Energy Conservation by Geometric Numerical Integrators

Thomas Sales

University of Bath
2020/2021

Project presented for MMath in Mathematics, under supervision of Dr Pranav Singh.

Acknowledgements

Firstly, I'd like to thank my friends and family for putting up with me every time I said "My project is actually pretty cool". I only hope they retained some interest, and that I haven't worsened the stigma that mathematics seems to have with the general population. I'd also like to thank my supervisor, Dr Pranav Singh, for his general guidance with this project, and for making sure that I didn't make any obvious errors in writing this report.

Contents

Chapter 1

The Schrödinger Equation

1.1 Introduction

Partial differential equations (PDEs) are complicated, and their complexity is only matched by their utility. Apart from some simple examples, PDEs must be solved numerically, and such numerical schemes are of interest in this dissertation. Specifically, we will be investigating exponential splitting methods for the time-dependent Schrödinger equation[1],

$$i\frac{\partial u}{\partial t} = (-\Delta + V)\, u = Hu, \qquad u(0) = u_0 \in L^2(\mathbb{R}^d; \mathbb{C}). \tag{1.1}$$

The operator H, as seen above, is the Hamiltonian operator for this system, composed of a Laplacian (corresponding to the kinetic energy of the system), and a multiplicative term (corresponding to the potential energy of the system). The function $V(x,t)$ is called the potential, and if this is identically zero then we call (1.1) the free Schrödinger equation. Similarly, if the potential has no time dependence, then we say the Hamiltonian is time-independent. In this dissertation, we assume the Hamiltonian is time-independent, unless stated otherwise, as exponential splittings aren't immediately applicable in the presence of time-dependence.

The aim of this dissertation is to introduce the reader to relevant concepts for the analysis of geometric numerical integrators. A brief introduction to each section is given at the beginning the appropriate chapter, and a more detailed overview of the content is as follows.

1.2 Dissertation Overview

The remainder of this first chapter discusses properties of the exact solution of (1.1), and of some selected numerical solutions. Section 1.3 focuses on the exact solution, introducing some conserved quantities of this system. We refer the reader to [3] for

[1] This formulation is the abstract Schrödinger equation, where we have chosen convenient physical units e.g. $\hbar = 1$.

details similar to this. Then, section 1.4 introduces some numerical schemes of interest, and provides some examples illustrating both why they are needed, and how they preserve the conserved quantities from the previous section.

The second chapter discusses a variant of the problem (1.1) in the form of an ordinary differential equation (ODE). Section 2.1 introduces this problem, and relates it back to (1.1). The remaining sections closely follows the content of [1]. Section 2.2 introduces many of the ideas to be used throughout, namely the Baker-Campbell-Hausdorff formula and the adjoint operator. Section 2.3 then utilises these to analyse symplectic methods, which have similar properties to the exponential splittings we are interested in. This problem is well understood, and the analysis serves as a nice introduction for that of the PDE. Despite this, the techniques from this section cannot be used in the analysis of splitting methods applied to (1.1), due to the unbounded operators defining the Hamiltonian, H. We refer the reader to [1], [2] for further reading relevant to this chapter.

Chapter 3 investigates why the analysis for splitting methods applied to (1.1) is difficult. Section 3.1 begins by illustrating some problems with the kind of arguments used in the previous chapter, in particular we look at why one must be careful when defining the exponential of an unbounded operator. Sections 3.2/3.3 introduce some functional analysis from [7]. This is required for us to give a more thoughtful approach to the analysis, introducing self adjointness and the spectral theorem for self adjoint operators. These are important tools for the analysis of these methods, and are used throughout the final two chapters. Section 3.4 follows the error analysis presented by Erwan Faou in [2], and discusses how it works where the analysis presented in the previous chapter doesn't. This proof is expanded upon in section 3.5, where we consider extending Faou's technique to a higher order splitting.

Lastly, chapter 4 considers (1.1) with a time dependent potential, for which we no longer know the exact solution and cannot use exponential splittings. Section 4.1 introduces the Magnus expansion (see [17]), and Magnus based methods, as a way for us to approximate the solution. Section 4.2 gives an error bound, from [5], for such a method, and investigates the long term behaviour of Magnus based methods. Finally, section 4.3 looks at a new approach to analysing exponential splittings, utilising the Magnus expansion, and techniques from [4] and [5]. We refer the reader to [15] for background information on Magnus based methods.

1.3 Solution Structure

The solution of (1.1) is given explicitly by

$$u(x,t) = e^{-itH}u_0(x), \tag{1.2}$$

where the operator e^{-itH} is defined using the functional calculus for a self adjoint linear operator (see section 3.3). In this section we discuss some of the properties of

this solution.

The physical significance of the Schrödinger equation imposes some ideal properties on an integrator. This is similar to a classical mechanical system, which may be more familiar to the reader. It is well known that classical mechanical systems, with no external forces, can be shown to have a conservation of energy. As such, when one resorts to numerical solutions for this system, they may want a similar conservation law. This motivates the idea of geometric numerical integration, which uses the underlying geometry (see [15]) of these problems to create useful numerical schemes. A subclass of geometric numerical integrators are symplectic methods, of which a well known example is the Störmer-Verlet method (see [11]). This method is known have an approximate energy conservation for classical Hamiltonian systems. Similarly, the Schrödinger equation has the following conserved quantities, which one may want an integrator to conserve.

Firstly, for a solution of (1.1) the Born interpretation says that the square of the absolute value corresponds to a probability density function. More specifically we have the following:

Definition 1.3.1 *Let $u_0 \in L^2(\mathbb{R}^d;\mathbb{C})$ be such that $\|u_0\|_{L^2} = 1$, and let u be the solution, (1.2), for this u_0. Then for a Borel set $E \subseteq \mathbb{R}^d$, the probability that the particle described by (1.1) is in E at time t is given by*

$$\int_E |u(x,t)|^2 dx$$

In particular we see $\|u\|_{L^2} = 1$ for all time. This is justified mathematically by the fact that the operator e^{-itH} is unitary for all time, and thus,

$$\|u(t)\|_{L^2}^2 = \langle e^{-itH}u_0, e^{-itH}u_0\rangle = \langle u_0, u_0\rangle = \|u_0\|_{L^2}^2 = 1.$$

Definition 1.3.2 (Energy) *Let u be the solution of (1.1). Then the energy of the system is $\langle u, Hu\rangle$. This is a constant determined by u_0.*

Similarly to the above, we prove that the energy along solutions of (1.1) is constant. Our energy is a function of time $E(t) = \langle u(t), Hu(t)\rangle = \langle e^{-itH}u_0, He^{-itH}u_0\rangle$. We now note that H and e^{-itH} are commutative operators, see chapter 3, and hence we may write $\langle e^{-itH}u_0, He^{-itH}u_0\rangle = \langle e^{-itH}u_0, e^{-itH}Hu_0\rangle = \langle u_0, Hu_0\rangle$, after again using the fact that e^{-itH} is a unitary operator. Thus we have lost all time dependence, and the energy is a constant determined by u_0, as claimed.

These can be proven by use of a stronger result for general observables, which is given in [3].

Theorem 1.3.3 (Heisenberg Equation) *Let $A : D_A \to L^2(\mathbb{R}^d;\mathbb{C})$ be a self adjoint operator (we call this an observable), with domain $D_A \subseteq L^2(\mathbb{R}^d;\mathbb{C})$. Let $u(t)$ be a solution of (1.1). Then,*

$$\frac{d}{dt}\langle u(t), Au(t)\rangle = -ie^{itH}[A,H]e^{-itH},$$

where square brackets denote the commutator of operator, $[A, B] = AB - BA$.

From the Heisenberg equation, we can obtain the conservation of norm by taking $A = \mathrm{id}$, and the conservation of energy by taking $A = H$. In both cases it is clear that A and H will commute, i.e. $[A, H] = 0$, and hence $\langle u(t), Au(t)\rangle$ has no time dependence.

The conservation of energy is a well known result in the context of classical mechanics where the Hamiltonians are instead functions (see [12]). Another well known property of classical Hamiltonian flows is that of symplecticity.

Definition 1.3.4 (Symplecticity) *Let* $A : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ *be a linear map. We say* A *is symplectic if*

$$A^T J A = J = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix},$$

where I_n *is the* $n \times n$ *identity matrix.*

Let $U \subseteq \mathbb{R}^{2n}$ *be open. We say a differentiable function* $g : U \to \mathbb{R}^{2n}$ *is symplectic if its derivative is symplectic, in the above sense, at all points.*

This result is often interpreted geometrically as a conservation of area [1]. As mentioned above, the flow of a classical Hamiltonian system is symplectic, which imposes another property we might like our numerical integrator to have.

1.4 Numerical Methods

In the previous section we gave a solution, (1.2), for (1.1). This was in terms of an operator, e^{itH}, which we have not yet defined. For the free Schrödinger equation, we can give a more concrete solution via the Fourier transform [3].

$$u(x,t) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} e^{i(k\cdot x - |k|^2 t)} \hat{u_0}(k)\, dk,$$

where $\hat{u_0}(k) = \mathcal{F}[u_0](k)$ is the Fourier transform of u_0. While this is useful for the free Schrödinger equation, such a closed form solution is not applicable when we have a non-zero potential. Hence, in practical applications, one resorts to some sort of numerical method.

Before resorting to any spatial discretization, we first think about simplifying the operator e^{-itH}. Suppose that we have operators T and V on a Hilbert space, $\mathcal{H}$, such that $H = T + V$. We would like to be able to split the exponential operator in a way similar to the exponential of complex numbers, $e^{w+z} = e^w e^z$. However, this property does not extend to operators[2], which can easily be illustrated by considering the

[2]In fact, this will hold if, and only if, the operators commute.

exponential of matrices. Despite this, we have the following result. If T, V, H are all self adjoint on their respective domains (see chapter 3), then we have the Lie-Trotter formula:

$$e^{-itH}\phi = \lim_{n\to\infty} \left(e^{-itT/n}e^{-itV/n}\right)^n \phi,$$

where $t \in \mathbb{R}$ and $\phi \in \mathcal{H}$.

