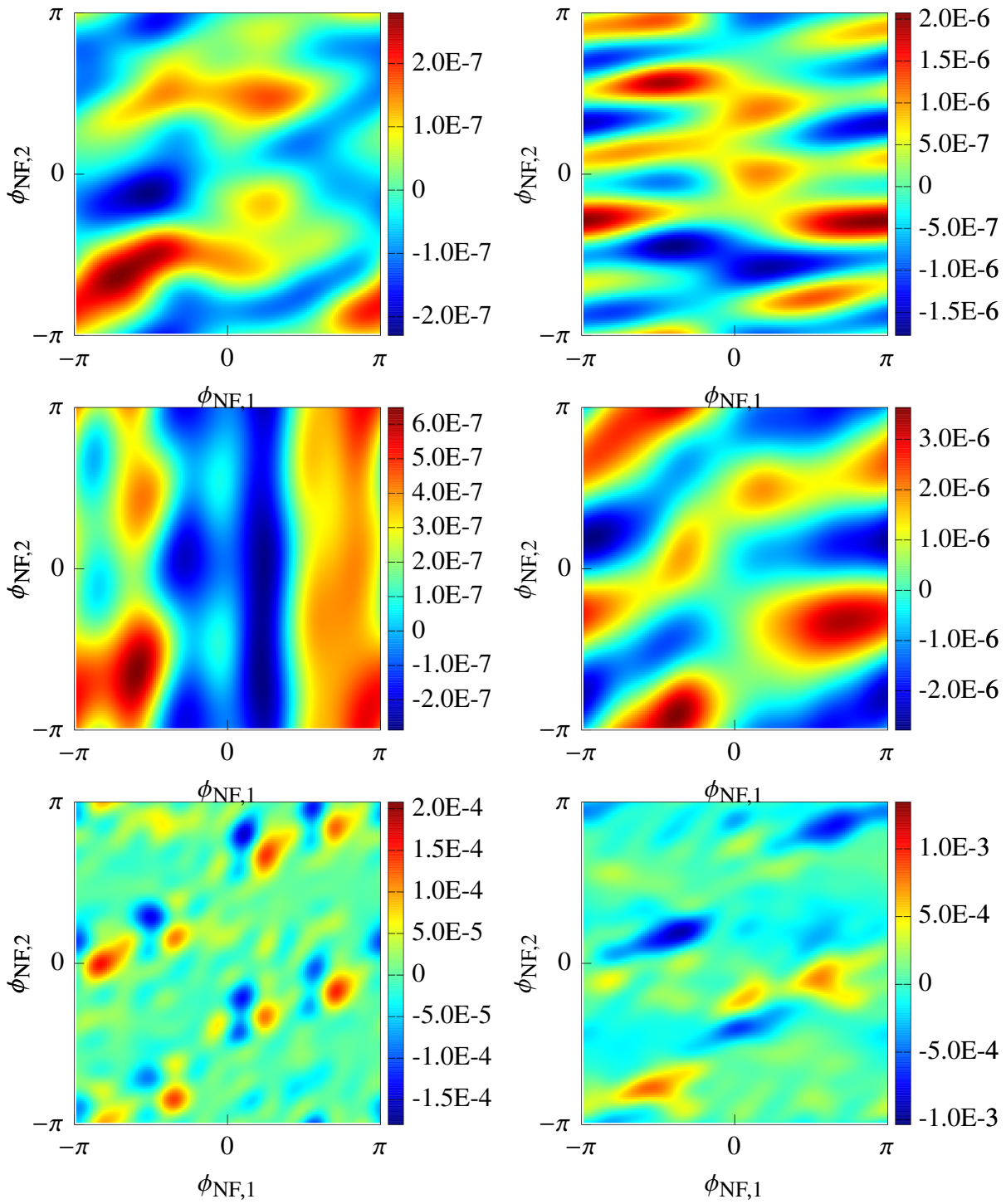

Figure 6.26: The normal form defect landscape of the radial (left) and vertical (right) phase space for multiple onion layers of zero thickness, which are characterized by $(r_{NF,1}, r_{NF,2}, \delta p)$. The top row corresponds to $(0.1, 0.2, 0.24\%)$, the middle row corresponds to $(0.2, 0.05, 0.24\%)$, and the bottom row corresponds to $(0.56, 0.72, 0.04\%)$. The underlying map considers an ESQ voltage of 18.3 kV.

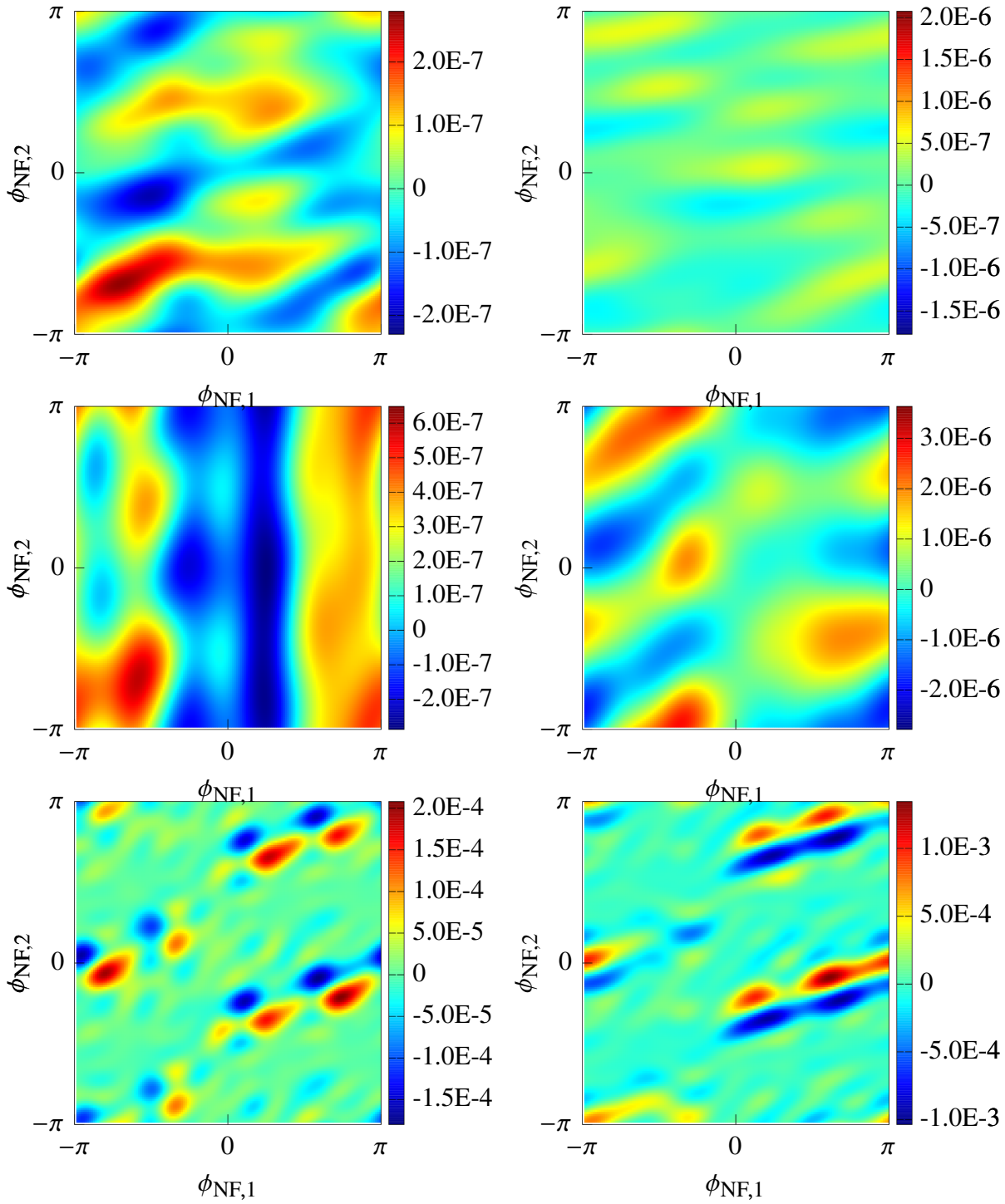

Figure 6.27: The normal form defect landscape of the radial (left) and vertical (right) phase space for multiple onion layers of zero thickness, which are characterized by $(r_{\mathrm{NF},1}, r_{\mathrm{NF},2}, \delta p)$. The top row corresponds to $(0.1, 0.2, 0.24\%)$, the middle row corresponds to $(0.2, 0.05, 0.24\%)$, and the bottom row corresponds to $(0.56, 0.72, 0.04\%)$. The underlying map considers an ESQ voltage of 17.5 kV.

Comparing the normal form defect of the radial and vertical phase space clearly shows the

different orders of magnitude at play for those particular onion layers of zero thickness. The normal form defect of the vertical phase space is about 1.5 orders of magnitude larger than the normal form defect of the radial phase space.

The comparison between Fig. 6.27 and Fig. 6.26 shows something rather fascinating. Even though the normal form defect landscapes change so drastically for different phase space positions, they are very similar for the two maps at the same normal form positions. The magnitude of the normal form defect is usually higher for the 18.3 kV, but the example in the bottom row shows that there are also normal form phase space regions where it is the other way round.

The top row and middle row of Fig. 6.27 and Fig. 6.26 show phase space points with the same momentum offset and roughly the same overall normal form radius. While the magnitude of the normal form defects in the radial and vertical direction is roughly the same, the shape of the normal form defect landscape differs tremendously. For the global optimization, this means that the objective function looks vastly different for each of the onion layer search domains.

6.3.6 The Results of the Verified Global Optimization of the Normal Form Defect

As mentioned in Sec. 6.3.4, we partition the search space into onion layers of the size $0.04 \times 2\pi \times 0.04 \times 2\pi \times 0.04\%$ in $(r_{\mathrm{NF},1}, \phi_{\mathrm{NF},1}, r_{\mathrm{NF},2}, \phi_{\mathrm{NF},2}, \delta p)$. We cover the entire relevant δp space from -0.22% to $+0.42\%$ (see Fig. 5.10), which yields 16 partitions of size 0.04%.

For each of those 16 pieces, we additionally partition the $(r_{\mathrm{NF},1}, r_{\mathrm{NF},2})$ space into boxes of size 0.04×0.04. To determine which of those boxes represent phase space behavior within the collimator region, we probe the bottom left corner of each box, namely, the point with the lowest amplitudes $(r_{\mathrm{NF},1,\mathrm{min}}, r_{\mathrm{NF},2,\mathrm{min}})$ and check if those lowest amplitudes are already outside the collimator region in the original phase space coordinates. For the probing, we take $30 \times 30 \times 2$ testing points in $\phi_{\mathrm{NF},1}, \phi_{\mathrm{NF},2}, \delta p$ and map them back into the original phase space (x, a, y, b) using the inverse normal form transformation $\mathcal{A}^{-1}$. A box is only analyzed if all of the 1800 probing points satisfy $\sqrt{x^2 + y^2} < 0.045$ mm.

To benchmark the verified analysis, we also present a nonverified normal form defect analysis of

the same onion layers. The nonverified analysis is based on probing using 3600 well-chosen probing points for each onion layer. This probing approach is used in a verified form as a method to obtain a good cutoff value for the global optimizer. Accordingly, the nonverified analysis provides a lower bound on the maximum normal form defect, while the verified analysis constitutes an upper bound.

The results on the following pages (see Fig. 6.28 to Fig. 6.31) are ordered such that the nonverified probing analysis can be compared to the verified global optimization by switching back and forth between pages. Additionally, the two verified normal form defect analyses for the map with an ESQ voltage of 18.3 kV and 17.5 kV can be compared the same way.

The tune shifts of the 17.5 kV map in Fig. 6.32 to Fig. 6.34 are of similar magnitude and complexity as the tune shifts of the 18.3 kV map in Fig. 5.7 to Fig. 5.9. However, their absolute values are in lower vertical tune ranges and therefore further away from the vertical low-order $1/3$ resonance tune. Even under the combined influence of both the radial and vertical amplitude, as well as the momentum offset, none of the tunes of the 17.5 kV map cross the vertical $1/3$ resonance tune. In contrast, almost for every momentum offset there is a combination of radial and vertical amplitudes that crosses the vertical $1/3$ resonance tune for the 18.3 kV map.

Accordingly, both the tune analysis as well as the normal form defect analysis could show that the map with an ESQ voltage of 18.3 kV yields more potential divergence and instability.

The inner white onion layers have a maximum normal form defect below 10^{-5}. Accordingly, even in the worst case, it takes at least 4000 turns to cross the respective onion layer. It takes at least 400 turns to cross a yellow onion layer by the same measure and at least 40 turns to cross an orange onion layer. Red onion layers take at least 12 turns to cross, and black onion layers can, in the worst case, be crossed in fewer turns.

However, those turn numbers are a verified underestimation of the minimum number of turns it takes to cross a respective onion layer. The estimation assumes that the maximum normal form defect of the onion layer is encountered in every turn. To put this underestimation in perspective we take a look at the island patterns from Fig. 5.30, which only differ in their vertical amplitude ($r_{\mathrm{NF},1} \approx 0, \delta p = 0.126\%$). Out of the five island sizes, we consider the largest, the smallest, and the

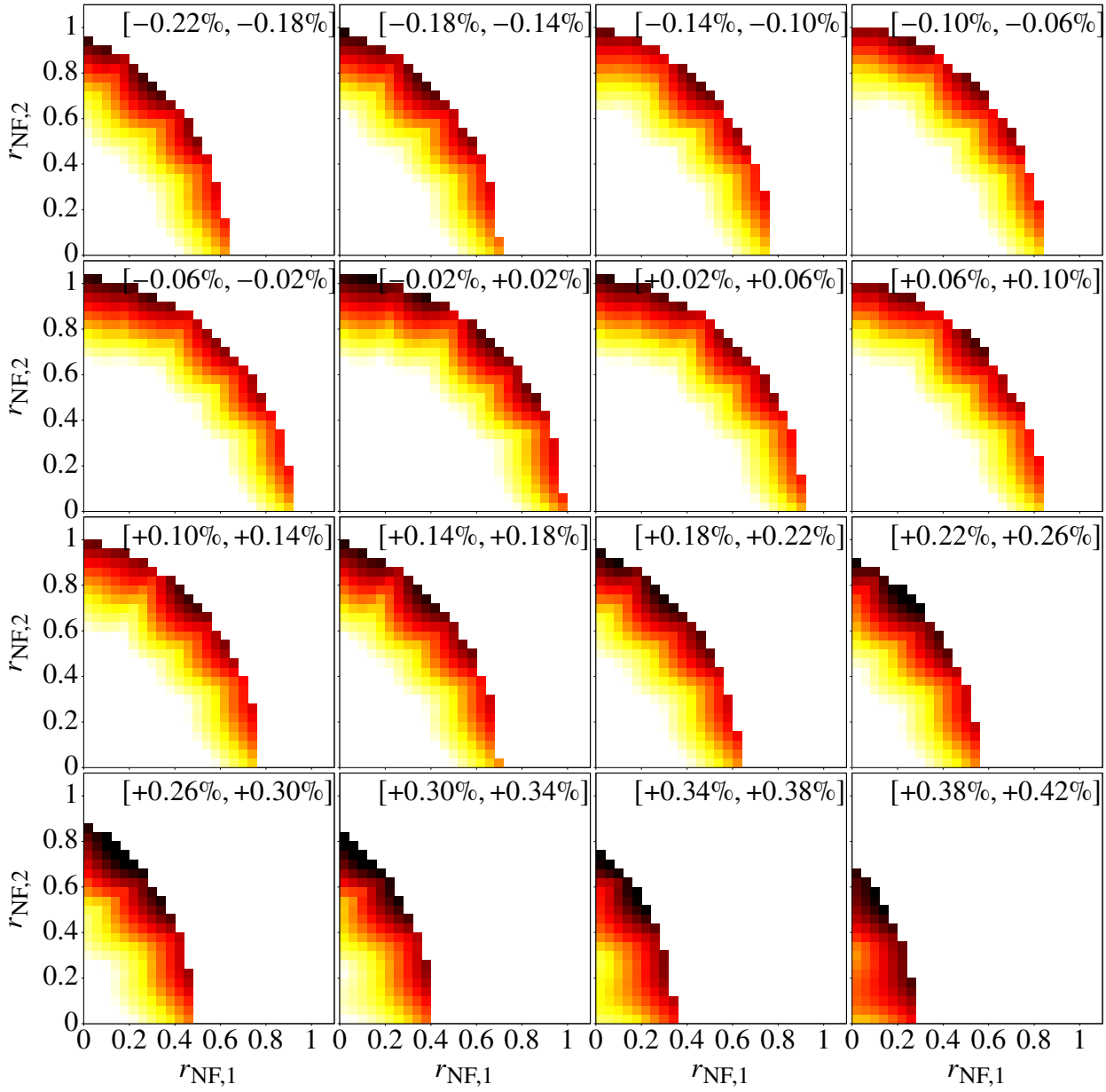

Figure 6.28: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 17.5 kV. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

