Oleksii Beznosov

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Deborah Sulsky , Chairperson

Daniel Appelö

Desmond Barber

James Ellison

Klaus Heinemann

Stephen Lau

From Wave Propagation to Spin Dynamics: Mathematical and Computational Aspects

by

Oleksii Beznosov

B.S., Taras Shevchenko National University of Kyiv, 2012
M.S., Taras Shevchenko National University of Kyiv, 2014

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Mathematics

The University of New Mexico

Albuquerque, New Mexico

December, 2020

Dedication

To the memory of my grandparents

Acknowledgments

When you work on a problem, try to be creative.
My mother

I started to work on projects for my PhD with Daniel Appelö, Desmond Barber, Jim Ellison, and Klaus Heinemann. With their attitude towards our studies, that I can now proudly present in this manuscript, they have taught me the most important lesson: as long as we approach the problem with creativity it brings the most joy. This group, where everyone carries an enormous amount of expertise in mathematics and physics and puts in so much effort, has set up the creative environment for me where I wanted to understand every single bit of mathematics and science, and I am very grateful for that. Also I would like to thank Deborah Sulsky for her expert insights in numerical analysis, suggesting the literature and providing her written notes, that I have used to implement the algorithms developed in Chapter 7. I would like to thank Stephen Lau for providing his fast code that I used as a part of the solver for the six-dimensional reduced Bloch equation, and his excellent suggestions regarding it. I would like to thank Bill Henshaw for his input on overset grid methods presented in the Chapter 10 of this thesis. I would like to thank the inviting accelerator physics community, especially the United States Particle Accelerator School (USPAS), its organizers and invited teachers for giving me the necessary background in the field. I would like to thank our collaborators David Sagan and Fanglei Lin for many fruitful discussions. I also would like to thank Etienne Forest for an introduction to his library for analysis and particle tracking, and for the work he put into our spin-tracking project that I plan to continue in the nearest future.

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under Award Numbers DE-SC0018008 and DE-SC0018370. The HEP and accelerator stewardship grants made it possible for me to work on this thesis without interruption. Therefore, I thank Eric Colby (DOE), Jie Gao (CEPC), Georg Hoffstaetter (Cornell), L.K. Len (DOE), Vadim Ptitsyn (BNL), and Frank Zimmermann (CERN) for their support of the grant awards.

Every part of this work, from code to manuscript and figures, was written in an open source text editor, running on an open source operating system, and was compiled using an open source software (Latex, gcc, gfortran, gnuplot, etc.). So, I would like to thank the huge community of the open source developers for putting a lot of work into their software packages and freely distributing them among everyone to use.

I also would like to thank my family: Olga, Valerii, Anastasiia, Oleksandra, Boris, my little nieces, and two very best boys, Roy and Spin, for being there for me and supporting me regardless of distance separating us.

And last, but not least, I would like to thank you, the reader, reading this manuscript. Without your interest, this work might never leave these pages and I am very happy to share it with you. From now on, in every chapter, I will refer to you and me as *we*, and so let us start *our* discussion.

From Wave Propagation to Spin Dynamics: Mathematical and Computational Aspects

by

Oleksii Beznosov

B.S., Taras Shevchenko National University of Kyiv, 2012
M.S., Taras Shevchenko National University of Kyiv, 2014

Ph.D., Mathematics, University of New Mexico, 2020

Abstract

In this work we concentrate on two separate topics which pose certain numerical challenges. The first topic is the spin dynamics of electrons in high-energy circular accelerators. We introduce a stochastic differential equation framework to study spin depolarization and spin equilibrium. This framework allows the mathematical study of known equations and new equations modelling the spin distribution of an electron bunch. A spin distribution is governed by a so-called Bloch equation, which is a linear Fokker-Planck type PDE, in general posed in six dimensions. We propose three approaches to approximate solutions, using analytical and modern numerical techniques. We also present simple models that carry all computational difficulties of those modelling the realistic accelerators, to demonstrate the effectiveness of our framework and the approximations. In the second part of this work we present a high-order accurate numerical method for the wave equation posed on a domain with complex boundary. The method combines efficient Hermite methods with the geometrically flexible Discontinuous Galerkin method by using overset grids. Near boundaries we use thin boundary-fitted curvilinear grids and inside the volume we use Cartesian grids so that the computational complexity of the method approaches that of a structured Cartesian Hermite method.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern numerical analysis finds its applications in all the existing sciences including the physical sciences and engineering, biology and the medical sciences, business and data science. Modern numerical algorithms for approximately solving differential equations have been applied to complex dynamical systems modelling many phenomena. In this work we concentrate on two separate topics. The first topic is the dynamics of a property of electrons and positrons known as spin in circular accelerators that store electron or positron bunches. The techniques we propose aim at improving the simulation and optimization of that dynamics in order to increase the utility of the particle bunches. The second part of this work concentrates on a new approach for solving wave propagation problems posed on domains with complex boundaries. The techniques we propose find their applications in seismic wave propagation and acoustics.

1.1 Spin dynamics in modern electron storage rings

In the first part of this work we describe analytical and numerical aspects of our work on spin dynamics and so-called spin polarization of bunches in high-energy electron and positron storage rings. The results of this work are relevant for high-energy electron storage rings in general but in particular they are relevant for the Electron Ion Collider (EIC) [1], the e^-e^+ option of the proposed Future Circular Collider (FCC-ee) [2], and the proposed Circular Electron Positron Collider (CEPC) [3]. Thus our ultimate goal is to provide a framework that helps deepen the understanding of spin dynamics of ultra relativistic particles in storage rings and accelerators. These particles do not uniformly fill the vacuum chamber of an accelerator but travel in one or more bunches of 10^{10} or more particles.

Electrons (positrons) carry an intrinsic quantum angular momentum and in a classical picture it is convenient to imagine them to be tiny gyroscopes spinning on their axes. They also carry an intrinsic magnetic moment parallel to the angular momentum vector. For a full description of the quantum mechanics of these particles the reader is referred to standard text books, see for example [4]. However, this work focusses on the above-mentioned polarization and for that it is sufficient to concentrate on the so-called single-particle spin expectation value. This vector lies parallel to the intrinsic angular momentum vector and we normalize its magnitude to unity to obtain a vector which we call the *spin* and denote by $\hat{\mathcal{S}}$. The polarization of an ensemble is the ensemble average of the spin, denoted by $\langle\hat{\mathcal{S}}\rangle$. Since the spin dynamics of electrons and positrons is the same, we concentrate here mostly on electrons.

As indicated, the spin polarization of a bunch of particles is our main quantity of interest. This quantity defines the quality of a bunch for conducting certain spin-

sensitive collision experiments and for bunch energy measurements [5]. As an example, highly polarized electron or positron bunches are helpful for understanding the structure of protons and nuclei via electron(positron)-proton or electron(positron)-nuclei collision experiments. A high polarization means a high alignment of the spins in the bunch without which the spin related measurements do not deliver sufficient information. Thus, the main questions for any proposed high-energy ring are: (Q1) Can one get high polarization? (Q2) What are the theoretical limits of the polarization?

Let us now look at the dynamics. The electric and magnetic fields in a storage ring couple to the magnetic moments of electrons and exert a torque on the intrinsic angular momenta, causing the spins to precess. Moreover, electrons moving in the magnetic fields in storage rings emit streams of photons, in the direction of the particle's momentum, known as synchrotron radiation.

The spin precession is described by the Thomas-Bargmann-Michel-Telegdi equation (Thomas-BMT) [6] and this will be presented later. The photon emission in synchrotron radiation affects the orbital motion of electrons in a storage ring and this can lead to an equilibrium particle distribution in phase space of a bunch. This is modeled by adding noise and damping to the particle motion [7, 8]. The photon emission also affects the spin motion and this can lead to the build-up of spin polarization which can reach an equilibrium resulting from a balance of three factors, namely the so-called Sokolov-Ternov process, depolarization and the so-called kinetic polarization effect.

The Sokolov-Ternov process [9] causes a build up of the polarization due to an asymmetry in the spin-flip transition rates for spin up and spin down along a spin-quantization axis. This effect was originally derived from the Dirac equation. The depolarization can be viewed as a consequence of the fact that photon emission is stochastic and puts noise into the particle trajectories which then feeds through to

the spin motion via the spin-orbit coupling embodied in the Thomas-BMT equation. This causes the spins to spread out randomly ("spin diffusion") so that the there is a tendency for the polarization to fall. The kinetic polarization is also a result of spin-orbit coupling. Without the synchrotron radiation, the spin motion would be deterministic along trajectories.

We present three approaches (A1, A2 and A3) to these effects of synchrotron radiation, where A1 and A2 go back to the 1970s and A3 is new and is studied here in detail. A1 is based on formalism in [10] by Derbenev and Kondratenko (see also [11]) and A2 is based on results in [12], also by Derbenev and Kondratenko. Here we discuss these two approaches and then introduce A3, formulated via stochastic differential equations (SDEs).

So far, analytic and resulting numerical estimates of the attainable polarization have been based on the aforementioned formalism in [10] via the so-called *Derbenev-Kondratenko formulas* [13]. This is A1. A recent overview of these estimates can be found in [14]. The assumption is that the polarization across phase space is aligned parallel to a field of spin-quantization axes, the so-called *invariant spin field* (ISF), [15]. Thus the assumption is that the polarization local to a point in phase space, which we call $\vec{P}_{\text{loc}}$, is parallel to the ISF at that point (see also Remark 4 in Chapter 2).

For the future, a third question (Q3) for high-energy rings like the FCC-ee and CEPC is: are the Derbenev-Kondratenko formulas complete? We believe that the approach A1, based on the Derbenev-Kondratenko formulas, is an approximation of A2 from [12], mentioned above, which is, in turn, based on the phase-space density f of a bunch and the so-called *polarization density* of the bunch. The polarization density at a point in phase space is, by definition, the product of $\vec{P}_{\text{loc}}$ at that point and the phase-space density f at that point and it is proportional to the spin angular-momentum density in phase space at that point. The integral over phase

space of the polarization density is the polarization vector of the bunch. In this approach one studies the evolution of the phase-space density by solving (analytically or numerically) the orbital Fokker-Planck equation. The corresponding equation for spin is the evolution equation for the polarization density as introduced by Derbenev and Kondratenko in 1975 [12] as a generalization to the whole phase space (with its noisy trajectories) of the Baier-Katkov-Strakhovenko (BKS) equation which just describes the evolution of polarization by spin flip along a single deterministic trajectory [16, 14]. We call this the Bloch equation (BE) to reflect the analogy with equations for magnetization in condensed matter, [17]. [1] The BE is a system of three Fokker-Planck-like equations for the three components of the polarization density which is coupled by a Thomas-BMT term and the BKS terms but uncoupled within the Fokker-Planck terms. In particular, in addition to the Thomas-BMT motion, it takes into account effects on spin due to synchrotron radiation including the depolarization effect, the Sokolov-Ternov effect with its Baier-Katkov (BK) correction, as well as the kinetic-polarization effect. Thus in A2 we study the initial-value problem for the coupled system consisting of the orbital Fokker-Planck and the Bloch equation. See [14] and [19] for recent reviews of polarization history and phenomenology.

A3 is based on a system of coupled spin-orbit SDEs and their associated Fokker-Planck equation which governs the evolution of the (joint) spin-phase-space probability density. The SDEs of A3 lead to the orbital Fokker-Planck equation and the Bloch equation of A2, i.e. the ones based on [12]. Therefore no information from A2 is lost, but we believe that the third approach is more amenable to analysis.

[1] Note that, for example, in [18] we use the term "Full Bloch equation" instead of simply "Bloch equation".

1.2 Overview

Chapters 2-10 are organized as follows. In Chapter 2 we present, in the laboratory frame (the lab frame), the equations underlying the second and the third approach (all these equations were established in the lab frame). Thus Chapter 2 contains the spin-orbit SDEs, their associated spin-orbit Fokker-Planck equation, the orbital Fokker-Planck equation and the Bloch equation. As a final result of Chapter 2, we state the *reduced* Bloch equation which is instrumental for computing the depolarization time. Note that we say that equations which do not include the Sokolov- Ternov effect, its BK correction and the kinetic polarization are "reduced". Chapter 2 includes the definitions of all necessary quantities. Note that *lab frame* entails the use of laboratory Cartesian coordinates for the orbital variables from special relativity, where the independent variable is time.

In Chapter 3, we introduce the beam frame. This is a generalization of a Frenet-Serret frame which is more convenient for describing the spin-orbit dynamics. In the beam frame the underlying reference curve is a closed curve in $\mathbb{R}^3$ located in the middle of the vacuum chamber of a storage ring. So the beam frame is the most suitable for our analysis since the phase-space coordinates of particles are small, facilitating approximations. In this frame the independent variable is the accelerator azimuth $\theta = 2\pi s/C$, where C is the length of the closed reference curve and where s is the path length variable associated with the reference curve. This involves the transformation from t to θ, see [20, 21]. In Chapter 3 we do our first approximation: we linearize the spin-orbit SDEs with respect to (w.r.t.) the orbital beam-frame variables (but not w.r.t. the spin variables!). All numerical calculations will be performed using the beam frame and related frames and at this stage we focus the numerical calculations on the depolarization. So for that purpose we ignore the terms associated with the other effects, e.g, the Sokolov-Ternov effect. In fact Chapter 3 sets up the main framework for studying the depolarization and defines the

main quantities of interest: the polarization and the polarization density introduced informally earlier. There is also some supplementary analysis of the equilibrium orbital distribution used in later chapters.

In Chapter 4 we use our framework to consider an approximation to the polarization density which is inspired by [12] and which is based on what we call the ISF approximation [10, 14, 19]. Chapter 4 formalizes the assumptions made to obtain the approximation, and also gives an equation for the error which can be used to obtain the correction to the ISF approximation. This approximation supplemented by the correction provides an excellent background for validation of the numerical algorithm that we develop in Chapter 7. Note that Chapter 4 combines A1 and A2.

In Chapter 5, we develop approximations based on the so-called method of averaging (MOA), that can be applied to the systems modeling real accelerators and provides models which we call effective models and which are suitable for our numerical approach. Ultimately, by interfacing with modern accelerator software, like Bmad [22], through the MOA, we will extend the work of this thesis to study the spin dynamics in high energy storage rings and design optimal solutions to the problems linked to depolarization.

In Chapter 6 we define two simple models that are both interesting physically and numerically: a one-degree-of-freedom model (SM1) and its extention to three degrees of freedom (SM3). These models are inspired by the *single resonance model* [15, 23], describing the spin motion in the presence of vertical betatron motion. See also the original work on the so-called rotating wave approximation [24]. We call these models *simple* to emphasise that the ISF approximation can be applied directly to them and that, in addition, the associated reduced Bloch equations are similar to the effective Bloch equation obtained via the MOA. Nevertheless, from the point of view of numerical analysis these models are challenging w.r.t. the anticipated computational cost. At the end of Chapter 6 we assign realistic parameters for the

SM1 and demonstrate its behaviour using the ISF approximation.

In Chapter 7 we present a numerical method for solving the Bloch equation and demonstrate it by evolving the simple models from Chapter 6. Our method computes the spectral approximation to the polarization density by numerically solving the reduced Bloch equation. The term *spectral* means that the polarization density is approximated by a finite sum of the polynomials in the phase space. The number of polynomials in each space dimension is the order of the approximation. For these types of methods, the convergence rate gradually increases with the increase of the polynomial degree if the solution is smooth. In this chapter, the results from Chapter 4 for SM1 and a simple 1-degree-of-freedom model described via the effective Bloch equation are used to test the accuracy of the method in one degree of freedom. SM3 is used to demonstrate the accuracy in three degrees of freedom.

In Chapter 8 we present our study of SM1 taking into account the ISF approximation of Chapter 4 and using the (model-independent) spectral method of Chapter 7. We perform numerical experiments where we compare the results of the ISF approximation for SM1 with the results obtained via the numerical solution of the reduced Bloch equation and investigate the behaviour of the error of the ISF approximation with a set of model parameters. We use the results in the last summarizing experiment to address the sensitivity of the ISF approximation for the SM1 to the spin-precession rate.

In Chapter 9 we summarize our work on spin dynamics in modern electron storage rings and discuss planned extensions to this work.

Chapter 10 deals with a very different topic, namely a high-order accurate numerical method for the wave equation that combines efficient Hermite methods with geometrically flexible discontinuous-Galerkin methods by using overset grids. Near boundaries we use thin boundary-fitted curvilinear grids and in the volume we use

Cartesian grids so that the computational complexity of the solvers approaches that of a structured Cartesian Hermite method. In contrast to many other overset methods we do not need to add artificial dissipation since we find that the built-in dissipation of the Hermite and discontinuous-Galerkin methods is sufficient to maintain stability. Using numerical experiments we demonstrate the stability, accuracy, efficiency and applicability of the methods to forward and inverse problems.

Chapter 2

Spin-orbit motion in the laboratory frame

In this chapter we present the lab-frame equations underlying the second and the third approaches, A2 and A3, mentioned in Chapter 1. We proceed as follows. In Section 2.1 we introduce the third approach which is based on a system of spin-orbit stochastic differential equations. In Section 2.2 we introduce the second approach and show its close relation to the third approach.

2.1 Spin-orbit dynamics via random processes

In A3 a charged particle with position $\vec{r}$ and momentum $\vec{p}$ obeys a system of stochastic differential equations (SDEs) of Itô type. Using the units as in [12], the Itô SDEs

and their initial conditions can be written informally in Langevin form as in [18],

$$\dot{\vec{r}} = \frac{1}{m\gamma(\vec{p})}\vec{p}, \tag{2.1}$$

$$\dot{\vec{p}} = e\vec{E}(t,\vec{r}) + \frac{e}{m\gamma(\vec{p})}(\vec{p}\times\vec{B}(t,\vec{r})) + \vec{F}_{\text{rad}}(t,\vec{r},\vec{p}) + \vec{Q}_{\text{rad}}(t,\vec{r},\vec{p}) + \vec{\mathcal{B}}^{\text{orb}}(t,\vec{r},\vec{p})\xi(t), \tag{2.2}$$

$$\vec{r}(0) = \vec{r}_0, \tag{2.3}$$

$$\vec{p}(0) = \vec{p}_0. \tag{2.4}$$

where $\gamma(\vec{p}) = \frac{1}{m}\sqrt{|\vec{p}|^2 + m^2}$ is the Lorenz factor, e and m are the charge and rest mass of the electron or positron with $\vec{E}, \vec{B}$ being the external electric and magnetic fields. As usual, since it is minuscule compared to all other forces, the effect of the spin on the orbit, i.e the Stern-Gerlach effect, is neglected in (2.1)–(2.2). The initial values $\vec{p}_0$ and $\vec{r}_0$ are random vectors with a joint probability density function describing the initial bunch distribution and ξ is a scalar white noise process, accounting for the quantum fluctuations due to photon emission in the synchrotron radiation. The synchrotron-radiation contribution to the particle motion is taken into account via the terms

$$\vec{\mathcal{B}}^{\text{orb}}(t,\vec{r},\vec{p}) := \vec{p}\sqrt{\frac{55}{24\sqrt{3}}\lambda(t,\vec{r},\vec{p})},$$

$$\vec{F}_{\text{rad}}(t,\vec{r},\vec{p}) := -\frac{2}{3}\frac{e^4}{m^5\gamma(\vec{p})}|\vec{p}\times\vec{B}(t,\vec{r})|^2\vec{p},$$

$$Q_{\text{rad},i}(t,\vec{r},\vec{p}) := \frac{55}{48\sqrt{3}}\sum_{j=1}^{3}\frac{\partial[\lambda(t,\vec{r},\vec{p})p_ip_j]}{\partial p_j},$$

$$\lambda(t,\vec{r},\vec{p}) := \frac{\hbar|e|^5}{m^8\gamma(\vec{p})}|\vec{p}\times\vec{B}(t,\vec{r})|^3,$$

where the $\hbar$ in $\lambda(t,\vec{r},\vec{p})$ reveals the quantum nature of the synchrotron-radiation effects. Also $\vec{F}_{\text{rad}}$ is the classical radiation reaction force due to the synchrotron radiation and $\vec{Q}_{\text{rad}}$ is a quantum correction to $\vec{F}_{\text{rad}}$. The initial value problem (2.1),

(2.2), (2.3) and (2.4) can be written concisely as

$$\dot{Z} = F(t, Z) + G(t, Z)\xi(t), \quad Z(0) = Z_0. \tag{2.5}$$

To be precise, the stochastic process $Z = (\vec{r}, \vec{p})^{\mathrm{T}}$ evolves according to the integral equation

$$Z(t) = Z(0) + \int_0^t F(\tau, Z(\tau))d\tau + \int_0^t G(\tau, Z(\tau))d\mathcal{W}(\tau), \tag{2.6}$$

where the second integral in (2.6) is the so-called Itô integral and $\mathcal{W}$ is a Wiener process.

Note that in (2.5), and from now on, the dependent variables in the SDEs are denoted by capital letters. In contrast, independent variables are denoted by lowercase letters. We note that (2.5) is ambiguous. It is common to interprete (2.5) as either an Itô system of SDEs or a Stratonovich system of SDEs, leading to different Fokker-Planck equations if G depends on z. In this work all SDEs are to be interpreted in the Itô sense. Helpful discussions about Itô SDEs can be found, for example, in [25, 26, 27]. However it is sufficient for the reader to know that there is a unique Fokker-Planck equation associated with a system of Itô SDEs.

As mentioned in the Introduction, in the absence of the effects of synchrotron radiation, a spin $\hat{\mathcal{S}}$ precesses according to the Thomas-BMT equation [6] which we write in the form

$$\dot{\hat{\mathcal{S}}} = \vec{W}(\vec{p}, \vec{B}(t, \vec{r}), \vec{E}(t, \vec{r})) \times \hat{\mathcal{S}}\,. \tag{2.7}$$

This displays the dependence of $\vec{W}$ on the external electric and magnetic fields and the velocity and energy (via $\vec{p}$).

We now present our recently discovered SDE for describing spin motion in the presence of the Sokolov-Ternov effect, its Baier-Katkov correction and the kinetic

polarization. For this we introduce the vector $\vec{S}$ which we define as the three-dimensional stochastic process governed by

$$\dot{\vec{S}} = M(t,Z)\vec{S} + \vec{\mathcal{D}}^{\,\mathrm{spin}}(t,Z) + \vec{\mathcal{B}}^{\,\mathrm{kin}}(t,Z)\xi(t), \quad \vec{S}(0) = \vec{S}_0, \tag{2.8}$$

where

$$\begin{aligned}
M(t,z) &:= W(t,z) - \lambda(t,z)\frac{5\sqrt{3}}{8}[I_{3\times 3} - \frac{2}{9m^2\gamma^2(\vec{p})}\vec{p}\vec{p}^{\mathrm{T}}], \qquad (2.9)\\
\vec{a}(t,z) &:= \frac{e}{m^2\gamma^2(\vec{p})}(\vec{p}\times\vec{B}(t,\vec{r})),\\
\vec{\mathcal{D}}^{\,\mathrm{spin}}(t,z) &:= -\lambda(t,z)\frac{1}{m\gamma(\vec{p})}\frac{\vec{p}\times\vec{a}(t,z)}{|\vec{a}(t,z)|},\\
\vec{\mathcal{B}}^{\,\mathrm{kin}}(t,z) &:= \frac{1}{m\gamma(\vec{p})}\frac{\vec{p}\times\vec{a}(t,z)}{|\vec{a}(t,z)|}\sqrt{\frac{24\sqrt{3}}{55}\lambda(t,z)}.
\end{aligned}$$

The skew-symmetric matrix $W(t,z)$ is equivalent to the $\vec{W}$ in (2.7) and accounts for the Thomas-BMT spin-orbit coupling and thereby the depolarization as mentioned in the Introduction. The terms $M(t,Z)$, $\vec{\mathcal{B}}^{\,\mathrm{kin}}(t,z)$, $\vec{\mathcal{D}}^{\,\mathrm{spin}}(t,z)$ in (2.8) are chosen so that they deliver the required BE to be described in Section 2.2. The terms $-\lambda(t,Z)\frac{5\sqrt{3}}{8}\vec{S}$ and $\vec{\mathcal{D}}^{\,\mathrm{spin}}(t,Z)$ will account for spin flips due to synchrotron radiation and encapsulate the Sokolov-Ternov effect. The term proportional to $2/9$ in (2.9) will account for the Baier-Katkov correction, and the white-noise term $\vec{\mathcal{B}}^{\,\mathrm{kin}}(t,Z)\xi(t)$ will account for the kinetic-polarization effect. The latter motivates the use of the superscript "kin". As the notation suggests, the white-noise process $\xi(t)$ in (2.8) is the same as the white-noise process $\xi(t)$ in (2.2).

The system of SDEs for the joint process $(Z,\vec{S})$ is written as

$$\begin{pmatrix} \dot{Z} \\ \dot{\vec{S}} \end{pmatrix} = H(t,Z,\vec{S}) + N(t,Z)\xi(t), \tag{2.10}$$

where

$$H(t,Z,\vec{S}) = \begin{pmatrix} F(t,Z) \\ M(t,Z)\vec{S} + \vec{\mathcal{D}}^{\,\mathrm{spin}}(t,Z) \end{pmatrix}, \quad N(t,Z) = \begin{pmatrix} G(t,Z) \\ \vec{\mathcal{B}}^{\,\mathrm{kin}}(t,Z) \end{pmatrix},$$

and we remind the reader that the SDE is to be interpreted as an Itô SDE. Note that (2.10) is equivalent to the combined SDEs (2.1), (2.2) and (2.8).

Remark 1. *Note that* $|\vec{S}(t)|$ *in* (2.8) *is not conserved in time. So* $\vec{S}(t)$ *in* (2.8) *is not the spin vector of a single particle, but is an average. In particular* $|\vec{S}(t)| \leq 1$*. Nevertheless,* $\vec{S}(t)$ *can be related to familiar quantities and we generally stick to the terminology "spin vector".*

In fact, as we shall see below the polarization vector of the bunch at time t *is the expected value of the random vector* $\vec{S}(t)$*, i.e.,* $\vec{P}(t) = \langle \vec{S}(t) \rangle$ *with* $\vec{S}(t)$ *defined by* (2.8)*. Thus, and since* $|\vec{P}(t)| \leq 1$*, we obtain* $|\langle \vec{S}(t) \rangle| \leq 1$*. In particular the constraint on the initial condition is:* $|\langle \vec{S}(0) \rangle| \leq 1$*.*

As an alternative of trying to analyze the SDEs analytically, one can use Monte-Carlo spin-orbit tracking. The conventional Monte-Carlo spin tracking algorithms simulate stochastic photon emission and concentrate on computing the rate of the radiative depolarization. For example SLICKTRACK by D.P. Barber which is used in [28, 29] (see the sections on polarization), SITROS by J. Kewisch [30], Zgoubi by F. Meot [31], PTC/FPP by E. Forest [32], and Bmad by D. Sagan [22] simulate the spin diffusion and they are based on, or closely related to, the so-called reduced SDEs that are obtained from (2.10) by removing the kinetic polarization and the Sokolov-Ternov effect, [13, 31, 33, 34]. In fact, Monte-Carlo algorithms have been used since the 1980s and as just mentioned they are in effect based on the reduced SDEs.

Remark 2. *One can use* (2.10) *(and thus:* (2.5) *and* (2.8)*) as the basis for a Monte-Carlo spin tracking algorithm for* $\vec{P}(t)$ *to extend the standard Monte-Carlo spin tracking algorithms by taking into account all physical effects described, like the Sokolov-Ternov effect, the Baier-Katkov correction, the kinetic-polarization effect and, of course, spin diffusion. This will be an important part of our future work.*

Remark 3. *In our study we ignore collective effects such as the beam-beam interaction and coherent synchrotron radiation, that may be a subject of the future extensions to* (2.1) *and* (2.2)*.*

The Fokker-Planck equation associated with the process $(Z, \vec{S})$ in (2.10) evolves the (joint) density $\mathcal{P} = \mathcal{P}(t, z, \vec{s})$ and thus reads as

$$\begin{aligned}\partial_t \mathcal{P} =& L_{\mathrm{FP}}\mathcal{P} - \sum_{i=1}^{3} \frac{\partial}{\partial s_i}\left(\left(M(t,z)\vec{s} + \vec{\mathcal{D}}^{spin}(t,z)\right)_i \mathcal{P}\right) \\ &+ \sum_{i,j=1}^{3} \frac{\partial^2}{\partial s_i \partial p_j}\left((\vec{\mathcal{B}}^{kin}(t,z))_i (\vec{\mathcal{B}}^{orb}(t,z))_j \mathcal{P}\right) \\ &+ \frac{1}{2}\sum_{i,j=1}^{3} \frac{\partial^2}{\partial s_i \partial s_j}\left((\vec{\mathcal{B}}^{kin}(t,z))_i (\vec{\mathcal{B}}^{kin}(t,z))_j \mathcal{P}\right), \qquad (2.11) \\ \mathcal{P}(0,z,s) =& \mathcal{P}_0(z,s),\end{aligned}$$

where $\mathcal{P}_0$ is an initial joint spin-phase-space density of the bunch. The Fokker-Planck operator L_{FP} is defined by

$$\begin{aligned}L_{\mathrm{FP}} :=& -\nabla_{\vec{r}} \cdot \frac{1}{m\gamma(\vec{p})}\vec{p} - \nabla_{\vec{p}} \cdot [e\vec{E}(t,\vec{r}) + \frac{e}{m\gamma(\vec{p})}(\vec{p} \times \vec{B}(t,\vec{r})) \\ &+ \vec{F}_{\mathrm{rad}}(t,z) + \vec{Q}_{\mathrm{rad}}(t,z)] + \frac{1}{2}\sum_{i,j=1}^{3} \partial_{p_i}\partial_{p_j}\mathcal{E}_{ij}(t,z),\end{aligned}$$

where

$$\mathcal{E}_{ij}(t,z) := \frac{55}{24\sqrt{3}}\lambda(t,z)p_i p_j.$$

The terms $\mathcal{E}_{i,j}$ are the so-called parabolic Fokker-Planck terms.

2.2 The orbital Fokker-Planck equation. The Bloch equation for polarization density

The phase-space density f is given by the integral of the joint density over the spin variable, i.e

$$f(t,z) := \int_{\mathbb{R}^3} \mathcal{P}(t,z,\vec{s})\, d\vec{s}. \tag{2.12}$$

The Fokker-Planck equation for the orbit is decoupled from the spin so that by integrating (2.11) w.r.t $\vec{s}$ and taking into account (2.12) we obtain the orbital Fokker-Planck equation for the phase-space density f

$$\begin{aligned} \partial_t f &= L_{\mathrm{FP}} f, \\ f(0,z) &= f_0(z), \end{aligned} \tag{2.13}$$

where f_0 is the probability density function of Z_0. Consistency requires this to be the Fokker-Planck equation for the SDE in (2.5) and it is since (2.5) does not depend on spin.

The Fokker-Planck operator L_{FP} whose explicit form is taken from [12] is a linear second-order partial differential operator and, with some additional approximations, is commonly used in formalizations of the orbital motion in electron synchrotrons and storage rings, see [7] and [35].

We write the polarization density mentioned in the Introduction as

$$\vec{\eta}(t,z) = \int \vec{s}\,\mathcal{P}(t,z,\vec{s})\, d\vec{s}, \tag{2.14}$$

and it can be shown that it evolves via the lab-frame Bloch equation (BE) originally given in [12]

$$\begin{aligned} \partial_t \vec{\eta} &= L_{\mathrm{FP}}\vec{\eta} + M(t,z)\vec{\eta} - [1 + \nabla_{\vec{p}} \cdot \vec{p}]\lambda(t,z)\frac{1}{m\gamma(\vec{p})}\frac{\vec{p}\times\vec{a}(t,z)}{|\vec{a}(t,z)|} f(t,z), \\ \vec{\eta}(0,z) &= \int_{\mathbb{R}^3} \vec{s}\,\mathcal{P}_0(z,\vec{s})\, d\vec{s}. \end{aligned} \tag{2.15}$$

In fact differentiating (2.14) w.r.t. t and using (2.11), (2.12) and (2.14) results in (2.15). Thus f written as in (2.12) and $\vec{\eta}$ written as in (2.14) in terms of $\vec{s}$ are indeed what we need for the orbital Fokker-Planck equation and BE. This in fact shows that the SDEs (2.1) and (2.2) contain the information of (2.13) and (2.15), and are thus consistent with [12]. Also, (2.14) implies that the polarization vector of a bunch reads as

$$\vec{P}(t) = \int_{\mathbb{R}^6} \vec{\eta}(t,z)\, dz\ .$$

Remark 4. *Recall from Remark 1, that $|\vec{S}(t)| \leq 1$. So if $\mathcal{P}$ is a physically meaningful density, then $\mathcal{P}(t,z,\vec{s}) = 0$ if $|\vec{s}| > 1$, so that by* (2.12) *and* (2.14)

$$|\vec{\eta}(t,z)| \leq \int_{\mathbb{R}^3} |\vec{s}|\mathcal{P}(t,z,\vec{s})\, d\vec{s} \leq \int_{\mathbb{R}^3} \mathcal{P}(t,z,\vec{s})\, d\vec{s} = f(t,z).$$

Thus the local polarization field defined as $\vec{P}_{\mathrm{loc}} := \vec{\eta}/f$ satisfies $|\vec{P}_{\mathrm{loc}}(t,z)| \leq 1$. Indeed, by the meaning of f and $\vec{\eta}$, the quantity $\vec{P}_{\mathrm{loc}}(t,z)$ is the bunch polarization at (t,z). So its size has to be less or equal to 1.

Since $\vec{\eta}$ is proportional to the probability density function of Z, we consider the solutions that rapidly decay as z approaches infinity

$$\lim_{z\to\infty} \vec{\eta}(t,z)e^{\alpha|z|^2} = 0,$$

for some $\alpha > 0$. Recall that the quantum aspects are embodied in the terms in (2.15) containing λ since λ is proportional to $\hbar$. The term

$$-\lambda(t,z)\frac{5\sqrt{3}}{8}\vec{\eta},$$

in $M(t,z)$ and the term $\lambda(t,z)\frac{1}{m\gamma(\vec{p})}\frac{\vec{p}\times\vec{a}(t,z)}{|\vec{a}(t,z)|}f(t,z)$ take into account spin flips due to synchrotron radiation and encapsulate the Sokolov-Ternov effect. The term

$$\lambda(t,z)\frac{5\sqrt{3}}{8}\frac{2}{9m^2\gamma^2(\vec{p})}\vec{p}\vec{p}^{\mathrm{T}}\vec{\eta},$$

encapsulates the Baier-Katkov correction, and the term

$$\nabla_{\vec{p}} \cdot \vec{p}\, \lambda(t,z) \frac{1}{m\gamma(\vec{p})} \frac{\vec{p} \times \vec{a}(t,z)}{|\vec{a}(t,z)|} f(t,z) = \sum_{1}^{3} \partial_{p_i} [p_i\ \lambda(t,z) \frac{1}{m\gamma(\vec{p})} \frac{\vec{p} \times \vec{a}(t,z)}{|\vec{a}(t,z)|} f(t,z)],$$

encapsulates the kinetic-polarization effect. Each of these is proportional to λ and hence proportional to $\hbar$.

If we ignore the spin flip terms and the kinetic-polarization term in the BE then (2.15) simplifies to

$$\partial_t \vec{\eta} = L_{\mathrm{FP}} \vec{\eta} + W(t,z)\vec{\eta}. \tag{2.16}$$

We refer to (2.16) as the reduced Bloch equation (RBE).

The RBE models spin diffusion due to the spin-orbit coupling. The RBE is sufficient for computing the depolarization time and it shares the terms with the BE that are most challenging to discretize with our numerical algorithm.

Remark 5. *The equations* (2.13) *and* (2.15) *were derived in [12] from quantum electrodynamics, using the semiclassical approximation of the Foldy-Wouthuysen transformation of the Dirac Hamiltonian and finally by making a Markov approximation (see also [36]). The SDEs* (2.5), (2.8) *contain the whole information about* (2.13) *and* (2.15). *In fact* (2.5), (2.8) *were obtained in [18] via reverse engineering of* (2.13) *and* (2.15). *In the special case where one neglects all spin flip effects and the kinetic-polarization effect the corresponding SDEs (and thus the RBE) can be derived purely classically as in [37].*

When the particle motion is governed just by a Hamiltonian, as in the case of protons and other heavy particles where one neglects all synchrotron radiation effects, the phase-space density is conserved along a trajectory. Then, since $\vec{P}_{\mathrm{loc}}$ obeys the Thomas-BMT equation along each trajectory the polarization density does too. In other words, since $t \mapsto \vec{P}_{\mathrm{loc}}(t, z(t))$ obeys the Thomas-BMT equation so does

$t \mapsto \vec{\eta}(t, z(t))$. Moreover $|\vec{s}(t)|$ is constant. In fact in this case the equations of motion (2.1)–(2.4) become

$$\begin{aligned}
\dot{\vec{r}} &= \frac{1}{m\gamma(\vec{p})}\vec{p}, \quad \vec{r}(0) = \vec{r}_0 \\
\dot{\vec{p}} &= q(\vec{E}(t, \vec{r}) + \frac{1}{m\gamma(\vec{p})}(\vec{p} \times \vec{B}(t, \vec{r}))), \quad \vec{p}(0) = \vec{p}_0 \\
\dot{\vec{s}} &= W(t, \vec{r}, \vec{p})\vec{s}, \quad \vec{s}(0) = \vec{s}_0,
\end{aligned}$$

where $\vec{p}_0, \vec{r}_0, \vec{s}_0$ are random vectors with probability density functions describing the initial bunch and spin distributions.

Note that since in this case $|\vec{s}(t)|$ is constant, it can be interpreted as the unit vector $\hat{\mathcal{S}}$. In fact, it is these equations, involving $\hat{\mathcal{S}}$, that are the basis of the Monte-Carlo methods discussed earlier.

Chapter 3

Spin-orbit motion in the beam frame

Next, and for the remainder of this thesis, we work in the so-called beam frame, namely a Frenet-Serret coordinate system following the (closed) design orbit of the ring. For this we change the independent variable from the time t to the azimuthal position on the ring, θ. In the beam frame, i.e., in accelerator coordinates y, centered at the reference particle at azimuth θ, particle position, momentum and spin are governed by the system of SDEs

$$Y' = f_{\mathrm{b}}(\theta, Y) + g_{\mathrm{b}}(\theta, Y)\xi(\theta), \tag{3.1}$$

$$\vec{S}' = \underbrace{W_{\mathrm{b}}(\theta, Y)\vec{S}}_{\text{T-BMT}} + \underbrace{M_{\mathrm{b}}(\theta, Y)\vec{S} + G_{\mathrm{b}}(\theta, Y) + H_{\mathrm{b}}(\theta, Y)\xi(\theta)}_{\text{ST effect, BK correction, kinetic polarization}}, \tag{3.2}$$

$$Y(0) = Y_0,$$

$$\vec{S}(0) = \vec{S}_0,$$

where the primes are used to denote the derivative w.r.t. θ, $Y \in \mathbb{R}^{2d}, \vec{S} \in \mathbb{R}^3$, and $f_{\mathrm{b}}(\theta, Y) \in \mathbb{R}^{2d}$, $g_{\mathrm{b}}(\theta, Y) \in \mathbb{R}^{2d\times m}$, $W_{\mathrm{b}}(\theta, Y), M_{\mathrm{b}}(\theta, Y) \in \mathbb{R}^{3\times 3}, G_{\mathrm{b}}(\theta, Y), H_{\mathrm{b}}(\theta, Y) \in \mathbb{R}^{3\times m}$ are 2π-periodic in θ, with $W_{\mathrm{b}}(\theta, Y)$ being skew-symmetric and $d = 1, 2$ or 3

being the number of degrees of freedom. As in [38], if $d = 3$ then the first four components of Y are the transverse positions and transverse momenta of a particle, the fifth component is the longitudinal position, and the sixth component of Y is $(\gamma - \gamma_r)/\gamma_r$ where γ is the Lorenz factor and γ_r is the reference value of γ. Here the ξ is an m-dimensional white noise process (for $m = 1$ it is a scalar white noise process). For $d = 3$ and $m = 1$, (3.1) and (3.2) are obtained by transforming (2.10) from the lab frame to the beam frame. The cases $d \neq 3$, $m \neq 1$ are needed for the simple models in later chapters.

In this work, when it comes to the beam frame, we will focus on the so-called reduced system, i.e., the case where one neglects the Sokolov-Ternov effect, its BK correction and the kinetic polarization effect. Thus $M_{\mathrm{b}}, G_{\mathrm{b}}, H_{\mathrm{b}}$ will be neglected, and they will be the subject of future work. In fact this thesis sets the framework for future extensions.

3.1 The reduced stochastic differential equations

The reduced SDEs in the beam frame are obtained from (3.1) and (3.2) as was mentioned, by neglecting $M_{\mathrm{b}}, G_{\mathrm{b}}, H_{\mathrm{b}}$ to obtain

$$Y' = f_{\mathrm{b}}(\theta, Y) + g_{\mathrm{b}}(\theta, Y)\xi(\theta), \tag{3.3}$$

$$\vec{S}' = W_{\mathrm{b}}(\theta, Y)\vec{S}, \tag{3.4}$$

By linearizing (3.3) and (3.4) w.r.t. Y we obtain

$$Y' = (A(\theta) + \varepsilon\delta A(\theta))Y + \sqrt{\varepsilon}B(\theta)\xi(\theta), \tag{3.5}$$

$$\vec{S}' = \Omega(\theta, Y)\vec{S}, \quad \Omega(\theta, Y) = \Omega_0(\theta) + \sum_{j=1}^{2d} Y_j\Omega_j(\theta), \quad d = 1, 2 \text{ or } 3, \tag{3.6}$$

subject to the initial conditions

$$Y(0) = Y_0,$$
$$\vec{S}(0) = \vec{S}_0.$$

Here $A(\theta), \delta A(\theta) \in \mathbb{R}^{2d\times 2d}$, $B(\theta) \in \mathbb{R}^{2d\times m}$ and $\Omega(\theta, Y) \in \mathbb{R}^{3\times 3}$. Since $f_{\rm b}(\theta, Y)$, $g_{\rm b}(\theta, Y)$ and $W_{\rm b}(\theta, Y)$ in (3.3) and (3.4) are 2π-periodic in θ so are $A(\theta)$, $\delta A(\theta)$, $B(\theta)$ and $\Omega(\theta, Y)$ in (3.5) and (3.6). Without loss of generality we assume Y to be $\mathcal{O}(1)$ (since (3.5) can always be rescaled) and the parameter ε is chosen such that δA and B are $\mathcal{O}(1)$. $A(\theta)$ is a Hamiltonian matrix, i.e.

$$(J_{2d}A(\theta))^{\rm T} = J_{2d}A(\theta), \quad J_{2d} := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \otimes I_d,$$

so that

$$\mathrm{Tr}[A(\theta)] = 0. \tag{3.7}$$

Here I_d is the $d \times d$ identity matrix and $\otimes$ denotes the Kronecker product. Equation (3.5) describes the orbital motion, which similarly to (2.1)–(2.2) can be separated into three fundamental components. First, the matrix $A(\theta)$ is the Hamiltonian matrix associated with the motion of a nonradiative particle, which is deterministic. Second, the non-Hamiltonian matrix $\varepsilon\delta A$ is associated with the motion due to classical synchrotron radiation effects (also deterministic). In particular $\varepsilon\delta A$ contains the damping terms associated with synchrotron radiation (damping w.r.t. the reference orbit) see, e.g., (5.3) in [38]. The third component of the orbital motion contains the quantum fluctuations from synchrotron radiation encapsulated in $\varepsilon B(\theta)$ multiplying the white noise $\xi(\theta)$. In (3.5) the δA terms and the B terms are balanced at $O(\varepsilon)$ and so can be treated together in first order perturbation theory, see Chapters 4–6. This is the reason for the $\sqrt{\varepsilon}$ in (3.5). However this balance is also physical since the damping and diffusion come from the same source.

Equation (3.6) describes the spin motion due to Thomas-BMT precession.

$\Omega(\theta, Y)$ is skew-symmetric and it is (affinely) linear in Y (as in [38]). It is important to note that, although $\Omega(\theta, Y)$ is linear in Y, the right hand side of (3.6) is bilinear in Y and $\vec{S}$, so that the spin-orbit motion is nonlinear (although the orbital motion is linear).

The process defined by the Itô system (3.5) is a Gaussian process if Y_0 is a Gaussian random variable. See [25, 26], where (3.5) is called a "narrow-sense linear" SDE since B is independent of Y. As an aside, a further consequence of the latter is that the Itô and Stratonovich interpretations of (3.5) and (3.6) are the same.