This motivates the Lie-Trotter Splitting,

$$\exp(-ihH)u_0 \approx \exp(ih\Delta)\exp(-ihV)u_0, \tag{1.3}$$

where h is some sufficiently small timestep. One can then compose this method in such a way to obtain a more accurate splitting, known as the Strang splitting,

$$\exp(-ihH)u_0 \approx \exp\left(\frac{-ihV}{2}\right)\exp(ih\Delta)\exp\left(\frac{-ihV}{2}\right)u_0. \tag{1.4}$$

These are aptly named exponential splittings, and are examples of Lie group methods [15]. It can be shown (see [6]) that (1.3) is a first order method, and that (1.4) is a second order method. The analysis of these methods is similar to that of symplectic methods, which we will see later.

We now consider spatially discretizing the system in some way, so that we may find an approximate solution using the above. An example of such a discretization is to use finite difference methods to approximate the Laplacian and potential operator on some spatial grid. We discuss this in more detail for $d = 1$.

Firstly, we have to choose a finite domain, as we cannot perform this computation over $\mathbb{R}$ immediately. This "truncation of domain" will be dependent on the initial data. In the following examples we took our spatial domain to be $[-10, 10] \subset \mathbb{R}$. This was sufficient for our purposes, as we consider Gaussian initial data in a confining potential, and so there is minimal loss of information here. We then approximate our infinite dimensional operators as follows. For our spatial step size Δx, we approximate the Laplacian using a central finite difference scheme

$$\mathcal{K}_2 = \frac{1}{\Delta x^2}\begin{pmatrix} -2 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 1 & -2 & 1 \\ 0 & \dots & 0 & 1 & -2 \end{pmatrix}.$$

Similarly, we approximate the multiplicative operator $V(x)$ by the diagonal matrix $D = \mathrm{diag}(V(X_0), V(X_0 + \Delta x), ..., V(X_1 - \Delta x), V(X_1))$, where X_0, X_1 are the end points of the "truncated domain".

Given these approximations, we could just immediately approximate the exponential operator as $\exp(-itH) \approx \exp(it(\mathcal{K}_2 - D))$. In practise, this can be computed

using MATLAB's expm function, which use Padé methods [19]. This leads to higher accuracy, but at a higher computational cost. However, as we incur some error in the spatial discretization, it isn't unreasonable for us to pick a method with the same error, at a minor cost. This is where the exponential splittings become relevant as these are more computationally feasible than the direct exponentiation. If our discretization yields $N \times N$ matrices, then direct exponentiation using Padé methods will be $\mathcal{O}(N^3)$ in time. However, exponentiation of $\mathcal{K}_2$ will be $\mathcal{O}(N \log(N))$, and exponentiation of D will be $\mathcal{O}(N)$, since $\mathcal{K}_2$ can be diagonalised using the fast Fourier transform, and D is a diagonal matrix.

In figures (1.1) and (1.2) we compare the accuracy of these methods, where we considered our spatial domain to be $[-10, 10] \subset \mathbb{R}$, a potential function $V(x) = x^4 - 10x^2$, and initial data $u_0(x) = \exp\left(\frac{-(x+2.5)^2}{0.2}\right)$.

The plots in (1.1) and (1.2) illustrate the need for more specialised methods, such as the Lie-Trotter splitting (1.3) and the Strang splitting (1.4), as the most straightforward choices of integrators perform very poorly. In fact, the forward Euler solution is so unstable that it doesn't complete the full time integration, and quickly blows up to infinity.

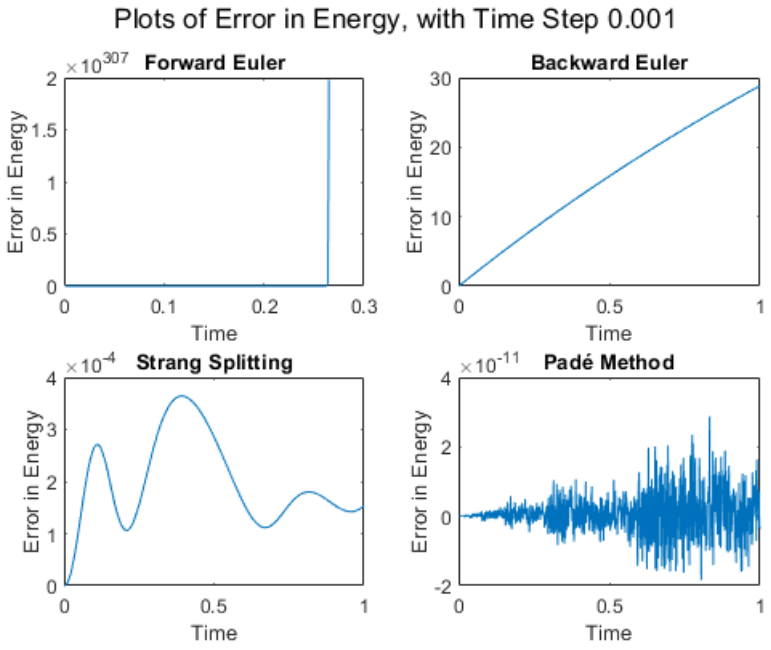

Figure 1.1: Comparison of how accurately various numerical methods preserved the energy of the system.

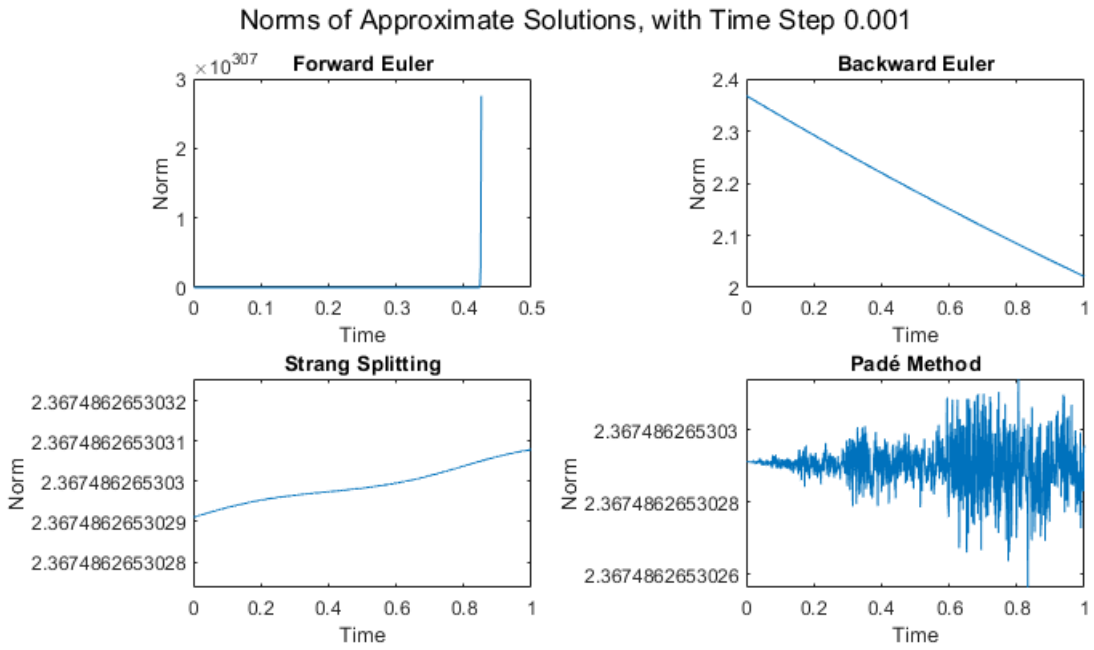

Figure 1.2: Comparison of how accurately various numerical methods preserved the norm of the system.

Chapter 2

Finite Dimensional Case

In this chapter, we consider a simplification of the system (1.1), and the associated analysis. Section 2.1 introduces the problem, and relates it back to (1.1). Section 2.2 introduces the necessary components for the analysis, in particular we introduce the Baker-Campbell-Hausdorff formula and the adjoint operator. Section 2.3 utilises the previous section and presents the analysis of symplectic methods, as in [1], which is very similar to that of the exponential splittings we're interested in.

2.1 The Finite Dimensional Problem

We now consider the autonomous ODE,

$$\frac{dy}{dt} = f(y) = f^{[1]}(y) + f^{[2]}(y), \qquad y(0) = y_0. \tag{2.1}$$

In a way not dissimilar to the previously mentioned splitting methods, we would like to relate the solution to this equation to the solutions of the equations

$$\frac{dy}{dt} = f^{[1]}(y), \qquad \frac{dy}{dt} = f^{[2]}(y). \tag{2.2}$$

We denote the respective flows of these differential equations as $\varphi_t^{[1]}, \varphi_t^{[2]}$. Then for a small time step h, we have our splitting methods given by:

$$\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]},$$

the Lie-Trotter splitting, and

$$\Phi_h^{[S]} = \varphi_{\frac{h}{2}}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{\frac{h}{2}}^{[1]},$$

the Strang splitting.

In particular, we will look at a finite dimensional Schrödinger equation, which is of the form

$$i\frac{dy}{dt} = Ay + By, \tag{2.3}$$

for some $A, B \in M_{n,n}(\mathbb{C})$. In this case it is well known that the flow will be $\exp(-it(A+B))$, and we have similar results for the separated equations. This is particularly relevant, as after considering some discretization in space, this is precisely the form we use for computing numerical solutions of (1.1). For example, taking $A = -\mathcal{K}_2$ and $B = D$ from section 1.4. yields precisely the system we solved for our numerical examples.

We will consider the analysis of more complicated examples, but will frequently refer to this example of interest.

2.2 The Baker-Campbell-Hausdorff Formula

Typically when one investigates the properties of a numerical integrator we would consider the relation between the actual solution and the numerical solution. This is the idea of *forward error analysis*. An alternate approach, which we take, is to consider what differential equation is solved by the method in question. This is known as *backward error analysis*. A concrete example of this is to consider the problems in the previous section and instead ask what ODE is solved by the flow $\Phi_t = \varphi_t^{[1]} \circ \varphi_t^{[2]}$. In order to investigate this we require some definitions and results, which comprise this section.