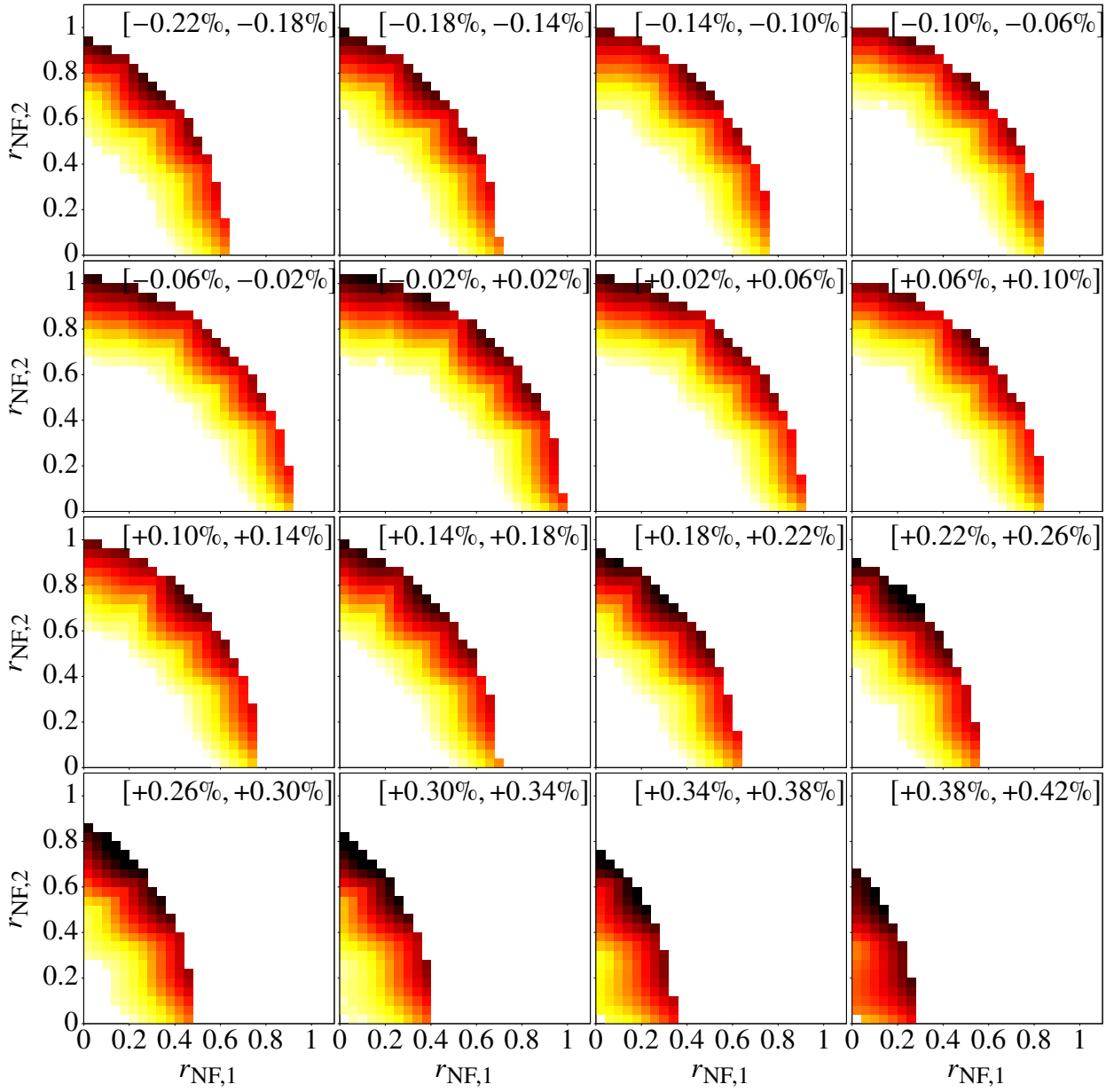

Figure 6.29: Verified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 17.5 kV. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

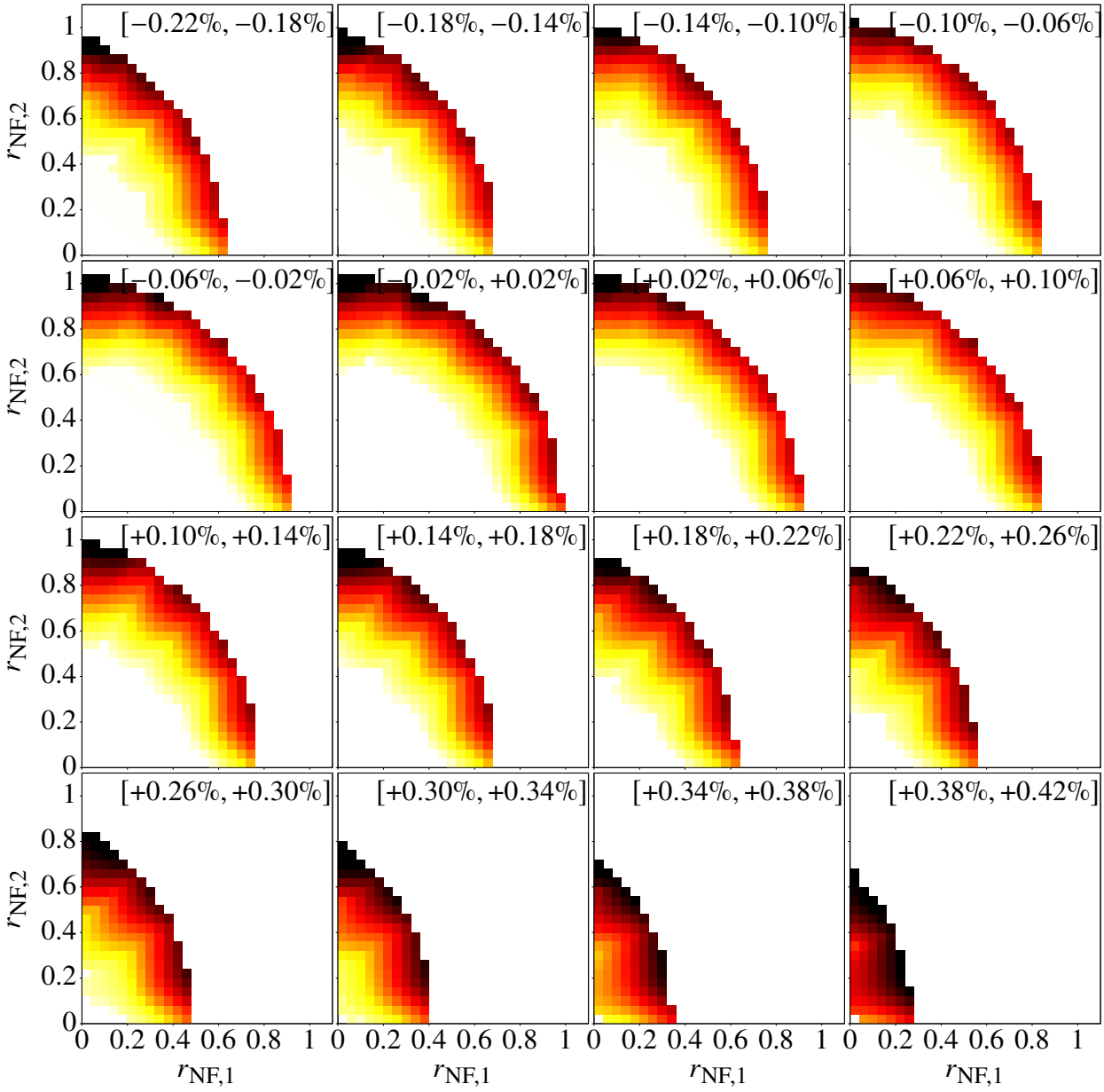

Figure 6.30: Verified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

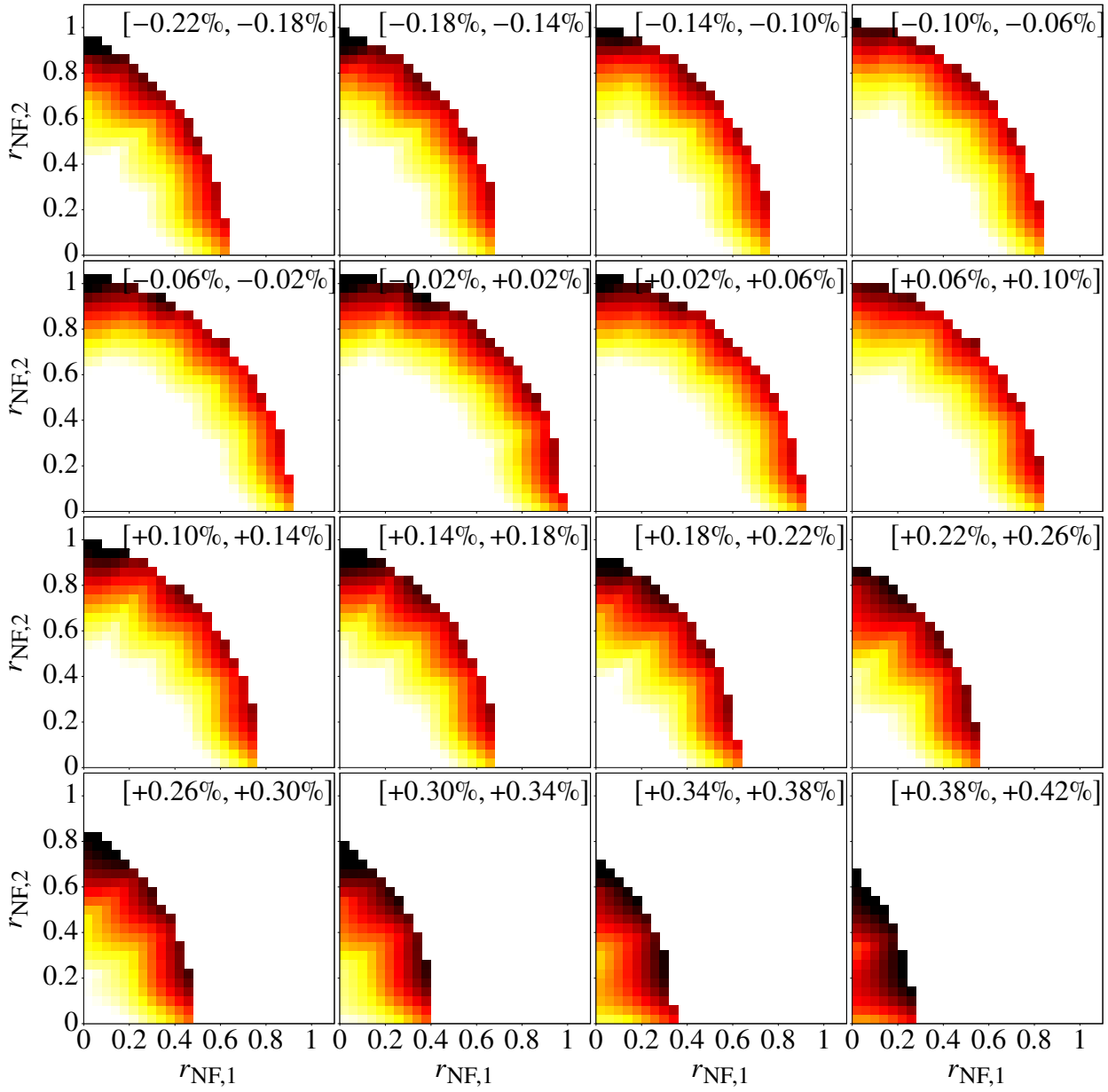

Figure 6.31: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV. The individual plots show different momentum ranges which are clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layers corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

third largest/smallest referred to as the medium sized one.

The smallest island is very close to the period-3 fixed point. Its normal form radius only varies slightly between 0.925 and 0.932. It takes roughly 250 turns to get from the lowest normal form radius to the largest and another 250 turns to get back. This corresponds to an average normal form defect of 2.8E-5. The corresponding verified analysis of the respective onion layer yields a maximum normal form defect of 8.8E-3, which is roughly larger by a factor of 300.

The medium sized island varies between normal form radii of 0.820 and 1.028 every roughly 300 turns, which corresponds to an average normal form defect of 6.9E-4. The corresponding verified analysis of the respective onion layers yield maximum normal form defects between 1.3E-3 and 1.6E-2, which are roughly larger by a factor of two to 23.

The largest sized island varies between normal form radii 0.773 and 1.066 every roughly 660 turns, which corresponds to an average normal form defect of 4.4E-4. The verified analysis guarantees that the onion layer $[0.8, 0.84]$ in $r_{\mathrm{NF},2}$ can not be crossed in less than 29.7 turns. the largest island crosses this onion layer in 301 turns. For the onion layer $[0.84, 0.88]$ in $r_{\mathrm{NF},2}$, the verified analysis predicts it can not be crossed in less than 15.7 turns. The largest island crosses it in 104 turns. The onion layer $[0.88, 0.92]$ in $r_{\mathrm{NF},2}$ is crossed in 36 turns while the verified analysis predicts a minimum of 8.3 turns. For the onion layer $[0.92, 0.96]$ in $r_{\mathrm{NF},2}$, it is 23 turns which is more than the predicted minimum of 4.5 turns, and for the onion layer $[0.96, 1]$ in $r_{\mathrm{NF},2}$, it takes the large island 20 turns to cross, while the verified analysis predicts a minimum of 2.5 turns.

As we saw in this analysis, the dynamics in a single onion layer can vary significantly. Some orbits remain in an onion layer indefinitely, like the smallest island. In contrast, others are transported through it with sometimes less than a factor ten between the worst case divergence and the actual rate of divergence. However, for those island patterns, the same onion layer does not only transport the particle outward but also transports it back inwards at the same rate. In short, it is possible to relate the quantitative aspects of the normal form defect analysis to the actual dynamics within the onion layer, in particular, the potential worst case dynamics.

The global normal form defect analysis is also very powerful for the qualitative stability analysis

of different storage ring configurations. A comparison of Fig. 6.29 and Fig. 6.30 yields obvious differences between the verified normal form defect of the map with an ESQ voltage of 17.5 kV and the map with an ESQ voltage of 18.3 kV. There are clearly more diverging regions with a larger maximum normal form defect for 18.3 kV in Fig. 6.30 than there are for the 17.5 kV map in Fig. 6.29.

Note that the individual onion layers of the 18.3 kV map in Fig. 6.30 and the 17.5 kV map in Fig. 6.29 do not necessarily correspond to the same phase space regions in the (x, a, y, b) phase space. Because we calculate a normal form transformation for each map, the representation of the relevant (x, a, y, b) phase space in normal form space can be slightly different for the two maps. However, each of the 16 plots show the exact same viable (x, a, y, b) phase space in the normal form coordinates just with a slightly different scaling in $r_{\mathrm{NF},1}$ and $r_{\mathrm{NF},2}$. Accordingly, comparing the color distributions for each of the 16 plots between the two maps is a valid measure to compare the stability of the two storage ring configurations.

As already discussed in Chapter 5, the vertical $1/3$-resonance tune and its associated period-3 fixed point structures for the simulation using an ESQ voltage of 18.3 kV were a major loss and instability factor. For the map with an ESQ voltage of 17.5 kV, the tunes a further away from the vertical $1/3$-resonance tune as Fig. 6.32 to Fig. 6.34 illustrate.

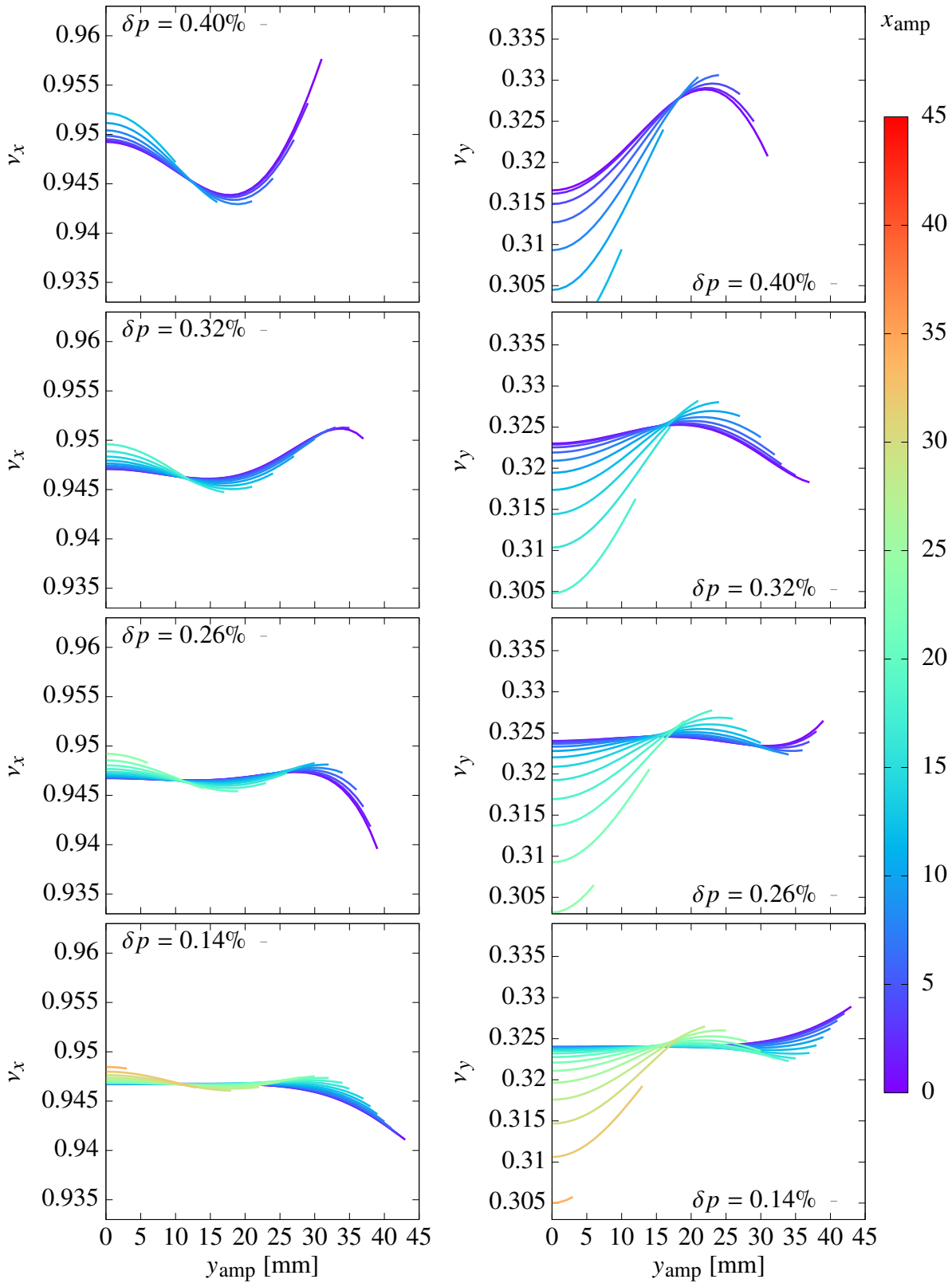

Figure 6.32: Behavior of combined amplitude dependent tune shifts at multiple momentum offsets.

