Inertial Manifolds and Nonlinear Galerkin Methods

Denis C. Kovacs

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Mathematics

Jeff Borggaard, Chair
Christopher Beattie
Serkan Gugercin
Traian Iliescu

December 9, 2005
Blacksburg, Virginia

Inertial Manifolds and Nonlinear Galerkin Methods

Denis C. Kovacs

(ABSTRACT)

Nonlinear Galerkin methods utilize approximate inertial manifolds to reduce the spatial error of the standard Galerkin method. For certain scenarios, where a rough forcing term is used, a simple postprocessing step yields the same improvements that can be observed with nonlinear Galerkin. We show that this improvement is mainly due to the information about the forcing term that is neglected by standard Galerkin. Moreover, we construct a simple postprocessing scheme that uses only this neglected information but gives the same increase in accuracy as nonlinear or postprocessed Galerkin methods.

To the interested reader.

Acknowledgments

I want to express my thanks to my advisors, Prof. Beattie and Prof. Borggaard and my committee for a topic that led me through many fascinating fields of mathematics, including functional analysis, the theory of PDEs and (infinite-dimensional) dynamical systems, numerical analysis (especially spectral methods and efficient time integrators), and reduced order modeling. They always had time for my questions and did a tremendous job in encouraging me whenever nonlinear Galerkin gave disappointing results. I also want to thank them for granting me a Research Assistantship for the Summer and Fall 2005, which gave me the possibility to focus on my thesis work, and for the opportunity to visit the SIAM Annual Meeting 2005 in New Orleans.

Thanks to Prof. Adjerid, Prof. Hagedorn, Prof. Haskell and Prof. Renardy for answering random questions that I had along the way, and to Jeff, Alexey and Weston for proofreading.

The student atmosphere at the Math Department is amazing, I have never experienced such a friendly school environment before. It made studying math a pleasure and I made true friends here. I will miss D2 and the great discussions we had over lunch, biking, ATHF [27], the DVD nights until the natural threshold, geek talk etc.

TeXmacs (`www.texmacs.org`) greatly reduced the time I spent typing and debugging the thesis.

Last but not least, I want to thank my family for their moral support from a distance and apologize for not keeping in touch as much as I should have.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Dissipative evolution equations are time-dependent, typically nonlinear and autonomous partial differential equations (PDEs) that include a linear damping term. We can write them as an abstract initial value problem

$$u_t + Au + N(u) \quad = \quad 0,$$

where $u \in \mathcal{H}$ and $\mathcal{H}$ a (separable) Hilbert space. A is a linear dissipative term and $N(u)$ is the nonlinear part. Possible boundary conditions are assumed to be enforced by $\mathcal{H}$.

Many PDEs obtained from models of chemical processes or physical phenomena can be written in the above form: popular examples are reaction-diffusion-equations, pattern formation equations like the Cahn-Hilliard equation, the Kuramoto-Sivashinsky equation and Burgers' and the Navier-Stokes equation.

The numerical solution of a dissipative evolution equation is challenging in several ways. To be numerically tractable, the equation has to be **discretized**, both in space and time:

- **Spatial Discretization** After introducing a basis of the underlying Hilbert space, the PDE can be written as an infinite-dimensional (abstract) ODE. The standard approach of Galerkin approximations is to project the equation onto a finite-dimensional subspace. There are many competing criteria in the choice of this subspace. A natural criterion is the approximation quality: clearly it is desirable to approximate the solution accurately by as few basis functions as possible. However, for fast solvers, other criteria like orthonormality and little overlap of the domains of basis functions can be just as important.

- **Temporal Discretization** The ODEs obtained by spatial discretization are solved numerically by approximating the solution on a discrete temporal grid. When integrating ODEs obtained by discretizations of dissipative PDEs, one is faced with the problem of stiffness. Explicit time integrators are impractical in this case, as the time

> step they require for stability is too small. The remedy is either to transform the ODE in some way to get rid of stiffness, or to use implicit solvers, that do not suffer from the severe restrictions on the time step size, but are more costly than explicit methods (Chapter 6.3).

The main focus of this thesis is **spatial discretization**. Standard Galerkin methods project the PDE onto a “flat” subspace spanned by a finite number of basis functions. Due to this projection the component in the orthogonal complement is completely neglected.

The central idea of nonlinear Galerkin methods is not to ignore the orthogonal complement, but rather tie it to the finite-dimensional component. The effect of this is a projection of the equations onto a manifold instead of a flat subspace.

The motivation for this approach comes from infinite-dimensional dynamical systems theory. Under certain conditions (Section 4.3) it can be shown that a finitely parametrized manifold (an *inertial manifold*) exists with two remarkable properties: trajectories starting on the manifold stay on it for all positive times (the manifold is said to be **invariant** under the flow), and all trajectories starting off the manifold are attracted to it at an exponential rate.

Unfortunately, the existence proof is non-constructive, and we are left clueless on where exactly the inertial manifold is located. However, one can construct *approximate inertial manifolds*, that have provably better quality (Chapter 4.7) than the flat projection.

The original thesis topic was to combine nonlinear Galerkin with a basis coming from a model reduction technique called *proper orthogonal decomposition* (POD) on Burgers’ equation as a way to improve the quality of the reduced model. The results were disappointing, so we ran nonlinear Galerkin on the full model (with a finite element basis) to observe the improvements over standard Galerkin. Again, the results did not match with the promising results found in literature. At this point, we decided to take a closer look on the test scenarios in which nonlinear Galerkin was reported to outperform the standard Galerkin method.

There are several issues that arise with the implementation of nonlinear Galerkin methods (Chapter 5.2-5.6). The (approximate) inertial manifold theory can be handled most easily in an eigenbasis of the dissipative operator (Chapter 2.4), which might not be known explicitly. In the cases where it is known, it is typically a Fourier basis. The resulting spectral methods are known to have exponential decay of coefficients for smooth solutions. One has to create very special circumstances for the nonlinear Galerkin method to provide a noticably higher accuracy than the standard Galerkin method. Also, the computational cost is higher for nonlinear Galerkin methods. Thus an increase in accuracy alone is not enough to establish an advantage over standard Galerkin.

A variation on the idea of nonlinear Galerkin is postprocessing the solution of standard Galerkin by lifting it onto the approximate inertial manifold at final time. Indeed, some of the examples that showed a significant improvement in the rate of convergence of nonlinear Galerkin indicated the same improvements with this postprocessing step. The main reason

(Chapter 6.5) is the combination of trivial dynamics (into an equilibrium) and the choice of a rough forcing term, that is badly approximated in the finite-dimensional subspace spanned by the first n Fourier modes. However, as the forcing term is known in advance, its Fourier coeffiecients are known to arbitrary accuracy, and the approximate inertial manifolds exploit this additional information to achieve a higher accuracy. In Chapter 6.4 we construct a flat manifold (i.e. an affine subspace) that uses *only* the additional information on the forcing term, and show that postprocessing with this manifold gives the same improvements observed in the literature.

Chapter 2

Functional Analysis and PDE Toolbox

2.1 Definitions and Conventions

We use the standard definitions of vector spaces, norms, inner products, bounded sets, closures, normed spaces, completeness, Banach and Hilbert spaces and linear operators.

A bounded linear operator A between two normed spaces X and Y is a linear operator for which there is an $M > 0$ such that $\|Ax\|_Y \leq M\|x\|_X \quad \forall x \in X$. For unbounded linear operators there is no such M.

Throughout this thesis all vector spaces are $\mathbb{R}$-vector spaces, i.e. scalars are always be real-valued. We also assume Ω to be an open, bounded, connected set in $\mathbb{R}^m$ with a smooth boundary.

A functional maps a Banach space X into the real numbers. The space of all bounded linear functionals on a normed space X is called its dual space, denoted by X^*. One can identify elements of a Hilbert space with elements of its dual space, in a way that keeps their norms identical. This is the result of the famous Riesz lemma:

Lemma 1 ***Riesz Lemma***

Let $\mathcal{H}$ a Hilbert space. For each $l \in \mathcal{H}^*$ there is a unique $y \in \mathcal{H}$ such that

$$l(x) \;\; = \;\; (y, x) \quad \forall x \in \mathcal{H}.$$

Furthermore, $\|y\|_{\mathcal{H}} = \|l\|_{\mathcal{H}^*}$.

2.2 Compactness

A set U in a normed space is compact if every open cover has a finite subcover. In this case every sequence in U has a convergent subsequence. In finite dimensions a set is compact if and only if it is closed and bounded. In infinite dimensions this is no longer true. So compactness theorems are among the most important tools in the infinite dimensional setting.

A compact linear operator is an operator that maps bounded sets into precompact sets, i.e. into sets whose closure is compact. Put differently, a compact operator transforms bounded sequences into sequences that have a converging subsequence. If the range of a bounded operator is finite-dimensional, then it is compact. Moreover, A is compact if and only if there is a sequence $\{A_n\}$ of finite rank operators that converge to A in the operator norm.

Among the most frequently used compact operators are compact embeddings: A space X is compactly embedded in another space Y (or in short $X \subset\subset Y$), if the injection of X into Y, $\iota : X \to Y, \iota(x) = x$ is compact. One of the most frequently used compact embeddings are those of Sobolev spaces (most notably $H_0^1(\Omega) \subset\subset L^2(\Omega)$), these results are cited in Section 2.5.

There are several compactness theorems that guarantee converging subsequences under certain conditions. One of the most well-known is the Arzela-Ascoli theorem.

Theorem 1 ***Arzela-Ascoli***

Let X be a compact subset of $\mathbb{R}^p$, and let $\{f_n\}$ be a sequence of continuous functions from X into $\mathbb{R}^q$. If f_n is uniformly bounded, (there exists an $M > 0$ such that $\|f_n\| \leq M \, \forall n$) and equicontinuous ($\forall \varepsilon > 0 \quad \exists \delta > 0$ such that $|x - y| \leq \delta \Rightarrow |f_n(x) - f_n(y)| \leq \varepsilon$ independent of n), then there is a subsequence that is uniformly converging on X.

We need an extension of this theorem ([32]) to continuous functions on Banach spaces for the existence proof of inertial manifolds in Chapter 4:

Theorem 2 ***Arzela-Ascoli, Banach Space Version***

Let (X, d_1) be a compact metric space, and (Y, d_2) a complete metric space. Define (C, d_3) to be the (complete) metric space of continuous functions from X to Y. Let $A \subset (C, d_3)$. The following are equivalent:

- *$\overline{A}$ is compact*
- *A is pointwise compact ($\overline{\{f(x) : f \in A\}}$ is compact $\forall x \in X$) and equicontinuous.*

2.3 Outline of the rest of the chapter

Virtually every paper on inertial manifolds or nonlinear Galerkin methods assumes the following:

> "Let A be a positive, self-adjoint, unbounded operator with compact inverse."

Some mention the (negative) Laplacian $-\Delta$ (or more precisely its closure) as an example, and in fact most dissipative PDEs considered include either the Laplacian or the Bi-Laplacian as the dissipative linear term. The above characterization guarantees an orthonormal basis of the underlying Hilbert space consisting of eigenfunctions ω_k of A, corresponding to positive, increasing eigenvalues $\lambda_k \to \infty$. This basis "diagonalizes" A:

$$A \quad = \quad \sum_{k=1}^{\infty} \lambda_k |\omega_k\rangle\langle\omega_k|.$$

This remarkable property can be exploited both on the theoretical and on the computational side. First and foremost, it is easy to see which subspaces are left invariant by A: All spaces $X = \text{span}\{\omega_k : k \in S\}$ with $S \subset \mathbb{N}$ are A-invariant. Secondly, it makes it easy to find bounds on the norm of Au for certain u: In Chapter 4 we split $\mathcal{H}$ into two orthogonal subspaces: $P_n\mathcal{H} = \text{span}\{\omega_k : k \leq n\}$ and its orthogonal complement $Q_n\mathcal{H}$. If $p \in P_n\mathcal{H}$, we can easily obtain an upper bound on $\|Ap\|$ by

$$\|Ap\| \quad = \quad \| \sum_{k=1}^{n} \lambda_k |\omega_k\rangle\langle\omega_k|p\rangle \| \leq \|\lambda_n \sum_{k=1}^{n} |\omega_k\rangle\langle\omega_k|p\rangle \| = \lambda_n \|p\|.$$

Similarly, if $q \in Q_n\mathcal{H}$, we can easily find a lower bound on $\|Aq\|$ by

$$\|Aq\| \quad \geq \quad \lambda_{n+1}\|q\|.$$

These bounds are applied e.g. in Section 4.4 for the derivation of the spectral gap condition.

Another advantage of a basis that diagonalizes the dissipative term is connected to the stiffness of discretizations of disspative PDEs. The diagonal form of the finite-dimensional approximation of A enables us to exactly integrate the dissipative linear part by a change of variables. This **integrating factor** method allows us to compute solutions using large numbers of modes. The diagonal form can also be used to speed up implicit integrators like `ode15s` by using diagonal approximations of the Jacobian.

The rest of this chapter is concerned with understanding how the above assumptions guarantee the existence of an orthonormal basis of eigenvectors, and why the Laplacian satisfies these assumptions.

2.4 Eigenfunction Expansions

In finite dimensions we can define eigenpairs of a matrix by the identity

$$Aw_j \;=\; \lambda_j w_j, \tag{2.1}$$

We can characterize the eigenvalues of A equivalently by requiring $\det(A - \lambda I) = 0$ or $\mathrm{Ker}(A - \lambda I) \neq \{0\}$ or $\mathrm{Ran}(A - \lambda I) \neq \mathbb{R}^m$. In infinite dimensions we can not give meaning to the determinant, and the other three criteria are also no longer equivalent. For a bounded operator $A : \mathcal{H} \to \mathcal{H}$, this problem is solved by introducing a new set, the **spectrum** $\sigma(A)$, which in turn is the complement of the **resolvent**. The resolvent is the set of λ for which $(A - \lambda I)$ has a bounded inverse. Of course every λ that satisfies (2.1) is in $\sigma(A)$. But in general there are other elements of $\sigma(A)$ that do not satisfy (2.1). The spectrum can be separated into the **point spectrum** (the eigenvalues), the **residual spectrum** and the **continuous spectrum**. (In the following we are only interested in very special operators, and since self-adjoint bounded operators do not have a residual spectrum, and compact self-adjoint operators have only point spectrum, we will not go into more detail here.)

A bounded operator A in a Hilbert space $\mathcal{H}$ is called **self-adjoint** if

$$(Au, v) \;=\; (u, Av) \quad \forall u, v \in \mathcal{H}.$$

In finite dimensions these can be represented by symmetric matrices. We know that in this case there exists an orthonormal basis of eigenvectors of A in which A is diagonal. In infinite dimensions a self-adjoint operator might not have any eigenvalues/eigenvectors at all! However, if the operator is compact, the famous Hilbert-Schmidt theorem guarantees that the spectrum consists *solely* of eigenvalues, and furthermore there is an orthonormal basis of the Hilbert space consisting of eigenvectors of A:

Theorem 3 ***Hilbert-Schmidt***

Let A be a self-adjoint compact operator on a Hilbert space $\mathcal{H}$. Then there is a complete orthonormal basis $\{\omega_k\}$ for $\mathcal{H}$ so that $A\omega_k = \lambda_k \omega_k$, and if one orders λ_k in a way that $|\lambda_{k+1}| \leq |\lambda_k|$ then $\lambda_k \to 0$ for $k \to \infty$. In this orthonormal basis A can be written as

$$A \;=\; \sum_{k=0}^{\infty} \lambda_k |\omega_k\rangle\langle\omega_k|.$$

In this sense, compact, self-adjoint operators behave the most like “infinite-dimensional matrices” and can in fact be well approximated by finite-dimensional matrices (as stated above) since the λ_k decrease to 0.

Unfortunately, the classical differential operator, and thus the Laplacian and Bi-Laplacian, are unbounded operators on $L^2(\Omega)$. The Hellinger-Toeplitz theorem implies that an unbounded self-adjoint operator A on $\mathcal{H}$ can not be defined everywhere on $\mathcal{H}$. In general it is defined only on a dense subset of $\mathcal{H}$, the **domain** $D(A)$. For unbounded operators there is a distinction between **symmetric** and **self-adjoint** operators.

Symmetric unbounded operators satisfy

$$(Au, v) \;=\; (u, Av) \quad \forall u, v \in D(A).$$

This means that the adjoint $A^* = A$ on the domain of A, and thus $D(A) \subset D(A^*)$.

A self-adjoint unbounded operator is a symmetric operator with the additional restriction that $D(A) = D(A^*)$ (since $D(A) = \mathcal{H}$ if A is bounded, a symmetric bounded operator is always self-adjoint). Self-adjoint operators are the ones for which one can extend the spectral theory from the bounded to the unbounded case.

Proving that an operator is actually self-adjoint (or finding its self-adjoint closure) is a more difficult task than proving that it is symmetric. Luckily, our goal here is only to find eigenfunction expansions, and it turns out that for this it is enough to require the operator to be symmetric.

The key idea is to apply the Hilbert-Schmidt theorem to A^{-1}. For the inverse to be well-defined, $A : D(A) \to \mathcal{H}$ has to be one-to-one and onto. Now, let A be symmetric and assume A^{-1} exists and is compact. It is easy to see that A^{-1} is self-adjoint:

Since A is onto, for any arbitrary $x, y \in \mathcal{H}$ there are $u, v \in D(A)$ such that $x = Au$, $y = Av$. Thus:

$$(A^{-1}x, y) = (A^{-1}Au, Av) = (u, Av) = (Au, v) = (x, A^{-1}y).$$

We can apply the Hilbert-Schmidt theorem, and it gives us an orthonormal basis of eigenfunctions of A^{-1}. An eigenfunction of A is also an eigenfunction of A^{-1}:

$$A\omega = \lambda\omega \quad \Longleftrightarrow \quad A^{-1}\omega = \lambda^{-1}\omega,$$

so the same basis is also an orthonormal basis of eigenfunctions of A, and we proved the following:

Corollary 1 *Let A be an unbounded, symmetric operator on a Hilbert space $\mathcal{H}$, with compact inverse. Then there is a complete orthonormal basis $\{\omega_k\}$ for $\mathcal{H}$ so that $A\omega_k = \lambda_k\omega_k$. If one orders the λ_k in a way that $|\lambda_{k+1}| < |\lambda_k|$, then $|\lambda_k| \to \infty$ for $k \to \infty$. In this orthonormal basis A can be written as*

$$A \;=\; \sum_{k=0}^{\infty} \lambda_k |\omega_k\rangle\langle\omega_k|. \tag{2.2}$$

2.5 Sobolev Spaces

Corollary 1 gives us conditions, under which we are able to diagonalize an unbounded operator. The next step is to show that there is a suitable extension of the Laplacian that satisfies these conditions. As the name suggests, an extension $\hat{A}$ of A extends A to a larger domain, i.e. $\hat{A} = A$ on $D(A)$ and $D(\hat{A}) \supset D(A)$.

The inverse of the negative Laplacian is the solution operator to the elliptic PDE $-\Delta u = f$, subject to some boundary conditions, where the spaces for u and f are yet to be determined. Showing that under suitable conditions on u and f the solution operator is compact requires existence and uniqueness results for the PDE. For this we need to sufficiently generalize the notion of a derivative. Spaces that ensure the existence of these “weak” derivatives are called Sobolev spaces.

Definitions

Suppose that $u \in C^1(\mathbb{R})$. Integrating against smooth test functions with compact support (denoted by the space $C_c^\infty(\mathbb{R})$), and performing a simple integration by parts gives:

$$\int_{\mathbb{R}} \frac{du}{dx}\phi dx \quad = \quad -\int_{\mathbb{R}} u\frac{d\phi}{dx}dx \quad \forall \phi \in C_c^\infty(\mathbb{R}).$$

The weak derivative is based on this fundamental equality.

Definition 1 *If for any $u \in L^1_{\text{loc}}(\Omega)$, there exists a $v \in L^1_{\text{loc}}(\Omega)$ such that*

$$\int_\Omega v\phi dx \quad = \quad -\int_\Omega u\frac{d\phi}{dx_j}dx \quad \forall \phi \in C_c^\infty(\Omega),$$

then v is called the weak derivative of u with respect to x_j, written $v = D_j u$.