Definition 2.2.1 (Adjoint Operator) *Let $A, B \in M_{n,n}(\mathbb{C})$. We define the adjoint operator for A as*

$$ad_A(B) = [A, B] = AB - BA.$$

We use this to recursively define the composition of adjoint operators by

$$ad_A^k(B) = ad_A(ad_A^{k-1}(B)),$$

for some $k \in \mathbb{N}$. We also have the convention that $ad_A^0(B) = B$.

This will be used throughout this section. Next we state, and prove, a lemma relating to the derivative of an exponential.

Lemma 2.2.2 (Derivative of Exponential) *Let $Z : \mathbb{R} \to M_{n,n}(\mathbb{C})$ be a differentiable map. Then we have the following:*

$$\frac{d}{dt}\exp(Z(t)) = \left[\left(\frac{\exp(ad_{Z(t)}) - 1}{ad_{Z(t)}}\right)\frac{dZ}{dt}\right]\exp(Z(t)),$$

where

$$\frac{\exp(ad_{Z(t)}) - 1}{ad_{Z(t)}} := \sum_{k=0}^{\infty} \frac{1}{(k+1)!} ad_{Z(t)}^k.$$

We mimic the proof presented in [2], starting by defining a function,

$$U(s,t) = \left(\frac{d}{dt}\exp(sZ(t))\right)\exp(-sZ(t)).$$

We then compute the partial derivative in the first argument as

$$\frac{\partial U}{\partial s} = \left(\frac{d}{dt}Z(t)\exp(sZ(t))\right)\exp(-sZ(t)) - \left(\frac{d}{dt}\exp(sZ(t))\right)Z(t)\exp(-sZ(t)),$$

which can then be simplified by using the product rule on the first term, and noting that $Z(t)$ and $\exp(-sZ(t))$ commute. This yields

$$\frac{\partial U}{\partial s} = \frac{dZ}{dt} + ad_{Z(t)}(U(s,t)).$$

From this, we solve using the variation of constants formula, obtaining

$$U(s,t) = \exp(s\,ad_{Z(t)})U(0,t) + \int_0^s \exp((s-\sigma)ad_{Z(t)})\frac{dZ}{dt}\,d\sigma.$$

One then notes that $U(0,t) = 0$, and performs a change of variables to see that

$$U(1,t) = \int_0^1 \exp(\sigma\,ad_{Z(t)})\frac{dZ}{dt}\,d\sigma.$$

Finally, we evaluate this integral and use the definition of $U(s,t)$ to obtain the result,

$$\frac{d}{dt}\exp(Z(t)) = \left(\sum_{k=0}^{\infty}\frac{1}{(k+1)!}ad_{Z(t)}^k\frac{dZ}{dt}\right)\exp(Z(t)),$$

as claimed. ■

In related literature this may also be written using the dexp map,

$$\frac{d}{dt}\exp(Z(t)) = \left(\sum_{k=0}^{\infty}\frac{1}{(k+1)!}ad_{Z(t)}^k\frac{dZ}{dt}\right)\exp(Z(t)) =: \mathrm{dexp}_Z\left(\frac{dZ}{dt}\right)\exp(Z(t)).$$

We also give a lemma relating to the inverse of the dexp map.

Lemma 2.2.3 *We define the Bernoulli numbers via the generating function:*

$$\frac{z}{e^z-1} = \sum_{k=0}^{\infty}\frac{B_k}{k!}z^k,$$

where we have convergence for $|z| < 2\pi$. We then find similarly that for $Z \in M_{n,n}(\mathbb{C})$ with $\|Z\| < 2\pi$ that

$$\mathrm{dexp}_Z^{-1} = \left(\frac{\exp(ad_Z)-1}{ad_Z}\right)^{-1} = \sum_{k=0}^{\infty}\frac{B_k}{k!}ad_Z^k.$$

These results are then used when we consider the following problem.

Given two matrices $A, B \in M_{n,n}(\mathbb{C})$, we would like to know if we may write the product of their exponentials as $\exp(tA)\exp(tB) = \exp(C(t))$ for some matrix valued function $C : \mathbb{R} \to M_{n,n}(\mathbb{C})$. As mentioned in section 1.4, this is not a trivial problem, as it is when considering scalars instead of matrices. This is due to the fact that, in general, products of matrices do not commute. In fact, it will become clear that when the matrices A, B do commute the above problem yields the same answer as it does for scalars, i.e. $\exp(tA)\exp(tB) = \exp(t(A+B))$.

One can use the results given above to obtain an ODE which $C(t)$ solves, and derive the power series of $C(t)$,

$$C(t) = t(A+B) + \frac{t^2}{2}[A,B] + \frac{t^3}{12}\left(\Big[A,[A,B]\Big] + \Big[B,[B,A]\Big]\right) + ... \tag{2.4}$$

This is known as the Baker-Campbell-Hausdorff (BCH) formula. We prove this, following the proof from [1].

We consider, for $s, t \in \mathbb{R}$ sufficiently small, a smooth matrix function $Z(s,t)$ such that

$$\exp(sA)\exp(tB) = \exp(Z(s,t)).$$

Using Lemma 2.2.2, we may differentiate this with respect to s to obtain

$$A\exp(sA)\exp(tB) = \mathrm{dexp}_{Z(s,t)}\left(\frac{\partial Z}{\partial s}\right)\exp(Z(s,t)),$$

from which we use Lemma 2.2.3, which yields

$$\frac{\partial Z}{\partial s} = A - \frac{1}{2}[Z,A] + \sum_{k=2}^{\infty}\frac{B_k}{k!}ad_Z^k(A).$$

We note, from properties of the matrix exponential, that we also have $\exp(-tB)\exp(-sA) = \exp(-Z(s,t))$. *We repeat the same process as above, but now differentiating with respect to t, to obtain*

$$\frac{\partial Z}{\partial t} = B + \frac{1}{2}[Z,B] + \sum_{k=2}^{\infty}\frac{B_k}{k!}ad_Z^k(B),$$

as $ad_{-Z}^k(B) = (-1)^k ad_Z^k(B)$, and $B_k = 0$ for odd $k > 2$. Now noting that $C(t) = Z(t,t)$, it is clear that

$$C'(t) = \frac{\partial Z}{\partial s}(t,t) + \frac{\partial Z}{\partial t}(t,t).$$

We then expand the function $C(t)$ in terms of Taylor coefficients, $C(t) = tC_1 + t^2C_2 + +t^3C_3 + ...$, and using the above ODE, we can compare powers of t to obtain the Taylor coefficients and hence the Baker-Campbell-Hausdorff formula. ■

This does not provide a wholly satisfying answer, as there are simple examples for which no such matrix $C(t)$ exists, as shown in example 3.4.1 of [14]. Moreover, this series is known to not converge for all matrices (see [13]). For our applications, we will note that the series converges if $\|tA\| < 1$, and $\|tB\| < 1$.

Analogously, one may be interested in finding a matrix valued function $S(t)$, such that $\exp(\frac{t}{2}A)\exp(tB)\exp(\frac{t}{2}A) = \exp(S(t))$. This problem has a similar solution, which is a power series of the form

$$S(t) = t(A+B) + \frac{t^3}{24}\left(-[A,\Big[A,B\Big]] + 2\Big[B,[B,A]\Big]\right) + ... \tag{2.5}$$

This is known as the symmetric Baker-Campbell-Hausdorff (sBCH) formula. This can be derived by applying (2.4) firstly to $\exp(\frac{t}{2}A)\exp(\frac{t}{2}B) = \exp(C(t))$, and then again to $\exp(C(t))\exp(-C(-t)) = \exp(S(t))$. These formulae will be used in the next section to construct our modified differential equations.

2.3 Backward Error Analysis

This section follows closely from [1]. In particular, we direct the reader to chapters III and IX for details, such as proofs, which are omitted from this report.

It should be clear from the previous section that we can, given sufficient conditions for convergence, apply BCH-like formulas to construct a modified differential equation which is solved by our splitting method. For example, we can consider our problem (2.3) with the Lie-Trotter splitting, $\Phi_h(y_0) = \exp(hA)\exp(hB)y_0$, where h is some sufficiently small timestep. We can then apply (2.4) to see $\exp(hA)\exp(hB) = \exp(h\tilde{C})y_0$, for

$$\tilde{C} = A + B + \frac{h}{2}[A,B] + \frac{h^2}{12}\left(\Big[A,[A,B]\Big] + \Big[B,[B,A]\Big]\right) + ... = \frac{C(h)}{h}.$$

That is to say, Φ_h is the flow of $\dot{y} = \tilde{C}y$ at time h. Similarly, one can consider the Strang splitting $\Phi_h^{[S]}(y_0) = \exp(\frac{h}{2}A)\exp(hB)\exp(\frac{h}{2}A)y_0$, and use (2.5) to construct the modified differential equation which this composition solves exactly. It is worth noting that one does not just apply (2.4) for some arbitrary time, but for a fixed timestep, so that we have equality at $t = h$.

In the following analysis will consider the more general form of this problem, (2.1), presented at the beginning of this chapter. To begin this, we introduce the Lie derivative.

Definition 2.3.1 (Lie Derivative) *For our functions, $f^{[i]} : \mathbb{R}^n \to \mathbb{R}^n$, $i = 1, 2$, we introduce corresponding Lie derivative*

$$D_i = \sum_{j=1}^{n} f_j^{[i]}(y)\frac{\partial}{\partial y_j},$$

where $f_j^{[i]}(y)$ is the 'j'th component of $f^{[i]}(y)$.