3.2 The spin-orbit Fokker-Planck equation and orbital Fokker-Planck equation

With (3.5) and (3.6) the evolution equation for the joint spin-phase-space probability density $\mathcal{P}_{YS}$ is the following Fokker-Planck equation

$$\begin{aligned} \partial_\theta \mathcal{P}_{YS} &= L_Y \mathcal{P}_{YS} - \sum_{j=1}^{3} \partial_{s_j} \left(\left[\Omega(\theta, y) \vec{s} \right]_j \mathcal{P}_{YS} \right), \\ \mathcal{P}_{YS}(0, y, \vec{s}) &= \mathcal{P}_{Y_0 S_0}(y, \vec{s}), \end{aligned} \tag{3.8}$$

where $P_{Y_0 S_0}$ is the joint probability density of Y_0 and $\vec{S}_0$, L_Y is the Fokker-Planck operator defined by

$$L_Y := -\sum_{j=1}^{2d} \partial_{y_j} \left[\mathcal{A}(\theta) y \right]_j + \frac{1}{2} \sum_{j,k=1}^{2d} \left[\mathcal{B}(\theta) \mathcal{B}^{\mathrm{T}}(\theta) \right]_{j,k} \partial^2_{y_j y_k},$$

and where, for convenience, we used the abbreviations $\mathcal{A}(\theta) := A(\theta) + \varepsilon \delta A$ and $\mathcal{B}(\theta) := \varepsilon B(\theta)$. The Fokker-Planck equation for the density of the process Y of (3.5) is

$$\partial_\theta \mathcal{P}_Y = L_Y \mathcal{P}_Y, \tag{3.9}$$

which is consistent with obtaining it by integrating both sides of (3.8) w.r.t. $\vec{s}$ since $\mathcal{P}_Y$ is related to $\mathcal{P}_{YS}$ by

$$\mathcal{P}_Y(\theta, y) = \int_{\mathbb{R}^3} \mathcal{P}_{YS}(\theta, y, \vec{s})\, d\vec{s}. \tag{3.10}$$

The polarization density $\vec{\eta}_Y$ corresponding to $\mathcal{P}_{YS}$ is given by

$$\vec{\eta}_Y(\theta, y) = \int_{\mathbb{R}^3} \vec{s}\mathcal{P}_{YS}(\theta, y, \vec{s})\, d\vec{s}. \tag{3.11}$$

Note that (3.10) and (3.11) are analogous to (2.12) and (2.14).

From Chapter 2 we recall that the relation between a system of SDEs and its Fokker-Planck equation is standard, see, e.g., [25, 26, 27]. In the next section we obtain the reduced Bloch equation from (3.8) by differentiating (3.11) w.r.t. θ.

3.3 The reduced beam-frame Bloch equation

The reduced Bloch equation (RBE) for the beam-frame polarization density $\vec{\eta}_Y$ is obtained by differentiating (3.11) w.r.t θ and using (3.8)

$$\begin{aligned} \partial_\theta \vec{\eta}_Y &= L_Y \vec{\eta}_Y + \Omega(\theta, y)\vec{\eta}_Y, \\ \vec{\eta}_Y(0, y) &= \int_{\mathbb{R}^3} \vec{s}\mathcal{P}_{Y_0,S_0}(y, s)\, d\vec{s}, \end{aligned} \tag{3.12}$$

In analogy to the lab frame Bloch equation, the boundary condition is

$$\lim_{y\to\infty} \vec{\eta}_Y(\theta, y) e^{\alpha y^{\mathrm{T}} y} = 0,$$

for some $\alpha > 0$.

Given the beam-frame polarization density $\vec{\eta}_Y$, the beam-frame polarization vector $\vec{P}(\theta)$ of the bunch at azimuth θ is given by

$$\vec{P}(\theta) = \int_{\mathbb{R}^{2d}} \vec{\eta}_Y(\theta, y)\, dy.$$

Our central computational focus is the RBE (3.12) with $\vec{P}(\theta)$ being the quantity of interest. Note that $\vec{P}(\theta)$ and $\vec{P}(t)$ defined in Chapter 2 are related via the transformation from the lab frame to the beam frame.

Remark 6. *In analogy to $\mathcal{P}$ in the lab frame (see Remark 4 in Section 2.2), physically meaningful densities $\mathcal{P}_{YS}$ satisfy $\mathcal{P}_{YS}(\theta, y, \vec{s}) = 0$ if $|\vec{s}| > 1$. So by* (3.10) *and* (3.11)

$$|\vec{\eta}_Y(\theta, y)| \leq \int_{\mathbb{R}^3} |\vec{s}| \mathcal{P}_{YS}(\theta, y, \vec{s})\, d\vec{s} \leq \int_{\mathbb{R}^3} \mathcal{P}_{YS}(\theta, y, \vec{s})\, d\vec{s} = \mathcal{P}_Y(\theta, y).$$

Thus the local polarization field $\vec{P}^Y_{\text{loc}}$ defined via $\vec{\eta}_Y =: \mathcal{P}_Y \vec{P}^Y_{\text{loc}}$ satisfies $|\vec{P}^Y_{\text{loc}}(\theta, y)| \leq 1$. By the meaning of $\mathcal{P}_Y$ and $\vec{\eta}_Y$, the quantity $\vec{P}^Y_{\text{loc}}(\theta, y)$ is the polarization at (θ, y). Hence its size has to be less or equal to 1.

3.4 Equilibrium orbital dynamics

For the next chapters it is important to obtain the exact form of the orbital equilibrium as a 2π-periodic solution to the orbital Fokker-Planck equation. It is obtained here by using the principal solution matrix for $\mathcal{A}(\theta)$, the equilibrium mean vector, $m(\theta)$, and the equilibrium covariance matrix, $K(\theta)$.

Recall that since (3.5) is narrow-sense linear and Y_0 is a Gaussian random variable, the process $Y(\theta)$ is Gaussian. Thus the solution to the corresponding Fokker-Planck equation is completely determined by its mean and covariance which are defined by

$$m' = \mathcal{A}(\theta) m, \tag{3.13}$$

$$K' = \mathcal{A}(\theta) K + K \mathcal{A}^{\mathrm{T}}(\theta) + \mathcal{B}(\theta) \mathcal{B}^{\mathrm{T}}(\theta), \tag{3.14}$$

$$m(0) = \langle Y_0 \rangle, \quad K(0) = \langle (Y_0 - \langle Y_0 \rangle)(Y_0 - \langle Y_0 \rangle)^{\mathrm{T}} \rangle,$$

i.e.

$$\mathcal{P}_Y(\theta, y) = \frac{\exp\left(\frac{1}{2}(y - m(\theta))^{\mathrm{T}} K^{-1}(\theta)(y - m(\theta))\right)}{\sqrt{(2\pi)^{2d} \det K(\theta)}}.$$

Let $\Phi(\theta)$ be the principal solution matrix for $\mathcal{A}$, i.e.

$$\Phi' = \mathcal{A}(\theta)\Phi, \quad \Phi(0) = I_{2d},$$

where I_{2d} is the $2d$-dimensional identity matrix. [1] Then the solutions to (3.13) and (3.14) can be written in terms of Φ as

$$\begin{aligned} m(\theta) &= \Phi(\theta)m(0), \\ K(\theta) &= \Phi(\theta)\left[K(0) + \int_0^\theta \Phi^{-1}(\tau)\mathcal{B}(\tau)\mathcal{B}^{\mathrm{T}}(\tau)\Phi^{-\mathrm{T}}(\tau)\,d\tau\right]\Phi^{\mathrm{T}}(\theta), \end{aligned} \tag{3.15}$$

as is easily checked. The Floquet decomposition of Φ takes the form

$$\Phi(\theta) = P(\theta)e^{Q\theta},$$

where $P(\theta + 2\pi) = P(\theta)$, $P(0) = I$. Because of the damping in $\mathcal{A}(\theta)$ we assume that $e^{Q\theta} \to 0$ as $\theta \to \infty$.

Thus, the eigenvalues of the monodromy matrix, $\Phi(2\pi) = e^{Q2\pi}$, are assumed to have negative real parts. Clearly $m(\theta) \to 0$ as $\theta \to \infty$, so that we focus on the initial value problem for K.

Theorem 1. The unique 2π-periodic solution to (3.14) is given by

$$\begin{aligned} K(\theta) &= \Phi(\theta)\int_{-\infty}^{\theta} e^{-Q\tau}\mathfrak{B}(\tau)e^{-Q^{\mathrm{T}}\tau}\,d\tau\,\Phi^{\mathrm{T}}(\theta) \\ &= P(\theta)\int_{-\infty}^{0} e^{-Q\tau'}\mathfrak{B}(\tau'+\theta)e^{-Q^{\mathrm{T}}\tau'}\,d\tau'\,P^{\mathrm{T}}(\theta), \end{aligned} \tag{3.16}$$

where $\mathfrak{B}(\theta) = P^{-1}(\theta)\mathcal{B}(\theta)\mathcal{B}^{\mathrm{T}}(\theta)P^{-\mathrm{T}}(\theta) = \mathfrak{B}(\theta + 2\pi)$.

Proof. Differentiating the first equality in (3.16) gives us

$$K' = \mathcal{A}(\theta)K + K\mathcal{A}^{\mathrm{T}}(\theta) + \Phi(\theta)e^{-Q\theta}\mathfrak{B}(\theta)e^{-Q^{\mathrm{T}}\theta}\Phi^{\mathrm{T}}(\theta),$$

[1] The PSM $\Phi(\theta)$ in Appendices A and B is the PSM for a Hamiltonian $\mathcal{A}(\theta)$, e.g $A(\theta)$ as in (3.5).

and we note that the last term is clearly $\mathfrak{B}(\theta)$. The first expression in (3.16) can be written as

$$P(\theta)\int_{-\infty}^{\theta} e^{-Q(\tau-\theta)}\mathfrak{B}(\tau)e^{-Q^{\mathrm{T}}(\tau-\theta)}\,d\tau\,P^{\mathrm{T}}(\theta).$$

Changing the variables to $\tau' = \tau - \theta$ gives the second expression which is clearly 2π-periodic. □

Remark 7. *Consistent with [25], the initial covariance matrix*

$$K(0) = \int_{-\infty}^{0} e^{-Q\tau'}\mathfrak{B}(\tau')e^{-Q^{\mathrm{T}}\tau'}\,d\tau',$$

is the solution to the matrix equation

$$K(0) = K(2\pi) = e^{Q2\pi}\left[K(0) + \int_0^{2\pi}\Phi^{-1}(\tau)\mathcal{B}(\tau)\mathcal{B}^{\mathrm{T}}(\tau)\Phi^{-\mathrm{T}}(\tau)\,d\tau\right]\,e^{Q^{\mathrm{T}}2\pi},$$

and thus it is the initial value for (3.14) *giving the* 2π*-periodic solution* (3.16)*.*

Theorem 2. The 2π-periodic function in (3.16) is a globally asymptotically stable solution of (3.14).

Proof. Consider (3.15) with $\theta = 2\pi N + \theta^*$, $\theta^* \in [0, 2\pi)$, so that (3.15) becomes

$$K(2\pi N + \theta^*) = \Phi(2\pi N + \theta^*)[K(0) + \int_0^{2\pi N+\theta^*} e^{-Q\tau}\mathfrak{B}(\tau)e^{-Q^{\mathrm{T}}\tau}\,d\tau]\,\Phi^{\mathrm{T}}(2\pi N + \theta^*).$$

Clearly the term with $K(0)$ goes to 0 as $\theta \to \infty$. The rest can be written as

$$\begin{aligned} &P(\theta^*)\int_0^{2\pi N+\theta^*} e^{-Q(\tau-2\pi N-\theta^*)}\mathfrak{B}(\tau)e^{-Q^{\mathrm{T}}(\tau-2\pi N-\theta^*)}\,d\tau\,P^{\mathrm{T}}(\theta^*) \\ &\quad = P(\theta^*)\int_{-2\pi N-\theta^*}^{0} e^{-Q\tau'}\mathfrak{B}(\tau'+\theta^*)e^{-Q^{\mathrm{T}}\tau'}\,d\tau\,P^{\mathrm{T}}(\theta^*), \\ &\quad \to P(\theta^*)\int_{-\infty}^{0} e^{-Q\tau'}\mathfrak{B}(\tau'+\theta^*)e^{-Q^{\mathrm{T}}\tau'}\,d\tau\,P^{\mathrm{T}}(\theta^*), \quad N\to\infty, \end{aligned}$$

which is the second equality in (3.16). □

Therefore the asymptotically stable equilibrium solution of the Fokker-Planck equation (3.10), reads as

$$\mathcal{P}_Y(\theta, y) = \frac{\exp\left(\frac{1}{2} y^T K^{-1}(\theta) y\right)}{\sqrt{(2\pi)^{2d} \det K(\theta)}} = \mathcal{P}_Y(\theta + 2\pi, y),$$

where $K(\theta)$ is given by (3.16).

Remark 8. *We have proven this even for the case where the Fokker-Planck equation is not uniformly parabolic, see §11.9 in [39].*

Resonance phenomena are hidden in a factor of (3.16),

$$\int_{-\infty}^{0} e^{-Q\tau'} \mathfrak{B}(\tau' + \theta) e^{-Q^{\mathrm{T}}\tau'} d\tau', \tag{3.17}$$

in particular, in the eigenvalues of the monodromy matrix $\Phi(2\pi) = e^{Q2\pi}$. In the most interesting case, $d = 3$, we assume that $\Phi(2\pi)$ has 6 linearly independent eigenvectors with eigenvalues

$$\rho = \exp\left(2\pi(\mu_k + \mathbf{i}\nu_k)\right), \quad 0 < \nu_1 < \nu_3 < \nu_5 < \frac{1}{2}, \quad \nu_{2l} = -\nu_{2l-1}, \quad \mu_k < 0,$$

using the notation of Appendix A and [8][2]. Thus, (3.17) can be analyzed off resonance by looking at the eigen-structure of $\Phi(2\pi)$. This is work for the future.

3.5 The non-radiative problem

Since we treat synchrotron radiation as a perturbation it is important to understand the dynamics without the perturbation. We start with (3.5) and (3.6) and remove the radiation effects by setting $\varepsilon = 0$. Then we have

$$Y' = A(\theta) Y, \tag{3.18}$$

$$\vec{S}' = \Omega(\theta, Y) \vec{S}. \tag{3.19}$$

[2]Reference [8] discusses $\Phi(2\pi)$ in the Hamiltonian case as does Appendix A.

Recall that $A(\theta)$ is a Hamiltonian matrix and from (3.7) that $\mathrm{Tr}(A(\theta)) = 0$. Then, from Section 3.2, $L_Y\big|_{\varepsilon=0} = -\sum_{j=1}^{2d}[A(\theta)y]_j\partial_{y_j}$ and the joint probability density function satisfies

$$\begin{aligned}\partial_\theta \mathcal{P}_{YS} &= -\sum_{j=1}^{2d}[A(\theta)y]_j\partial_{y_j}\mathcal{P}_{YS} - \sum_{j=1}^{3}\partial_{s_j}\left([\Omega(\theta,y)\vec{s}]_j\,\mathcal{P}_{YS}\right), \qquad (3.20)\\ \mathcal{P}_{YS}(0,y,\vec{s}) &= \mathcal{P}_{Y_0S_0}(y,\vec{s}),\end{aligned}$$

and similarly the phase-space density satisfies

$$\partial_\theta \mathcal{P}_Y = -\sum_{j=1}^{2d}[A(\theta)y]_j\partial_{y_j}\mathcal{P}_Y. \qquad (3.21)$$

Thus, to no surprise, the Fokker-Planck equations for $\mathcal{P}_{YS}$ and $\mathcal{P}_Y$ reduce to what we call the Liouville [3] equations for (Y,S) and Y respectively. Following Section 3.3, the non-radiative RBE is

$$\partial_\theta \vec{\eta}_Y = -\sum_{j=1}^{2d}\left[A(\theta)y\right]_j\partial_{y_j}\vec{\eta}_Y + \Omega(\theta,y)\vec{\eta}_Y. \qquad (3.22)$$

It is easy to show that $\vec{\eta}_Y(\theta, Y(\theta))$ is a solution of (3.19) if Y satisfies (3.18) and if $\vec{\eta}_Y$ satisfies (3.22). Differentiating $\mathcal{P}_{YS}(\theta, Y(\theta), \vec{\eta}(\theta, Y(\theta)))$ and $\mathcal{P}_Y(\theta, Y(\theta))$, and using (3.20)–(3.22) and (3.18), leads to

$$\begin{aligned}\mathcal{P}_{YS}(\theta, Y(\theta), \vec{\eta}(\theta, Y(\theta))) &= \mathcal{P}_{Y_0S_0}(Y_0, \vec{\eta}(0, Y_0)),\\ \mathcal{P}_Y(\theta, Y(\theta)) &= \mathcal{P}_{Y_0}(Y_0),\end{aligned}$$

[3] Consider the random initial value problem $X' = f(t,X)$, $X(0) = X_0$, where the only randomness is in X_0. Clearly $X(t)$ is a random process and its probability density $p(t,x)$ evolves via $\partial_t p = -\nabla\cdot[f(t,x)p]$ which we call the Liouville equation.

which are constant in time. For example, using (3.18) and (3.21) we obtain

$$\begin{aligned}\frac{d}{d\theta}\mathcal{P}_Y(\theta, Y(\theta)) =&(\partial_\theta \mathcal{P}_Y)(\theta, Y(\theta)) + \sum_{j=1}^{2d} Y_j'(\theta)(\partial_{y_j}\mathcal{P}_Y)(\theta, Y(\theta)) \\ =& -\sum_{j=1}^{2d}[A(\theta)Y(\theta)]_j \partial_{y_j}\mathcal{P}_Y(\theta, Y(\theta)) \\ &+ \sum_{j=1}^{2d}[A(\theta)Y(\theta)]_j \partial_{y_j}\mathcal{P}_Y(\theta, Y(\theta) = 0.\end{aligned}$$

We define the notion of spin-orbit equilibrium by $\mathcal{P}_{YS}$ being 2π-periodic in θ. It follows that for spin-orbit equilibrium $\vec{\eta}_Y$ and $\mathcal{P}_Y$ are 2π-periodic. Thus we are interested in 2π-periodic solutions to (3.22). From Remark 6 we have $\vec{P}^Y_{\text{loc}}$ as

$$\vec{\eta}_Y(\theta, y) =: \mathcal{P}_Y(\theta, y)\vec{P}^Y_{\text{loc}}(\theta, y).$$

Assuming there exists a 2π-periodic P_{YS} we write

$$\vec{\eta}_{\text{eq}}(\theta, y) = \mathcal{P}_{\text{eq}}(\theta, y)\vec{P}_{\text{loc,eq}}(\theta, y), \tag{3.23}$$

where $\vec{P}_{\text{loc,eq}}(\theta, y)$ must be 2π-periodic, and $\vec{\eta}_{\text{eq}}(\theta, y)$ and $\mathcal{P}_{\text{eq}}(\theta, y)$ denote the equilibrium densities. To keep the argumentation simple we assume that $\mathcal{P}_{\text{eq}} > 0$ and $|\vec{P}_{\text{loc,eq}}| > 0$, so that $\vec{P}_{\text{loc,eq}} = \mathcal{P}_{\text{eq}}^{-1}\vec{\eta}_{\text{eq}}$. Note that ${\mathcal{P}_{\text{eq}}}^{-1}$ satisfies (3.21) and recall that $\vec{\eta}_{\text{eq}}$ satisfies (3.22). Hence $\vec{P}_{\text{loc,eq}}$ satisfies (3.22), so that $|\vec{P}_{\text{loc,eq}}|$ (and thus $|\vec{P}_{\text{loc,eq}}|^{-1}$) satisfies (3.21). Then the direction of $\vec{P}_{\text{loc,eq}}$ defined as $\hat{n} := \frac{\vec{P}_{\text{loc,eq}}}{|\vec{P}_{\text{loc,eq}}|}$ satisfies (3.22). $\hat{n}$ defines the so-called invariant spin field (ISF) mentioned in the Introduction and which is a central quantity for depolarization studies [15]. The invoking of the ISF calls for a formal definition as follows.

Definition 3.5.1. The *invariant spin field* (ISF) is a normalized 2π-periodic solution to the non-radiative Bloch equation

$$\partial_\theta \hat{n} = -\sum_{j=1}^{2d}\left[A(\theta)y\right]_j \partial_{y_j}\hat{n} + \Omega(\theta, y)\hat{n}, \tag{3.24}$$

where

$$\hat{n}(\theta, y) = \hat{n}(\theta + 2\pi, y), \quad |\hat{n}(\theta, y)| = 1, \quad y \in \mathbb{R}^{2d}, \theta \in \mathbb{R}.$$

If $|P_{\mathrm{loc,eq}}|$ is independent of y, then the equilibrium solution, (3.23), to (3.22) can be written in the form

$$\vec{\eta}_{\mathrm{eq}}(\theta, y) = c\mathcal{P}_{\mathrm{eq}}(\theta, y)\hat{n}(\theta, y), \tag{3.25}$$

where $c = |P_{\mathrm{loc,eq}}|$ is a constant. Note that, by Remark 6 and (3.25), $c \leq 1$. Equation (3.25) defines a 2π-periodic polarization density for an unperturbed problem in w.r.t. the ISF. This will be the starting point for including the influence of synchrotron radiation as a perturbation to this problem in the next chapter.

Remark 9. *For single particles we can use Hamiltonian systems like* (3.18) *as well. We assume that the solutions of* (3.18) *are bounded. This is the case if and only if the monodromy matrix* $\Phi(2\pi)$ *has* $2d$ *linearly independent eigenvectors,* w_k*, and its eigenvalues (characteristic multipliers) are of the form* $\rho_k = e^{\mathbf{i}2\pi\nu_k}$*, i.e. have modulus* 1 *(the characteristic exponents,* ν_k*, are the orbital tunes), see [8] and Appendix A. Then, the trajectories of particles lie on* $(d+1)$*-dimensional non-intersecting (distinct) closed tubes. Each* θ *cross-section of a tube is a subset of the phase space,* $\mathcal{T}_\theta$*, homeomorphic to a* d*-dimensional torus (Cartesian product of* d *circles). An example for one degree of freedom is the simple phase-space ellipse of Courant-Snyder theory, [40]. The ISF can be evaluated separately on each tube. Moreover, if, as is naturally the case, the particles are distributed uniformly on their respective* $\mathcal{T}_\theta$*, the phase-space density on each tube is uniform and* 2π*-periodic, and thus* $\mathcal{P}_Y$ *is* 2π*-periodic, i.e., in equilibrium. An analogous insight leads to the conclusion that for* $\vec{\eta}$ *to be* 2π*-periodic,* $|P_{\mathrm{loc}}|$ *must be a constant all over each tube. However,* $|P_{\mathrm{loc}}|$ *can be different for different tubes and this has been observed experimentally, [41]. Then we need a more general form for* (3.25) *namely* (3.23)*.*

Also, it is worthwhile to mention that the equations in this section are also impor-

tant in their own right since they provide a good description to the motion of realistic particles for which the radiation can be neglected, e.g. protons.

Remark 10. *In this section we presented a non-radiative description of spin-orbit dynamics. This is well motivated, since the radiation is considered to be a perturbation in the following chapters. In contrast to the situation detailed in Remark 9 the stochastic motion in phase space means that the electrons do not stay on distinct subsets of the phase space but diffuse through phase space with the result that* $\mathcal{P}_Y$ *is a Gaussian. At the same time there is mixing of the* $|P_{\mathrm{loc}}(\theta)|$*. Then after a few damping times we expect a common* $|P_{\mathrm{loc}}(\theta)|$ *for all points in phase space so that* (3.25) *indeed becomes an appropriate form for* $\vec{\eta}$ *in the radiative case.*

Remark 11. *Equation* (3.22) *is consistent with the non-radiative reduced Bloch equation derived from first principles in [37], i.e., without reference to* (3.12)*.*

Chapter 4

The ISF approximation of the polarization density

As the rest of the thesis is relevant to the beam-frame only from now on we change the notation for beam-frame polarization density from $\vec{\eta}_Y$ to $\vec{\eta}$. Here we consider the effect of synchrotron radiation as a small perturbation to a Hamiltonian system driving the particle motion. In Section 3.5 we discussed the unperturbed problem, where in equilibrium the polarization density has a direction, called the ISF denoted as $\hat{n}$, see Definition 3.5.1. Then in Remark 10, in considering the effect of synchrotron radiation, we motivated the form for $\vec{\eta}$ given given in (3.25), namely

$$\vec{\eta}(\theta, y) \approx \vec{\eta}_{\mathrm{ISF}}(\theta, y) := P_{\mathrm{ISF}}(\theta)\mathcal{P}_{\mathrm{eq}}(\theta, y)\hat{n}(\theta, y), \tag{4.1}$$

We call $\vec{\eta}_{\mathrm{ISF}}$ the *ISF approximation* of $\vec{\eta}$. Now $\mathcal{P}_{\mathrm{eq}}$ is the periodic phase–space density as discussed in Section 3.4 which includes synchrotron radiation. We expect $P_{\mathrm{ISF}}(\theta)$ to be an exponentially decaying function, since we expect diffusion to drive $\vec{\eta}$ to 0. In this chapter we calculate $P_{\mathrm{ISF}}(\theta)$, so that the residual error of the ISF approximation is orthogonal to the ISF on average. This leads to a well-defined PDE for the error in (4.1) and an ODE for P_{ISF}, showing that indeed $P_{\mathrm{ISF}}(\theta)$ is a decaying exponential

function.

4.1 ISF approximation

Consider (3.5) and (3.6) and the orbital Fokker-Planck equation (3.9) written in the form

$$\begin{aligned}\partial_\theta \mathcal{P}_Y &= L_Y \mathcal{P}_Y \\ &= -\sum_{j=1}^{2d} \partial_{y_j} \left(\left[A(\theta) y + \varepsilon \delta A(\theta) y \right]_j \mathcal{P}_Y \right) + \frac{\varepsilon}{2} \sum_{j,k=1}^{2d} \left[B(\theta) B^{\mathrm{T}}(\theta) \right]_{j,k} \partial^2_{y_j y_k} \mathcal{P}_Y, \quad (4.2) \\ \mathcal{P}_Y(0, y) &= \mathcal{P}_{Y_0}(y).\end{aligned}$$

A phase-space density is, by definition, a nonnegative solution $\mathcal{P}_Y$ of (4.2) for which

$$\int_{\mathbb{R}^{2d}} \mathcal{P}_Y(\theta, y)\, dy = 1. \quad (4.3)$$

Recall, the Bloch equation associated with (3.5) and (3.6) is

$$\partial_\theta \vec{\eta} = L_{\mathrm{Bloch}} \vec{\eta} = L_Y \vec{\eta} + \Omega(\theta, y) \vec{\eta}, \quad (4.4)$$

where $\vec{\eta}$ was defined in (3.11). For some work on the ISF see [42] and [15]. The ISF on the closed orbit is denoted by $\hat{n}_0(\theta)$, i.e. $\hat{n}_0(\theta) = \hat{n}(\theta, 0)$. It is easily obtained as an eigenvector of the one-turn spin-transport map on the closed orbit [13]. There are many methods for computing the ISF but none are trivial, see e.g. [43], [44], references in [42] and for a recent technique see [45]. In fact the existence, in general, of the invariant spin field is a mathematical issue which is only partially resolved, see, e.g., [42].

In our approach the real valued function P_{ISF} in (4.1) will be determined by the minimal residual method, i.e., by minimizing the residual in a certain way. The residual $\Delta\vec{r}(\theta, y)$ of $\vec{\eta}_{\mathrm{ISF}}$ w.r.t. the Bloch equation is defined by

$$\Delta\vec{r}(\theta, y) := \partial_\theta \vec{\eta}_{\mathrm{ISF}} - L_{\mathrm{Bloch}} \vec{\eta}_{\mathrm{ISF}}. \quad (4.5)$$

Since $\vec{\eta}_{\mathrm{ISF}}$ points into the direction of $\hat{n}$, one ideally would like to have

$$\Delta\vec{r}(\theta, y) \cdot \hat{n}(\theta, y) = 0. \tag{4.6}$$

However this condition is too strong, i.e., no function P_{ISF} exists such that (4.6) holds (see Remark 12 after Theorem 3 below). Perhaps surprisingly, if one weakens (4.6) to

$$\int_{\mathbb{R}^{2d}} \Delta\vec{r}(\theta, y) \cdot \hat{n}(\theta, y)\ dy = 0, \tag{4.7}$$

then a function P_{ISF} exists, i.e., the minimal residual condition (4.7) can be satisfied. In fact the following theorem states, that the minimal residual condition (4.7) is satisfied if and only if P_{ISF} satisfies the first-order ODE

$$\begin{aligned} &P'_{\mathrm{ISF}} = -\varepsilon q(\theta) P_{\mathrm{ISF}}, \\ &P_{\mathrm{ISF}}(0) = \int_{\mathbb{R}^{2d}} \vec{\eta}(0, y)\ \cdot \hat{n}(0, y) dy = \int_{\mathbb{R}^{2d}} \int_{\mathbb{R}^3} \mathcal{P}_{Y_0, \vec{S}_0}(y, \vec{s})\ \vec{s} \cdot \hat{n}(0, y)\ d\vec{s} dy, \end{aligned} \tag{4.8}$$

where

$$q(\theta) = \frac{1}{2} \sum_{j,k=1}^{2d} \left[B(\theta) B^{\mathrm{T}}(\theta)\right]_{j,k} \int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta, y) [\partial_{y_j} \hat{n}(\theta, y)] \cdot [\partial_{y_k} \hat{n}(\theta, y)]\ dy. \tag{4.9}$$

Choosing P_{ISF} in (4.1) as a solution of (4.8) completes the definition of $\vec{\eta}_{\mathrm{ISF}}$. For the generalization of (4.1), (4.7) and (4.8) see Section 5.2.

We now state and prove the theorem.

Theorem 3. P_{ISF} satisfies the first-order ODE (4.8), if and only if (4.7) holds.

Proof. We first compute, by (4.1), (4.4) and (4.5),

$$\begin{aligned} \Delta\vec{r} &= (\partial_\theta - L_{\mathrm{Bloch}})(P_{\mathrm{ISF}} \mathcal{P}_{\mathrm{eq}} \hat{n}) \\ &= P'_{\mathrm{ISF}} \mathcal{P}_{\mathrm{eq}} \hat{n} + P_{\mathrm{ISF}} \partial_\theta (\mathcal{P}_{\mathrm{eq}} \hat{n}) - P_{\mathrm{ISF}} L_{\mathrm{Bloch}} (\mathcal{P}_{\mathrm{eq}} \hat{n}). \end{aligned} \tag{4.10}$$

To get insight into (4.10) we define

$$L_1^A := -\sum_{j=1}^{2d}[A(\theta)y]_j\partial_{y_j}, \tag{4.11}$$

$$\delta L_1^A := -\sum_{j=1}^{2d}[\varepsilon\delta A(\theta)y]_j\partial_{y_j}, \tag{4.12}$$

$$L_{1,l}^B := \sqrt{\varepsilon}\sum_{j=1}^{2d} B_{j,l}(\theta)\partial_{y_j}, \quad l=1,\dots m, \tag{4.13}$$

where the operators $L_1^A, \delta L_1^A, L_{1,l}^B$ can act on scalar and vector functions. Moreover we define the multiplication operator L_0 to be the multiplication by the function

$$-\sum_{j=1}^{2d}\partial_{y_j}[A(\theta)y+\varepsilon\delta A(\theta)y]_j = -Tr[\varepsilon\delta A(\theta)], \tag{4.14}$$

where in (4.14) we used (3.7). With the operators $L_1^A, \delta L_1^A, L_1^B$ and L_0 at hand we obtain from (4.11), (4.12), (4.13) and (4.14)

$$\begin{aligned} L_Y\vec{\eta} =& -\sum_{j=1}^{2d}\partial_{y_j}\left([A(\theta)+\varepsilon\delta A(\theta)y]_j\,\vec{\eta}\right) \\ &+\frac{\varepsilon}{2}\sum_{l=1}^{m}\sum_{j=1}^{2d}B_{j,l}(\theta)\partial_{y_j}\sum_{k=1}^{2d}B_{k,l}(\theta)\partial_{y_k}\vec{\eta} \\ =&(L_0+L_1^A+\delta L_1^A+\frac{1}{2}\sum_{l=1}^{m}L_{1,l}^BL_{1,l}^B)\vec{\eta}. \end{aligned}$$

Then by (4.4) and (4.10)

$$\begin{aligned} \Delta\vec{r} =& P'_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}}\hat{n} \\ &+P_{\mathrm{ISF}}\left(\partial_\theta(\mathcal{P}_{\mathrm{eq}}\hat{n})-(L_0+L_1^A+\delta L_1^A+\frac{1}{2}\sum_{l=1}^{m}L_{1,l}^BL_{1,l}^B)(\mathcal{P}_{\mathrm{eq}}\hat{n})\right) \\ &-P_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}}\Omega\hat{n}. \end{aligned} \tag{4.15}$$

Note that $\partial_\theta, L_1^A, \delta L_1^A, L_{1,l}^B$ are first-order differential operators and L_0 is a multiplication operator, i.e., a zeroth-order differential operator. It is easy to see that these

operators satisfy the product rule for differential operators and this will facilitate our task. In fact using the product rule for $\partial_\theta, L_1^A, \delta L_1^A, L_{1,l}^B$ and L_0 we get

$$
\begin{aligned}
&\partial_\theta(\mathcal{P}_{\rm eq}\hat{n}) = \hat{n}\partial_\theta\mathcal{P}_{\rm eq} + \mathcal{P}_{\rm eq}\partial_\theta\hat{n},\\
&L_1^A(\mathcal{P}_{\rm eq}\hat{n}) = \hat{n}(L_1^A\mathcal{P}_{\rm eq}) + \mathcal{P}_{\rm eq}(L_1^A\mathcal{P}_{\rm eq}),\\
&\delta L_1^A(\mathcal{P}_{\rm eq}\hat{n}) = \hat{n}(\delta L_1^A\mathcal{P}_{\rm eq}) + \mathcal{P}_{\rm eq}(\delta L_1^A\mathcal{P}_{\rm eq}),\\
&L_{1,l}^B(\mathcal{P}_{\rm eq}\hat{n}) = \hat{n}(L_{1,l}^B\mathcal{P}_{\rm eq}) + \mathcal{P}_{\rm eq}(L_{1,l}^B\mathcal{P}_{\rm eq}),\\
&L_{1,l}^B L_{1,l}^B(\mathcal{P}_{\rm eq}\hat{n}) = L_{1,l}^B\Big(\hat{n}(L_{1,l}^B\mathcal{P}_{\rm eq}) + \mathcal{P}_{\rm eq}(L_{1,l}^B\mathcal{P}_{\rm eq})\Big)\\
&= \hat{n}(L_{1,l}^B L_{1,l}^B\mathcal{P}_{\rm eq}) + \mathcal{P}_{\rm eq}(L_{1,l}^B L_{1,l}^B\hat{n}) + 2(L_{1,l}^B\mathcal{P}_{\rm eq})(L_{1,l}^B\hat{n}),\\
&L_0(\mathcal{P}_{\rm eq}\hat{n}) = \hat{n}(L_0\mathcal{P}_{\rm eq}),
\end{aligned}
$$

hence the expression in the large bracket of (4.15) becomes

$$
\begin{aligned}
&\partial_\theta(\mathcal{P}_{\rm eq}\hat{n}) - (L_0 + L_1^A + \delta L_1^A + \frac{1}{2}\sum_{l=1}^m L_{1,l}^B L_{1,l}^B)(\mathcal{P}_{\rm eq}\hat{n}) - \mathcal{P}_{\rm eq}\Omega\hat{n}\\
&= \hat{n}\partial_\theta\mathcal{P}_{\rm eq} + \mathcal{P}_{\rm eq}\partial_\theta\hat{n} - \hat{n}(L_0\mathcal{P}_{\rm eq}) - \hat{n}(L_1^A\mathcal{P}_{\rm eq}) - \hat{n}(\delta L_1^A\mathcal{P}_{\rm eq})\\
&- \mathcal{P}_{\rm eq}(L_1^A\hat{n}) - \mathcal{P}_{\rm eq}(\delta L_1^A\hat{n}) - \frac{1}{2}\hat{n}\left(\sum_{l=1}^m L_{1,l}^B L_{1,l}^B\mathcal{P}_{\rm eq}\right) - \frac{1}{2}\mathcal{P}_{\rm eq}\left(\sum_{l=1}^m L_{1,l}^B L_{1,l}^B\hat{n}\right)\\
&- \sum_{k=1}^m (L_{1,l}^B\mathcal{P}_{\rm eq})(L_{1,l}^B\hat{n}) - \mathcal{P}_{\rm eq}\Omega\hat{n}\\
&= \hat{n}\Big(\partial_\theta\mathcal{P}_{\rm eq} - L_0\mathcal{P}_{\rm eq} - L_1^A\mathcal{P}_{\rm eq} - \delta L_1^A\mathcal{P}_{\rm eq} - \frac{1}{2}\sum_{l=1}^m L_{1,l}^B L_{1,l}^B\mathcal{P}_{\rm eq}\Big)\\
&+ \mathcal{P}_{\rm eq}\left(\partial_\theta\hat{n} - L_1^A\hat{n} - \delta L_1^A\hat{n} - \frac{1}{2}\sum_{l=1}^m L_{1,l}^B L_{1,l}^B\hat{n} - \Omega\hat{n}\right) - \sum_{k=1}^m (L_{1,l}^B\mathcal{P}_{\rm eq})(L_{1,l}^B\hat{n}),
\end{aligned}
$$

so that (4.15) can be written as

$$\begin{aligned}\Delta\vec{r} =& P'_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}}\hat{n} \\ &+ P_{\mathrm{ISF}}\Bigg\{\hat{n}\Bigg(\partial_\theta\mathcal{P}_{\mathrm{eq}} - L_0\mathcal{P}_{\mathrm{eq}} - L_1^A\mathcal{P}_{\mathrm{eq}} - \delta L_1^A\mathcal{P}_{\mathrm{eq}} - \frac{1}{2}\sum_{l=1}^{m} L_{1,l}^B L_{1,l}^B\mathcal{P}_{\mathrm{eq}}\Bigg) \\ &+ \mathcal{P}_{\mathrm{eq}}\Bigg(\partial_\theta\hat{n} - L_1^A\hat{n} - \delta L_1^A\hat{n} - \frac{1}{2}\sum_{l=1}^{m} L_{1,l}^B L_{1,l}^B\hat{n} - \Omega\hat{n}\Bigg) \\ &- \sum_{k=1}^{m}(L_{1,l}^B\mathcal{P}_{\mathrm{eq}})(L_{1,l}^B\hat{n})\Bigg\}. \end{aligned} \tag{4.16}$$

With the operators $L_1^A, \delta L_1^A, L_{1,l}^B$ and L_0 the PDEs (4.2) and (3.24) for $\mathcal{P}_{\mathrm{eq}}$ and $\hat{n}$ can be written as

$$\begin{aligned}\partial_\theta\mathcal{P}_{\mathrm{eq}} &= (L_0 + L_1^A + \delta L_1^A + \frac{1}{2}\sum_{l=1}^{m} L_{1,l}^B L_{1,l}^B)\mathcal{P}_{\mathrm{eq}}, \\ \partial_\theta\hat{n} &= L_1^A\hat{n} + \Omega(\theta, y)\hat{n}.\end{aligned}$$

Then $\Delta\vec{r}$ in (4.16) simplifies to

$$\begin{aligned}\Delta\vec{r} =& P'_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}}\hat{n} \\ &- P_{\mathrm{ISF}}\left(\mathcal{P}_{\mathrm{eq}}(\delta L_1^A\hat{n}) + \frac{1}{2}\mathcal{P}_{\mathrm{eq}}\left(\sum_{l=1}^{m} L_{1,l}^B L_{1,l}^B\hat{n}\right) + \sum_{k=1}^{m}(L_{1,l}^B\mathcal{P}_{\mathrm{eq}})(L_{1,l}^B\hat{n})\right),\end{aligned}$$

so that since $|\hat{n}(\theta, y)| = 1$

$$\begin{aligned}\Delta\vec{r}\cdot\hat{n} &= P'_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}} - P_{\mathrm{ISF}}\left(\frac{1}{2}\mathcal{P}_{\mathrm{eq}}\hat{n}\cdot\left(\sum_{l=1}^{m} L_{1,l}^B L_{1,l}^B\hat{n}\right) + \sum_{l=1}^{m}(L_{1,l}^B\mathcal{P}_{\mathrm{eq}})\hat{n}\cdot(L_{1,l}^B\hat{n})\right) \\ &= P'_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}} - \frac{1}{2}P_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}}\hat{n}\cdot\left(\sum_{l=1}^{m} L_{1,l}^B L_{1,l}^B\hat{n}\right),\end{aligned} \tag{4.17}$$

which implies, by (4.3), that

$$\begin{aligned}&\int_{\mathbb{R}^{2d}}\Delta\vec{r}(\theta, y)\cdot\hat{n}(\theta, y)\, dy = P'_{\mathrm{ISF}}(\theta) \\ &- \frac{1}{2}P_{\mathrm{ISF}}(\theta)\int_{\mathbb{R}^{2d}}\mathcal{P}_{\mathrm{eq}}(\theta, y)\hat{n}(\theta, y)\cdot\left[\left(\sum_{l=1}^{m} L_{1,l}^B L_{1,l}^B\hat{n}\right)(\theta, y)\right]\, dy.\end{aligned} \tag{4.18}$$

To simplify (4.18) we compute by (4.13) and since $|\hat{n}(\theta, y)| = 1$

$$\begin{aligned}
&\int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta, y)\hat{n}(\theta, y) \cdot \left[\left(\sum_{l=1}^{m} L_{1,l}^{B} L_{1,l}^{B} \hat{n} \right)(\theta, y) \right] \, dy \\
&= \varepsilon \sum_{l=1}^{m} \sum_{j,k=1}^{2d} B_{j,l}(\theta) B_{k,l}(\theta) \int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta, y)\hat{n}(\theta, y) \cdot [\partial_{y_j} \partial_{y_k} \hat{n}(\theta, y)] \, dy \\
&= -\varepsilon \sum_{l=1}^{m} \sum_{j,k=1}^{2d} B_{j,l}(\theta) B_{k,l}(\theta) \int_{\mathbb{R}^{2d}} \partial_{y_j} \left[\left(\mathcal{P}_{\mathrm{eq}} \hat{n} \right)(\theta, y) \right] \cdot \partial_{y_k} \hat{n}(\theta, y) \, dy \\
&= -\varepsilon \sum_{l=1}^{m} \sum_{j,k=1}^{2d} B_{j,l}(\theta) B_{k,l}(\theta) \int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta, y) [\partial_{y_j} \hat{n}(\theta, y)] \cdot [\partial_{y_k} \hat{n}(\theta, y)] \, dy \\
&= -\varepsilon \sum_{j,k=1}^{2d} \left[B(\theta) B^{\mathrm{T}}(\theta) \right]_{j,k} \int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta, y) [\partial_{y_j} \hat{n}(\theta, y)] \cdot [\partial_{y_k} \hat{n}(\theta, y)] \, dy
\end{aligned}$$

hence, by (4.9) and (4.18),

$$\begin{aligned}
&\int_{\mathbb{R}^{2d}} \Delta \vec{r}(\theta, y) \cdot \hat{n}(\theta, y) \, dy = P'_{\mathrm{ISF}}(\theta) \\
&+ \frac{\varepsilon}{2} P_{\mathrm{ISF}}(\theta) \sum_{j,k=1}^{2d} \left[B(\theta) B^{\mathrm{T}}(\theta) \right]_{j,k} \int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta, y) [\partial_{y_j} \hat{n}(\theta, y)] \cdot [\partial_{y_k} \hat{n}(\theta, y)] \, dy \\
&= P'_{\mathrm{ISF}}(\theta) + q(\theta) P_{\mathrm{ISF}}(\theta). \qquad (4.19)
\end{aligned}$$

It follows from (4.19) that P_{ISF} satisfies (4.8) if and only if (4.7) holds. □

Remark 12. *In general* $\hat{n} \cdot \left(\sum_{l=1}^{m} L_{1,l}^{B} L_{1,l}^{B} \hat{n} \right)$ *is not independent of* y *so that by* (4.17) *one cannot satisfy* (4.6) *except in the uninteresting case where* P_{ISF} *is the zero function.*

We now make some remarks on the error of $\vec{\eta}_{\mathrm{ISF}}$ which is defined by

$$\Delta \vec{\eta}(\theta, y) = \vec{\eta}(\theta, y) - \vec{\eta}_{\mathrm{ISF}}(\theta, y).$$

It follows from the Bloch equation for $\vec{\eta}$ that

$$\begin{aligned}\partial_\theta \Delta\vec{\eta} &= \partial_\theta \vec{\eta} - \partial_\theta \vec{\eta}_{\mathrm{ISF}} = L_{\mathrm{Bloch}}\vec{\eta} - P'_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}}\hat{n} - P_{\mathrm{ISF}}\partial_\theta(\mathcal{P}_{\mathrm{eq}}\hat{n}) \\ &= L_{\mathrm{Bloch}}\vec{\eta}_{\mathrm{ISF}} + L_{\mathrm{Bloch}}\Delta\vec{\eta} - P'_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}}\hat{n} - P_{\mathrm{ISF}}\partial_\theta(\mathcal{P}_{\mathrm{eq}}\hat{n}) \\ &= P_{\mathrm{ISF}}L_{\mathrm{Bloch}}(\mathcal{P}_{\mathrm{eq}}\hat{n}) + L_{\mathrm{Bloch}}\Delta\vec{\eta} - P'_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}}\hat{n} - P_{\mathrm{ISF}}\partial_\theta(\mathcal{P}_{\mathrm{eq}}\hat{n}),\end{aligned}$$

so that by the above theorem we get the following PDE for $\Delta\vec{\eta}$

$$\partial_\theta \Delta\vec{\eta} = L_{\mathrm{Bloch}}\Delta\vec{\eta} + P_{\mathrm{ISF}}\left[L_{\mathrm{Bloch}}(\mathcal{P}_{\mathrm{eq}}\hat{n}) + \varepsilon q(\theta)\mathcal{P}_{\mathrm{eq}}\hat{n} - \partial_\theta(\mathcal{P}_{\mathrm{eq}}\hat{n})\right]. \tag{4.20}$$

Note that this is a non-homogeneous RBE and that the item in brackets satisfies the residual condition i.e.

$$\int_{\mathbb{R}^3} \hat{n}\cdot\left[L_{\mathrm{Bloch}}(\mathcal{P}_{\mathrm{eq}}\hat{n}) + \varepsilon q(\theta)\mathcal{P}_{\mathrm{eq}}\hat{n} - \partial_\theta(\mathcal{P}_{\mathrm{eq}}\hat{n})\right]\, dy = 0.$$

4.2 The polarization vector and its approximation. The depolarization time and its approximation

The polarization vector is defined by

$$\begin{aligned}\vec{P}(\theta) &= \int_{\mathbb{R}^{2d}} \vec{\eta}(\theta, y)\, dy \\ &= P_{\mathrm{ISF}}(\theta)\int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta, y)\hat{n}(\theta, y)\, dy + \int_{\mathbb{R}^{2d}} \Delta\vec{\eta}(\theta, y)\, dy,\end{aligned} \tag{4.21}$$

and the polarization is its size (Euclidean norm), i.e., $|\vec{P}(\theta)|$. Using the ISF approximation of the polarization density the polarization vector for small $\Delta\vec{\eta}$ becomes

$$\vec{P}(\theta) \approx P_{\mathrm{ISF}}(\theta)\int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta, y)\hat{n}(\theta, y)\, dy.$$

Hence in the ISF approximation the polarization satisfies

$$\begin{aligned}|\vec{P}(\theta)| &\approx P_{\mathrm{ISF}}(\theta)\left|\int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta,y)\hat{n}(\theta,y)\,dy\right| \\ &\le P_{\mathrm{ISF}}(\theta)\int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta,y)|\hat{n}(\theta,y)|\,dy = P_{\mathrm{ISF}}(\theta),\end{aligned}$$

since $|\hat{n}| = 1$ and $\int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}\,dy = 1$.