A convenient way to extend this to higher order, mixed derivatives in $\mathbb{R}^n$ is to introduce multi-indices: let $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ and define $|\alpha| = \sum_{j=1}^n \alpha_j$, then the α-th weak derivative $v = D_\alpha u = D_1^{\alpha_1} D_2^{\alpha_2} \cdots D_n^{\alpha_n} u$ has to satisfy the equality

$$\int_\Omega v\phi dx \quad = \quad (-1)^{|\alpha|}\int_\Omega u D^\alpha \phi dx, \quad \forall \phi \in C_c^\infty(\mathbb{R}). \tag{2.3}$$

With this notion we can now define Sobolev spaces:

Definition 2 *The **Sobolev space** $W^{k,p}(\Omega)$ is defined as*

$$W^{k,p}(\Omega) \quad = \quad \{u : D^\alpha u \in L^p(\Omega), \quad 0 \le |a| \le k\},$$

with the norm

$$\|u\|_{W^{k,p}} \quad = \quad \left\{ \sum_{0 \le |\alpha| \le k} \|D^\alpha u\|_{L^p}^p \right\}^{1/p}.$$

Contrary to the spaces $C^\alpha(\Omega)$, Sobolev spaces are “nice” spaces:

Theorem 4 *$W^{k,p}(\Omega)$ is a separable Banach space, i.e. it is complete and has a countable dense subset.*

Thus limits of convergent sequences in $W^{k,p}$ stay inside $W^{k,p}$. We will only use the Sobolev spaces $H^k = W^{k,2}$. Since L^2 is a Hilbert space, we can define an inner product on H^k and make it into a Hilbert space by defining the following inner product:

$$((u, v))_{H^k} \;=\; \sum_{0\leq|\alpha|\leq k} (D^\alpha u, D^\alpha v)_{L^2}.$$

Obviously this inner product is consistent with the norm defined above for $H^k(\Omega)$. The separability property of H^k gives us the possibility to define a Hilbert basis.

Differentiability almost everywhere is not enough to guarantee the existence of weak derivatives (in the 1D case, the relation does not hold for any piecewise differentiable function that has jumps). An even greater generalization deals with this problem: We can view both integrals in (2.3) as functionals on $\mathcal{D}(\Omega) = C_c^\infty(\Omega)$, i.e. as distributions in $\mathcal{D}'(\Omega)$:

$$L_u(\phi) \;=\; \int_\Omega u\phi dx \quad \forall\phi \in \mathcal{D}(\Omega). \tag{2.4}$$

If a functional can be written this way, it is called a regular distribution. Every $u \in L^1_{\text{loc}}(\Omega)$ can be embedded in $\mathcal{D}'(\Omega)$ by identifying it with the regular distribution generated by (2.4). We can then generalize the notion of the weak derivative to *distributional derivatives* by extending (2.3) to all distributions:

Definition 3 *For $u \in \mathcal{D}'(\Omega)$, the distributional derivative is defined by*

$$\langle D^\alpha u, \phi\rangle \;=\; (-1)^\alpha \langle u, D^\alpha \phi\rangle.$$

The beautiful thing about distributional derivatives is the fact that they are defined for every distribution, which means we can take the α-th derivative for any multi-index α. It also means that we can make sense of a derivative for every $u \in L^1_{\text{loc}}(\Omega)$, for instance functions with jumps mentioned above (their distributional derivative has a component that is the well-known “delta-distribution”)

Boundary values and the trace theorem

One difficulty that naturally arises with Sobolev spaces is that strictly speaking one always consideres equivalence classes of functions that differ only on sets of measure zero. Solutions of PDEs usually have to satisfy some sort of boundary condition. The boundary of the domain Ω is a set of measure zero, so the classical way of satisfying boundary conditions does not make sense. The **trace theorem** provides a work-around: It states that we can extend the notion of boundary values from equivalence classes with continuous representatives to all of H^1 in a unique way:

Theorem 5 ***Trace theorem***

Let Ω be a C^1-domain. Then there exists a unique bounded linear operator (the trace operator)

$$T : H^1(\Omega) \quad \to \quad L^2(\partial\Omega),$$

such that

$$Tu \quad = \quad u|_{\partial\Omega} \quad \forall u \in H^1(\Omega) \cap C^0(\overline{\Omega}).$$

We can now automatically enforce boundary conditions on Sobolev functions by restricting them to a subspace of a Sobolev space. A common type of boundary conditions are homogeneous Dirichlet boundary conditions, that enforce the trace to be zero: $Tu = 0$. The subspaces of $H^k(\Omega)$-functions with such restrictions are called $H^k_0(\Omega)$. $C^\infty_c(\Omega)$ is dense in $H^k_0(\Omega)$:

Lemma 2 *$H^k_0(\Omega)$ is the completion of $C^\infty_c(\Omega)$ in the H^k-norm.*

It is possible to define a simpler norm in $H^1_0(\Omega)$ that is equivalent to the $H^1(\Omega)$-norm by using Poincaré's inequality (recall that Ω is a bounded domain).

Proposition 1 ***Poincaré's inequality:*** *There is a $C > 0$ such that*

$$|u| \quad \leq \quad C|Du| \quad \forall u \in H^1_0(\Omega).$$

Thus we can omit the terms with $|\alpha| = 0$ in the definitions of the norm and the inner product:

$$\begin{aligned} \|u\|_{H^1_0} \quad &= \quad \sum_{|\alpha|=1} |D^\alpha u|, \\ ((u, v))_{H^1_0} \quad &= \quad \sum_{|\alpha|=1} (D^\alpha u, D^\alpha v). \end{aligned}$$

Since we can take the k-th weak derivatives and $C^\infty_c(\Omega)$ is dense in $H^k_0(\Omega)$, we can view its dual space as the space in which one can take the k-th distributional derivative:

$$\langle D^\alpha u, \phi\rangle \quad = \quad (-1)^{|\alpha|}\langle u, D^\alpha \phi\rangle \quad \forall \phi \in H^k_0(\Omega), \quad |\alpha| \leq k.$$

Definition 4 *$H^{-k}(\Omega)$ is the dual space of $H_0^k(\Omega)$.*

The Riesz lemma gives an isometry between $L^2(\Omega)$ and $L^2(\Omega)^*$ and therefore provides a method to embed $L^2(\Omega)$ in $H^{-1}(\Omega)$:

$$\iota : L^2(\Omega) \to H^{-1}(\Omega), \quad \langle \iota(f), v\rangle = (f, v) \quad \forall v \in H_0^1(\Omega),$$

which gives us a whole sequence of nested spaces $H^{k+1}(\Omega) \subset H^k(\Omega) \quad \forall k \in \mathbb{Z}$. In fact their distributional derivatives obey the following lemma:

Lemma 3 *If $u \in H^k(\Omega) \quad k \in \mathbb{Z}$, then $D^\alpha u \in H^{k-|\alpha|}(\Omega)$.*

To prove that the inverse of the Laplacian is in fact a compact operator, we need the following compact embedding theorem:

Theorem 6 ***Rellich-Kondrachov compactness theorem***

Let Ω be a bounded C^1 domain. Then $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$.

Periodic Functions

Apart from Dirichlet boundary conditions we can also look at functions that are L-periodic in each direction on the hypercube $Q = [0, K_1] \times \ldots \times [0, K_m]$:

$$u(x + K_j e_j) \quad = \quad u(x), \quad j = 1, \ldots m, \quad K \in \mathbb{Z}^m.$$

The functions $C_p^\infty(Q)$ refer to those in $C^\infty(\mathbb{R})$ satisfying this periodicity condition.

We can define periodic Sobolev spaces as follows:

Definition 5 *The Sobolev space $H_p^s(Q)$ is the completion of $C_p^\infty(Q)$ with respect to the H^s norm*

$$\|u\|_{H^s} \quad = \quad \left(\sum_{0 \le |\alpha| \le s} \|D^\alpha u\|_{L^2(Q)}^2 \right)^{1/2}.$$

Since we can write a periodic function as a formal Fourier series, it is interesting to describe $H_p^s(Q)$ by the behavior of its Fourier coefficients:

Proposition 2

$$H_p^s(Q) \quad = \quad \left\{ u : \quad u = \sum_{k \in \mathbb{Z}^m} c_k e^{2\pi i k \cdot x / L}, \quad \overline{c_k} = c_{-k}, \quad \sum_{k \in \mathbb{Z}^m} |k|^{2s} |c_k|^2 < \infty \right\}.$$

The analogon to $H_0^k(\Omega)$ is the space of functions in $H_p^s(Q)$ with zero mean:

Definition 6

$$\dot{H}_p^s(Q) \;=\; \left\{ u \in H_p^s(Q) : \quad \int_Q u(x)dx = 0 \right\}.$$

For these functions one can prove a Poincaré inequality:

Lemma 4 ***Poincaré inequality***

If $u \in \dot{H}_p^s(Q)$, then

$$|u| \;\leq\; \left(\frac{L}{2\pi}\right)|Du|.$$

For negative powers, $H_p^{-s}(Q)$ is defined to be the dual space of $\dot{H}_p^s(Q)$, and we get similar nested spaces as for Dirichlet boundary conditions. We also get the same compactness result:

Theorem 7 ***Rellich-Kondrachov compactness theorem***

$H^1(Q)$ is compactly embedded in $L^2(Q)$.

Thus we have the same tools for periodic boundary conditions as we have for homogeneous Dirichlet conditions. The following section only deals with the Laplacian subject to Dirichlet conditions, but the same results can be derived analogously for the periodic case.

2.6 $-\Delta$: A Positive, Symmetric, Unbounded Operator with Compact Inverse

The notion of the weak derivative enables us to generalize the Laplacian to a symmetric operator that is invertible on $\mathcal{H} = L^2(\Omega)$. We assume Dirichlet boundary conditions. Then the inverse of the negative Laplacian (mapping f to u) is the solution to the Poisson equation:

$$-\Delta u \;=\; f, \quad u = 0 \text{ on } \partial\Omega.$$

One can take this equation literally, which leads to *classical* solutions, formally stated as:

Definition 7 Classical Solution:

- *For a given* $f \in C^0(\Omega)$*,*
- *find a* $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$
- *with* $u = 0$ *on* $\delta\Omega$

such that $-\Delta u(x) = f(x) \quad \forall x \in \Omega.$

Obviously, this is too restrictive, since f has to be continuous, but we want Δ to be invertible on all of $L^2(\Omega)$. By taking weak derivatives instead of classical derivatives, and requiring the equation to hold in the L^2-sense, we get *strong* solutions:

Definition 8 *Strong Solution:*

- *For a given* $f \in L^2(\Omega)$*,*
- *find a* $u \in H^2(\Omega) \cap H_0^1(\Omega)$

such that $-\Delta u = f$ *in the* L^2*-sense.*

We incorporated the boundary condition into the solution space, but it is still a complicated space. Finally, for *weak* solutions we multiply by "appropriate" test functions and integrate:

$$-\int_\Omega \Delta u(x)\phi(x)dx \quad = \quad \int_\Omega f(x)\phi(x)dx.$$

An integration by parts makes it possible to choose u merely to be in $H_0^1(\Omega)$ (again incorporating boundary conditions) for the left-hand side to make sense. For this we have to require the test functions ϕ to be in $C_c^1(\Omega)$, and thus in $H_0^1(\Omega)$, since $C_c^1(\Omega)$ is dense in $H_0^1(\Omega)$:

$$\int_\Omega \nabla u(x) \cdot \nabla\phi(x)dx \quad = \quad \int_\Omega f(x)\phi(x)dx \quad \forall\phi \in H_0^1(\Omega).$$

In general, a function $u \in H_0^1(\Omega)$ has second derivatives in $H^{-1}(\Omega)$ (cf. Lemma 3). The right-hand side does not need to be a regular distribution, but in fact any element in $H^{-1}(\Omega)$. We can also write the (generalized) negative Laplacian as the linear operator

$$A : H_0^1(\Omega) \to H^{-1}(\Omega) \qquad Au(\phi) = \int_\Omega \nabla u(x) \cdot \nabla\phi(x)dx.$$

We write the standard pairing $y(x)$ of a functional $y \in H^{-1}(\Omega) = (H_0^1(\Omega))^*$ applied to a function $x \in H_0^1(\Omega)$ as $\langle y, x\rangle$. With this notation we can state the weak solution of Poisson's equation as follows:

Definition 9 ***Weak Solution:***

- *For $f \in H^{-1}(\Omega)$,*
- *find $u \in H_0^1(\Omega)$*

such that $\langle Au, \phi \rangle \;=\; \langle f, \phi \rangle \quad \forall \phi \in H_0^1(\Omega)$ *, or in short: $Au = f$ in $H^{-1}(\Omega)$.*

The reason weak solutions are so useful from a theoretical point of view is the fact that the existence/uniqueness theory reduces to a simple application of the Riesz lemma:

Theorem 8 *For $f \in H^{-1}(\Omega)$, the weak form of the Poisson-equation has a unique solution $u \in H_0^1(\Omega)$. Moreover, $\|u\|_{H_0^1(\Omega)} = \|f\|_{H^{-1}(\Omega)}$.*

Proof

With Poincaré's inequality $((u, v)) = \int_\Omega \nabla u \cdot \nabla v dx$ defines an equivalent inner product on $H_0^1(\Omega)$. By definition, $\langle Au, \phi \rangle = ((u, \phi))$, so we need to solve

$$((u, \phi)) \;=\; \langle f, \phi \rangle \quad \forall \phi \in H_0^1(\Omega).$$

This is exactly what the Riesz lemma guarantees: every bounded linear functional f on a Hilbert space H can be written as the inner product with an element $u \in H_0^1(\Omega)$ that has the same norm as f. □

We now know that the negative Laplacian can be inverted on all of $H^{-1}(\Omega)$, and the solutions are elements of $H_0^1(\Omega)$. Since we can embed $L^2(\Omega)$ into $H^{-1}(\Omega)$, the solution operator maps $L^2(\Omega)$ *into* $H_0^1(\Omega)$, and by Rellich-Kondrachov compactness theorem $H_0^1(\Omega)$ is compactly embedded in $L^2(\Omega)$. Putting these facts together, we have that the solution operator A^{-1} is a compact operator from $L^2(\Omega)$ into $L^2(\Omega)$. It is self-adjoint (since $A = -\Delta$ is symmetric) and thus has an orthonormal basis of eigenfunctions. Due to Corollary 1, the same basis is also an eigenbasis of $-\Delta$.

2.7 Fractional Power Spaces of Positive Operators

A **positive** operator satisfies the following inequality:

$$(Au, u) \;\geq\; C\|u\|^2.$$

From Corollary 1 it is easy to see that A is positive if all eigenvalues are positive. In this case, we can define fractional powers of A by

$$A^\alpha \;=\; \sum_{k=0}^{\infty} \lambda_k^\alpha |\omega_k\rangle\langle\omega_k|.$$

The domain of A^α is $D(A^\alpha) = \{u \in \mathcal{H} : \|A^\alpha u\| < \infty\}$. $D(A^\alpha)$ becomes a Hilbert space under the inner product $(u,v)_{D(A^\alpha)} = (A^\alpha u, A^\alpha v)$. Furthermore, $D(A^{\alpha+\varepsilon})$ is compactly embedded in $D(A^\alpha)$.

We need this notion to specify the smoothness of the nonlinear term (i.e. the forcing function) in Chapter 6, where we look at $Au = -u_{xx}$ on $[0,\pi]$ with homogeneous Dirichlet boundary conditions. In this case the orthonormal L^2-basis of eigenfunctions is given by $\omega_k = \sqrt{\frac{2}{\pi}}\sin(kx)$, the corresponding eigenvalues are $\lambda_k = k^2$. The forcing term will be the following hat function:

$$f(x) \quad = \quad \begin{cases} 0 & x \in \left[0, \frac{\pi}{4}\right) \\ \frac{4}{\pi}\left(x - \frac{\pi}{4}\right) & x \in \left[\frac{\pi}{4}, \frac{\pi}{2}\right) \\ \frac{4}{\pi}\left(\frac{3\pi}{4} - x\right) & x \in \left[\frac{\pi}{2}, \frac{3\pi}{4}\right) \\ 0 & x \in \left[\frac{3\pi}{4}, \pi\right). \end{cases}$$

The eigenfunction expansion of f is easily computed as $\sum_{k=1}^\infty c_k \omega_k$ where the coefficients are

$$c_k \quad = \quad \begin{cases} 0 & k \quad \text{even} \\ \frac{8}{\pi k^2}\sqrt{\frac{2}{\pi}}\left(\sin\left(\frac{k\pi}{2}\right) - \sin\left(\frac{k\pi}{4}\right)\right) & k \quad \text{odd}. \end{cases}$$

The requirement $\|A^\alpha u\| < \infty$ means

$$\sum_{k=1}^\infty \lambda_k^{2\alpha} c_k^2 = \sum_{k=1}^\infty k^{4\alpha} c_k^2 = M \sum_{k=1}^\infty k^{4\alpha - 4} \quad < \quad \infty$$

(since the other terms are bounded). This is true as long as $4\alpha - 4 < -1$ or $\alpha < 3/4$. Thus we have shown that $f \in D(A^\alpha)$ for $\alpha \in [0, 3/4)$.

Chapter 3

Dynamical Systems

In this and the following chapter we look at the asymptotic behavior of solutions to dissipative PDEs. Naturally, this requires knowledge about at least existence of solutions for all positive times. However, existence and uniqueness results are not our main concern, and we assume in the following that the PDE in question is known to have unique solutions for all $t > 0$.

In this chapter we introduce important concepts of the theory of dynamical systems that we need for the theory of inertial manifolds.

3.1 The Semigroup operator

Let $u(t)$ be the unique solution to an autonomous PDE, with initial condition $u(0) = u_0$. Then we can define a **semigroup operator** $S(t) : H \to H$ for all $t \geq 0$ by

$$S(t)u_0 \quad = \quad u(t),$$

where $u(t)$ is the solution of the initial value problem to the initial value u_0. The semigroup operator has the following properties

$$\begin{aligned} S(0) &= I \\ S(t)S(s) &= S(s+t). \end{aligned}$$

Moreover, if the solutions to the PDE depend continuously on the initial condition, then $S(t)u_0$ is also continuous in u_0 and t. Such a semigroup is called **strongly continuous**. A semigroup is **injective** (or **backwards unique**), if, whenever two trajectories coincide at some time t_1, they are in fact equal:

$$S(t_1)u_0 = S(t_1)v_0 \quad \text{for}\, t_1 > 0 \quad \Rightarrow \quad u_0 = v_0.$$

A semidynamical system is the phase space H together with the semigroup operator

$$(H, \{S(t)\}_{t \geq 0}).$$

Our main goal in this chapter is to explore the behavior of a semidynamical system for $t \to \infty$. If the system is *dissipative* (see below), all trajectories eventually wander into a compact set B and then stay inside B for all times: B is said to *absorb* all trajectories. In this case it is possible to prove the existence of a set $\mathcal{A}$, that *attracts* all bounded sets: every trajectory approaches this set arbitrarily closely, although it might never reach it, and the rate of attraction might be arbitrarily slow.

Losely speaking, qualitative information about the asymptotic behavior of trajectories is concentrated to $\mathcal{A}$. So, in theory, we can reduce the complexity of our task by focusing only on these sets. Of course this comes with a price: Each trajectory has a *transient phase* before it gets sufficiently close to $\mathcal{A}$ for $\mathcal{A}$ to determine its behavior. By restricting our attention to $\mathcal{A}$ (and to inertial manifolds in the next chapter) we can not obtain any information about this transient phase.

3.2 Dissipativity and Invariance

Definition 10 *$S(t)$ is called* ***(bounded) dissipative****, if there exists a compact set B such that for each* bounded set *X there exists a time $t_1(X)$ for which the following holds:*

$$S(t)X \quad \subset \quad B \quad \forall t \geq t_1(X).$$

B is then called an ***absorbing set****.*

The existence of compact absorbing sets for the K-S equation as well as many other PDEs (including 2D Navier-Stokes and many reaction-diffusion equations) is proved in e.g. [46].

A related definition is that of a **positively invariant** set. Such a set Y is mapped into itself by $S(t)$:

$$S(t)Y \quad \subseteq \quad Y \quad \forall t \geq 0.$$

Similarly, Y is **negatively invariant**, if

$$S(t)Y \quad \supseteq \quad Y \quad \forall t \geq 0.$$

An **invariant** set then is a set that is both positively and negatively invariant, i.e. it is mapped *onto* itself:

$$S(t)Y \quad = \quad Y \quad \forall t \geq 0. \tag{3.1}$$

If the semigroup operator is backwards-unique on an invariant set, we can define its inverse $S(-t)$. In this case

$$S(t)Y = Y \quad \forall t \in \mathbb{R}.$$

An absorbing set is not necessarily positively invariant. This can easily be seen by observing that every compact set X that includes an absorbing set B is also an absorbing set, but trajectories in X might first leave it first before they come back and stay in B for all $t > t_1$.