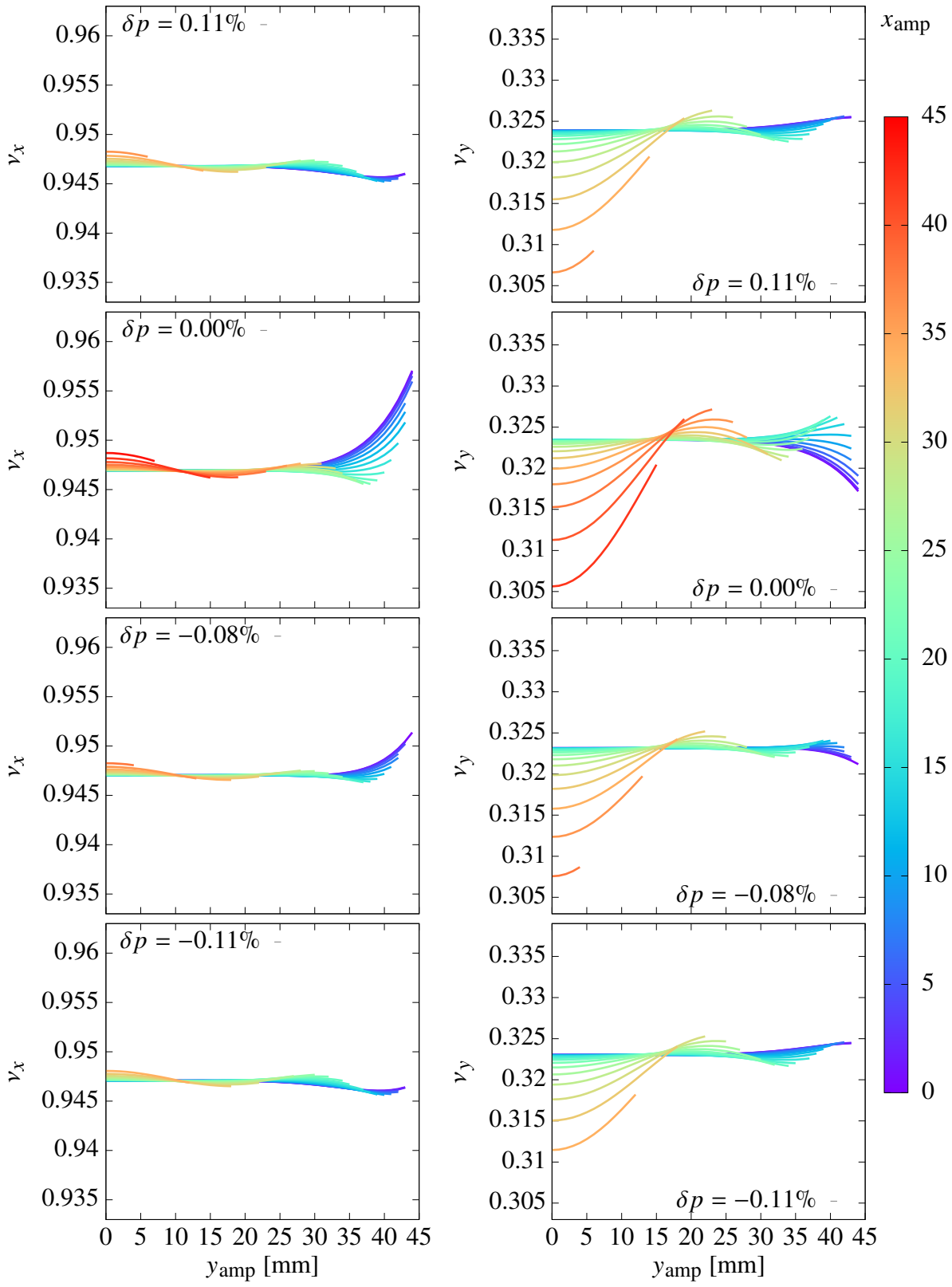

Figure 6.33: Behavior of combined amplitude dependent tune shifts at multiple momentum offsets.

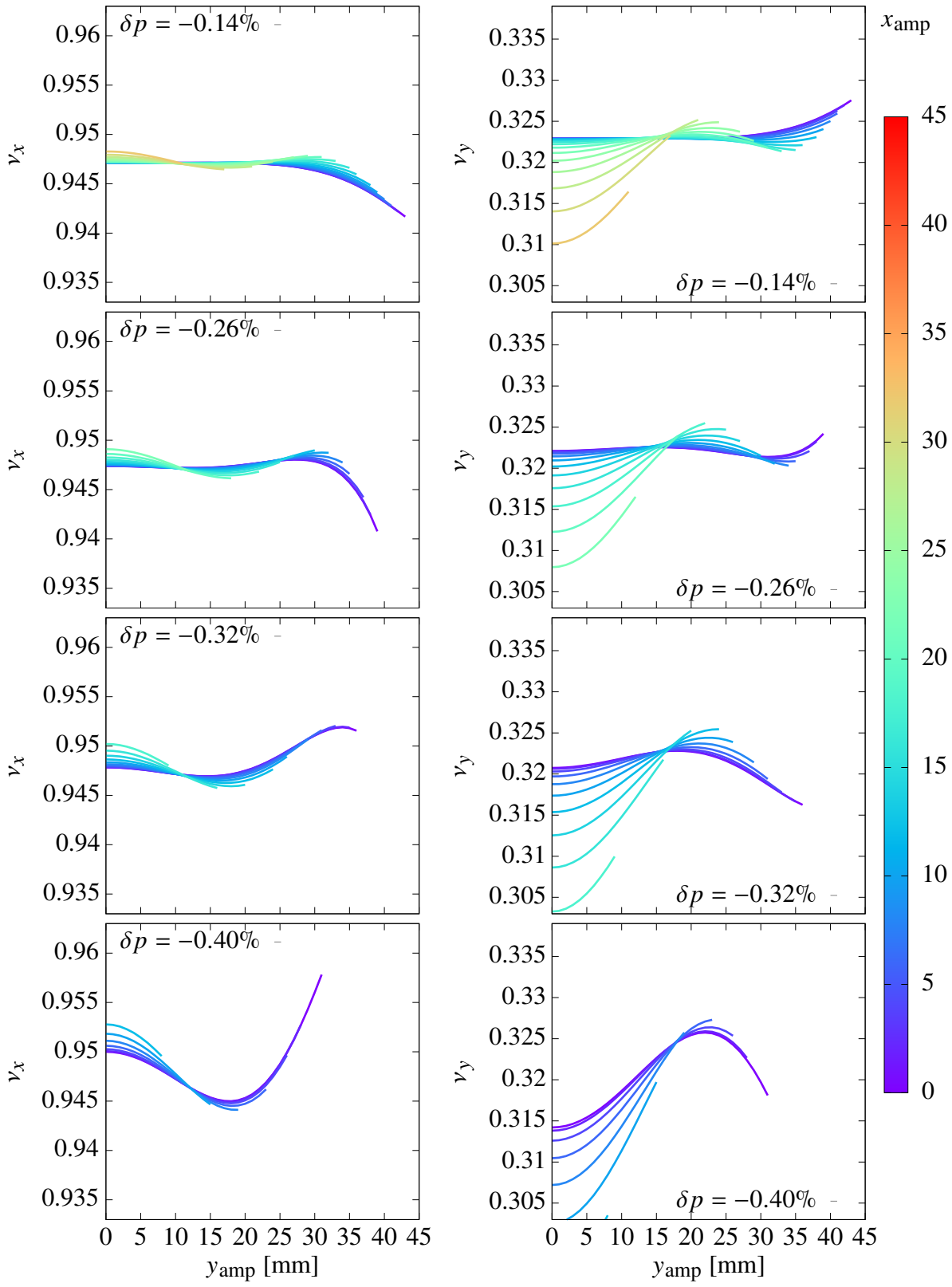

Figure 6.34: Behavior of combined amplitude dependent tune shifts at multiple momentum offsets.

6.3.7 Comparison of Nonverified and Verified Normal Form Defect Analysis

Fig. 6.28 and Fig. 6.29 respectively show the nonverified and verified analysis for the map with an ESQ voltage of 17.5 kV, while Fig. 6.30 and Fig. 6.31 respectively show the verified and the nonverified analysis for the map with an ESQ voltage of 18.3 kV.

The differences between the nonverified and verified computations are small but visible if one switches back and forth between the pages. To emphasize the differences between the verifed and nonverified computations onion layer by onion layer, Fig. 6.36 and Fig. 6.35 illustrate those differences for 17.5 kV and 18.3 kV, respectively. The differences show the importance of a verified method to capture each onion layer's maximum normal form defect, especially for the more diverging regions.

To show that this difference is not just an artifact of back bounding by the global optimizer, Fig. 6.37 and Fig. 6.38 illustrates the difference between the upper and the lower bound bound on the maximum normal form defect for the 17.5 kV map and 18.3 kV map, respectively. Because they only consists of white boxes, those differences are all smaller than 1E-5 and therefore do not influence the calculation of the differences between the nonverified and verified evaluation.

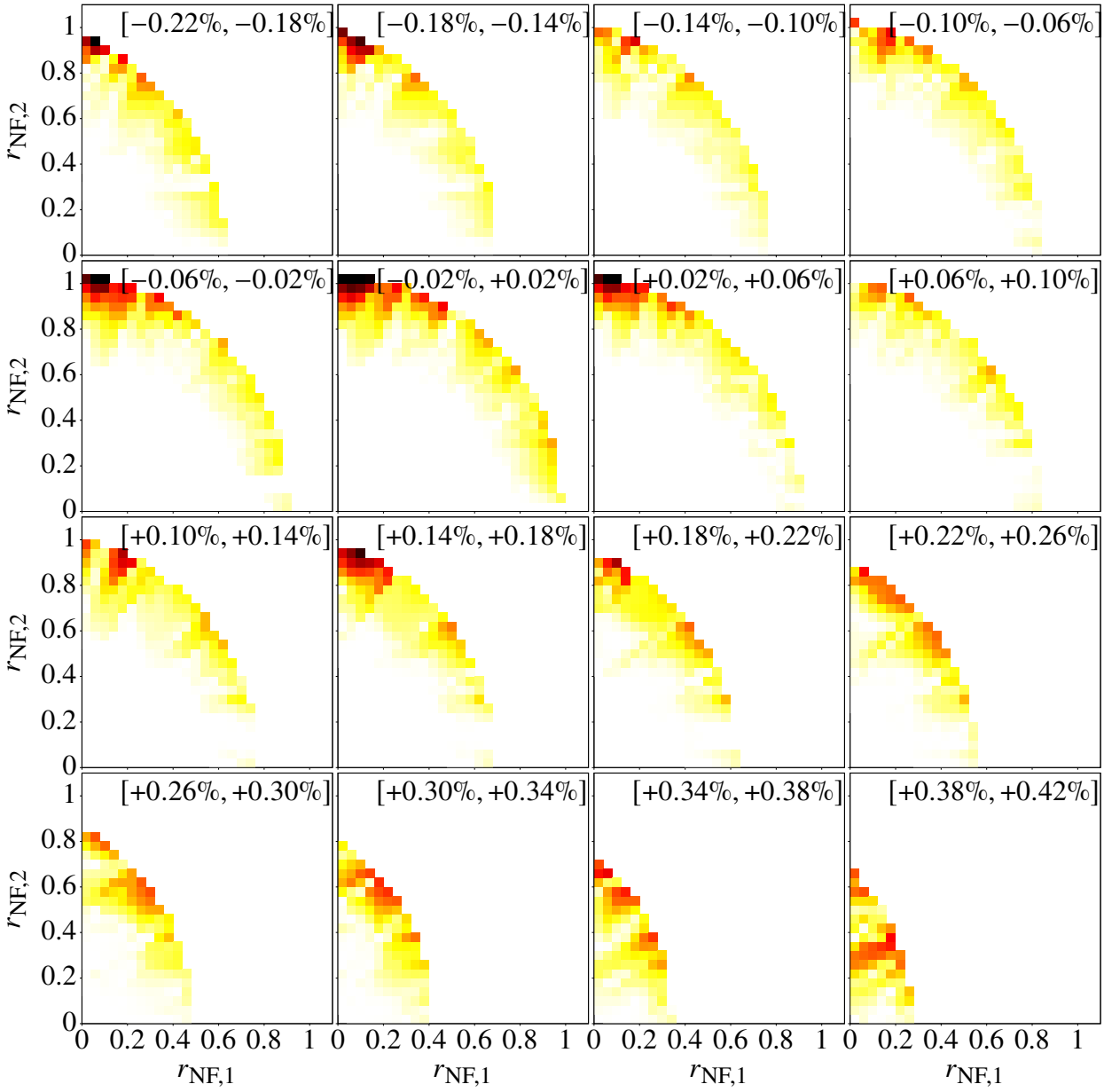

Figure 6.35: Difference between verified normal form defect analysis and nonverified normal form defect analysis for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the difference of the evaluated normal form defects of the specific onion layer. The white boxes for lower normal form radii indicate a difference below 10^{-5}. The yellow boxes denote differences up to 10^{-4}. The orange boxes correspond to differences up to 10^{-3}. The red boxes denote differences up to $10^{-2.5}$, and the black boxes indicate differences larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

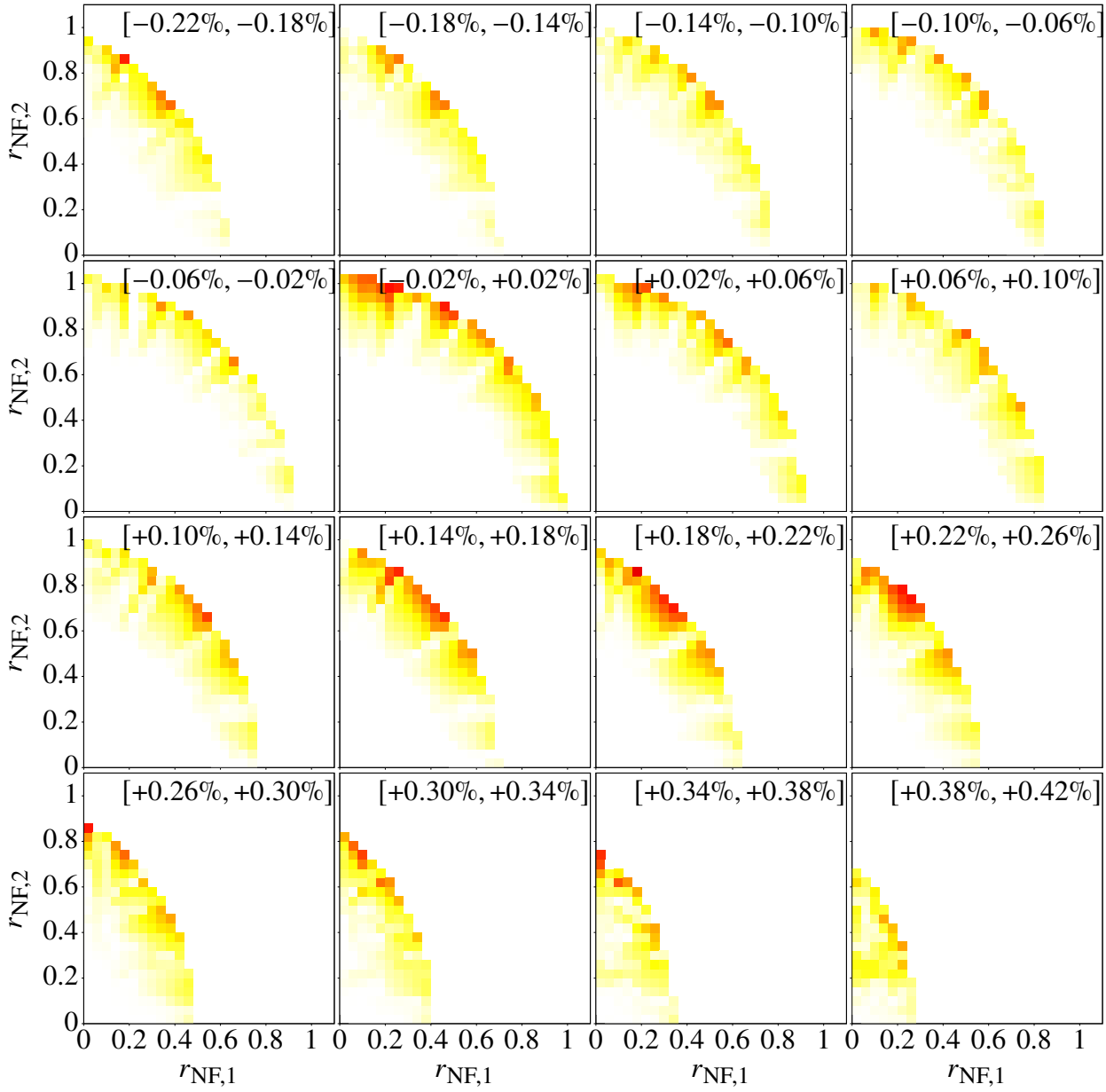

Figure 6.36: Difference between verified normal form defect analysis and nonverified normal form defect analysis for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 17.5 kV. The individual plots show different momentum ranges which are clarified by the label at the top of each graph. The color scheme corresponds to the difference of the evaluated normal form defects of the specific onion layer. The white boxes for lower normal form radii indicate a difference below 10^{-5}. The yellow boxes denote difference up to 10^{-4}, the orange boxes correspond to a differences up to 10^{-3}, the red boxes denote differences up to $10^{-2.5}$ and the black boxes indicate differences larger than that. Each onion layers corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