The solution to the ODE (4.8) is

$$P_{\mathrm{ISF}}(\theta) = P_{\mathrm{ISF}}(0)\exp\left(-\varepsilon\int_0^\theta q(\theta')\,d\theta'\right).$$

$q(\theta)$ is 2π-periodic. So let $q(\theta) = \bar{q} + \tilde{q}(\theta)$, where $\bar{q}$ is the average of q and $\tilde{q}$ is its zero-mean part. Then

$$\int_0^\theta q(\tau)\,d\tau = \bar{q}\theta + \int_0^\theta \tilde{q}(\tau)\,d\tau = \bar{q}(\theta) + r(\delta).$$

where $\delta \in [0, 2\pi)$ is defined by $\theta = 2\pi N + \delta$ and $r(\delta) = \int_0^\delta \tilde{q}(\tau)\,d\tau$. It follows that there exist $\alpha > 0$, such that

$$|P_{\mathrm{ISF}}(\theta) - P_{\mathrm{ISF}}(0)e^{-\varepsilon\bar{q}\theta}| = P_{\mathrm{ISF}}(0)e^{-\varepsilon\bar{q}\theta}|1 - e^{-\varepsilon r(\delta)}| \le P_{\mathrm{ISF}}(0)\alpha\varepsilon e^{-\varepsilon\bar{q}\theta}$$

This is a trivial averaging theorem where the error is not only $\mathcal{O}(\varepsilon)$ for $0 \le \theta < \mathcal{O}(1/\varepsilon)$ but it is also $\mathcal{O}(\varepsilon)$ for $0 \le \theta < \infty$. Furthermore the error decays to zero as $\theta \to \infty$. Thus for $\Delta\vec{\eta}$ and ε small,

$$\begin{aligned}\vec{P}(\theta) &\approx P_{\mathrm{ISF}}(0)\exp\left(-\int_0^\theta q(\tau)\,d\tau\right)\int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta,y)\hat{n}(\theta,y)\,dy \\ &\approx P_{\mathrm{ISF}}(0)\exp\left(-\varepsilon\bar{q}\theta\right)\int_{\mathbb{R}^{2d}} \mathcal{P}_{\mathrm{eq}}(\theta,y)\hat{n}(\theta,y)\,dy =: \vec{P}_a(\theta). \qquad (4.22)\end{aligned}$$

It is believed that in general the polarization decays, for large times, exponentially, i.e.,

$$\frac{|\vec{P}(2\pi n + \theta_0)|}{|\vec{P}(\theta_0)|} \approx e^{-\frac{C}{c}\frac{n}{\tau_{\mathrm{dep}}}}, \qquad (4.23)$$

where n is a large positive integer, C is circumference, c is the speed of light and θ_0 is of the order of $1/\varepsilon$ ($\equiv$ orbital damping time in radians) with τ_{dep} defining the depolarization time in seconds. From (4.22)

$$\frac{|\vec{P}_a(2\pi n + \theta_0|}{|\vec{P}_a(\theta_0)|} = e^{-2\pi n\varepsilon\bar{q}},$$

which consistent with (4.23) where

$$\begin{aligned}\tau_{\text{dep}}^{-1} &= \frac{c}{C} 2\pi\varepsilon\bar{q} \\ &= \frac{\pi\varepsilon c}{C} \sum_{j,k=1}^{2d} \overline{[B(\theta)B^{\mathrm{T}}(\theta)]_{j,k} \int_{\mathbb{R}^{2d}} \mathcal{P}_{\text{eq}}(\theta, y)[\partial_{y_j}\hat{n}(\theta, y)] \cdot [\partial_{y_k}\hat{n}(\theta, y)]\, dy}. \quad (4.24)\end{aligned}$$

Note that this is consistent with the so-called Derbenev–Kondratenko formulas, [12], and we will refer to (4.24) as a generalized Derbenev-Kondratenko formula for the depolarization time. If $\Delta\vec{\eta}$ is not small then the computation of τ_{dep} needs to involve $\Delta\vec{\eta}$. Then we need to use (4.21) whence the PDE in (4.20) for $\Delta\vec{\eta}$ becomes important.

Chapter 5

The averaging approximation of the reduced Bloch equation

The method of averaging (MOA), as in the ISF approximation, considers the effect of synchrotron radiation as a small perturbation to a Hamiltonian system driving the particle motion. In contrast to the ISF approximation, which minimizes the residual for the approximation of polarization density w.r.t. the reduced Bloch equation, the MOA leads to an effective Bloch equation for the polarization density, that can be practically integrated numerically. The exact reduced Bloch equation has time dependent coefficients, and the diffusion operator may not be fully elliptic. Therefore the reduced Bloch equation is difficult to understand analytically and difficult for a numerical method. The effective Bloch equation, removes some of the time dependence, leading to a time-independent fully elliptic Fokker-Planck operator which is more viable for the numerical analysis.

5.1 The averaging approximation. The effective Bloch equation

The RBE is derivable from the associated SDEs, (3.5) and (3.6), for which approximation methods are better developed. As a matter of fact, here we focus on the difficulties above in the SDEs, rather than in the RBE. For this purpose we again consider (3.5)

$$Y' = (A(\theta) + \varepsilon\delta A(\theta))Y + \sqrt{\varepsilon}B(\theta)\xi(\theta), \quad Y(0) = Y_0,$$

where $A(\theta)$ is a Hamiltonian matrix. Recall that Y is considered to be $\mathcal{O}(1)$ and that ε is chosen so that $\delta A(\theta)$ and $B(\theta) \in \mathbb{R}^{2d\times m}$ are of order 1. The term $\sqrt{\varepsilon}B(\theta)$ corresponds to the quantum noise and the square root is needed for the balance of radiation damping and quantum noise. As in Chapter 4, the synchrotron radiation has a small effect in the SDE so that ε is small.

Equation (3.5) can be approximated using the MOA to eliminate the θ-dependent coefficients in (3.5). When (3.5) is combined with (3.6) the θ independent coefficients will allow for a numerical method which can integrate the resultant RBE efficiently over long times. This has the added benefit of deepening our analytical understanding just as a perturbation analysis usually does. We will find the effective Bloch equation by refining the averaging technique presented in Section 2.1.4 in [46].

Because the process Y is Gaussian, if Y_0 is Gaussian, all the information is in its mean m and covariance K and they evolve by the ODEs

$$m' = (A(\theta) + \varepsilon\delta A(\theta))m, \tag{5.1}$$

$$K' = (A(\theta) + \varepsilon\delta A(\theta))K + K(A(\theta) + \varepsilon\delta A(\theta))^{\mathrm{T}} + \varepsilon B(\theta)B^{\mathrm{T}}(\theta). \tag{5.2}$$

In (5.2) the δA terms and the B terms are balanced at $O(\varepsilon)$ and so can be treated together in first order perturbation theory. As mentioned in Section 3.1 this is also

the reason for the $\sqrt{\varepsilon}$ in (3.5). We do not include the spin equation (3.6) in the averaging because the joint $(Y, \vec{S})$ process is not Gaussian. As mentioned before, the spin equation (3.6) has a quadratic nonlinearity since it is bilinear in Y and $\vec{S}$ so that the joint-moment equations do not close. Thus here we will apply averaging to the Y process only and discuss the spin motion after that. However, see Remark 15 below which outlines a plan for an approach where the spin equation is included in the averaging.

To apply the MOA to (5.1) and (5.2), we must transform them to a standard form for averaging. We do this by using a fundamental solution matrix (FSM) Ψ of the unperturbed $\varepsilon = 0$ part of (5.1), i.e.,

$$\Psi' = A(\theta)\Psi.$$

A convenient FSM will be discussed later and in Appendix A. Next we transform Y, m and K into U, m_U and K_U via

$$Y = \Psi(\theta)U, \quad m = \Psi(\theta)m_U, \quad K = \Psi(\theta)K_U\Psi^{\mathrm{T}}(\theta), \tag{5.3}$$

and (3.5), (5.1) and (5.2) are transformed to

$$U' = \varepsilon\mathcal{D}(\theta)U + \sqrt{\varepsilon}\Psi^{-1}(\theta)B(\theta)\xi(\theta), \tag{5.4}$$

$$m_U' = \varepsilon\mathcal{D}(\theta)m_U, \tag{5.5}$$

$$K_U' = \varepsilon(\mathcal{D}(\theta)K_U + K_U\mathcal{D}^{\mathrm{T}}(\theta)) + \varepsilon\mathcal{E}(\theta). \tag{5.6}$$

Here $\mathcal{D}(\theta)$ and $\mathcal{E}(\theta)$ are defined by

$$\mathcal{D}(\theta) = \Psi^{-1}(\theta)\delta A(\theta)\Psi(\theta), \tag{5.7}$$

$$\mathcal{E}(\theta) = \Psi^{-1}(\theta)B(\theta)B^{\mathrm{T}}(\theta)\Psi^{-T}(\theta). \tag{5.8}$$

Of course, since the transformation (5.3) is exact, (5.4)–(5.6) carry the same information as (3.5), (5.1) and (5.2).

Now, applying the MOA to (5.5) and (5.6), we obtain a Gaussian process V with mean and covariance matrix

$$m_V' = \varepsilon\overline{\mathcal{D}}m_V, \tag{5.9}$$

$$K_V' = \varepsilon(\overline{\mathcal{D}}K_V + K_V\overline{\mathcal{D}}^{\mathrm{T}}) + \varepsilon\overline{\mathcal{E}}, \tag{5.10}$$

where the bar denotes θ-averaging, i.e., the operation $\lim_{T\to\infty}(1/T)\int_0^{\mathrm{T}} d\theta\cdots$. For a calculation of $\overline{\mathcal{D}}$ and $\overline{\mathcal{E}}$ using the FSM of Appendix A, see Appendix B. For a physically reasonable A, Ψ is a quasiperiodic function whence $\mathcal{D}$ and $\mathcal{E}$ are quasiperiodic functions so that their θ averages $\overline{\mathcal{D}}$ and $\overline{\mathcal{E}}$ exist. By averaging theory the averaging errors are $\mathcal{O}(\varepsilon)$, i.e

$$|m_U(\theta) - m_V(\theta)| \leq C_1(T)\varepsilon,$$
$$|K_U(\theta) - K_V(\theta)| \leq C_2(T)\varepsilon,$$

on an interval $0 \leq \theta \leq T/\varepsilon$ where T is a constant (also see [47, 48, 49, 50]) and ε is small. However, we expect to be able to show that these estimates are uniformly valid on $[0,\infty)$ so that an accurate estimate of the orbital equilibrium can be found. A trivial example of this is the analysis of (4.8) at the end of Chapter 4 where we approximate (4.8) by the averaged equation $P'_{\mathrm{ISF,a}} = -\varepsilon\bar{q}P_{\mathrm{ISF,a}}$. For more details on the ISF approximation see Section 5.2 below.

A key point now is that every Gaussian process V, whose mean m_V and covariance matrix K_V satisfy the ODEs (5.9) and (5.10), also satisfies the SDE

$$V' = \varepsilon\overline{\mathcal{D}}V + \sqrt{\varepsilon}C(\xi_1,...,\xi_k)^{\mathrm{T}}. \tag{5.11}$$

Here $\xi_1,...,\xi_k$ are statistically independent versions of the white noise process and C is a $2d\times k$ matrix which satisfies $CC^{\mathrm{T}} = \overline{\mathcal{E}}$ with $k = \mathrm{rank}(\overline{\mathcal{E}})$ (note that in general $k \neq m$). Since $m_U(\theta) = m_V(\theta) + O(\varepsilon)$ and $K_U(\theta) = K_V(\theta) + O(\varepsilon)$, and U and V are the Gaussian processes determined by their respective means and covariances, we get $U(\theta) \approx V(\theta)$ in the sense that their density functions are close. In particular

$Y(\theta) \approx \Psi(\theta)V(\theta) = \Phi(\theta)U(\theta)$. Conversely, the mean vector m_V and covariance matrix K_V of every V in (5.11) satisfy the ODEs (5.9) and (5.10).

Remark 13. *It is likely that stochastic averaging techniques can be applied directly to* (5.4) *giving* (5.11) *as an approximation (see [51] and references therein). However, because* (5.4) *is linear and defines a Gaussian process, the theory for getting to* (5.11) *from the ODEs for the moments could not be simpler, even though it is indirect.*

To proceed with an analysis of (5.11) and its associated Fokker-Planck equation we need an appropriate Ψ and we note that $\Psi(\theta) = \Phi(\theta)R$ is an FSM, if R is an arbitrary invertible $2d \times 2d$ matrix and Φ is the principal solution matrix (PSM), i.e., $\Phi' = A(\theta)\Phi$, $\Phi(0) = I_{2d}$. Thus choosing Ψ boils down to choosing a good R (note that $R = \Psi(0)$). For the relevant discussion, see Appendix A.

Choosing R as in Appendix A and using Appendix B we calculate $\overline{\mathcal{D}}$ and $\overline{\mathcal{E}}$. From Appendix B it follows that $\overline{\mathcal{D}}$ has block diagonal form and $\overline{\mathcal{E}}$ has diagonal form. Explicitly, for $d = 3$,

$$\overline{\mathcal{D}} = \begin{pmatrix} \mathcal{D}_1 & 0_{2\times 2} & 0_{2\times 2} \\ 0_{2\times 2} & \mathcal{D}_3 & 0_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} & \mathcal{D}_5 \end{pmatrix},$$

$$\mathcal{D}_\alpha = \begin{pmatrix} a_\alpha & b_\alpha \\ -b_\alpha & a_\alpha \end{pmatrix}, \quad (\alpha = 1, 3, 5),$$

and $\overline{\mathcal{E}} = \mathrm{diag}(\mathcal{E}_1, \mathcal{E}_1, \mathcal{E}_3, \mathcal{E}_3, \mathcal{E}_5, \mathcal{E}_5)$ with $a_\alpha \leq 0$ and $\mathcal{E}_1, \mathcal{E}_3, \mathcal{E}_5 \geq 0$.

To include the spin note that, under the transformation $Y \mapsto U$, (3.5) and (3.6) become

$$U' = \varepsilon\mathcal{D}(\theta)U + \sqrt{\varepsilon}\Psi^{-1}(\theta)B(\theta)\xi(\theta), \tag{5.12}$$

$$\vec{S}' = \Omega(\theta, \Psi(\theta)U)\vec{S}. \tag{5.13}$$

Now, as we have just mentioned, U is well approximated by V, i.e., $\mathcal{P}_U \approx \mathcal{P}_V$, so that we expect $(U, \vec{S})$ to be well approximated by $(V, \vec{T})$ where

$$V' = \varepsilon \overline{\mathcal{D}} V + \sqrt{\varepsilon} C(\xi_1, ..., \xi_k)^{\mathrm{T}}, \tag{5.14}$$

$$\vec{T}' = \Omega(\theta, \Psi(\theta)V)\vec{T}, \tag{5.15}$$

and where (5.14) is a repeat of (5.11). Hence we expect $\mathcal{P}_{U\vec{S}}$ to be well approximated by $\mathcal{P}_{V\vec{T}}$ and this is work in progress.

With (5.14) and (5.15) the evolution equation for the spin-orbit probability density $\mathcal{P}_{V\vec{T}} = \mathcal{P}_{V\vec{T}}(\theta, v, \vec{t})$ is the following Fokker-Planck equation:

$$\partial_\theta \mathcal{P}_{V\vec{T}} = L_V(v)\mathcal{P}_{V\vec{T}} - \sum_{j=1}^{3} \partial_{t_j} \left(\left(\Omega(\theta, \Psi(\theta)v)\vec{t} \right)_j \mathcal{P}_{V\vec{T}} \right), \tag{5.16}$$

where

$$L_V = -\varepsilon \sum_{j=1}^{2d} \partial_{v_j} (\overline{\mathcal{D}} v)_j + \frac{\varepsilon}{2} \sum_{j=1}^{2d} \overline{\mathcal{E}}_{jj} \partial_{v_j}^2 \tag{5.17}$$

The degrees of freedom are uncoupled in L_V. For example if $d = 3$ then by (5.17),

$$L_V = L_{V,1} + L_{V,3} + L_{V,5},$$

where each $L_{V,\alpha}$ is an operator in one degree of freedom (two dimensions) and is determined by $\mathcal{D}_\alpha$ and $\mathcal{E}_\alpha$ via (5.17) ($\alpha = 1, 3, 5$). This is important for our numerical approach.

The polarization density $\vec{\eta}_V$ corresponding to $\mathcal{P}_{V\vec{T}}$ is defined by

$$\vec{\eta}_V(\theta, \mathrm{v}) = \int_{\mathbb{R}^3} \vec{t}\, \mathcal{P}_{V\vec{T}}(\theta, v, \vec{t})\, d\vec{t}, \tag{5.18}$$

so that by (5.16), the effective RBE is by definition

$$\partial_\theta \vec{\eta}_V = L_V \vec{\eta}_V + \Omega(\theta, \Psi(\theta)v)\vec{\eta}_V. \tag{5.19}$$

The operator L_V is θ-independent this is what we need for our numerical method described in Chapter 7 to be the most efficient.

We now have $Y(\theta) = \Psi(\theta)U(\theta) \approx Y_a(\theta) := \Psi(\theta)V(\theta)$ and it follows that $\vec{\eta}_Y$ in (3.11) is given approximately by

$$\vec{\eta}_Y(\theta, y) \approx \vec{\eta}_{Y,a}(\theta, y) = \det(\Psi^{-1}(0))\vec{\eta}_V(\theta, \Psi^{-1}(\theta)y) = \det(R^{-1})\vec{\eta}_V(\theta, \Psi^{-1}(\theta)y).$$

Now (5.19) and the effective RBE for $\vec{\eta}_{Y,a}$ carry the same information. However in general the effective RBE for $\vec{\eta}_{Y,a}$ does not have the nice feature of L_V being θ-independent. Hence we discretize (5.19) rather than the effective RBE for $\vec{\eta}_{Y,a}$.

5.2 Comments

We first mention a feature of $\vec{\eta}_V$ which is helpful for finding an appropriate numerical phase-space domain for $\vec{\eta}_V$. The orbital probability density $\mathcal{P}_V$ corresponding to $\mathcal{P}_{V\vec{T}}$ is defined by

$$\mathcal{P}_V(\theta, v) = \int_{\mathbb{R}^3} \mathcal{P}_{V\vec{T}}(\theta, v, \vec{t})\, d\vec{t},$$

and in analogy to the beam frame (See Remark 6 in Section 3.3) physically meaningful densities $\mathcal{P}_{V\vec{T}}$ have the property that $\mathcal{P}_{V\vec{T}}(\theta, v, \vec{t}) = 0$ if $|\vec{t}| > 1$, whence by (5.18),

$$\begin{aligned} |\vec{\eta}_V(\theta, v)| &= \left| \int_{\mathbb{R}^3} \vec{t}\,\mathcal{P}_{V\vec{T}}(\theta, v, t)\, d\vec{t} \right| \leq \int_{\mathbb{R}^3} |\vec{s}|\mathcal{P}_{V\vec{T}}(\theta, v, t)\, d\vec{t} \\ &\leq \int_{\mathbb{R}^3} \mathcal{P}_{V\vec{T}}(\theta, v, t)\, d\vec{t} = \mathcal{P}_V(\theta, v), \end{aligned}$$

so that the numerical phase space domain for $\vec{\eta}_V$ can be identified with the numerical phase space domain for $\mathcal{P}_V$. The latter is easy to find since we generally use exact expressions of $\mathcal{P}_V$, e.g., the one for orbital equilibrium.

We now make several remarks on the validity of the approximation leading to (5.14) and (5.15) and thus to (5.19).

Remark 14. *The averaging which leads to* (5.19) *affects only the orbital variables. It was justified by using the fact that* (5.12) *is linear so that it defines a Gaussian process when the initial condition is Gaussian. This allows us to apply the MOA to the first and second moments rather than the SDEs themselves.*

Remark 15. *We cannot include the spin equation* (5.13) *in the averaging because* (5.13) *has a quadratic nonlinearity and the system of spin-orbit moment equations do not close. However in future work, we will pursue approximating the system* (5.12) *and* (5.13) *using stochastic averaging as in [51]. We will split* Ω *into two pieces:* $\Omega(\theta, y) = \Omega_0(\theta) + \tilde{\varepsilon}\omega(\theta, y)$ *where* Ω_0 *is the reference-orbit contribution to* Ω *and* $\tilde{\varepsilon}$ *is chosen so that* ω *is* $O(1)$*. Then, in the case where* $\tilde{\varepsilon} = \varepsilon$*,* (5.13) *becomes* $\vec{S}' = \Omega_0(\theta)\vec{S} + \varepsilon\omega(\theta, \Psi(\theta)U)\vec{S}$*. By letting* $\vec{S}(\theta) = X(\theta)\vec{T}(\theta)$ *where* $X' = \Omega_0(\theta)X$ *we obtain* $\vec{T}' = \varepsilon\mathfrak{D}(\theta, U)\vec{T}$ *where* $\mathfrak{D}(\theta, U) = X^{-1}(\theta)\omega(\theta, \Psi(\theta)U)X(\theta)$ *and where from* (3.6) $\mathfrak{D}(\theta, U) = \sum_{j,k=1} U_k \Psi_{j,k}(\theta) X^{-1}(\theta)\varepsilon^{-1}\Omega_j(\theta)X(\theta) =: \sum_{j,k=1} U_k H_{j,k}(\theta)$*. Our system is now*

$$\begin{aligned} U' &= \varepsilon\mathcal{D}(\theta)U + \sqrt{\varepsilon}\Psi^{-1}(\theta)B(\theta)\xi(\theta), \\ \vec{T}' &= \varepsilon\mathfrak{D}(\theta, U)\vec{T}, \end{aligned} \tag{5.20}$$

where we assume $\varepsilon^{-1}\Omega_j(\theta) = \mathcal{O}(1)$*. The associated averaged system consists of* (5.14) *and of the averaged form of* (5.20)*, i.e.,*

$$V' = \varepsilon\overline{\mathcal{D}}V + \sqrt{\varepsilon}C(\xi_1, ..., \xi_k)^{\mathrm{T}}, \tag{5.21}$$

$$\vec{T}_a' = \varepsilon\overline{\mathfrak{D}}(V)\vec{T}_a. \tag{5.22}$$

where $\overline{\mathfrak{D}}(V) = \sum_{k=1} V_k \overline{H}_{j,k}$*. It seems likely that* $\vec{S}(\theta) \approx X(\theta)\vec{T}_a(\theta)$ *for* $0 \leq \theta < \mathcal{O}(1/\varepsilon)$ *in the sense that their probability density functions are close, which we hope to prove as a part of our future work.*

We now make a remark on future work with regard to a generalization of the ISF approximation. To keep the remark simple we focus on the system of SDEs

(5.21) and (5.22) of Remark 15 and restrict our discussion to the case where $d = 1$ and $a_1 = -\mathcal{E}_1$ (to address the general case of (5.21) and (5.22) is more tedious but straightforward). We write the Fokker-Planck equation for the phase space density and the RBE associated with the system SDEs (5.21) and (5.22) as

$$\partial_\theta \mathcal{P}_V = L_V \mathcal{P}_V \ , \tag{5.23}$$

$$\partial_\theta \vec{\eta}_V = L_{\mathrm{Bloch},V} \vec{\eta}_V \equiv L_V \vec{\eta}_V + \varepsilon \overline{\mathfrak{D}}(v) \vec{\eta}_V \ . \tag{5.24}$$

We define the function $\mathcal{P}_{V,\mathrm{eq}} : \mathbb{R}^2 \to \mathbb{R}$ by

$$\mathcal{P}_{V,\mathrm{eq}}(v) := \frac{1}{\pi} e^{-v^{\mathrm{T}} v}. \tag{5.25}$$

Note that $\mathcal{P}_{V,\mathrm{eq}}$ is a stationary solution of (5.23). We denote by $\vec{L}^2_{\mathrm{eq}}$ the set of functions $\vec{f} : \mathbb{R}^2 \to \mathbb{C}^3$ for which $f_1, f_2, f_3 \in L^2$ and

$$\int_{\mathbb{R}^2} \frac{1}{\mathcal{P}_{V,\mathrm{eq}}(v)} |\vec{f}(v)|^2 \, dv < \infty,$$

where L^2 is the Hilbert space of complex valued square integrable functions. As is common we identify functions in L^2 which are equal almost everywhere (same for $\vec{L}^2_{\mathrm{eq}}$). We define the function $\langle \cdot, \cdot \rangle_{\mathrm{eq}} : \vec{L}^2_{\mathrm{eq}} \times \vec{L}^2_{\mathrm{eq}} \to \mathbb{C}$ by

$$\langle \vec{f}, \vec{g} \rangle_{\mathrm{eq}} := \int_{\mathbb{R}^2} \frac{1}{\mathcal{P}_{V,\mathrm{eq}}(v)} \left(\vec{f}(v) \right)^\dagger \vec{g}(v) \, dv, \tag{5.26}$$

where $\dagger$ denotes Hermitian conjugation. Note that $\langle \cdot, \cdot \rangle_{\mathrm{eq}}$ is an inner product on $\vec{L}^2_{\mathrm{eq}}$ and that the complex vector space $\vec{L}^2_{\mathrm{eq}}$ is a Hilbert space w.r.t. this inner product. The Hilbert space norm $\| \cdot \|_{\mathrm{eq}}$ is determined via (5.26) by

$$\|\vec{f}\|_{\mathrm{eq}} := \sqrt{\langle \vec{f}, \vec{f} \rangle_{\mathrm{eq}}}. \tag{5.27}$$

Let $\vec{\eta}_{\mathrm{ISF}}$ be the ISF approximation associated with the RBE (5.24) for the orbital density $\mathcal{P}_{V,\mathrm{eq}}$, i.e.,

$$\vec{\eta}_{\mathrm{ISF}}(\theta, v) = P_{\mathrm{ISF}}(\theta) \mathcal{P}_{V,\mathrm{eq}}(v) \hat{n}(v), \tag{5.28}$$

where $\hat{n}$ is an ISF solving the nonradiative part of (5.24). [1] Note by Chapter 4 that P_{ISF} satisfies the ODE

$$P'_{\mathrm{ISF}} = -\varepsilon q_V(\theta) P_{\mathrm{ISF}}, \tag{5.29}$$
$$q_V(\theta) = \bar{q}_V = \frac{1}{2}\mathcal{E}_1 \sum_{j=1}^{2} \int_{\mathbb{R}^2} \mathcal{P}_{V,\mathrm{eq}}(v) |\partial_{v_j} \hat{n}(v)|^2 \, dv.$$

It follows from (5.25) that $\int_{\mathbb{R}^2} \mathcal{P}_{V,\mathrm{eq}}(v)\, dv = 1$ hence by (5.26) and (5.27)

$$\|\mathcal{P}_{V,\mathrm{eq}} \hat{n}\|_{\mathrm{eq}} = 1. \tag{5.30}$$

Note, by (5.28),(5.30), that $\|\vec{\eta}_{\mathrm{ISF}}(\theta, \cdot)\|_{\mathrm{eq}} = |P_{\mathrm{ISF}}(\theta)|$.

The key fact here is that, with the inner product $\langle \cdot, \cdot \rangle_{\mathrm{eq}}$ at our disposal, the minimal residual condition (4.7) from Chapter 4 can be written as

$$\langle \Delta \vec{r}(\theta, \cdot), \mathcal{P}_{V,\mathrm{eq}} \hat{n} \rangle_{\mathrm{eq}} = 0, \tag{5.31}$$

which is a condition known from Galerkin's method. Since (5.31) is a Galerkin condition it can be generalized to (5.35) using $\vec{\eta}_{\mathrm{ISF},N}$, to be defined below, which approximates $\vec{\eta}_V$ arbitrarily well for sufficiently large N and whose leading part, $\vec{\eta}_{\mathrm{ISF},1}$, is the ISF approximation. The approximation $\vec{\eta}_{\mathrm{ISF},N}$ of $\vec{\eta}_V$ is based on an orthonormal basis $b_1, b_2, ...$ of $\vec{L}^2_{\mathrm{eq}}$ whose first basis vector is $b_1 = \mathcal{P}_{V,\mathrm{eq}} \hat{n}$ and whose remaining basis vectors $b_2, b_3, ...$ are obtained by a Gram-Schmidt procedure based on any natural and convenient orthonormal basis $\tilde{b}_1, \tilde{b}_2, ...$ of $\vec{L}^2_{\mathrm{eq}}$ with the property that $b_1, \tilde{b}_1, \tilde{b}_2, ...$ are linearly independent. [2] Thus a solution $\vec{\eta}_V$ of (5.24) for which $\vec{\eta}_V(\theta, \cdot) \in \vec{L}^2_{\mathrm{eq}}$, can be written as

$$\vec{\eta}_V(\theta, v) = \sum_{j=1}^{\infty} P_{\mathrm{ISF},j}(\theta) b_j(v), \tag{5.32}$$

[1] Since the coefficients of (5.24) are θ-independent we assume that a θ-independent ISF exists. We also assume that $\hat{n}$ is smooth enough such that $\mathcal{P}_{V,\mathrm{eq}} \hat{n} \in \vec{L}^2_{\mathrm{eq}}$ (this is for example the case when $\hat{n}$ is continuous).

[2] Most convenient is a basis $\tilde{b}_1, \tilde{b}_2, ...$ of eigenfunctions of L_V. These eigenfunctions are easy to compute, see, e.g. [52].

where $P_{\mathrm{ISF},j}(\theta) \in \mathbb{C}$. The truncation of (5.32) gives us

$$\vec{\eta}_V(\theta, v) \approx \vec{\eta}_{\mathrm{ISF},N}(\theta, v) := \sum_{j=1}^{N} P_{\mathrm{ISF},j}(\theta) b_j(v), \tag{5.33}$$

where N is a positive integer. Note for $N = 1$ that

$$\vec{\eta}_{\mathrm{ISF},1} = \vec{\eta}_{\mathrm{ISF}}, \quad P_{\mathrm{ISF},1} = P_{\mathrm{ISF}}.$$

The residual of $\vec{\eta}_{\mathrm{ISF},N}$ w.r.t. (5.24) is the function $\Delta\vec{r}_N = \Delta\vec{r}_N(\theta, v)$ defined by

$$\Delta\vec{r}_N := (\partial_\theta - L_{\mathrm{Bloch},V})\vec{\eta}_{\mathrm{ISF},N}. \tag{5.34}$$

Note that, for $N = 1$,

$$\Delta\vec{r}_1 = \Delta\vec{r}.$$

With $\Delta\vec{r}_N$ at hand the minimal residual condition (5.31) generalizes via Galerkin's method to

$$0 = \langle \Delta\vec{r}_N(\theta, \cdot), b_1\rangle_{\mathrm{eq}} = \cdots = \langle \Delta\vec{r}_N(\theta, \cdot), b_N\rangle_{\mathrm{eq}}, \tag{5.35}$$

where $N = 1, 2, ...$ In the special case, $N = 1$, (5.35) becomes (5.31). The ODE (5.29) generalizes to an ODE system for the functions $P_{\mathrm{ISF},1}, ..., P_{\mathrm{ISF},N}$. To obtain this system we compute by (5.33) and (5.34)

$$\begin{aligned}\Delta\vec{r}_N(\theta, \cdot) &= (\partial_\theta - L_{\mathrm{Bloch},V})\vec{\eta}_{\mathrm{ISF},N} \\ &= \sum_{j=1}^{N} \Big(P'_{\mathrm{ISF},j}(\theta) b_j - P_{\mathrm{ISF},j}(\theta) L_{\mathrm{Bloch},V} b_j \Big),\end{aligned}$$

hence, by the orthonormality of the b_j,

$$\langle \Delta\vec{r}_N(\theta, \cdot), b_k\rangle_{\mathrm{eq}} = P'_{\mathrm{ISF},k}(\theta) - \sum_{j=1}^{N} P_{\mathrm{ISF},j}(\theta) \langle L_{\mathrm{Bloch},V} b_j, b_k\rangle_{\mathrm{eq}}, \tag{5.36}$$

where $k = 1, ..., N$. It follows from (5.36) that (5.35) results in the following system of ODEs for the functions $P_{\mathrm{ISF},1}, ..., P_{\mathrm{ISF},N}$:

$$P'_{\mathrm{ISF},k} = \sum_{j=1}^{N} P_{\mathrm{ISF},j} \langle L_{\mathrm{Bloch},V} b_j, b_k \rangle_{\mathrm{eq}}, \tag{5.37}$$

where $k = 1, ..., N$. In the special case, $N = 1$, (5.37) becomes (5.29).

The above outline of generalizing the ISF approximation was focused on the system of SDEs (5.21) and (5.22). However our future work on generalizing the ISF approximation will not be restricted to (5.21) and (5.22). For example we will address the models SM1 and SM3 to be introduced in Chapter 6 where the ISF is θ-independent as for (5.21) and (5.22). Moreover we will modify the above outline for situations where the ISF is θ-dependent, e.g., for the system of SDEs (5.14) and (5.15).

Chapter 6

Simple models

In this chapter we describe two simple models that are useful for testing the numerical method presented in the next chapter. The first model has the structure of a one-degree-of-freedom model (SM1) and the second is its extension to three degrees of freedom (SM3). We will see that these two models capture some of the effects on polarization in realistic machines. SM1 is adapted from the so-called single resonance model, [15, 23].

The single resonance model is frequently used to describe the spin motion of protons in the presence of vertical betatron motion in a storage ring. SM3, is used in Chapter 7 to verify the numerical method for the reduced Bloch equation posed is six phase space dimensions by comparing the numerical solution to the numerical solution obtained for the first model.

We begin with SM1. In a simple ring with bending magnets which bend just in the horizontal plane, and quadrupoles and drift spaces, the vector of the ISF on the reference (design) orbit, $\hat{n}_0(\theta) := \hat{n}(\theta, y)|_{y=0}$, is vertical. Spins on the reference orbit precess around $\hat{n}_0(\theta)$ and the number of precessions per turn is called the design spin tune and denoted by ν_0. According to the Thomas-BMT equation the angle

between a spin and $\hat{n}_0(\theta)$ is constant in θ. However, spins of particles which execute vertical betatron oscillations, experience radial magnetic fields in the quadrupoles so that the aforementioned angle is no longer constant. The number of vertical betatron oscillations per turn is called the vertical betatron tune and we denote it by ν_v here. Intuition suggests that if the frequency of unperturbed spin precession is related to the frequency of the perturbation from the quadrupoles by the relationship $\nu_0 \approx \nu_\mathrm{v} + k$ where k is an integer, the angle between a spin and $\hat{n}_0(\theta)$ can vary largely. For an ensemble of spins, which are initially parallel to $\hat{n}_0$, their average projection on $\hat{n}_0$ then falls, i.e., the ensemble becomes depolarized. The polarization might even oscillate. The condition $\nu_0 = \nu_\mathrm{v} + k$ is called spin-orbit resonance. Stability of spin motion on the reference orbit requires that ν_0 is not an integer. Stability of the vertical orbital motion requires that ν_v is not an integer.

Consider a particle undergoing vertical betatron motion without damping and photon emission. Owing to the discontinuities (in θ) of the quadrupole fields and non-periodicity of the orbital motion, the θ dependence of the radial fields seen by a particle is complicated. However, the θ dependence of the radial fields can be represented as a sum of Fourier harmonics with "tunes" $\nu_\mathrm{v} + m$ with m being an integer. Each harmonic describes a sinusoidal radial field component. Moreover, each sinusoidal radial field can be represented as the sum of two horizontal fields counter-rotating around $\hat{n}_0$ with tune $\nu_\mathrm{v} + m$. One of the rotating fields of a Fourier component rotates in the same sense as the natural precession of a spin around $\hat{n}_0$ and when its tune is close to ν_0, i.e., close to resonance, so that it would rotate almost in step with an unperturbed spin, it produces a large disturbance to the spin motion. The amplitude of this component is represented in the Thomas-BMT equation by the so-called *resonance strength*, denoted here by $\sigma_0 > 0$. The companion counter-rotating field is permanently out of phase with the unperturbed spin motion and its effect averages away. Likewise the other Fourier components can be neglected and we have reduced the phenomenology to that of a single resonance. A justification for

this approximation might be also possible using the MOA, [53]. The above approach is called the rotating wave approximation and is used to describe the effects of (i) external radio-frequency magnetic fields in the measurement of beam energy, (ii) the measurement of magnetic field strength, (iii) spin-resonance tomography in medicine and (iv) investigations in condensed matter physics. See [24] for the original work. As we shall see, an important feature of the SM1 is that the ISF $\hat{n}$ is known so that it is easy to check the numerical method with the ISF approximation, introduced in Chapter 4.

To build SM1 we first consider the Thomas-BMT equation for a spin $\vec{S}$ in the form

$$\vec{S}' = \Omega(Y(\theta))\vec{S} \qquad \text{with} \qquad \Omega(Y) = \nu_0 \mathcal{J}_0 + \sigma_0 \sum_{j=1}^{2} \mathcal{J}_j Y_j, \tag{6.1}$$

with

$$\mathcal{J}_0 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathcal{J}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathcal{J}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix},$$

where $Y_1 = \cos(b\theta + \chi)$, $Y_2 = \sin(b\theta + \chi)$ for $b = \nu_v + m$, and χ is an arbitrary phase. Here Y_1 and Y_2 are considered to be the components of a rotating horizontal unit vector representing the direction of the rotating field and σ_0 is the (constant) resonance strength. See (7.1) in [15] too.

Now we need to build in the effects on electrons of radiation with the aim of simulating radiative depolarization and, in particular, resonant radiative depolarization and we would like to estimate the depolarization time using (4.24), which can be viewed as a generalized DK formula from [13] to compute the equilibrium polarization.

SM1 is centered on vertical betatron motion and in real rings quantum noise

and damping, together with the so-called vertical dispersion, lead to an equilibrium (i.e. one-turn periodic) distribution of the vertical positions and vertical momenta and the individual trajectories carry the noise which leads to spin diffusion and depolarization. The spread of vertical particle positions leads to a spread in the resonance strength and this is also noisy. So in effect the resonance strength becomes a stochastic quantity. Then, to calculate the depolarization we could construct an SDE for the resonance strength. Instead, we keep σ_0 constant and inject noise into (Y_1, Y_2). Then the reduced SDE is structurally similar to that of the familiar one-degree-of-freedom reduced SDE for orbital motion and toke the form of (3.5) and (3.6). The vertical dispersion is usually mainly due to vertical misalignments of the ring but since we are not dealing with a realistic ring with its realistic misalignments, we have no physical model on which to base the strength of the noise in the vertical particle position and subsequently in (Y_1, Y_2). Thus we first set up the problem in purely mathematical terms and return to the specification of realistic parameters in Section 6.4 below. Then we carry out various numerical experiments, e.g., a study of the effects of varying ν_0 in Section 6.4 and Chapter 8.

6.1 Simple model in one degree of freedom (SM1)

Following the motivation just given we now set up the SDEs for the combined motion of the vector $\vec{S}$ and (Y_1, Y_2), in the form of (3.5) and (3.6) with $A = -bJ_2$, $\delta A = -I_2$ and $B = I_2$ where I_2 is the (2×2) identity matrix and

$$J_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Thus

$$\begin{aligned} Y' &= (-bJ_2 - \varepsilon I_2)Y + \sqrt{\varepsilon}(\xi_1(\theta), \xi_2(\theta))^{\mathrm{T}}, & Y(0) &= Y_0 \\ \vec{S}' &= \Omega(Y)\vec{S}, & \vec{S}(0) &= \vec{S}_0, \end{aligned}$$

The components of the vector $(\xi_1(\theta), \xi_2(\theta))$ are statistically independent white noise processes. The factor $\sqrt{\varepsilon}$ is the noise strength and ε is the damping constant in rad^{-1}.

For this model the joint probability density satisfies the Fokker-Planck equation

$$\begin{aligned} \partial_\theta \mathcal{P}_{YS} &= L_Y \mathcal{P}_{YS} - \sum_{j=1}^{3} \partial_{s_j} \left([\Omega(y)\vec{s}]_j \mathcal{P}_{YS} \right), \\ \mathcal{P}_{YS}(0, y, \vec{s}) &= \mathcal{P}_{Y_0 S_0}(y, \vec{s}), \end{aligned} \tag{6.2}$$

where L_Y is the Fokker-Planck operator, written as a sum of the Hamiltonian Liouville term and the radiative term

$$\begin{aligned} L_Y \mathcal{P}_{YS} :=& b \left(\partial_{y_1}(y_2 \mathcal{P}_{YS}) - \partial_{y_2}(y_1 \mathcal{P}_{YS}) \right) \\ &+ \varepsilon \left(\partial_{y_1}(y_1 \mathcal{P}_{YS}) + \partial_{y_2}(y_2 \mathcal{P}_{YS}) \right) + \frac{\varepsilon}{2} \nabla^2 \mathcal{P}_{YS}. \end{aligned}$$

The orbital Fokker-Planck equation is

$$\partial_\theta \mathcal{P}_Y = L_Y \mathcal{P}_Y, \quad \mathcal{P}_Y(0, y) = \mathcal{P}_{Y_0}(y), \tag{6.3}$$

consistent with integration of (6.2) over y. Following Section 3.4 we obtain the equilibrium periodic solution to (6.3)

$$\mathcal{P}_{\text{eq}}(y) = \frac{1}{\pi} e^{-y^{\mathrm{T}} y}, \tag{6.4}$$

i.e., a Gaussian density with mean 0 and covariance $\frac{1}{2} I_2$. Assuming that the beam is initially at orbital equilibrium, then

$$Y_0 \sim \mathcal{N}_2 \left(0, \frac{1}{2} I_2 \right),$$

i.e. it has the probability density function (6.4). The reduced Bloch equation for this model reads as $\partial_\theta \vec{\eta} = L_{\text{Bloch}} \vec{\eta} := L_Y \vec{\eta} + \Omega(y) \vec{\eta}$, and so

$$\begin{aligned} \partial_\theta \vec{\eta} =& b \left(\partial_{y_1}(y_2 \vec{\eta}) - \partial_{y_2}(y_1 \vec{\eta}) \right) \\ &+ \varepsilon \left(\partial_{y_1}(y_1 \vec{\eta}) + \partial_{y_2}(y_2 \vec{\eta}) \right) + \frac{\varepsilon}{2} \nabla^2 \vec{\eta} + \Omega(y) \vec{\eta}, \end{aligned} \tag{6.5}$$

subject to the initial and boundary conditions

$$\vec{\eta}(0,y) = \int_{\mathbb{R}^3} \vec{s}\,\mathcal{P}_{Y_0S_0}(y,\vec{s})\,d\vec{s},$$
$$\lim_{y\to\infty} \vec{\eta} e^{\alpha y^{\mathrm{T}} y} = 0, \quad \text{for some } \alpha > 0.$$

This completes the construction of SM1.