3.3 Limit Sets and Attractors

If we want to describe the asymptotic behavior of trajectories, it is natural to look at limit points of the trajectory. The set of all limit points of a set X is called the **ω-limit set of X**:

$$\omega(X) = \{y : \exists \quad t_n \to \infty, \quad x_n \in X \,\mathrm{with}\, S(t_n)x_n \to y\}.$$

An equivalent characterization is

$$\omega(X) = \bigcap_{t\geq 0} \overline{\bigcup_{s\geq t} S(s)X}.$$

Definition 11 *The* ***global attractor*** $\mathcal{A}$ *is a maximal compact invariant set (i.e. a compact invariant subset that includes every other compact invariant subset),*

$$S(t)\mathcal{A} = \mathcal{A}$$

, that attracts all bounded sets:

$$\mathrm{dist}(S(t)X, \mathcal{A}) \to 0 \quad \mathrm{for}\, t \to \infty$$

, where $\mathrm{dist}(X, Y) = \sup_{x\in X} \inf_{y\in Y} |x - y|$.

For a dissipative system there is an easy way to describe the global attractor.

Theorem 9 *If* $S(t)$ *is dissipative and* B *is a compact absorbing set then there exists a global attractor* $\mathcal{A} = \omega(B)$*. If* H *is connected then so is* $\mathcal{A}$*.*

The global attractor $\mathcal{A}$ has some important structural properties: Obviously it contains all fixed points and periodic orbits. In fact all complete bounded orbits (i.e. bounded orbits that are defined for all $t \in \mathbb{R}$) are contained in $\mathcal{A}$. Moreover, if $S(t)$ is backwards unique on

$\mathcal{A}$, then $\mathcal{A}$ consists only of complete bounded orbits. In this case, the semidynamical system restricted to $\mathcal{A}$ is actually a dynamical system: $(\mathcal{A}, S(t)_{t\in\mathbb{R}})$.

The restriction to the (semi-)dynamical system on the global attractor yields a somewhat "smaller" system (even though the attractor itself might have a complicated geometry). This raises the question of how much information we loose about trajectories that are not contained in the global attractor, or how well the trajectories on the attractor mimic the behavior of solutions that start far away from $\mathcal{A}$.

Each trajectory approaches the global attractor as a whole. But there is an even more interesting "shadowing" relationship between an arbitrary trajectory and the trajectories on the global attractor: As time passes, a trajectory will be closer and closer to some trajectory on $\mathcal{A}$ (thus "shadowing" it) for longer and longer time intervals. There are no trajectories that move "perpendicular" to every trajectory on $\mathcal{A}$ for all times.

Theorem 10 *Let $u(t)$ be a trajectory of a semidynamical system with a global attractor $\mathcal{A}$. Then there is*

- *a sequence $\{\varepsilon_n\}_{n=1}^{\infty}$ with $\varepsilon_n \to 0$,*
- *a sequence of "switching" times $\{t_n\}_{n=1}^{\infty}$ with $t_{n+1} - t_n \to \infty$ for $n \to \infty$,*
- *a sequence of initial values $\{v_n\}_{n=1}^{\infty}$ with $v_n \in \mathcal{A}$,*

such that

$$|u(t) - S(t-t_n)v_n| \quad \leq \quad \varepsilon_n \qquad \text{for} \qquad t_n \leq t \leq t_{n+1}.$$

Moreover, $|v_{n+1} - S(t_{n+1}-t_n)v_n| \to 0$ for $n \to \infty$.

We will encounter this shadowing property again with inertial manifolds, in fact we will see an even stronger version of shadowing, where it is not necessary to switch trajectories to "track" a trajectory $u(t)$ from the manifold.

As mentioned, there is no statement about the rate of convergence of an arbitrary trajectory towards the global attractor. Under stronger assumptions, one can show the existence of *exponential attractors*, where the rate of attraction is guaranteed to be exponential.

The main issue with global attractors, however, is their difficult geometry. One way to measure their geometric complexity is to obtain estimates on their Hausdorff- or fractal dimension. Many semigroups (e.g. that of the K-S equation) have indeed a finite-dimensional attractor, which suggests that the asymptotic behavior of solutions is governed by only a finite number of "degrees of freedom". Finding a finite-dimensional parametrization of the attractor is a difficult task and subject of current research. One of the most successful methods embeds the attractor in a Lipschitz manifold over a finite-dimensional domain. These manifolds are called inertial manifolds.

Chapter 4

Inertial Manifolds

Our aim in this chapter is to find conditions under which the global attractor of a dissipative semidynamical system

$$\dot{u} + Au + F(u) \;=\; 0 \qquad (4.1)$$

can be embedded in a Lipschitz manifold $\mathcal{M}$ (the inertial manifold). A gentle introduction into the theory of inertial manifolds can be found in [36], [44] and [40].

We split the phase space $\mathcal{H}$ into a finite dimensional component $P\mathcal{H}$ and its orthogonal complement $Q\mathcal{H} = (I - P)\mathcal{H}$, P and Q being the orthogonal projectors onto the repective subspaces. In the following we denote the projections of an arbitrary trajectory $u(t) = S(t)u_0$ onto $P\mathcal{H}$ and $Q\mathcal{H}$ by $p(t) = Pu(t), \quad q(t) = Qu(t)$.

The linear dissipative operator $A : \mathcal{H} \to \mathcal{H}$ is assumed to be an unbounded, self-adjoint, positive operator with compact inverse. From Section 2.4 we know that there is an orthonormal basis of $\mathcal{H}$ consisting of eigenvectors of A. We choose $P\mathcal{H}$ to be the span of the eigenvectors corresponding to the first n eigenvalues, and $Q\mathcal{H}$ its orthogonal complement. This has the advantage that A is diagonalized by this basis. Therefore, A leaves both $P\mathcal{H}$ and $Q\mathcal{H}$ invariant, or put differently, the projection operators commute with A:

$$PA = AP, \qquad QA = AQ.$$

For obtaining inequalities later on, the two eigenvalues $\lambda = \lambda_n$ and $\Lambda = \lambda_{n+1}$ are essential.

The nonlinear term $F(u)$ in (4.1) might in general contain derivatives and thus decrease smoothness. A convenient way to express this is to use fractional power spaces of A. Here we assume that $F : D(A^\alpha) \to D(A^\beta)$ is locally Lipschitz with $\alpha - \beta \leq 1/2$. The condition on $\alpha - \beta$ allows us to easily derive bounds in the proof of the spectral gap condition. It can be relaxed to $\alpha - \beta < 1$.

Furthermore, we assume a positively invariant, absorbing set B. After preparing the equation appropriately (essentially cutting off the nonlinear term in a way that does not affect it in B, see below), we can construct the inertial manifold as the graph of a function ϕ that maps $P\mathcal{H}$ into $Q\mathcal{H}$. Since the prepared equation is identical to the original in B, its inertial manifold restricted to B is also a positively invariant inertial manifold of the original equation (recall that by definition, all asymptotic behavior takes place inside an absorbing set).

4.1 Preparing the Equation

Assume B is contained in a sphere Ω_ρ of radius ρ. Then we can define a **cut-off function**

$$\begin{aligned} \theta &\in C^\infty(\mathbb{R}, [0,1]) \\ \theta(r) &= 1 \quad \forall r \leq 1 \\ \theta(r) &= 0 \quad \forall r \geq 2 \\ |\theta(r)'| &\leq 2, \end{aligned}$$

and replace $F(u)$ with

$$R(u) = \theta(|u|/\rho) F(u).$$

This will not affect the asymptotics of (4.1) since for $|u| < \rho$, $R(u) = F(u)$. Furthermore, since $du/dt + Au = 0$ is dissipative, B will still be an absorbing set. However, $R(u)$ now is globally Lipschitz with a uniform Lipschitz constant C_L. In the following we assume that - if necessary - an appropriate preparation has been made and we have a dissipative evolution equation of the form

$$du/dt + Au + R(u) = 0, \tag{4.2}$$

where $R(u)$ is globally Lipschitz. Together with global boundedness, $R(u)$ has the following properties:

$$\begin{aligned} |R(u)|_\beta &\leq C_0 \quad \forall u \in D(A^\alpha) \\ |R(u_1) - R(u_2)|_\beta &\leq C_L |u_1 - u_2|_\alpha \quad \forall u_1, u_2 \in D(A^\alpha) \\ \mathrm{supp}(R) &\subset \Omega_\rho \equiv \{u \in D(A^\alpha) : |u|_\alpha \leq 2\rho\}. \end{aligned}$$

Equation (4.2) generates a semigroup $S(t)$ on $D(A^\alpha)$, the solutions are both forwards and backwards unique, and for all $t > 0$, $u(t) \in D(A^{1+\beta})$ is more regular than the initial condition $u_0 \in D(A^\alpha)$, with $\frac{d}{dt}u(t) \in D(A^\beta)$ (see [38], [19]).

4.2 Inertial Manifold, Asymptotic Completeness

Definition 12 *An inertial manifold is a*

1. *finite-dimensional Lipschitz manifold, constructed as a graph of a Lipschitz function* $\phi : P\mathcal{H} \to Q\mathcal{H} \cap D(A^\alpha)$ *:*

$$\mathcal{M} = \mathcal{G}[\phi] \equiv \{p + \phi(p) : p \in P\mathcal{H}\},\ |\phi(p_1) - \phi(p_2)|_\alpha \leq l|p_1 - p_2|_\alpha, \tag{4.3}$$

2. *that is invariant:*

$$S(t)\mathcal{M} \ = \ \mathcal{M}, \tag{4.4}$$

3. *and attracts all orbits at an exponential rate:*

$$\operatorname{dist}(S(t)u_0, \mathcal{M}) \ \leq \ Ce^{-kt}. \tag{4.5}$$

The inertial manifold of the *prepared* equation (4.2) is shown to be invariant. Restricting the manifold to $\mathcal{M} \cap B$ yields a positively invariant inertial manifold (since B was assumed positively invariant) for the *original* equation (4.1). Bearing this in mind, we focus on the global inertial manifold of the prepared equation from now on.

The invariance property makes it possible to restrict the dynamics onto $\mathcal{M}$ and still get a closed semidynamical system $(\mathcal{M}, S(t))$. In fact, it reduces the original infinite dimensional evolution equation to finite dimensions: Each trajectory $u(t)$ on $\mathcal{M}$ can be written as $u(t) = p(t) + q(t)$ using the decomposition of $\mathcal{H}$ and then $u(t) = p(t) + \phi(p(t))$ using the definition of $\mathcal{M}$. One can thus write the evolution equation in its finite dimensional form as

$$\dot{p} + Ap + PF(p + \phi(p)) \ = \ 0. \tag{4.6}$$

This form is frequently called the **inertial form**. Note that even though the unknown function $p(t)$ is finite dimensional, ϕ maps into the infinite dimensional space $Q\mathcal{H}$, and the nonlinear operator F has infinite dimensional domain. So the number of degrees of freedom is indeed finite dimensional, but (4.6) does not immediately lend itself towards numerical computations.

The question of what space to choose for ϕ is also partially answered by (4.6): For ODEs, Lipschitz continuity of the right hand side is needed for existence and uniqueness of solutions. Thus it makes sense to require ϕ to be at least Lipschitz continuous.

The rate of attraction towards the global attractor can be arbitrarily slow. The inertial manifolds we construct in this chapter attract all trajectories exponentially, i.e. the distance between any trajectory and $\mathcal{M}$ decreases as $e^{-\mu t}$ for some $\mu > 0$ (μ depends on the concrete trajectory). Note that (exponential) attraction and invariance imply that the global attractor has to be a subset of $\mathcal{M}$.

The measure of distance toward the manifold is greatly simplified if one can additionally prove asymptotic completeness:

Definition 13 *Asymptotic completeness*

An inertial manifold is said to be ***asymptotically complete*** *if for any $u(t)$ with $u_0 \in D(A^\alpha)$ there is a point $v_0 \in \mathcal{M}$ s.th.*

$$|u(t) - S(t)v_0|_\alpha \quad \rightarrow \quad 0 \quad \text{for} \quad t \rightarrow \infty. \tag{4.7}$$

If the rate of convergence is exponential, this is called exponential tracking.

Asymptotic completeness is a much stronger version of the "shadowing" property that we encountered in Chapter 3 for global attractors. It means that for each trajectory there is a trajectory on the manifold that has the same asymptotic behavior, i.e. the asymptotic dynamics on the manifold capture the asymptotic dynamics of the whole dynamical system.

4.3 The Strong Squeezing Property

For the existence proof we have to make two additional important assumptions. Together they are called the *Strong Squeezing Property* (see Figure 4.1). Roughly speaking, the *Cone Invariance Property* makes sure that the evolution of a certain manifolds under the semigroup, $S(t)\mathcal{M}$, stays a manifold. The second part, called the *Squeezing Property*, ensures that the distance between points, that are not on the same manifold, is decreasing exponentially. Both properties together also imply asymptotic completeness.

Let ψ be some Lipschitz function as in (4.3), and let $u_1(t)$ and $u_2(t)$ be two trajectories *on* $\mathcal{G}[\psi]$ for all $t > 0$. Then for their difference

$$w(t) = u_1(t) - u_2(t) \quad = \quad p_1(t) - p_2(t) + \phi(p_1(t)) - \phi(p_2(t)),$$

the following is true *for all* $t > 0$ (again, due to (4.3)):

$$|Qw(t)|_\alpha \quad \leq \quad l|Pw(t)|_\alpha. \tag{4.8}$$

The **Cone Invariance Property** generalizes this inequality to trajectories that do not necessarily lie on some $\mathcal{G}[\psi]$ for all $t > 0$.

Definition 14 *Cone Invariance Property*

If the Lipschitz condition (4.8) holds for any two trajectories at some $t_0 > 0$, then it holds for all $t > t_0$:

$$|Qw(t_0)|_\alpha \leq l|Pw(t_0)|_\alpha \quad \Rightarrow \quad |Qw(t)|_\alpha \leq l|Pw(t)|_\alpha \quad \forall t > t_0. \tag{4.9}$$

Put differently, if we define a cone $\mathcal{C}_l = \{w = u_1 - u_2 : |Qw|_\alpha \leq l|Pw|_\alpha\}$, then (4.9) means that $\mathcal{C}_l$ is positively invariant under $S(t)$:

$$S(t)\mathcal{C}_l \quad \subset \quad \mathcal{C}_l \quad \forall t > 0,$$

where by $S(t)\mathcal{C}_l$ we mean the set $\{S(t)u_1 - S(t)u_2 : (u_1 - u_2) \in \mathcal{C}_l\}$. As $S(t)$ is not linear in general, this is not necessarily equal to $\{S(t)w : w \in \mathcal{C}_l\}$.

The **Squeezing Property** concerns the difference of two trajectories $w(t) = u_1(t) - u_2(t)$ for which at some $t_0 > 0$ their difference is not in the interior of $\mathcal{C}_l$:

Definition 15 *Squeezing Property*

If for any two trajectories at some $t_0 > 0$ their difference $w(t_0)$ satisfies $|Qw(t_0)|_\alpha \geq l|Pw(t_0)|_\alpha$, then $w(t)$ has been "squeezed together" exponentially in the past:

$$|Qw(t_0)|_\alpha \geq l|Pw(t_0)|_\alpha \quad \Rightarrow \quad \exists k > 0 : |Qw(t)|_\alpha \leq |Qw(0)|_\alpha \, e^{-kt} \quad \forall 0 < t \leq t_0. \tag{4.10}$$

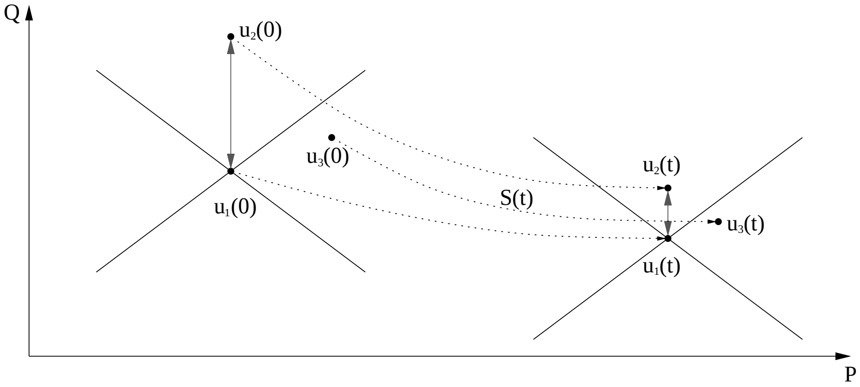

Figure 4.1: The Strong Squeezing Property

4.4 Existence of Inertial Manifolds - The Graph Transform Method

The *Hadamard* or *Graph Transform method* ([37], [38]) is an elegant geometrical way to prove the existence of inertial manifolds under the hypotheses made above. This is done in two stages: First, one proves the existence of a manifold that is invariant under $S(t)$. In a second step that manifold is shown to attract all trajectories exponentially.

Theorem 11 *If the Strong Squeezing Property holds, then there exists an asymptotically complete inertial manifold for equation (4.2).*

Existence of an Invariant Manifold

The idea for the proof of an invariant manifold ($S(t)\mathcal{M} = \mathcal{M}$), is to look at the evolution of a manifold, as a graph of a Lipschitz function ϕ, under the semigroup $S(t)$. The evolved manifold $\mathcal{M}_t$ is still the graph of a (different) Lipschitz function ϕ_t under the assumptions we made above. Thus we can define operators $T(t)$ on an appropriate space of Lipschitz functions, and show that they inherit the semigroup properties of $S(t)$. We then use a fixed point theorem on Banach spaces to conclude that there is a ϕ such that $T(t)\phi = \phi$. The graph of this ϕ is an invariant manifold.

The appropriate space $\mathcal{F}_l^n$ mentioned above consists of all

$$\phi : P\mathcal{H} \to Q\mathcal{H} \cap D(A^\alpha)$$

with the additional constrains:

1. ϕ is globally Lipschitz: $|\phi(p_1) - \phi(p_2)|_\alpha \leq l|p_1 - p_2|_\alpha$
2. ϕ is globally bounded: $\|\phi\| := \sup_{p\in P_n H} |\phi(p)|_\alpha < \infty$
3. ϕ has bounded support: $\text{supp}(\phi) \subset P_n\Omega_\rho$.

Clearly, $\mathcal{F}_l^n$ is a convex subset of the Banach space of continuous functions $C_0(P_n\mathcal{H}, Q_n\mathcal{H} \cap D(A^\alpha))$ and as such a complete metric space.

We look at the evolution of each point $u_0 = p_0 + \phi(p_0)$ on the graph of some $\phi \in \mathcal{F}_l^n$. The Cone Invariance Property guarantees that under evolution, each $u(t) = S(t)u_0$ can still be written as $u(t) = p_t + \psi(p_t)$ with $p_t \in P\mathcal{H}$ and ψ a Lipschitz function with the same Lipschitz bound l as ϕ. We can thus define a semiflow $T(t)$ on $\mathcal{F}_l^n$ that maps ϕ into ψ by:

$$\mathcal{G}[\psi] = S(t)\mathcal{G}[\phi] = \mathcal{G}[T(t)\phi]. \tag{4.11}$$

To show that $T(t)$ actually maps $\mathcal{F}_l^n$ into itself, we first check the constraints on functions in $\mathcal{F}_l^n$. Since $\text{supp}(R) \subset \Omega_\rho$, $\psi(p) = 0$ for $|p|_\alpha > \rho$. This also implies the bound on $\|\psi\|$. The same Lipschitz bound l is the consequence of the Cone Invariance Property.

To conclude that $T(t)$ maps into $\mathcal{F}_l^n$, we have to show that the domain of ψ is actually $P\mathcal{H}$, i.e. there are no "holes" in $\mathcal{G}[\psi]$, or $P\mathcal{M}_t = P\mathcal{H}$. This is done using a topological argument: if $P\mathcal{M}_t$ had a hole, one could construct a retraction of a ball onto its boundary and hence get a contradiction (A retraction is a continuous map of a space onto a subspace that leaves the subspace fixed):

We assume that there is a hole in $\mathcal{M}_t$, i.e. there is a $y \in P\mathcal{H}$ with $y \neq P\mathcal{M}_t$. Define a map $g_\phi(t) : P\mathcal{H} \to P\mathcal{H}$ with $g_\phi(t)(p) = P(S(t)(p + \phi(p)))$. $g_\phi(t)$ is continuous since ϕ is continuous and $S(t)$ is strongly continuous. Moreover, $g_\phi(t) = e^{-At}p$ for $|p|_\alpha \geq 2e^{At}\rho$ due to the preparation of $R(u)$ in equation (4.2), and thus $e^{At}g_\phi(t)$ is a continuous map that fixes every p with $|p|_\alpha \geq 2e^{At}\rho$ and by assumption introduces a hole y. One can then construct a continuous map that retracts $\Omega_{2\rho}$ onto its boundary, by composing $e^{At}g_\phi(t)$ with the map that maps each point $x \in \Omega_{2\rho}$ onto the boundary of $\Omega_{2\rho}$ along the line through x and y, a contradiction.