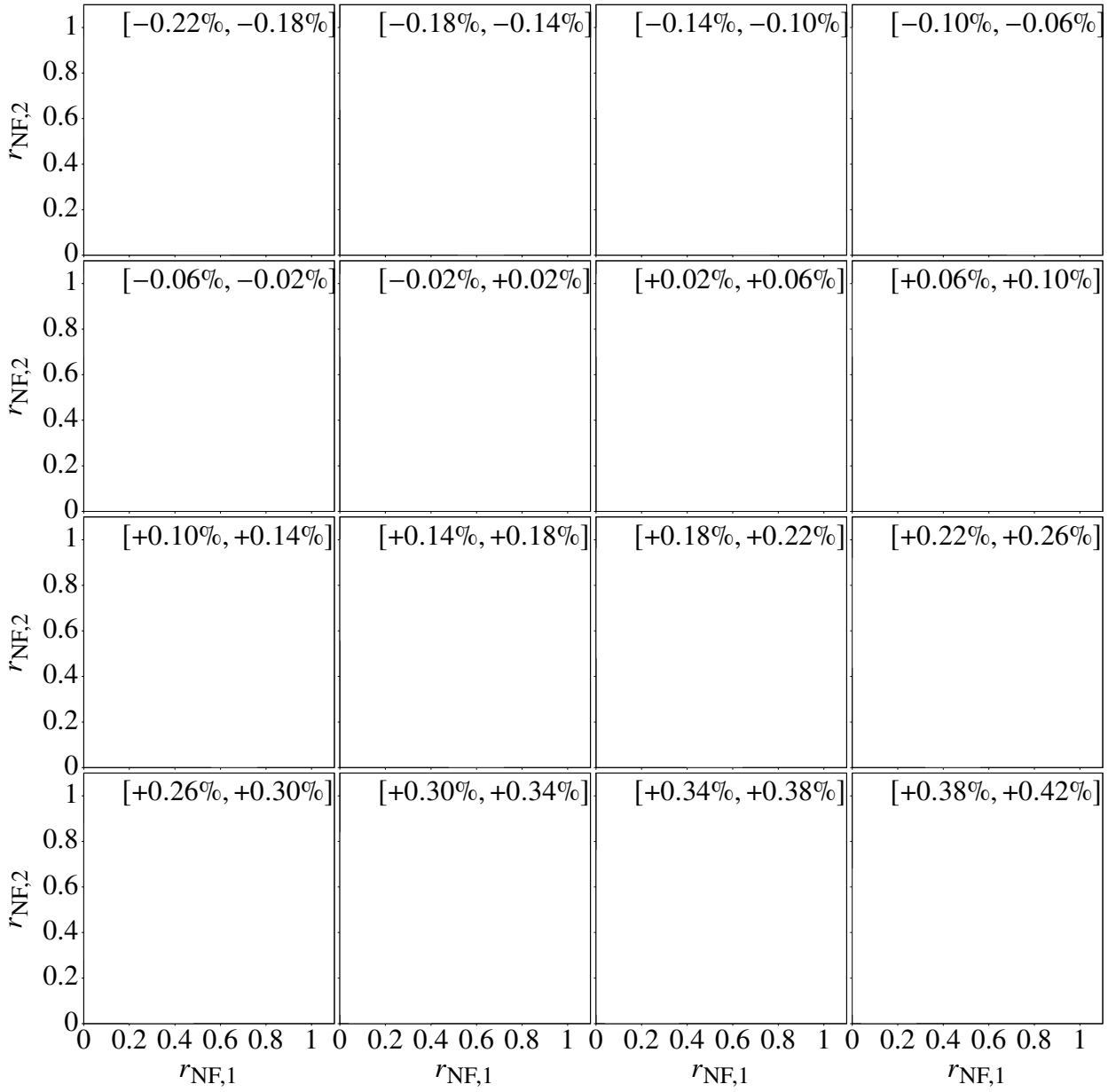

Figure 6.37: Difference between the rigorously guaranteed upper bound and the lower bound of the maximum normal form defect using Taylor Model based verified global optimization. The analysis is for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 17.5 kV. The individual plots show different momentum ranges which are clarified by the label at the top of each graph. The color scheme corresponds to the difference between the upper bound and the lower bound of the maximum normal form defect of the specific onion layer. The white boxes indicate a difference below 10^{-5}. Each onion layers corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

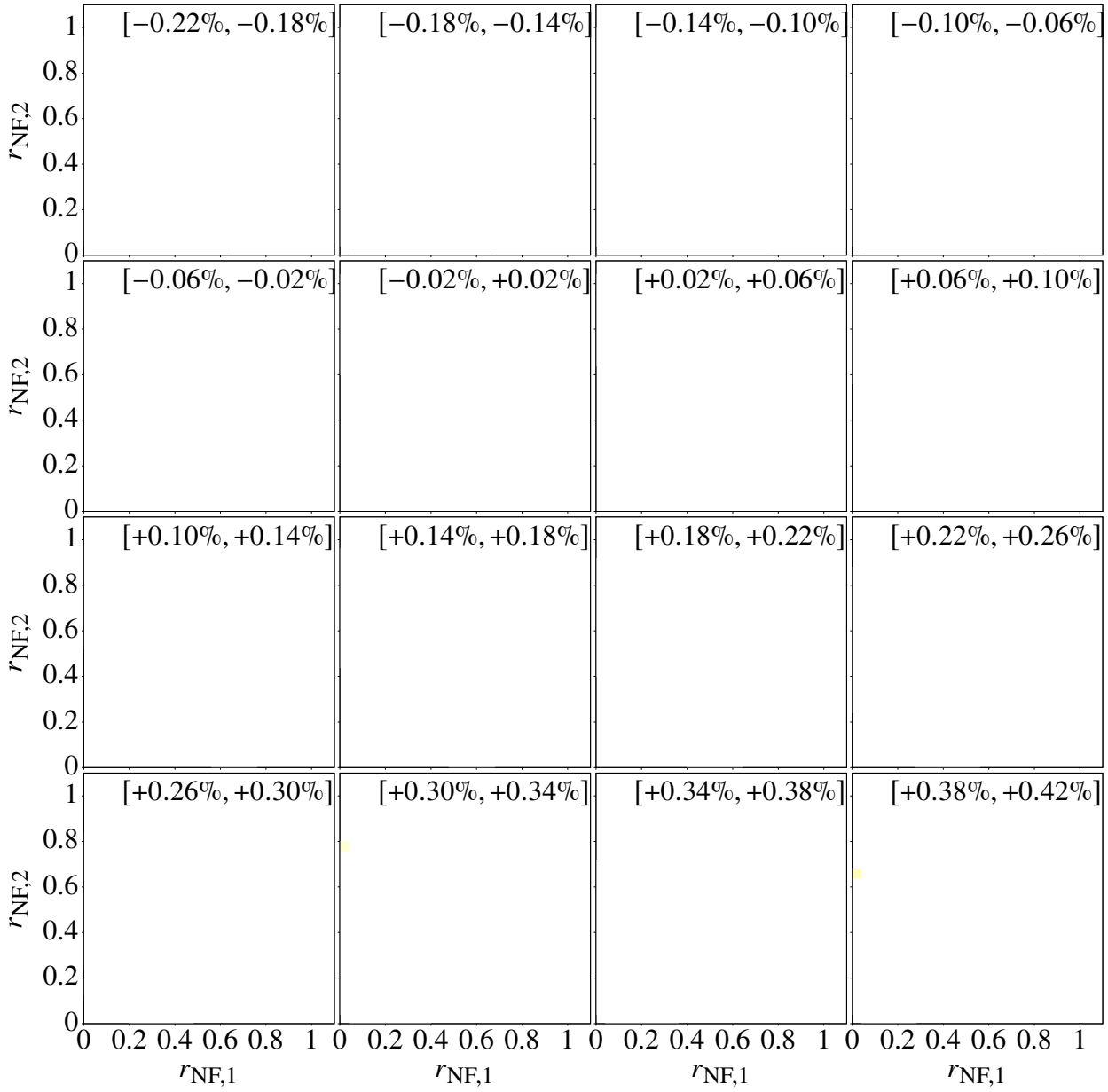

Figure 6.38: Difference between the rigorously guaranteed upper bound and the lower bound of the maximum normal form defect using Taylor Model based verified global optimization. The analysis is for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV. The individual plots show different momentum ranges which are clarified by the label at the top of each graph. The color scheme corresponds to the difference between the upper bound and the lower bound of the maximum normal form defect of the specific onion layer. The white boxes indicate a difference below 10^{-5}. Each onion layers corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

6.3.8 The Analysis of the Effect of Normal Form Transformations of Different Order on the Normal Form Defect

We use the normal form transformation as a function that provides pseudo-invariants of the motion, i.e., the normal form radii. By using the normal form transformation up to different orders, we can analyze the influence of the respective map orders on the dynamics of the system. In Fig. 6.39 to Fig. 6.31, the nonverified normal form defect analysis is performed for the tenth order map with a ESQ voltage of 18.3 kV using normal form transformations from order one to order ten.

The normal form defect pictures for a normal form transformation of order five, six, and seven look identical even when carefully switching between pages. The larges improvement occurs with the ninth order normal form transformation because it captures large parts of the strong ninth order nonlinearities of the map caused by the 20th-pole of the ESQ potential.

To further analyze if the tenth order map does indeed capture most of the relevant dynamics, we produce an eleventh order map and calculate its normal form defect using the tenth order normal form transformation (see Fig. 6.49). This kind of order increasing analysis is known from nonverified integrators with step size control. Compared to the tenth order map evaluation with the tenth order normal form transformation in Fig. 6.31, the eleventh order of the map leads to no visible difference, which is a good sign and suggests that a tenth order map is sufficient to capture the critical dynamics. However, this heuristic approach can not guarantee that even higher order maps would also not yield a significant change. To capture this uncertainty of unknown higher order terms a verified map is required that includes all higher order errors in its Taylor Model remainder bound.

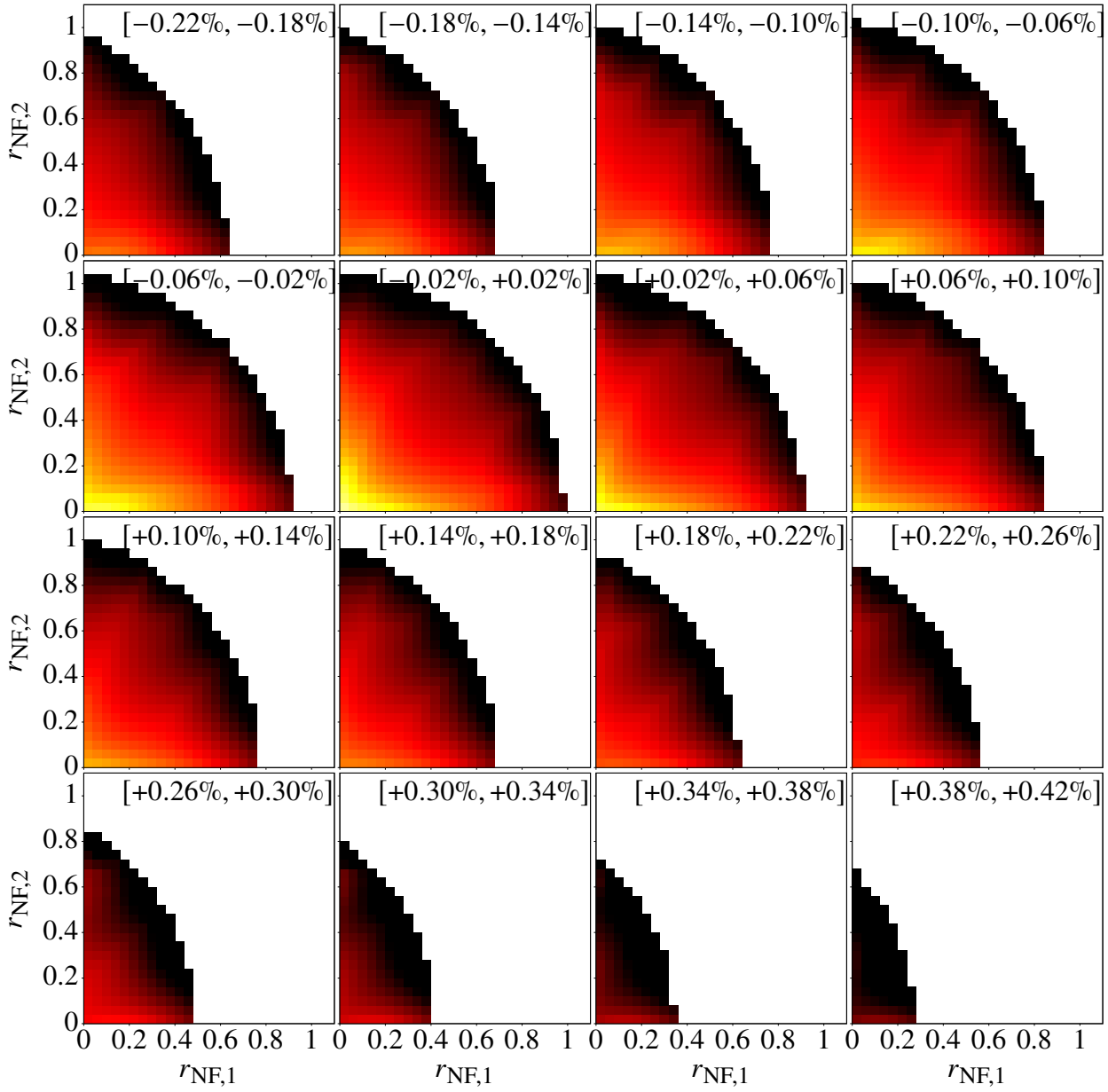

Figure 6.39: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to order 1 instead of the full tenth order. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

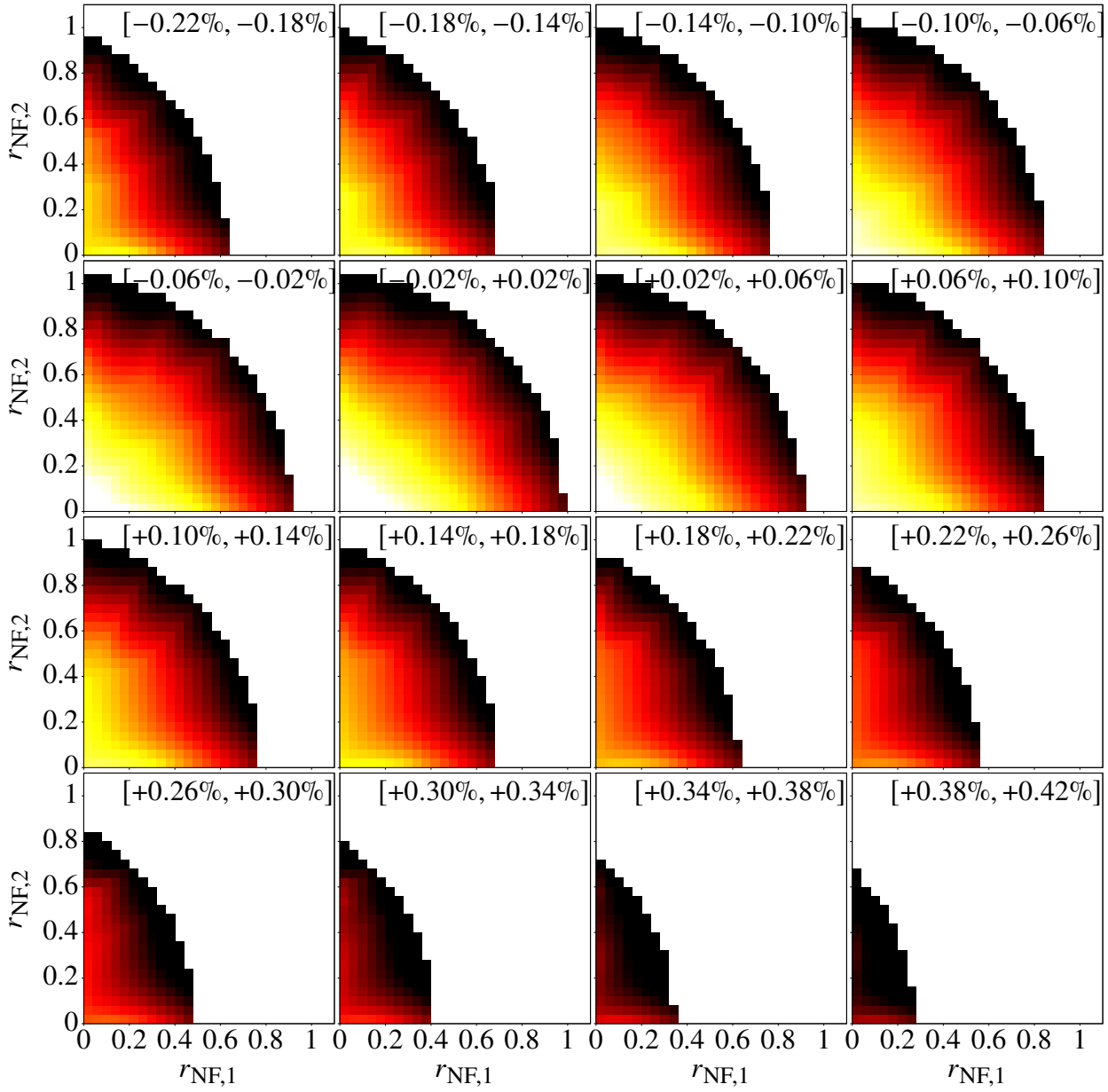

Figure 6.40: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to order 2 instead of the full tenth order. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