These are relevant, as it can be shown, assuming our $f^{[i]}$ are analytic, that the respective flows of the separated equations (2.2) are

$$\varphi_t^{[i]}(y_0) = \exp(tD_i)\mathrm{id}(y_0), \quad i = 1, 2, \tag{2.6}$$

where we have considered the operator $\exp(tD_i)$ as a formal series of differential operators acting on the identity function, id, at the point y_0. This result is visually very similar to the case for a matrix, but now D_i is an unbounded differential operator. We were able to avoid these differential operators in the matrix case as the above equation holds for matrices in a more simple, and better behaved, sense. This then relates to our splitting method, as it can be shown that

$$\left(\varphi_t^{[2]} \circ \varphi_s^{[1]}\right)(y_0) = \exp(sD_1)\exp(tD_2)\mathrm{id}(y_0).$$

From this, we would like to apply (2.4) in the same way that we did at the beginning of the chapter. However, as these are unbounded operators, it is unexpected that one would recover a convergent series. However, we can consider this as a formal series, and ignore the issues of convergence for $\exp(sD_1)\exp(tD_2) = \exp(D(s,t))$, with

$$D(s,t) = sD_1 + tD_2 + \frac{st}{2}[D_1, D_2] + \frac{st}{12}\left(t\Big[D_1, [D_1, D_2]\Big] + s\Big[D_2, [D_2, D_1]\Big]\right) + ...$$

From this formal series, we can continue in the same way we did at the beginning of the chapter to see $\Phi_h = \exp(hD_2)\exp(hD_1)\mathrm{id} = \exp(h\tilde{D})\mathrm{id}$, where

$$\tilde{D} = D_1 + D_2 + \frac{h}{2}[D_2, D_1] + \frac{h^2}{12}\left(\Big[D_2, [D_2, D_1]\Big] + \Big[D_1, [D_1, D_2]\Big]\right) + ... = \frac{D(h,h)}{h}.$$

Similarly to (2.6), it follows that Φ_h is formally the flow, at time $t = h$, of the modified equation

$$\dot{y} = \tilde{D}\,\mathrm{id}(y) =: \tilde{f}(y)$$

. One can perform these calculations to obtain an equivalent formulation with

$$\tilde{f}(y) = f(y) + hf_2(y) + h^2 f_3(y) + ...\,,$$

where the first two terms are

$$f = f^{[1]} + f^{[2]}, \quad f_2 = \frac{1}{2}\left(f^{[1]'}f^{[2]} - f^{[2]'}f^{[1]}\right).$$

This is how the modified equation is constructed for the Lie-Trotter splitting, and one can apply the same method using (2.5) to obtain a similar modified equation for the Strang splitting. We will later make use of this by truncating this formal series, to obtain a practical method which one can then produce error bounds for. Before this however, we consider this construction with a Hamiltonian system.

If the functions are $f^{[i]}(y) = J^{-1}\nabla H^{[i]}(y)$ for some functions $H^{[i]}$, then we say they are Hamiltonian vector fields. In this case we are assuming $n = 2d$ for some $d \in \mathbb{N}$, so we may write $y = (q_1, ..., q_d, p_1, ...p_d)$, and recall that we have $J = \begin{pmatrix} 0 & I_d \\ -I_d & 0 \end{pmatrix}$. In this case our Hamiltonian is then $H(y) = H^{[1]}(y) + H^{[2]}(y)$. It can be shown that for two Hamiltonian vector fields, F and G, that their associated Lie derivative operators have the property that $[D_F, D_G] = D_{\{G,F\}}$, where the curly brackets denote the Poisson bracket,

$$\{G, F\} = \sum_{k=1}^{d} \left(\frac{\partial G}{\partial q_i} \frac{\partial F}{\partial p_i} - \frac{\partial G}{\partial p_i} \frac{\partial G}{\partial q_i} \right).$$

This then lets us see that our modified equation may be written in terms of the Hamiltonians, $H^{[1]}$ and $H^{[2]}$, with $f_j(y) = J^{-1}\nabla H_j(y)$, where the first non-trivial two terms become

$$H_2 = \frac{1}{2}\{H^{[1]}, H^{[2]}\}, \quad H_3 = \frac{1}{12}\left(\{\{H^{[1]}, H^{[2]}\}, H^{[2]}\} + \{\{H^{[2]}, H^{[1]}\}, H^{[1]}\}\right).$$

This is in fact the construction of the modified Hamiltonian for our system,

$$\tilde{H}(y) = H(y) + hH_2(y) + h^2H_3(y) + h^3H_4(y) + ...,$$

which is again considered as a formal series. One can perform a construction like this for symplectic methods, which we now define.

Definition 2.3.2 (Symplectic Method) *A numerical method is said to be symplectic if the one step map, $y_1 = \Phi_h(y_0)$ is symplectic in the sense of definition 1.3.4.*

Exponential splitting methods are not always symplectic methods. This can easily be verified by considering the Lie-Trotter splitting applied to (2.3), for diagonal matrices A, B. Regardless, the analysis, and properties, of symplectic methods are still very similar to those of splitting methods. For example, symplectic methods can also be shown to exactly solve a modified equation,

$$\dot{y} = f(y) + hf_2(y) + h^2f_3(y) + h^3f_4(y) + ... \tag{2.7}$$

The following result, from [1], expands upon this.

Theorem 2.3.3 (Preservation of Hamiltonian Structure) *If a symplectic method $\Phi_h(y)$ is applied to a Hamiltonian system with smooth Hamiltonian $H : \mathbb{R}^{2d} \to \mathbb{R}$, then the associated modified equation is also Hamiltonian. More precisely, there exist smooth functions $H_j : \mathbb{R}^{2d} \to \mathbb{R}$ such that $f_j(y) = J^{-1}\nabla H_j(y)$. Moreover, if H is defined on some open set $D \subseteq \mathbb{R}^{2d}$ the functions H_j can be shown to be defined on the same set.*

We note that for our example of interest, (2.3), that we can write $Ay = J^{-1}\nabla H^{[1]}(y)$, where we define $H^{[1]}(y) = y^T JAy$. We can do the same for the map $y \mapsto By$ to find our corresponding Hamiltonian.

For the rest of the section, we will be considering symplectic methods with their modified equations of the form (2.7). Although the exponential splittings aren't symplectic, similar analysis can be done in the finite case, and can be found in [2]. We have now gone through the first part of the backward error analysis by constructing our modified equations. This then allows one to begin analysis on the methods in question. To begin to do so, one must begin by addressing the issues of convergence. In [1] a simple example, considering the trapezoidal rule, is given, which shows that the modified differential equation (2.7) will not converge for any timestep $h \neq 0$. Thus it is apparent that we must consider truncations of any formal series.

For the following estimates, we assume that that the function $f(y)$ is analytic on a complex neighbourhood of y_0, the initial data. Moreover, we assume that we have constants $M, R > 0$ such that

$$\|y - y_0\| \leq 2R \Rightarrow \|f(y)\| \leq M.$$

An immediate question that arises when one considers truncations of infinite series is where should one truncate from? We refer the reader to [1] for how the answer to this is motivated, and indeed for all proofs and further details. We consider the truncated modified equation,

$$\dot{y} = F_N(y) = f(y) + hf_2(y) + ... + h^{N-1}f_N(y), \tag{2.8}$$

where $hN \leq \frac{R}{e\eta M} =: h_0$, for η some constant which depends only on the method. Similarly, we assume that we may expand our numerical method in the form

$$\Phi_h(y) = y + hf(y) + h^2 d_2(y) + h^3 d_3(y) + ...,$$

where this is analytic on a neighbourhood of $h = 0$ and $y \in B_R(y_0)$ This analyticity of both the method and the equation allows use of Cauchy estimates. One can then show under these conditions that

$$\|F_N(y) - f(y)\| \leq cMh^p,$$

where c is a constant depending on the method. Moreover, one can also obtain the following error result.

Theorem 2.3.4 *Let $f(y)$ be analytic in $B_{2R}(y_0)$, and the coefficients of the method, $d_j(y)$, be analytic in $B_R(y_0)$. Moreover, we assume the above boundedness condition on f and that*

$$\|d_j(y)\| \leq \mu M \left(\frac{2\kappa M}{R}\right)^{j-1},$$

for $\|y-y_0\| \leq R$, where μ is some constant depending on the method. If $h \leq h_0/4$, then there is some integer $N = N(h)$ (namely the largest integer N such that $Nh \leq h_0$) then if $y_1 = \Phi_h(y_0)$, and $\tilde{\varphi}_{N,t}(y_0)$ is the exact solution of (2.8), then we have

$$\|y_1 - \tilde{\varphi}_{N,h}(y_0)\| \leq h\gamma M e^{-h_0/h},$$

where γ is some constant depending on the method.

This justifies our truncation of the formal series, as we can make our truncation arbitrarily accurate. Thus, we may proceed, without any issues of convergence, using the truncated series (2.8).

Finally, this can be used to obtain a result about the long term energy conservation. We consider our method to be of order p, which means the truncated modified Hamiltonian will be of the form

$$\tilde{H}(y) = H(y) + h^p H_{p+1}(y) + ... + h^{N-1} H_N(y). \tag{2.9}$$

Theorem 2.3.5 (Approximate Energy Conservation) *Consider a Hamiltonian system with analytic $H : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^{2d}$ is open. We apply a symplectic method, $\Phi_h(y)$, for some timestep $h > 0$. If the numerical solution stays in the compact set $K \subset D$, then there exists h_0 and $N = N(h)$, as in the previous result, such that:*

$$\tilde{H}(y_n) = \tilde{H}(y_0) + \mathcal{O}(e^{-h_0/2h}),$$

$$H(y_n) = H(y_0) + \mathcal{O}(h^p),$$

over exponentially long time intervals, $nh \leq e^{h_0/2h}$.

Thus, as we have a conservation of energy, we see that a symplectic method of order p will conserve the energy of the system with a $\mathcal{O}(h^p)$ error for an exponentially long time interval. Although this is a result for symplectic methods, this is consistent with our observations of near energy conservation for exponential splittings. Moreover, this shows that as the timestep size decreases, the interval for which we have near energy conservation increases. This is an important feature as one may be interested in integration over a very long time period. An example of this is [18], where the authors are interested in long term behaviour of orbits.

The explicit analysis for splitting methods in this finite case is presented in [2], and yields a result very similar to the above. These tools will not be applicable in the analysis of splitting methods for (1.1). The analysis in this case is considered in the following chapter.