6.2 ISF approximation

Following Chapter 4 we need the ISF for our simple model, namely the ISF of the single resonance model. This is given by [15] which we write as

$$\hat{n}(y) = \frac{1}{\sigma(y)} \begin{pmatrix} y_1 \\ y_2 \\ \zeta \end{pmatrix}, \quad \sigma(y) := \sqrt{y^{\mathrm{T}} y + \zeta^2}, \quad \zeta := \frac{\nu_0 - b}{\sigma_0}, \quad \sigma_0 \neq 0.$$

At spin-orbit resonance $\hat{n}$ is not defined at $y = 0$, thus we assume the non-resonant case $\nu_0 \neq b$. Resonance can be studied in SM1 by the spectral method of Chapter 7 or the MOA. As required $\hat{n}$ satisfies the non-radiative Bloch equation

$$\partial_\theta \hat{n} = \left(L_Y\big|_{\varepsilon=0} + \Omega(y)\right)\hat{n} = b\left(\partial_{y_1}(y_2\hat{n}) - \partial_{y_2}(y_1\hat{n})\right) + \Omega(y)\hat{n}. \tag{6.6}$$

Next we write $\vec{\eta}$ via Section 4.1 as

$$\vec{\eta}(\theta,y) = P_{\mathrm{ISF}}(\theta)\mathcal{P}_{\mathrm{eq}}(y)\hat{n}(y) + \Delta\vec{\eta}(\theta,y). \tag{6.7}$$

Here $P_{\mathrm{ISF}}(\theta)$ is the solution to

$$P'_{\mathrm{ISF}} = -\varepsilon q P_{\mathrm{ISF}},\ q = \frac{1}{2}\sum_{j=1}^{2}\int_{\mathbb{R}^2} \mathcal{P}_{\mathrm{eq}}(y)|\partial_{y_j}\hat{n}(y)|^2\,dy,$$
$$P_{\mathrm{ISF}}(0) = \int_{\mathbb{R}^2} \vec{\eta}(0,y)\cdot\hat{n}(y)\,dy,$$

from Chapter 4. After simplification, q can be written using the so-called exponential integral

$$q = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{\sigma^2(y)+\zeta^2}{e^{y^{\mathrm{T}}y}\sigma^2(y)}\, dy = \frac{1}{2}\left(\zeta^2-1\right)e^{\zeta^2}\mathrm{Ei}\left(-\zeta^2\right)+\frac{1}{2},$$

where Ei(x) is an exponential integral function

$$\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{e^t}{t}\, dt.$$

Plugging (6.7) into the Bloch equation we get the left hand side of (6.5) becomes

$$\partial_\theta \vec{\eta} = \partial_\theta \Delta\vec{\eta} - P_{\mathrm{ISF}}(\theta)\left(\varepsilon q(\theta)\mathcal{P}_{\mathrm{eq}}\hat{n} - \partial_\theta(\mathcal{P}_{\mathrm{eq}}\hat{n})\right). \tag{6.8}$$

For the right hand side of (6.5) we obtain

$$\begin{aligned} L_{\mathrm{Bloch}}\vec{\eta} &= L_{\mathrm{Bloch}}\Delta\vec{\eta} + P_{\mathrm{ISF}}(\theta)\varepsilon\sum_{j=1}^{2} y_j\mathcal{P}_{\mathrm{eq}}\partial_{y_j}\hat{n} \\ &\quad + P_{\mathrm{ISF}}(\theta)\left(\left(L_{\mathrm{Y}}\mathcal{P}_{\mathrm{eq}}\right)\hat{n} + \mathcal{P}_{\mathrm{eq}}L_Y\big|_{\varepsilon=0}\hat{n} + \Omega(y)\mathcal{P}_{\mathrm{eq}}\hat{n}\right) \\ &\quad + \frac{\varepsilon}{2}P_{\mathrm{ISF}}(\theta)\sum_{j=1}^{2}\left(\mathcal{P}_{\mathrm{eq}}\partial_{y_j}^2\hat{n} + 2\partial_{y_j}\mathcal{P}_{\mathrm{eq}}\partial_{y_j}\hat{n}\right) \\ &= L_{\mathrm{Bloch}}\Delta\eta + \frac{\varepsilon}{2}P_{\mathrm{ISF}}\mathcal{P}_{\mathrm{eq}}\sum_{j=1}^{2}\left(\partial_{y_j}^2\hat{n} - 2y_j\partial_{y_j}\hat{n}\right). \end{aligned} \tag{6.9}$$

Using (6.3) and (6.6), the relations (6.8) and (6.9) deliver an equation for $\Delta\vec{\eta}$, i.e. the RBE with an inhomogeneous term

$$\begin{aligned} &\partial_\theta\Delta\vec{\eta} = L_{\mathrm{Bloch}}\Delta\vec{\eta} + \varepsilon P_{\mathrm{ISF}}(\theta)P_{\mathrm{eq}}(y)\vec{f}(y), \\ &\Delta\vec{\eta}(0,y) = \vec{\eta}(0,y) - P_{\mathrm{ISF}}(0)\mathcal{P}_{\mathrm{eq}}(y)\hat{n}(y), \\ &\vec{f}(y) = \frac{1}{2}\sum_{j=1}^{2}\left(\partial_{y_j}^2\hat{n}(y) - 2y_j\partial_{y_j}\hat{n}(y)\right) + q\hat{n}(y). \end{aligned} \tag{6.10}$$

For this model, $\vec{f}(y)$ can be written explicitly in terms of q and $\hat{n}$ as

$$\vec{f}(y) = \left(\frac{1}{\sigma^2(y)}+1\right)\begin{pmatrix}0\\0\\\zeta/\sigma(y)\end{pmatrix} - \hat{n}(y)\left(\frac{\zeta^2}{\sigma^2(y)} + \frac{\sigma^2(y)+3\zeta^2}{2\sigma^4(y)}\right) + q\hat{n}(y) \tag{6.11}$$

From Chaper 4, the residual $\partial_\theta \Delta\vec{\eta} - L_{\text{Bloch}}\Delta\vec{\eta}$ must not have a component along the ISF on average, i.e. $\int_{\mathbb{R}^2} \mathcal{P}_{\text{eq}}(y)\vec{f}(y)\cdot\hat{n}(y)\,dy = 0$. To verify, we compute as follows,

$$\begin{aligned}
&\int_{\mathbb{R}^2} \mathcal{P}_{\text{eq}}\vec{f}\cdot\hat{n}\,dy \\
&\quad= \int_{\mathbb{R}^2} \mathcal{P}_{\text{eq}}(y)\left[\frac{\zeta^2}{\sigma^2(y)}\left(\frac{1}{\sigma^2(y)}+1\right) - \left(\frac{\zeta^2}{\sigma^2(y)} + \frac{\sigma^2(y)+3\zeta^2}{2\sigma^4(y)}\right)\right]\,dy \\
&\quad+ q\int_{\mathbb{R}^2} \mathcal{P}_{\text{eq}}(y)\,dy \\
&\quad= \int_{\mathbb{R}^2} \mathcal{P}_{\text{eq}}(y)\left[-\frac{\sigma^2(y)+\zeta^2}{2\sigma^4(y)}\right]\,dy + q = 0.
\end{aligned}$$

6.3 Extension to three degrees of freedom (SM3)

Our 3-degree-of-freedom model is based on SM1 described in Section 6.1. It is obtained by adding two independent blocks to A and δA, and also extending the noise term. The T-BMT term affects only the first degree of freedom. This model was constructed to test the numerical method for the reduced Bloch equation and also for demonstrating the spin dynamics in the presence of orbital motion in three degrees of freedom.

Let $Y \in \mathbb{R}^6$ and $\vec{S} \in \mathbb{R}^3$ be a spin vector. Consider the system of SDE's

$$\begin{aligned}
Y' &= [A + \varepsilon\delta A]Y + \sqrt{\varepsilon}B\xi(\theta), \\
\vec{S}' &= \Omega(Y_1, Y_2)\vec{S},
\end{aligned}$$

where

$$A = -\mathrm{diag}(b_1, b_1, b_3, b_3, b_5, b_5)J_6,$$
$$\delta A = -\mathrm{diag}(a_1, a_1, a_3, a_3, a_5, a_5),$$
$$B = I_6,$$
$$J_6 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \otimes I_3,$$

and I_3 being the (3×3) identity matrix. Here ξ is a vector of six statistically independent white noise processes and $\Omega(Y_1, Y_2)$ is the same as in one degree of freedom (see also Section 6.1). For this model the joint probability density satisfies the Fokker-Planck equation

$$\partial_\theta \mathcal{P}_{YS} = L_Y \mathcal{P}_{YS} - \sum_{j=1}^{3} \partial_{s_j} \left([\Omega(y_1, y_2)\vec{s}]_j \mathcal{P}_{YS} \right),$$

where L_Y is the 3-degree-of-freedom Fokker-Planck operator, written as a sum

$$L_Y \mathcal{P}_{YS} := -\sum_{j=1}^{6} \partial_{y_j} \left([(A + \varepsilon \delta A)y]_j \mathcal{P}_{YS} \right) + \frac{\varepsilon}{2} \nabla^2 \mathcal{P}_{YS}.$$

The orbital Fokker-Planck equation then becomes

$$\partial_\theta \mathcal{P}_Y = L_Y \mathcal{P}_Y. \tag{6.12}$$

Following Section 3.4 we obtain the equilibrium periodic solution to (6.12)

$$\mathcal{P}_{\mathrm{eq}}(y) = \frac{a_1 a_3 a_5}{\pi^3} e^{-y^{\mathrm{T}} y}.$$

The reduced Bloch equation for this model reads as

$$\partial_\theta \vec{\eta} = L_{\mathrm{Bloch}} \vec{\eta} = L_Y \vec{\eta} + \Omega(y_1, y_2)\vec{\eta} \tag{6.13}$$

subject to the initial and boundary conditions

$$\vec{\eta}(0, y) = \int_{\mathbb{R}^3} \vec{s}\, \mathcal{P}_{Y_0 S_0}(y, s)\, d\vec{s},$$
$$\lim_{y \to \infty} \vec{\eta} e^{\alpha y^{\mathrm{T}} y} = 0, \quad \text{for some } \alpha > 0.$$

Although we called this model "simple", it is challenging for numerical analysis because of the high dimensionality. In Section 6.4 we again consider SM1, and leave the discussion of the numerical issues for the next chapter.

6.4 Results of the ISF approximation for HERA

We now return to the matter of choosing realistic parameters for our SM1 of Section 6.1 with the intention, using the generalized DK formula (4.24), of mimicking the depolarization seen in a real ring and thereby demonstrating the efficacy and consistency of our approach.

For this we consider the 6.3 km electron-positron storage ring of the Hadron-Electron Ring Accelerator (HERA) [54]. This is sufficiently representative of a high energy machine for our purposes. In particular, the time scales are typical. Moreover we do not need to know the detailed layout of the ring because that has been subsumed into the Fourier decomposition as discussed in the introduction to this chapter. We choose the usual running energy of HERA, around $E = 27.54$ GeV. At this energy the design spin tune ν_0 is 62.5. For this ring the bending magnets have a bending radius of $\rho = 600$ m. The orbital damping time for HERA is 619 turns, so we set the damping constant to $\varepsilon = 1/(619 \cdot 2\pi)[\text{rad}^{-1}]$ [1] and proceed as follows.

Plugging (6.11) in (4.24) we obtain the depolarization time τ_{dep} for SM1 written as a function of $\zeta = (\nu_0 - b)/\sigma_0$

$$\tau_{\text{dep}}^{-1}(\zeta) = \frac{\pi c \varepsilon}{C} \left(\left(\zeta^2 - 1\right) e^{\zeta^2} \text{Ei}\left(-\zeta^2\right) + 1 \right).$$

Next we need the Sokolov-Ternov polarization time τ_0 given by [13]

$$\tau_0^{-1}[\text{s}^{-1}] \approx \frac{2\pi}{99} \frac{E[\text{GeV}]^5}{C[m]\rho[m]^2},$$

[1] With $b = 0$, and no noise $Y = Y_0 e^{-\varepsilon\theta} = Y_0 e^{-1}$ for $\theta = 1/\varepsilon = 619 \cdot 2\pi$.

where C is a circumference of the ring and we then define

$$\tau_{\text{tot}}^{-1} = \tau_0^{-1} + \tau_{\text{dep}}^{-1}.$$

This is the polarizing rate in the presence of both polarization and depolarization. Then, following [37], the equilibrium polarization P_{eq}, being the result of the balance of polarization build up and depolarization, is given by

$$P_{\text{eq}} = P_{\text{bks}} \frac{\tau_{\text{tot}}}{\tau_0}, \tag{6.14}$$

where P_{bks} is the Baier-Katkov-Strakhovenko polarization which for a flat ring is also the Sokolov-Ternov polarization, namely 92.38%. With the HERA parameters τ_0 is about 40 mins.

We now choose $b = 62.45$ so that $\nu_0 - b = 0.05$. We want to mimic the behaviour of a real ring without having the details of a real ring. So we choose the resonance strength as $\sigma_0 = \varepsilon \cdot 0.6405918611 = 1.647065609 \times 10^{-4}$ (recall that $\sigma_0 \propto \varepsilon$), a value chosen to give a depolarization time of about 20 minutes, namely about half of the Sokolov-Ternov polarization time. This, in turn, should give a P_{eq} of about 30%.

In Figure 6.1 we display the equilibrium polarization P_{eq} computed with resonance strengths σ_0, $2\sigma_0$ and $3\sigma_0$ for SM1. Each curve shows the equilibrium polarization from (6.14) as a function of ν_0. As we hoped, we see a family of polarization curves similar to the kind seen in measurements and Monte-Carlo simulations. As we expect, the minimum equilibrium polarization is at the spin-orbit resonance, when $\nu_0 = b$ and the "width" of the resonance is proportional to the resonance strengths. Note that a scan of ν_0 implies a scan of beam energy and that implies that σ_0 varies. However, for this narrow range of ν_0 the variation of σ_0 is small and it suffices to keep σ_0 constant at the chosen value.

In Chapter 8 we repeat this experiment using the numerical solutions to the Bloch equation and also compute the size of $\Delta\eta$ to estimate the accuracy of these

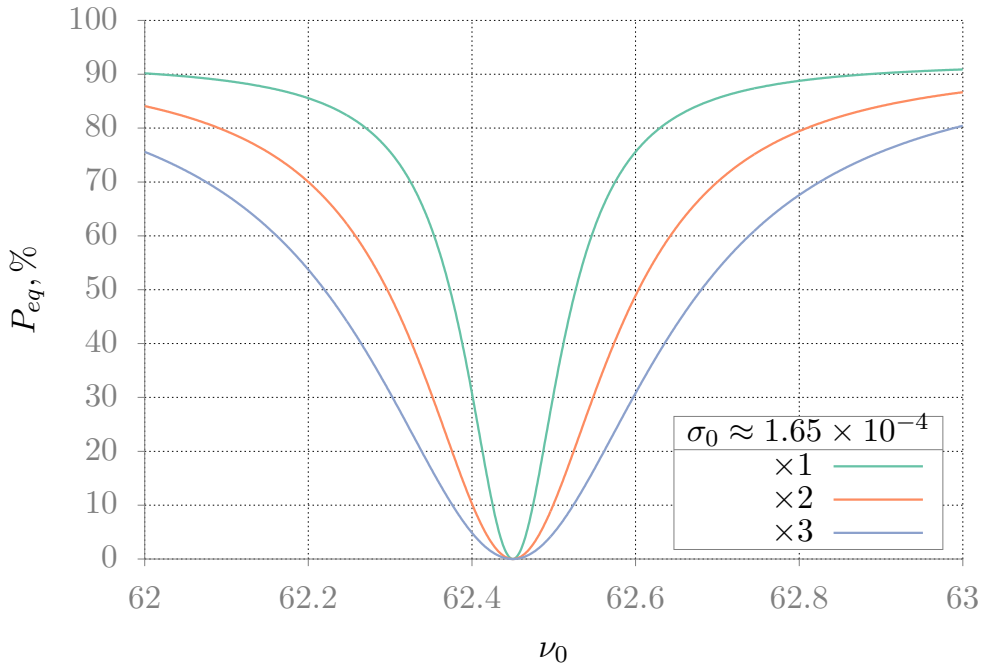

Figure 6.1: A spin-tune scan of polarlization for SM1, with HERA parameters. The curves show the equilibrium polarization as a function of ν_0. Each curve corresponds to the resonance strengths σ_0 set to base value $1.647065609 \times 10^{-4}$ and base value multiplied by 2 and 3.

results. We find that even close to resonance the error in equilibrium polarization computation is below 0.002%.

Chapter 7

The spectral method for the Bloch equation

In this chapter we develop a numerical method for solving the reduced beam frame Bloch equation presented in Chapter 3. To our knowledge this PDE has not been studied numerically in the past. As we have seen, the reduced beam-frame Bloch equation can be viewed as a system of Fokker-Planck equations with a coupling term coming from the Thomas-BMT precession. The Fokker-Planck equation, or the forward Kolmogorov equation, [55, 56, 39], is a parabolic PDE used in physics, ecology, engineering, biology, psychology, neuroscience and economics [57]. Numerical simulations of the Fokker-Planck equation are important for studying phenomena modeled with Markov diffusion processes.

The Fokker-Planck equation has been long studied and numerical methods for solving it have been actively developed since the 1970s. Examples include finite difference methods, [58, 59, 60], finite element methods, [61, 62, 58, 59] and spectral methods for both unbounded and bounded domains, [63, 64, 65]. In [63] the solutions of the Fokker-Planck equation were approximated with Chebyshev polynomials,

utilizing sparse matrices to speed up calculations.

An advantage of spectral methods is their high accuracy and efficient memory usage. With spectral methods the numerical solution is represented as a sum of N basis functions (for example, Chebyshev or Hermite polynomials, or trigonometric functions). The difference, between finite element methods and spectral methods, is that with spectral methods the basis functions take nonzero values on the entire domain, while with finite element methods the basis functions are non-zero locally on each element and zero outside. As a result, spectral methods achieve a high order of accuracy with a small number of numerical degrees of freedom (polynomial coefficients) compared to finite element methods. However, spectral differentiation matrices are typically dense. Hence spectral methods have high arithmetic intensity.

Memory usage becomes a concern for solving high-dimensional Fokker-Planck equations ($4, 6$, or more dimensions). Also, the finite difference and finite element meshes must be sufficiently refined for stability [66], and that also leads to high computational cost. Nevertheless, although computationally expensive, spectral methods are preferable for smooth, high dimensional problems posed on simple domains.

In [67] a general-purpose "τ-method" for spectral approximation of a wide class of problems is presented. In τ-methods the basis functions themselves are not required to respect whatever auxiliary conditions, for example boundary conditions, are imposed on the solution. Rather the auxiliary conditions are imposed directly on the numerical solution (basis function expansion) as so-called τ-conditions, [68]. In [67] the application of integration matrices is used, both to achieve banded representations (of "bulk" operators) and to organize the placement of the τ-conditions into the resulting systems. The ultimate goal is to achieve well-conditioned approximations, and the technique is referred to as "integration preconditioning". However, especially for higher dimensional settings, for many standard problems it is not clear that, by itself, integration preconditioning actually improves the condition number. Nonethe-

less, the technique does afford a systematic way to achieve *sparse approximations.* We focus on this aspect and think of the technique as *integration sparsification*, with the realization that further work may be needed. Typically, this will mean the use of sparse direct solvers or iterative methods with further (and genuine) preconditioning.

For Chebyshev-based polynomial approximations of equations which are second order in the derivatives, "integration preconditioning" relies on application of the double Chebyshev spectral integration matrix (along various dimensions). This is a bandwidth-5 matrix, with vanishing first superdiagonal and first subdiagonal elements. This structure, coupled with the fact that integration undoes differentiation, is the mechanism for achieving sparsity for operators with polynomial coefficients, as multiplication by a polynomial is also a banded operation in the modal space. Results for polynomial coefficients yield results for rational coefficients.

The Fourier based spectral methods utilize a finite series of trigonometric functions to approximate the solution. Fourier differential matrices are sparse, containing $\mathcal{O}(N)$ non-zero elements, but the methods are limited to problems posed on periodic domains. In [69], Fourier and Chebyshev pseudospectral (or nodal) methods are combined for solving the Laplace equation posed on a disk, in polar coordinates. In our work this technique is generalized to modal methods, to utilize the Fourier spectral method for angular variables and the Chebyshev-based method in the radial variables.

In this chapter we present the Fourier-Chebyshev method for the reduced Bloch equation (6.13) posed on a truncated six dimensional phase space. Each degree of freedom, being represented by a pair of phase space variables y_{2d-1}, y_{2d}, $d = 1, 2, 3$, is transformed into a polar coordinate pair. The method approximates the solution with tensor-product Chebyshev and Fourier basis functions in the radial and angle variables respectively. For the time evolution we use the high order additive Runge-Kutta method, [70]. The time evolution algorithm computes spectral coefficients

of the numerical solution at each time step by solving linear systems. The second order differentiation matrix for the Fourier-based method is diagonal, so, considering the Bloch equation, the computational work for the time evolution can be efficiently distributed between parallel processes. The parallel algorithm requires only local communication. Sparsity of the time-evolution linear systems is increased by using integration preconditioning corresponding to the radial directions to further reduce the computational cost.

The efficiency of our 3-degree-of-freedom algorithm for the reduced Bloch equation strongly depends on the efficiency of the inversion of the Laplace operator in the radial variables. In our implementation we use a three-dimensional code provided by S. R. Lau [71] that is designed for the three-dimensional Helmholtz equation on a rectangular block, approximated with a sparse spectral-τ method as described above. Lau's direct method has a provable start-up cost of $\mathcal{O}(n^2)$, followed by an $\mathcal{O}(n^{4/3})$ reuse cost, where $n = N^3$ is a total number of unknowns and N is the number of modes in each dimension. However, through the introduction of an iterative component, a sub-quadratic solve cost is observed empirically. This work is ongoing, but on the basis of these developments we describe inversion of the Helmholtz operator as "fast". This inversion must be performed for each Fourier mode and multiple times within the context of a time-stepping scheme. It bears mentioning that if one considers the reduced Bloch equation in Cartesian coordinates, the anticipated complexity of (reuse) solves would be $\mathcal{O}(n^{(d+1)/d})$ where d is the number of dimensions. It is unclear at this point whether an $\mathcal{O}(n^2)$ startup bottleneck would be at issue.

We first describe the method for the simple model in one degree of freedom presented in Section 6.1. Then we describe how to extend the method to three degrees of freedom. Finally we test the accuracy of the spectral method for the reduced Bloch equation in several numerical experiments.

7.1 The spectral method for the Bloch equation

Consider the reduced Bloch equation for the SM1, (6.5) of Section 6.1

$$\begin{aligned}\partial_\theta \vec{\eta} =& b\left(\partial_{y_1}(y_2\vec{\eta}) - \partial_{y_2}(y_1\vec{\eta})\right) \\ &+ \varepsilon\left(\partial_{y_1}(y_1\vec{\eta}) + \partial_{y_2}(y_2\vec{\eta})\right) + \frac{\varepsilon}{2}\nabla^2\vec{\eta} + \Omega(y)\vec{\eta},\end{aligned} \tag{7.1}$$

with the initial condition

$$\vec{\eta}(0, y) = \vec{g}(y).$$

Here $\vec{\eta} \in C^\infty\left([0,T]\times\mathbb{R}^2, \mathbb{R}^3\right)$ for optimal convergence and $\Omega(y) = \Omega_0 + \Omega_1 y_1 + \Omega_2 y_2$.

Naturally, we consider solutions to (7.1) that rapidly decay as y approaches infinity, since they are connected to the Gaussian beam distribution around the reference orbit of the lattice

$$\lim_{|y|\to\infty} \vec{\eta}(\theta, y) e^{\alpha y^{\mathrm{T}} y} = 0,$$

for some $\alpha > 0$. Following Chapter 11 in [69], we transform (7.1) into a special frame (r, ϕ), which we refer to as the frame of *symmetric polar coordinates*. The symmetric polar coordinates are polar coordinates, where the domain of the radial variable r is extended to negative values so that $r = 0$ is an interior point. Under this transformation the single-valuedness of the solution is guaranteed by a simple trigonometric property

$$\vec{\eta}(\theta, (-r)\cos\phi, (-r)\sin\phi) = \vec{\eta}(\theta, r\cos(\phi+\pi), r\sin(\phi+\pi).$$

that later we refer to as the symmetry condition. The symmetric polar coordinate transformation is

$$y_1 = r\cos\phi,\ \ y_2 = r\sin\phi,\ \ r\in(-\infty,\infty),\ \ \phi\in[0,2\pi),$$

Thus the Bloch equation for $\vec{\eta}(\theta, r\cos\phi, r\sin\phi)$ becomes

$$\partial_\theta\vec{\eta} = \varepsilon a\left(2\vec{\eta} + r\partial_r\vec{\eta}\right) - b\partial_\phi\vec{\eta} + \frac{\varepsilon}{2}\Delta\vec{\eta} + \Omega(r\cos\phi, r\sin\phi)\vec{\eta}, \tag{7.2}$$

with the initial condition

$$\vec{\eta}(0, r\cos\phi, r\sin\phi) = \vec{g}(r\cos\phi, r\sin\phi).$$

Next we use the substitution $\vec{u}(\theta, r, \phi) = r^2\vec{\eta}(\theta, r\cos\phi, r\sin\phi)$ and (7.2) becomes

$$\partial_\theta\vec{u} = \varepsilon a\left(\partial_r(r\vec{u}) - \vec{u}\right) - b\partial_\phi\vec{u} + \frac{\varepsilon}{2}\left(\partial_r^2\vec{u} + (1+\partial_\phi^2)\frac{\vec{u}}{r^2} - 3\partial_r\frac{\vec{u}}{r}\right) + \Omega(r,\phi)\vec{u}, \quad (7.3)$$

with the initial condition

$$\vec{u}(0, r, \phi) = \vec{h}(r, \phi) \equiv r^2\vec{g}(r\cos\phi, r\sin\phi),$$

and where, in an abuse of notation,

$$\Omega(r, \phi) = \Omega_0 + \Omega_1 r\cos\phi + \Omega_2 r\sin\phi.$$

Note that, $\vec{u}/r$ and $\vec{u}/r^2$ are smooth since $\vec{\eta}$ is smooth. Also note that all differential operators have been placed to the left-most positions. This is done to obtain the cancellation of the discretized differential operators by integration at a later stage, when integration preconditioning is performed.

Instead of posing (7.3) on the infinite domain, we impose it on a disk $\mathcal{D}$

$$\mathcal{D} = [-r_{\max}, r_{\max}] \times [0, 2\pi),$$

The boundary conditions are periodic in ϕ and we take $r_{\max}$ large enough to impose homogeneous Dirichlet boundary conditions

$$u(\theta, r_{\max}, \phi) \approx 0.$$

The domain is discretized with an $N \times M$ grid, that is the tensor product of Chebyshev-Gauss-Legendre nodes r_i and equidistant angles ϕ_j

$$\left\{r_i = r_{\max}\cos\left(\frac{(i-1)\pi}{N-1}\right),\ i = 1, \ldots, N\right\} \times \left\{\phi_j = \frac{(j-1)2\pi}{M},\ j = 1, \ldots, M\right\}.$$

The origin is a special point of the domain $\mathcal{D}$ and so we avoid it in our discretization by requiring N to be an even number. Furthermore to impose the symmetry we require M to be even. Thus, the points (r_i, ϕ_j) and $(r_{N-i+1}, \phi_{j+M/2})$ are the same points on the Cartesian plane and thus the symmetry condition becomes

$$u(\theta, r_i, \phi_j) = u(\theta, r_{N-i+1}, \phi_{j+M/2}). \tag{7.4}$$

The θ-domain (note that we treat θ as time) is discretized by a uniform grid with increments Δt, i.e.,

$$t_\nu = \nu \Delta t, \ \nu = 0, 1, \dots \ .$$

At each time step ν the approximation of the solution is represented by its coefficients in the Chebyshev-Fourier expansion that approximates the values of u. Equivalently, the approximations to u are represented as truncated Fourier expansions in the ϕ variable, with coefficients that are the sums of Chebyshev polynomials in r centered at $r = 0$. At the initial azimuth $\theta = 0$ we represent the initial condition for u as

$$\vec{h}(r, \phi) \approx \sum_{i=0}^{N-1} \sum_{k=-M/2+1}^{M/2} \hat{u}^0_{i,k} T_i(r/r_{\max}) e^{\mathbf{i}k\phi} =: \vec{p}(0, r, \phi), \tag{7.5}$$

where

$$T_i(r/r_{\max}) = \cos(i \cos^{-1}(r/r_{\max})),$$

are the Chebyshev polynomials of the first kind. We use a projection or an interpolation procedure on the grid and a discrete Fourier transform to find the coefficients in (7.5). The approximation for $\vec{u}$ at each time step can be then written as

$$\vec{u}(t_\nu, r, \phi) \approx \sum_{i=0}^{N-1} \sum_{k=-M/2+1}^{M/2} \hat{u}^\nu_{i,k} T_i(r/r_{\max}) e^{\mathbf{i}k\phi} =: \vec{p}(t_\nu, r, \phi), \quad \nu = 0, 1, \dots, \tag{7.6}$$

where the coefficients are computed via the time evolution algorithm described in Section 7.2.

Remark 16. *Due to the symmetry condition* (7.4) *half of the coefficients must be zero, and thus they are not evolved in time. That becomes clear after plugging* (7.5)

or (7.6) *into* (7.4) *and noting that if* i *and* k *are of different parity (for example, if* i *is odd and* k *is even) then* $\hat{u}^{\nu}_{i,k}$ *is 0.*

The coefficients $\hat{u}^{\nu}_{i,k}$ are three-dimensional vectors. Each of the three components of $\hat{u}^{\nu}_{i,k}$ and each Fourier mode is treated independently. So for convenience we represent the k-th Fourier mode as a three-vector of Chebyshev coefficients, U^{ν}_k, V^{ν}_k and W^{ν}_k,

$$\begin{pmatrix} (U^{\nu}_k)_i \\ (V^{\nu}_k)_i \\ (W^{\nu}_k)_i \end{pmatrix} \equiv \hat{u}^{\nu}_{i,k}.$$

Every time step is executed via an additive Runge-Kutta method. The solutions of the Bloch equation express the transient behavior initially, up to n damping times of the ring, where $\tau_{\text{damp}} = 1/(\varepsilon a)$ and $n \approx 5$. After a few damping times the polarization density varies much less in time. Thus we favor embedded schemes for the time evolution to allow adaptive time steps for good performance. In the next section, for simplicity we describe the evolution of the Bloch equation for a fixed time step. For a general reference to such methods see [70].

7.2 Time evolution

Additive Runge-Kutta methods are semi-implicit. Runge-Kutta methods are often summarized by a matrix $\{\alpha^{\kappa,s}\}$ and two vectors $\{\gamma_s\}$ and $\{\delta_s\}$ (these are called Butcher tableau). Additive Runge-Kutta methods are a pair of Runge-Kutta methods, one explicit (with triangular coefficient matrix) and one implicit (with full coefficient matrix), which share the same stage times δ_s and stage expansions γ_s. In our work we use ARK56, which combines an explicit (ERK) and a diagonally implicit Runge–Kutta (DIRK) scheme with $S = 8$ stages. Conventionally, the full Laplace

operator is treated implicitly, but here we only treat the second derivative term in r implicitly. The rest of the equation (including the other two terms in the Laplace operator, the Liouville terms, and the Thomas–BMT term) will be treated explicitly. The remarkable features of this approach are, first, that the matrix in the implicit part is easy to invert using any factorization, second that the matrix is block diagonal, where each block corresponds to a single Fourier mode, and that all blocks are identical. Furthermore, since the inversions of the operator are mode–independent, the implicit stage can be split between parallel processes, each handling one mode.

As input, the time evolution algorithm takes vectors from the previous timestep $U_k^\nu, V_k^\nu, W_k^\nu$ and returns the vectors for the next timestep, $U_k^{\nu+1}, V_k^{\nu+1}, W_k^{\nu+1}$. The formula to compute these vectors is

$$U_k^{\nu+1} = U_k^\nu + \Delta t \sum_{s=1}^{S} \gamma_s U_k^{(s)},\ V_k^{\nu+1} = V_k^\nu + \Delta t \sum_{s=1}^{S} \gamma_s V_k^{(s)},\ W_k^{\nu+1} = W_k^\nu + \Delta t \sum_{s=1}^{S} \gamma_s W_k^{(s)},$$

where $U_k^{(s)}, V_k^{(s)}, W_k^{(s)}$ are computed sequentially in S stages. The first stage is fully explicit. For $s = 1$ the vectors are computed by applying the discretized Bloch operator to $U_k^\nu, V_k^\nu, W_k^\nu$ using

$$\begin{aligned}
U_k^{(1)} &= \left(\frac{\varepsilon}{2} D_r^2 + C_k\right) U_k^\nu - \nu_0 V^\nu + \frac{\sigma_0}{2} A_r \left(W_{k-1}^\nu + W_{k+1}^\nu\right), \\
V_k^{(1)} &= \left(\frac{\varepsilon}{2} D_r^2 + C_k\right) V_k^\nu + \nu_0 U^\nu - \frac{\sigma_0}{2} A_r \left(W_{k-1}^\nu - W_{k+1}^\nu\right), \\
W_k^{(1)} &= \left(\frac{\varepsilon}{2} D_r^2 + C_k\right) W_k^\nu - \frac{\sigma_0}{2} A_r \left(U_{k-1}^\nu + U_{k+1}^\nu - V_{k-1}^\nu + V_{k+1}^\nu\right),
\end{aligned}$$

where

$$C_k = (\varepsilon a D_r A_r - I) - \mathbf{i} k b I + \frac{\varepsilon}{2}(1 - k^2)(A_r^{-1})^2 - \frac{3\varepsilon}{2} D_r A_r^{-1}.$$

Here D_r^1, D_r^2 are $N \times N$ Chebyshev spectral differentiation matrices of the first and second order, the matrices A_r and A_r^{-1} represent multiplication and division by r in a Chebyshev basis and I is the $N \times N$ identity matrix. The $k-1$ and $k+1$ terms incorporate the multiplication by $\sin\phi$ and $\cos\phi$ in Fourier space. Stages $2, \ldots S$ are

semi-implicit and computed by solving $3(M/2+1)$ linear systems (because Fourier modes come in conjugate pairs)

$$\begin{aligned} U_k^{(s)} = {} & \frac{\varepsilon}{2} D_r^2 \left(U_k^\nu + \Delta t \sum_{\kappa=1}^{s} \alpha^{\kappa,s} U_k^{(\kappa)} \right) + C_k \left(U_k^\nu + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} U_k^{(\kappa)} \right) \\ & - \nu_0 \left(V_k^\nu + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} V_k^{(\kappa)} \right) \\ & + \frac{\sigma_0}{2} A_r \left(W_{k-1}^\nu + W_{k+1}^\nu + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} \left(W_{k-1}^{(\kappa)} + W_{k+1}^{(\kappa)} \right) \right), \end{aligned} \tag{7.7}$$

$$\begin{aligned} V_k^{(s)} = {} & \frac{\varepsilon}{2} D_r^2 \left(V_k^\nu + \Delta t \sum_{\kappa=1}^{s} \alpha^{\kappa,s} V_k^{(\kappa)} \right) + C_k \left(V_k^\nu + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} V_k^{(\kappa)} \right) \\ & + \nu_0 \left(U_k^\nu + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} U_k^{(\kappa)} \right) \\ & - \frac{\sigma_0}{2\mathbf{i}} A_r \left(W_{k-1}^\nu - W_{k+1}^\nu + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} \left(W_{k-1}^{(\kappa)} - W_{k+1}^{(\kappa)} \right) \right), \end{aligned} \tag{7.8}$$

$$\begin{aligned} W_k^{(s)} = {} & \frac{\varepsilon}{2} D_r^2 \left(W_k^\nu + \Delta t \sum_{\kappa=1}^{s} \alpha^{\kappa,s} W_k^{(\kappa)} \right) + C_k \left(W_k^\nu + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} W_k^{(\kappa)} \right) \\ & - \frac{\sigma_0}{2} A_r \left(U_{k-1}^\nu + U_{k+1}^\nu + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} \left(U_{k-1}^{(\kappa)} + U_{k+1}^{(\kappa)} \right) \right) \\ & + \frac{\sigma_0}{2\mathbf{i}} A_r \left(V_{k-1}^\nu - V_{k+1}^\nu + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} \left(V_{k-1}^{(\kappa)} - V_{k+1}^{(\kappa)} \right) \right). \end{aligned} \tag{7.9}$$

where $\alpha^{\kappa,s}$ and $\beta^{\kappa,s}$ are coefficients from Butcher tableaux of the DIRK and ERK schemes respectively. Thus we seek an efficient method of solving the systems with matrices to the left of the form $I_N - \frac{\Delta t \varepsilon}{2} D_r^2$ and also an efficient method to impose the boundary conditions.

7.3 Integration preconditioning

The systems (7.7), (7.8) and (7.9) are analogous to systems for numerically solving the boundary value problem

$$\begin{aligned} &v - \lambda \partial_r^2 v = g, \quad r \in (-r_{\max}, r_{\max}), \\ &v(-r_{\max}) = v(r_{\max}) = 0. \end{aligned} \tag{7.10}$$

As mentioned in the Introduction, to solve (7.10) one can use so–called integration preconditioning described in [67]. Consider modal coefficients V from a spectral expansion of v in terms of Chebyshev polynomials. Similarly, G is the vector of modal coefficients for g. Let D_r^2 represent the second order differentiation operation in the Chebyshev basis, so that the equation above is approximated by

$$(I_N - \lambda D_r^2)U = G, \tag{7.11}$$

To "sparsify" the system (7.11), we multiply it by a matrix $B_{r[2]}^2$ for double integration, where the [2] means that the first two rows have been set to zero. This then gives

$$\widetilde{M}U = B_{r[2]}^2 G,$$

where $\widetilde{M} = B_{r[2]}^2 - \lambda I_{[2]}$ and it has a structure of band–width 5 with gaps. The first two rows of $\widetilde{M}$ are comprised of zeros. We then fill these two rows with Dirichlet vectors

$$\begin{aligned} M(1,:) &= [T_0(-1), T_1(-1), ..., T_{N-1}(-1)], \\ M(2,:) &= [T_0(+1), T_1(+1), ..., T_{N-1}(+1)], \\ M(3:N,:) &= \widetilde{M}(3:N,:), \end{aligned}$$

to obtain the system $MU = B_{r[2]}G$ where the first two elements of the right hand side are kept 0 since the boundary conditions are homogeneous.

Remark 17. *The matrix M is sparse except of the first two rows. So the LU decomposition of it is dense. To solve the system using only sparse LU decomposition in our implementation we use the Woodbury matrix identity*

$$M^{-1} = (M_{\mathrm{bw5}} + RS)^{-1} = M_{\mathrm{bw5}}^{-1} - M_{\mathrm{bw5}}^{-1} R (I_2 - S M_{\mathrm{bw5}}^{-1} R)^{-1} S M_{\mathrm{bw5}}^{-1}$$

where M_{bw5} is a band–width 5 part of matrix M, and RS consists of first two rows of M with some elements removed.

7.4 Higher dimensions

In three degrees of freedom the pairs of phase–space variables $(y_1, y_2), (y_3, y_4), (y_5, y_6)$ are transformed to the pairs of polar coordinates (r_α, ϕ_α), $\alpha = 1, 3, 5$. The approximations to $u = (r_1 r_3 r_5)^2 \eta$ take the form of the tensor product of Chebyshev and Fourier expansions. In six dimensions (plus time) the coefficients would be of the form $U_{i_1,i_3,i_5,k_1,k_3,k_5}$, with the three first indices representing the orders of Chebyshev polynomials in radial variables r_1, r_3, r_5, and the other three indices representing the Fourier modes.

For the reduced Bloch equation for SM3 introduced in Section 6.3 written in symmetric polar coordinates we have

$$\begin{aligned} \partial_\theta \vec{u} = \sum_{\alpha=1,3,5} \Bigg[& \varepsilon a_\alpha \left(\partial_{r_\alpha} (r_\alpha \vec{u}) - \vec{u} \right) - b_\alpha \partial_{\phi_\alpha} \vec{u} \\ & + \frac{\varepsilon}{2} \left(\partial_{r_\alpha}^2 \vec{u} + (1 + \partial_{\phi_\alpha}^2) \frac{\vec{u}}{r_\alpha^2} - 3 \partial_{r_\alpha} \frac{\vec{u}}{r_\alpha} \right) \Bigg] \\ & + \Omega(r_1, \phi_1) \vec{u}. \end{aligned}$$

Note that in SM3 the Thomas–BMT precession is considered only in the first degree of freedom. It can be easily generalized to three degrees of freedom. The time evolution is a straightforward generalization of the 1–degree–of–freedom case. For

example the s–stage equation for U becomes

$$\begin{aligned} U^{(s)}_{k_1,k_3,k_5} = \sum_{\alpha=1,3,5} \Bigg[& \frac{\varepsilon}{2} D^2_{r_\alpha} \left(U^{\nu}_{k_1,k_3,k_5} + \Delta t \sum_{\kappa=1}^{s} \alpha^{\kappa,s} U^{(\kappa)}_{k_1,k_3,k_5} \right) \\ & + C_{r_\alpha,k_\alpha} \left(U^{\nu}_{k_1,k_3,k_5} + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} U^{(\kappa)}_{k_1,k_3,k_5} \right) \Bigg] \\ & - \nu_0 \left(V^{\nu}_{k_1,k_3,k_5} + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} V^{(\kappa)}_{k_1,k_3,k_5} \right) \\ & + \frac{\sigma_0}{2} A_{r_1} \Bigg[W^{\nu}_{k_1-1,k_3,k_5} + W^{\nu}_{k_1+1,k_3,k_5} \\ & + \Delta t \sum_{\kappa=1}^{s-1} \beta^{\kappa,s} \left(W^{(\kappa)}_{k_1-1,k_3,k_5} + W^{(\kappa)}_{k_1+1,k_3,k_5} \right) \Bigg], \end{aligned} \tag{7.12}$$

where $D^2_{r_\alpha}$, A_{r_α} and C_{r_α} are expressed with Kronecker products with the identity matrices

$$\begin{aligned} & D^2_{r_1} = D^2_r \otimes I \otimes I, \quad D^2_{r_3} = I \otimes D^2_r \otimes I, \quad D^2_{r_5} = I \otimes I \otimes D^2_r, \\ & C_{r_1,k_1} = C_{k_1} \otimes I \otimes I, \quad C_{r_3,k_3} = I \otimes C_{k_3} \otimes I, \quad C_{r_5,k_3} = I \otimes I \otimes C_{k_5}, \\ & A_{r_1} = A_r \otimes I \otimes I. \end{aligned}$$

Integration preconditioning follows the same path, with both sides of the equation (7.12) multiplied by $B^2_{r_1[2]} \otimes B^2_{r_3[2]} \otimes B^2_{r_5[2]}$.

7.5 Numerical experiments

In this section we verify the accuracy of the spectral method for the Bloch equation using 3 numerical experiments. First we test the method on a problem, where the exact solution of the 1–degree–of–freedom reduced Bloch equation is known. In the second test we study the accuracy of the method applied to the 1–degree–of–freedom simple model (SM1), described in Chapter 6, where the exact solution in unknown. We evolve the reduced Bloch equation for SM1 (6.5), compute the ISF approximation

(6.7) and evolve the PDE for the error term $\Delta\vec{\eta}$ in the ISF approximation of SM1 (6.10). We estimate the error of our numerical method for SM1 as a discrepancy between the computed polarization density and its ISF approximation corrected by the solution to (6.10). In the third test we do a preliminary study of the method applied to the 3–degree–of–freedom simple model (SM3), where we evolve the 3–degree–of–freedom reduced Bloch equation and compare results to the results for SM1.