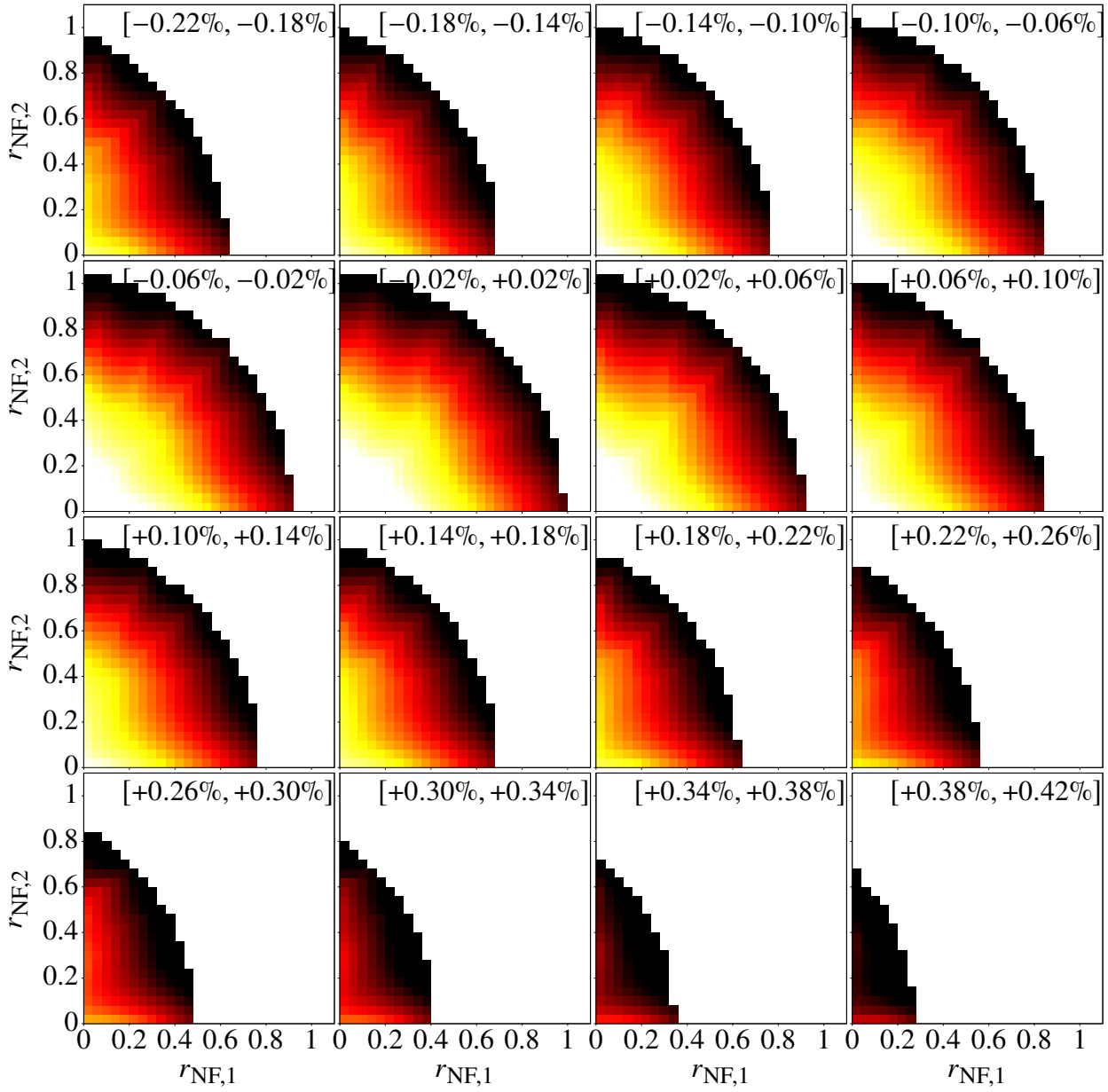

Figure 6.41: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to order 3 instead of the full tenth order. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

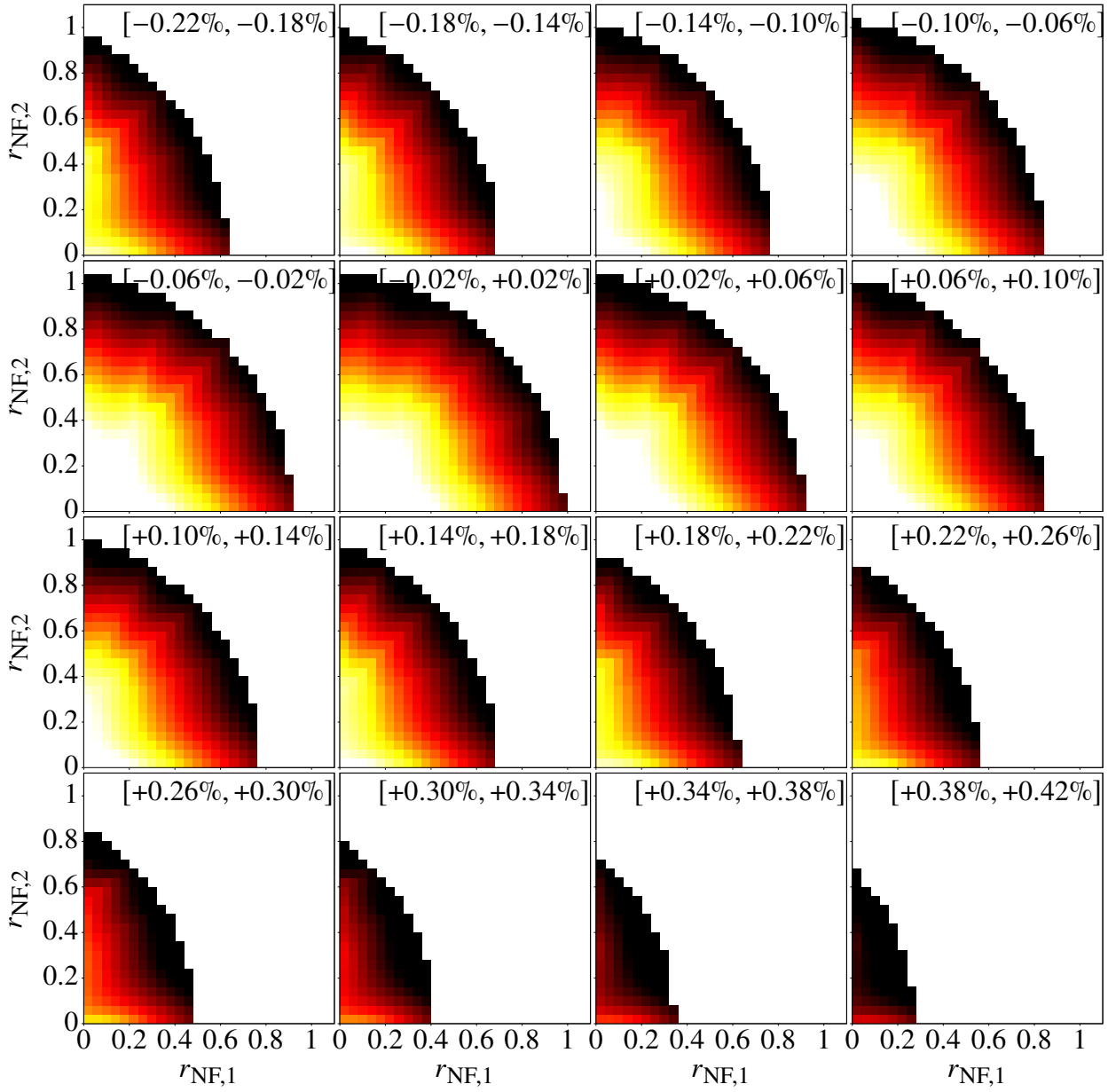

Figure 6.42: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to order 4 instead of the full tenth order. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

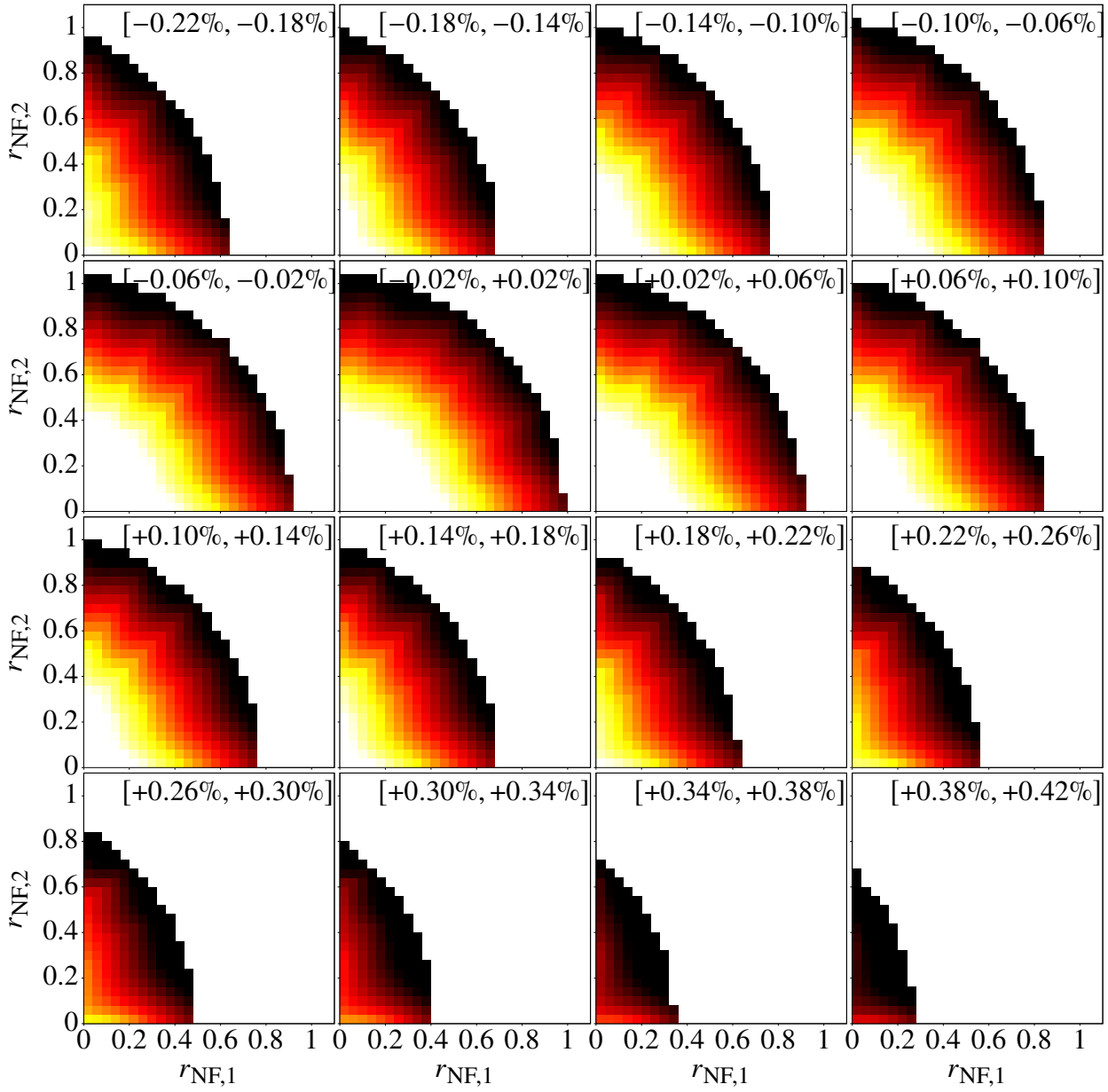

Figure 6.43: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to order 5 instead of the full tenth order. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

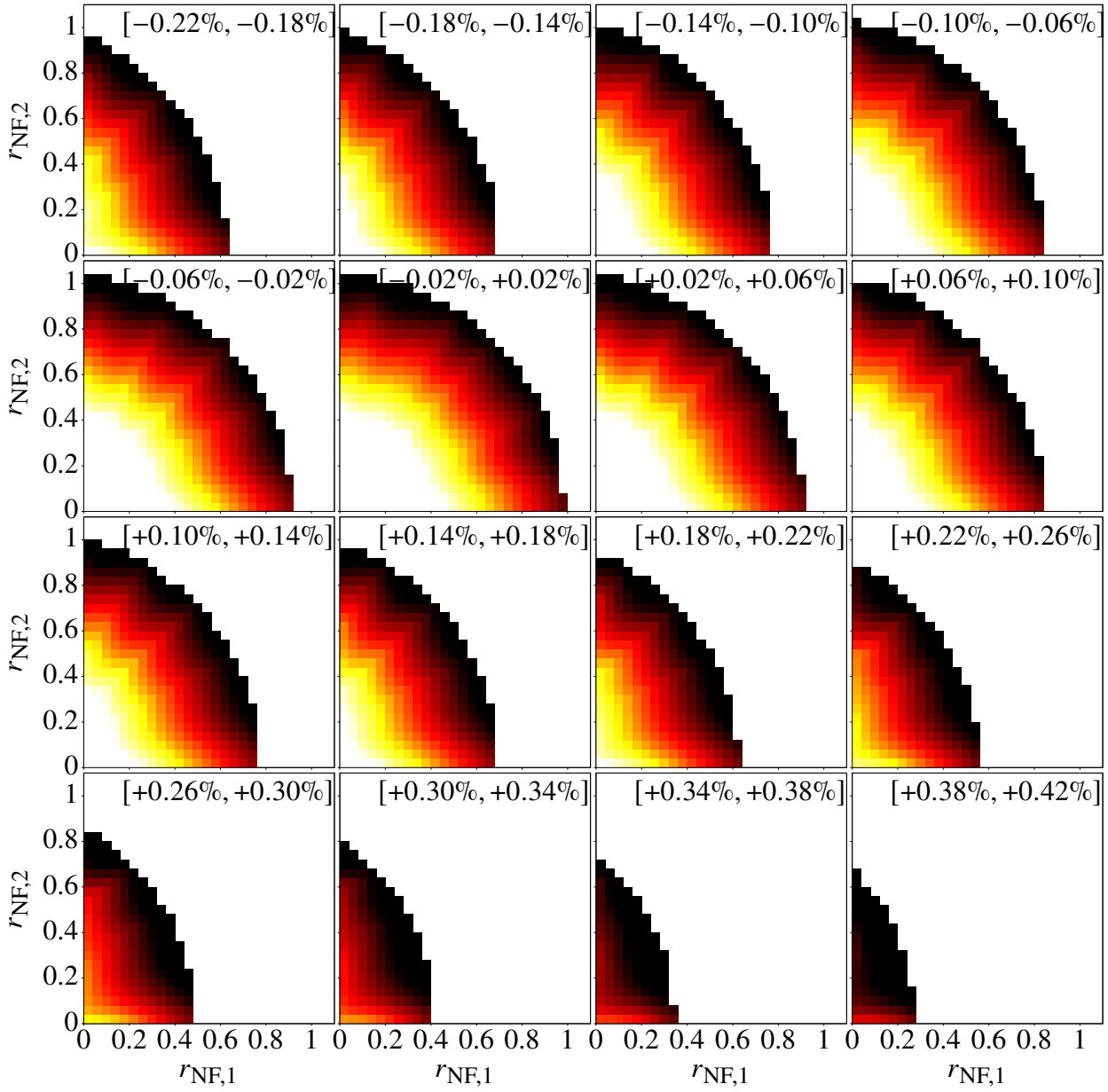

Figure 6.44: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to order 6 instead of the full tenth order. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

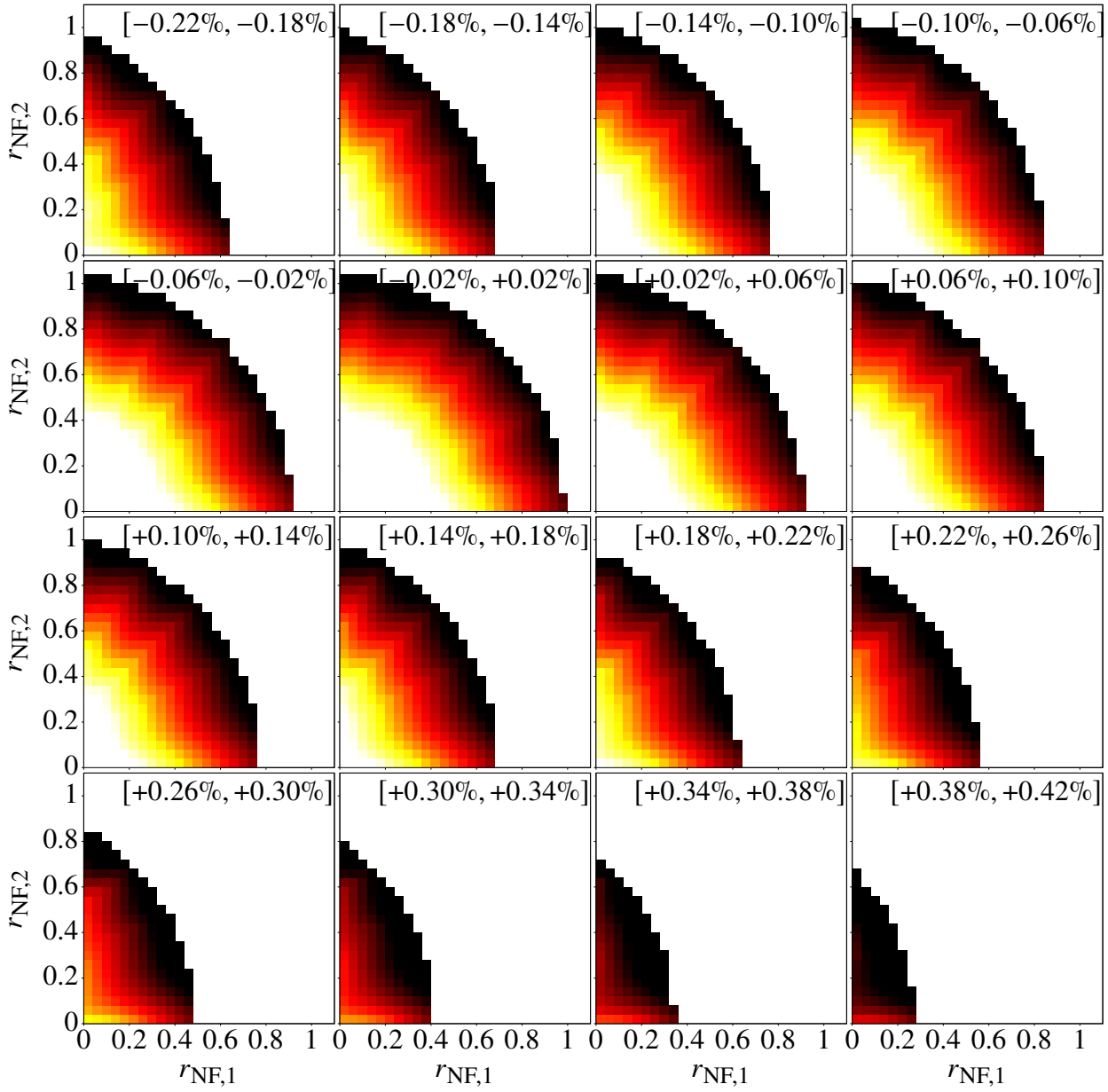

Figure 6.45: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to order 7 instead of the full tenth order. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