Thus $T(t)$ maps $\mathcal{F}_l^n$ into itself. In fact, $T(t)$ inherits all semigroup properties from $S(t)$. For this, we write $[T(t)\phi](p)$ as the function that lifts p onto the evolved manifold $\mathcal{M}_t$:

$$[T(t)\phi](p) \;=\; QS(t)([g_\phi(t)]^{-1}(p) + \phi([g_\phi(t)]^{-1}p)),$$

(backwards uniqueness is crucial for $g_\phi(t)^{-1}$ to be well-defined). Obviously, $T(0) = I$. It also inherits continuity in both t and ϕ. Now, we can rewrite

$$S(t)(p + \phi(p)) \;=\; p(t) + (T(t)\phi)(p(t)).$$

With $u_\phi(p) = p + \phi(p)$ this becomes

$$u_{T(t)\phi}(PS(t)u_\phi(p)) \;=\; S(t)u_\phi(p).$$

Using this relation, it is straightforward to prove $T(t)T(s) = T(t+s)$, i.e.

$$u_{T(t+s)\phi}(P(S(t+s)u_\phi(p))) \;=\; u_{T(t)T(s)\phi}(PS(t+s)u_\phi(p)).$$

To show that $T(t)$ actually has a fixed point, we use the following theorem from Hale [18]:

Theorem 12 *Let X be a complete metric space and $T(t) : X \to X$ a continuous semigroup which is completely continuous and dissipative. Then $T(t)$ has a fixed point.*

Complete continuity means that $T(t)B$ is precompact for any bounded set B and all $t > 0$.

As a closed convex subset of the continuous functions, $\mathcal{F}_l^n$ is a complete metric space. To verify dissipativity and complete continuity, we need to find bounds $\|T(t)\phi\|$ and $\|A^\varepsilon T(t)\phi\|$ for some $\varepsilon > 0$. As $\phi(p)$ is the Q-projection of some trajectory $u(t)$, this amounts to finding bounds on $|q(t)|_\alpha$ and $|q(t)|_{\alpha+\varepsilon}$. $q(t)$ is governed by the following evolution equation:

$$\dot{q} + Aq + QR(u) \;=\; 0.$$

Using the variation of constants formula then gives

$$q(t) \;=\; e^{-At}q(0) - \int_0^t e^{-A(t-s)}QR(u(s))\,\mathrm{ds},$$

from which one can obtain bounds on $|q(t)|_{\alpha+\varepsilon}$ (with $\alpha - \beta + \varepsilon < 1$) of the form

$$|T(t)\phi|_{\alpha+\varepsilon} \;\leq\; K_2(\varepsilon)\|\phi\|e^{-\Lambda t} + K_3. \tag{4.12}$$

This implies that the set $\{\phi \in \mathcal{F}_l^n : \phi\| \leq 2K_3\}$ is absorbing and thus proves dissipativity of $T(t)$.

To prove complete continuity, we have to show $X = \overline{T(t)B}$ is compact for every bounded set $B \in \mathcal{F}_l^n$. If we can show that X is equicontinuous and pointwise compact, we can apply the Arzela-Ascoli theorem for Banach spaces (Theorem 2). Since all $\phi \in \mathcal{F}_l^n$ have the same Lipschitz bound l, equicontinuity is obvious. For pointwise compactness, i.e. to prove that the sets $\{\phi(p),\, \phi \in X\}$ are compact, we can again use (4.12) to find a uniform bound on ϕ in $D(A^{\alpha+\varepsilon})$. We can then use the compact embedding of $D(A^{\alpha+\varepsilon})$ into $D(A^{\alpha})$ to conclude that X is indeed compact.

All assumptions of Theorem 12 are thus satisfied, and there exists a fixed point ϕ of $T(t)$. Its graph is an invariant manifold.

Exponential Attraction, Asymptotic Completeness

To prove that the invariant manifold obtained above is in fact an asymptotically complete inertial manifold, we need to show that for any trajectory $u(t)$ there is a trajectory $v(t) \subset \mathcal{M}$ such that the distance $|u(t) - v(t)|_\alpha$ is exponentially decreasing. The Squeezing Property is of central importance here, since it defines a cone at each point of $u(t)$ inside which distances decrease exponentially. The key idea is to prove the existence of a trajectory $v(t)$ on $\mathcal{M}$ that stays inside this moving cone for all times. This allows us to conclude on the asymptotic completeness and thus exponential attraction of $\mathcal{M}$.

Let $C_s(u) = \{v \in D(A^\alpha) : |Qw|_\alpha \geq l|Pw|_\alpha,\ w = u - v\}$ be the cone attached to a some point $u \in D(A^\alpha)$, inside which the Squeezing Property holds. The cone casts a “shadow” $V(u) = \{v \in \mathcal{M} : |Qw|_\alpha \geq l|Pw|_\alpha,\ w = u - v\}$ onto $\mathcal{M}$. Let $V_t = V(u(t))$ be the evolving cone along an arbitrary trajectory $u(t)$. If there is a $v_0 \in \mathcal{M}$ such that its trajectory $v(t)$ stays inside V_t for all $t > 0$, then the Squeezing Property guarantees that $u(t)$ is attracted exponentially towards $v(t)$ and thus towards $\mathcal{M}$.

The complement of $C_s(u(t))$ is invariant under $S(t)$ by the Cone Invariance Property, $\mathcal{M}$ is invariant as we proved above, so trajectories on $\mathcal{M}$ can only leave the shadows V_t, but never enter it. Thus we can define for $t_n \to \infty$ a sequence of points v_n inside the shadow V_{t_n}, such that their backwards-evolution $S(t - t_n)v_n$ will be in V_t for all $t \in [0, t_n]$, in particular $S(-t_n)v_n \in V_0$. Thus we have a sequence of points inside V_0 of which we know by construction that their trajectories stay longer and longer in the shadows V_t. As soon as we know that V_0 is compact, we can extract a subsequence that is converging inside V_0, say towards v_0. The trajectory $S(t)v_0$ will be inside V_t for all $t > 0$, for if there was a $t_0 > 0$ with $S(t)v_0 \notin V_t \quad \forall t > t_0$, there would be an open neighborhood $N \ni v_0$ such that $S(t)v \notin V_t \quad \forall t > t_0, \quad v \in N$. But then v_0 was not the limit point of the subsequence above, a contradiction.

To prove compactness of $V(u)$, note that since $V(u)$ is by definition closed and contained in the finite-dimensional manifold $\mathcal{M}$, it is enough to show that $V(u)$ is bounded, furthermore since all $u \in \mathcal{M}$ can be written as $p + \phi(p)$ with continuous ϕ, it is enough to show $PV(u)$ is bounded in $PD(A^\alpha)$. Let $u = p_u + q_u$, $v = p_v + \phi(p_v)$. As $v \in V(u)$

$$|p_u - p_v|_\alpha \leq l^{-1}|q_u - \phi(p_v)|_\alpha \leq l^{-1}(|q_u|_\alpha + \|\phi\|).$$

Together with $|p_u - p_v|_\alpha \geq |p_v|_\alpha - |p_u|_\alpha$ this gives the required bound on $PV(u)$:

$$|p_v|_\alpha \quad \leq \quad |p_u|_\alpha + l^{-1}(|q_u|_\alpha + \|\phi\|).$$

Assured of the existence of a trajectory $S(t)v_0$ that is shadowed by $u(t)$ for all $t > 0$, we can use the Squeezing Property to prove exponential decay of $|S(t)v_0 - u(t)|_\alpha$, which in turn proves exponential attraction and asymptotic completeness for all $u(t) = S(t)u_0$ with $u_0 \in \Omega_\rho$:

$$\begin{aligned}
|u(t) - S(t)v_0|_\alpha &\leq |P(u(t) - S(t)v_0)|_\alpha + |Q(u(t) - S(t)v_0)|_\alpha \\
&\leq (1 + l^{-1})|Q(u(t) - S(t)v_0)|_\alpha \\
&\leq (1 + l^{-1})|Qu(0) - Qv_0|_\alpha e^{-kt} \\
&\leq (1 + l^{-1})(|Qu(0)|_\alpha + \|\phi\|)e^{-kt} \\
&\leq (1 + l^{-1})(\rho + \|\phi\|)e^{-kt}.
\end{aligned}$$

This concludes the proof of Theorem 11.

A Spectral Gap Condition

It is possible to derive a general spectral gap condition that implies the Strong Squeezing Property and depends only on the gap between the largest eigenvalue λ of PA and the smallest eigenvalue Λ of QA.

Observe that the Cone Invariance Property (4.9) is assured if for $w(t) = u_1(t) - u_2(t)$

$$\frac{d}{dt}(|Qw(t)|_\alpha - l|Pw(t)|_\alpha) \quad < \quad 0. \tag{4.13}$$

The evolution equation is autonomous, so it is sufficient to ensure (4.13) for $t = 0$.

Starting from $\frac{1}{2}\frac{d}{dt}|A^\alpha p|^2 = (A^\alpha \dot{p}, A^\alpha p)$ and $\frac{1}{2}\frac{d}{dt}|A^\alpha q|^2 = (A^\beta \dot{q}, A^{2\alpha-\beta}q)$, some technical calculations allow us to derive a lower bound for $\frac{d}{dt}|Qw(t)|_\alpha$ and an upper bound for $\frac{d}{dt}|Pw(t)|_\alpha$, given by

$$\left(\frac{d}{dt}|q|_\alpha\right)_{t=0} \quad \leq \quad -\left(\Lambda - C_L\Lambda^{\alpha-\beta}(1 + l^{-1})\right)|q|_\alpha \tag{4.14}$$

$$\left(\frac{d}{dt}|p|_\alpha\right)_{t=0} \quad \geq \quad -(\lambda + C_L\lambda^{\alpha-\beta}(1 + l))|q|_\alpha/l. \tag{4.15}$$

Recall that C_L was the global Lipschitz constant of the nonlinear operator $R(u)$, and l the (arbitrary) Lipschitz constant for $\mathcal{F}_l^n$.

Thus we find the following upper bound for $\frac{d}{dt}(|Qw(t)|_\alpha - l|Pw(t)|_\alpha)$:

$$\frac{d}{dt}\left(|q|_\alpha - l|p|_\alpha\right)_{t=0} \;\leq\; -(\Lambda - \lambda - C_L\Lambda^{\alpha-\beta}(1+l^{-1}) - C_L\lambda^{\alpha-\beta}(1+l))|q|_\alpha,$$

which is negative provided the following **spectral gap condition** holds:

$$\Lambda - \lambda \;>\; C_L\Lambda^{\alpha-\beta}(1+l^{-1}) + C_L\lambda^{\alpha-\beta}(1+l). \tag{4.16}$$

The Squeezing Property follows from (4.14): Let $\mu = -\left(\Lambda - C_L\Lambda^{\alpha-\beta}(1+l^{-1})\right)$, then

$$\frac{d}{dt}|q|_\alpha \;\leq\; \mu|q|_\alpha.$$

The spectral gap (4.16) ensures $\mu < 0$. Integration then gives the Squeezing Property:

$$|q(t)|_\alpha \;\leq\; |q(0)|_\alpha e^{\mu t}.$$

One can minimize the right hand side of (4.16) with respect to l (which was an arbitrary constant up to now). The term is minimized by $l = (\Lambda/\lambda)^{(\alpha-\beta)/2}$, which gives:

$$\Lambda - \lambda \;>\; C_L\left(\lambda^{\frac{\alpha-\beta}{2}} + \Lambda^{\frac{\alpha-\beta}{2}}\right)^2. \tag{4.17}$$

This spectral gap condition requires that there have to be **large enough** gaps between eigenvalues of A, and that they have to occur **soon enough** if $\alpha \neq \beta$, since the size of the necessary gaps increases with the eigenvalues.

4.5 Other Methods of Proof

There are several other methods of proof for the existence of inertial manifolds ([44]). We outline their general ideas here for completeness and because each existence proof could potentially be used for the construction of *approximate inertial manifolds* (Section 4.7).

The **Lyapunov-Perron** ([12], [46], [11]) method is similar in spirit to the Graph Transform method: the inertial manifold is shown to be the fixed point of a particular map. This method starts off with the projections of the evolution equation on $P\mathcal{H}$ and $Q\mathcal{H}$:

$$\dot{p} + Ap + PR(p+q) \;=\; 0 \tag{4.18}$$

$$\dot{q} + Aq + QR(p+q) \;=\; 0. \tag{4.19}$$

Now, on the manifold $q(t) = \phi(p(t))$. Plugging this identity into the nonlinear part yields a nonlinear ODE in p with unique solutions for all t (since $P\mathcal{H}$ is finite dimensional and R is assumed Lipschitz) and a linear abstract ODE in q that depends on $p(t)$:

$$\begin{aligned} \dot{p} &= -Ap - PR(p + \phi(p)) \\ \dot{q} &= -Aq - QR(p + \phi(p)). \end{aligned} \tag{4.20}$$

Using the variation of constants formula, one can write the solution for $q(0)$ at $t = 0$ explicitly as:

$$q(0) = T\phi(p(0)) = -\int_{-\infty}^{0} e^{-A(t-s)} QR(p(s) + \phi(p(s)))ds. \tag{4.21}$$

If the graph of ϕ is an invariant manifold, $q(0) = \phi(p(0))$, which makes ϕ a fixed point of the operator T. Under the Strong Squeezing Property, T can be shown to be a contraction and the inertial manifold $\mathcal{G}[\phi]$ a fixed point.

The **Cauchy method** ([39], [2]) constructs the manifold as the evolution of a large enough sphere $\Gamma = \{p \in P\mathcal{H} : |p| = R\}$ in $P\mathcal{H}$:

$$\mathcal{M} = \overline{\bigcup_{t>0} S(t)\Gamma}.$$

The **Sacker method** ([26]) starts from (4.19) and uses $q(t) = \phi(p(t))$:

$$[\phi'(p)](\dot{p}) + A\phi(p) + QR(p + \phi(p)) = 0, \tag{4.22}$$

where $\phi'(p)$ denotes the Fréchet differential of ϕ at p. Solving (4.18) for $\dot{p}$ and substituting that into (4.22) yields the following abstract hyperbolic equation in ϕ:

$$[\phi'(p)](-Ap - PR(p + \phi(p))) + A\phi(p) + QR(p + \phi(p)) = 0.$$

4.6 Inertial Manifolds for Common PDEs

The existence of inertial manifolds has been proven for a wide variety of dissipative evolution equations, for instance certain reaction-diffusion equations ([46], [2]), the Kuramoto-Sivashinsky equation ([2], [40]), the nonlocal Burgers' equation ([2]), the Cahn-Hillard equation ([2]) and the Ginzburg-Landau equation ([46]).

Despite this success, the primary motivation for the development of the theory of inertial manifolds, the 2D Navier-Stokes equation, is still not known to have an inertial manifold. The existence of an inertial manifold could shed light on whether turbulence is a finite dimensional phenomenon. For the 3D Navier-Stokes equation, even existence and uniqueness of solutions is an open question.

4.7 Approximate Inertial Manifolds

So far, all known existence proofs for inertial manifolds have been nonconstructive - we do not gain any information about the exact location of the inertial manifold. To make the theory of inertial manifolds useful for both further theoretical examination and computational use, we have to find constructive ways to at least approximate ϕ, and then to measure the quality of the approximation.

Much research has been done on *approximate inertial manifolds* (AIMs). These manifolds that are no longer invariant but attract all solutions into a thin layer around the manifold in finite time. AIMs can be shown to exist under weaker conditions than e.g. the spectral gap condition and are thus available even for dissipative PDEs for which an exact inertial manifold is not known to exist (e.g. 2D Navier-Stokes). In this case the AIMs approximate the global attractor directly. AIMs are not necessarily constructed as graphs of functions: some are defined implicitly (see for example the AIM ϕ^s of Section 4.7).

The general approach for constructing AIMs is to start off with the projection of the original evolution equation onto $P\mathcal{H}$ and $Q\mathcal{H}$:

$$\begin{aligned} \dot{p} + Ap + PR(p+q) &= 0 && (4.23) \\ \dot{q} + Aq + QR(p+q) &= 0. && (4.24) \end{aligned}$$

We can then use the fact that on the inertial manifold, $q = \phi(p)$ satisfies (4.24), to obtain a formula for the AIM, either by neglecting or approximating $\dot{q}$. Some existence proofs that rely on a fixed-point argument give rise to iterative constructions that yield sequences of AIMs with ever-increasing accuracy ([4], [46], [3]).

An important issue is the quality of an AIM. Essentially, one wants to measure the distance between the inertial manifold $\mathcal{M}$ and an approximate inertial manifold $\mathcal{M}_\phi$:

$$\operatorname{dist}(\mathcal{M}, \mathcal{M}_\phi),$$

which is difficult to handle in general, since we do not know more about $\mathcal{M}$ than its existence. One possible measure is given by a uniform *asymptotic* bound on the difference between a solution $u(t) = p(t) + q(t)$ and its projection onto the AIM: $u_\phi(t) = p(t) + \phi(p(t))$, given by

$$|u(t) - u_\phi(t)| = |q(t) - \phi(p(t))|.$$

Typically, estimates on the **rate of convergence** in the spatial dimension n are derived:

$$|q(t) - \phi(p(t))| \quad \leq \quad K\lambda_{n+1}^{-\gamma} \quad \forall t > t_*,$$

where K is independent of n, but usually dependent on t and on the particular AIM.

It is then shown that the value of γ for a certain AIM construction of a certain PDE in a certain norm is larger than the value of γ for the zero manifold.

For dissipative PDEs, we can immediately construct an AIM: the trivial AIM $\phi = 0$. Since in this case $|u(t) - u_\phi(t)| = |q(t) - 0| = |q(t)|$, the size of an absorbing set gives an estimate on the neighborhood size. For nontrivial AIMs to be useful, they have to attract solutions into a thinner neighborhood, i.e. its value of γ has to be larger than the corresponding value for the zero manifold. The size of the neighborhood depends on smoothness assumptions (e.g. the forcing term) and also on the particular structure of the PDE in question.

In the following, we review the most common AIM constructions. The primary motivation is to obtain explicit formulae that approximate existing inertial manifolds, which we can turn into numerical schemes. The actual implementation raises further issues (for example, all AIMs constructed below map into an infinite-dimensional space $Q\mathcal{H}$), which is discussed in the next chapter.

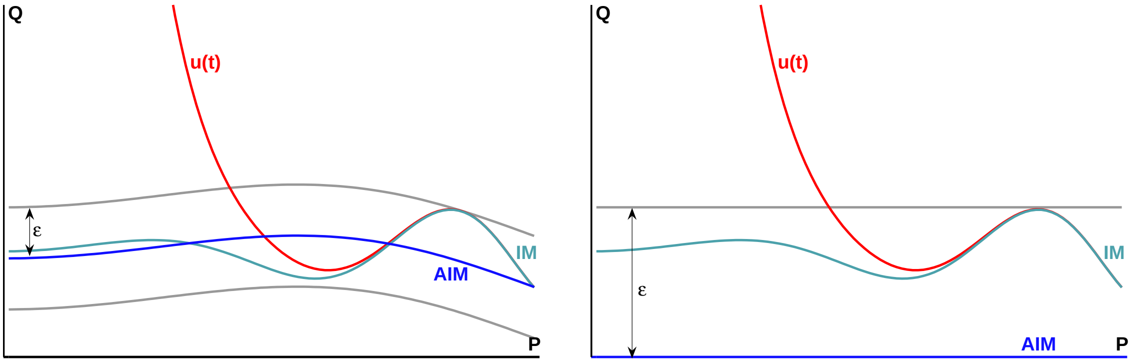

Figure 4.2: Approximate inertial manifold and zero-manifold

Steady and Pseudo-Steady AIMs

One of the first published and most popular AIM constructions is the **steady AIM** ϕ^s ([47], [7]), derived from (4.24) by neglecting the time-derivative:

$$Aq + QR(p+q) \quad = \quad 0.$$

A justification for this step is the fact that for n sufficiently large, q and $\dot{q}$ remain small compared to p. For all equilibrium solutions the derivative is indeed 0, which means that these points lie on the exact inertial manifold and the global attractor as well as on the AIM, thus “threading” them together. This AIM turns the coupled system of equations (4.23),(4.24) into a differential-algebraic equation:

$$\begin{aligned} \dot{p} + Ap + PR(p+q) &= 0 \\ Aq + QR(p+q) &= 0. \end{aligned}$$

One can use a sequence of Picard iterations to approximate ϕ^s. This yields a sequence of **pseudo-steady AIMs** defined by:

$$\phi_{k+1}(p) = q_{k+1} = T_p(q_k) = -A^{-1}QR(p + q_k).$$

For n (the dimension of $P\mathcal{H}$) large enough, T_p is a contraction on some bounded set ([47]) and ϕ^s is the unique fixed point. Applying only one Picard iteration with $q_0 = 0$ yields the probably most popular AIM in the literature, the Foias-Manley-Temam AIM (**FMT AIM**)

$$\phi(p) \;=\; -A^{-1}QR(p). \tag{4.25}$$

The FMT AIM is discussed for the Navier Stokes equation e.g. in [10], [31], [20], [47], for the Kuramoto-Sivashinsky equation in [23] and [24], for reaction-diffusion equations in high space dimensions in [29] and the Cahn-Hilliard equation in [30]. For the forced Burgers' equation, $\gamma = \frac{3}{2} + \alpha$ as opposed to $\gamma = 1 + \alpha$ for the zero manifold, if the forcing term $f \in D(A^\alpha)$ ([24]).