7.5.1 Analytical solution in one degree of freedom. Rates of convergence

For our first numerical experiment we consider a model obtained via the method of averaging of Chapter 5. We present this model in some detail here and refer the reader to our work on how the following Bloch equation was obtained and for discussion of our previous numerical approach [72]. For this model the reduced Bloch equation can be solved analytically. We use this model here to demonstrate the accuracy of the numerical method.

Consider the 1–degree–of–freedom effective Bloch equation that evolves the two–dimensional polarization density $\eta = (\eta_1, \eta_2)^{\mathrm{T}}$

$$\partial_\theta \eta = \varepsilon\Big(\partial_{y_1}(y_1\eta) + \partial_{y_2}(y_2\eta)\Big) + \frac{\varepsilon}{4}\nabla^2\eta - \varepsilon g y_1 J_2 \eta - \frac{\varepsilon}{2} g J_2 \partial_{y_1}\eta - \frac{\varepsilon}{4} g^2 \eta, \quad (7.13)$$

where ε is the perturbation parameter, g is a positive constant,

$$J_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

and the last 3 terms are from the Thomas–BMT precession. With the initial condition

$$\eta(0, y) = \frac{2}{\pi}\begin{pmatrix} \cos(\psi_0) \\ \sin(\psi_0) \end{pmatrix} e^{-2(y_1^2+y_2^2)}, \quad (7.14)$$

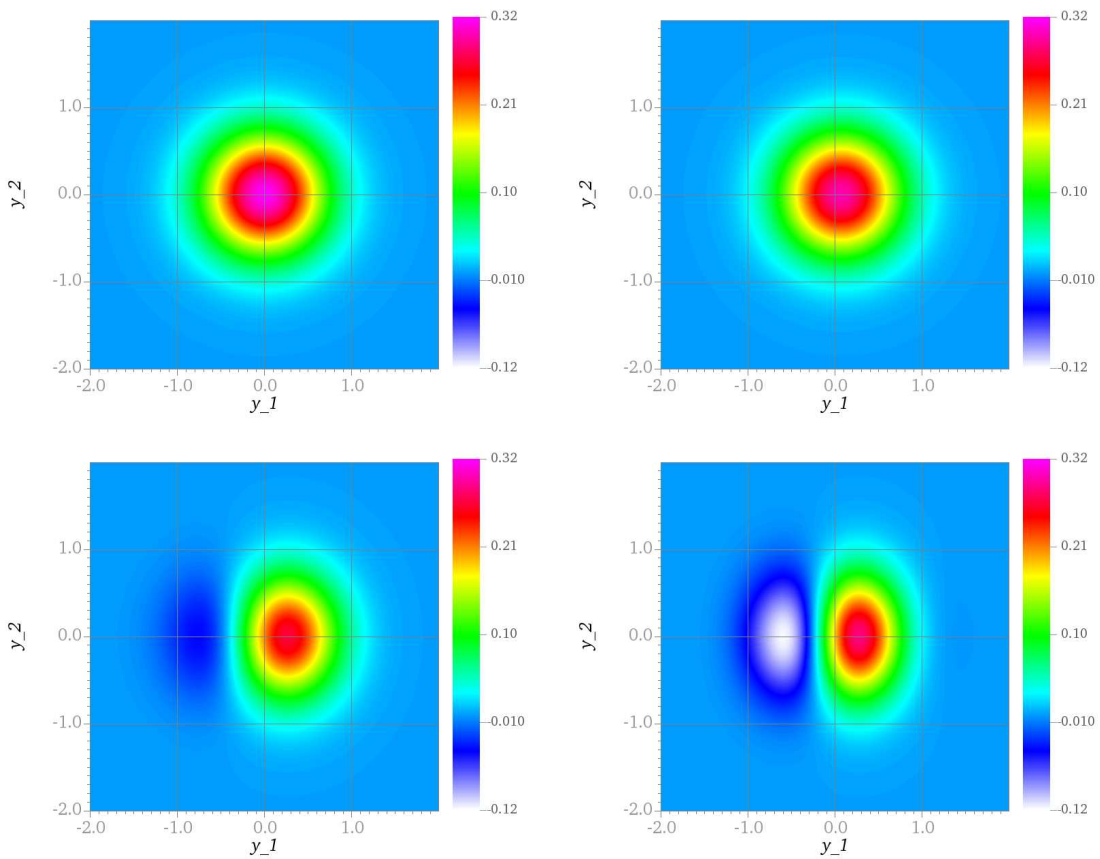

Figure 7.1: Numerical solution for the first component, η_1, of the polarization density of (7.13). The upper left subfigure displays the discretized initial condition. The upper right subfigure displays the numerical solution at time $\theta = 10$. The lower subfigures display the numerical solution at time $\theta = 100$, and $\theta = 600$. For display the numerical solution is evaluated on the fine polar grid oversampled with 100 grid points in both angular and radial dimensions.

the exact solution can be expressed as

$$\eta(\theta, y) = \frac{2}{\pi} e^{\Sigma_2(\theta)} \begin{pmatrix} \cos(\psi_0 + \Sigma_1(\theta) y_1) \\ \sin(\psi_0 + \Sigma_1(\theta) y_1) \end{pmatrix} e^{-2(y_1^2 + y_2^2)},$$

$$\Sigma_1(\theta) = -g(1 - e^{-\varepsilon\theta}), \quad \Sigma_2(\theta) = \frac{g^2}{8}(e^{-2\varepsilon\theta} - 1),$$

as can be checked by substitution in (7.13). In Figure 7.1 we display the snapshots

of the numerical solution for η_1 obtained with $\Delta t = 0.002$, $M = 32$, $N = 64$, taken at the initial time and at $\theta = 10, 100$ and 600. The transition from $\theta = 0$ to $\theta = 10$ is a slight shift of the peak to the right. In transition from $\theta = 10$ to $\theta = 100$ the polarization density starts to show more complex structure. At $\theta = 600$ the solution approaches the equilibrium given by

$$\eta_{\text{eq}}(y) = \frac{2}{\pi} e^{-\frac{g^2}{8}} \begin{pmatrix} \cos(\psi_0 - g y_1) \\ \sin(\psi_0 - g y_1) \end{pmatrix} e^{-2(y_1^2 + y_2^2)},$$

as demonstrated in the lower right subfigure in Figure 7.1.

To confirm the spectral convergence in r and ϕ we evolve (7.13) with the initial data given by (7.14). The error for the numerical solution $\theta = t_\nu$, $\nu = 0, 1, \ldots$, is

$$\epsilon(t_\nu, r, \phi) := \frac{1}{r^2} p(t_\nu, r, \phi) - \eta(t_\nu, r\cos\phi, r\sin\phi),$$

were p approximates $u = r^2\eta$ similarly to (7.6).

Remark 18. *As mentioned in Section 7.1, for the numerical method we consider the reduced Bloch equation on a bounded domain, which is a disk of radius* $r_{\max}$*, large enough so that the solution is close to* 0 *on the boundary. Thus we compute the size for the error on the bounded domain.*

Recall that η, as well as p and ϵ are vector functions with 2 components. Thus to obtain the size of the error we compute the Euclidean norm of the error evaluated on a fine polar grid oversampled with $N_0 = 100$, $M_0 = 100$ grid points in the radial and angular dimensions respectively and then take the l_2 norm of the result

$$|\epsilon|_\theta := \left(\sum_{i=1}^{N_0} \sum_{j=1}^{M_0} |\epsilon(\theta, r_i, \phi_j)|^2 \right)^{\frac{1}{2}}$$

The error relative to the size of the numerical solution solution is $|\epsilon|_\theta / |p|_\theta$. If one neglects the truncation error, the error of our spectral method can be written informally as a sum

$$|\epsilon|_\theta \sim C_r(\theta) e^{-\alpha N} + C_\phi(\theta) e^{-\beta M}$$

where C_r and C_ϕ are increasing functions of θ that illustrate the error growth over time. Here α and β are constants that are related to the regularity of the problem in the radial and angular dimensions respectively. If $N \to \infty$ and M is fixed the error is dominated by the error of the angular discretization, but when M large enough, and $C_r(\theta)e^{-\alpha N} \gg C_\phi(\theta)e^{-\beta M}$, we expect to see exponential decay of the error as a function of N. Figure 7.2 is the convergence plot. The relative errors at time $\theta = 20$

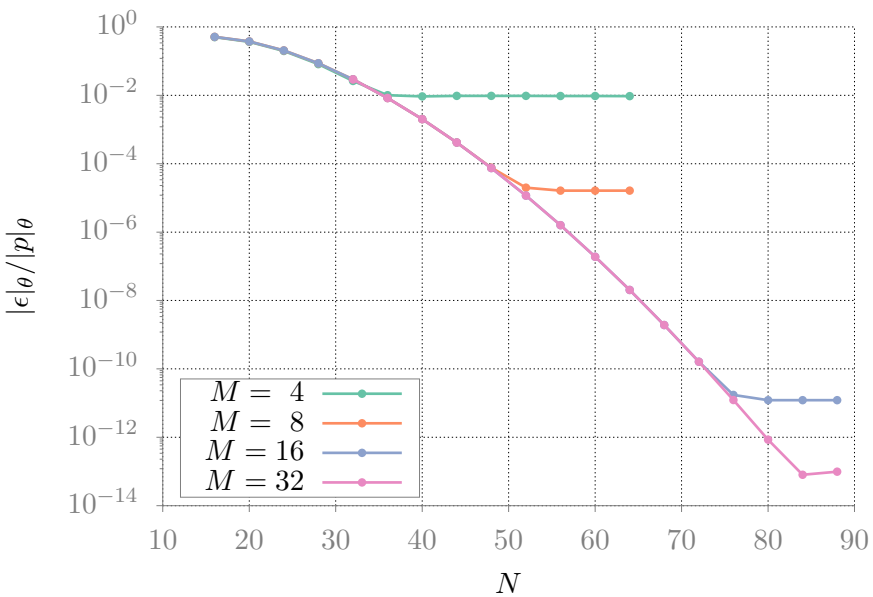

Figure 7.2: Convergence for the 1-degree-of-freedom effective Bloch equation (7.13). The error is measured at time $\theta = 20$ on the fine polar grid oversampled with 100 grid points in each dimension. Each curve corresponds to a different order of the angular discretization of the method.

as functions of N are displayed with lines. Each color corresponds to a different order of angular discretization $M = 4, 8, 16$ and 32. By looking at the green curve we see that for $M = 4$, the error in the Fourier discretization dominates the error of Chebyshev discretization when $N \geq 36$. Similarly, by looking at the orange and blue curves, we see that the Fourier discretization error dominates for $M = 8, N \geq 52$ and $M = 16, N \geq 76$. Finally by looking at the pink curve, we observe that for $M = 32$

and $N = 84$ the error of the method becomes dominated by the truncation error. As can be seen, the expected increase of order with each increment of N is observed as the slope of each curve gradually increases.

7.5.2 SM1. Rates of convergence

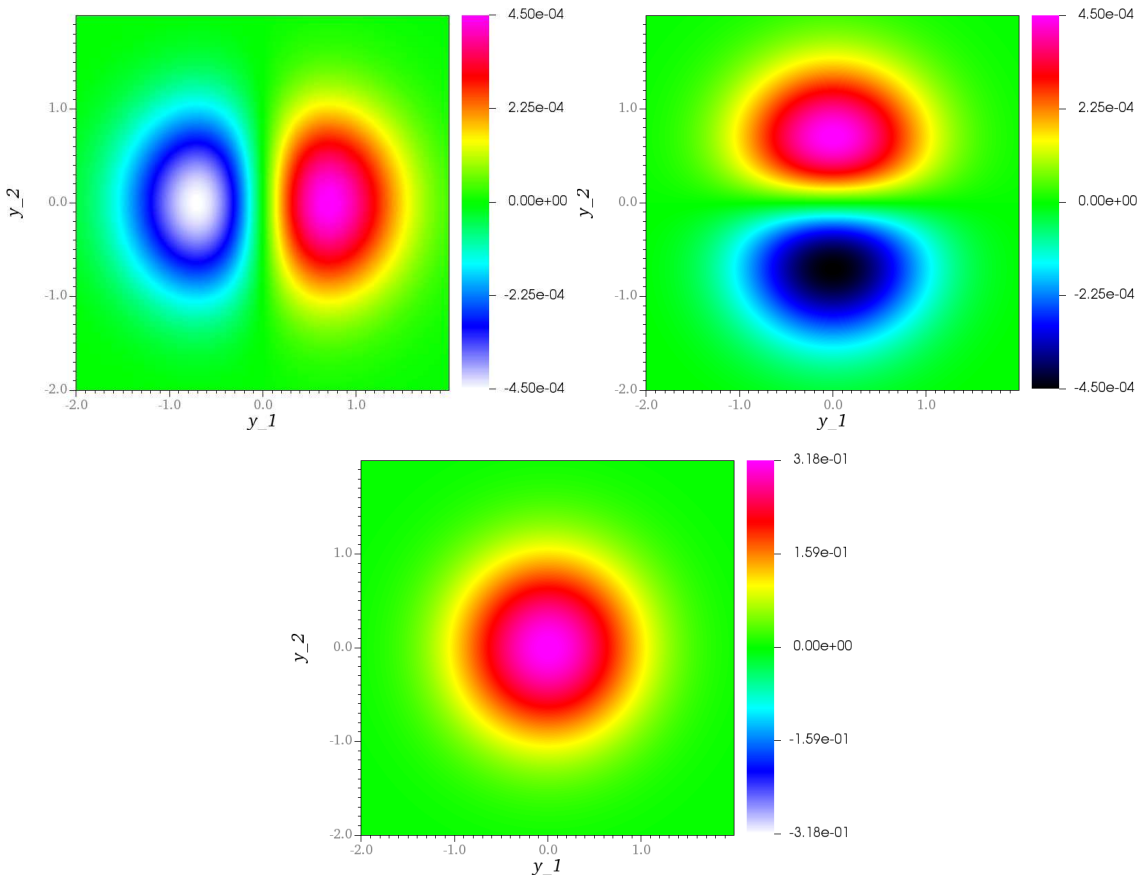

Figure 7.3: The initial value of the polarization density $\vec{\eta}$ aligned with $\hat{n}$, (7.15). The upper figures displays η_1 (left) and η_2 (right), and the lower figure displays η_3.

We now consider the reduced Bloch equation (7.1) with parameters from HERA, (see Chapter 6) and the Bloch equation for the correction term $\Delta\vec{\eta}$, (6.10), posed on

the same domain. Using the symmetric polar coordinate transformation we initialize the polarization density at $\theta = 0$ with the vector field parallel to the ISF $\hat{n}$ and proportional to the equilibrium phase–space density function $\mathcal{P}_{\rm eq}(y) = e^{-y^{\rm T}y}/\pi$ as

$$\begin{aligned}\vec{\eta}(0, r\cos\phi, r\sin\phi) &= \frac{e^{-r^2}}{\pi}\hat{n}(r\cos\phi, r\sin\phi) \\ &= \frac{e^{-r^2}}{\pi\sqrt{r^2+\zeta^2}}\begin{pmatrix} r\cos\phi \\ r\sin\phi \\ \zeta \end{pmatrix}, \quad \zeta = \frac{\nu_0 - b}{\sigma_0}. \end{aligned} \tag{7.15}$$

The initial condition is displayed in Figure 7.3. We initialize the correction term as zero

$$\Delta\vec{\eta}(0, r\cos\phi, r\sin\phi) = 0.$$

As discussed in Chapter 4 and Chapter 6, solutions to (6.10) and (7.1) satisfy

$$\vec{\eta} = \vec{\eta}_{\rm ISF} + \Delta\vec{\eta},$$

where

$$\begin{aligned}\vec{\eta}_{\rm ISF}(\theta, r\cos\phi, r\sin\phi) &= P_{\rm ISF}(\theta)\mathcal{P}_{\rm eq}(r\cos\phi, r\sin\phi)\hat{n}(r\cos\phi, r\sin\phi) \\ &= e^{-\varepsilon q\theta - r^2}\hat{n}(r\cos\phi, r\sin\phi), \quad q = \frac{1}{2}\left(\zeta^2 - 1\right)e^{\zeta^2}\mathrm{Ei}\left(-\zeta^2\right) + \frac{1}{2}.\end{aligned}$$

Hence, the numerical error for the solution at time $\theta = t_\nu$ is

$$\epsilon(t_\nu, r, \phi) := \frac{1}{r^2}\left(\vec{p}(t_\nu, r, \phi) - \Delta\vec{p}(t_\nu, r, \phi)\right) - \vec{\eta}_{\rm ISF}(t_\nu, r\cos\phi, r\sin\phi),$$

where Δp is the numerical approximation to $r^2\Delta\vec{\eta}$. In Figure 7.4 we display the relative size of the error as a function of time for the method with $N = 32, 48, 64$ and 80. $\Delta t = 0.002$ and $M = 8$ are fixed. The error grows linearly in time, indicating that the solutions are numerically stable in long–time simulations. In Figure 7.5 we display the error at time $\theta = 2\pi \times 100$ as a function of N. The plot shows that the method has a spectral convergence in the radial discretization: with each increment

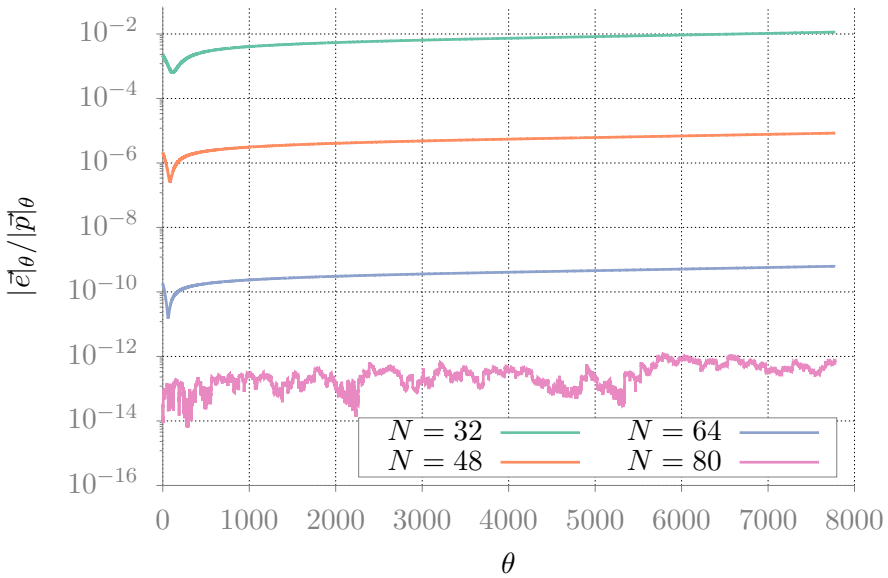

Figure 7.4: Relative error of the solution in time of SM1. The curves correspond to different orders of radial discretization. Each curve displays the relative error of the solution as a function θ. For $N = 64$, 9 significant digits of the solution are computed accurately and for $N = 80$ the 12 significant digits of the solution are obtained after two damping times of the ring ($\tau_{\text{damp}} \equiv 619 \times 2\pi \approx 3889$).

of N the slope of the error curve increases. For $N = 80$, $M = 8$ and $\Delta t = 0.002$ the error is dominated by a truncation error. Indeed, since the depolarization rate is much smaller than the orbital damping rate and since it has a small number of Fourier harmonics in the angular variable compared to the radial variable, we obtain accurate results without refining in time and angle discretization.

7.5.3 SM3. Accuracy in six space dimensions

The 3–degree–of–freedom simple model is developed in Section 6.3. Here we present the preliminary results of our numerical method for that six–dimensional problem. In this test for given parameters, we set up the initial condition for (6.13) such that the

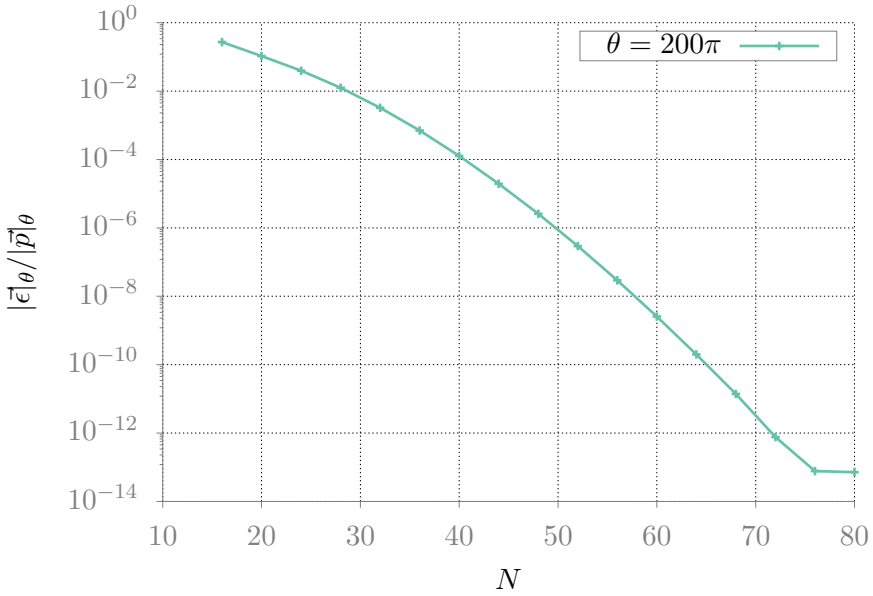

Figure 7.5: Convergence for the 1–degree–of–freedom reduced Bloch equation from SM1. The error is measured after 100 turns ($\theta = 2\pi \times 100$) on the fine polar grid oversampled with 100 grid points in each dimension. For $N > 76$ the error is dominated by the truncation error.

exact solution is comparable to the 1–degree–of–freedom case. Let $a_1 = a_3 = a_5 = 1$, $b_1 = 1$, $b_3 = b_5 = 0$ and $\varepsilon = 0.01$. Just to do the test $\sigma_0 = 0.0031415$ and $\nu_0 = 3.1415$. Consider the initial condition

$$\vec{\eta}_{\mathrm{III}}(0, r, \phi) = \hat{n}(r_1 \cos\phi_1, r_1 \sin\phi_1) \prod_{\alpha=1,3,5} \mathcal{P}_{\mathrm{eq}}(r_\alpha \cos\phi_\alpha, r_\alpha \sin\phi_\alpha),$$

where $r = (r_1, r_3, r_5)$ and $\phi = (\phi_1, \phi_3, \phi_5)$, $\hat{n}$ and $\mathcal{P}_{\mathrm{eq}}$ are defined as in the 1–degree–of–freedom case, see (7.15). Thus, since the Thomas–BMT term in (6.13) does not depend on the second and the third degree of freedom and the Fokker–Planck operator is uncoupled, the solution is stationary in the second and third degree of freedom. Further, since $\int_{\mathbb{R}^2} \mathcal{P}_{\mathrm{eq}}\, dy = 1$ we obtain that

$$\int_0^\infty \int_0^{2\pi} \int_0^\infty \int_0^{2\pi} r_3 r_5 \vec{\eta}_{\mathrm{III}}(\theta, r, \phi)\, dr_3 d\phi_3 dr_5 d\phi_5 = \vec{\eta}_{\mathrm{I}}(\theta, r_1, \phi_1)$$

where $\vec{\eta}_{\mathrm{I}}$ is the solution to the initial value problem in one degree of freedom given by (6.5) and (7.15) with parameters $a = a_1$ and $b = b_1$. We compute the size of the error by comparing the values of the polarization vector obtained in one and three degrees of freedom, namely $\vec{P}_{\mathrm{I}}(\theta) = \int_{\mathbb{R}^2} \vec{\eta}_{\mathrm{I}}\ dy_1 dy_2$ and $\vec{P}_{\mathrm{III}}(\theta) = \int_{\mathbb{R}^6} \vec{\eta}_{\mathrm{III}}\ dy_1 \dots dy_6$, approximated via spectral integration of the numerical solution. For this comparison we use the data from the simulation of the 1–degree–of-freedom reduced Bloch equation with $N = 64$, $M = 8$ and $\Delta t = 0.002$. For the 3–degree–of–freedom simulation we fix $M_1 = M_2 = M_3 = 8$ and $\Delta t = 0.002$ and vary the radial discretization sizes. We have found that for stability in 6 dimensions the timestep should be reduced to 0.0001 for the radial discretization size $N \times N \times N = 48 \times 48 \times 48$. In Figure 7.6 we display the discrepancy $|\vec{P}_{\mathrm{III}} - \vec{P}_{\mathrm{I}}|$ as a function of θ over a small interval. The increments in N lead to the rapid decrease of the error similarly to the 1–degree–of-freedom case, see Figure 7.5. This error also incorporates the error of the spectral integration of $\vec{\eta}$, which is of the same order as spatial discretization.

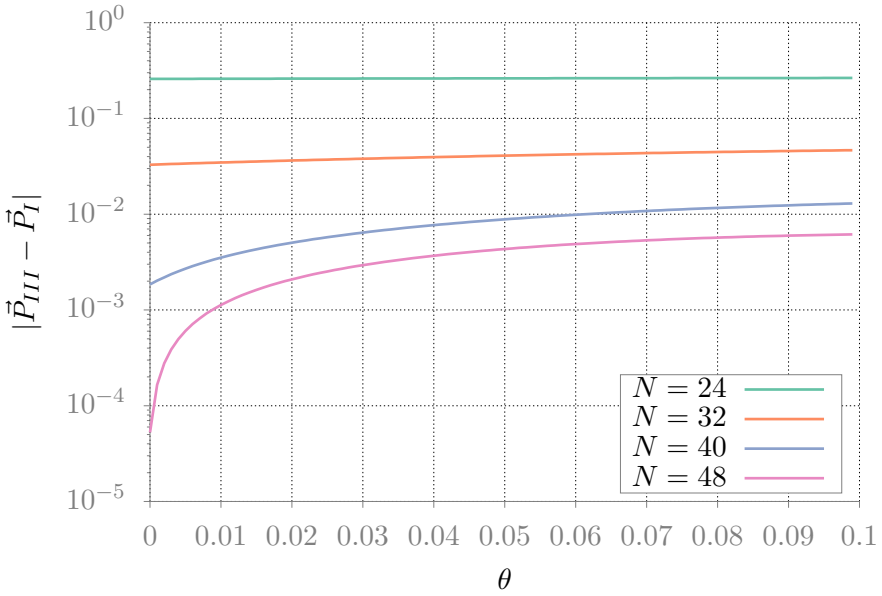

Figure 7.6: Convergence of the numerical method for SM3. The error is the difference of the polarization measurements in one degree of freedom (SM1), refined at $N = 64$ and SM3. Different curves display the error for different radial discretization sizes $(24 \times 24 \times 24, 32 \times 32 \times 32, \ldots, 48 \times 48 \times 48)$.

To conclude, the 3–degree–of–freedom code gives promising results in terms of accuracy. We will study the numerical stability of the method in six dimensions in future extensions to this work.

Chapter 8

Study of the 1–degree–of–freedom simple model

We now continue further with SM1 in numerical experiments using the ISF approximation described in Chapter 4 and the spectral method for the reduced Bloch equation (S-RBE) developed in Chapter 7. SM1 is an interesting physical model for which, as we have seen in Section 6.4, the ISF approximation gives a result in line with expectations. So here we would like to make initial further checks of this approximation in a model-independent way by means of the spectral method of Chapter 7. As explained in Chapter 4 the ISF approximation means writing the polarization density at a point in phase space as a product of an exponent that decays in time (capturing the effect of depolarization), the equilibrium phase-space density at that point and an alignment axis, namely the vector $\hat{n}$ of the ISF at that point. As we shall see, the spectral method shows that this approximation works well for SM1 when the beam is initially at equilibrium with $\vec{S}_0$ aligned with the ISF, i.e. $\vec{S}_0 = \hat{n}(Y_0)$. So that the error $\Delta\vec{\eta}$ in the ISF approximation is initially zero.

Here we perform three numerical experiments. In the first experiment we will

test the ISF approximation for SM1 close to spin-orbit resonance. In the second experiment we will look at the behavior of SM1 far from the spin-orbit resonance. In the third experiment we demonstrate the behavior of SM1 close to the "design" value of the spin tune that is used to set the parameters for HERA in Section 6.4. These three experiments compare the polarization calculated via the ISF approximation with the polarization estimated using S-RBE. In this chapter the spatial discretization parameters for S-RBE are set to $N = 80,\ M = 8$ and the timestep is set to $\Delta t = 0.002$. The results obtained in Section 7.5.2 have shown that these discretization parameters are sufficient to obtain at least 12 significant digits of the numerical solution to the RBE after a few damping times.

The three numerical experiments from this chapter are summarized in the last, fourth, experiment where we perform a spin-tune scan as in Section 6.4 using both the ISF approximation and S-RBE (for 3 values of spin tune). To estimate the equilibrium polarization in this experiment we again follow [13], but for the data obtained from S-RBE we use the depolarization time obtained by a least-squares fit to the polarization with an exponential function.

For convenience we now summarize SM1 from Section 6.1 and introduce the initial condition for this study. The SDEs and initial conditions are

$$Y' = (-bJ_2 - \varepsilon I_2)Y + \sqrt{\varepsilon}(\xi_1(\theta), \xi_2(\theta))^{\mathrm{T}}, \qquad Y(0) = Y_0 \sim \mathcal{N}\left(0, \frac{1}{2}I_2\right),$$

$$\vec{S}' = \Omega(Y)\vec{S} = \nu_0 \mathcal{J}_0 \vec{S} + \sigma_0 \sum_{j=1}^{2} \mathcal{J}_j Y_j \vec{S}, \qquad \vec{S}(0) = \hat{n}(Y_0),$$

with

$$\mathcal{J}_0 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathcal{J}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathcal{J}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix},$$

Name	Notation	Value
Damping constant	ε	$1/(619 \cdot 2\pi)$
Vertical tune	b	62.45
Spin tune	ν_0	62–63
Resonance strength	σ_0	1.65×10^{-4}

Table 8.1: HERA parameters for SM1.

and the ISF is given by

$$\hat{n}(y) = = \frac{1}{\sqrt{y^{\mathrm{T}} y + \zeta^2}} \begin{pmatrix} y_1 \\ y_2 \\ \zeta \end{pmatrix}, \quad \zeta = \frac{\nu_0 - b}{\sigma_0}.$$

The RBE for SM1 that we aim to study is

$$\begin{aligned} \partial_\theta \vec{\eta} =& b \left(\partial_{y_1}(y_2 \vec{\eta}) - \partial_{y_2}(y_1 \vec{\eta}) \right) \\ &+ \varepsilon \left(\partial_{y_1}(y_1 \vec{\eta}) + \partial_{y_2}(y_2 \vec{\eta}) \right) + \frac{\varepsilon}{2} \nabla^2 \vec{\eta} + \Omega(y) \vec{\eta} =: L_{\mathrm{Bloch}} \vec{\eta}, \\ \vec{\eta}(0, y) =& \int_{\mathbb{R}^3} \vec{s}\, \mathcal{P}_{\mathrm{eq}}(y) \delta(\vec{s} - \hat{n}(y))\, d\vec{s} = \mathcal{P}_{\mathrm{eq}}(y) \hat{n}(y), \end{aligned}$$

where $\mathcal{P}_{\mathrm{eq}}(y) = \frac{1}{\pi} e^{-y^{\mathrm{T}} y}$. Here b is the orbital tune and ν_0 is the spin tune and we expect a resonant behavior for $b \approx \nu_0$. The ISF approximation to $\vec{\eta}$ is given by

$$\vec{\eta}(\theta, y) = P_{\mathrm{ISF}}(\theta) \mathcal{P}_{\mathrm{eq}}(y) \hat{n}(y) + \Delta \vec{\eta}(\theta, y),$$

where $P_{\mathrm{ISF}}(\theta) = e^{-\varepsilon q \theta}$, $q = \frac{1}{2} \left(\zeta^2 - 1 \right) e^{\zeta^2} \mathrm{Ei}\left(-\zeta^2 \right) + \frac{1}{2}$. $\vec{\eta}$ is computed by S-RBE from Chapter 7 and thus the error of the ISF approximation is computed by $\Delta \vec{\eta} = \vec{\eta} - \vec{\eta}_{\mathrm{ISF}}$.

To observe the behavior of this model on realistic time scales we use the parameters from HERA provided in Table 8.1. The damping constant is $\varepsilon = 1/(619 \cdot 2\pi)$ so the damping time $\tau_{\mathrm{damp}} = 619$ turns.

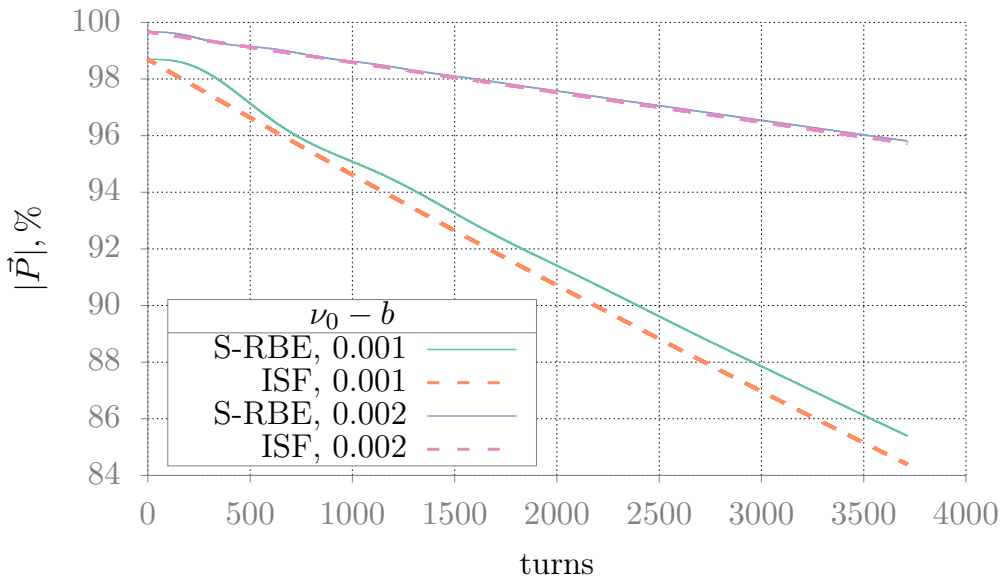

Figure 8.1: The polarization in SM1 close to resonance ($\nu_0 \approx b$). The lines display the polarization as a function of turn number $= \theta/2\pi$. The solid lines display the polarization in (8.1) computed using S-RBE. The dashed lines display the polarization in (8.2) computed via the ISF approximation.

8.1 The behavior close to the spin-orbit resonance

First we consider the case where $0 < \nu_0 - b \ll 1$. Recall from Section 4.2 that the polarization is the size (Euclidean norm) of the phase-space average of the polarization density,

$$|\vec{P}(\theta)| = \left| \int_{\mathbb{R}^2} \vec{\eta}(\theta, y)\, dy \right|. \tag{8.1}$$

In Section 4.2 we derived an estimate of the polarization via the ISF approximation

$$\left| \int_{\mathbb{R}^2} \vec{\eta}_{\mathrm{ISF}}(\theta, y)\, dy \right| = e^{-\varepsilon q \theta} \left| \int_{\mathbb{R}^2} \mathcal{P}_{\mathrm{eq}}(y) \hat{n}(y)\, dy \right| . \tag{8.2}$$

The comparison of (8.1), where $\vec{\eta}$ is computed with S-RBE, and the approximation (8.2) as functions of turn number, is displayed in Figure 8.1. The case $\nu_0 - b = 0.001$ ($\nu_0 = 62.451$) is displayed with the orange dashed line and the cyan solid line. The

case $\nu_0 - b = 0.002$ is displayed with the pink dashed line and the solid blue line. Dashed lines display the polarization estimated using the ISF approximation and solid lines show the polarization computed using S-RBE. This clearly shows a more rapid depolarization near resonance. Also the ISF approximation underestimates the polarization and is less accurate closer to resonance. Also note that the initial transient behavior of the polarization can be seen in the cyan curves for the first three damping times (3×619 turns). Similar behavior can be seen for the first 1-2 damping times on the blue curve.

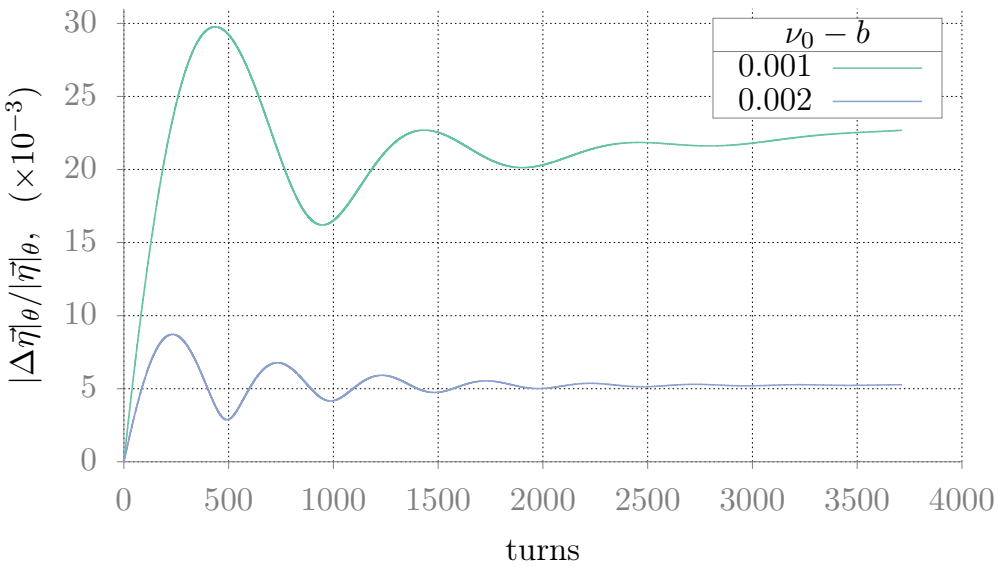

Figure 8.2: The error in the ISF approximation for SM1 close to resonance ($\nu_0 \approx b$). The solid lines display the size of $\Delta\vec{\eta}$ relative to the size of $\vec{\eta}$ for SM1 close to resonance as a function of turn number.

To study the error quantitatively we measure the size of $\Delta\vec{\eta}$ for the above cases, in a same way we compute the numerical error in Section 7.5, namely as the l_2 norm

$$|\Delta\vec{\eta}|_\theta = \left(\sum_{i=1}^{N_0} \sum_{j=1}^{M_0} |\Delta\eta(\theta, y_{i,j})|^2 \right)^{\frac{1}{2}}$$

where $\Delta\vec{\eta}$ is evaluated on the fine polar grid $\{y_{i,j}\}$ oversampled with $N_0 \times M_0$ grid points. In Figure 8.2, $|\Delta\vec{\eta}|_\theta/|\vec{\eta}|_\theta$ is displayed as a function of turn number. Clearly the relative error is significantly larger for the more resonant case. After roughly 3 orbital damping times for $\nu_0 = 62.451$, the size $\Delta\vec{\eta}$ is above 2% of the size of $\vec{\eta}$ and for $\nu_0 = 62.452$ this fraction is about 0.5%. Another important observation is that $|\Delta\vec{\eta}|_\theta/|\vec{\eta}|_\theta$ oscillates and the amplitude of the oscillations decreases. This is due to the transient behavior of the polarization mentioned above. In the second half, the simulation exposes the growth of $|\Delta\vec{\eta}|_\theta/|\eta|_\theta$ and this growth faster for the more resonant case (cyan line). One possible explanation for these phenomena is that there exists an alignment axis of $\vec{\eta}$, i.e, an ISF for the radiative problem, different from the $\hat{n}$. Perhaps $\hat{n}$ and such a radiative ISF diverge more strongly as the system comes closer to resonance. This is clearly an interesting topic for future work.

8.2 The behavior away from resonance

We now look at the behavior away from resonance. We first consider the far-from-resonance case, i.e. $\nu_0 - b = 0.3$ and 0.31 and then at the intermediate values $\nu_0 - b = 0.05$ and 0.06.

The far-from-resonance case is shown in Figure 8.3. Here we see very little depolarization. For $\nu - b = 0.31$ at 1000 turns the polarization is 99.99994% vs. 95% in close-to-resonance case ($\nu - b = 0.001$). The difference between $\nu - b = 0.31$ and 0.3 is small and the ISF approximation result is indistinguishable from S-RBE result by the eye.

Now we look at the above intermediate values of the spin tune. This is shown in Figure 8.4. Again we see the ISF approximation result is indistinguishable from the S-RBE result by the eye. The depolarization is still very small at 1000 turns in comparison to Figure 8.1 but we can see a significant change from Figure 8.3.

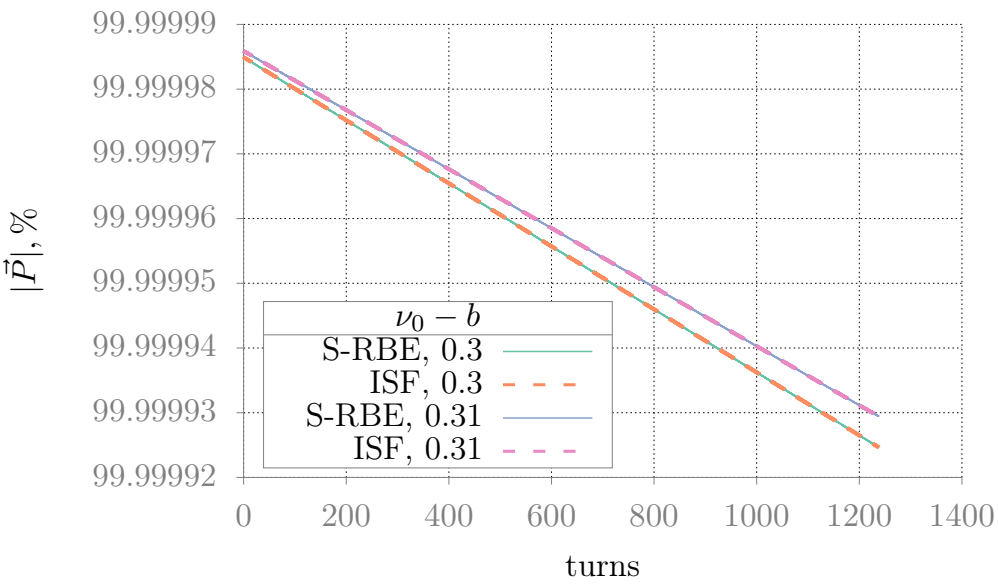

Figure 8.3: The polarization in SM1 far from resonance ($\nu_0 = 62.75,\ 62.76$). The lines display the polarization as a function of turn number $= \theta/2\pi$. The solid lines display the polarization in (8.1) computed using S-RBE. The dashed lines display the polarization computed in (8.2) via the ISF approximation.

Next we compute the size of $\Delta\vec{\eta}$ as a function of turn number for these two cases, see Figures 8.5 and 8.6. As we see, the size of $\Delta\vec{\eta}$ is $\sim 10^{-6}$ for the design value of ν_0 (Figure 8.6). Far from resonance (Figure 8.6) the size of the $\Delta\vec{\eta}$ is $\sim 10^{-7}$. As we noted at the end of Section 8.1, there is an initial transient behaviour of the polarization density, that we also see in Figures 8.5 and 8.6. But in contrast to the close-to-resonance case, the amplitude of this oscillation is miniscule, and the oscillations are not seen in Figures 8.3 and 8.4. Also note that the frequency of the oscillations of $|\Delta\vec{\eta}|_\theta/|\vec{\eta}|_\theta$ is larger the larger $\nu_0 - b$ is. Referring back to the discussion about the radiative ISF in the end of Section 8.1, these observations show that $\hat{n}$ must be close to the radiative ISF in the far-from-resonance case, and simply off by a few higher temporal harmonics.