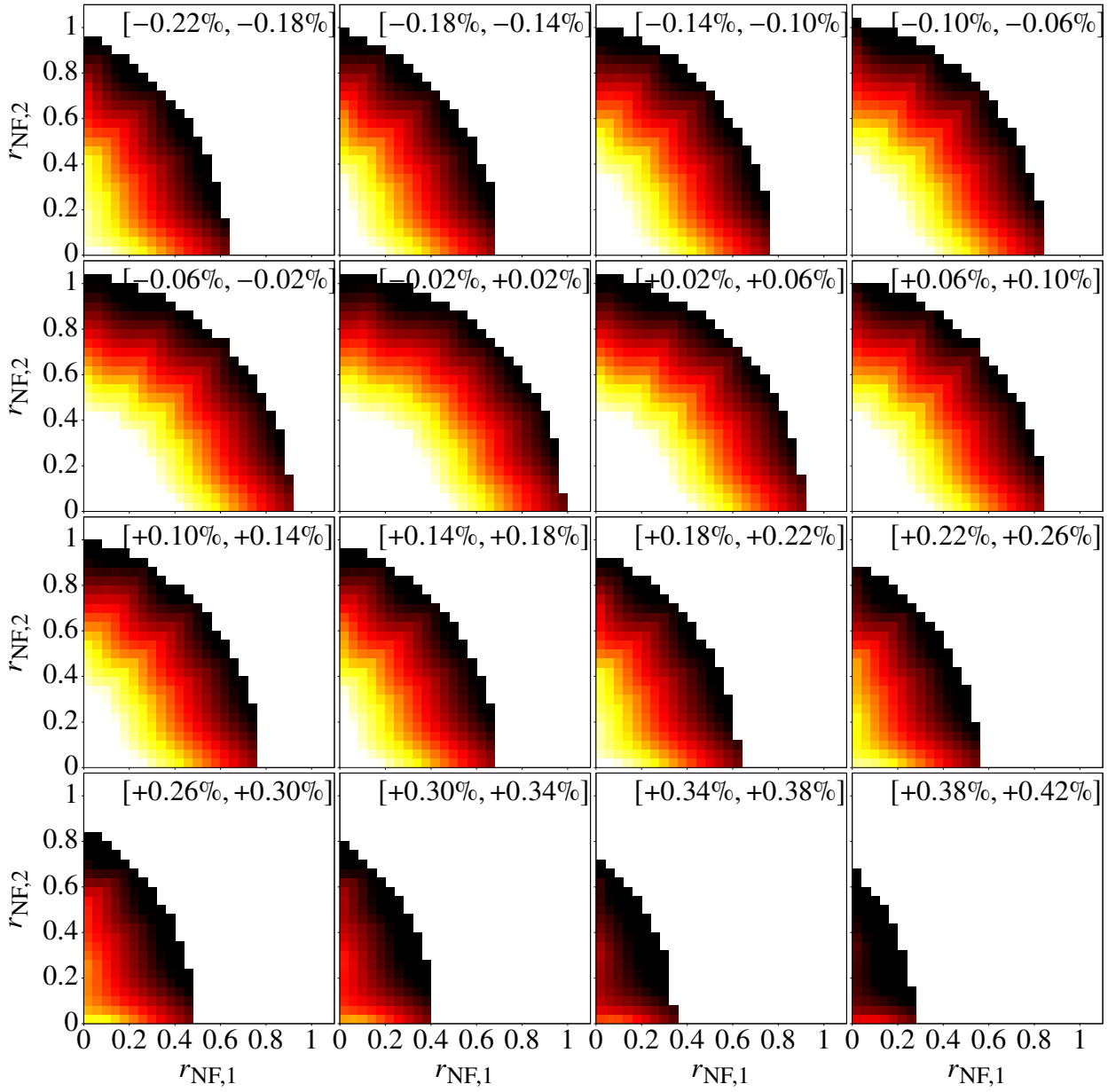

Figure 6.46: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to order 8 instead of the full tenth order. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

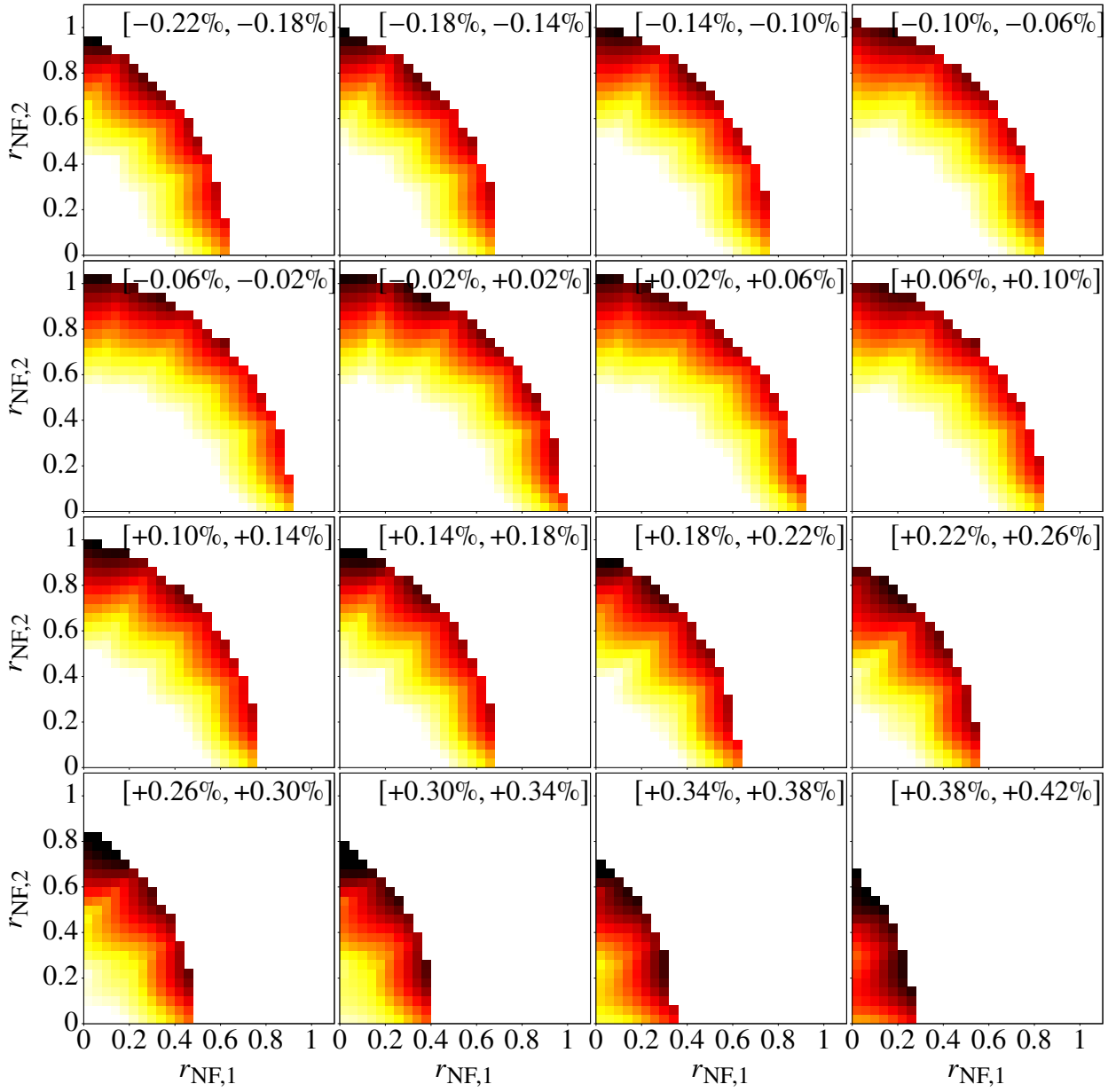

Figure 6.47: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to order 9 instead of the full tenth order. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

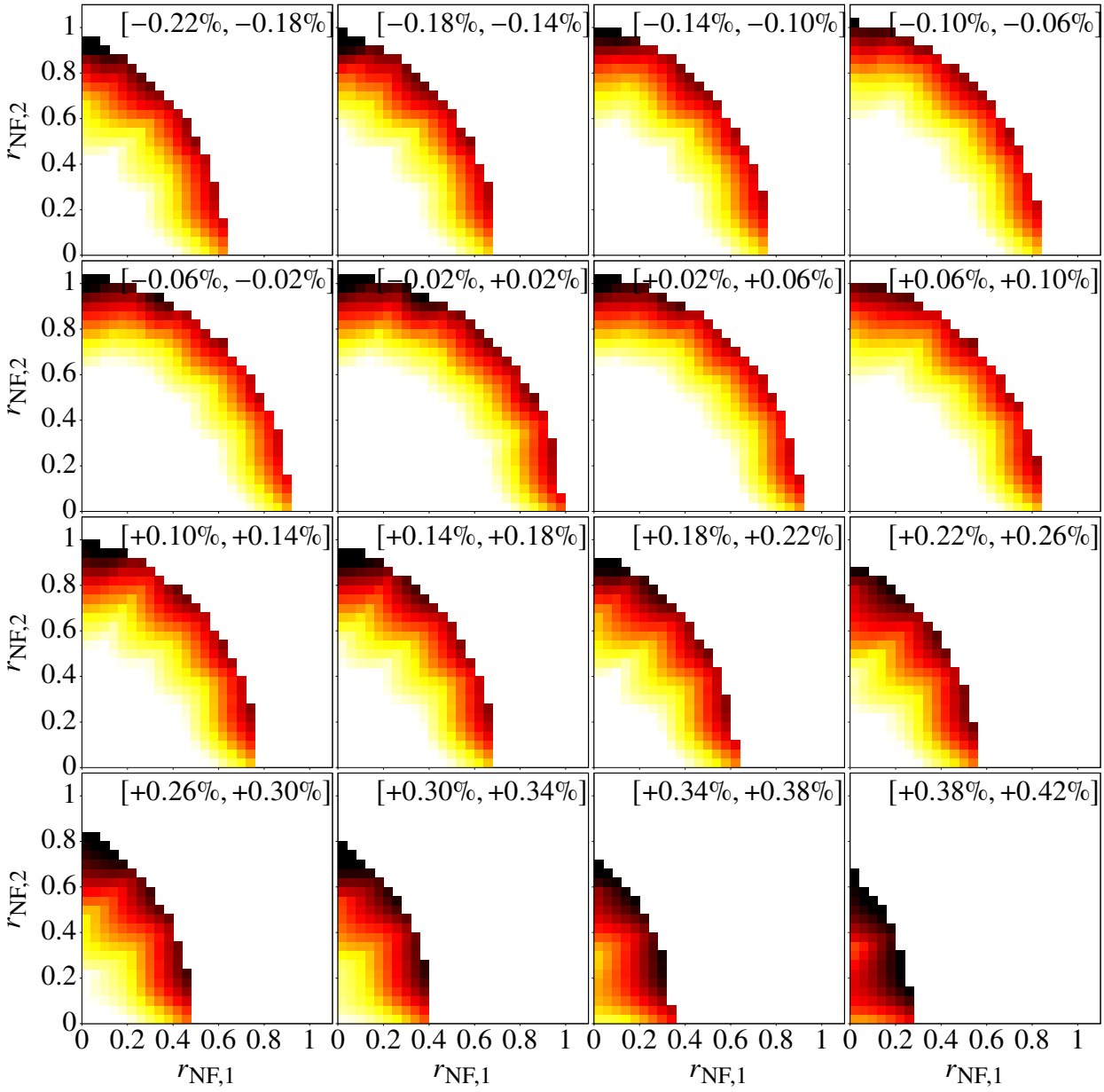

Figure 6.48: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to the full tenth order. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

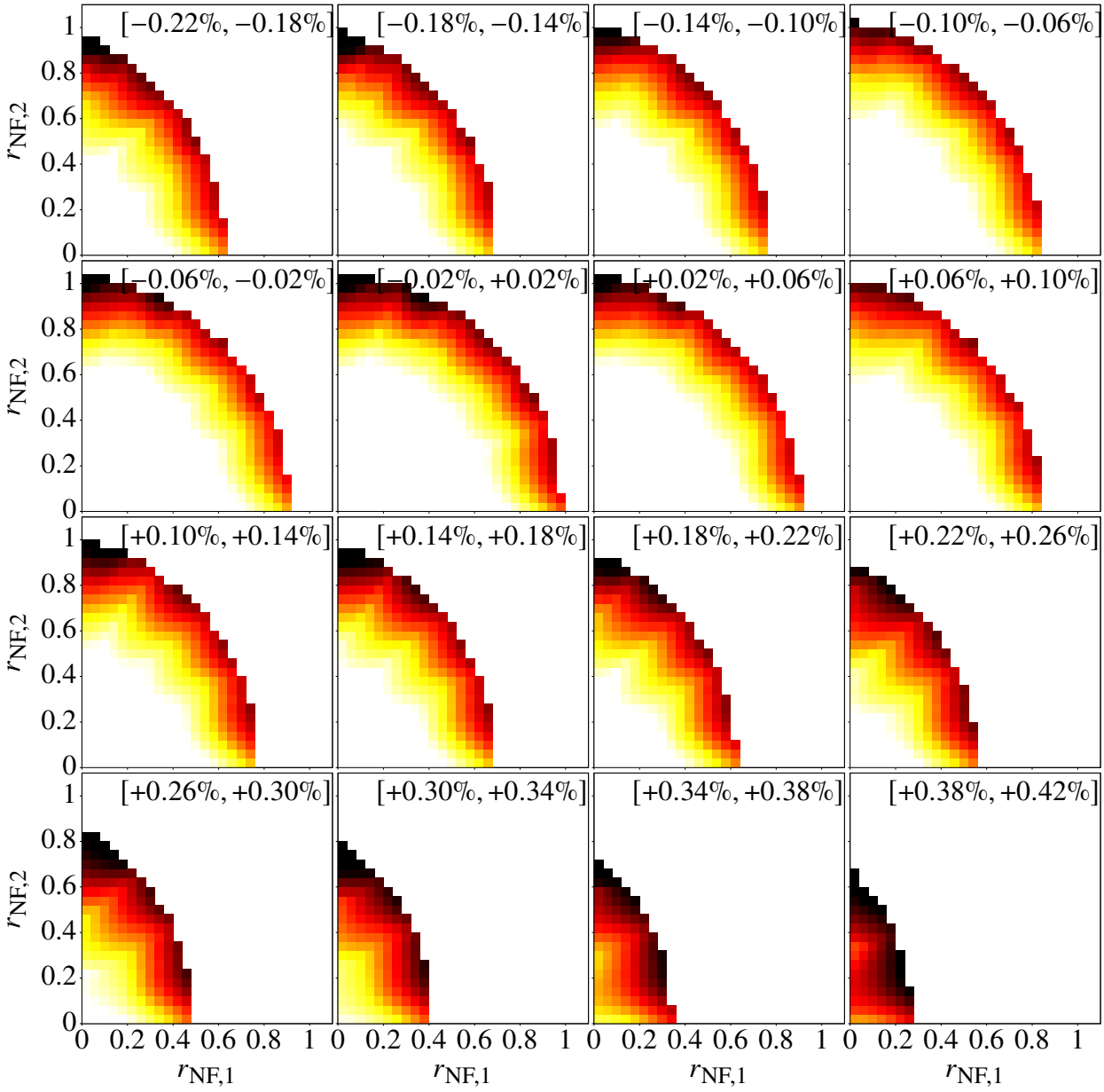

Figure 6.49: Nonverified normal form defect for the phase space storage regions of the muon g-2 storage ring simulation with an ESQ voltage of 18.3 kV using the normal form transformation up to tenth order and an eleventh order map. The individual plots show different momentum ranges, clarified by the label at the top of each graph. The color scheme corresponds to the normal form defect of the specific onion layer. The white boxes for lower normal form radii indicate a normal form defect below 10^{-5}. The yellow boxes denote normal form defects up to 10^{-4}. The orange boxes correspond to normal form defects up to 10^{-3}. The red boxes denote normal form defects up to $10^{-2.5}$, and the black boxes indicate normal form defects larger than that. Each onion layer corresponds to a 0.04×0.04 box in normal form space with a thickness of 0.04% in δp.

CHAPTER 7

CONCLUSION

We investigated a diverse set of nonlinear systems using normal forms and rigorous differential algebra methods. The differential algebra framework implemented in COSY INFINITY served as the backbone of all the methods and techniques in this thesis, and allowed us to establish algorithms and solutions up to arbitrary order and with floating point accuracy.

The basis of our analysis constituted map representations of the various systems based on the underlying equations of motion. These stroboscopic descriptions of the dynamics were expanded around a fixed point corresponding to an equilibrium state of the motion. Using Poincaré projections, the dimensionality of the system was reduced to the essential components of the system's dynamics.

For the bounded motion problem in the zonal gravitational field of the Earth in Chapter 4, the motion was considered within a four dimensional Poincaré surface capturing all ascending node states. In Chapter 5, the dynamics within the muon g-2 storage ring were analyzed in transverse cross sections of the storage ring at multiple azimuthal locations.