Chapter 3

Infinite Dimensional Case

Section 3.1 provides some motivation as to why the analysis of splitting methods may be difficult. In particular, we investigate the exponential of a linear operator, which is crucial in defining the solution (1.2) for (1.1). Sections 3.2 and 3.3 follow very closely from Brian C. Hall's book, [7], introducing some concepts of functional analysis, which may not be seen as an undergraduate, that underlie a lot of the later error analysis. As such, all omitted proofs can be found in this book, in chapters 7-10, unless stated otherwise. Section 3.4 discusses why one cannot easily extend the error analysis from the previous chapter, and looks at how Erwan Faou deals with this in [2], which considers the Lie-Trotter splitting (1.3). Finally, section 3.5 considers extending Faou's technique to the Strang splitting (1.4).

We assume some prior knowledge of functional analysis and measure theory in this chapter. For background in functional analysis we refer the reader to [9], and to [8] for background in measure theory. Throughout we will consider $\mathcal{H}$ to be a separable Hilbert space, a concrete example of interest is $\mathcal{H} = L^2(\mathbb{R}^d; \mathbb{C})$. As a convention we take our inner product, $\langle \cdot, \cdot \rangle$, to be linear in the second slot.

3.1 Linear Operators and the Exponential Map

We have expressed the solution of (1.1) as (1.2), an exponential of an (unbounded) linear operator. We will make some sense of this later on, but for now we consider why this could be problematic.

Consider the matrix exponential $\exp : M_{n,n}(\mathbb{C}) \to M_{n,n}(\mathbb{C})$ which is defined by the following,

$$\exp(tA) = \sum_{k=0}^{\infty} \frac{t^k A^k}{k!},$$

and is defined for all time. To see this formally, one can fix $t \in \mathbb{R}$ and consider the sequence of partial sums $S_n = \sum_{k=0}^{n} \frac{t^k A^k}{k!}$. Then, noting that for $n, m \in \mathbb{N}$ such that $n > m$,

$$\left\| \sum_{k=m+1}^{n} \frac{t^k A^k}{k!} \right\| \leq \sum_{k=m+1}^{n} \frac{(|t| \cdot \|A\|)^k}{k!} \leq \sum_{k=m+1}^{\infty} \frac{(|t| \cdot \|A\|)^k}{k!},$$

which follows from the triangle inequality and some basic inequalities involving operator norms. It is then clear that as $\|A\| < \infty$ that the right hand of the inequality is the tail of a convergent series, and thus can be made arbitrarily small. Hence the sequence $(S_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $M_{n,n}(\mathbb{C})$, a Banach space, and hence has a unique limit, $\exp(tA)$.

This map is useful when we consider linear ODEs. In particular, it is relatively easy to show that the initial value problem,

$$\frac{dx}{dt} = Ax, \qquad x(0) = x_0,$$

is solved (uniquely) by the function $x(t) = \exp(tA)x_0$.

Now, consider how we might similarly define the exponential map of the linear operator $\frac{d}{dx}$. Naturally, we might want to write this operator as

$$\mathscr{L}_t = \sum_{k=0}^{\infty} \frac{t^k}{k!} \frac{d^k}{dx^k},$$

for some $t \in \mathbb{R}$. Notice that we have the domain consisting of smooth functions, instead of merely differentiable functions, as we might have hoped. Even with this naive definition, $\mathscr{L}_t$ isn't as nicely behaved as the matrix exponential, which we illustrate with the following example.

Example 3.1.1 (Naive Exponential) *Fix a value of $t \in \mathbb{R}$. Consider the operator $\mathscr{L}_t : C^\infty((-1,1)) \to C^\infty((-1,1))$, as defined above, and the function $u \in C^\infty((-1,1))$ given by $u(x) = \frac{1}{1-x}$. We then see that*

$$\mathscr{L}_t u(x) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \frac{d^k}{dx^k} \left(\sum_{n=0}^{\infty} x^n \right) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \left(\sum_{n=k}^{\infty} \frac{n!}{(n-k)!} x^{n-k} \right).$$

It's clear that for $n \geq k$ that $\frac{n!}{(n-k)!} \geq k!$, so if we reindex the second sum with $m = n - k$ we obtain,

$$\mathscr{L}_t u(x) \geq \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} t^k x^m,$$

which is clearly not defined for all $t \in \mathbb{R}$.

Indeed for analytic[1] functions $u \in C^\infty(\mathbb{R})$ we find that this operator gives us $\mathscr{L}u(x,t) = u(x+t)$. The idea is that we can obtain a Taylor series about x, i.e.

$$\mathscr{L}_t u(x) = \sum_{k=0}^{\infty} \frac{t^k}{k!} u^{(k)}(x) = \sum_{k=0}^{\infty} \frac{((x+t)-x)^k}{k!} u^{(k)}(x) = u(x+t)$$

This emphasises that the issue in the above example is that the function isn't analytic on $\mathbb{R}$. This should inspire some idea that the way we define the domain for these operators is significant, and indeed we will expand upon this idea later.

We now consider how one might define the exponential of $\frac{d}{dx}$, and specifically what property we would like this to have. To begin this, we again consider the matrix exponential as defined at the beginning of this section. We first defined the matrix exponential as a power series, which happened to solve a specific ODE. However, we could have equivalently started with the ODE and defined the matrix exponential as the solution. This is how we will define the exponential of $\frac{d}{dx}$.

Example 3.1.2 (Exponential of $\frac{d}{dx}$) *Consider the differential equation that should be solved by* $\exp\left(t\frac{d}{dx}\right) u(x)$*. As motivated by the above, a sensible choice would be the following:*

$$\frac{\partial U}{\partial t} = \frac{\partial U}{\partial x}, \qquad U(x,0) = u(x),$$

where we think of the function $U(x,t)$ as $exp(t\frac{d}{dx})u(x)$. To then solve this Cauchy problem, we use the Fourier transform in x. We apply standard properties of the Fourier transform to see that this problem becomes

$$\frac{\partial \hat{U}}{\partial t} = \frac{-ik}{\sqrt{2\pi}} \hat{U}, \qquad \hat{U}(k,0) = \hat{u}(k), \tag{3.1}$$

where hats denote Fourier transforms. The solution of (3.1) is an exponential function in t, multiplied by $\hat{u}(k)$, and then taking inverse Fourier transform gives us our function U via

$$U(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ikx - \frac{ikt}{\sqrt{2\pi}}} \hat{u}(k)\, dk =: exp\left(t\frac{d}{dx}\right) u(x)$$

This gives a more meaningful way to define the exponential of $\frac{d}{dx}$, and we will revisit this later on when addressing the exponential e^{-itH} which defined our solution (1.2). It is worth noting that we did not specify a domain when defining this operator. This will be addressed later, as we must choose a sufficiently nice domain for this to work. The takeaway from this section should be that our finite dimensional intuition fails us in the infinite dimensional case, and that we must resort to more sophisticated machinery.

[1]It is worth noting that, unlike smooth complex functions, not all smooth real functions are analytic, but all analytic functions are smooth.

3.2 Unbounded Operators and Self Adjointness

In the previous section, we showed that the matrix exponential is well defined for all time. This relied on the fact that any element of $M_{n,n}(\mathbb{C})$ is bounded. In more general Hilbert spaces, this isn't true, as we will demonstrate later. Firstly, recall that a linear operator $A : \mathcal{H} \to \mathcal{H}$ is bounded if there exists some constant, $C > 0$, such that $\|A\phi\|_{\mathcal{H}} \leq C\|\phi\|_{\mathcal{H}}$. We will denote the space of such operators as $\mathcal{B}(\mathcal{H})$. We will now be considering unbounded (linear) operators, which means that this property may not hold, but it includes the possibility that it does.

Given some linear operator A on a Hilbert space $\mathcal{H}$, a natural question one might ask is "Is there a linear operator B such that $\langle A\phi, \psi\rangle = \langle \phi, B\psi\rangle$, for all $\phi, \psi \in \mathcal{H}$?". For bounded A, the answer is fairly obvious – yes, as shown by the Riesz representation theorem. Fix $\phi \in \mathcal{H}$, and consider the linear functional $L(\psi) = \langle \phi, A\psi\rangle$. Then for bounded A we may apply the Riesz representation theorem to see that $\langle \phi, A\psi\rangle = \langle \xi, \psi\rangle$ for some $\xi \in \mathcal{H}$ which depends on ϕ. We use this to define a map, $A^* : \mathcal{H} \to \mathcal{H}$ by $A^*\phi = \xi$. It is then easy to show that A^* is a bounded linear operator, which we call the adjoint of A.

For an unbounded operator, this isn't immediately obvious as we cannot apply Riesz's representation theorem. Moreover, we will show later that if a linear operator, A, is not bounded then it cannot be defined on all of $\mathcal{H}$. Hence we have to be careful with the domain on which we define A^*, and indeed A itself. Because of this issue with domains, although the obvious notion of a self adjoint operator, $\langle \phi, A\psi\rangle = \langle A\phi, \psi\rangle$, works for bounded A, it won't be enough if A is unbounded.

Definition 3.2.1 (Adjoint for Unbounded Operator) *Let A be an unbounded operator with domain $D_A \subseteq \mathcal{H}$. We set D_{A^*} to be the space of $\phi \in \mathcal{H}$ such that the linear function $L(\psi) = \langle \phi, A\psi\rangle$ is bounded. We let $A^*\phi$ be the unique vector such that $\langle A^*\phi, \psi\rangle = \langle \phi, A\psi\rangle$.*

We define D_{A^*} in this way so that we have a bounded extension to $\mathcal{H}$, by the bounded linear transformation theorem, and hence we may apply Riesz's representation theorem in the same way as we did for bounded operators. We now discuss why the intuitive definition of self adjointness isn't sufficient for unbounded operators.

Definition 3.2.2 (Symmetric Operators and Extensions) *For $A : D_A \to \mathcal{H}$ an unbounded operator, we say A is symmetric if $\langle \phi, A\psi\rangle = \langle A\phi, \psi\rangle$ for all $\phi, \psi \in D_A$.*

We say an unbounded operator $B : D_B \to \mathcal{H}$ is an extension of A if $D_A \subseteq D_B$ and $A = B$ on D_A.