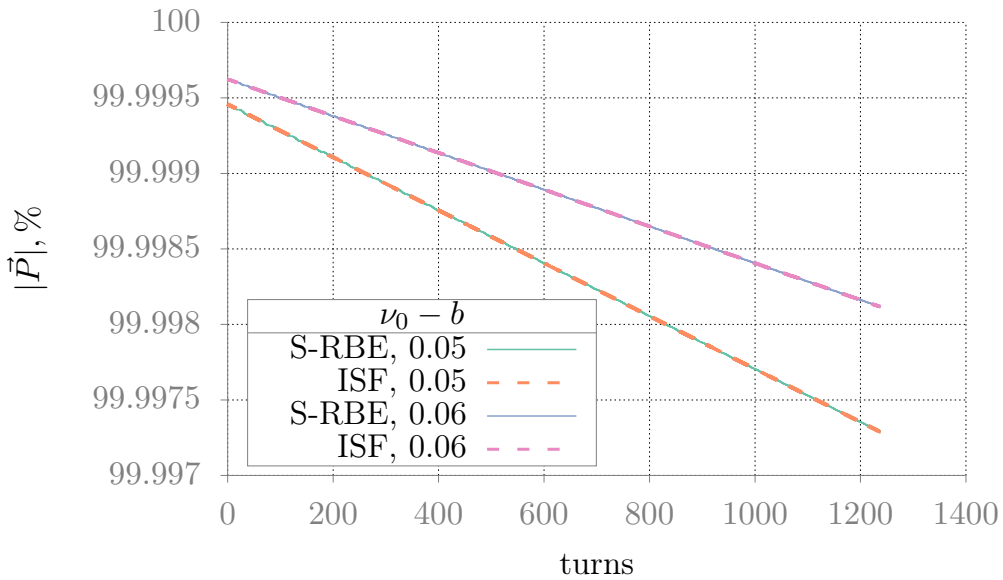

Figure 8.4: The polarization in SM1 for design values of the spin tune ($\nu_0 = 62.75,\ 62.76$). The lines display the polarization as a function of turn number $= \theta/2\pi$. The solid lines display the polarization in (8.1) computed using S-RBE . The dashed lines display the polarization in (8.2) computed via the ISF approximation.

8.3 Spin tune scan

To conclude, we perform the following spin-tune scan. To estimate the depolarization time from S-RBE we evolve $\vec{\eta}$ over 2 damping times (6 damping times for $\nu_0 = 62.451$) and record the values of the polarization in the second half of the simulation to avoid the initial transient behavior. The least-squares fit of the polarization with an exponential $\exp\left(\tau_{\text{dep}}^{-1}\frac{C}{2\pi c}\theta\right)$ then gives our estimate of the depolarization time in seconds. Then using (6.14) we estimate the equilibrium polarization in the presence of the Sokolov–Ternov effect. Figure 8.7 displays the comparison between the ISF approximation (solid line) and S-RBE results (crosses) for equilibrium polarization. We note that overall, the ISF approximation agrees well with the numerical results from the reduced Bloch equation. The detailed comparison is shown in Ta-

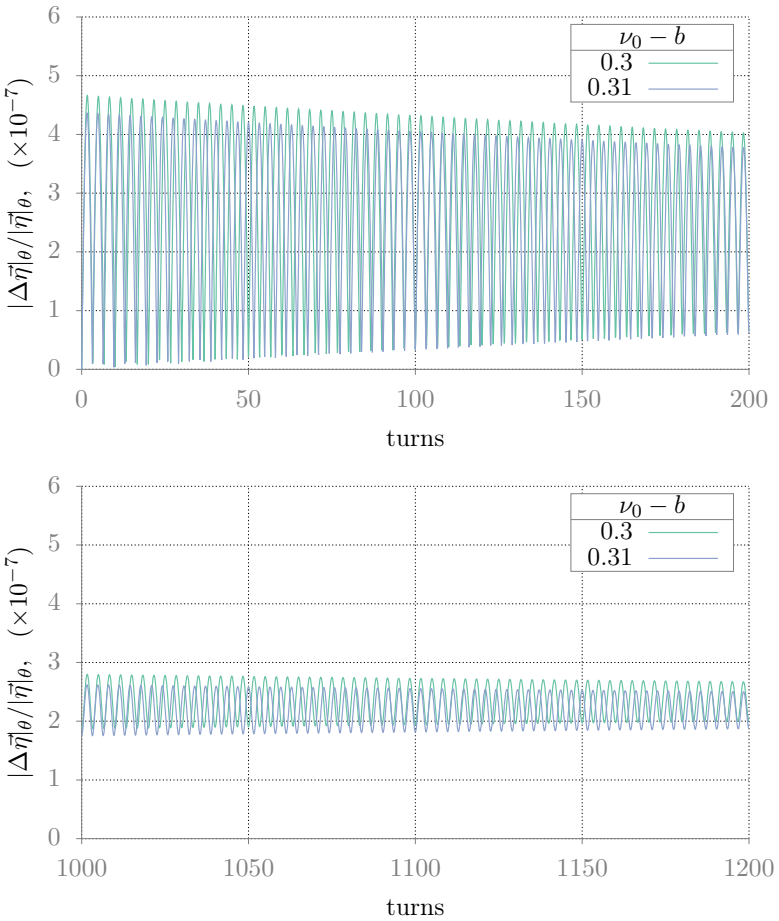

Figure 8.5: The error in the ISF approximation in SM1 far away from resonance ($\nu_0 = 62.75,\ 62.76$). The solid lines display the size of $\Delta\vec{\eta}$ relative to the size of $\vec{\eta}$ for SM1 as a function of turn number. The upper figure shows the start of the simulation, the lower figure shows the end of the simulation.

ble 8.2. The right column shows the error with respect to S-RBE of the equilibrium polarization computed using the ISF approximation. As expected, close to resonance the error (with respect to the value of polarization) is larger than far from resonance. Overall, the error of ISF approximation in SM1 is below 0.002%. This is clearly negligible in most design purposes.

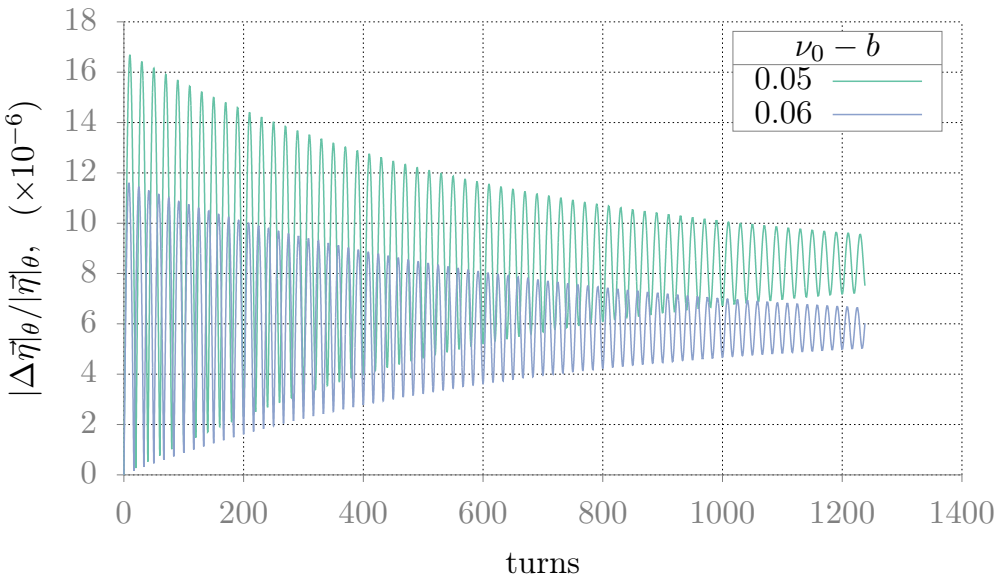

Figure 8.6: The error of the ISF approximation error in SM1 for design values of the spin tune ($\nu_0 = 62.5,\ 62.51$). The solid lines display the size of $\Delta\vec{\eta}$ relative to the size of $\vec{\eta}$ for SM1 as a function of turn number.

	ν_0	$P_{\text{eq,ISF}}$, %	$P_{\text{eq,S-RBE}}$, %	Error,%
Close	62.451	0.019205959300680	0.020371910615795	1.1660E-3
	62.452	0.074604686524295	0.075200754094522	5.9607E-4
Des.	62.50	30.79999999352302	30.79905543833404	-9.4456E-4
	62.51	38.67895792093078	38.67883861846531	-1.1930E-4
Far	62.75	87.53676919842549	87.53677652176667	7.3233E-6
	62.76	87.83019819790047	87.83019768696403	-5.1094E-7

Table 8.2: Comparison of the ISF approximation and S-RBE. The results close to spin-orbit resonance are displayed in the upper block. The results for design values of ν_0 are displayed in the middle block. The results far from resonance are displayed in the lower block. The last column displays the error with respect to S-RBE of the equilibrium polarization computed using the ISF approximation.

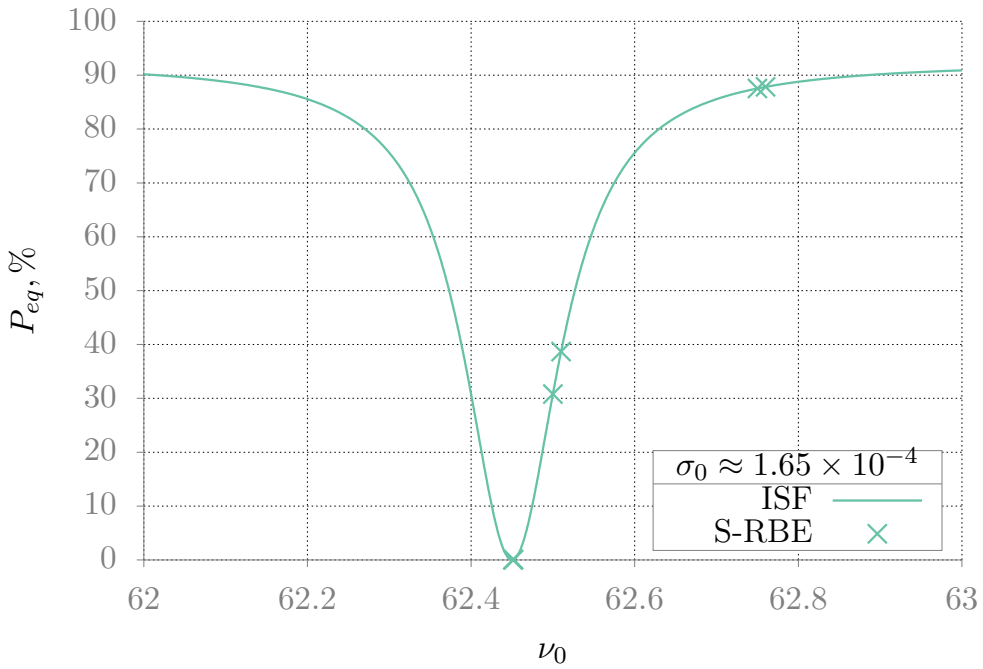

Figure 8.7: The spin tune scan for the simple model with parameters from HERA. The solid curve displays the equilibrium polarization estimated using the ISF approximation. The points correspond to the equilibrium polarization estimated using the spectral method for the reduced Bloch equation.

Chapter 9

Summary and future work

In Chapters 2-8 we have built a framework for studying spin polarization in modern electron storage rings based on stochastic differential equations. In Chapter 2 we presented the lab-frame equations starting from the SDE (2.1)-(2.4) for position and momentum and the SDE (2.8) for the spin vector $\vec{S}$. These 9 random processes characterize a bunch of electrons (or positrons) in a storage ring. The joint probability density function of these random processes obeys the Fokker-Planck equation (2.11) that leads to the Bloch equation (BE) (2.15) for our main quantity of interest, the polarization density $\vec{\eta}$. Without the Sokolov-Ternov self-polarization effect and without the kinetic polarization effect, (2.15) is reduced to (2.16) that we refer to as the reduced Bloch equation (RBE).

In Chapter 3 we introduced the beam-frame reduced linearized SDEs for the orbit variable, Y, and their spin variable $\vec{S}$, (3.5) and (3.6). The orbital SDE (3.5) is a narrow-sense linear SDE, while the SDE (3.6), describing spin, is nonlinear. These two equations were our main focus throughout the work. Following the same outline as in Chapter 2 we then derived the PDE for the joint density of Y and $\vec{S}$, (3.8), and the beam-frame RBE (3.12) for the polarization density. Next, in Section 3.4,

we focused on the orbital dynamics and obtained the equilibrium solution for the mean and covariance of Y resulting in a Gaussian equilibrium phase-space density. Further, in Section 3.5 we discussed the equilibrium dynamics for the non-radiative problem that is of importance for the non-radiative particle dynamics and serves as an introduction to the ISF approximation to the polarization density described in Chapter 4.

The beam-frame RBE is a system of three Fokker-Planck equations with time dependent coefficients, coupled by the Thomas-BMT term. It is important to note that, when posed in six-dimensional phase space, the RBE is an extremely challenging problem with respect to the anticipated computational and memory cost of the numerical algorithms. This in fact motivated two different techniques of approximation that we presented in Chapters 4 and 5.

In Chapter 4 we formalized the ISF approximation for the polarization density. The ISF approximation (4.1) considers the synchrotron radiation as a perturbation to the problem presented in Section 3.5. The ISF approximation represents $\vec{\eta}_Y$ as a product of an exponentially decaying function P_{ISF}, an equilibrium phase-space density for the radiative problem, and the invariant spin field $\hat{n}$, which is the normalized, unique periodic solution of the non-radiative RBE. We obtained P_{ISF} in the ISF approximation by writing the equation for the residual of the RBE and constraining the residual to be on average orthogonal to the ISF. Plugging the ISF approximation to the RBE with radiation led to a very simple ODE for P_{ISF} (4.8), and P_{ISF} as anticipated was a decaying exponential. The PDE for the error of the ISF approximation denoted as $\Delta\vec{\eta}$ is an RBE with an inhomogeneous force term. The inhomogeneous term is by construction orthogonal to the ISF on average and therefore the ISF approximation is expected to be accurate when the directions of $\vec{\eta}$ and $\hat{n}$ are close. This was also verified in Chapter 8 using the model presented in Chapter 6. In Section 4.2 we obtained a formula from the ISF approximation for the so-called

depolarization time, a well known quantity that can even be used to compute the equilibrium polarization in the presence of the Sokolov-Ternov effect. This quantity can be found in most technical reports which discuss optimizing lattice design for obtaining the maximum attainable polarization in particle accelerators and storage rings. The main strength of the ISF approximation is that it avoids solving the RBE entirely, but as a trade-off it relies on the ISF to be known a priori and that is not a trivial quantity to compute. There are algorithms for computing the ISF, described for example in [42, 43, 44, 45]. So an application of the ISF approximation to a realistic lattices is possible, moreover the ISF-based approaches have been widely used in the polarization studies of modern spin-polarized electron storage rings.

In Chapter 5 we presented the method of averaging. Also this method considers radiation as a perturbation but, in contrast to the ISF approximation, provides a new PDE (5.18) which we call the effective RBE. The method of averaging is first applied to the system of ODEs for the mean and covariance of Y transformed into slowly varying forms (5.5) and (5.6). The resulting averaged system (5.9) and (5.10) for the mean and covariance then governs the averaged process V satisfying the SDE (5.11). The joint probability density for processes V and $\vec{T}$, where $\vec{T}$ denotes the spin vector after the transformation, satisfies the Fokker-Planck equation (5.16). This then leads to the effective RBE (5.18). Compared to the RBE (3.12) the effective RBE (5.18) has time independent coefficients in the Fokker-Planck operator and thus is adaptable to numerical analysis as discussed in Chapter 7. It is worthwhile to mention that for realistic lattices only the MOA may provide effective models for which the Bloch equation can be solved numerically in reasonable time.

In Chapter 6 we presented two simple models, SM1 and SM3. In these models the RBE has θ-independent coefficients, and thus is very similar to the effective RBE provided by the MOA. Also, for SM1 the ISF is known. Hence SM1 was useful for testing our numerical method in Chapter 7 to evaluate its applicability to an

effective RBE and also to verify the ISF approximation in Chapter 8. SM3 was developed for testing the 6-dimensional numerical algorithm in the end of Chapter 7. Our main model, SM1, has shown physically meaningful results in Section 6.4 with parameters from HERA where we performed a spin tune scan experiment using the ISF approximation to explore the resonance structure of SM1. This model shows massive polarization loss close to a spin-orbit resonance.

In Chapter 7 we developed the numerical techniques, applicable to an RBE with a time independent radial diffusion operator in up to 3 degrees of freedom. With minor modifications this method should also be applicable to the (full) BE. For realistic lattices our techniques are certainly applicable to the effective RBE. Our numerical method is a Fourier-Chebyshev method for the RBE (6.13) posed on a truncated six dimensional phase space. Each degree of freedom is transformed into a polar coordinate pair. The method approximates the solution with tensor-product Chebyshev and Fourier basis functions in the radial and angle variables respectively. For the time evolution we use a high order additive Runge-Kutta method, [70]. The time evolution algorithm computes spectral coefficients of the numerical solution at each time step by solving linear systems. The second order differentiation matrix for the Fourier-based method is diagonal, so, considering the RBE, the computational work for the time evolution can be efficiently distributed between parallel processes. The parallel algorithm requires only local communication. Sparsity of the time-evolution linear systems is increased by using integration preconditioning corresponding to the radial directions to further reduce the computational cost. We demonstrated the spectral accuracy of the method and its numerical stability in numerical experiments of Section 7.5. We tested the method for solving the effective RBE (7.13) where the exact solution is known and also for SM1, where the exact solution is unknown. SM3 was used to test the accuracy of the method in three degrees of freedom.

In Chapter 8 we extensively used our framework to study the SM1. For this

model the ISF approximation gave a high quality result even close to the spin orbit resonance but, as we found from the RBE, the polarization density appeared to be not exactly aligned with the ISF. This led to the speculation that there is an axis of polarization density alignment for a radiative problem which is different from the ISF. Note that such a "radiative" ISF was used in [73, 74]. The (full) BE may lead to a different answer. In the presence of the Sokolov-Ternov polarizing effect, the polarization density is non-zero in the long term and it approaches a non-zero equilibrium which may have another equilibrium alignment vector field. Ultimately this study may lead to a novel extension of the ISF approximation for the polarization density, but currently it is only a hypothesis that we find very promising and we are eager to look into this matter in the future.

The developments in every chapter will be extended in our future work. As we mentioned in a few remarks in Chapters 2 and 3, the SDEs with Sokolov-Ternov and kinetic polarization effects could find applications in state-of-the-art spin tracking programs like those in Bmad and PTC. Furthermore the full system of SDEs (3.1), (3.2) (that include the Sokolov-Ternov effect and kinetic polarization) will be studied in the further advancements of this work. Here the goal is to find the equilibrium polarization if it exists.

In Section 5.2 we outlined two main future advances for the ISF approximation and the MOA approximation. In Section 5.2 we have shown how to use the stochastic averaging for the system (5.12) and (5.13) for including spin in the MOA. For example, with this averaging one can analytically describe the resonance structure of a model, like SM1 or SM3. A generalization of the ISF approximation would be to expand the polarization density w.r.t. to an orthonormal basis as described at the end of Section 5.2. In this expansion the ISF approximation is the leading term and the other terms are obtained from the eigenfunctions of the Fokker-Planck operator via Gram-Schmidt orthonormalization. This method relies on the MOA, using the

θ-independence of the averaged Fokker-Planck operator. Our findings in Chapter 8 led to a speculation about the existence of the radiative ISF for SM1. In future work this will be considered for formalizing an alternative ISF approximation for the polarization density, based on the radiative ISF instead of $\hat{n}$.

One possible extension of the spectral method for the Bloch equation is the implicit treatment of the full Laplace operator in the time evolution scheme, that may result in better stability properties. The current method may then be used as a preconditioner to speed up the inversions of the discretized Laplacian with an iterative method.

Finally, SM1 is an interesting physical model for which, as we have seen in Chapter 4 and Chapter 8, the ISF approximation gives a result in line with expectations but we intend to invent further checks. As a follow-up project we would like to treat σ_0 as the stochastic variable and keep Y as a unit-length rotating vector in the same way as it is used for protons in (6.1). This would be closer to the spirit of the physics outlined in Section 6.1. We assume that this would lead to the same conclusions. As another future extension we would like to develop a version of SM1 where the spin tune oscillates around the design value with the frequency of synchrotron motion and thus generates the so-called synchrotron sideband resonances. All the numerical techniques developed in this work can be easily applied to such a model.

Chapter 10

Hermite-Discontinuous Galerkin Overset Grid Methods for the Scalar Wave Equation

Accurate and efficient simulation of waves is important in many areas in science and engineering due to the ability of waves to carry information over large distances. This ability stems from the fact that waves do not change shape in free space. On the other hand when the background medium is changing this induces a change in the wave forms that propagate through the medium and the waves can be used for probing the interior material properties of objects.

In order to preserve the properties of waves from the continuous setting it is preferable to use high order accurate discretizations that are able to control dispersive errors. The development of high order methods for wave propagation problems has been an active area of research for a long time and there are by now many attractive methods. Examples include (but are not limited to) finite difference methods, [75, 76, 77, 78, 79], embedded boundary finite differences, [80, 81, 82, 83, 84], element

based methods like discontinuous Galerkin (DG) methods, [85, 86, 87, 88, 89, 90], hybridized discontinuous Galerkin (HDG) methods, [91, 92], cut-cell finite elements [93, 94] and Galerkin-difference methods [95].

An advantage of summation-by-parts finite differences and Galerkin type methods is that stability is guaranteed, however this guarantee also comes with some drawbacks. For diagonal norm summation-by-parts finite differences the order of accuracy is reduced to roughly half of that in the interior near boundaries. Further the need for multi-block grids also restricts the geometrical flexibility.

As DG and HDG methods are naturally formulated on unstructured grids they have good geometric flexibility. However, Galerkin based polynomial methods often have the drawback that they require small timesteps (the difference Galerkin and cut-cell finite element methods are less affected by this) when combined with explicit timestepping methods, but on the other hand they preserve high order accuracy all the way up to the boundary and it is easy to implement boundary conditions independent of the order of the method.

The pioneering work by Henshaw and co-authors, see for example [96], describes techniques for generating overset grids as well as how they can be used to solve elliptic and first order time-dependent partial differential equations (PDE) by second order accurate finite differences. In an overset grid method the geometry is discretized by narrow body-fitted curvilinear grids while the volume is discretized on one or more Cartesian grids. The generation of such body-fitted grids is local and typically produces grids of very high quality, [97]. The grids overlap (we say that they are overset) so that the solution on an interior (often referred to as non-physical or ghost) boundary can be transferred from the interior of another grid. In [96] and in most other overset grid methods the transfer of solutions between grids is done by interpolation. Since the bulk of the domain can be discretized on a Cartesian grid the efficiency asymptotically approaches that of a Cartesian solver but still retains

the geometrical flexibility of an unstructured grid method.

We note that the same type of efficiency can be expected for embedded boundary and cut-cell finite elements. A difference is that overset grid methods typically have smoother errors near physical boundaries and this may be important if quantities that include derivatives of the solution, such as traction or strain, are needed.

Here we are concerned with the approximation of the scalar wave equation on overset grids. To our knowledge, high order overset grid methods for wave equations in second order form have been restricted to finite difference discretizations. For example, in [98] high order centered finite difference approximations to Maxwell's equations (written as a system of second order wave equations) was introduced. More recently, in [99], the upwind discretizations by Banks and Henshaw introduced in [100] were generalized to overset grids. In [79] convergence at 11th order for a finite difference method is demonstrated. A second order accurate overset grid method for elastic waves can be found in [101].

We use the recently introduced dissipative Hermite methods for the scalar wave equation in second order form, [102], for the approximation on Cartesian grids. To handle geometry we use the energy based DG methods of [86] on thin grids that are grown out from physical boundaries. We use projection to transfer the solutions between grids rather than interpolation.

Both the Hermite and DG methods we employ increase the order of accuracy by increasing the number of degrees of freedom on an element or cell. This has practical implications for grid generation as a single grid with minimal overlap can be used independent of order, reducing the complexity of the grid generation step. This can be important for example in problems like optimal shape design, where the boundary changes throughout the optimization. This is different from the finite difference methods where, due to the wider finite difference stencils, the overlap must

grow as the order is increased.

The transfer of solutions between overset grids typically causes a perturbation to the discrete operators which, especially for hyperbolic problems, results in instabilities, see [101] for example. These instabilities are often weak and can thus be suppressed by a small amount of artificial dissipation. There are two drawbacks of this added dissipation, first it is often not easy to determine the suitable amount needed, i.e. big enough to suppress instabilities but small enough not to reduce the accuracy or timestep too severely. Second, in certain cases the instabilities are strong enough that the dissipation must scale with the discretization parameter (the grid size) in such a way that the order of accuracy of the overall method is reduced by one.

Similar to [99], we use a dissipative method that has naturally built-in damping that is sufficient to suppress the weak instabilities caused by the overset grids. The order of the hybrid overset grid method is the design order of the Hermite method or DG method, whichever is the smallest.

In the hybrid H-DG overset grid method the Hermite method is used on a Cartesian grid in the interior of the domain, and the discontinuous Galerkin method on another, curvilinear grid at the boundary. The numerical solution is evolved independently on these grids for one timestep of the Hermite method. By using the Hermite method in the interior the strict timestep constraints of the DG method are relaxed by a factor that grows with the order of the method. Asymptotically, as discussed above, the complexity of the hybrid H-DG solver approaches that of the Cartesian Hermite solver [102].

The paper is organized as follows. The Hermite method is described in the next section. We first explain the method in a simple one dimensional case and then explain how the method generalizes to two dimensions. The DG method is described

in Section 10.2. The details of the overset grids and a hybridization of the DG and the Hermite methods are described in Section 10.3. We illustrate the hybrid H-DG method with numerical simulations in the Section 10.4.

10.1 Dissipative Hermite method for the scalar wave equation

We present the Hermite method in some detail here and refer the reader to the original work [102] for convergence analysis and error estimates.

Consider the one dimensional wave equation in second order form in space and first order in time

$$\begin{aligned} u_t &= v, \\ v_t &= c^2 u_{xx} + f, \quad x \in \Omega, \quad t \in (0,T). \end{aligned}$$

Here $u \in C^{2m+3}(\Omega \times [0,T])$, $v \in C^{2m+1}(\Omega \times [0,T])$ and $f \in C^{2m+1}(\Omega \times [0,T])$ for optimal convergence. We refer to u as the displacement, and v as the velocity. The speed of sound is c. We consider boundary conditions of Dirichlet or Neumann type

$$\begin{aligned} u(t,x) &= h_0(t,x), \quad x \in \partial\Omega_D, \\ u_x(t,x) &= h_1(t,x), \quad x \in \partial\Omega_N, \end{aligned}$$

and initial conditions

$$\begin{aligned} u(0,x) &= g_0(x), \\ v(0,x) &= g_1(x). \end{aligned}$$

Let the spatial domain be $\Omega = [a,b]$. The domain will be discretized by a primal grid

$$x_i = a + ih, \quad h = (b-a)/N, \quad i = 0, \ldots, N,$$

and a dual grid

$$x_i = a + ih, \;\; i = \frac{1}{2}, \ldots, N - \frac{1}{2}.$$

The use of staggered grids allow us to evaluate the derivatives of the polynomial approximations to derivatives at the cell center, rather than throughout the cell as in most other element based methods. The slow growth with polynomial degree of the derivative approximations near the cell centers (see [103]) allows us to use timesteps that are bounded by the speed of sound and not by the degree of the polynomial. In time we discretize using a uniform grid with increments $\Delta t/2$, that is

$$t_n = n\Delta t, \;\; n = 0, 1/2, 1, \ldots.$$

At each grid point x_i the approximation to the solution is represented by its degrees of freedom (DOF) that approximate the values and spatial derivatives of u and v. Equivalently, the approximations to u and v can be represented as polynomials centered at grid points x_i. The Taylor coefficients of these polynomials are scaled versions of the degrees of freedom. To achieve the optimal order of accuracy $(2m+1)$ we require the $(m+1)$ and m first derivatives of u and v respectively to be stored at each grid point.

At the initial time (which we take to be $t = 0$) these polynomials are approximations to the initial condition on the primal grid

$$u(x,0) \approx \sum_{l=0}^{m+1} \hat{u}_l \left(\frac{x - x_i}{h} \right)^l \equiv p_i(x), \quad i = 0, \ldots, N,$$
$$v(x,0) \approx \sum_{l=0}^{m} \hat{v}_l \left(\frac{x - x_i}{h} \right)^l \equiv q_i(x), \quad i = 0, \ldots, N.$$

The coefficients $\hat{u}_l$ and $\hat{v}_l$ are assumed to be accurate approximations to the scaled Taylor coefficients of the initial data. If expressions for the derivatives of the initial data are known we simply set

$$\hat{u}_l = \frac{h^l}{l!} \frac{d^l g_0}{dx^l}\bigg|_{x=x_i}, \quad \hat{v}_l = \frac{h^l}{l!} \frac{d^l g_1}{dx^l}\bigg|_{x=x_i}.$$

Alternatively, if only the functions g_0 and g_1 are known, we may use a projection or interpolation procedure to find the coefficients in (10.1).

The numerical algorithm for a single timestep consists of two phases, an interpolation step and an evolution step. First, during the interpolation phase the spatial piecewise polynomials are constructed to approximate the solution at the current time. Then, in the time evolution phase we use the spatial derivatives of the interpolation polynomials to compute time derivatives of the solution using the PDE. We compute new values of the DOF on the next time level by evaluating the obtained Taylor series. We now describe each step separately.

10.1.1 Hermite interpolation

At the beginning of a timestep at time t_n (or at the initial time) we consider a cell $[x_i, x_{i+1}]$ and construct the unique local Hermite interpolant of degree $(2m+3)$ for the displacement and degree $(2m+1)$ for the velocity. The interpolating polynomials are centered at the dual grid points $x_{i+\frac{1}{2}}$ and can be written in Taylor form

$$\begin{aligned} p_{i+\frac{1}{2}}(x) &= \sum_{l=0}^{2m+3} \hat{u}_{l,0} \left(\frac{x - x_{i+\frac{1}{2}}}{h} \right)^l, \quad x \in [x_i, x_{i+1}], \quad i = 0, \ldots, N-1, \\ q_{i+\frac{1}{2}}(x) &= \sum_{l=0}^{2m+1} \hat{v}_{l,0} \left(\frac{x - x_{i+\frac{1}{2}}}{h} \right)^l, \quad x \in [x_i, x_{i+1}], \quad i = 0, \ldots, N-1. \end{aligned}$$

The interpolants $p_{i+\frac{1}{2}}$ and $q_{i+\frac{1}{2}}$ are determined by the local interpolation conditions:

$$\begin{aligned} \frac{d^l p_{i+\frac{1}{2}}}{dx^l} &= \left. \frac{d^l p_i}{dx^l} \right|_{x=x_i}, \quad \frac{d^l p_{i+\frac{1}{2}}}{dx^l} = \left. \frac{d^l p_{i+1}}{dx^l} \right|_{x=x_{i+1}}, \quad l = 0, \ldots, m+1, \\ \frac{d^l q_{i+\frac{1}{2}}}{dx^l} &= \left. \frac{d^l q_i}{dx^l} \right|_{x=x_i}, \quad \frac{d^l q_{i+\frac{1}{2}}}{dx^l} = \left. \frac{d^l q_{i+1}}{dx^l} \right|_{x=x_{i+1}}, \quad l = 0, \ldots, m. \end{aligned}$$

We find the coefficients in (10.3) and (10.4) by forming a generalized Newton table as described in [104].

10.1.2 Time evolution

To evolve the solution in time we further expand the coefficients of $p_{i+\frac{1}{2}}$ and $q_{i+\frac{1}{2}}$. At each point on the dual grid, $x_{i+\frac{1}{2}}$ we seek temporal Taylor series

$$\begin{aligned} p_{i+\frac{1}{2}}(x,t) &= \sum_{l=0}^{2m+3}\sum_{s=0}^{\kappa_p} \hat{u}_{l,s}\left(\frac{x-x_{i+\frac{1}{2}}}{h}\right)^l\left(\frac{t}{\Delta t}\right)^s, \\ q_{i+\frac{1}{2}}(x,t) &= \sum_{l=0}^{2m+1}\sum_{s=0}^{\kappa_q} \hat{v}_{l,s}\left(\frac{x-x_{i+\frac{1}{2}}}{h}\right)^l\left(\frac{t}{\Delta t}\right)^s, \end{aligned}$$

where $\kappa_p = (2m+3-2\lceil\frac{l}{2}\rceil)$ and $\kappa_q = (2m+1-2\lfloor\frac{l}{2}\rfloor)$. The coefficients $\hat{u}_{l,0}$ and $\hat{v}_{l,0}$ are given by the coefficients of (10.3) and (10.4). At this time the scaled time derivatives, $\hat{u}_{l,s}$ and $\hat{v}_{l,s}$ $s>0$, are unknown and must be determined. Once they are determined we may simply evaluate (10.5) and (10.6) at $t = t_n + \Delta t/2$ to find the solution at the next half timestep.

In Hermite methods the coefficients of temporal Taylor polynomials are determined by collocating the differential equation, [105, 102, 104]. In particular, by differentiating (10.1) and (10.2) in space and time the time derivatives of the solution can be directly expressed in terms of spatial derivatives

$$\begin{aligned} \frac{\partial^{s+1+r}u}{\partial t^{s+1}\partial x^r} &= \frac{\partial^{s+r}v}{\partial t^s \partial x^r}, \\ \frac{\partial^{s+1+r}v}{\partial t^{s+1}\partial x^r} &= c^2\frac{\partial^{s+r+2}u}{\partial t^s\partial x^{r+2}} + \frac{\partial^{s+r}f}{\partial t^s\partial x^r}. \end{aligned}$$

Substituting (10.5) and (10.6) into (10.7) and (10.8) and evaluating at $x = x_{i+\frac{1}{2}}$ and $t = t_n$, we can match the powers of the coefficients to find the recursion relations

$$\begin{aligned} \hat{u}_{l,s+1} &= \frac{\Delta t}{s}\hat{v}_{l,s}, \\ \hat{v}_{l,s+1} &= c^2\frac{(l+1)(l+2)}{h^2}\frac{\Delta t}{s}\hat{u}_{l+2,s} + \frac{\Delta t}{s}\hat{f}_{l,s}. \end{aligned}$$

Here $\hat{f}_{l,s}$ are the coefficients of the Taylor expansion of f, or of the polynomial which interpolates $f(t, x_{i+1/2})$ in time around $t = t_n$. Note that since there are a

finite number of coefficients, representing the spatial derivatives at the time t_n, the recursions truncate and only κ_p and κ_q terms need to be considered.

To complete a half timestep we evaluate the approximation at $t = t_n + \frac{\Delta t}{2}$ for the $(m+1)$ and m first derivatives

$$\frac{\partial^l u}{\partial x^l}(x_{i+\frac{1}{2}}, t_{n+\frac{1}{2}}) \approx \frac{\partial^l p_{i+\frac{1}{2}}}{\partial x^l}(x_{i+\frac{1}{2}}, t_{n+\frac{1}{2}}) = \frac{l!}{h^l} \sum_{s=0}^{\kappa_p} \frac{\hat{u}_{l,s}}{2^s}, \quad l = 0, \ldots, m+1,$$
$$\frac{\partial^l v}{\partial x^l}(x_{i+\frac{1}{2}}, t_{n+\frac{1}{2}}) \approx \frac{\partial^l q_{i+\frac{1}{2}}}{\partial x^l}(x_{i+\frac{1}{2}}, t_{n+\frac{1}{2}}) = \frac{l!}{h^l} \sum_{s=0}^{\kappa_q} \frac{\hat{v}_{l,s}}{2^s}, \quad l = 0, \ldots, m.$$

Remark 19. *A remarkable feature of Hermite methods is that (independent of order of accuracy) since the initial data for each cell is a polynomial the time evolution is exact whenever the following conditions are met: 1.) The recursion relations (10.9) and (10.10) are run until they truncate, 2.) The forcing is zero (or a polynomial of degree $2m+1$), 3.) Each cell $[x_i, x_{i+1}]$ includes the base of the domain of dependence of the solution at a dual grid point $x_{i+\frac{1}{2}}$ at time $t = \frac{\Delta t}{2}$ (see e.g. [102]). The latter condition can also be stated as a CFL condition*

$$c\frac{\Delta t}{2} \leq \frac{h}{2}.$$

In the present method we do not quite achieve this optimal CFL condition but have verified numerically that our solvers of orders of accuracy 3, 5 and 7 are stable for $c\Delta t \leq 0.75h$.

Variable coefficients

For problems with a variable wave speed the acceleration is governed by

$$v_t = \underbrace{(c^2(x)u_x)_x}_{s(x)}.$$

To compute v_t, v_{tt}, v_{ttt}, etc., needed to evolve the solution by a Taylor series method, we must evaluate the right hand side of (10.1.2). This is done in sequence by form-

ing the polynomial $s(x)$ by operations on polynomials. Let $p \approx u$ and $a \approx c^2$ be polynomials approximating $u(t,x)$ and $c^2(x)$ then

$$s(x) = \mathcal{D}(a(x) \otimes (\mathcal{D}p(t,x))).$$

Here $\mathcal{D}$ denotes polynomial differentiation and $\otimes$ represents polynomial multiplication with degree truncation to the degree of p. Computation of v_t, v_{tt} etc. can now be done by forming $s(x)$ with $p \approx u, u_t, u_{tt}$, etc.

10.1.3 Imposing boundary conditions for the Hermite method

In the hybrid Hermite-DG overset grid method, physical boundary conditions can be imposed on any grid that discretizes the boundary. For example, in the numerical experiments in Section 10.4.6, the boundary conditions are imposed on both grids. In this section we explain how physical boundary conditions are imposed for the Hermite method and a Cartesian grid.

Physical boundary conditions are enforced at the half time level, i.e. when the solution on the dual grid is to be advanced back to the primal grid. As there are many degrees of freedom that are located on the boundary and the physical boundary condition must be augmented by the differential equation to generate more independent conditions so that the degrees of freedom can be uniquely determined. The basic principle, often referred to as compatibility boundary conditions (see e.g. [98]), is to take tangential derivatives of the boundary conditions and combine these with the PDE.

For example, assume we want to impose the boundary condition

$$u(t,0) = g(t). \tag{10.11}$$

Then, as $x_0 = 0$ is a boundary grid point the Taylor polynomials (10.5)-(10.6), centered at x_0, should satisfy the boundary condition (10.11) and compatibility conditions, (i.e. conditions for the derivatives), that one obtains by differentiating (10.11) in time and then replace time derivatives of u in favor of spatial derivatives by using the wave equation. We thus seek a polynomial outside the domain which together with the polynomial just inside the boundary forms a Hermite interpolant that satisfies the boundary and compatibility conditions.

Precisely, to evolve the solution on the boundary we must determine the $2(m+2)$ and $2(m+1)$ coefficients of the polynomials approximating u and v at the boundary. For example for u, this polynomial must interpolate the $(m+2)$ data describing the current approximation of u at dual grid point next to the boundary, this yields $(m+2)$ independent linear equations. The remaining $(m+2)$ independent linear equations can be obtained by requiring that the polynomial satisfies with the boundary condition $u(0,t) = g(t)$ and its time derivatives as described above.

Once the interpolant is determined on the boundary we evolve it as in the interior (see Section 10.1.2).

Remark 20. *We note that in the special case of a flat boundary and homogeneous Dirichlet or Neumann boundary conditions then enforcing the boundary conditions reduces to enforcing that the polynomial on the boundary is either odd or even, respectively, in the normal direction. Then the correct odd polynomial can be obtained by constructing the polynomial outside the domain Ω (often referred as ghost-polynomial) by mirroring the coefficients corresponding to even powers in the normal coordinate variable with a negative sign and the coefficients corresponding to odd powers with the same sign.*

Boundary conditions at interior overset grid boundaries are supplied by projection of the known solutions from other grids and will be discussed below.

10.1.4 Higher dimensions

In higher dimensions the approximations to u and v take the form of centered tensor product Taylor polynomials. In two dimensions (plus time) the coefficients would be of the form $\hat{u}_{k,l,s}$, with the two first indices representing the powers in the two spatial directions, and the third representing time.

For the scalar wave equation

$$\begin{aligned} u_t &= v, \\ v_t &= c^2(u_{xx} + u_{yy}), \quad (x,y) \in \Omega, \quad t > 0, \end{aligned}$$

the recursion relations for computing the time derivatives are a straightforward generalization of the one dimensional case

$$\begin{aligned} \hat{u}_{k,l,s} &= \frac{\Delta t}{s} \hat{v}_{k,l,s-1}, \\ \hat{v}_{k,l,s} &= c^2 \frac{(k+2)(k+1)}{s} \frac{\Delta t}{h_x^2} \hat{u}_{k+2,l,s-1} + c^2 \frac{(l+2)(l+1)}{s} \frac{\Delta t}{h_y^2} \hat{u}_{k,l+2,s-1}. \end{aligned}$$

As noted in [102], using this recursion for all the time derivatives does not produce a method with order independent CFL condition but a method whose time-step size decrease slightly as the order increases. For optimally large timesteps it is necessary to use the special start up procedure

$$\begin{aligned} \hat{u}_{k,l,1} &= \Delta t \, \hat{v}_{k,l,0}, \\ \hat{v}_{k,l,1} &= \Delta t c^2 \left(\frac{(k+2)(k+1)}{h_x^2} \hat{u}^X_{k+2,l} + \frac{(l+2)(l+1)}{h_y^2} \hat{u}^Y_{k,l+2} \right). \end{aligned}$$

Here $\hat{u}^X_{k,l}$ are the $(2m+4) \times (2m+2)$ coefficients of the interpolating polynomial of degree $(2m+3)$ in x and degree $(2m+1)$ in y and $\hat{u}^Y_{k,l}$ are the $(2m+4) \times (2m+2)$ coefficients of the interpolating polynomial of degree $(2m+3)$ in y and degree $(2m+1)$ in x. For the remaining coefficients $s = 2 \ldots, 4m+3$ we use (10.1.4) and (10.1.4) with $k, l = 0, \ldots, 2m+1$. Further details of the two dimensional method can be found in [102].

10.2 Energy based discontinuous Galerkin methods for the wave equation

Our spatial discontinuous Galerkin discretization is a direct application of the energy based formulation described for general second order wave equations in [86, 106, 107]. Here, our energy based DG method starts from the energy of the scalar wave equation

$$H(t) = \int_\Omega \frac{v^2}{2} + G(x, y, \nabla u)\, d\Omega,$$

where $G(x, y, \nabla u) = \frac{c^2(x,y)}{2}|\nabla u|^2$ is the potential energy density, v is the velocity or the time derivative of the displacement, $v = u_t$.

Now, the wave equation, written as a second order equation in space and first order in time takes the form

$$u_t = v, \quad v_t = -\delta G,$$

where δG is the variational derivative of the potential energy

$$\delta G = -\nabla \cdot (c^2(x, y)\nabla u).$$

For the continuous problem the change in energy is

$$\frac{dH(t)}{dt} = \int_\Omega v v_t + u_t \left[\nabla \cdot (c^2(x, y)\nabla u\right]\, d\Omega = [u_t(n \cdot (c^2(x, y)\nabla u))]_{\partial\Omega},$$

where the last equality follows from integration by parts together with the wave equation.

A variational formulation that mimics the above energy identity can be obtained if the equation $v - u_t = 0$ is tested with the variational derivative of the potential energy. Let Ω_j be an element and $(\Pi^{q_u}(\Omega_j))^2$ and $(\Pi^{q_v}(\Omega_j))^2$ be the spaces of tensor product polynomials of degrees q_u and $q_v = q_u - 1$. Then, the variational formulation on that element is:

Problem 1. Find $v^h \in (\Pi^{q_v}(\Omega_j))^2$, $u^h \in (\Pi^{q_u}(\Omega_j))^2$ such that for all $\psi \in (\Pi^{q_v}(\Omega_j))^2$, $\phi \in (\Pi^{q_u}(\Omega_j))^2$

$$\int_{\Omega_j} (c^2 \nabla \phi) \cdot \left(\frac{\partial \nabla u^h}{\partial t} - \nabla v^h \right) d\Omega = [(c^2 \nabla \phi) \cdot n \left(v^* - v^h \right)]_{\partial \Omega_j}, \tag{10.12}$$

$$\int_{\Omega_j} \psi \frac{\partial v^h}{\partial t} + c^2 \nabla \psi \cdot \nabla u^h \, d\Omega = [\psi \, (c^2 \, \nabla u \cdot n)^*]_{\partial \Omega_j}. \tag{10.13}$$

Let $[[f]]$ and $\{f\}$ denote the jump and average of a quantity f at the interface between two elements, then, choosing the numerical fluxes as

$$v^* = \{v^h\} - \tau_1 [[c^2 \, \nabla u^h \cdot n]],$$
$$(c^2 \, \nabla u \cdot n)^* = \{c^2 \, \nabla u^h \cdot n\} - \tau_2 [[v^h]],$$

yields a contribution $-\tau_1([[c^2 \, \nabla u^h \cdot n]])^2 - \tau_2([[v^h]])^2$ from each element face to the change of the discrete energy, guaranteeing that

$$\frac{dH^h(t)}{dt} \equiv \frac{d}{dt} \sum_j \int_{\Omega_j} \frac{(v^h)^2}{2} + G(x, y, \nabla u^h) \leq 0.$$

Physical boundary conditions are enforced through the numerical fluxes, see [86] for details.