The origin preserving maps were then analyzed using high order normal forms to calculate a description of the phase space dynamics that is rotationally invariant up to calculation order. In Chapter 3, the normal form algorithm was discussed in full detail using the illustrative example of the centrifugal governor. In this particular case, the normal form radii, which constitute the (pseudo-)invariants of the motion up to calculation order produced by the normal form algorithm, were directly related to the energy of the system up to calculation order. Additionally, the normal form produced high order functional descriptions of the period of oscillation of the centrifugal governor arms around their equilibrium angle depending on the amplitude of oscillation and changes in the rotation frequency of the governor.

For the bounded motion problem, this rotational invariant representation of the phase space motion provided by the normal form was used to transform the system into action-angle like coordinates. This allowed us to average the bounded motion quantities while maintaining their

functional dependence on the constants of motion. DA inversion methods were then used to enforce the bounded motion conditions and produce parameterized descriptions of the constants of motion, which yielded entire continuous sets of bounded motion orbits. We illustrated that the resulting sets of orbits remained bounded for decades and far beyond the practically relevant distances of formation flying missions.

Our approach can possibly be advanced to the fully gravitationally perturbed case. However, the associated break of the rotational symmetry makes this already complex system even more complex. The introduced longitudinal dependence and the loss of the angular momentum component as a constant of motion increase the dimensionality of the problem by two. Accordingly, pseudo-circular orbits of the full state are required to expand the fixed point map around. Only further research can answer if and how the approach can be adjusted to compensate for the loss of a known constant of motion and the increase in dimensionality.

In our analysis of the dynamics in the muon g-2 storage ring in Chapter 5, we studied the oscillation frequencies of particles in the radial and vertical transfers direction also known as the betatron tunes. The normal form transformation allowed us to calculate the functional dependence of the tunes on the momentum offset of the particles and their amplitude of oscillation. A major insight of this investigation was that particles over the entire momentum offset range could cross the vertical $1/3$-resonance frequency for certain vertical and radial amplitude combinations.

This closeness to the low order resonance triggered intensive lost muon tracking studies, which revealed period-3 fixed point structures in the vertical phase space. Particles caught around those period-3 fixed points experienced significant vertical amplitude modulations, which drastically increased their risk of hitting a collimator and getting lost in the process.

Throughout the analysis, the strong ninth order nonlinearities of the map caused by the 20th-pole of the ESQ potential were prominent. They could be found as eighth order dependencies in the amplitude and momentum dependent tune shifts and be visualized by the drastic change in the tune footprint when comparing eighth order to tenth order results.

To further assess the stability of the muon g-2 storage ring rigorously, we utilized Taylor

Model based verified global optimization in Chapter 6. The abilities of Taylor Model based global optimization was presented using the objective functions of different example problems. The generalized Rosenbrock function served as an example to illustrate different effects that can sometimes influence the optimization like the dependency problem and the cluster effect. We illustrated that Taylor Models and their associated advanced bounding techniques could drastically suppress those effects compared to other commonly used approaches.

The Lennard-Jones problem was used to illustrate the many intricacies that have to be solved for rigorous global optimization of some complex systems. While the Lennard-Jones problem is easily formulated, its formal description with optimization variables and bounding to a rigorous initial search domain are far from trivial. Our discussion of the problem also illustrated the struggle associated with not being able to exclude manifolds from the search domain for which the objective function is not defined.

For the rigorous stability analysis of the muon g-2 storage ring, we calculated verified upper bounds on the rate at which particles can escape the storage region. To get a detailed understanding of the stability properties of the storage ring, we partitioned the five dimensional storage region into more than 8000 sections using the onion layer approach. We used Taylor Model based verified global optimization to calculate the maximum rate of divergence in the form of the normal form defect for each one of those partitions. The verified normal form defect results from the map with the closeness to the vertical $1/3$-resonance from Chapter 5 were compared to the results of a map with a different ESQ voltage, which yielded tunes further away from this vertical low order resonance. The comparison illustrated significant differences in the stability of phase space regions close to the collimators, confirming that the low order resonance noticeably impairs the system's long-term stability. The normal form defect analysis was also able to identify the strong ninth order nonlinearities of the map caused by the 20th-pole of the ESQ potential.

APPENDIX

APPENDIX

VERIFIED GLOBAL OPTIMIZATION RESULTS OF LENNARD-JONES PROBLEM

Table A.1: Verified global optimization results for configurations of k particles in 1D, where the pairwise particle interaction is modeled by the Lennard-Jones potential (see Sec. 6.2.7). The variable v_l is the distance between two adjacent particles p_l and p_{l+1}. The table shows the results for $k = 3$ to $k = 13$.

k	l	$v_l^\star$
3	1	0.998724^{21}_{06}
3	2	0.998724^{21}_{06}
4	1	0.99864^{313}_{295}
4	2	0.997396^{51}_{34}
4	3	0.99864^{313}_{295}
5	1	0.998632^{43}_{22}
5	2	0.997306^{87}_{66}
5	3	0.997306^{87}_{66}
5	4	0.998632^{43}_{22}
6	1	0.9986^{3023}_{2999}
6	2	0.997294^{27}_{03}
6	3	0.99721^{522}_{498}
6	4	0.997294^{27}_{03}
6	5	0.9986^{3023}_{2999}
7	1	0.99862^{964}_{935}
7	2	0.997291^{53}_{25}
7	3	0.99720^{205}_{177}
7	4	0.99720^{205}_{177}
7	5	0.997291^{53}_{25}
7	6	0.99862^{964}_{935}
8	1	0.998629^{45}_{13}
8	2	0.997290^{74}_{43}
8	3	0.99719^{911}_{880}
8	4	0.99718^{868}_{837}
8	5	0.99719^{911}_{880}
8	6	0.997290^{74}_{43}
8	7	0.998629^{45}_{13}

k	l	$v_l^\star$
9	1	0.998629^{39}_{03}
9	2	0.997290^{49}_{13}
9	3	0.99719^{826}_{790}
9	4	0.997185^{67}_{31}
9	5	0.997185^{67}_{31}
9	6	0.99719^{826}_{790}
9	7	0.997290^{49}_{13}
9	8	0.998629^{39}_{03}
10	1	0.99862^{938}_{897}
10	2	0.9972^{9040}_{8999}
10	3	0.997197^{97}_{57}
10	4	0.997184^{78}_{38}
10	5	0.997182^{63}_{22}
10	6	0.997184^{78}_{38}
10	7	0.997197^{97}_{57}
10	8	0.9972^{9040}_{8999}
10	9	0.99862^{938}_{897}
11	1	0.99862^{938}_{893}
11	2	0.9972^{9037}_{8992}
11	3	0.997197^{86}_{41}
11	4	0.997184^{48}_{03}
11	5	0.997181^{72}_{27}
11	6	0.997181^{72}_{27}
11	7	0.997184^{48}_{03}
11	8	0.997197^{86}_{41}
11	9	0.9972^{9037}_{8992}
11	10	0.99862^{938}_{893}

k	l	$v_l^\star$
12	1	0.99862^{941}_{888}
12	2	0.9972^{9038}_{8986}
12	3	0.997197^{84}_{32}
12	4	0.99718^{437}_{385}
12	5	0.99718^{142}_{090}
12	6	0.997180^{82}_{30}
12	7	0.99718^{142}_{090}
12	8	0.99718^{437}_{385}
12	9	0.997197^{84}_{32}
12	10	0.9972^{9038}_{8986}
12	11	0.99862^{941}_{888}
13	1	0.99862^{944}_{885}
13	2	0.9972^{9040}_{8981}
13	3	0.997197^{84}_{26}
13	4	0.99718^{434}_{376}
13	5	0.99718^{131}_{073}
13	6	0.9971^{8051}_{7993}
13	7	0.9971^{8051}_{7993}
13	8	0.99718^{131}_{073}
13	9	0.99718^{434}_{376}
13	10	0.997197^{84}_{26}
13	11	0.9972^{9040}_{8981}
13	12	0.99862^{944}_{885}

Table A.2: Verified global optimization results for configurations of k particles in 1D, where the pairwise particle interaction is modeled by the Lennard-Jones potential (see Sec. 6.2.7). The variable v_l is the distance between two adjacent particles p_l and p_{l+1}. The table shows the results for $k = 14$ and $k = 15$.

k	l	$v_l^\star$
14	1	0.99862_{881}^{947}
14	2	0.9972_{8977}^{9043}
14	3	$0.99719_{21}^{7}86$
14	4	0.99718_{369}^{434}
14	5	0.99718_{063}^{128}
14	6	0.9971_{7975}^{8040}
14	7	0.9971_{7955}^{8020}
14	8	0.9971_{7975}^{8040}
14	9	0.99718_{063}^{128}
14	10	0.99718_{369}^{434}
14	11	$0.99719_{21}^{7}86$
14	12	0.9972_{8977}^{9043}
14	13	0.99862_{881}^{947}

k	l	$v_l^\star$
15	1	0.99862_{877}^{950}
15	2	0.9972_{8973}^{9046}
15	3	$0.99719_{16}^{7}89$
15	4	0.99718_{364}^{436}
15	5	0.99718_{056}^{128}
15	6	0.9971_{7965}^{8037}
15	7	0.9971_{7937}^{8009}
15	8	0.9971_{7937}^{8009}
15	9	0.9971_{7965}^{8037}
15	10	0.99718_{056}^{128}
15	11	0.99718_{364}^{436}
15	12	$0.99719_{16}^{7}89$
15	13	0.9972_{8973}^{9046}
15	14	0.99862_{877}^{950}

Table A.3: Results for the calculated lower bounds r_{LB} on the minimum distance between particles in a 1D configuration of k particles (see Eq. (6.11) and Sec. 6.2.7).

k	r_{LB}	k	r_{LB}
3	0.892064059	10	0.799735218
4	0.864104625	11	0.793892293
5	0.846285971	12	0.788630919
6	0.833085500	13	0.783845621
7	0.822566681	14	0.779457384
8	0.813810435	15	0.775405451
9	0.806305067		

Table A.4: Verified global optimization results for symmetric configurations of k particles in 1D, where the pairwise particle interaction is modeled by the Lennard-Jones potential (see Sec. 6.2.8). The variable v_l is the distance between two adjacent particles p_l and p_{l+1}. The table shows the results for $k = 3$ to $k = 18$.

k	l	$v_l^\star$
3	1	0.998724^{19}_{08}
4	1	0.99864^{310}_{297}
4	2	0.997396^{51}_{34}
5	1	0.998632^{40}_{24}
5	2	0.997306^{84}_{69}
6	1	0.998630^{19}_{02}
6	2	0.997294^{23}_{06}
6	3	0.99721^{522}_{498}
7	1	0.998629^{59}_{39}
7	2	0.997291^{49}_{29}
7	3	0.99720^{201}_{181}
8	1	0.998629^{40}_{18}
8	2	0.997290^{70}_{48}
8	3	0.99719^{907}_{885}
8	4	0.997188^{68}_{37}
9	1	0.998629^{34}_{08}
9	2	0.997290^{44}_{18}
9	3	0.99719^{821}_{796}
9	4	0.997185^{62}_{36}
10	1	0.998629^{32}_{03}
10	2	0.997290^{34}_{05}
10	3	0.997197^{91}_{63}
10	4	0.997184^{72}_{44}
10	5	0.997182^{62}_{23}
11	1	0.99862^{932}_{899}
11	2	0.9972^{9030}_{8999}
11	3	0.997197^{80}_{48}
11	4	0.997184^{41}_{09}
11	5	0.997181^{66}_{34}
12	1	0.99862^{933}_{896}
12	2	0.9972^{9030}_{8994}
12	3	0.997197^{76}_{40}
12	4	0.997184^{429}_{393}
12	5	0.997181^{134}_{098}
12	6	0.997180^{82}_{31}
13	1	0.99862^{935}_{893}
13	2	0.9972^{9031}_{8990}
13	3	0.997197^{76}_{34}
13	4	0.997184^{425}_{384}
13	5	0.997181^{122}_{081}
13	6	0.997180^{43}_{01}
14	1	0.99862^{937}_{891}
14	2	0.9972^{9033}_{8987}
14	3	0.997197^{76}_{30}
14	4	0.997184^{425}_{379}
14	5	0.997181^{118}_{072}
14	6	0.9971^{8031}_{7985}
14	7	0.9971^{8020}_{7955}
15	1	0.99862^{939}_{888}
15	2	0.9972^{9035}_{8984}
15	3	0.997197^{78}_{27}
15	4	0.997184^{425}_{375}
15	5	0.997181^{117}_{066}
15	6	0.9971^{8026}_{7975}
15	7	0.997179^{98}_{47}
16	1	0.99862^{942}_{886}
16	2	0.9972^{9037}_{8982}
16	3	0.997197^{80}_{24}
16	4	0.997184^{426}_{371}
16	5	0.997181^{117}_{062}
16	6	0.9971^{8025}_{7970}
16	7	0.997179^{93}_{38}
16	8	0.997179^{97}_{19}
17	1	0.99862^{942}_{885}
17	2	0.9972^{9037}_{8981}
17	3	0.997197^{80}_{24}
17	4	0.997184^{426}_{370}
17	5	0.997181^{117}_{061}
17	6	0.9971^{8024}_{7967}
17	7	0.997179^{90}_{34}
17	8	0.997179^{79}_{23}
18	1	0.99862^{944}_{883}
18	2	0.9972^{9039}_{8979}
18	3	0.997197^{81}_{22}
18	4	0.997184^{428}_{368}
18	5	0.997181^{118}_{058}
18	6	0.9971^{8024}_{7964}
18	7	0.997179^{90}_{31}
18	8	0.997179^{77}_{18}
18	9	0.997179^{86}_{02}

Table A.5: Verified global optimization results for symmetric configurations of k particles in 1D, where the pairwise particle interaction is modeled by the Lennard-Jones potential (see Sec. 6.2.8). The variable v_l is the distance between two adjacent particles p_l and p_{l+1}. The table shows the results for $k = 19$ to $k = 25$.

k	l	$v_l^\star$
19	1	0.99862_{880}^{947}
19	2	0.9972_{8976}^{9042}
19	3	0.997197_{19}^{84}
19	4	0.99718_{365}^{430}
19	5	0.99718_{055}^{121}
19	6	0.9971_{7961}^{8026}
19	7	0.997179_{27}^{92}
19	8	0.997179_{13}^{78}
19	9	0.997179_{08}^{73}
20	1	0.99862_{877}^{950}
20	2	0.9972_{8973}^{9045}
20	3	0.997197_{15}^{87}
20	4	0.99718_{362}^{433}
20	5	0.99718_{052}^{123}
20	6	0.9971_{7958}^{8029}
20	7	0.997179_{23}^{95}
20	8	0.997179_{09}^{80}
20	9	0.997179_{03}^{74}
20	10	0.99717_{886}^{987}
21	1	0.99862_{874}^{953}
21	2	0.9972_{8970}^{9048}
21	3	0.997197_{12}^{90}
21	4	0.99718_{359}^{436}
21	5	0.99718_{049}^{126}
21	6	0.9971_{7954}^{8032}
21	7	0.997179_{19}^{97}
21	8	0.997179_{05}^{83}
21	9	0.99717_{899}^{976}
21	10	0.99717_{896}^{974}
22	1	0.99862_{870}^{956}
22	2	0.9972_{8966}^{9051}
22	3	0.997197_{09}^{94}
22	4	0.99718_{355}^{440}
22	5	0.99718_{045}^{130}
22	6	0.9971_{7950}^{8035}
22	7	0.997179_{16}^{100}
22	8	0.997179_{01}^{86}
22	9	0.99717_{894}^{979}
22	10	0.99717_{891}^{976}
22	11	0.99717_{873}^{993}
23	1	0.99862_{858}^{969}
23	2	0.9972_{8954}^{9063}
23	3	0.99719_{697}^{806}
23	4	0.99718_{343}^{452}
23	5	0.99718_{033}^{142}
23	6	0.9971_{7938}^{8047}
23	7	0.9971_{7904}^{8012}
23	8	0.99717_{889}^{997}
23	9	0.99717_{882}^{991}
23	10	0.99717_{879}^{987}
23	11	0.99717_{877}^{986}
24	1	0.99862_{853}^{974}
24	2	0.9972_{8949}^{9069}
24	3	0.99719_{691}^{811}
24	4	0.99718_{337}^{457}
24	5	0.99718_{027}^{147}
24	6	0.9971_{7933}^{8053}
24	7	0.9971_{7898}^{8018}
24	8	0.9971_{7883}^{8003}
24	9	0.99717_{876}^{996}
24	10	0.99717_{873}^{993}
24	11	0.99717_{871}^{991}
24	12	0.9971_{7846}^{8015}
25	1	0.99862_{848}^{979}
25	2	0.9972_{8944}^{9074}
25	3	0.99719_{686}^{816}
25	4	0.99718_{333}^{462}
25	5	0.99718_{023}^{152}
25	6	0.9971_{7928}^{8057}
25	7	0.9971_{7893}^{8022}
25	8	0.9971_{7878}^{8008}
25	9	0.9971_{7871}^{8001}
25	10	0.99717_{868}^{997}
25	11	0.99717_{866}^{995}
25	12	0.99717_{865}^{995}

Table A.6: Verified global optimization results for symmetric configurations of k particles in 1D, where the pairwise particle interaction is modeled by the Lennard-Jones potential (see Sec. 6.2.8). The variable v_l is the distance between two adjacent particles p_l and p_{l+1}. The table shows the results for $k = 26$ and $k = 27$.

k	l	$v_l^\star$	k	l	$v_l^\star$
26	1	0.99862^{984}_{843}	27	1	0.99862^{989}_{838}
26	2	0.9972^{9078}_{8939}	27	2	0.9972^{9084}_{8934}
26	3	0.99719^{821}_{682}	27	3	0.99719^{826}_{676}
26	4	0.99718^{467}_{328}	27	4	0.99718^{472}_{323}
26	5	0.99718^{157}_{018}	27	5	0.99718^{162}_{012}
26	6	0.9971^{8062}_{7923}	27	6	0.9971^{8067}_{7918}
26	7	0.9971^{8027}_{7888}	27	7	0.9971^{8032}_{7883}
26	8	0.9971^{8012}_{7873}	27	8	0.9971^{8017}_{7868}
26	9	0.9971^{8005}_{7866}	27	9	0.9971^{8010}_{7861}
26	10	0.9971^{8002}_{7863}	27	10	0.9971^{8007}_{7857}
26	11	0.9971^{8000}_{7861}	27	11	0.9971^{8005}_{7855}
26	12	0.99717^{999}_{860}	27	12	0.9971^{8004}_{7854}
26	13	0.9971^{8027}_{7831}	27	13	0.9971^{8004}_{7854}

Table A.7: Results for the calculated lower bounds r_{LB} on the minimum distance between particles in a 1D symmetric configuration of k particles (see Eq. (6.11) and Sec. 6.2.8).

k	r_{LB}	k	r_{LB}
3	0.892064059	16	0.771642074
4	0.864104625	17	0.768129039
5	0.846285971	18	0.764835276
6	0.833085500	19	0.761735176
7	0.822566681	20	0.758807381
8	0.813810435	21	0.756033888
9	0.806305067	22	0.753399380
10	0.799735218	23	0.750890717
11	0.793892293	24	0.748496539
12	0.788630919	25	0.746206959
13	0.783845621	26	0.744013320
14	0.779457384	27	0.741907996
15	0.775405451		

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Kyle T. Alfriend, Srinivas R. Vadali, Pini Gurfil, Jonathan P. How, and Louis S. Breger. *Spacecraft Formation Flying: Dynamics, Control and Navigation*, pages 144–146. Batterworth-Heinemann, Oxford, 2010.