AIMs based on the Graph Transform Method

Another approach in constructing AIMs is based on the Graph Transform method of proof described in Section 4.4: Starting out with a simple manifold ϕ_0 (e.g. the flat manifold with $q = 0$), we construct better approximations ϕ_t of the inertial manifold by applying the semigroup $G[\phi_t] = S(t)G[\phi_0]$. Under the conditions discussed in Section 4.3 this mapping is a contraction. Thus there is hope that the evolution by a short time τ will improve the accuracy of the AIM.

The simplest of these AIMs was introduced by Foias, Sell, Titi [12]. To obtain a computationally inexpensive explicit form, $\dot{q}$ in (4.24) is discretized using a simple implicit-explicit Euler scheme (implicit in the linear term, explicit in the nonlinear term):

$$\frac{q_{k+1} + q_k}{\tau} + Aq_{k+1} + QR(p + q_k) \;=\; 0,$$

which yields the following explicit AIM rule:

$$q_{k+1} \;=\; -(I + \tau A)^{-1}(\tau QR(p + q_k) + q_k).$$

Starting with $q_0 = 0$ we get the following simple AIM (the **Euler-Galerkin AIM**) after one Euler-step:

$$q \;=\; -\tau(I + \tau A)^{-1}QR(p).$$

Of course one needs to address the question of how to choose a "good" τ for practical purposes. For estimates on the rate of convergence for the Euler-Galerkin AIM, see e.g. [12], [26], [9], [42], [23].

This scheme can be extended e.g. by considering other time discretization schemes ([41])

From a numerical point of view there does not seem to be much difference over using a simple time discretization scheme for the $Q\mathcal{H}$ part of the evolution equation, except for the fact that at each time step the previous q is assumed to be 0.

AIMs based on the Lyapunov-Perron Method

In [40] and [4] a sequence of converging AIMs is constructed, based on the Lyapunov-Perron existence proof. Similar to the Graph Transform method, the inertial manifold is shown to be the fixed point of a particular map. We start off with the projection of the evolution equation on $Q\mathcal{H}$, with $q = \phi(p)$:

$$\frac{d\phi(p)}{dt} + A\phi(p) + QR(p + \phi(p)) \quad = \quad 0.$$

Integrating over $[-\infty, 0]$ using the variation of constants formula yields:

$$\phi(p_0) = T\phi(p_0) \quad = \quad -\int_{-\infty}^{0} e^{-A(t-s)} QR(p(s) + \phi(p(s)))ds. \tag{4.26}$$

Under the spectral gap condition the operator T is a contraction.

To derive an explicit form for ϕ, we need an approximation of $p(s)$ for $s \in [-\infty, 0]$. We construct a discrete-time approximation $p_k \approx p(-k\tau)$ for using a simple Euler scheme on the time-derivative in (4.23):

$$\frac{p_{k+1} - p_k}{-\tau} + Ap_k \quad + PR(p_k + \phi(p_k)) = \quad 0$$

to get:

$$p_{k+1} \quad = \quad (I + \tau A)p_k + \tau P_n R(p_k + \phi(p_k))$$

We extend this discrete-time sequence to a continuous-time step function $p_\tau(t)$ by

$$\begin{aligned} p_\tau(t) = p_k \quad & \text{for} \quad -(k+1)\tau < t < -k\tau, \quad k = 0, \ldots, N-1 \\ p_\tau(t) = p_N \quad & \text{for} \quad t \leq -N\tau. \end{aligned}$$

With p_τ instead of p, (4.26) now becomes:

$$\begin{aligned} T_N^\tau(\phi(p_0)) &= -\int_{-\infty}^{0} e^{As} QR(p_\tau(s) + \phi(p_\tau(s)))ds \\ &= -\sum_{k=0}^{N-1} \int_{-(k+1)\tau}^{-k\tau} e^{As} QR(p_k + \phi(p_k))ds - \int_{-\infty}^{-N\tau} e^{As} QR(p_N + \phi(p_N))ds \\ &= -A^{-1}(I - e^{-A\tau}) \sum_{k=0}^{N-1} e^{-k\tau A} QR(p_k + \phi(p_k)) - A^{-1} e^{-N\tau A} QR(p_N + \phi(p_N)). \end{aligned}$$

Under some additional conditions it can be shown that T_N^τ maps $\mathcal{F}_{l,b}$ into itself and that it is a contraction.

The AIMs are now constructed by:

$$\begin{aligned}\phi_0 &\equiv 0\\ \phi_{N+1} &= T_N^{\tau_N}(\phi_N).\end{aligned}$$

It is interesting to note that $\phi_1(p_0) = -A^{-1}QR(p_0)$, the well-known FMT AIM of Section 4.7. In [46] it is shown that if the spectral condition holds, then $\mathcal{M}_N = \mathcal{G}[\phi_N]$ converges to the exact inertial manifold for $\tau_N \to 0, \quad N\tau_N \to \infty$, thus being a *convergent* sequence of AIMs. It can also be shown that for N sufficiently large, $\mathcal{M}_N$ approximate the global attractor at an exponential rate.

Since this construction involves backwards integration of p, it is not suitable for numerical implementation, unless one uses the time discretization p_τ coming from the numerical integration of $p(t)$.

AIMs approximating $\dot{q}$

In [43], Temam developed a method of constructing a sequence of AIMs that not only approximate q but also its higher-order derivatives $q^{(i)}$. In the following we assume that the nonlinear term R and the solution q are sufficiently smooth.

We start off with the projection onto $Q\mathcal{H}$:

$$\dot{q} + Aq + QR(p+q) = 0.$$

The simplest AIM with $q_1 = \phi_1(p)$ is defined to be the FMT AIM:

$$Aq_1 + QR(p+0) = 0.$$

We define $q_2 = \phi_2(p)$ through

$$Aq_2 + QR(p+q_1) = 0.$$

For the j-th AIM with $j > 2$, we no longer neglect $\dot{q}$, but approximate it. We define q_j by:

$$q_{j-2}^1 + Aq_j + QR(p+q_{j-1}) = 0.$$

The approximation q_{j-2}^1 to the time derivative $\dot{q}$ is not yet known. It is constructed by taking time-derivatives of (4.24). This in turn introduces higher time derivatives of p and q, which are approximated by taking higher time-derivatives of (4.23) and (4.24). To solve all dependencies, the following clever recursion is used: two sequences $\{q_j^i\}$ and $\{p_j^i\}$ are defined, where p_j^i and q_j^i are the respective approximations of the j-th AIM to the i-th derivative. The term q_{j-i}^i is defined by

$$q_{j-i-1}^{i+1} + Aq_{j-i}^i + Q\left[R(p_{j-i-1} + q_{j-i-1})\right]^{(i)} \quad = \quad 0 \quad i = 0, \ldots j-1,$$

and p_{j-i-1}^i by

$$p_{j-i-2}^{i+1} + Ap_{j-i-1}^i + P\left[R(p_{j-i-1} + q_{j-i-1})\right]^{(i)} \quad = \quad 0 \quad i = 0, \ldots, j-2.$$

Observe that the j-th AIM uses approximations to time derivatives of p and q coming from earlier AIMs. Thus to evaluate the k-th AIM, one has to compute all j-th AIMs for $j = 0 \ldots k$. The recursion is stopped for the k-th AIM by setting

$$\begin{aligned} q_0^k &= 0 \\ p_k^0 &= p \\ q_k^0 &= q_k. \end{aligned}$$

This construction is another example of sequences of AIMs with ever-increasing accuracy. For details, see [43], [6] and [33].

Chapter 5

Nonlinear Galerkin Methods

Nonlinear Galerkin methods were first introduced in [31]. While the standard Galerkin method solves the PDE by restricting it to a “flat”, finite-dimensional subspace, nonlinear Galerkin methods solve the PDE on an approximate inertial manifold.

There has been much confusion in the literature about the advantages of the nonlinear Galerkin method, both in accuracy and computational time. For example, [8] reported huge computational savings of 40 to 60% over standard Galerkin. However, [14] found exactly the opposite with a more efficient implicit time integrator. With respect to accuracy, [20] reports that nonlinear Galerkin shows an improvement if the basis functions are incompatible with the forcing term at the boundary, whereas experiments in [24] indicate that this is certainly not the only case: if the forcing term can not be approximated well by the basis functions, then nonlinear Galerkin shows the same advantages as in the case of boundary incompatibility.

In the following, we first set up the general framework in which we describe both standard and nonlinear Galerkin methods. We then look at various issues concerning their application.

5.1 General Framework, Standard Galerkin Method

The setting is identical to the previous chapters: We look at a dissipative PDE of the form

$$u_t + Au + N(u) \quad = \quad 0, \tag{5.1}$$

with $u(0) = u_0$ and subject to some boundary conditions. A is a linear, self-adjoint, positive unbounded operator on a domain $D(A)$ which is dense in some Hilbert space $\mathcal{H}$, and N is the nonlinear part. Again, we use an orthonormal basis $\{w_k\}$ of eigenfunctions of A, and write the solution $u(t)$ in this basis as:

$$u(t) \quad = \quad \sum_{k=1}^{\infty} c_k(t) w_k, \quad A w_k = \lambda_k w_k, \quad \langle w_k, w_m \rangle = \delta_{km}.$$

We split $\mathcal{H}$ into the two A-invariant subspaces $P\mathcal{H} = P_n\mathcal{H}$, spanned by the first n eigenfunctions of A, and its infinite dimensional orthogonal complement $Q\mathcal{H} = Q_n\mathcal{H}$ and rewrite (5.1) again as the projections onto these subspaces:

$$\begin{aligned} p_t + Ap + PN(p+q) &= 0 \\ q_t + Aq + QN(p+q) &= 0. \end{aligned} \tag{5.2}$$

The standard Galerkin method (SGM) now computes a solution $y(t)$ inside $P\mathcal{H}$, setting $q \equiv 0$:

$$y_t + Ay + PN(y) \quad = \quad 0. \tag{5.3}$$

Comparing this with (5.2), we see that the **spatial discretization** error committed by the SGM can be split into two orthogonal components (see Figure 5.1): The **subspace approximation error** (SAE) is the error component in $Q\mathcal{H}$ and is identical to $q(t)$. The **subspace integration error** (SIE) is the component in $P\mathcal{H}$, that comes from neglecting $q(t)$ in the nonlinear term during the integration of (5.3). These terms might be non-standard in this context. They are borrowed from [22] where they are used in the setting of model reduction.

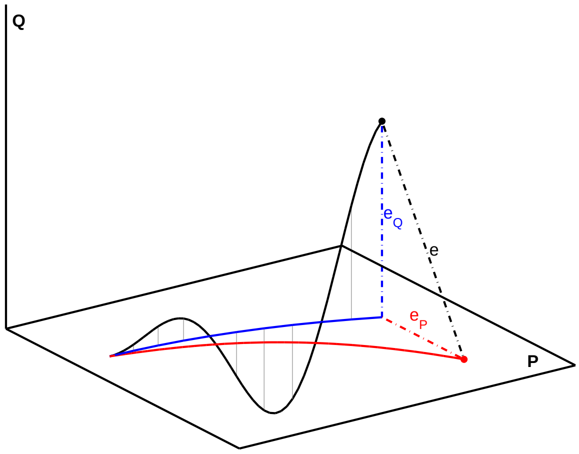

Figure 5.1: Spatial error decomposition

5.2 The Nonlinear Galerkin Method

The nonlinear Galerkin method (NLG) does not neglect $q(t)$ completely, but approximates it using the AIM as a functional relation between $p(t)$ and $q(t)$:

$$q(t) \approx \phi(p(t)),$$

assuming that the AIM is given as the graph of a function ϕ. Put differently, computing a trajectory in $P\mathcal{H}$ (the *flat manifold*), the trajectory is computed on the AIM. As with inertial manifolds, using the (approximate) inertial form, we obtain the following ODE in $P\mathcal{H}$:

$$y_t + Ay + PN(y + \phi(y)) \;=\; 0. \tag{5.4}$$

The solution computed by the NLG is then given by $z(t) = y(t) + \phi(y(t))$.

Observe that evaluating $\phi(y)$ and a higher-dimensional $N(y + \phi(y))$ *during* time integration only serves the purpose to reduce the subspace integration error; if this error were negligible, a single evaluation at the final time $t = T$ would be enough to obtain the improvement in the subspace approximation error, since the $Q\mathcal{H}$-component of the solution is functionally dependent on the component in $P\mathcal{H}$. However, the two error components are *not* independent of each other.

5.3 Rates of Convergence

We have seen in Chapter 4.7 that the quality of an AIM is measured by an estimate on

$$|q(t) - \phi(p(t))|.$$

Here the trajectory $u(t)$ is known in advance and unaffected by the error committed in the AIM. For the NLG, this is different: since ϕ is used to compute $y(t)$, errors caused by the approximate ϕ affect both $y(t)$ and $\phi(y(t))$:

$$|u(t) - z(t)| \;\leq\; \underbrace{|p(t) - y(t)|}_{\text{SIE}} + \underbrace{|q(t) - \phi(y(t))|}_{\text{SAE}}.$$

Similar to the rate of convergence for AIMs, bounds on the rate of convergence of the NLG have been derived in the literature and have the form:

$$|u(t) - z(t)| \;\leq\; C(t)\lambda_{m+1}^{-\gamma} \tag{5.5}$$

(see e.g. [7] for 2D Navier Stokes, [24] for forced Burgers and K-S). The rate of convergence depends on two factors: the regularity of the solution (which in turn depends on the regularity of initial conditions, and the PDE itself, including possible forcing terms), and the quality of the AIM.

For instance, $\gamma = 1 + \alpha$ for the SGM applied to the forced Burgers' equation with a forcing term in $D(A^\alpha)$, but $\gamma = 3/2 + \alpha$ for NLG on the same equation, using the FMT AIM ([24]), and so it is expected that the NLG outperforms the SGM in terms of accuracy, using a fixed number of modes.

Even with these theoretical results in favor of NLG, there are several issues with NLG that one has to be aware of and that we address next.

5.4 Truncating ϕ

The error decay in the previous section assumes that we evaluate ϕ exactly. This is not possible in practice, since ϕ maps into the infinite dimensional space $Q\mathcal{H}$. An implementation has to truncate $\phi(p)$ after a number of modes, say m, i.e. it works with $P_m\phi(p)$ instead of $\phi(p)$. The number m should not be too small so that the gain in accuracy is not affected significantly by the truncation, but it should not be too big either, because the computational cost of evaluating ϕ and the nonlinear term $N(p + \phi(p))$ increases with m. Marion et al. suggested in [31] $m = 2n$, i.e. the NLG scheme finds a solution in $P_{2n}\mathcal{H}$, with the first n modes (primary modes) computed directly, and the second n modes (secondary modes) "enslaved" to the primary modes by ϕ. A more careful analysis, e.g. in [47] or [24], shows that this might not be enough to sustain the faster rate of convergence. The difference between $\phi(p(t))$ and $P_m\phi(p(t))$ can be bounded by $|Q_m q(t)|$, which is the error of the flat manifold spanned by m instead of n basis vectors. For the above example with Burgers' equation, we get a bound of the form:

$$|\phi(p) - P_m\phi(p)| \;\; \leq \;\; K\lambda_{m+1}^{-(1+\alpha)}.$$

Using the triangle inequality, the difference between $q(t)$ and $P_m\phi(p(t))$ is bounded by:

$$|q(t) - P_m\phi(p(t))| \;\; \leq \;\; |q(t) - \phi(p(t))| + |\phi(p(t)) - P_m\phi(p(t))|.$$

Therefore, to keep the total error in the order of (5.5), we need to make sure that

$$\lambda_{m+1}^{-(1+\alpha)} \;\; \sim \;\; \lambda_{n+1}^{-(3/2+\alpha)}$$

and, since $\lambda_m \sim m^2$,

$$m \;\; \sim \;\; (n+1)^{(3/2+\alpha)/(1+\alpha)}.$$

If we choose the forcing term f of Section 2.7, for which $\alpha < 3/4$, we get $m \sim n^{9/7}$.

The issue of truncation described above is in most cases entirely an issue with the forcing function: The most common nonlinear dissipative PDEs have a "weak nonlinearity," for example uu_x for Burgers' equation. For the coefficients $c_k(t)$ this becomes a convolution in Fourier space (see Section 6.1). In this case, if we compute $N(p)$, for example to evaluate the FMT AIM, at most $2n$ coefficients are nonzero. All other coefficients are only nonzero due to the forcing function.

5.5 Appropriate test scenarios for NLG

By design, NLG methods require initial conditions that are sufficiently close to the AIM, since all the dynamics are restricted to the AIM. During the transient phase, in which a solution is far away from the AIM, the NLG method can be expected to perform poorly. This is true for the SGM's flat AIM as well (this just means that one needs a certain number of modes to properly resolve the dynamics of the system)

If the eigenbasis coefficients drop too quickly with $n \to \infty$ (e.g. exponentially, which is typical for (pseudo)spectral methods as long as the solutions are smooth), the NLG will not improve the solution significantly enough, since the improvement in the rate of convergence is only algebraic. In other words, the solutions have to be sufficiently irregular. This can be achieved in two ways:

Firstly, one can look at PDEs that produce steep gradients (shocks). In this case, the Fourier coefficients decay slowly enough, until enough basis functions are used to fully resolve the steep gradient. Burgers' equation produces such shocks for small viscosities.

The second possibility used in numerical experiments is to (artificially) introduce a rough forcing term, such as a spike or discontinuity, that cannot be approximated well by the eigenbasis functions. However, knowing *in advance* that a *given* forcing function can not be well approximated by the *chosen* basis functions indicates that the wrong method is used for the problem. Nevertheless, this method is often used to prove NLG's superiority, as it gives very convincing rate of convergence plots.

5.6 Computational Cost

The evaluation of ϕ and $N(p + P_m\phi(p))$ instead of $N(p)$ increases the computational cost of the NLG. So in practice, the usefulness of NLG depends not only on a higher rate of convergence, but also on the computational cost with which this increase in accuracy is achieved. As soon as there is an SGM implementation using more than n modes, that achieves the same accuracy in less time than an NLG implementation with n primary modes, that particular NLG implementation is rendered useless. A measurement of the computational cost depends on many different factors, including the choice of the time integration method. A study on the computational cost of both SGM and NLG, using an efficient time integrator, has been performed in [14] for the K-S and Allen-Cahn equations. The authors came to the conclusion that “for these problems, the nonlinear Galerkin method is not competitive with either pure spectral Galerkin or pseudospectral discretizations”.

5.7 Postprocessed Galerkin

We mentioned in Section 5.2 that the increased cost of NLG during time integration is used to decrease the subspace integration error. If this error component is negligible for some reason, it is not necessary to evaluate ϕ during integration: here the SGM can be used, and a single ϕ evaluation at the final time $t = T$ gives the same improvement in the subspace approximation error. Since the solution is “lifted” onto the AIM only at the very end, this method is called “Postprocessed Galerkin.” It was first introduced in [15]. Since postprocessed Galerkin is identical to the SGM on $P\mathcal{H}$, it can only improve the subspace approximation error. Interestingly, the authors show several scenarios where such postprocessing yields the same improvements as the NLG, at practically no additional cost to SGM. All experiments involve a rough forcing term. With this method, even very expensive AIM constructions seem to become attractive for practical purposes [33].

Chapter 6

Numerical Experiments

In this chapter we let the SGM, NLG and postprocessed Galerkin (PPG) methods compete on a problem that is tailored to meet all the restrictions we described in the previous chapter, so that we can expect a good performance of NLG.