Note that the above energy estimate follows directly from the formulation (10.12) - (10.13) but as the energy is invariant to constants equation (10.12) must be supplemented by the equation

$$\int_{\Omega_j} \left(\frac{\partial u^h}{\partial t} - v^h \right) d\Omega = 0.$$

Our implementation uses quadrilaterals and approximations by tensor product Chebyshev polynomials of the solution on the reference element $(r, s) \in [-1, 1]^2$. That is, on each quadrilateral we have approximations on the form

$$u(x(r,s), y(r,s), t_n) \approx \sum_{l=0}^{q_u} \sum_{k=0}^{q_u} c_{lk} T_l(r) T_k(s),$$
$$v(x(r,s), y(r,s), t_n) \approx \sum_{l=0}^{q_v} \sum_{k=0}^{q_v} d_{lk} T_l(r) T_k(s).$$

We choose $\tau_1 = \tau_2 = 1/2$ (so called upwind or Sommerfeld fluxes) which result in methods where u is observed to be $q_u + 1$ order accurate in space [86]. We note that another basis like Legendre polynomials could also be used. In fact we have repeated some of the long time computations in the numerical experiments section below to confirm that a change of basis to Legendre polynomials does not effect the stability or accuracy properties of the method.

10.2.1 Taylor series time-stepping

In order to match the order of accuracy in space and time for the DG method we employ Taylor series time-stepping. Assuming that all the degrees of freedom have been assembled into a vector $\mathbf{w}$ we can write the semi-discrete method as $\mathbf{w}_t = A\mathbf{w}$ with A being the matrix representing the spatial discretization. If we know the discrete solution at the time t_n we can advance it to the next time step $t_{n+1} = t_n + \Delta t$ by the simple formula

$$\begin{aligned} \mathbf{w}(t_n + \Delta t) &= \mathbf{w}(t_n) + \Delta t \mathbf{w}_t(t_n) + \frac{(\Delta t)^2}{2!}\mathbf{w}_{tt}(t_n)\dots \\ &= \mathbf{w}(t_n) + \Delta t A\mathbf{w}(t_n) + \frac{(\Delta t)^2}{2!}A^2\mathbf{w}(t_n)\dots \end{aligned}$$

As we use dissipative fluxes this timestepping method is stable as long as the number of stages in the Taylor series is greater than the order of accuracy in space and with the timestep small enough.

10.3 Overset grid methods

In this section we explain how we use the two discretization techniques described above on overset grids to approximate solutions to the scalar wave equation.

The idea behind the overset grid methods is to cover the bulk of the domain with a Cartesian grid, where efficient methods can be employed, and to discretize the geometry with narrow body-fitted grids. In Figure 10.1 we display two overset grids, a blue Cartesian grid, which we denote a, and a red curvilinear grid, which we denote b, that are used to discretize a geometry consisting of a circular hole cut out from a square region. Note that the grids overlap, hence the name overset grids. Also, note that the annular grid cuts out a part of the Cartesian grid. This cut of the Cartesian grid creates an internal, non-physical boundary in the blue grid.

Here physical boundary conditions are enforced on the red grid at the black boundary which defines the inner circle and on the outermost boundary on the blue grid.

In order to use the Hermite or DG methods on the grids we will need to supply boundary conditions at the interior boundaries. In the example in Figure 10.1 this means that we would have to specify the solution on the outer part of the annular grid and on the staircase boundary (marked with filled black circles) that has been cut out from the Cartesian grid.

In most methods that use overset grids, in particular those using finite differences, the communication of the solution on the interior boundaries is done by interpolation, see e.g. [96]. For the methods we use here we have found that the stability properties are greatly enhanced if we instead transfer volumetric data (numerical solution) in the elements / gridpoints near the internal boundaries by projection rather than by interpolation. In fact, when we use volume data the resulting methods are stable without adding artificial dissipation, when we use interpolation they are not. At the end of this section we discuss a possible reason why the projection behaves better than interpolation.

As mentioned above, in a Hermite method, we can think of the degrees of free-

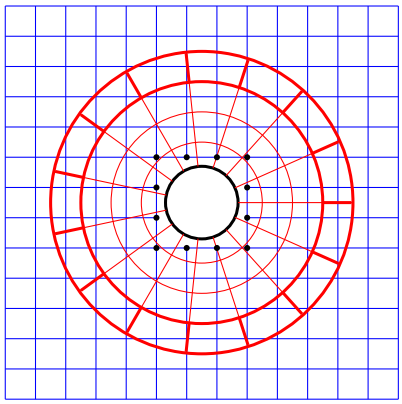

Figure 10.1: An example of an overset grid for a circular boundary inside a square. The red grid is curvilinear and the blue grid is Cartesian (in a realistic problem the red grid would be significantly thinner). The black filled circles indicate the cut out domain boundary.

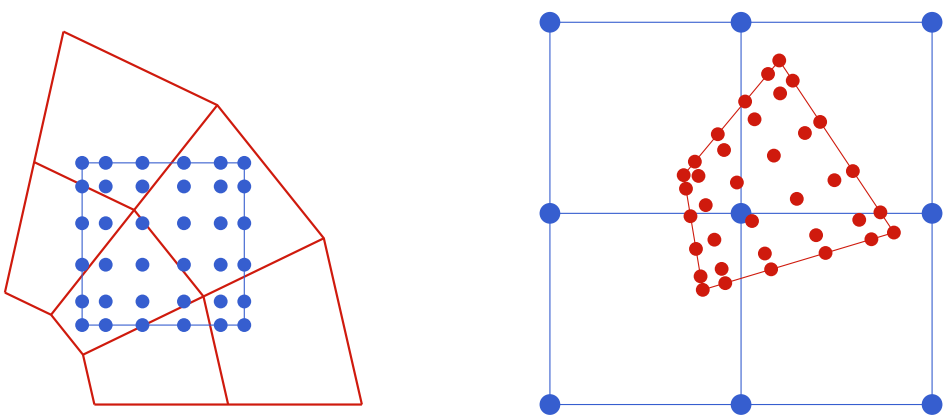

Figure 10.2: Typical setup for communication. In the left subfigure the local tensor product GLL grid around a Hermite grid point is marked with filled blue circles. The points in the GLL grid may be covered by different DG elements. In the right subfigure the tensor product grid inside the DG element is marked with filled red circles. The points in the GLL grid may be contained in different Hermite cells.

dom as either being nodal data, consisting of function and derivative values, or as coefficients in a Taylor polynomial. Thus, when transferring data to a grid where a Hermite method is used (like the example in the left subfigure of Figure 10.2) we

must determine a tensor product polynomial centered around a gridpoint local to that grid (the points we would center around are indicated by black points in Figure 10.1). Below we will explain in detail how we determine this polynomial.

For elements with an internal boundary face (denoted by thick red lines in Figure 10.1) we could in principle transfer the solution by specifying a numerical flux on that face, however we have found that this approach results in weakly unstable methods. Instead we transfer volumetric data to each element that has an internal boundary face, we give details below. Given the timestep constraints of DG methods we must march the DG solution using much smaller timesteps than those used for the Hermite method. This necessitates the evaluation of the Hermite data not only at the beginning of a Hermite timestep but at many intermediate times.

10.3.1 Determining internal boundary data for the Hermite solver

We first consider the problem of determining internal boundary data required by the Hermite method. An example of how to compute solution data at the gridpoints (x_i, y_j) at the boundary of Cartesian grid (filled black circles) is depicted in Figure 10.1.

In general, the tensor product polynomial centered around (x_i, y_j) is found by a two step procedure. First we project into a local L_2 basis spanned by Legendre polynomials and perform a numerically stable and fast change of basis into the monomial basis. Then we truncate the monomial to the degree required by the Hermite method.

To carry out the L_2 projection we introduce a local tensor product Gauss-Legendre-Lobatto (GLL) grid centered around (x_i, y_j). These points are marked

as filled blue circles in the left subfigure of Figure 10.2. The number of grid points in the local grids are determined by the order of the projection. To maintain the order of the method, the order of the projection should be at least the same as the order of the spatial discretization, thus it is sufficient to have $2m+4$ points in each direction. The GLL quadrature nodes are defined on the reference element $(r,s) \in [-1,1]^2$ that maps to a cell defined by the dual gridpoints closest to (x_i, y_j).

Let $\tilde{u}$ be the numerical solution on the red grid. In the first step of the communication we compute the coefficients of a polynomial $\tilde{p}$ approximating $\tilde{u}$ by projecting $\tilde{u}$ on the space of tensor product Legandre polynomials $P_l P_k$, that is

$$\tilde{p}(r,s) = \sum_{l=0}^{2m+3} \sum_{k=0}^{2m+3} c_{lk} P_l(r) P_k(s), \;\; c_{lk} = \frac{(\tilde{u}, P_l P_k)}{\|P_l P_k\|^2}.$$

Here (f,g) denotes the L_2 inner product on $(r,s) \in [-1,1]^2$ and $\|f\|_2^2 = (f,f)$ is the norm induced by the inner product. Note that the expression (10.3.1) is particularly simple since the Legendre polynomials are orthogonal on the domain of integration. To do this we evaluate $\tilde{u}$ at the underlying blue quadrature points in the left subfigure of Figure 10.2.

Once the polynomial (10.3.1) has been found we perform a change of basis into the local monomial used by the Hermite method. Such a change of basis can be done by the fast Vandermonde techniques by Björk and Pereyra, see e.g. [108, 109]. At this stage the polynomial is of total degree $2m+3$ so the final step is to truncate it to total degree m or $m+1$ depending on whether we are considering the displacement or the velocity. With the $(m+1)^2$ and $(m+2)^2$ degrees of freedom determined everywhere on a Hermite grid we may evolve the solution as described in Section 10.1.

10.3.2 Determining data for DG elements with internal boundary faces

We now consider the problem of determining the data required by the DG method. Here we show how to obtain the data at a single DG element with at least one internal boundary face. As the timesteps of the DG method are significantly smaller than for the Hermite method we must repeat the transfer of data many times. We must also explicitly transfer time derivative data in order to use a Taylor series timestepping approach.

The tensor product polynomials in our implementation of the DG method are composed by the product of Chebyshev polynomials $T_j(z) = \cos(j \cos^{-1}(z))$ that are expressed on the reference element $(r, s) \in [-1, 1]^2$. Precisely we seek

$$p(r, s) = \sum_{l=0}^{q} \sum_{k=0}^{q} c_{lk} T_l(r) T_k(s).$$

To determine such polynomials we perform a projection of the solution u, i.e the solution on Cartesian grid,

$$c_{lk} = \frac{(\tilde{u}, T_l T_k)_{\mathrm{C}}}{\|T_l T_k\|_{\mathrm{C}}^2},$$

but in this case the weighted inner product is

$$(f, g)_{\mathrm{C}} = \int_{-1}^{1} \int_{-1}^{1} \frac{f(r, s) g(r, s)}{\sqrt{1 - r^2}\sqrt{1 - s^2}} \, dr ds,$$

where the Chebyshev polynomials are orthogonal. To carry out this projection we use local tensor product Chebyshev quadrature nodes, $2m + 2$ in each dimension, as shown in right subfigure of Figure 10.2.

The local time levels used by the DG solver nth Hermite timestep are defined to be

$$t_{n,\nu} = t_{n,0} + \nu \Delta t_b, \ \nu = 0, \ldots N_{\mathrm{DG}},$$

where Δt_b, and similarly Δt_a, are timesteps taken on grids b (curvilinear) and a (Cartesian) respectively. For simplicity the starting local time level and the final local time level are equal to consequent timesteps on the Hermite grid, t_n and t_{n+1}

$$t_{n,0} = t_n, \quad t_{n,N_{\mathrm{DG}}} = t_{n+1}.$$

To transfer the solution values and the time derivatives needed at each of the quadrature points and at each $t_{n,\nu}$ we carry out the following "start up" procedure at $t_{n,0}$. For each of the quadrature points we re-center the Hermite interpolants closest to it and compute the time derivatives precisely by the recursion relations described in Section 10.1. We note that this is an inexpensive computation as the interpolants have already been found as a step in the evolution of the Hermite solution, the only added operation is the re-centering.

10.3.3 Discussion of projection and interpolation

One of the differences in the present method and a finite difference method is that during the transfer of data to the Hermite method there is a degree truncation of (in one dimension and for u) a polynomial of degree $(2m+3)$ to a polynomial of degree $(m+1)$. It is natural to ask how the truncated polynomial depends on whether projection or interpolation was used to find the un-truncated polynomial.

Suppose the same $(2m+4)$ data has been used to determine two polynomials

$$p_{\mathrm{interp.}}(z) = \sum_{l=0}^{2m+3} a_l z^l, \quad z \in [-1/2, 1/2],$$

and

$$p_{\mathrm{proj.}}(z) = \sum_{l=0}^{2m+3} \tilde{b}_l P_l(2z) = \sum_{l=0}^{2m+3} b_l z^l, \quad z \in [-1/2, 1/2].$$

Then due to the orthogonality of the projected polynomial it is clear that the trun-

cated polynomial satisfies

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\sum_{l=0}^{m+1} \tilde{b}_l P_l(2z) \right)^2 dx \leq \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(p_{\text{proj.}}(z) \right)^2 dx.$$

The polynomial determined by interpolation does not satisfy a similar inequality. In fact the truncation can cause a significant increase in the L_2-energy. To investigate this we find

$$\{a_0^*, \ldots, a_{2m+3}^*\} = \text{argmax} \frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\sum_{l=0}^{m+1} a_l z^l \right)^2 dx}{\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\sum_{l=0}^{2m+3} a_l z^l \right)^2 dx},$$

for 10000 randomly selected initial data and for $m = 1, 2, 3$. The largest ratio between square of the L_2-norms of the truncated and un-truncated polynomials were 19, 657 and 3555 for $m = 1$, $m = 2$ and $m = 3$ respectively. While this does not conclusively rule out that it could be possible to use interpolation it does indicate that a projection based approach is to be preferred. We stress that the cause of the problem is the combination of the truncation and interpolation and that there is therefore not obvious that there is any advantage to use projection rather than interpolation for methods that do not have truncation (like finite difference methods).

10.4 Numerical experiments

The hybrid H-DG method is empirically stable and accurate, and here we demonstrate it with numerical experiments. To test the stability of the method in one dimension we first define the amplification matrix and compute its spectral radius. To test the stability in two dimensions, where the amplification matrix will take too long to compute, we provide the long time simulation and estimate the error growth for multiple refinements. Convergence tests in one and two dimensions are done for the domains where the exact solution is known. In the second half of this section we apply the method to the domain with complex curvilinear boundary in an ex-

periment with wave scattering from a smooth pentagonal object. Finally, in the end of this section we apply the method as the forward solver in the inverse problem of locating underground cavities.

10.4.1 Numerical stability test

Unlike the Hermite and DG methods, stability of the hybrid H-DG method cannot easily be shown analytically. As a weaker alternative, the stability can be investigated numerically by looking at the spectrum of the amplification matrix associated with the method, [110].

To construct the amplification matrix we apply the method to initial data composed of the unit vectors. The vector that is returned after one timestep is then placed as columns in a square matrix. If the spectral radius of the amplification matrix is smaller than 1, or if the eigenvalues with magnitude one correspond to no-trivial Jordan blocks, then the amplification matrix is power-bounded.

We consider the wave equation (10.1)-(10.2) on the unit interval $x \in [0, 1]$ with homogeneous Dirichlet and Neumann boundary conditions at $x = 0$ and $x = 1$ respectively. We introduce two uniform Cartesian grids which overlap inside a small interval close to one of the boundaries. Precisely, the grids are

$$\Omega_a = \{x_i^a = ih_a, \quad i = 0, \ldots, n_a\},$$
$$\Omega_b = \{x_i^b = 1 - (n_b - i)h_b, \quad i = 0, \ldots, n_b\}.$$

The Hermite method is used on a grid a and the DG method is used on grid b. The grids thus overlap inside the interval $[x_0^b, x_{n_a}^a]$. Here the ratio of the overlap size and the discretization width is $(x_{n_a}^a - x_0^b)/h_a$. This ratio is fixed for all values of h_a and h_b. We also fix n_b so that the amount of work done on grid b is constant per timestep for all refinements. Fixing the ratio $(x_{n_a}^a - x_0^b)/h_a$ and n_b makes the efficiency of the

overall method asymptotically determined by the efficiency the Hermite method.

Let $\mathbf{w}^n$ be a vector holding the degrees of freedom of both methods at nth timestep, then we may express the complete timestep evolution as $\mathbf{w}^{n+1} = \mathcal{H}\mathbf{w}^n$ where $\mathcal{H}$ incorporates timestepping and projection. $\mathcal{H}$ can be expressed as the matrix H that can be computed column by column via

$$H_k = \mathcal{H}e_k, \tag{10.14}$$

where e_k is the kth unit vector. The equation

$$\mathbf{w}^n = H^n\mathbf{w}^0,$$

is equivalent to the n timesteps of the hybrid H-DG method. Let $\lambda = \rho(H)$ be the spectral radius of H. If $|\lambda| < 1$ then $\|H^n\|_2$ will tend to zero for large n. Of course this only means that this particular discretization of this particular problem is stable and does (in principle) not tell us anything about other grid configurations.

We consider the case $c = 1$ and take the parameters to be

$$n_a = 10, 20, \ldots, 60, \quad n_b = 5, \quad \frac{h_b}{h_a} = 0.9.$$

Other parameters are q_u, q_v for the DG method and n_{DG}, the number of timesteps done by the DG method during one step of the Hermite method. The parameters q_u and q_v are set so the methods used have the same order of accuracy as the approximation of v for the Hermite method

$$q_u = 2m + 2, \; q_v = 2m + 1.$$

To get an optimal n_{DG}, we take the largest possible timestep for the energy based DG method (empirically determined in [86]), so that

$$\frac{\Delta t_b}{h_b} \leq 0.15/q_u,$$

and

$$n_{\mathrm{DG}} = \frac{\Delta t_a}{\Delta t_b},$$

is an integer. Equivalently, if the Hermite method CFL number is set, we get

$$n_{\mathrm{DG}} = \frac{\Delta t_a}{\Delta t_b} = \left\lceil \mathrm{CFL} \frac{q_u}{0.15} \frac{h_a}{h_b} \right\rceil .$$

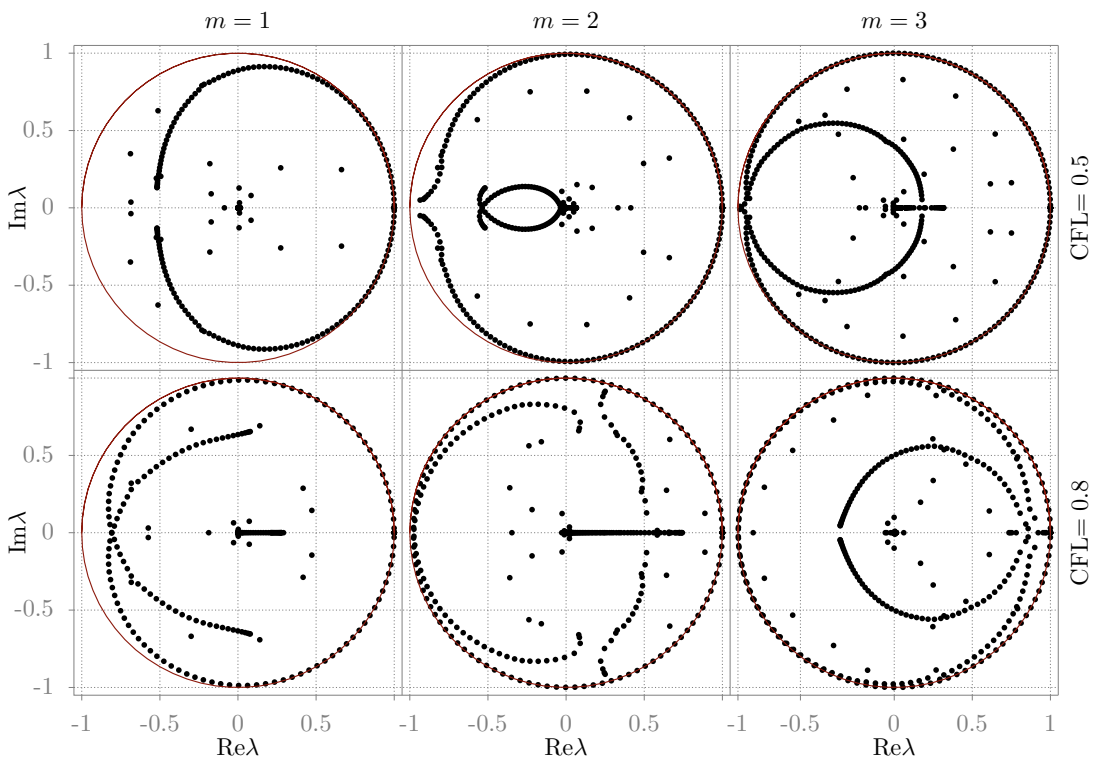

Figure 10.3: Spectrum of the amplification matrix H for CFL numbers $\Delta t_a/h_a = 0.5, 0.8$, orders of accuracy $3, 5, 7$, and $n_a = 40,\ n_b = 5$. No eigenvalues are outside the unit circle.

Following the column-by-column construction process (10.14) described above we compute the amplification matrix H. The spectrum of H is shown in Figure 10.3 for $m = 1, 2, 3$. Displayed results are for the cases $n_a = 40$ and $n_b = 5$. The CFL numbers set for Hermite method are $\Delta t_a/h_a = 0.5$ and 0.8. The absolute value of

eigenvalues does not exceed 1. We note that if interpolation is used some eigenvalues of the amplification matrix shift outside of the unit circle. Such unstable modes can possibly be stabilized by numerical dissipation / hyperviscosity but we do not pursue such stabilization here. Instead we observe that when projection is used all eigenvalues are inside the unit circle and the method is stable. Although we only display the results for one problem here the same results were obtained for other grid sizes, various overlap sizes to grid spacing ratios and different CFL numbers set for the Hermite method. We stress that it is possible to make the method unstable if we take the CFL number close to one and if we take m to be larger than 3 and thus we only claim that the methods of orders of accuracy up to 7 are stable.

10.4.2 Convergence to an exact solution

Using the same grid setup and boundary conditions as in the example above we test the method for the wave equation (10.1)-(10.2), $c = 1$ and initial conditions

$$u(x,0) = \sin\left(\frac{15\pi}{2}x\right),$$
$$v(x,0) = 0.$$

A solution to this problem is the standing wave

$$u(x,t) = \sin\left(\frac{15\pi}{2}x\right)\cos\left(\frac{15\pi}{2}t\right).$$

The errors for the solution on the grids are

$$\varepsilon_a(x,t) = p_{i+\frac{1}{2}}(x,t) - u(x,t), x \in x_i^a, x_i^a + 1,\ i = 0,\ldots,n_a,$$

for the Hermite grid and

$$\varepsilon_b(x,t) = u^{h_b}(x,t) - u(x,t),$$

for the DG grid. The maximum error for the total method is

$$\max\left(\max_{x\in[x_0^a,x_{n_a}^a]}|\varepsilon_a(x,t)|),\ \max_{x\in[x_0^b,x_{n_b}^b}|\varepsilon_b(x,t)|)\right).$$

In Figure 10.4 we display computed maximum errors as functions of time for the method with $m = 3$ (i.e. the order of accuracy is 7). In the left subfigure the CFL number for the Hermite method is set to be 0.5 and in the right subfigure the CFL number is set to be 0.75. For all Hermite grid sizes, the error growth is linear in time (dashed lines display a least squares fit of a linear function), indicating that the solution is stable for long time computations.

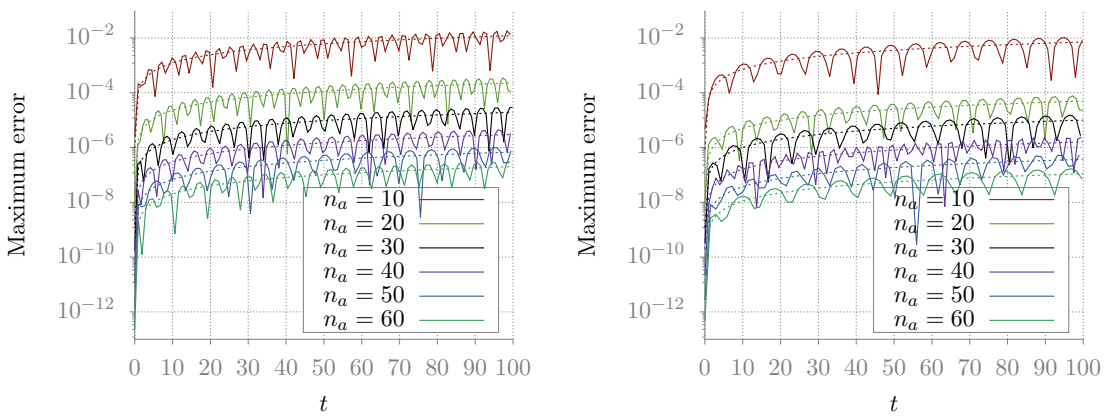

Figure 10.4: Maximum error of the solution as a function of time. The curves correspond to different refinements for $m = 3$. In the left subfigure CFL number for Hermite method is set to 0.5. In the right subfigure CFL number for Hermite method is set to 0.75. Dashed lines display lines αt.

In the left subfigure of Figure 10.5 the numerical solution and the absolute error are shown for the 7th order accurate method at time $t = 2$. As can be seen in the lower left subfigure in Figure 10.5 the error is rather smooth across the overlap indicating that the projection is highly accurate.

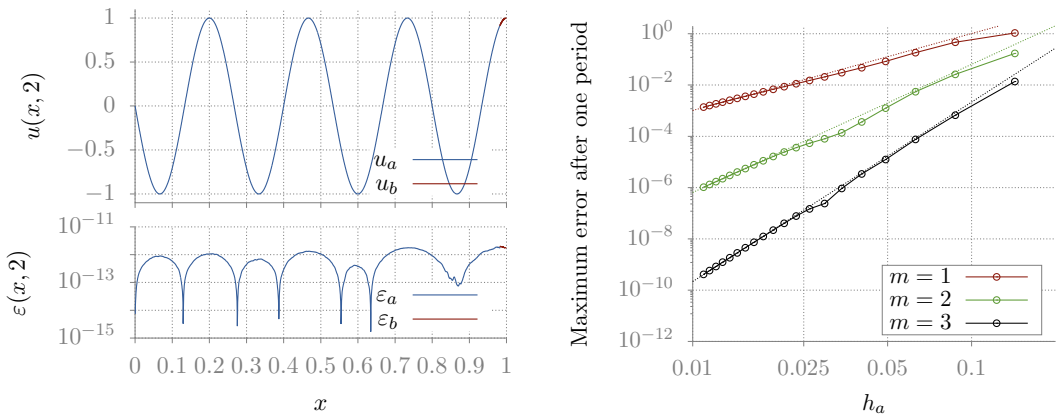

Figure 10.5: The upper left subfigure displays the solution at time $t = 2$. The error at time $t = 2$ is shown in the lower left subfigure. The number of grid points are $n_a = 200$, $n_b = 5$ and $m = 3$, the Hermite CFL number is set to 0.75. Red curves indicate the solution and the error on the DG grid. Blue curves indicate the solution and the error on the Hermite grid. (The solution and the error were computed on finer grid, 10 grid points per cell/element). In the right subfigure we display a convergence plot for $m = 1, 2, 3$. Dashed lines show the least squares fit of $C_m h_a^q$, $q = 3, 5, 7$.

To the right in Figure 10.5 the error at the final time $t = 2$ is shown as a function $h = h_a$. The dashed lines show the least squares fit with polynomial functions of h_a of order $3, 5$ and 7 respectively. The results indicate that the orders of accuracy of the methods are $2m + 1$ as expected. The parameters (n_a, n_b, n_{DG}, etc.) are the same is in previous example.

10.4.3 Analytical solution in a disk. Rates of convergence

Consider the solution of (10.1)-(10.2) with $f(x, y, t) \equiv 0$ on the unit disk, $(x, y) \in x^2 + y^2 \leq 1$, with homogeneous Dirichlet boundary conditions. Then the analytical solution can be expressed in polar coordinates as a composition of modes

$$u_{\mu\nu}(r, \theta, t) = J_\mu(r\kappa_{\mu\nu}) \cos(\mu\theta) \cos \kappa_{\mu\nu} t.$$

Here $J_\mu(z)$ is the Bessel function of the first kind of order μ and $\kappa_{\mu\nu}$ is the νth zero of J_μ. In the following experiment we set $\mu = \nu = 7$, $\kappa_{77} = 31.4227941922$. The initial condition $u_{77}(x, y, 0)$ is displayed in the left subfigure of Figure 10.6.

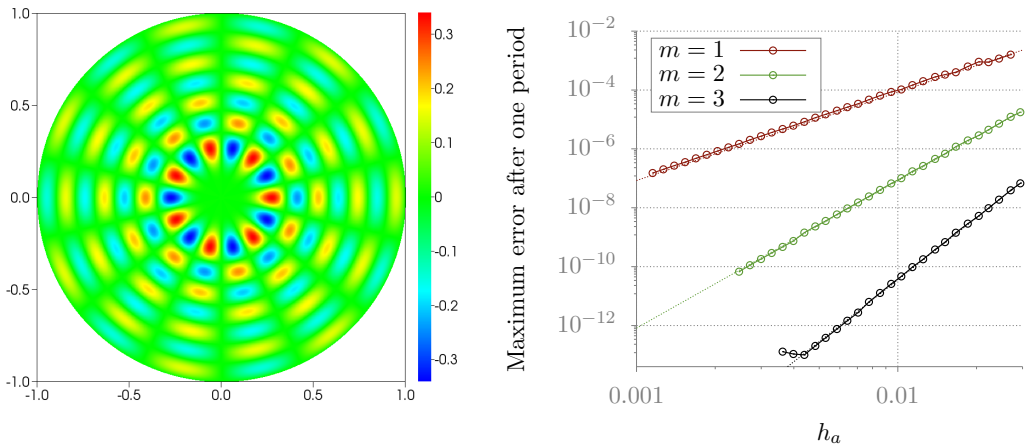

Figure 10.6: The left subfigure displays the initial condition. In the right subfigure the max-error at time $t = 2\pi/\kappa 77$ as a function of grid spacing of the Hermite method. Solid curves correspond the methods with $m = 1, 2, 3$ and dashed lines display the expected the convergence rates i.e. $\mathcal{O}(h_a^{2m+1})$.

We setup overset grids as displayed in Figure 10.7. Grid a is a Cartesian grid discretizing a square domain with $2n_a + 1$ grid points in each direction and grid spacing $h_a = 1/n_a$. Grid b is a curvilinear grid discretizing a thin annulus with radial grid spacing $1.1h_a$. For all refinements Grid b has 7 elements in the radial direction thus the number of elements (or equivalently the number of DOFs of the DG method) will grow linearly with the reciprocal of the discretization size h_a. In contrast the number of grid points in the Cartesian grid where the Hermite method will be used grows quadratically with $1/h_a$.

To measure the error we evaluate the solution on a finer grid, oversampled with 20 grid points inside each Hermite cell and DG element. The convergence is displayed in the right subfigure of Figure 10.6. The errors at time $t = 2\pi/\kappa_{77}$ as functions of h_a

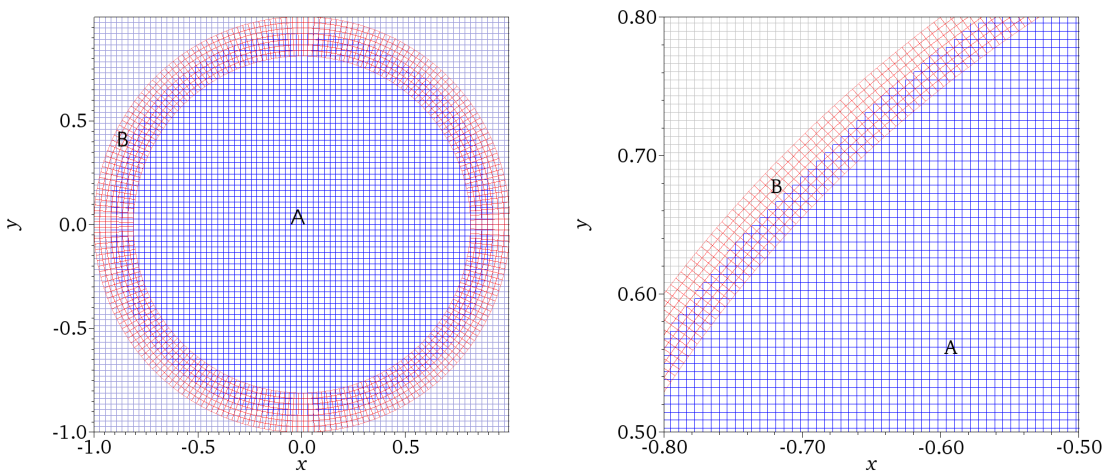

Figure 10.7: Overset grid set up for two different discretization widths. The Hermite grid is blue and DG grid is red. The Hermite grid is truncated at radius $1 - h_a$, i.e one Hermite grid spacing smaller than the computational domain. This creates a stair sharped interior boundary. The solution at that boundary is imposed by by the projection described above. The curvilinear grid has 7 elements in the radial direction, thus the number of elements grows linearly with n_a. The number of grid points in the Hermite grid grows as n_a^2.

for $m = 1, 2, 3$ are displayed as solid lines. The dashed lines show the polynomials in h_a of order $2m+1$. We use $h_a = 1/34, 1/36, ..., 1/94$ in the computations. As can be seen the expected orders of accuracy (3,5 and 7) are observed. To test the stability of the method we evolve the solution until time $t = 60\pi/7$ which is roughly 130 periods of the solution. We set $h_a = 1/54$ and test methods with orders of accuracy $3, 5$ and 7. The error growth appears to be linear in time as indicated by dashed lines in the right subfigure of Figure 10.8.

To test the performance of the method we evolve the method over one time period of the solution and measure the CPU time, see the left subfigure of Figure 10.8. The red curve, displaying the error of the 3rd, order accurate method only reaches the error 10^{-6} in about 1000 seconds while the 5th and 7th order accurate methods, using

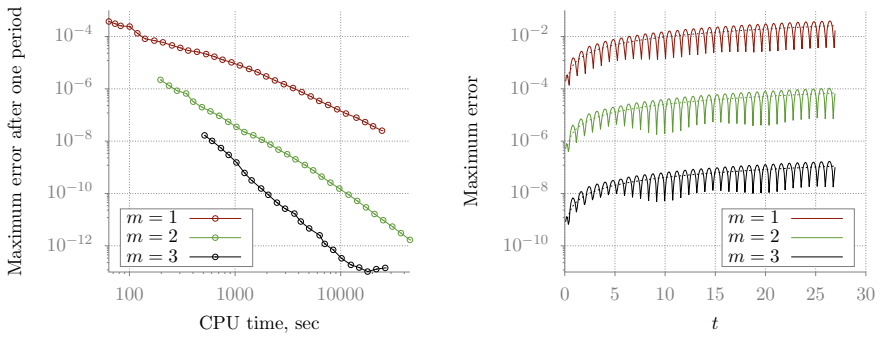

Figure 10.8: In the left subfigure the error at fixed time $t = 2\pi/\kappa_{77}$ is shown as a function of CPU time. The right subfigure displays the error as a function of time t. Dashed lines are the linear functions αt formed by a least squares fit.

the same compute time, yield errors on the order of 10^{-8} and 10^{-10} respectively. Clearly the higher order methods are more efficient.

Table 10.1 displays a breakdown of time spent in the various parts of the code. As can be seen from the timing results the largest time is spent in the DG solver even for the finest grid. The increase in time does grow approximately quadratically and linearly for the Hermite and DG respectively so that eventually the complexity of the Hermite solver will dominate but practically speaking this may not happen for practical refinements for this problem. The large computational cost of the DG method is, in part, due to the small timestep requirement but also due to our implementation.

10.4.4 A wave scattering of a smooth pentagon

In this experiment we study the scattering of a smooth pentagon in free-space. In addition to the use of non-reflecting boundary conditions, this experiment demonstrates the hybrid Hermite-DG method for a solution which is propagated over many wave-

	HERMITE	DG	DG per step	H→ DG	DG → H
TIME	0.34	70.33	1.56289	10.00	11.32
DOF	300448	222992	222992	15744	55748
TIME / DOF	1.13(-6)	3.15(-4)	7.00(-6)	6.35(-4)	2.03(-4)
TIME	1.61	159.67	3.39	22.32	23.77
DOF	1244760	451052	451052	32144	112763
TIME / DOF	1.29(-6)	3.53(-4)	7.51(-6)	6.94(-4)	2.10(-4)
TIME	4.86	183.45	4.08	32.74	41.07
DOF	5066944	905724	905724	64944	226431
TIME/DOF	9.59(-7)	2.02(-4)	4.45(-6)	5.04(-4)	1.8(-4)

Table 10.1: Timing of the 7th order accurate hybrid Hermite-DG method for the disk experiment. The table contains timings for three different numbers of degrees of freedom. TIME denotes average time in seconds per 1 Hermite timestep of Hermite timestepping, DG timestepping and communication stages with the exception of the fourth column which displays the time per 1 DG timestep for the DG method. The TIME/DOF row in each block displays the time per degree of freedom computed by time evolution or communication.

lengths. The geometry of the pentagon is defined as the smooth closed parametric curve:

$$\begin{aligned} x(s) &= \frac{1}{10}\left(1+\frac{1}{10}\cos(10s)\right)\cos(s), \\ y(s) &= \frac{1}{10}\left(1+\frac{1}{10}\cos(10s)\right)\sin(s), \quad s\in[0,2\pi). \end{aligned}$$

The pentagon is placed in a square domain $(x,y)\in[-2,2]^2$ discretized by a Cartesian grid with grid spacing $1/n$, $n=40$. The curvilinear grid has 10 elements in the radial direction and the outer boundary is a circle of radius $0.1+20/n$. The overlap width is at most 5 DG elements.

On the boundary of the body we set Dirichlet data

$$u(x,y,t)=\sin(\omega t), \quad (x,y)\in\Gamma, \quad t\geq 0, \quad \omega=250.$$

The exterior boundary condition is modeled by truncating the domain using perfectly

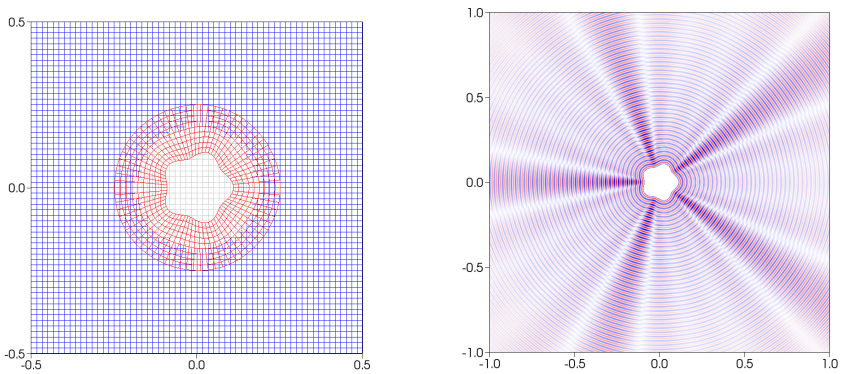

Figure 10.9: Left: Overset grid set up around the body. The overlapping DG grid and background Cartesian grid shown on the domain $[-0.5, 0.5]^2$. Right: Snapshot of $u(x, y, 10)$.

matched layers governed by the equations, (see [80] for derivation)

$$u_{tt} = \frac{\partial}{\partial x}\left(u_x + \sigma^x \phi^{(1)}\right) + \frac{\partial}{\partial y}\left(u_y + \sigma^y \phi^{(2)}\right)\sigma^{(x)}\phi^{(3)} + \sigma(y)\phi^{(4)}, \tag{10.17}$$

where the auxiliary variables satisfy the equations

$$\phi_t^{(1)} + (\alpha + \sigma(x))\phi^{(1)} = -u_x,$$

$$\phi_t^{(2)} + (\alpha + \sigma(y))\phi^{(2)} = -u_y,$$

$$\phi_t^{(3)} + (\alpha + \sigma(x))\phi^{(3)} = -u_{xx} - \tfrac{\partial}{\partial x}\left(\sigma^{(x)}\phi^{(1)}\right),$$

$$\phi_t^{(4)} + (\alpha + \sigma(y))\phi^{(4)} = -u_{yy} - \tfrac{\partial}{\partial y}\left(\sigma^{(y)}\phi^{(2)}\right).$$

The damping profiles $\sigma^{(z)}$, $z = x, y$ are taken as

$$\sigma^{(z)}(z) = \sigma_s\left(\tanh\left(\frac{z - z_1}{0.7w_{\text{lay}}}\right) - \tanh\left(\frac{z - z_2}{0.7w_{\text{lay}}}\right)\right).$$

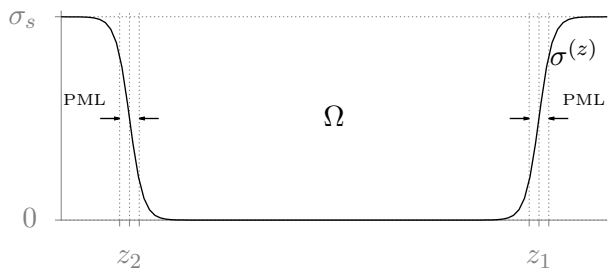

Figure 10.10: The damping profile $\sigma^{(z)}$ for PML with damping strength σ_s. The damping is zero at the center of the domain Ω and rapidly increases in the PML on the both edges of the domain.

Here σ_s is a damping strength, w_{lay} is layer width and z_1 and z_2 control the location of the damping profile of the PML. The shape of $\sigma^{(z)}$ is displayed in Figure 10.10. In experiments involving PML we discretize the modified equations (10.17) with the Hermite method. The order of accuracy of the methods is set to be 7, i.e. $q_u = q_v = 6$, and $m = 3$. The solution is evolved to $t = 10$. A snapshot of the solution at the final time is displayed in the right subfigure of Figure 10.9. The proposed algorithm clearly is able to accurately propagate waves in complex domains.

Table 10.2 displays a breakdown of time spent in the various parts of the code. As can be seen from the timing for this problem the largest time is now spent in the Hermite solver. Here, due to the geometry being an interior object, the relative number of degrees of freedom in the DG solver is small and we see the asymptotic behavior more clearly than for the disc experiment.

	HERMITE	DG	DG per step	H→ DG	DG → H
TIME	23.87	14.15	0.25	1.98	0.27
DOF	1972100	87010	87010	3280	7910
T/DOF	0.12(-4)	0.16(-3)	0.29(-5)	0.60(-3)	0.34(-4)
TIME	40.10	17.52	0.37	2.52	0.42
DOF	4170520	125543	125543	4920	11413
T/DOF	0.96(-5)	0.13(-3)	0.30(-5)	0.51(-3)	0.37(-4)
TIME	76.18	27.38	0.60	3.94	0.70
DOF	11007352	203852	203852	8200	18532
T/DOF	0.69(-5)	0.13(-3)	0.29(-5)	0.48(-3)	0.38(-4)
TIME	162.28	43.08	0.87	7.30	1.38
DOF	34433112	287811	287811	15088	31979
T/DOF	0.47(-5)	0.14(-3)	0.31(-5)	0.48(-3)	0.43(-4)

Table 10.2: Timing of the 7th order accurate hybrid Hermite-DG method for the smooth pentagon experiment. The table contains timings for three different numbers of degrees of freedom. TIME denotes average time in seconds per 1 Hermite timestep of Hermite timestepping, DG timestepping and communication stages with the exception of the fourth column which displays the time per 1 DG timestep for the DG method. The T/DOF row in each block displays the time per degree of freedom computed by time evolution or communication.

10.4.5 Wave scattering of many cylinders in free space

As another demonstration of the method we simulate a domain with multiple circular holes. Precisely we consider the infinite domain $\Omega \in [-\infty, \infty] \times [-\infty, 1.33]$ with homogeneous Neumann boundary condition at $y = 1.33$. The computational domain is a rectangle $[-1, 1] \times [-1.33]$ with PML $|x| > 1$ and $y < -1$. Inside the computational domain there are 5 cylinders of radii 0.1 and centers at $(x_k, y_k), k = 1, \ldots 5$. We impose the homogeneous Dirichlet boundary conditions on the boundary of all cylinders except first. On the first cylinder we impose a time dependent boundary condition

$$u(t, x, y) = (t - 0.1) \exp\left(-918(t - 0.1)^2\right), \ (x, y) \in \{(x - x_1)^2 + (y - y_1)^2 = 0.01\}.$$

The initial solution is at rest.