[2] Tatsumi Aoyama, Nils Asmussen, Maurice Benayoun, Johan Bijnens, Thomas C. Blum, et al. The anomalous magnetic moment of the muon in the standard model. *Physics reports*, 2020.

[3] Babak Abi et al. (Muon g-2 Collaboration). Measurement of the positive muon anomalous magnetic moment to 0.46 ppm. *Phys. Rev. Lett.*, 126:141801, 2021.

[4] Nicola Baresi, Zubin P. Olikara, and Daniel J. Scheeres. Fully numerical methods for continuing families of quasi-periodic invariant tori in astrodynamics. *The Journal of the Astronautical Sciences*, 65(2):157–182, 2018.

[5] Nicola Baresi and Daniel J. Scheeres. Bounded relative motion under zonal harmonics perturbations. *Celestial Mechanics and Dynamical Astronomy*, 127(4):527–548, 2017.

[6] Nicola Baresi and Daniel J. Scheeres. Design of bounded relative trajectories in the earth zonal problem. *Journal of Guidance, Control, and Dynamics*, 40(12):3075–3087, 2017.

[7] Sonja Berner. Parallel methods for verified global optimization practice and theory. *Journal of Global Optimization*, 9(1):1–22, 1996.

[8] Martin Berz. Private communication.

[9] Martin Berz. The method of power series tracking for the mathematical description of beam dynamics. *Nuclear Instruments and Methods A*, 258(3):431–436, 1987.

[10] Martin Berz. Differential algebraic description of beam dynamics to very high orders. *Part. Accel.*, 24(SSC-152):109–124, 1988.

[11] Martin Berz. High-order computation and normal form analysis of repetitive systems. In *AIP Conference Proceedings*, volume 249, pages 456–489, 1992.

[12] Martin Berz. Differential algebraic formulation of normal form theory. In *Conference series - Institute of Physics*, volume 131, pages 77–77. IOP Publishing LTD, 1993.

[13] Martin Berz. Differential algebraic description and analysis of spin dynamics. *AIP CP*, 343, 1995.

[14] Martin Berz. *Modern Map Methods in Particle Beam Physics*. Academic Press, 1999.

[15] Martin Berz and Georg Hoffstätter. Computation and application of Taylor polynomials with interval remainder bounds. *Reliable Computing*, 4(1):83–97, 1998.

[16] Martin Berz and Kyoko Makino. Constructive generation and verification of Lyapunov functions around fixed points of nonlinear dynamical systems. *International Journal of Computer Research*, 12,2:235–244, 2003.

[17] Martin Berz and Kyoko Makino. Suppression of the wrapping effect by Taylor model- based verified integrators: Long-term stabilization by shrink wrapping. *International Journal of Differential Equations and Applications*, 10,4:385–403, 2005.

[18] Martin Berz and Kyoko Makino. COSY INFINITY Version 9.2 programmer's manual. Technical Report MSUHEP-151102, Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, 2015. See also http://cosyinfinity.org.

[19] Martin Berz and Kyoko Makino. COSY INFINITY 10.0 beam physics manual. MSU Report MSUHEP-151103, Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, 2017. See also http://cosyinfinity.org.

[20] Martin Berz and Kyoko Makino. COSY INFINITY Version 10.0 beam physics manual. Technical Report MSUHEP-151103-rev, Michigan State University, 2017.

[21] Martin Berz and Kyoko Makino. COSY INFINITY Version 10.0 programmer's manual. Technical Report MSUHEP-151102-rev, Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, 2017. See also http://cosyinfinity.org.

[22] Martin Berz, Kyoko Makino, and Youn-Kyung Kim. Long-term stability of the Tevatron by validated global optimization. *Nuclear Instruments and Methods*, 558(1):1–10, 2006.

[23] Roger A. Broucke. Numerical integration of periodic orbits in the main problem of artificial satellite theory. *Celestial Mechanics and Dynamical Astronomy*, 58(2):99–123, 1994.

[24] Owen Brown and Paul Eremenko. Fractionated space architectures: A vision for responsive space. Technical report, DEFENSE ADVANCED RESEARCH PROJECTS AGENCY ARLINGTON VA, 2006.

[25] Ernest D. Courant and Hartland S. Snyder. Theory of the alternating-gradient synchrotron. *Annals of Physics*, 3(1):1 – 48, 1958.

[26] Simone D'Amico, Jean Sebastien Ardaens, and Robin Larsson. Spaceborne autonomous formation flying experiment on the prisma mission. *Journal of Guidance, Control and Dynamics*, 35(3):834–850, 2012.

[27] Simone D'Amico and Oliver Montenbruck. Proximity operations of formation-flying spacecraft using an eccentricity/inclination vector separation. *Journal of Guidance Control and Dynamics*, 29(3):554–563, 2006.

[28] Paul Adrien Maurice Dirac. The quantum theory of the electron. *Proc. R. Soc. Lond. A*, 117(778):610–624, 1928.

[29] Paul Adrien Maurice Dirac. The quantum theory of the electron. part II. *Proc. R. Soc. Lond. A*, 118(779):351–361, 1928.

[30] Kaisheng Du and Ralph B. Kearfott. The cluster problem in multivariate global optimization. *J. Global Optim.*, 5:253–265, 1994.

[31] Etienne Forest, John Irwin, and Martin Berz. Normal form methods for complicated periodic systems. *Part. Accel.*, 24:91–107, 1989.

[32] Gerald W. Bennett et al. (Muon g-2 Collaboration). Final report of the E821 muon anomalous magnetic moment measurement at BNL. *Physical Review D*, 73(7):072003, 2006.

[33] Graziano Venanzoni (on behalf of the Fermilab E989 collaboration). The new muon g-2 experiment at Fermilab. *Nuclear and Particle Physics Proceedings*, 273:584–588, 2016.

[34] Johannes Grote, Martin Berz, and Kyoko Makino. High-order representation of Poincaré maps. *Nuclear Instruments and Methods A*, 558(1):106–111, 2006.

[35] Yanchao He, Roberto Armellin, and Ming Xu. Bounded relative orbits in the zonal problem via high-order poincaré maps. *Journal of Guidance, Control, and Dynamics*, 42(1):91–108, 2018.

[36] Joe M. Grange et al. (Muon g-2 Collaboration). Muon (g-2) technical design report. *arXiv preprint: 1501.06858*, 2015.

[37] Ralph B. Kearfott. *Rigorous Global Search: Continuous Problems*. Kluwer, Dordrecht, 1996.

[38] Ralph B. Kearfott and Kaisheng Du. The cluster problem in global optimization: The univariate case. *Computing*, 9 (Supple):117–127, 1992.

[39] Ellis Robert Kolchin. *Differential algebra & algebraic groups*. Academic press, 1973.

[40] Wang Sang Koon, Jerrold E. Marsden, Richard M. Murray, and Josep Masdemont. J_2 dynamics and formation flight. In *AIAA Guidance, Navigation, and Control Conference and Exhibit*, Montreal, Canada, 2001. AIAA.

[41] Ulrich W. Kulisch and Willard L. Miranker. *Computer Arithmetic in Theory and Practice*. Academic Press, New York, 1981.

[42] Vangipuram Lakshmikantham, Vladimir M. Matrosov, and Seenith Sivasundaram. *Vector Lyapunov Functions and Stability Analysis of Nonlinear Systems*. Kluwer Academic Publishers, Dordrecht, Netherlands, 1991.

[43] John E. Lennard-Jones. On the determination of molecular fields. ii. from the equation of state of gas. *Proceedings of the Royal Society of London A*, 106:463–477, 1924.

[44] John E. Lennard-Jones. Cohesion. *Proceedings of the Physical Society (1926-1948)*, 43(5):461, 1931.

[45] Aleksandr M. Lyapunov. *The General Problem of the Stability of Motion*. Taylor and Francis, London, 1992.

[46] Kyoko Makino. *Rigorous Analysis of Nonlinear Motion in Particle Accelerators*. PhD thesis, Michigan State University, East Lansing, Michigan, USA, 1998. Also MSUCL-1093.

[47] Kyoko Makino and Martin Berz. Remainder differential algebras and their applications. In M. Berz, C. Bischof, G. Corliss, and A. Griewank, editors, *Computational Differentiation: Techniques, Applications, and Tools*, pages 63–74, Philadelphia, 1996. SIAM.

[48] Kyoko Makino and Martin Berz. Efficient control of the dependency problem based on Taylor model methods. *Reliable Computing*, 5(1):3–12, 1999.

[49] Kyoko Makino and Martin Berz. Effects of kinematic correction on the dynamics in muon rings. *AIP CP*, 530:217–227, 2000.

[50] Kyoko Makino and Martin Berz. Verified global optimization with Taylor model methods. In N. Mastorakis, editor, *Problems in Modern Applied Mathematics*, pages 253–258. World Scientific and Engineering Society Press, 2000.

[51] Kyoko Makino and Martin Berz. Taylor models and other validated functional inclusion methods. *International Journal of Pure and Applied Mathematics*, 6,3:239–316, 2003.

[52] Kyoko Makino and Martin Berz. Suppression of the wrapping effect by Taylor model- based verified integrators: Long-term stabilization by preconditioning. *International Journal of Differential Equations and Applications*, 10,4:353–384, 2005.

[53] Kyoko Makino and Martin Berz. COSY INFINITY version 9. *Nuclear Instruments and Methods*, 558(1):346–350, 2006.

[54] Kyoko Makino and Martin Berz. Suppression of the wrapping effect by Taylor model- based verified integrators: The single step. *International Journal of Pure and Applied Mathematics*, 36,2:175–197, 2006.

[55] Kyoko Makino and Martin Berz. Optimal correction and design parameter search by modern methods of rigorous global optimization. *Nuclear Instruments and Methods*, 645:332–337, 2011. doi:10.1016/j.nima.2010.12.185.

[56] Kyoko Makino and Martin Berz. The LDB, QDB, and QFB bounders. Technical Report MSUHEP-40617, Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, June 2004.

[57] Vladimir Martinusi and Pini Gurfil. Closed-form solutions for satellite relative motion in an axially-symmetric gravitational field. *Advances in the Astronautical Sciences*, 140:1525–1544, 2011.

[58] Oliver Montenbruck, Michael Kirschner, Simone D'Amico, and Srinivas Bettadpur. E/I-vector separation for safe switching of the grace formation. *Aerospace Science and Technology*, 10(7):628–635, 2006.

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

[60] Ramon E. Moore. *Methods and Applications of Interval Analysis*. SIAM, 1979.

[61] Ramon E. Moore, Eldon Hansen, and Anthony Leclerc. Rigorous methods for global optimization. In *Recent Advances in Global Optimization (Princeton, NJ, 1991)*, Princeton Ser. Comput. Sci., pages 321–342. Princeton Univ. Press, 1992.

[62] John E. Nafe, Edward B. Nelson, and Isidor I. Rabi. The hyperfine structure of atomic hydrogen and deuterium. *Physical Review*, 71(12):914, 1947.

[63] Darragh E. Nagle, Renne S. Julian, and Jerrold R. Zacharias. The hyperfine structure of atomic hydrogen and deuterium. *Physical Review*, 72(10):971, 1947.

[64] Nikolai N. Nekhoroschev. An exponential estimate of the time of stability of nearly integrable Hamiltonian systems. *Uspekhi Mat. Nauk 32:6, English translation Russ. Math. Surv.*, 32:6,5:1, 1977.

[65] Henri Poincaré. *Les méhodes nouvelles de la mécanique céleste*, volume I-III. Gauthier-Villars it fils, 1892, 1893, 1899.

[66] Nathalie Revol, Kyoko Makino, and Martin Berz. Taylor models and floating-point arithmetic: Proof that arithmetic operations are validated in COSY. *Journal of Logic and Algebraic Programming*, 64/1:135–154, 2004.

[67] Joseph F. Ritt. *Differential equations from the algebraic standpoint*, volume 14. American Mathematical Soc., Washington, D.C., 1932.

[68] Joseph F. Ritt and Joseph Liouville. *Integration in finite terms: Liouville's theory of elementary methods*. Columbia Univ. Press, 1948.

[69] Howard H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *The Computer Journal*, 3(3):175–184, 1960.

[70] Hanspeter Schaub and Kyle T. Alfriend. J_2 invariant relative orbits for spacecraft formations. *Celestial Mechanics and Dynamical Astronomy*, 79(2):77–95, 2001.

[71] Julian Schwinger. Quantum electrodynamics. I. A covariant formulation. *Physical Review*, 74(10):1439, 1948.

[72] Julian Schwinger. Quantum electrodynamics. III. The electromagnetic properties of the electron—radiative corrections to scattering. *Physical Review*, 76(6):790, 1949.

[73] Yannis K. Semertzidis, Gerald Bennett, Efstratios Efstathiadis, Frank Krienen, Richard Larsen, et al. The Brookhaven muon (g-2) storage ring high voltage quadrupoles. *Nuclear Instruments and Methods A*, 503(3):458–484, 2003.

[74] Diktys Stratakis, Mary E. Convery, Carol Johnstone, John Johnstone, James P. Morgan, et al. Accelerator performance analysis of the fermilab muon campus. *Physical Review Accelerators and Beams*, 20(11):111003, 2017.

[75] Diktys Stratakis, Brian Drendel, James P. Morgan, Michael J. Syphers, and Nathan S. Froemming. Commissioning and first results of the fermilab muon campus. *Physical Review Accelerators and Beams*, 22(1):011001, 2019.

[76] Michael J. Syphers. Long-term muon loss rates and an estimate of ω_a systematic uncertainty. Technical report, Muon g-2 Collaboration, Fermi National Accelerator Laboratory, 2020.

[77] David Tarazona. *Beam dynamics characterization and uncertainties in the Muon g-2 Experiment at Fermilab*. PhD thesis, Michigan State University, East Lansing, Michigan, USA, 2021.

[78] David Tarazona, Martin Berz, and Kyoko Makino. Muon loss rates from betatron resonances at the muon g-2 storage ring at Fermilab. *International Journal of Modern Physics A*, 34(36):1942008, 2019.

[79] David Tarazona, Martin Berz, Kyoko Makino, Diktys Stratakis, and Michael J. Syphers. Dynamical simulations of the muon campus at Fermilab. *International Journal of Modern Physics A*, 34(36):1942033, 2019.

[80] David Tarazona, Eremey Valetov, Adrian Weisskopf, Martin Berz, and Kyoko Makino. E989 note 265: Lost-muon studies. Technical report, Fermilab Muon g-2, 2021.

[81] Tareq Albahri et al. (Muon g-2 Collaboration). Beam dynamics corrections to the run-1 measurement of the muon anomalous magnetic moment at fermilab. *arXiv preprint arXiv:2104.03240*, 2021.

[82] Tareq Albahri et al. (Muon g-2 Collaboration). Magnetic-field measurement and analysis for the muon g-2 experiment at Fermilab. *Physical Review A*, 103(4):042208, 2021.

[83] Tareq Albahri et al. (Muon g-2 Collaboration). Measurement of the anomalous precession frequency of the muon in the Fermilab muon g-2 experiment. *Physical Review D*, 103(7):072002, 2021.

[84] Srinivas R. Vadali, Hanspeter Schaub, and Kyle T. Alfriend. Initial conditions and fuel-optimal control for formation flying of satellites. In *AIAA Guidance, Navigation, and Control Conference and Exhibit*, Portland, OR, 1999.

[85] Eremey Valetov, Martin Berz, and Kyoko Makino. Validation of transfer map calculation for electrostatic deflectors in the code COSY INFINITY. *International Journal of Modern Physics A*, 34(36):1942010, 2019.

[86] Adrian Weisskopf. Applications of the DA based normal form algorithm on parameter-dependent perturbations. Master's thesis, Michigan State University, East Lansing, Michigan, USA, 2016.

[87] Adrian Weisskopf. Introduction to the differential algebra normal form algorithm using the centrifugal governor as an example. Technical Report MSUHEP-190617, Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, 2019. arXiv:1906.10758[physics.class-ph].

[88] Adrian Weisskopf, Roberto Armellin, and Martin Berz. Bounded motion design in the earth zonal problem using differential algebra based normal form methods. *Celestial Mechanics and Dynamical Astronomy*, 132(14), 2020.

[89] Adrian Weisskopf, David Tarazona, and Martin Berz. Computation and consequences of high order amplitude- and parameter-dependent tune shifts in storage rings for high precision measurements. *International Journal of Modern Physics A*, 34(36):1942011, 2019.

[90] Ming Xu, Yue Wang, and Shijie Xu. On the existence of J_2 invariant relative orbits from the dynamical system point of view. *Celestial Mechanics and Dynamical Astronomy*, 112(4):427–444, 2012.