While for a bounded operator symmetry is equivalent to self adjointness, this isn't the case for unbounded operators, and in fact, generally, the adjoint of a symmetric operator, A, is an extension of A. We prove this as follows:

Let $A : D_A \to \mathcal{H}$ be a symmetric operator, and $\phi, \psi \in D_A$. By the Cauchy Schwarz inequality, we see

$$|\langle \phi, A\psi \rangle| \leq \|\phi\| \, \|A\psi\|,$$

and thus the linear functional $L(\psi) = \langle \phi, A\psi \rangle$ is bounded for all $\phi \in D_A$. Hence $D_A \subseteq D_{A^*}$. Moreover, as A is symmetric we see that we have

$$\langle A^*\phi, \psi \rangle = \langle \phi, A\psi \rangle = \langle A\phi, \psi \rangle,$$

that is, $A^* = A$ on D_A, and hence A^* is an extension of A. ■

It can be shown that these are equivalent statements. That is, an operator is symmetric if, and only if, its adjoint as an extension of it. From this we see that the obvious way to define self adjointness as follows:

Definition 3.2.3 (Self Adjoint Operator) *The operator $A : D_A \to \mathcal{H}$ is said to be self adjoint if it is symmetric, and $D_A = D_{A^*}$*

Now we come back to why the domain of definition is problematic. Firstly, we consider an example.

Example 3.2.4 (An unbounded operator on $L^2(\mathbb{R})$) *Consider the linear operator $\frac{d^2}{dx^2}$ on $L^2(\mathbb{R})$. It is should be immediately clear that this operator cannot be defined on all of $L^2(\mathbb{R})$ in the classical sense, but it is defined on a dense subspace. We note that this is a symmetric operator from the following:*

$$\int_{-\infty}^{\infty} \phi \frac{d^2\psi}{dx^2} dx = -\int_{-\infty}^{\infty} \frac{d\phi}{dx}\frac{d\psi}{dx} dx = \int_{-\infty}^{\infty} \frac{d^2\phi}{dx^2} \psi dx$$

which comes from repeated use of integration by parts, and noting that $\phi, \psi, \frac{d\phi}{dx}, \frac{d\psi}{dx} \in L^2(\mathbb{R})$ implies they vanish at infinity.

Moreover, this is an unbounded operator, as we may consider, for $n \in \mathbb{N}$ the functions defined by: $f_n(x) = e^{-\frac{nx^2}{2}}$. It is clear that these are smooth functions that belong to $L^2(\mathbb{R})$. We then calculate the appropriate norms of these functions as

$$\|f_n\|_{L^2}^2 = \int_{-\infty}^{\infty} e^{-nx^2} dx = \sqrt{\frac{\pi}{n}},$$

$$\left\| \frac{d^2 f_n}{dx^2} \right\|_{L^2}^2 = \int_{-\infty}^{\infty} (n^2 - 2n^3x^2 + n^4x^4) e^{-nx^2} dx = \frac{3}{4} n^{\frac{3}{2}} \sqrt{\pi}.$$

Then using the definition of the operator norm we see that $\sqrt{\frac{3n}{4}} \leq \left\| \frac{d^2}{dx^2} \right\|$ for any $n \in \mathbb{N}$, and thus we have an unbounded operator.

In fact, we can show that for general unbounded, symmetric operators that they cannot be defined on all of $\mathcal{H}$. We first recall the following definition.

Definition 3.2.5 (Closed Operators) *The operator $A : D_A \to \mathcal{H}$ is said to be closed if its graph is a closed set in $\mathcal{H} \times \mathcal{H}$.*

It can then be shown that if $A : D_A \to \mathcal{H}$ is an unbounded linear operator, then the graph of A^* is closed. We see this as follows:

Let $(\psi_n, A^*\psi_n)_{n\in\mathbb{N}}$ be a sequence in $Graph(A^*) := \{(x, A^*x) \mid x \in D_{A^*}\}$ converging to some limit $(\psi, \phi) \in \mathcal{H} \times \mathcal{H}$. This then implies $\psi_n \to \psi$, and $A^*\psi_n \to \phi$ in $\mathcal{H}$ as $n \to \infty$. Then for an arbitrary $\gamma \in D_A$ we see

$$\langle \psi, A\gamma \rangle = \lim_{n\to\infty} \langle \psi_n, A\gamma \rangle = \lim_{n\to\infty} \langle A^*\psi_n, \gamma \rangle = \langle \phi, \gamma \rangle$$

which implies that $\psi \in D_{A^*}$, and $A^*\psi = \phi$. Hence, we see that $Graph(A^*)$ is closed in $\mathcal{H} \times \mathcal{H}$, and so A^* is a closed operator. ■

This now lets us show that an unbounded, symmetric operator defined on $\mathcal{H}$ is in fact bounded. Assume that A is an unbounded, symmetric operator with $D_A = \mathcal{H}$. Then we have shown earlier that A^* is an extension of A, but as A has maximal domain we must have that $A^* = A$. Moreover, we have just shown that $A = A^*$ is closed. Thus we may apply the closed graph theorem to A to see that A is in fact a bounded operator[2].

Hence for a (symmetric) operator, A, which is not bounded, for example $\frac{d^2}{dx^2}$, we cannot have $D_A = \mathcal{H}$. One can apply a similar argument with skew symmetric operators, that is A such that $\langle \phi, A\psi \rangle = -\langle A\phi, \psi \rangle$, to see these similarly must be bounded if they are defined on $\mathcal{H}$.

Another complication arising from infinite dimensional operators is to do with eigenvalues and eigenvectors. In particular, it isn't guaranteed that an eigenvector exists, as we will demonstrate. Firstly, we generalise our notion of an eigenvalue.

Definition 3.2.6 (Spectrum of an Operator) *Let $A : D_A \to \mathcal{H}$ be a linear operator. We define the resolvent set of A, $\rho(A)$, as the set of $\lambda \in \mathbb{C}$ such that the operator $A - \lambda \mathrm{id}_{\mathcal{H}}$ has a bounded inverse. We then define the spectrum of A, $\sigma(A)$, to be the complement of this set in $\mathbb{C}$.*

We note that in the case of matrices the above reduces to the usual definition of whether the kernel of $A - \lambda I$ is trivial. Now we consider an example of an operator for which the spectrum is non-empty, but no eigenvectors exist.

Example 3.2.7 *Let $\mathcal{H} = L^2(\mathbb{R}; \mathbb{C})$. Let $X : D_X \to \mathcal{H}$ be the operator defined by $X\phi(x) = x\phi(x)$. We claim that the spectrum of this operator is $\sigma(X) = [0, 1]$. This is proven in more detail in [7]. However, what we would like to note is that X is*

[2]This formalises the comment made earlier, that symmetry is equivalent to self adjointness for bounded operators.

a bounded operator, and it can be shown that for bounded operators that $\sigma(X)$ is a closed, bounded, non-empty subset of $\mathbb{C}$.

Now let us assume that for a given $\lambda \in \sigma(X)$ that we have some $\psi \in \mathcal{H}$ such that $X\psi = \lambda\psi$. For such a ψ, we note that $(x-\lambda)\psi(x) = 0$, and thus $\psi(x) = 0$ for $x \in [0,1]\backslash\{\lambda\}$. Thus we see that such a ψ is zero almost everywhere, and hence we have $\psi = 0$ from the definition of $L^2(\mathbb{R};\mathbb{C})$.

As noted above, the spectrum of a bounded operator is closed, bounded and non-empty. Our objects of interest are unbounded self adjoint operators, which have a similar property that their spectrum is closed, and contained in the real numbers. In particular, this closedness is important as we will be taking integrals over the spectrum of such operators, and closedness implies that these sets are Borel.

3.3 Spectral Theorem for Unbounded Self Adjoint Operators

We begin this section by recalling the notion of a projection on a Hilbert space. A projection, on $\mathcal{H}$, is a linear map $P : \mathcal{H} \to \mathcal{H}$ such that $P^2 = P$. Moreover, we say this is an orthogonal projection if for all $\phi, \psi \in \mathcal{H}$ we have that $\langle \phi, P\psi\rangle = \langle P\phi, \psi\rangle$, i.e. P is symmetric. From this we can deduce that P is a bounded operator as,

$$\|P\phi\|^2 = \langle P\phi, P\phi\rangle = \langle \phi, P^2\phi\rangle = \langle \phi, P\phi\rangle \leq \|P\phi\|\|\phi\|,$$

where we have applied the Cauchy-Schwarz inequality in the final step.

An example of a projection in finite dimensions is the linear map represented by the matrix:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} : \mathbb{R}^3 \to \mathbb{R}^3,$$

which has the geometric interpretation of sending a point to its component in the x-direction. Orthogonal projections will be crucial for use in the spectral theorem, through the following definition.

Definition 3.3.1 (Projection Valued Measure) *Let (X, Ω) be a measurable space. A projection valued measure is a map $\mu : \Omega \to \mathcal{B}(\mathcal{H})$ such that the following hold:*

1. *For $E \in \Omega$ we have $\mu(E)$ is an orthogonal projection.*
2. *$\mu(\emptyset) = 0$ and $\mu(X) = \mathrm{id}_{\mathcal{H}}$*
3. *For pairwise disjoint $E_1, E_2, E_3, ... \in \Omega$ we have,*

$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=0}^{\infty} \mu(E_n)$$

4. *For any* $E_1, E_2 \in \Omega$ *we have* $\mu(E_1 \cap E_2) = \mu(E_1)\mu(E_2)$.

It is worth noting that while point (3) is identical to the case for a real valued measure, we are talking about a sum of linear operators. In fact, given $\phi \in \mathcal{H}$ we can recover a real valued measure by $\mu_\phi(E) = \langle \phi, \mu(E)\phi \rangle$.

Now we have this notion of a projection valued measure, we are interested with how one may define integration with it. This is the content of the following result.