The set up of the numerical method is similar to the previous experiment. The Cartesian grid covers the background domain and the PML. The 5 circular grids are placed around the bodies as shown in the upper left subfigure Figure 10.11. In this experiment we used the 7th order method. It can be noticed that as in all solution plots provided in this paper the solution is smooth across the overlap due to the high accuracy of methods used and the projection used for communication.

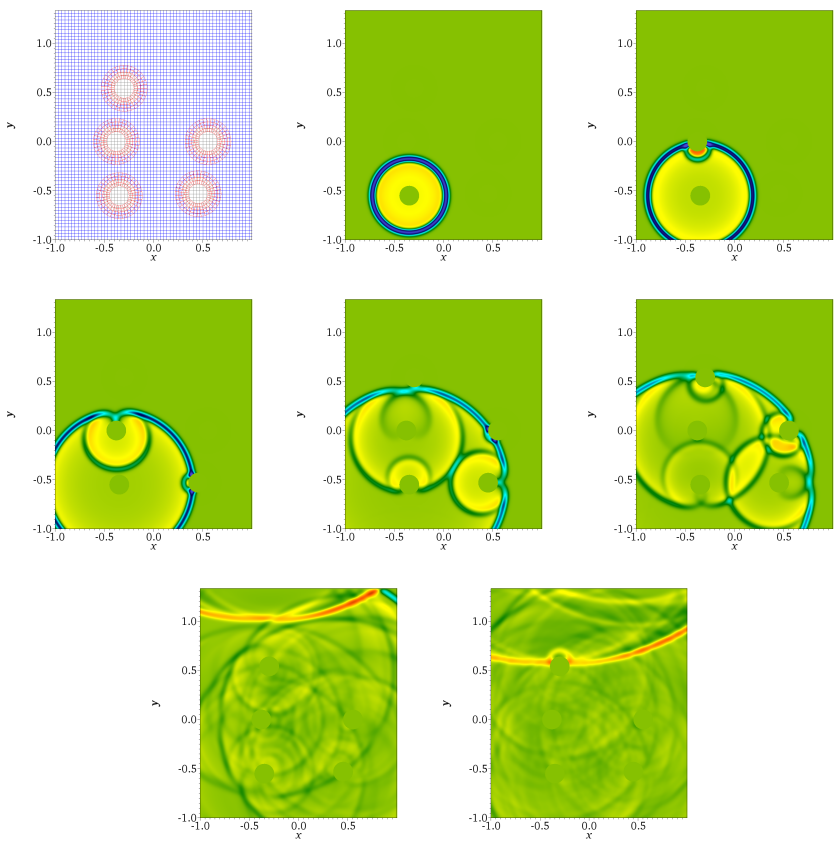

Figure 10.11: Overset grid setup and solution plots for 5 bodies in a free half space. In the upper left subfigure the grids are shown: DG gridlines plotted with red color, Cartesian grid lines inside the domain plotted with blue. Other figures are the solution plots at various increasing times.

10.4.6 An inverse problem, locating a body in free space

As a final experiment we solve the inverse problem of locating a cylindrical body in free space. An application of this problem could be a to locate a tunnel under the ground and determine its radius by sending waves from source devices buried at a relatively small distance from the surface and recording the solution near the surface. Waves will propagate from a source, reflect from an underground cavity and travel back to the surface to be captured by the recording devices. The underground cavity can be located by minimizing a cost functional, i.e misfit function of recorded data and data obtained from the numerical simulation in each iteration of the optimization process.

Consider a square region $\Omega \in [-1,1] \times [-1,1.25]$ with 3 circular bodies of radius $r = 0.1$ with centers at $x_1 = -0.7$, $x_2 = 0$, $x_3 = 0.7$ and $y_1 = y_2 = y_3 = -0.7$. On the boundary of the bodies we impose homogeneous Dirichlet boundary conditions. On the top boundary $y = 1.33$, that acts as a "ground surface" we impose homogeneous Neumann boundary conditions. The exterior boundary conditions at $x = \pm 1$ and $y = -1$ are imposed by truncating the domain using a PML. We discretize the domain with a Cartesian grid. Around each of the cavities we place annular DG grids that are 5 cells wide. An example of a complete set up with 4 receivers is shown in Figure 10.12.

First we create synthetic data by recording the displacement u at equidistant locations of the receivers

$$(0, 0.125),\ (0.25, 0.125),\ (0.25, 0.125),\ (0.25, 0.125),$$

to time $T = 2$. Let there be another circular body of radius A_1 and center at $(x, y) = (A_2, A_3)$ that we want to locate. In the right figure the first source is active, i.e. the initial condition is a smooth Gaussian centered at $\hat{x} = -0.25$, $\hat{y} = 1$, with

no initial velocity

$$u(0,x,y) = \exp\left(-40\left((x-\hat{x})^2 + (y-\hat{y})^2\right)\right), \quad v(0,x,y) \equiv 0.$$

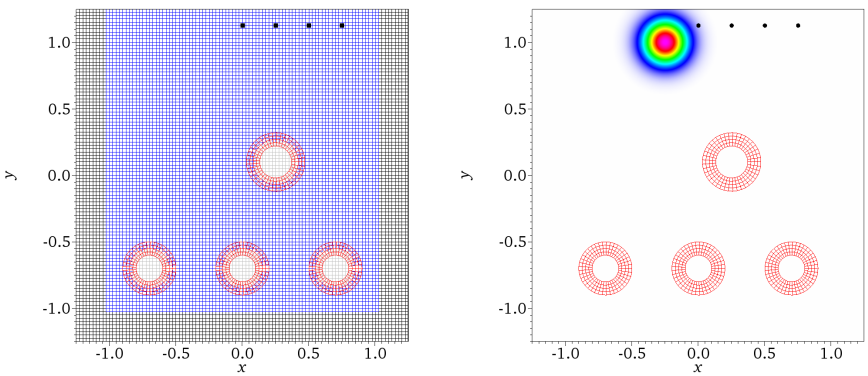

Figure 10.12: The inverse problem set up. Receivers are marked as dots. The left subfigure displays the complete overset grid set up, with 4 DG grids around bodies, and a Cartesian background grid. Blue and red grids discretize physical subdomain; the black grid is the PML layer; the gray grids indicate the subdomains covered by DG grids. The right subfigure displays of the initial condition, a smooth Gaussian centered at (-0.25, 1), receivers and the DG grids.

First we create the synthetic data for the exact location of the target, $A_1^* = 0.12$, $A_2^* = 0.25$ and $A_3^* = 0.1$. This gives us u^*. To locate the cavity we minimize the cost function that is a sum of squared L_2 norms of discrepancies between the output of the numerical simulation and synthetic data $u^*(t, \check{x}_l, 0.125)$

$$F(A_1, A_2, A_3) = \sum_{l=1}^{4} \int_0^T \left(u(t, \check{x}_l, 0.125) - u^*(t, \check{x}_l, 0.125)\right)^2.$$

During the minimization we impose the bounds $0.01 \leq A_1 \leq 0.2$, $|A_2| < 0.5$, $|A_3| < 0.2$. To recover A_1^*, A_2^* and A_3^* we use the L-BFGS-B algorithm, (see [111] for a description). Forward differences are used to compute the gradients, resulting in a

N iter.	A_1	A_2	A_3	F	$\|\nabla F\|$
0	0.10100	$0.25250E$	0.12120	$7.83582(-6)$	$4.41312(-3)$
1	0.10117	$0.25077E$	0.11944	$1.96547(-7)$	$2.88478(-4)$
2	0.10117	$0.25066E$	0.11955	$1.38952(-7)$	$2.32250(-4)$
3	0.10105	$0.25012E$	0.12000	$2.26171(-8)$	$4.32507(-5)$
4	0.10095	$0.25010E$	0.12001	$1.87633(-8)$	$3.97138(-5)$
5	0.10002	$0.24999E$	0.12001	$5.81672(-11)$	$5.80355(-6)$
6	0.10000	$0.25000E$	0.12000	$1.00121(-15)$	$4.21271(-8)$

Table 10.3: Convergence results of L-FBGS-B algorithm for the inverse problem for locating a body in free space. At each iteration the cost function F and its gradient ∇F is computed from the numerical solution of the wave equation. The forward solver is implemented using the 5th order accurate Hybrid Hermite-DG overset grid method.

total $1 + 3$ simulations per iteration. Table 10.3 displays the convergence results in detail for the initial values at 1% of the exact solution, that are 0.101, 0.2525 and 0.1212 respectively. At the 6th iteration the values computed were 0.1, 0.25 and 0.12, accurate to the 8th digit. For the initial guesses with larger the 1% deviation from the exact solution, it becomes harder to converge to a global minimum. The minimization process would become more robust if more data is recorded at the receivers, for example by increasing the number of receivers, recording longer data traces or adding simulations with different initial conditions.

Although during the minimization process before each simulation the grids have to be regenerated this is inexpensive since the grid generation is local. Precisely in each new iteration the DG grid is adjusted by regenerating an annular grid based on the updated center location and radius.

10.5 Summary

We have presented overset high order numerical methods for numerical solution of the wave equation. The hybrid H-DG overset grid method combines the highly efficient Hermite method on Cartesian grids with a DG method to treat complex boundaries. To combine the methods the overset grids were used. The advantage of using the overset grids for complex boundary problems is the low computational cost that asymptotically approaches the cost of the Cartesian solver.

In this work we communicate solutions via L_2 projection and this procedure combined with the dissipative nature of the methods was observed to be sufficient to guarantee stability without the need to add any artificial dissipation.

Stability, accuracy and efficiency of the method were tested numerically. To test the stability in 1 dimension, we looked at the spectrum of the amplification matrix associated with the method. For CFL numbers < 0.75 for the Hermite method, the overall method was stable in all tested settings for grid sizes and orders of accuracy $3, 5$ and 7. In 1 and 2 dimensions we also tested the stability by displaying the error growth as a function of time for long times.

Finally, three example applications of the methods were presented. First, the wave scattering of the pentagonal object in free space was shown, demonstrating the use of the method for the problem with curvilinear boundary and free space boundary conditions. Second, a simulation with five round objects in free space was demonstrated. Finally the method was used to solve the inverse problem of locating a cylindrical underground body.

A future extension could be to improve the efficiency of the DG method used on the curvilinear body fitted grids by the use of an implicit timestepping method. This would allow the timesteps to be commensurate to those of the Hermite method

at a relatively low cost since the linear systems needed to be inverted would be essentially one dimensional. Another natural extension of this work would be to apply the techniques presented here to the elastic wave equation.

Appendix A

Real form of Chao FSM

Here we construct the fundamental solution matrix (FSM), Ψ, mentioned after Remark 13 in Section 5.1. Let $\Phi(\theta)$ be the principal solution matrix (PSM) defined as

$$\Phi' = A(\theta)\Phi, \quad \Phi(0) = I_{2d}, \quad A^{\mathrm{T}} J_{2d} + J_{2d}^{\mathrm{T}} A(\theta) = 0,$$

Since A is Hamiltonian the PSM is symplectic, i.e.,

$$\Phi^{\mathrm{T}} J_{2d} \Phi = J_{2d}.$$

A Floquet form for Φ is

$$\Phi(\theta) = P(\theta) e^{Q\theta}, P(0) = I$$

where P is 2π-periodic and where Q is defined in terms of the monodromy matrix M as

$$M := \Phi(2\pi) = e^{Q2\pi}.$$

We assume that the orbital motion defined by A is stable. Thus M has a full set of linearly independent eigenvectors, w_k, with the eigenvalues on the unit circle in the complex plane [112]. More precisely,

$$M w_k = \rho_k w_k, \ \rho_k = \exp \mathbf{i} 2\pi \nu_k.$$

Further, to avoid a resonance we assume that the ρ_k are distinct and since M is real we can choose the ν_k such that

$$0 < \nu_1 < \nu_3 < \nu_5 < 1/2, \quad \nu_{2l} = \nu_{2l-1}, \quad l = 1, 2, 3,$$

and w_k such that $w_{2l} = w_{2l-1}^*$.

Let $w_{2l-1} = a_{2l-1} + \mathbf{i}b_{2l-1}$ then, as we show below, w_{2l-1} can be normalized such that

$$a_{2l-1}^T J b_{2l-1} = \gamma_{2l-1}, \quad \gamma_{2l-1} = \pm 1.$$

Let $R = [a_1, b_1, a_3, b_3, a_5, b_5]$, then R satisfies

$$MR = Re^{2\pi\Lambda}, \; R^{\mathrm{T}} J_{2d} R = \Gamma J_{2d},$$

where $\Lambda = \mathrm{diag}(\nu_1 J_2, \nu_3 J_2, \nu_5 J_2)$ and $\Gamma = \mathrm{diag}(\gamma_1, \gamma_1, \gamma_3, \gamma_3, \gamma_5, \gamma_5)$. Thus we can take $Q = R\Lambda R^{-1}$ and the PSM can be written

$$\Phi(\theta) = P(\theta) R e^{\Lambda\theta} R^{-1}.$$

As promised, the FSM in Section 5.1 becomes

$$\Psi(\theta) = \Phi(\theta) R = \hat{\Psi}(\theta) e^{\Lambda\theta}, \quad \hat{\Psi}(\theta) := P(\theta) R. \tag{A.1}$$

The importance of this form will be shown in the Appendix B where it will be shown that $\mathcal{D}$ becomes block diagonal and $\mathcal{E}$ becomes diagonal after averaging.

The analogue of symplecticity for Ψ is

$$\Psi^{\mathrm{T}} J \Psi = \Gamma J, \tag{A.2}$$

thus

$$\Psi^{-1}(\theta) = -\Gamma J \Psi(\theta)^{\mathrm{T}} J.$$

This is quite useful, in practice, since the inverse of Ψ is easily calculated from its transpose. The periodic part of Ψ, $\hat{\Psi}$, also satisfies the above.

A.1 Normalization of the w_k and their signatures

See [112] for comprehensive analysis of the eigen-structure of symplectic maps. See also [113] for applications.

Lemma 1. Let w be a vector in $\mathbb{C}^d$. There exists a normalization parameter $r \neq 0$ such that, for $\hat{w} = w/r$

$$\hat{w}^{\mathrm{H}}\mathbf{i}J\hat{w} = \sigma, \quad \sigma = \pm 1, \tag{A.3}$$

Proof. Let $w = a + \mathbf{i}b$, then

$$w^{\mathrm{H}}\mathbf{i}Jw = (a - \mathbf{i}b)\mathbf{i}J(a + \mathbf{i}b) = (b^{\mathrm{T}}Ja - a^{\mathrm{T}}Jb) = -2(a^{\mathrm{T}}Jb).$$

So, let $r = \sqrt{2|a^{\mathrm{T}}Jb|}$, then

$$\hat{w}^{\mathrm{H}}\mathbf{i}J\hat{w} = \frac{w^{\mathrm{H}}\mathbf{i}Jw}{2|a^{\mathrm{T}}Jb|} = \pm 1.$$

Thus, we can always normalize w_{2l-1} such that (A.3) is true. □

The quantity σ in Lemma 1 is referred to as the *signature* of w in [113]. We emphasize that each $w \in \mathbb{C}^d \setminus \{0\}$ has a unique signature.

Corollary 1. Let $w_{2l-1} = a_{2l-1} + \mathbf{i}b_{2l-1}$, then w_{2l-1} can be normalized such that

$$a_{2l-1}^{T}Jb_{2l-1} = \gamma_{2l-1}, \quad \gamma_{2l-1} = \pm 1.$$

This normalization differs from the one in Lemma 1 by a factor of $\sqrt{2}$.

From now on we consider w that are normalized.

Lemma 2. The eigenvectors of M satisfy

$$w_k^{\mathrm{H}}\mathbf{i}Jw_j = w_k^{*\mathrm{H}}\mathbf{i}Jw_j = 0, \quad k \neq j, \quad k, j = 1, 3, 5.$$

Proof.

$$w_k^{\mathrm{H}}\mathbf{i}Jw_j = \frac{(Mw_k)^{\mathrm{H}}\mathbf{i}JMw_j}{\rho_j\rho_k^*} = \frac{w_j^H M^{\mathrm{T}}\mathbf{i}JMw_k}{\rho_j\rho_k^*} = \frac{w_j^H\mathbf{i}Jw_k}{\rho_j\rho_k^*}.$$

But $\rho_j\rho_k^* \neq 1$ which proves the first equality. The second equality is obtained similarly by replacing w_k with w_k^* and ρ_k^* with ρ_k. □

Corollary 2. The real and imaginary parts of the eigenvectors of M

$$a_k^{\mathrm{T}}Jb_j = b_k^{\mathrm{T}}Ja_j = a_j^{\mathrm{T}}Jb_k = b_j^{\mathrm{T}}Ja_k = 0, \quad k \neq j,\ k,j = 1,3,5.$$

We define R in (A.1) as a matrix composed of real and imaginary parts of the eigenvectors of M, i.e $R = [a_1, b_1, a_3, b_3, a_5, b_5]$. From Corollary 1 and 2 it follows that $a_k^{\mathrm{T}}Jb_j = \pm\delta_{j,k}$, where $\delta_{j,k}$ is Kronecker delta, and thus

$$R^{\mathrm{T}}JR = \Gamma J, \quad \Gamma = \mathrm{diag}(\gamma_1, \gamma_1, \gamma_3, \gamma_3, \gamma_5, \gamma_5),$$

as we mentioned above. Hence (A.1) followed by (A.2) holds.

Remark 21. *Let $W = [w_1, w_2, \ldots w_6]$, it follows that the w_k can be normalized so that*

$$W^{\mathrm{H}}JW = \mathrm{diag}(\sigma_1, \ldots, \sigma_6),$$

and a complex FSM for $A(\theta)$ is given by

$$\Psi_W(\theta) = \Phi(\theta)W$$

This is the FSM used in [46].

For our purpose a real FSM is needed. [1]

[1]The real form is needed since we need a real effective FPE

Appendix B

Calculation of the averaged drift and diffusion matrices

Here we show how to obtain the drift and diffusion matrices $\overline{\mathcal{D}}$ and $\overline{\mathcal{E}}$ for averaging defined by (5.7) and (5.8).

We start with the averaging of $\mathcal{D}(\theta) = \Psi^{-1}(\theta)\delta A(\theta)\Psi(\theta)$. Using (A.1) and (A.2) the drift matrix is rewritten as

$$\mathcal{D}(\theta) = -\Gamma J_{2d} e^{-\Lambda\theta} G(\theta) e^{\Lambda\theta},$$

where $G(\theta) = \hat{\Psi}^{\mathrm{T}}(\theta) J_{2d} \delta A(\theta) \hat{\Psi}(\theta)$. Assuming there is no resonance, we average $\mathcal{D}$ as follows

$$\overline{\mathcal{D}} = -\Gamma J_{2d} \overline{e^{-\Lambda\theta} \overline{G} e^{\Lambda\theta}}.$$

where the bar denotes θ-averaging as in Section 5.1, e.g.

$$\overline{G} = \lim_{T\to\infty} \frac{1}{T} \int_0^{\mathrm{T}} G(\theta)\, d\theta,$$

Appendix B. Calculation of the averaged drift and diffusion matrices

element-wise. Let $H(\theta) = e^{-\Lambda\theta}\overline{G}e^{\Lambda\theta}$, then writing $\overline{G}$ in 2×2 block form

$$\overline{G} = \begin{pmatrix} \overline{G}_{1,1} & \overline{G}_{1,3} & \overline{G}_{1,5} \\ \overline{G}_{3,1} & \overline{G}_{3,3} & \overline{G}_{3,5} \\ \overline{G}_{5,1} & \overline{G}_{5,3} & \overline{G}_{5,5} \end{pmatrix}$$

and similarly for H we obtain

$$H_{j,k} = e^{-J_2\nu_j\theta}\overline{G}_{j,k}e^{J_2\nu_k\theta}, \quad j,k = 1,3,5, \quad J_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

It is easy to show that $\overline{H}_{j,k} = 0$ if $j \neq k$ and thus $\overline{H}$ is block diagonal with

$$\overline{H}_{j,j}(\theta) = \overline{e^{-J_2\nu_j\theta}\overline{G}_{j,j}e^{J_2\nu_j\theta}}.$$

Now, using the identity

$$e^{J_2\nu_j\theta} = \cos(\nu_j\theta)I + \sin(\nu_j\theta)J_2,$$

$H_{j,j}$ averages to $J_2\overline{G}_{j,j} + \overline{G}_{j,j}J_2$, and thus $\overline{\mathcal{D}}$ becomes the block diagonal matrix

$$\overline{\mathcal{D}} = -\frac{1}{2}\Gamma \operatorname{diag}\left(J_2\overline{G}_{j,j} + \overline{G}_{j,j}J_2\right) = \frac{1}{2}\Gamma \operatorname{diag}\left(\begin{pmatrix} a_j & b_j \\ -b_j & a_j \end{pmatrix},\ j = 1,3,5\right).$$

Now we proceed with averaging of $\mathcal{E} = \Psi^{-1}B(\theta)B^{\mathrm{T}}(\theta)\Psi^{-\mathrm{T}}$. It can be written using (A.1) and (A.2) as

$$\begin{aligned} \mathcal{E}(\theta) =& -\Gamma J\Psi^{\mathrm{T}}(\theta)JB(\theta)B^{\mathrm{T}}(\theta)(-J^{\mathrm{T}}\Psi(\theta)J^{\mathrm{T}}\Gamma) \\ =& \Gamma Je^{-\Lambda\theta}\hat{\Psi}^{\mathrm{T}}(\theta)JB(\theta)B^{\mathrm{T}}(\theta)J\hat{\Psi}(\theta)e^{\Lambda\theta}J\Gamma \\ =& \Gamma Je^{-\Lambda\theta}G(\theta)e^{\Lambda\theta}J\Gamma. \end{aligned}$$

As for $\mathcal{D}$ the average of $\mathcal{E}$ becomes

$$\overline{\mathcal{E}} = \Gamma J\overline{e^{-\Lambda\theta}\overline{G}e^{\Lambda\theta}}J\Gamma = \Gamma J\overline{H(\theta)}J\Gamma,$$

where the second equality defines H. Note that H is of the same form as in previous calculation except now $\overline{G} = \overline{G}^{\mathrm{T}}$. Recall that $\overline{H_{j,k}(\theta)} = 0$, so that

$$\begin{aligned}\overline{\mathcal{E}} = \Gamma J \overline{H} J \Gamma =& \Gamma \begin{pmatrix} J_2 & 0 & 0 \\ 0 & J_2 & 0 \\ 0 & 0 & J_2 \end{pmatrix} \overline{\begin{pmatrix} H_{1,1} & 0 & 0 \\ 0 & H_{3,3} & 0 \\ 0 & 0 & H_{5,5} \end{pmatrix}} \begin{pmatrix} J_2 & 0 & 0 \\ 0 & J_2 & 0 \\ 0 & 0 & J_2 \end{pmatrix} \Gamma \\ =& \Gamma \begin{pmatrix} J_2 \overline{H}_{1,1} J_2 & 0 & 0 \\ 0 & J_2 \overline{H}_{3,3} J_2 & 0 \\ 0 & 0 & J_2 \overline{H}_{5,5} J_2 \end{pmatrix} \Gamma,\end{aligned}$$

where

$$H_{j,j} = \frac{1}{2}(G_{j,j} - J_2 G_{j,j} J_2), \quad H_{j,j} = H_{j,j}^{\mathrm{T}}, \quad i = 1, 3, 5.$$

Let $\overline{G}_{j,j} = \begin{pmatrix} \alpha_j & \beta_j \\ \beta_j & \gamma_j \end{pmatrix}$, then $J_2 \overline{G}_{j,j} J_2 = \begin{pmatrix} -\gamma_j & \beta_j \\ \beta_j & -\alpha_j \end{pmatrix}$, and thus

$$\overline{H}_{j,j} = \frac{1}{2}(\alpha_j + \gamma_j) I.$$

So $\overline{\mathcal{E}}$ is a diagonal matrix that we write as

$$\overline{\mathcal{E}} = \mathrm{diag}(\mathcal{E}_1, \mathcal{E}_1, \mathcal{E}_3, \mathcal{E}_3, \mathcal{E}_5, \mathcal{E}_5), \quad i = 1, 3, 5,$$

where $\mathcal{E}_j = \frac{1}{2}(\alpha_j + \gamma_j)$.

References

[1] BNL, Electron–Ion Collider, https://www.bnl.gov/EIC, accessed on 2020-09-17 (2012).

[2] CERN, The FCC-ee design study, http://fcc-ee.web.cern.ch, accessed on 2020-09-17 (2014).

[3] IHEP, Circular Electron Positron Collider, http://cepc.ihep.ac.cn, accessed on 2020-09-17 (2012).

[4] E. Merzbacher, Quantum Mechanics, 3rd Edition, Wiley, 1998.

[5] A. Blondel, P. Janot, J. Wenninger, R. Aßmann, S. Aumon, P. Azzurri, D. P. Barber, M. Benedikt, A. V. Bogomyagkov, E. Gianfelice-Wendt, et al., Polarization and Centre-of-mass Energy Calibration at FCC-ee, arXiv preprint arXiv:1909.12245 (2019).

[6] J. Jackson, Classical Electrodynamics, 3rd Edition, Wiley, 1998.

[7] M. Sands, Physics of electron storage rings: An introduction., Tech. rep., Stanford Linear Accelerator Center, Calif. (1970).

[8] J. A. Ellison, H. Mais, G. Ripken, Orbital eigen-analysis for electron storage rings, in: A. W. Chao, K. H. Mess, M. Tigner, F. Zimmermann (Eds.), Handbook of Accelerator Physics and Engineering, 2nd Edition, World Scientific, 2013, pp. 68–71.

[9] A. Sokolov, I. Ternov, On polarization and spin effects in synchrotron radiation theory, Sov. Phys. Dokl. 8 (1964) 1203, see also [114].

[10] Y. Derbenev, A. Kondratenko, Polarization kinetics of particles in storage rings, Sov. Phys. JETP 37 (1973) 968.

[11] S. Mane, PTC SPIN: benchmark tests for analytically solvable models, Tech. Rep. KEK Report 2009-8, KEK (September 2009).

[12] Y. Derbenev, A. Kondratenko, Relaxation and equilibrium state of electrons in storage rings, Sov. Phys. Dokl. 19 (1975) 438.

[13] D. Barber, G. Ripken, Radiative Polarization, Computer Algorithms and Spin Matching in Electron Storage Rings, in: A. W. Chao, M. Tigner (Eds.), Handbook of Accelerator Physics and Engineering, 1st Edition, World Scientific, 2006, pp. 174–178, 3rd printing. See also `https://www.desy.de/~mpybar/`.

[14] K. Heinemann, D. Appelö, D. P. Barber, O. Beznosov, J. Ellison, Re-evaluation of Spin-Orbit Dynamics of Polarized e^+e^- Beams in High Energy Circular Accelerators and Storage Rings: Bloch equation approach, Int.J.Mod.Phys.A 35 (2041003) (2020).

[15] D. P. Barber, J. A. Ellison, K. Heinemann, Quasiperiodic spin–orbit motion and spin tunes in storage rings, Physical Review Special Topics-Accelerators and Beams 7 (12) (2004) 124002.

[16] V. Baier, V. Katkov, V. Strakhovenko, Kinetics of radiative polarization, Sov. Phys. JETP 31 (1970) 908.

[17] F. Bloch, Nuclear induction, Phys. Rev. 70 (1946) 460.

[18] K. Heinemann, D. Appelö, D. P. Barber, O. Beznosov, J. Ellison, The Bloch equation for spin dynamics in electron storage rings: computational and theoretical aspects, Int. J. Mod. Phys. A34 (1942032), see also `https://arxiv.org/abs/2007.14613` (2019).

[19] E. Gianfelice, Self polarization in storage rings, Proceedings of Science (2018).

[20] G. Bassi, J. A. Ellison, K. Heinemann, R. Warnock, Transformation of phase space densities under the coordinate changes of accelerator physics, Physical Review Special Topics-Accelerators and Beams 13 (10) (2010) 104403.

[21] K. Heinemann, D. P. Barber, O. Beznosov, J. A. Ellison, Cartesian to beam coordinates, `https://math.unm.edu/~ellison/GWpapers/DK75SDELab.pdf`, accessed on 2020-09-18. Work in progress (2020).

[22] D. Sagan, The BMAD Reference Manual (2005).

[23] S. Mane, Exact solution of the Derbenev-Kondratenko n axis for a model with one resonance, Tech. rep., Fermi National Accelerator Lab., Batavia, IL (USA) (1988).

[24] I. I. Rabi, N. F. Ramsey, J. Schwinger, Use of rotating coordinates in magnetic resonance problems, Reviews of Modern Physics 26 (2) (1954) 167.

[25] L. Arnold, Stochastic differential equations, New York (1974).

[26] T. Gard, Introduction to Stochastic Differential Equations, Dekker, 1988.

[27] C. Gardiner, Handbook of stochastic methods for physics, chemistry and the natural sciences, 4th Edition, Springer, 2009.

[28] O. Cakir, V. Cetinkaya, R. Ciftci, A. Ciftci, S. Turkoz, I. Cakir, H. Karadeniz, A. Akay, M. Sahin, S. Sultansoy, et al., A large hadron electron collider at cern, Journal of Physics G: Nuclear and Particle Physics 39 (7) (2012) 075001–075001.

[29] M. Farkhondeh, V. Ptitsyn, erhic zeroth–order design report, Tech. rep., Brookhaven National Laboratory (BNL) Relativistic Heavy Ion Collider (2004).

[30] J. Kewisch, Simulation of electron spin depolarisation with the computer code sitros, Tech. rep., Deutsches Elektronen-Synchrotron (DESY) (1983).

[31] F. Méot, The ray-tracing code Zgoubi, NIM A 427 (1999) 353–356, this code is a plain integrator for rings without special enforcement of symplecticity.

[32] E. Forest, Y. Nogiwa, The FPP and PTC Libraries, in: Proceedings of ICAP 2006, Joint Accelerator Conferences Website, Chamonix, France, 2006, p. 21.

[33] E. Forest, From Tracking Code to Analysis, Springer, 2016.

[34] D. Sagan, Bmad, a subroutine library for relativistic charged-particle dynamics. URL `https://www.classe.cornell.edu/bmad`

[35] J. M. Jowett, Non-linear Dissipative Phenomena in Electron Storage Rings, in: J. M. Jowett, M. Month (Eds.), Nonlinear Dynamics Aspects of Particle Accelerators, Springer-Verlag, Berlin, 1986, p. 343.

[36] K. Heinemann, D. Barber, The Semiclassical Foldy-Wouthuysen Transformation and the Derivation of the Bloch Equation for Spin-1/2 Polarised Beams Using Wigner Functions, in: P. Chen (Ed.), Advanced ICFA Beam Dynamics Workshop on Quantum Aspects of Beam Physics, World Scientific, Monterey, CA, 1998.

[37] K. Heinemann, D. P. Barber, Spin transport, spin diffusion and bloch equations in electron storage rings, Nucl. Instr. Meth. A463, A469 (1-2) (2001) 62–67, 294.

[38] D. Barber, K. Heinemann, H. Mais, G. Ripken, A Fokker-Planck treatment of stochastic particle motion within the framework of a fully coupled six-dimensional formalism for electron – positron storage rings including classical spin motion in linear approximation, DESY-91-146 (1991).

[39] A. Lasota, M. C. Mackey, Chaos, fractals, and noise: stochastic aspects of dynamics, Vol. 97, Springer Science & Business Media, 2013.

[40] E. D. Courant, H. S. Snyder, Theory of the alternating-gradient synchrotron, Annals of physics 3 (1) (1958) 1–48.

[41] T. Aniel, J. Laclare, G. Leleux, A. Nakach, A. Ropert, Polarized particles at saturne, Le Journal de Physique Colloques 46 (C2) (1985) C2–499.

[42] J. A. Ellison, K. Heinemann, Polarization fields and phase space densities in storage rings: Stroboscopic averaging and the ergodic theorem, Physica D: Nonlinear Phenomena 234 (2) (2007) 131–149.

[43] G. H. Hoffstaetter, High energy polarized proton beams: a modern view, Vol. 218, Springer, 2009.

[44] M. Vogt, Bounds on the maximum attainable equilibrium spin polarization of protons at high energy in hera, Ph.D. thesis, University of Hamburg/DESY (2000).

[45] D. Sagan, A superconvergent algorithm for invariant spin field stroboscopic calculations, in: 9th Int. Particle Accelerator Conf.(IPAC'18), Vancouver, BC, Canada, April 29-May 4, 2018, JACOW Publishing, Geneva, Switzerland, 2018, pp. 145–148.

[46] J. A. Ellison, H. Mais, G. Ripken, Orbital eigen-analysis for electron storage rings, in: A. W. Chao, M. Tigner (Eds.), Handbook of Accelerator Physics and Engineering, 1st Edition, World Scientific, 1999, pp. 69–71, 3rd printing.

[47] J. A. Ellison, H.-J. Shih, The method of averaging in beam dynamics, Accelerator Physics Lectures at the Superconducting Super Collider (326) (1995) 590–632.

[48] J. A. Ellison, K. A. Heinemann, M. Vogt, M. Gooden, Planar undulator motion excited by a fixed traveling wave: Quasiperiodic Averaging, normal forms and the FEL pendulum, Phys. Rev. ST Accel. Beams 16 (090702) (2013).

[49] J. Sanders, F. Verhulst, J. Murdock, Averaging Methods in Nonlinear Dynamical Systems, 2nd Edition, Springer, New York, 2007.

[50] J. Murdock, Perturbations: Theory and Methods, SIAM, Philadelphia, 1999.

[51] R. Cogburn, J. Ellison, A stochastic theory of adiabatic invariance, Communications of Mathematical Physics 148 (1992) 97–126.

[52] R. Graham, Solution of fokker planck equations with and without manifest detailed balance, Zeitschrift für Physik B Condensed Matter 40 (1) (1980) 149–155.

[53] J. A. Ellison, Personal communication (2020).

[54] D. P. Barber, M. Böge, H.-D. Bremer, R. Brinkmann, W. Brückner, M. Düren, E. Gianfelice-Wendt, C. Großhauser, H. Kaiser, R. Klanner, et al., The first achievement of longitudinal spin polarization in a high energy electron storage ring, Physics Letters B 343 (1-4) (1995) 436–443.

[55] C. W. Gardiner, et al., Handbook of stochastic methods, Vol. 3, springer Berlin, 1985.

[56] H. Risken, C. Braun, The Fokker-Planck Equation, ApOpt 28 (20) (1989) 4496–4497.

[57] T. D. Frank, Linear and nonlinear Fokker-Planck equations, Synergetics (2020) 149–182.

[58] P. Kumar, S. Narayanan, Solution of Fokker-Planck equation by finite element and finite difference methods for nonlinear systems, Sadhana 31 (4) (2006) 445–461.

[59] L. Pichler, A. Masud, L. A. Bergman, Numerical solution of the Fokker–Planck equation by finite difference and finite element methods—a comparative study, in: Computational Methods in Stochastic Dynamics, Springer, 2013, pp. 69–85.

[60] J. Chang, G. Cooper, A practical difference scheme for Fokker-Planck equations, Journal of Computational Physics 6 (1) (1970) 1–16.

[61] G. W. Harrison, Numerical solution of the Fokker Planck equation using moving finite elements, Numerical methods for Partial differential Equations 4 (3) (1988) 219–232.

[62] M. Dehghan, V. Mohammadi, The numerical solution of Fokker–Planck equation with radial basis functions (RBFs) based on the meshless technique of Kansa's approach and Galerkin method, Engineering Analysis with Boundary Elements 47 (2014) 38–63.

[63] J. Reif, R. Barakat, Numerical solution of the Fokker-Planck equation via Chebyschev polynomial approximations with reference to first passage time probability density functions, Journal of Computational Physics 23 (4) (1977) 425–445.

[64] D. J. Knezevic, E. Süli, Spectral Galerkin approximation of Fokker-Planck equations with unbounded drift, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique 43 (3) (2009) 445–485.

[65] J. Fok, B. Guo, T. Tang, Combined Hermite spectral-finite difference method for the Fokker-Planck equation, Mathematics of computation 71 (240) (2002) 1497–1528.

[66] S. Wojtkiewicz, L. Bergman, Numerical solution of high dimensional Fokker-Planck equations, in: 8th ASCE Specialty Conference on Probablistic Mechanics and Structural Reliability, Notre Dame, IN, USA, Citeseer, 2000.

[67] E. Coutsias, T. Hagstrom, J. Hesthaven, D. Torres, Integration preconditioners for differential operators in spectral-methods, in: Proceedings of the Third International Conference on Spectral and High Order Methods, Houston, TX, 1996, pp. 21–38.

[68] C. Lanczos, Applied analysis, Courier Corporation, 1988.

[69] L. N. Trefethen, Spectral methods in MATLAB, SIAM, 2000.

[70] C. A. Kennedy, M. H. Carpenter, Additive Runge–Kutta schemes for convection–diffusion–reaction equations, Applied numerical mathematics 44 (1-2) (2003) 139–181.

[71] S. R. Lau, Direct, low-memory, spectral solution of harmonic problems on a block at near optimal complexity, in preparation.

[72] O. Beznosov, J. A. Ellison, K. A. Heinemann, D. P. Barber, D. Appelö, Spin dynamics in modern electron storage rings: Computational aspects, in: Proceedings of 13th International Computational Accelerator Physics Conference, Key West, Florida, 2018, pp. 146–150.

[73] K. Heinemann, Some models of spin coherence and decoherence in storage rings, arXiv preprint physics/9709025 (1997).

[74] D. P. Barber, K. Heinemann, Spin decoherence in electron storage rings—more from a simple model, arXiv preprint arXiv:1508.05318 (2015).

[75] K. Mattsson, J. Nordström, Summation by parts operators for finite difference approximations of second derivatives, Journal of Computational Physics 199 (2004) 503–540.

[76] K. Virta, K. Mattsson, Acoustic wave propagation in complicated geometries and heterogeneous media, Journal of Scientific Computing 61 (1) (2014) 90–118.

[77] S. Wang, K. Virta, G. Kreiss, High order finite difference methods for the wave equation with non-conforming grid interfaces, Journal of Scientific Computing 68 (3) (2016) 1002–1028.

[78] N. A. Petersson, B. Sjögreen, High order accurate finite difference modeling of seismo-acoustic wave propagation in a moving atmosphere and a heterogeneous earth model coupled across a realistic topography, Journal of Scientific Computing 74 (1) (2018) 290–323.

[79] T. Hagstrom, G. Hagstrom, Grid stabilization of high-order one-sided differencing II: Second-order wave equations, Journal of Computational Physics 231 (23) (2012) 7907 – 7931.

[80] D. Appelö, N. Petersson, A fourth-order accurate embedded boundary method for the wave equation, SIAM Journal on Scientific Computing 34 (6) (2012) A2982–A3008.

[81] O. P. Bruno, M. Lyon, High-order unconditionally stable FC-AD solvers for general smooth domains I. Basic elements, Journal of Computational Physics 229 (6) (2010) 2009 – 2033.

[82] M. Lyon, O. P. Bruno, High-order unconditionally stable FC-AD solvers for general smooth domains II. Elliptic, parabolic and hyperbolic PDEs; theoretical considerations, Journal of Computational Physics 229 (9) (2010) 3358 – 3381.

[83] J.-R. Li, L. Greengard, High order marching schemes for the wave equation in complex geometry, Journal of Computational Physics 198 (1) (2004) 295 – 309.

[84] S. Wandzura, Stable, high-order discretization for evolution of the wave equation in 2 + 1 dimensions, Journal of Computational Physics 199 (2) (2004) 763 – 775.

[85] L. Wilcox, G. Stadler, C. Burstedde, O. Ghattas, A high-order discontinuous Galerkin method for wave propagation through coupled elastic-acoustic media, Journal of Computational Physics 229 (24) (2010) 9373–9396.

[86] D. Appelö, T. Hagstrom, A new discontinuous Galerkin formulation for wave equations in second order form, SIAM Journal On Numerical Analysis 53 (6) (2015) 2705–2726.

[87] C.-S. Chou, C.-W. Shu, Y. Xing, Optimal energy conserving local discontinuous Galerkin methods for second-order wave equation in heterogeneous media, Journal of Computational Physics 272 (2014) 88 – 107.

[88] E. T. Chung, B. Engquist, Optimal discontinuous Galerkin methods for wave propagation, SIAM Journal on Numerical Analysis 44 (5) (2006) 2131–2158.

[89] E. T. Chung, B. Engquist, Optimal discontinuous Galerkin methods for the acoustic wave equation in higher dimensions, SIAM Journal on Numerical Analysis 47 (5) (2009) 3820–3848.

[90] M. J. Grote, A. Schneebeli, D. Schötzau, Discontinuous Galerkin finite element method for the wave equation, SIAM Journal on Numerical Analysis 44 (6) (2006) 2408–2431.

[91] N. C. Nguyen, J. Peraire, B. Cockburn, High-order implicit hybridizable discontinuous Galerkin methods for acoustics and elastodynamics, Journal of Computational Physics 230 (10) (2011) 3695 – 3718.

[92] M. Stanglmeier, N. C. Nguyen, J. Peraire, B. Cockburn, An explicit hybridizable discontinuous Galerkin method for the acoustic wave equation, Computer Methods in Applied Mechanics and Engineering 300 (2016) 748 – 769.

[93] S. Sticko, G. Kreiss, A stabilized Nitsche cut element method for the wave equation, Computer Methods in Applied Mechanics and Engineering 309 (2016) 364 – 387.

[94] S. Sticko, G. Kreiss, Higher order cut finite elements for the wave equation, arXiv preprint arXiv:1608.03107 (2016).

[95] J. W. Banks, T. Hagstrom, On galerkin difference methods, Journal of Computational Physics 313 (2016) 310 – 327.

[96] G. Chesshire, W. D. Henshaw, Composite overlapping meshes for the solution of partial differential equations, Journal of Computational Physics 90 (1) (1990) 1–64.

[97] W. D. Henshaw, Ogen: An overlapping grid generator for Overture, Research Report UCRL-MA-132237, Lawrence Livermore National Laboratory (1998).

[98] W. D. Henshaw, A high-order accurate parallel solver for Maxwell's equations on overlapping grids, SIAM Journal on Scientific Computing 28 (5) (2006) 1730–1765.

[99] J. B. Angel, J. W. Banks, W. D. Henshaw, High-order upwind schemes for the wave equation on overlapping grids: Maxwell's equations in second-order form, Journal of Computational Physics 352 (2018) 534 – 567.

[100] J. W. Banks, W. D. Henshaw, Upwind schemes for the wave equation in second-order form, Journal of Computational Physics 231 (17) (2012) 5854 – 5889.

[101] D. Appelö, J. W. Banks, W. D. Henshaw, D. W. Schwendeman, Numerical methods for solid mechanics on overlapping grids: Linear elasticity, Journal of Computational Physics 231 (18) (2012) 6012–6050.

[102] D. Appelö, T. Hagstrom, A. Vargas, Hermite methods for the scalar wave equation, SIAM Journal on Scientific Computing 40 (6) (2018) A3902–A3927.

[103] T. Warburton, T. Hagstrom, Taming the CFL number for discontinuous Galerkin methods on structured meshes, SIAM J. Numerical Analysis 46 (6) (2008) 3151–3180.

[104] T. Hagstrom, D. Appelö, Solving PDEs with Hermite interpolation, in: Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2014, Springer, 2015, pp. 31–49.

[105] J. Goodrich, T. Hagstrom, J. Lorenz, Hermite methods for hyperbolic initial-boundary value problems, Mathematics of computation 75 (254) (2006) 595–630.

[106] D. Appelö, T. Hagstrom, An energy-based discontinuous Galerkin discretization of the elastic wave equation in second order form, Computer Methods in Applied Mechanics and Engineering 338 (2018) 362–391.

[107] D. Appelö, S. Wang, An energy-based discontinuous galerkin method for coupled elasto-acoustic wave equations in second-order form, International Journal for Numerical Methods in Engineering 119 (7) (2019) 618–638.

[108] Å. Björck, V. Pereyra, Solution of Vandermonde systems of equations, Mathematics of Computation 24 (112) (1970) 893–903.

[109] G. Dahlquist, Å. Björck, Numerical methods in scientific computing, Society for Industrial and Applied Mathematics, 2008.

[110] J. C. Strikwerda, Finite difference schemes and partial differential equations, Vol. 88, SIAM, 2004.

[111] R. H. Byrd, P. Lu, J. Nocedal, C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM Journal on Scientific Computing 16 (5) (1995) 1190–1208.

[112] K. Meyer, G. Hall, D. Offin, Introduction to Hamiltonian Dynamical Systems and the N-Body Problem, 2nd Edition, Springer, New York, 2009.

[113] A. J. Dragt, Lie methods for nonlinear dynamics with applications to accelerator physics, `http://www.physics.umd.edu/dsat/dsatliemethods.htm` (2011).

[114] A. A. Sokolov, I. M. Ternov, Radiation from relativistic electrons, AIP, 1986.