As the dissipative PDE we choose Burgers' equation, given by

$$u_t - \varepsilon u_{xx} + \frac{1}{2}(u^2)_x \;=\; f, \qquad (6.1)$$

subject to Dirichlet boundary conditions on the interval $[0, \pi]$:

$$u(0,t) = u(\pi,t) = 0 \quad \forall t \geq 0.$$

In the abstract framework the linear dissipative operator is $Au = -\varepsilon u_{xx}$, together with the boundary conditions from above. The orthonormal basis of eigenfunctions of A is given by

$$w_k \;=\; \sqrt{\frac{2}{\pi}}\sin(kx).$$

The PDE is solved in this Fourier basis; the standard Galerkin method is thus a spectral Galerkin method. We can control the decay of the Fourier coefficients of the solution by two parameters: The viscosity ε, decreasing which produces arbitrarily steep gradients (shocks), and the forcing term f, which can be chosen in a way that its Fourier coefficients decay algebraically. Observe that without a rough forcing term, the solutions to (6.1) would be smooth, and we should expect exponential decay of the coefficients, at least after enough basis functions so that the steep gradients are fully resolved.

The multiplication u^2 in physical space (that appears in the nonlinearities of both equations) becomes convolution in Fourier space, which is slow to evaluate. The remedy is FFT, which allows us to quickly ($\mathcal{O}(n \log n)$) switch between Fourier and physical space. Since our basis is not consisting of complex exponentials e^{ikx}, we need a fast sine transform as well as a fast cosine transform, based on FFT. Their construction is explained in Sections 6.1-6.1 below.

For NLG and PPG, we use the FMT AIM. This AIM is expected to perform well because in the above setting, the solutions approach an equilibrium at which the time derivative is actually 0.

As the forcing function we use

$$f(x) \;=\; \begin{cases} 0 & x \in \left[0, \frac{\pi}{4}\right) \\ \frac{4}{\pi}\left(x - \frac{\pi}{4}\right) & x \in \left[\frac{\pi}{4}, \frac{\pi}{2}\right) \\ \frac{4}{\pi}\left(\frac{3\pi}{4} - x\right) & x \in \left[\frac{\pi}{2}, \frac{3\pi}{4}\right) \\ 0 & x \in \left[\frac{3\pi}{4}, \pi\right), \end{cases} \tag{6.2}$$

for which we derived in Chapter 2.7 that $f \in D(A^\alpha)$ for $\alpha \in [0, 3/4)$. As the initial condition, we choose $u_0 = \sin(x)$.

This experiment is commonly found in literature. In fact, it has been presented as numerical evidence for the superiority of the NLG over SGM in [24], and then for the superiority of PPG over NLG in [15]. We give an explanation of why NLG and PPG have higher rates of convergence than SGM, as well as why NLG does not perform better than PPG. Moreover, we identify the main reason for the improvements over SGM and construct a simple static postprocessing scheme that adds the neglected information from f back into the solution and shows that as long as the forcing term is the reason for the better accuracy of NLG/PPG, it gives the same improvements.

6.1 Discrete Fourier Transforms

The discrete Fourier transform (DFT) computes the coefficients $c = [c_0, \ldots, c_{N-1}]$ of the complex Fourier polynomial that interpolates a vector of values $x = [x_0, \ldots, x_{N-1}]$ that are assumed to be values of a periodic function, sampled on a uniform grid over $[0, 1)$:

$$x_m \;=\; p\left(\frac{m}{N}\right) = \sum_{k=0}^{N-1} c_k e^{2\pi i k\left(\frac{m}{N}\right)}. \tag{6.3}$$

Let $\omega_N = e^{\frac{-2\pi i}{N}}$, the N-th root of unity. Then the DFT is given by $c = \mathcal{F}x$, with

$$\mathcal{F}_{mk} \;=\; \frac{1}{N} w_N^{-mk}.$$

This can be easily shown by verifying that $\mathcal{F}^H \mathcal{F} = \frac{1}{N} I_N$ and thus $\mathcal{F}^{-1}_{mk} = w_N^{mk}$ as required in (6.3): the entries of $E = \mathcal{F}^H \mathcal{F}$ are the inner products of the columns of $\mathcal{F}$:

$$E_{lm} \;=\; \frac{1}{N^2} \sum_{k=0}^{N-1} \overline{\omega}_N^{kl} \omega_N^{km} = \frac{1}{N^2} \sum_{k=0}^{N-1} \omega_N^{k(m-l)}.$$

Obviously, if $m = l$, $E_{lm} = 1/N$. If $m \neq l$, then $\omega_N^{k(m-l)} \neq 1$ and thus $E_{lm} = 0$, since:

$$(1 - \omega_N^{m-l})E_{lm} \;=\; \frac{1}{N^2}\left(\sum_{k=0}^{N-1} \omega_N^{k(m-l)} - \sum_{k=1}^{N} \omega_N^{k(m-l)}\right) = \frac{1}{N^2}\left(1 - \omega_N^{N(m-l)}\right) = 0.$$

The remarkable fact that $\mathcal{F}^H\mathcal{F}$ is the identity (up to a scalar factor) has the following interesting consequence: The complex exponentials are not only orthogonal in the L^2-inner product, but also in the $\mathbb{R}^n$-inner product when evaluated on an equispaced grid.

If the vector x is real, the coefficients c satisfy $c_{-k} = \overline{c_k}$ (with $k \in \mathbb{Z}_N$ due to the N-periodicity of ω_N^k):

$$\sum_{k=0}^{N-1} c_k \omega_N^{mk} = x_m \;=\; \overline{x_m} = \sum_{k=0}^{N-1} \overline{c_k}\omega_N^{-mk} = \sum_{k=0}^{N-1} \overline{c_{-k}}\omega_N^{mk}.$$

Therefore we loose a factor of 2 in efficiency by using complex FFT on real data.

Unfortunately, Matlab uses a different convention for FFT: The Fourier transform matrix is defined as $\mathcal{F}_{mk} = \omega_N^{-mk}$, the constant $1/N$ is here in the inverse matrix $\mathcal{F}^{-1}$. However, for us it is more convenient to have the constant inside $\mathcal{F}$, for the following reason: Suppose we want to evaluate the trigonometric polynomial in (6.3) on more than N nodes (say M). Then with our definition of $\mathcal{F}$, we simply pad the coefficients c with the appropriate number of zeros. In Matlab's definition, we would need to correct the constant, too, by multiplying by N/M.

Matlab provides an FFT implementation, but since the eigenfunctions of A are sine waves, we need fast sine/cosine transforms (FST, FCT). These are unfortunately not available in Matlab at this point (Matlab Version 7.0 R14). Luckily, there is a way to compute the DST and DCT using an existing FFT implementation. The two methods below are derived much more elegantly in [48] using block matrices, but the approach presented below seems more intuitive to the author from an interpolation point of view.

Exact Signal Reconstruction, Differentiation

The complex roots of unity satisfy $\omega^{N/2+k} = \omega^{-N/2-k}$. On the grid, frequencies higher than $N/2$ are therefore indistinguishable from corresponding lower frequencies and we could have written (6.3) also in the equivalent form:

$$x_m \;=\; \sum_{k=-N/2}^{N/2-1} c_k e^{2\pi i k\left(\frac{m}{N}\right)}. \tag{6.4}$$

For an exact reconstruction of a continuous, periodic signal from its values at the gridpoints, we must require it to be band-limited, i.e. to have frequencies only in a certain frequency band. As we have seen in the previous section, $c_{-k} = \overline{c_k}$ for a real-valued signal, so it is necessary to choose the frequency band symmetrically around 0. Unfortunately, (6.4) is still not symmetric around 0: the frequency $N/2$ is not included in the sum. This has a strange consequence: The discrete saw-tooth function $\cos(\pi m)$ is reconstructed as a complex-valued function $e^{-i\pi m}$ by (6.4). The problem can be overcome by enforcing symmetry in the following way:

$$\begin{aligned} x_m &= \left(\sum_{k=-N/2+1}^{N/2-1} c_k e^{2\pi i k \left(\frac{m}{N}\right)}\right) + \frac{1}{2} c_{N/2}\left(e^{2\pi i \frac{N}{2}\left(\frac{m}{N}\right)} + e^{-2\pi i \frac{N}{2}\left(\frac{m}{N}\right)}\right) \qquad (6.5) \\ &= \left(\sum_{k=-N/2+1}^{N/2-1} c_k e^{2\pi i k \left(\frac{m}{N}\right)}\right) + c_{N/2}\cos(\pi m). \end{aligned}$$

Note that this formula gives the same result on the grid as (6.3). However, if we need to consider values between gridpoints, this subtle correction becomes important. This is the case when taking derivatives of the Fourier polynomial: since $\sin(\pi m) \equiv 0$ on the grid, every coefficient c_k is multiplied by $2\pi i k$ *except* $c_{N/2}$, which is multiplied by 0. In fact, all odd derivatives have to be treated analogously: $c_{N/2}$ is always multiplied by 0. For even derivatives on the other hand, the cosine term does not disappear, and $c_{N/2}$ is multiplied by the appropriate power of $4\pi^2 k^2$.

Discrete Sine Transform

The discrete sine transform computes the coefficients b of a linear combination of sines that interpolate a given real vector $x = [0, x_1, \ldots, x_{N-1}, 0]$:

$$x_m = \sum_{k=1}^{N-1} b_k \sin\left(\pi k \frac{m}{N}\right) \qquad m = 1, \ldots, N-1. \qquad (6.6)$$

Note that the interpolating function has a period of $2N$ and that it automatically satisfies $x_0 = x_N = 0$. Since

$$\sin\left(\pi k \frac{m}{N}\right) = \frac{1}{2i}\left(e^{2\pi i k \frac{m}{2N}} - e^{-2\pi i k \frac{m}{2N}}\right),$$

(6.6) can be written as

$$x_m = \sum_{k=0}^{2N-1} \frac{1}{2i} b_k \omega_{2N}^{km} \qquad m = 1, \ldots, N-1,$$

with the constraints $b_{-k} = -b_k$ for $k \in \mathbb{Z}_{2N}$ (thus $b_N = 0$). We extend $x \in \mathbb{R}^N$ to $\tilde{x} \in \mathbb{R}^{2N}$ in a way that these constraints are satisfied. Since the interpolating function is $2N$-periodic, we can simply determine x_m for $m = -N+1, \ldots, -1$:

$$x_{-m} \;=\; \sum_{k=0}^{2N-1} \frac{1}{2i} b_k w_{2N}^{k(-m)} = \sum_{k=0}^{2N-1} \frac{1}{2i}(-b_{-k}) w_{2N}^{-km} = -\sum_{k=0}^{2N-1} \frac{1}{2i} b_k w_{2N}^{km} = -x_m. \qquad (6.7)$$

We now have a fast way of computing the DST: first extend x to length $2N$ according to (6.7), then use the FFT to compute the Fourier coefficients, scale the result by $2i$, and keep only the entries $m = 1, \ldots, N-1$.

It is easy to see that the inverse DST is equal to the DST scaled by $N/2$.

Discrete Cosine Transform

The discrete Cosine transform is the DST's counterpart, using cosine waves instead of sines:

$$x_m \;=\; \sum_{k=0}^{N} a_k \cos\left(\pi k \frac{m}{N}\right) \qquad m = 0, \ldots, N. \qquad (6.8)$$

Observe that since the endpoints are not forced to be zero ($x = [x_0, \ldots, x_N]$), one needs $N+1$ cosines, as opposed to $N-1$ sines for the DST. The interpolation function is $2N$-periodic, so x_0 is not necessarily equal to x_N.

Again, we express the cosines in terms of complex exponentials:

$$\cos\left(\pi k \frac{m}{N}\right) \;=\; \frac{1}{2}\left(e^{2\pi i k \frac{m}{2N}} + e^{-2\pi i k \frac{m}{2N}}\right),$$

which allows us to rewrite (6.8) as the Fourier series

$$x_m \;=\; \sum_{k=0}^{2N-1} \frac{1}{2} a_k \omega_{2N}^{km} \qquad m = 0, \ldots, N-1. \qquad (6.9)$$

with the restriction that $a_{-k} = a_k$ for $k \in \mathbb{Z}_{2N}$. Extending x_m for $m = -N+1, \ldots, -1$ using this restriction, we obtain $x_{-m} = x_m$, by the same computation as for the DST.

To compute the DCT quickly, first extend x to length $2N$ using $x_{-m} = x_m$, then use the FFT to compute the Fourier coefficients, scale the result by 2, and keep only the entries $m = 0, \ldots, N$.

The inverse DCT is equal to the DCT scaled by $N/2$.

Convolution by Collocation, Aliasing

Let $u(x)$ and $v(x)$ in $\operatorname{span}\{e^{2\pi i k x} : k = -N/2, \ldots, N/2\}$, and $\hat{u}_k$ and $\hat{v}_m$ their respective Fourier coefficients.

```
function b=FST(x)                  function a=FCT(x);
  b=iFST(x)*2/(size(x,1)+1);         a=iFCT(x)*2/(size(x,1)-1);
end                                end

function x=iFST(b)                 function x=iFCT(a)
  z=zeros(1,size(b,2));              x2=fft([a;a(end-1:-1:2,:)]);
  b2=fft([z;b;z;-flipud(b)]);        x=real(x2(1:end/2+1,:))/2;
  x=-imag(b2(2:end/2,:))/2;        end
end
```

Table 6.1: Fast sine/cosine transforms and their inverses

Then their product $w(x) = u(x)v(x)$ satisfies:

$$\begin{aligned} w(x) &= \left(\sum_{k=-N/2}^{N/2} \hat{u}_k e^{2\pi i k x} \right) \left(\sum_{m=-N/2}^{N/2} \hat{v}_m e^{2\pi i m x} \right) = \sum_{k+m=p} \hat{u}_k \hat{v}_m e^{2\pi i p x} \\ \hat{w}_p &= \sum_{k+m=p} \hat{u}_k \hat{v}_m . \end{aligned}$$

It is easy to see that $\hat{w}(p) = 0$ for $|p| > N$, or $w(x) \in S_{2N}(x)$. Usually, the result is then projected back to the same space S_N by truncating all $\hat{w}_p$ with $|p| > N/2$.

Computing $\hat{w}_k$ this way takes $\mathcal{O}(N^2)$ operations. Using three FFT evaluations, this can be done using $\mathcal{O}(N \log N)$ operations. However, a simple

$$\mathcal{F}_N \left(\mathcal{F}_N^{-1}(\hat{u}) \mathcal{F}_N^{-1}(\hat{v}) \right)$$

introduces **aliasing**: due to the low number of sample points (N), the discrete Fourier transform can not resolve frequencies higher than $N/2$. These appear as ghost-components in the Fourier coefficients with $|p| \leq N/2$. Such aliased spectral methods are called pseudospectral methods [1].

Aliasing can be avoided by **padding**: Before the inverse transform, the coefficients are padded with N zeros:

$$\mathcal{F}_{2N} \left(\mathcal{F}_{2N}^{-1}(\hat{u}) \mathcal{F}_{2N}^{-1}(\hat{v}) \right),$$

and the result is then truncated to the first N entries. This technique gives the same result as evaluating the convolution sum directly.

6.2 Spatial Discretization for Burgers' equation

We can now write up the ODE that is used by the SGM. In the following, $\mathcal{S}$ denotes the DST operator and $\mathcal{C}$ the DCT operator. The DST coefficients are denoted by $\hat{u}$.

The Standard Galerkin Method

The nonlinear term $\widehat{\frac{1}{2}(u^2)_x}$ evaluated in the following way: since u is a linear combination of sines, using the trigonometric equality

$$\sin(kx)\sin(mx) \quad = \quad \frac{1}{2}(\cos((k-m)x) - \cos((k+m)x)), \tag{6.10}$$

we see that u^2 is a sum of cosines, and the inverse DCT gives its coefficients. $\frac{d}{dx}\cos(kx) = k\sin(kx)$, and so the sine coefficients of the nonlinear term are given by

$$k/2 \cdot \left[\mathcal{C}_{2n+2}\left(\mathcal{S}_{2n}^{-1}(\hat{u})^2\right)\right]_{k+1} \quad k = 1, \ldots, n,$$

where $\hat{u}$ is padded to length $2n$.

The full ODE for Burgers' equation is

$$\frac{d}{dt}\hat{u}_k \quad = \quad -\varepsilon k^2 \hat{u}_k - k/2 \cdot \left[\mathcal{C}_{2n+2}\left(\mathcal{S}_{2n}^{-1}(\hat{u})^2\right)\right]_{k+1} + \hat{f}_k \quad k = 1, \ldots, n. \tag{6.11}$$

The Nonlinear Galerkin Method

The ODE obtained for the NLG has almost the same form as (6.11), except for the nonlinear part. Recall that in NLG, $QN(p + \phi(p))$ is evaluated instead of $QN(p)$. The FMT AIM has the form

$$\phi(p) \quad = \quad -A^{-1}QN(p),$$

which yields the coefficients of the secondary modes $k = n+1, \ldots m$, where $m = n^{9/7}$ as derived in Chapter 5.4:

$$\hat{u}_k \quad = \quad \left(\frac{1}{\varepsilon k^2}\right)\left(\hat{f}_k - \left[\mathcal{C}_{2n+2}\left(\mathcal{S}_{2n}^{-1}(\hat{u})^2\right)\right]_{k+1}\right) \quad k = n+1, \ldots, m. \tag{6.12}$$

The thus extended vector is fed into the nonlinear term, so that the inverse DST and DCT work on m- and $m+2$-dimensional vectors instead of $2n$ and $2n+2$:

$$\frac{d}{dt}\hat{u}_k \quad = \quad -\varepsilon k^2 \hat{u}_k - k/2 \cdot \left[\mathcal{C}_{m+2}\left(\mathcal{S}_{m}^{-1}(\hat{u})^2\right)\right]_{k+1} + \hat{f}_k \quad k = 1, \ldots, n. \tag{6.13}$$

6.3 Time Integration: Stiffness, Integrating Factors, BDF

With the spatial discretization settled, we have to deal with the temporal discretization of (6.11) and (6.13). The time integration has to be highly accurate so as not to affect the spatial error behavior that we are primarily interested in.

Table 6.2: Right-hand sides of Burgers' for SGM and NLG, as used by `ode15s`

```
function Su_t=SGMrhs(t,Sp)
  global PDE PDEin;
  p   =iFST([Sp;zeros(PDE.n,1)]);
  Cp2 =FCT([0;p.*p;0])/2;
  NL  =PDEin.N.*Cp2(2:PDE.n+1);
  Su_t=PDEin.A.*Sp-NL+PDEin.Pf;
end
```

```
function Sp_t=NLGrhs(t,Sp)
  global PDE PDEin;
  u   =iFST(PDEin.AIM(Sp,PDE.nQ));
  Cu2 =FCT([0;u.*u;0])/2;
  NL  =PDEin.N.*Cu2(2:PDE.n+1);
  Sp_t=PDEin.A.*Sp-NL+PDEin.Pf;
end
```

Table 6.3: FMT AIM for Burgers' equation

```
function Su=FMTAIM(Sp, nQ)
  global PDE PDEin;
  dQM  = ([(PDE.n+1):nQ]'/2)*2*pi/PDE.dom;
  dQN  = -dQM(1:PDE.n);
  dQAi = 1./(PDE.eps*(dQM.^2));
  p    = iFST([Sp;zeros(PDE.n,1)]);
  Cp2 = FCT([0;p.*p;0])/2;
  QRp = [dQN.*Cp2((PDE.n+2):end-1); zeros(nQ-2*PDE.n,1)];
  Sphi= dQAi.*(-QRp + PDEin.Qf(1:(nQ-PDE.n)));
  Su   = [Sp;Sphi];
end
```

The dissipative linear term$-\varepsilon k^2 \hat{u}_k$ is responsible for the **stiff** behavior of the above ODEs [5]. Explicit time integrators are forced to choose timesteps so small that they become unfeasible. There are several ways around this problem.

In our special case, where the dissipative linear term is diagonal, a change of variables trick frees us entirely from the stiffness problem: Let us write the ODE in question as

$$u_t = Au + N(u),$$

where A is the diagonal linear dissipative term, and $N(u)$ contains the remaining terms of the right hand side. Now, a simple change of variables $U = e^{-At}u$ eliminates the linear term:

$$\begin{aligned} U_t &= e^{-At}u_t - Ae^{-At}u \\ &= e^{-At}(Au + N(u)) - Ae^{-At}u \\ &= e^{-At}N(u) \\ &= e^{-At}N(e^{At}U). \end{aligned}$$

Table 6.4: Butcher table for fourth order Runge-Kutta method

$$\begin{array}{c|cccc}
0 & & & & \\
\frac{1}{2} & \frac{1}{2} & & & \\
\frac{1}{2} & 0 & \frac{1}{2} & & \\
1 & 0 & 0 & 1 & \\
\hline
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6}
\end{array}$$

The **integrating factor method** (IF) performs this change of variables during one time step, more precisely, at time t_n it uses $U = e^{-A(t-t_n)}u$. For the transformed ODE one can now use a regular non-stiff solver, e.g. an explicit Runge-Kutta method.