Lemma 3.3.2 (Operator Valued Integration) *Let* (X, Ω) *be a measurable space, with projection valued measure* $\mu : \Omega \to \mathcal{B}(\mathcal{H})$*, and* f *be a measurable, complex valued function on* X*. Then there is a unique linear map, written* $f \mapsto \int_X f \, d\mu$*, where* $\int_X f \, d\mu$ *is a linear operator defined on*

$$W_f = \left\{ \psi \in \mathcal{H} \,\middle|\, \int_X |f|^2 \, d\mu_\psi < \infty \right\},$$

such that

$$\left\langle \psi, \left(\int_X f \, d\mu \right) \psi \right\rangle = \int_X f \, d\mu_\psi,$$

where we have defined μ_ψ *above. Moreover, this has the following properties:*

1. *For* $E \in \Omega$ *we have*
$$\int_X 1_E \, d\mu = \mu(E).$$
2. *For all* f *in the domain, we have*
$$\left\| \int_X f \, d\mu \right\| \leq \sup_{x \in X} |f(x)|.$$
3. *For all* f, g *in the domain, we have*
$$\int_X fg \, d\mu = \left(\int_X f \, d\mu \right) \left(\int_X g \, d\mu \right).$$
4. *For all* f *in the domain, we have*
$$\int_X \bar{f} \, d\mu = \left(\int_X f \, d\mu \right)^*,$$
where bars denote the complex conjugate.

There are a few immediate results from this, for example:

1. $\int_X 1 \, d\mu = \mathrm{id}_{\mathcal{H}}$
2. If f is a bounded function, then $\int_X f \, d\mu$ is a bounded operator.

3. If we have operators of the form $A = \int_X f\,d\mu, B = \int_X g\,d\mu$, then A and B commute.

4. If f is real valued, then $\int_X f\,d\mu$ is a self adjoint operator.

We must specify the domain W_f, as generally (since f is unbounded) this operator is unbounded. The domain is motivated from the fact that when the function f is bounded we have:

$$\left\langle \left(\int_X f\,d\mu\right)\psi, \left(\int_X f\,d\mu\right)\psi \right\rangle = \left\langle \psi, \left(\int_X \bar{f}f\,d\mu\right)\psi \right\rangle = \int_X |f|^2\,d\mu_\psi.$$

This machinery gives us the basis for a form of the spectral theorem.

Theorem 3.3.3 (Spectral Theorem for Self Adjoint Operators) *Let $A : D_A \to \mathcal{H}$ be a self adjoint linear operator. Then there exists a unique projection valued measure $\mu^A : \sigma(A) \to \mathcal{B}(\mathcal{H})$ such that*

$$\int_{\sigma(A)} \mathrm{id}_{\sigma(A)}\,d\mu^A = A.$$

As with any generalisation, we would like to be able to recover the simpler versions of it. We do this now for finite dimensions, recovering the spectral theorem for Hermitian matrices as follows:

Let $A : M_{n,n}(\mathbb{C}) \to M_{n,n}(\mathbb{C})$ be a Hermitian matrix. This then says A is symmetric with respect to our inner product, which is equivalent to being self adjoint, since A is bounded as we are in finite dimensions. The spectrum of A is the set of $\lambda \in \mathbb{C}$ such that $(A - \lambda I_n)$ isn't invertible. We can calculate this, as one normally does with matrices, as the roots of the polynomial $p(\lambda) = det(A - \lambda I_n)$. A quick application of the fundamental theorem of algebra lets us see that this is a non-empty set, so we write $\sigma(A) = \{\lambda_1, ..., \lambda_m\}$ for some $m \leq n$.

The spectral theorem, as stated above, then implies that there exists a projection valued measure μ^A such that

$$\int_{\sigma(A)} \mathrm{id}_{\sigma(A)}\,d\mu^A = A,$$

and

$$\left\langle v, \left(\int_{\sigma(A)} \mathrm{id}_{\sigma(A)}\,d\mu^A\right)v \right\rangle = \int_{\sigma(A)} \mathrm{id}_{\sigma(A)}\,d\mu^A_v,$$

for $v \in \mathbb{C}^n$.

Now we pay closer attention to these integrals. Firstly, as $\sigma(A)$ is a finite set, we notice that $\mathrm{id}_{\sigma(A)}$ is a simple function, and our integrals will reduce into sums, for example:

$$\int_{\sigma(A)} \mathrm{id}_{\sigma(A)}\,d\mu^A_v = \sum_{k=1}^{m} \lambda_k \mu^A_v(\{\lambda_k\}).$$

It is worth nothing that the associated $\sigma-$algebra is the set of all Borel sets in $\mathbb{C}$ contained in $\sigma(A)$, which is exactly the power set, $\mathcal{P}(\sigma(A))$. Hence $\{\lambda_k\}$ is measurable for all k. We then use the above equations as follows:

$$\langle v, Av\rangle = \left\langle v, \left(\int_{\sigma(A)} \mathrm{id}_{\sigma(A)}\, d\mu_v^A\right) v \right\rangle = \sum_{k=1}^{m} \lambda_k \mu_v^A(\{\lambda_k\}) = \sum_{k=1}^{m} \lambda_k \langle v, \mu^A(\{\lambda_k\})v\rangle$$

Now using linearity and non-degeneracy of the inner product, it follows that

$$A = \sum_{k=1}^{m} \lambda_k \mu^A(\{\lambda_k\}),$$

which we will see is the familiar diagonalization of the matrix A. All that remains is to determine what $\mu^A(\{\lambda_k\})$ is. To do this, we use some further results from [7], in particular for $V_{\lambda_k} := Range(\mu^A(\{\lambda_k\}))$ we have:

1. *If $v \in V_{\lambda_k}$, then $Av \in V_{\lambda_k}$.*
2. *For distinct $1 \leq j, k \leq m$ we have $V_{\lambda_j} \perp V_{\lambda_k}$.*
3. *Moreover, these subspaces cover $\mathcal{H} = \mathbb{C}^n$, that is $\bigoplus_{k=1}^m V_{\lambda_k} = V_{\sigma(A)} = \mathbb{C}^n$.*

Hence for a given $v \in \mathbb{C}^n$ we can define $v_j = \mu^A(\{\lambda_j\})v$. From the above, this has the property that $v = \sum_{j=1}^m v_j$, and we see that

$$Av_j = \sum_{k=1}^{m} \lambda_k \mu^A(\{\lambda_k\})v_j = \lambda_j \mu^A(\{\lambda_j\})v_j = \lambda_j \mu^A(\{\lambda_j\})^2 v = \lambda_j v_j,$$

where we have used the defining property of a projection $\lambda_j \mu^A(\{\lambda_j\})^2 = \lambda_j \mu^A(\{\lambda_j\})$. The conclusion of this is that the spaces V_{λ_k} are the λ_k-eigenspaces, and the projections, $\mu^A(\{\lambda_k\})$, are the projections onto these eigenspaces. Thus, we can apply the Gram-Schmidt process to any bases of these spaces to obtain an orthonormal basis of eigenvectors of A. ■

This spectral theorem then motivates a way to define a wider class of operators in a way relevant to what was discussed in section 3.1.

Definition 3.3.4 (Functional Calculus) *For a measurable function, f, on $\sigma(A)$, we defined the operator $f(A)$ by*

$$f(A) = \int_{\sigma(A)} f\, d\mu^A,$$

where this operator is defined on domain

$$W_f = \left\{ \psi \in \mathcal{H} \,\middle|\, \int_{\sigma(A)} |f|^2\, d\mu_\psi^A < \infty \right\}.$$

One could apply the above to show that this functional calculus would return the our definition of $\exp(tA)$ for a Hermitian matrix A in a way not dissimilar to the above. This would be quite cumbersome, so we omit this. After this detour through some operator theory, we would like to return to the Schrödinger equation, and we make this link through the following:

Theorem 3.3.5 (Domains of Relevant Self Adjoint Operators) *The following operators are self adjoint on the following domains:*

1. *Multiplication by a potential function, $V(x)$, with domain:*

$$\left\{\psi \in L^2(\mathbb{R}^d;\mathbb{C}) \,|\, V\psi \in L^2(\mathbb{R}^d;\mathbb{C})\right\},$$

   where V is a measurable function $\mathbb{R}^d \to \mathbb{R}$.

2. *The Laplacian, Δ, with domain:*

$$\left\{\psi \in L^2(\mathbb{R}^d;\mathbb{C}) \,|\, |k|^2\hat{\psi}(k) \in L^2(\mathbb{R}^d;\mathbb{C})\right\},$$

   where $\hat{\psi}(k)$ is the Fourier transform of ψ. We then define

$$\Delta\psi = \mathcal{F}^{-1}(|k|^2\hat{\psi}(k)).$$

Moreover, these domains are both dense in $L^2(\mathbb{R}^d;\mathbb{C})$.

From this, we can now meaningfully define e^{-itH} for the free Schrödinger equation, and indeed this will give us the result presented at the beginning of section 1.3. However, when V isn't identically zero, we require the following result.

Theorem 3.3.6 (Kato-Rellich) *Let $A : D_A \to \mathcal{H}$, and $B : D_B \to \mathcal{H}$ be self adjoint operators. Suppose $D_A \subseteq D_B$. We define $A + B : D_A \to \mathcal{H}$ by $(A+B)\phi = A\phi + B\phi$. Then $A + B$ is self adjoint on D_A if there exist constants $0 < a < 1$ and $b > 0$ such that*

$$\|B\phi\| \leq a\|A\phi\| + b\|\phi\|,$$

for all $\phi \in D_A$.

Hence we now have suitable conditions on our potential function so that the Hamiltonian operator $H = -\Delta + V$ is self adjoint, and we may apply the functional calculus to define e^{-itH} accordingly.

As with our earlier examples, we would like this operator to have some meaningful properties, so that we may relate it back to our PDE. Specifically, we require that $\frac{d}{dt}e^{-itH}u_0 = -iHu_0$. To see this we look at strongly continuous unitary groups, and Stone's theorem.