We implemented a 4th order Runge-Kutta IF method based on the Butcher table in Figure 6.4 ([5]). Besides its simple structure it can be easily extended to embed a third-order Runge-Kutta method for step size control.

Suppose for simplicity we start at $t = 0$ with u_0, and we want to compute u_1 at $t_1 = \Delta t$. Let $f(t, U) = e^{-At}N(e^{At}U)$ be the right-hand side of the transformed ODE. Obviously $U_0 = u_0$. Then the IFRK4 scheme is given by:

$$\begin{array}{rcllcl}
a & = & f(0, & U_0) & = & N(u_0) \\
b & = & f(\Delta t/2, & U_0 + a/2) & = & e^{-A\Delta t/2}N(e^{A\Delta t/2}(u_0 + \Delta t \cdot a/2)) \\
c & = & f(\Delta t/2, & U_0 + b/2) & = & e^{-A\Delta t/2}N(e^{A\Delta t/2}(u_0 + \Delta t \cdot b/2)) \\
d & = & f(\Delta t, & U_0 + c) & = & e^{-A\Delta t}N(e^{A\Delta t}(u_0 + \Delta t \cdot c))
\end{array}$$

$$\begin{aligned}
U_1 &= U_0 + \frac{\Delta t}{6}\left(a + 2(b + c) + d\right) \\
u_1 &= e^{A\Delta t}\left(u_0 + \frac{\Delta t}{6}(a + 2(b + c) + d\right).
\end{aligned}$$

This can be further optimized by observing that some terms cancel out. We redefine the coefficients to be:

$$\begin{aligned}
a &= N(u_0) \\
b &= N(e^{A\Delta t/2}(u_0 + \Delta t \cdot a/2)) \\
c &= N(e^{A\Delta t/2}u_0 + \Delta t \cdot b/2) \\
d &= N(e^{A\Delta t}u_0 + \Delta t \cdot e^{A\Delta t/2}c))
\end{aligned}$$

$$\begin{aligned}
U_1 &= U_0 + \frac{\Delta t}{6}\left(a + 2e^{-A\Delta t/2}(b + c) + e^{-A\Delta t}d\right) \\
u_1 &= e^{A\Delta t}u_0 + \frac{\Delta t}{6}\left(e^{A\Delta t}a + 2e^{A\Delta t/2}(b + c) + d\right).
\end{aligned}$$

In view of the results in [25] for the IFRK4 method on Burgers' equation, we expected it to work well on our problem. We were surprised to find that for increasing numbers of modes, the time discretization error increased dramatically, to a point where the (fixed) step size, that is needed to meet our time discretization error tolerance, became so small that the method becomes infeasible. A step size control based on the embedded third-order Runge-Kutta with FSAL-trick did not reduce the necessary number of steps significantly.

In [25] this problem does not show up because the experiment is done on Burgers' equation without a rough forcing term and with a fixed, moderate number of modes (64). This is the normal setting in which spectral Galerkin methods are expected to perform well.

Later we found out that the poor performance of IFRK4 has been observed before: in [13] the same behavior is described; the author explains why for a fixed time step the error is increasing so rapidly with the number of modes.

Another way to cope with stiffness is to use implicit methods. Implicit methods do not suffer from the severe step size restrictions due to stiffness. They are computationally more expensive, however, since they require the solution of a system of equations at each time step. If the ODE is nonlinear, the system of equations is nonlinear as well, and is solved by a Newton iteration, which requires the Jacobian of the right-hand side.

We chose a BDF (*backward differentiation formula*) solver, for several reasons. First, this was the efficient time integrator that was also used in [14] to compare the computational performance of SGM and NLG, secondly, it is implemented in MATLAB (as `ode15s`), and thirdly, the MATLAB implementation provides control over the computation of the Jacobian, which proved to be essential.

The BDF method is a multi-step method. The idea behind a k-step BDF is the following (see [5], [17] for details): To compute y_{n+1}, a Lagrange interpolation polynomial $p(t)$ of degree k is constructed based on $y_{n-k+1}, \dots, y_{n+1}$. The values y_{n-k+1} to y_n are known. The additional constraint that defines the polynomial uniquely is to require $\frac{d}{dt}p(t_{n+1}) = f(t_{n+1}, y_{n+1})$. This results in a nonlinear system of equations, which is solved by a Newton iteration.

The Jacobian of the right hand side of (6.11) is a dense matrix due to the FFT in the nonlinear part. Computing the full Jacobian becomes unfeasible very quickly ($n > 100$). A natural option is to use sparse approximations of the Jacobian. MATLAB provides a way to specify the sparsity pattern (`odeset` option `JPattern`) To our dismay, the canonical choice of only computing the diagonal elements resulted in terrible convergence problems of the Newton iteration for small values for the error tolerance. We solved this problem by restricting the Jacobian to the *linear* term of (6.11) which is responsible for its stiffness. A is diagonal, and so we provided it as a diagonal approximation to the Jacobian (`odeset` option `Jacobian`). This resolved the convergence problem of the Newton iteration, and after the trouble with IFRK4, we deeply appreciate the efficiency of this time integration scheme. It is interesting to note that the very same IFRK4 scheme was used in [8], where a significant gain in efficiency of NLG over SGM was reported.

Table 6.5: Fourth order integrating-factor Runge-Kutta

```
tic;
  y=PDEin.y0; t0=PDE.T(1); T=PDE.T(end);
  PDEout.t=t0; PDEout.y=y;
  tau=PDE.tau; E=exp(PDEin.A*(tau/2)); E2=E.*E;
  nMax=round((T-t0)/tau);
  nPlot=ceil(nMax/100);  nSave=round(PDE.tts/tau);
  for k=1:nMax
    t = t0+(k-1)*tau;
    a = PDEin.rhs(t,y);
    b = PDEin.rhs(t+tau/2, E.*(y+a*tau/2));
    c = PDEin.rhs(t+tau/2, E.*y +b*tau/2);
    d = PDEin.rhs(t+tau,   E2.*y+E.*c*tau);
    y = E2.*y + (E2.*a + 2*E.*(b+c) + d)*tau/6;
    t  = t0+k*tau;
    if (mod(k,nSave)==0)
      PDEout.t(end+1)= t; PDEout.y(:,end+1)=y;
    end
  end
  if (PDEout.t(end)~=t)
      PDEout.t(end+1)= t; PDEout.y(:,end+1)=y;
  end;
  close(W);
PDE.comptime=toc; PDE.steps=nMax; PDEout.yT=y;
```

Table 6.6: Using MATLAB's BDF solver ode15s

```
PDEout = {}
JP = spdiags(PDEin.A, 0, PDE.n, PDE.n);
tic;
  odeopt = odeset('Stats','on',...
                  'AbsTol', PDE.TOL,...
                  'Jacobian',JP);
  [PDEout.t,PDEout.Sp]=ode15s(@PDEin.rhs,PDE.T,...
                               PDEin.Su0,odeopt);
  PDEout.Sp = PDEout.Sp';
PDE.comptime = toc;
```

6.4 Postprocessing Using Only Information of f

It is important to know how much of the increase in accuracy is solely due to the forcing term f, especially since f is known *a priori*, and NLG/PPG use more terms of its Fourier series than SGM. To examine this, we take another look at the FMT AIM, applied to a PDE with forcing term:

$$q \;=\; -A^{-1}QN(p) + A^{-1}Qf.$$

From this we can construct a simple "dummy AIM," that only uses information from f:

$$q \;=\; A^{-1}Qf. \tag{6.14}$$

In a sense, this AIM mimics only the energy that is fed directly into the secondary modes by the forcing term that are neglected by the SGM. We call it dummy AIM because it does not really deserve the name approximate inertial manifold, since it is completely independent of p. It merely lifts the solution onto a different flat manifold.

The dummy AIM provides an easy way to separate the total improvement of NLG or PPG into a component that exploits the information about f and a part that comes from p.

6.5 Experimental Results

In the following, we describe five interesting test scenarios, each varying the viscosity and the forcing term (i.e. either using exactly (6.2) or an approximation by truncating its Fourier series after a certain number of modes). For each set of parameters, we computed the solution at $T = 40$ obtained using SGM, NLG, PPG, and PPG using the dummy-AIM from Section 6.4 (denoted "DIM" in the plots). At this time the solution has practically reached the equilibrium. As explained in Chapter 5.4, we set the total number of modes considered by NLG and PPG to $\min(2n, n^{9/7})$.

For the time integration, a local error tolerance of 10^{-14} was specified. For each simulation, the global time discretization error was well below 10^{-12}. These values were obtained by running the same simulation with different error tolerances ranging from 10^{-9} to 10^{-14}.

To measure the spatial error, the solutions were compared against a "reference solution," computed using SGM with 8000 modes.

1. For the first set of simulations, we chose the viscosity $\varepsilon = 0.5$ and no truncation of the forcing term. The solution at $T = 40$ is shown in Figure 7.1. It has a rather smooth appearance and there are no steep gradients. Here the slow decay of the Fourier coefficients is caused by the forcing term alone. The NLG clearly outperforms the SGM both in accuracy and rate of convergence (similar to [24]). Moreover, the

PPG shows almost virtually the same convergence as NLG (as observed in [15]). What might be surprising at first is that postprocessing with the dummy AIM gives results that are practically identical to the NLG and PPG. Why doesn't NLG perform better than postprocessing? Is it unable to reduce the subspace integration error, i.e. is the effect of q on p unimportant in this test scenario? These questions can be answered by looking at the subspace integration error (SIE) and subspace approximation error (SAE) separately (Figure 7.2) As it turns out, the NLG does improve the SIE quite noticably. However, the SAE is around two orders of magnitude larger, and thus completely masks the improvement in the SIE. This is the reason why there is no noticable improvement in the total error, compared to a simple postprocessing step. Moreover, it is not necessary to evaluate any complicated AIM for the postprocessing, since the improvement is only due to the forcing term.

2. For the second set (Figures 7.3/7.4), we kept the same viscosity but truncated the Fourier series of the forcing term after 16 modes. As expected, we obtain spectral convergence, i.e. the spatial error was below observable accuracy for all $n \geq 32$.

3. Next we decreased the viscosity to $\varepsilon = 0.05$, using the untruncated f. The solution at $T = 40$ starts developing a steep gradient, which is fully resolved at $n = 160$ (Figures 7.5/7.6). Before this point, the error of DIM is identical to that of SGM. Afterwards, the forcing term takes over, and the situation is essentially the same as in the first case.

4. We decreased the viscosity even further to $\varepsilon = 0.01$, again leaving f untruncated. The solution develops now a noticable shock, which is resolved only for values of n greater than 512. Looking at the total error(Figure 7.7), we notice that the NLG no longer yields significantly better accuracy, and the rate of change is similar to that of SGM. Both SIE and SAE have comparable magnitudes (Figure 7.8). As expected, DIM gives results identical to SGM for $n \leq 512$. Probably the most surprising fact here, however, is that PPG has a larger total error than SGM for $n \leq 64$, the SAE is even less accurate. This is of course not in contradiction to the results on the rate of convergence of the FMT AIM, since the error constants for the FMT AIM and the flat AIM are not the same. But it is good to bear in mind that postprocessing a solution with an AIM might affect the accuracy in a negative way for low n.

5. We repeated the previous scenario but truncated the forcing term after 16 modes. As expected, the result is the same as in the previous scenario until $n = 512$. Afterwards, the errors quickly drop below observable accuracy (Figures 7.9/7.10).

Chapter 7

Conclusions and Outlook

Both NLG and PPG require a slow (i.e. algebraic) decay of the solution's Fourier coefficients. This is the case either if a faster decay is prevented externally by a forcing term that can not be well approximated by the basis functions in the subspace chosen for SGM or if the solution features steep gradients that are difficult to resolve by the basis functions.

We have shown that for Burgers' equation, the impressive gain in accuracy presented in [24] and [15] actually comes directly from the forcing term f. The same gain can be achieved by the dummy AIM of Section 6.4. Thus the improvement stems merely from using available information about f that the SGM neglects. It is arguable if such a scenario establishes the practicality of either PPG or NLG, since it is a clear indicator that maybe the choice an eigenbasis was not appropriate for the problem at hand.

For the case of steep gradients in the solution, the improvements when using NLG or PPG are far less spectacular. Moreover, postprocessing actually decreased the accuracy of the solution for a fairly large range of n.

NLG has been proposed as a method for “large time integration” [6]. However, in typical test scenarios the trajectory moves into a nearby equilibrium, which is not an example of interesting long-term dynamics. A good example of nontrivial dynamics in one space dimension is the Kuramoto-Sivashinsky equation. Numerical experiments in [14] showed no improvement of NLG over SGM in terms of computation time.

The quality of AIMs can be determined most easily using an eigenbasis of the dissipative term. However, an extension to other spatial discretization schemes like Finite Differences (Incremental Unknowns, [45]) or Finite Elements [28], where a slower decay of the coefficients can be expected even for smooth solutions, might prove to be a better setting for nonlinear Galerkin methods.

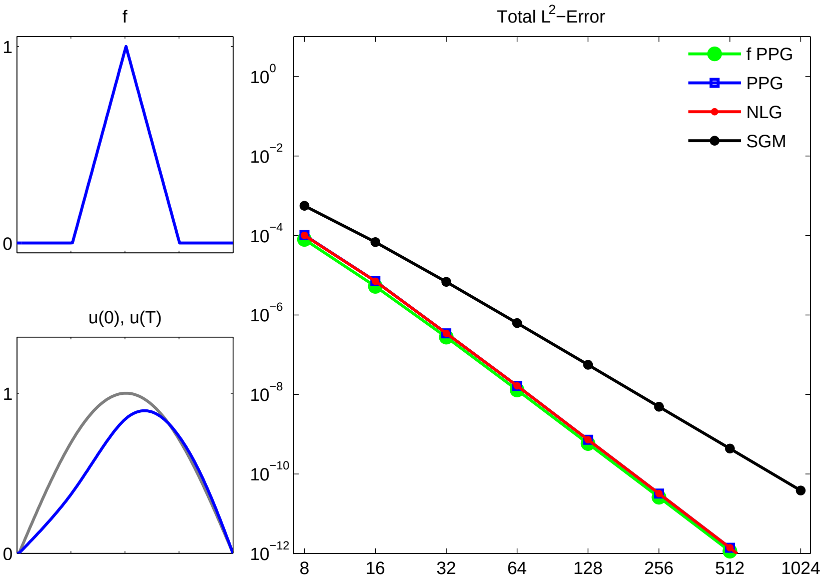

Figure 7.1: Total error for $\varepsilon = 0.5$ and no f-cutoff

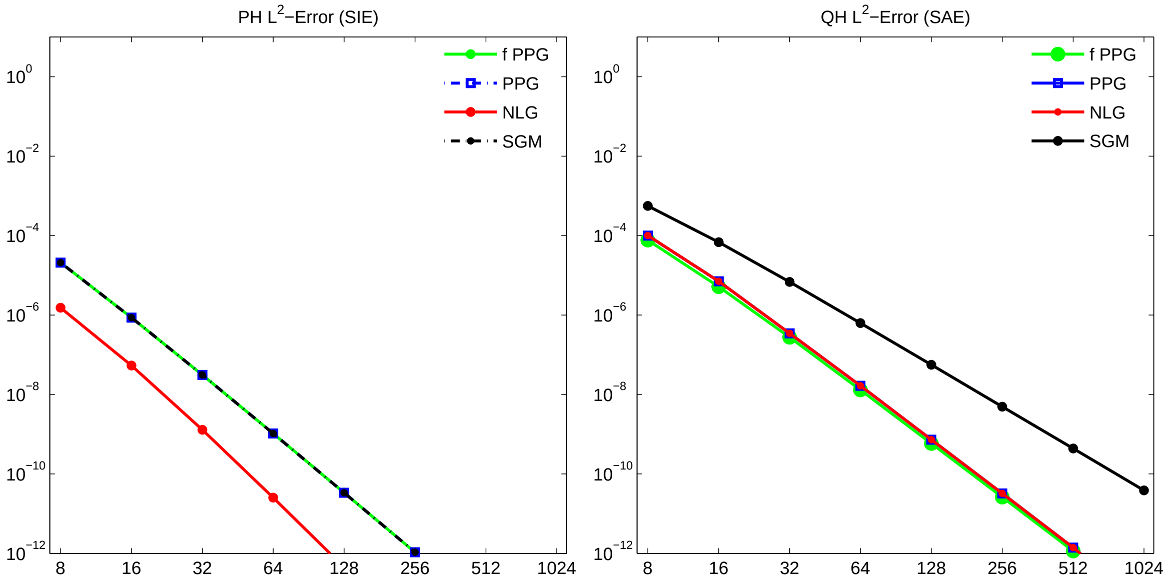

Figure 7.2: SIE and SAE for $\varepsilon = 0.5$ and no f-cutoff

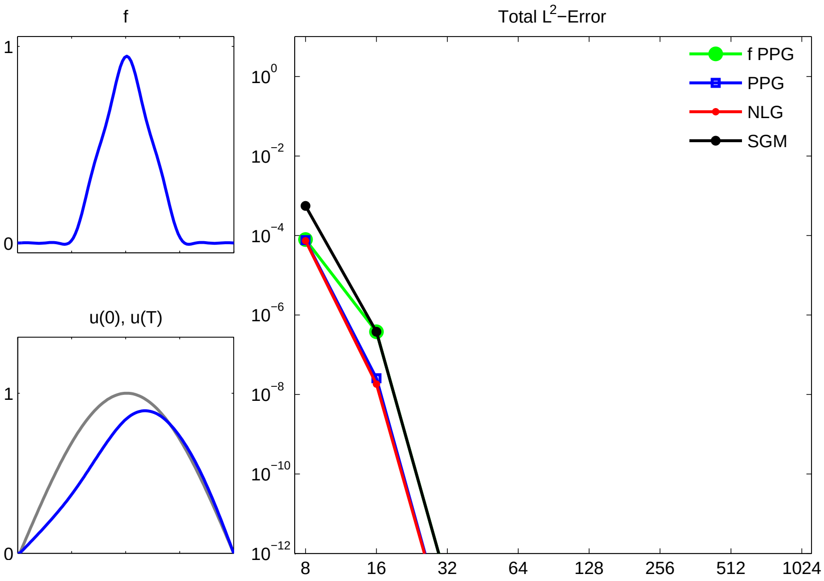

Figure 7.3: Total error for $\varepsilon = 0.5$ and f-cutoff after 16 modes

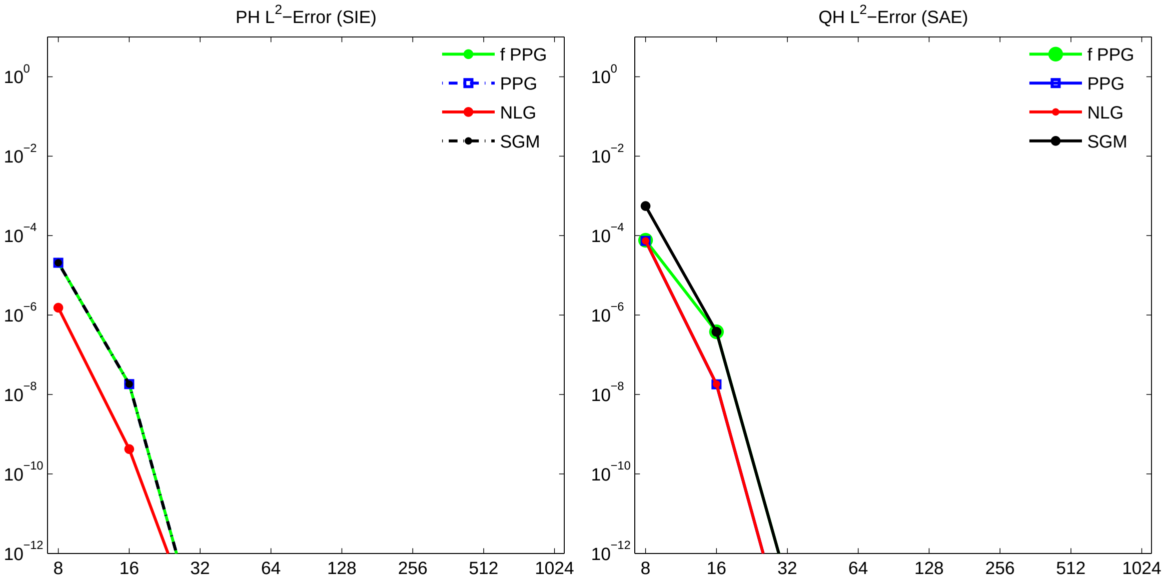

Figure 7.4: SIE and SAE for $\varepsilon = 0.5$ and f-cutoff after 16 modes

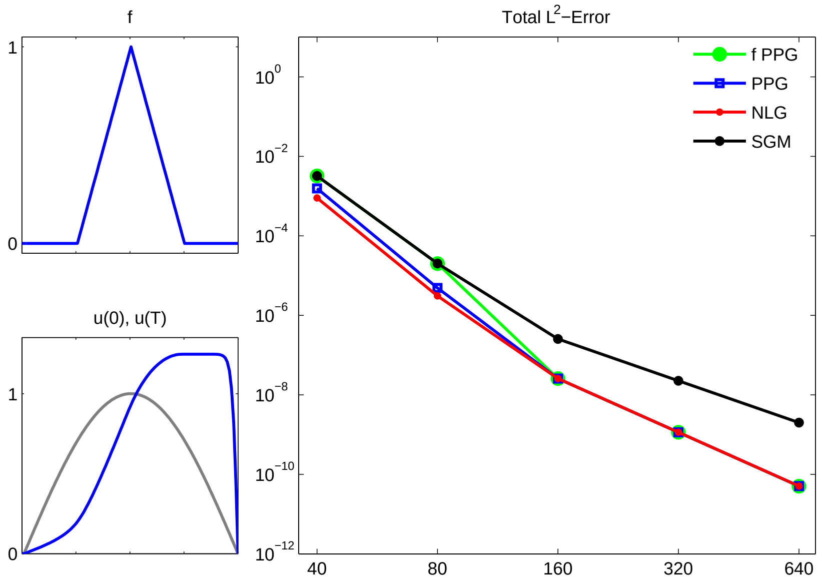

Figure 7.5: Total error for $\varepsilon = 0.05$ and no f-cutoff

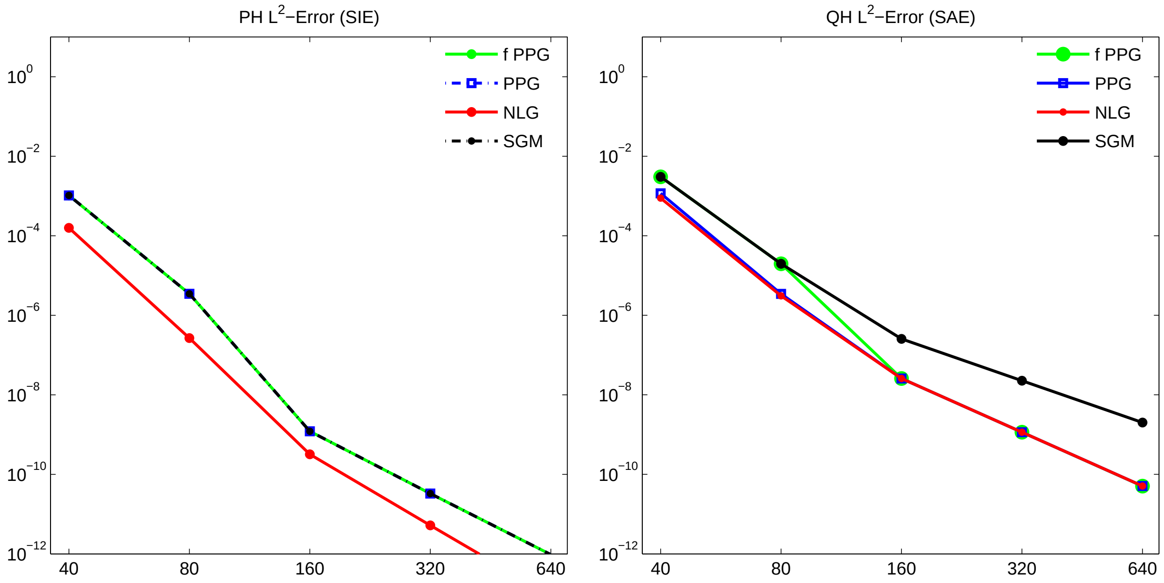

Figure 7.6: SIE and SAE for $\varepsilon = 0.05$ and no f-cutoff

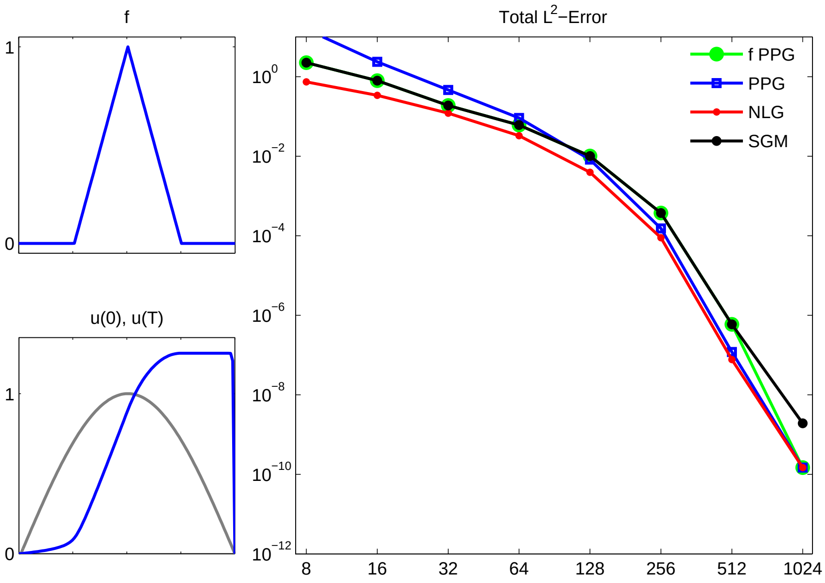

Figure 7.7: Total error for $\varepsilon = 0.01$ and no f-cutoff

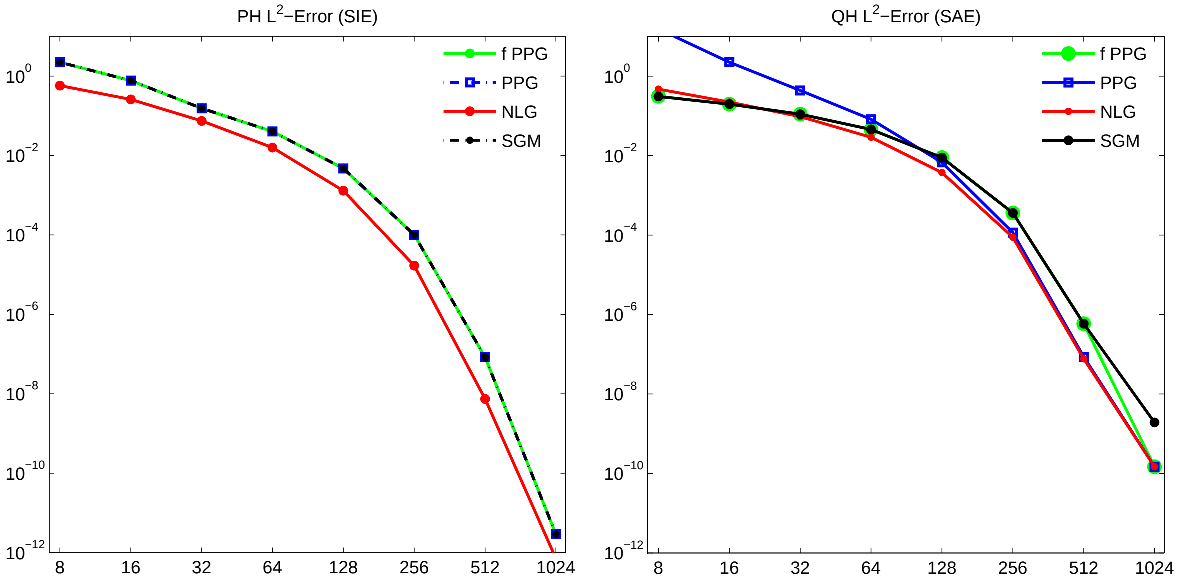

Figure 7.8: SIE and SAE for $\varepsilon = 0.01$ and no f-cutoff

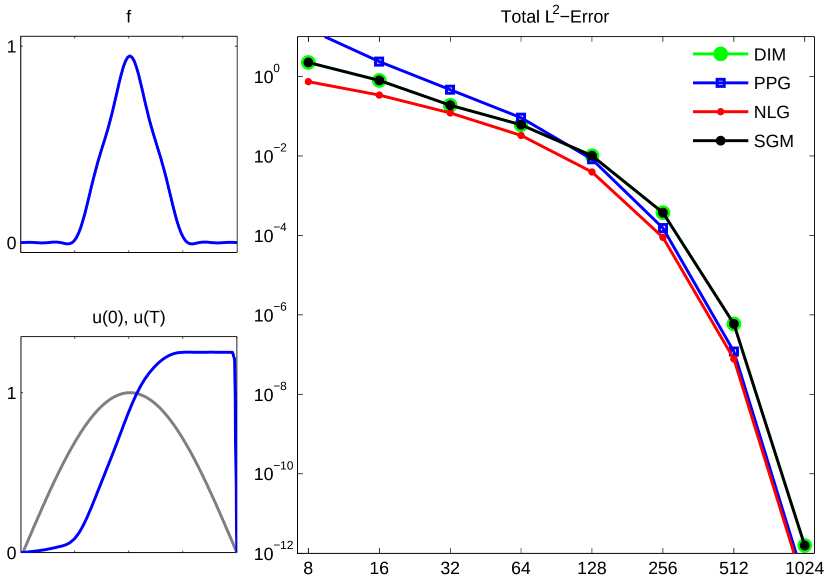

Figure 7.9: Total error for $\varepsilon = 0.01$ and f-cutoff after 16 modes

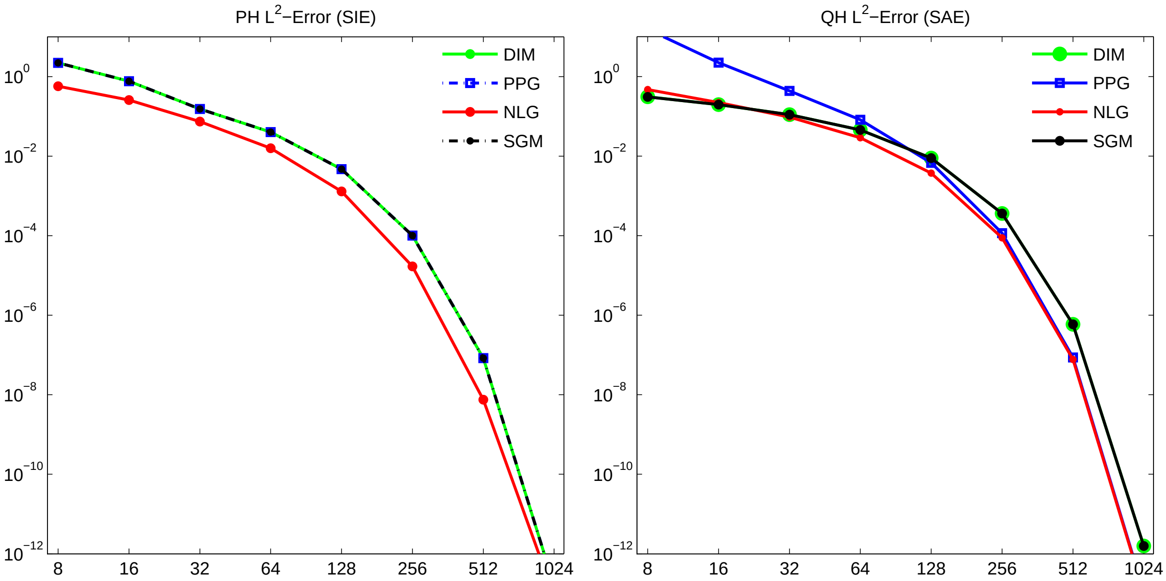

Figure 7.10: SIE and SAE for $\varepsilon = 0.01$ and f-cutoff after 16 modes

Bibliography

[1] Claudio Canuto, M. Yousuff Hussaini, Alfio Quarteroni, and Thomas A. Zang. *Spectral methods in fluid dynamics*. Springer Series in Computational Physics. Springer-Verlag, New York, 1988.

[2] P. Constantin, C. Foias, B. Nicolaenko, and R. Temam. *Integral manifolds and inertial manifolds for dissipative partial differential equations*, volume 70 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1989.

[3] A. Debussche and M. Marion. On the construction of families of approximate inertial manifolds. *J. Differential Equations*, 100(1):173–201, 1992.

[4] A. Debussche and R. Temam. Convergent families of approximate inertial manifolds. *J. Math. Pures Appl. (9)*, 73(5):489–522, 1994.

[5] Peter Deuflhard and Folkmar Bornemann. *Scientific computing with ordinary differential equations*, volume 42 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2002. Translated from the 1994 German original by Werner C. Rheinboldt.

[6] Christophe Devulder and Martine Marion. A class of numerical algorithms for large time integration: the nonlinear Galerkin methods. *SIAM J. Numer. Anal.*, 29(2):462–483, 1992.

[7] Christophe Devulder, Martine Marion, and Edriss S. Titi. On the rate of convergence of the nonlinear Galerkin methods. *Math. Comp.*, 60(202):495–514, 1993.

[8] T. Dubois, F. Jauberteau, M. Marion, and R. Temam. Subgrid modelling and the interaction of small and large wavelengths in turbulent flows. *Comput. Phys. Comm.*, 65(1-3):100–106, 1991.

[9] C. Foias, M. S. Jolly, I. G. Kevrekidis, G. R. Sell, and E. S. Titi. On the computation of inertial manifolds. *Phys. Lett. A*, 131(7-8):433–436, 1988.

[10] C. Foias, O. Manley, and R. Temam. Modelling of the interaction of small and large eddies in two-dimensional turbulent flows. *RAIRO Modél. Math. Anal. Numér.*, 22(1):93–118, 1988.

[11] Ciprian Foias, George R. Sell, and Roger Temam. Inertial manifolds for nonlinear evolutionary equations. *J. Differential Equations*, 73(2):309–353, 1988.

[12] Ciprian Foias, George R. Sell, and Edriss S. Titi. Exponential tracking and approximation of inertial manifolds for dissipative nonlinear equations. *J. Dynam. Differential Equations*, 1(2):199–244, 1989.

[13] Bosco García-Archilla. Some practical experience with the time integration of dissipative equations. *J. Comput. Phys.*, 122(1):25–29, 1995.

[14] Bosco García-Archilla and Javier de Frutos. Time integration of the non-linear Galerkin method. *IMA J. Numer. Anal.*, 15(2):221–244, 1995.

[15] Bosco García-Archilla, Julia Novo, and Edriss S. Titi. Postprocessing the Galerkin method: a novel approach to approximate inertial manifolds. *SIAM J. Numer. Anal.*, 35(3):941–972 (electronic), 1998.

[16] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1987. Nonstiff problems.

[17] E. Hairer and G. Wanner. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1991. Stiff and differential-algebraic problems.

[18] Jack K. Hale. *Asymptotic behavior of dissipative systems*, volume 25 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1988.

[19] Daniel Henry. *Geometric theory of semilinear parabolic equations*, volume 840 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1981.

[20] John G. Heywood and Rolf Rannacher. On the question of turbulence modeling by approximate inertial manifolds and the nonlinear Galerkin method. *SIAM J. Numer. Anal.*, 30(6):1603–1621, 1993.

[21] Philip Holmes, John L. Lumley, and Gal Berkooz. *Turbulence, coherent structures, dynamical systems and symmetry*. Cambridge Monographs on Mechanics. Cambridge University Press, Cambridge, 1996.

[22] C. Homescu, L. Petzold, and R. Serban. Error estimation for Reduced Order Models of Dynamical Systems. *SIAM J. Numer. Anal.*, 2005. To appear.

[23] M. S. Jolly, I. G. Kevrekidis, and E. S. Titi. Approximate inertial manifolds for the Kuramoto-Sivashinsky equation: analysis and computations. *Phys. D*, 44(1-2):38–60, 1990.

[24] D. A. Jones, L. G. Margolin, and E. S. Titi. On the effectiveness of the approximate inertial manifold - a computational study. *Theoret. Comput. Fluid Dynamics*, 7:243–260345, 1995.

[25] Aly-Khan Kassam and Lloyd N. Trefethen. Fourth-order time-stepping for stiff PDEs. *SIAM J. Sci. Comput.*, 26(4):1214–1233 (electronic), 2005.

[26] Mitchell Luskin and George R. Sell. Approximation theories for inertial manifolds. *RAIRO Modél. Math. Anal. Numér.*, 23(3):445–461, 1989. Attractors, inertial manifolds and their approximation (Marseille-Luminy, 1987).

[27] M. Maiellaro and D. Willis, dir. Aqua Teen Hunger Force: Season I-III, 2000-2005.

[28] M. Marion and R. Temam. Nonlinear Galerkin methods: the finite elements case. *Numer. Math.*, 57(3):205–226, 1990.

[29] Martine Marion. Approximate inertial manifolds for reaction-diffusion equations in high space dimension. *J. Dynam. Differential Equations*, 1(3):245–267, 1989.

[30] Martine Marion. Approximate inertial manifolds for the pattern formation Cahn-Hilliard equation. *RAIRO Modél. Math. Anal. Numér.*, 23(3):463–488, 1989. Attractors, inertial manifolds and their approximation (Marseille-Luminy, 1987).

[31] Martine Marion and Roger Temam. Nonlinear Galerkin methods. *SIAM J. Numer. Anal.*, 26(5):1139–1157, 1989.

[32] Arch W. Naylor and George R. Sell. *Linear operator theory in engineering and science*, volume 40 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition, 1982.

[33] Julia Novo, Edriss S. Titi, and Shannon Wynne. Efficient methods using high accuracy approximate inertial manifolds. *Numer. Math.*, 87(3):523–554, 2001.

[34] Michael Reed and Barry Simon. *Methods of modern mathematical physics. I. Functional analysis*. Academic Press, New York, 1972.

[35] Michael Renardy and Robert C. Rogers. *An introduction to partial differential equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2004.

[36] J. C. Robinson. Finite-dimensional behavior in dissipative partial differential equations. *Chaos*, 5(1):330–345, 1995.

[37] J. C. Robinson. Inertial manifolds and the strong squeezing property. In *Nonlinear evolution equations & dynamical systems: NEEDS '94 (Los Alamos, NM)*, pages 178–187. World Sci. Publishing, River Edge, NJ, 1995.

[38] James C. Robinson. Inertial manifolds and the cone condition. *Dynam. Systems Appl.*, 2(3):311–330, 1993.

[39] James C. Robinson. A concise proof of the "geometric" construction of inertial manifolds. *Phys. Lett. A*, 200(6):415–417, 1995.

[40] James C. Robinson. *Infinite-dimensional dynamical systems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2001. An introduction to dissipative parabolic PDEs and the theory of global attractors.

[41] James C. Robinson. Computing inertial manifolds. *Discrete Contin. Dyn. Syst.*, 8(4):815–833, 2002.

[42] Robert D. Russell, David M. Sloan, and Manfred R. Trummer. Some numerical aspects of computing inertial manifolds. *SIAM J. Sci. Comput.*, 14(1):19–43, 1993.

[43] R. Temam. Attractors for the Navier-Stokes equations: localization and approximation. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.*, 36(3):629–647, 1989.

[44] R. Temam. Inertial manifolds. *Math. Intelligencer*, 12(4):68–74, 1990.

[45] R. Temam. Inertial manifolds and multigrid methods. *SIAM J. Math. Anal.*, 21(1):154–178, 1990.

[46] Roger Temam. *Infinite-dimensional dynamical systems in mechanics and physics*, volume 68 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition, 1997.

[47] Edriss S. Titi. On approximate inertial manifolds to the Navier-Stokes equations. *J. Math. Anal. Appl.*, 149(2):540–557, 1990.

[48] Charles van Loan. *Computational frameworks for the fast Fourier transform*, volume 10 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

Vita

Denis Kovacs was born 04/13/1980 in Munich, Germany. He grew up in Burghausen and did his alternative national service in Altötting in 1999/2000 before he enrolled in Computer Science at Technical University of Munich in October 2000. In Fall 2001 he switched his major to Mathematics and received his Prediploma in Summer 2002. After Ferienakademie 2003, where his advisor Andreas Krahnke told him about his studies at Virginia Tech, Denis decided to apply to the Virginia Tech's Mathematics department for a Master's degree. He began his studies in August 2004 and is expected to obtain his Master's Degree in Winter 2005. He will start an internship at NVidia in January 2006 and plans to enroll as a PhD student in Fall 2006.