Definition 3.3.7 (Strongly Continuous Unitary Groups) *A one parameter unitary group on $\mathcal{H}$ is a family of unitary operators $\{U(t) \,|\, t \in \mathbb{R}\}$ such that $U(0) = \mathrm{id}_{\mathcal{H}}$, and $U(s+t) = U(s)U(t)$. This family is said to be strongly continuous if*

$$\lim_{s \to t} \|U(s)\phi - U(t)\phi\| = 0,$$

for all $\phi \in \mathcal{H}$ and $t \in \mathbb{R}$.
The infinitesimal generator of this group is the operator $A : D_A \to \mathcal{H}$ defined by

$$A\phi = \lim_{t \to 0} -i \left(\frac{U(t)\phi - \phi}{t} \right),$$

with D_A the set of all $\phi \in \mathcal{H}$ such that the limit exists in the norm topology on $\mathcal{H}$.

Theorem 3.3.8 (Stone's Theorem) *Let $\{U(t) \,|\, t \in \mathbb{R}\}$ be a strongly continuous one parameter unitary group on $\mathcal{H}$. Then the associated infinitesimal generator, A as defined above, is self adjoint and D_A is dense in $\mathcal{H}$. Moreover, $U(t) = e^{-itA}$ for all $t \in \mathbb{R}$.*

We also give a useful result which relates a strongly continuous one parameter unitary group to its derivative.

Lemma 3.3.9 *Let $\{U(t) \,|\, t \in \mathbb{R}\}$ and A be as given in Stone's theorem. If $\phi \in D_A$, then for all $t \in \mathbb{R}$ we have $U(t)\phi \in D_A$, and*

$$\lim_{h \to 0} \frac{U(t+h)\phi - U(t)\phi}{h} = iU(t)A\phi = iAU(t)\phi.$$

This limit is clearly $\frac{dU}{dt}\phi$, and so we combine these results with the earlier results about the self adjointness of the Hamiltonian operator, H. We see that if H is self adjoint, by use of the Kato-Rellich theorem, we see that if $u_0 \in D_H$ that

$$\frac{\partial u}{\partial t} = \frac{d}{dt} e^{-itH} u_0 = -iHe^{-itH} u_0 = -iHu,$$

that is, the map $u(x,t) = e^{-itH} u_0(x)$ indeed solves the problem (1.1).

3.4 Faou's Proof

Now we would like to begin our backward error analysis for exponential splitting methods applied to (1.1). A first question one might have is why can't we immediately use the same approach as in chapter 2? To answer this, we discuss a key part of the approach from chapter 2, the BCH formula (2.4). We considered this formula in the context of matrices, but it is in fact valid in the more general context of Lie groups/algebras. This means that our formal use of the BCH formula was indeed valid when we considered Lie derivatives, as we can consider the Lie derivative as an element of the Lie algebra corresponding to some appropriate Lie group. We refer

the reader to [14] for further details on Lie groups/algebras.

We could now consider undergoing the same process with our splitting methods (for example Lie-Trotter $e^{-itH} \approx e^{it\Delta}e^{-itV}$) applied to our solution, (1.2), of (1.1). Our goal is to find an operator $\tilde{H}(t)$ such that $e^{-i\tilde{H}(t)} = e^{it\Delta}e^{-itV}$. If we try the same method and construct a formal series from (2.4), we obtain

$$\tilde{H}(t) = t(-\Delta + V) + \frac{t^2}{2}[-\Delta, V] + \frac{t^3}{12}\left(\Big[-\Delta, [-\Delta, V]\Big] + \Big[V, [V, -\Delta]\Big]\right) + ...$$

This is an infinite series of unbounded operators which we will want to consider the exponential of. The content of the previous chapter should raise a few issues with this. In particular, what domain should this operator be defined on? We need this operator to be self adjoint[3] so that we may apply the spectral theorem and functional calculus to define the exponential. However, as we saw towards the end of the chapter the sum of two self adjoint operators isn't guaranteed to be self adjoint. This is complicated considerably by the fact we have an infinite sum of operators, and so one requires some kind of extension of the Kato-Rellich theorem to infinite sums. Even when considering a suitable truncation, as we did in chapter 2, then we will still obtain several terms which will again require a generalisation of Kato-Rellich. Similarly, the form of this modified operator imposes more smoothness on a solution, due to the nested commutators involving the Laplacian. For example, if one considered a truncation of this operator up to t^3, there would be a term involving Δ^2.

Thus a new method is required to construct our modified equation. This section examines how the proof that Erwan Faou presents in [2] counters these issues. We refer the reader to [2] for the full details of the proof, but we will look at some important aspects.

Firstly, we note that the proof considers a different domain, $\mathbb{T}^d$, the d-dimensional torus. This is justified in a practical sense because when one computes these numerical solutions they will consider some truncated domain. The initial data, u_0, is then assumed to be periodic on the boundary of the torus, allowing the use of Fourier series. This idea is central to the proof, as the operators involved become multiplicative operators in Fourier space. This is actually related to the spectral theorem, as there is an equivalent formulation which says that a self adjoint operator is unitarily equivalent to a multiplicative operator on some appropriate L^2 space.

In the Fourier series form, we write the following

$$u(x) = \sum_{a\in\mathbb{Z}^d} \xi_a e^{ia\cdot x}, \qquad \bar{u}(x) = \sum_{a\in\mathbb{Z}^d} \eta_a e^{-ia\cdot x}.$$

[3]This wasn't an issue we had to consider with the Lie derivatives, despite them being unbounded operators. This is because there is a very neat, and convenient relationship between the Lie derivative and the flow. We refer the reader to [1] to see this.

Similarly, if one considers a real function $W(x)$, we have an associated operator in Fourier space via $W_{ab} = W_{a-b}$, where W_c is the Fourier coefficient of W corresponding to $c \in \mathbb{Z}^d$. This is motivated by the form that the Schrödinger equation takes in the Fourier space. The operator (W_{ab}) acting on Fourier space corresponds to the operator of multiplication by W in L^2 space. We then define a norm to measure the decay of operators with respect to the diagonal level $|b-a|$, where for $\alpha > 1$, we set

$$\|A\|_\alpha = \sup_{a,b\in\mathbb{Z}^d} |A_{ab}|(1+|b-a|^\alpha),$$

and define the space

$$\mathscr{L}_\alpha = \{A = (A_{ab}) \,|A_{ba} = \bar{A}_{ab},\ \|A\|_\alpha < \infty\}.$$

For an operator, A, if $A_{ba} = \bar{A}_{ab}$ then we say A is symmetric. We now state some properties of this norm.

Lemma 3.4.1 (Properties of $\|\cdot\|_\alpha$) *The norm $\|\cdot\|_\alpha$ has the following properties:*

1. *Real functions correspond to a symmetric operator in Fourier space, and if $V(x)$ is a smooth function then $\|V\|_\alpha < \infty$.*
2. *If $\alpha > d$, then there is a constant C_α such that for all operators A, B:*

$$\|AB\|_\alpha \le C_\alpha \|A\|_\alpha \|B\|_\alpha,$$

where the product of operators is defined by

$$(AB)_{ab} = \sum_{c\in\mathbb{Z}^d} A_{ac}B_{cb}$$

This norm is used throughout the proof. Similarly, we assume our potential function $V(x)$ is smooth, and so $\|V\|_\alpha < \infty$.

We've seen that our potential function can be bounded in the $\|\cdot\|_\alpha$ norm, but this isn't true for the Laplacian. As discussed earlier, the Laplacian is the more problematic term in the operator, and so to help control the norm we introduce a filter function, $\beta(x)$. The idea is that we only want to consider the Laplacian as being relevant in some region, not dissimilar how we consider a truncated domain. Two such filter functions which Faou presents are $\beta(x) = x\mathbf{1}_{x\le c}$, and $\beta(x) = 2\arctan(\frac{x}{2})$.

With this in mind, we now come back to the first step of the error analysis, constructing our modified operator. We know that we cannot immediately appeal to the BCH formula (2.4), and so Faou presents a different plan, where we now decouple the terms involving the potential and the Laplacian. This is done as follows:

1. We fix our timestep $\tau > 0$, and define the operator $A_0 = -\beta(\tau\Delta)$ via the functional calculus for Δ.

2. The operator A_0 is now fixed, and so we look for a function $t \mapsto Z(t)$ such that $Z(0) = A_0$, and for $t \in [0, \tau]$ we have

$$\exp(-itV)\exp(-iA_0) = \exp(-iZ(t)).$$

3. Then we set $t = \tau$, and obtain the result.

We see this is different to the approach taken with BCH-like formulae, as here our term in Δ is now fixed. If such an expression exists, then differentiation yields the ODE

$$Z'(t) = \sum_{k=0}^{\infty} \frac{B_k}{k!}(-1)^k ad_{iZ(t)}^k(V).$$

From this, we define the formal series,

$$Z(t) = \sum_{l=0}^{\infty} t^l Z_l,$$

with $Z_0 = A_0 = -\beta(\tau\Delta)$. This forms the basis of Faou's argument.

It can be verified from properties of Fourier series that in Fourier space the operator A_0 is a diagonal operator with coefficients, $\lambda_a = (A_0)_{aa} = \beta(\tau|a|^2)$. This observation allows us to bound the norm of the operator ad_{A_0}, as follows:

Lemma 3.4.2 *Let A_0 be the diagonal operator with eigenvalues $\lambda_a = \beta(\tau|a|^2)$, and assume that for all $a \in \mathbb{Z}^d$ that $0 \leq \lambda_a \leq \pi$. Let $W = (W_{ab})$ be an operator in $\mathscr{L}_\alpha$ for some $\alpha > 1$. Then we have*

$$\|ad_{A_0}W\|_\alpha \leq \pi\|W\|_\alpha.$$

This condition on A_0 is satisfied if the filter function satisfies $0 \leq \beta(x) \leq \pi$ for all x. This holds for the previously mentioned filter functions, where we take $c = \pi$. Under these conditions, we now may construct the modified operator.

Theorem 3.4.3 (Modified Operator) *Let $\alpha > d$, and assume that $\|V\|_\alpha < \infty$. Assume that the eigenvalues, λ_a, of the operator A_0 satisfy $0 \leq \lambda_a \leq \pi$. Then there exists $\tau_0 > 0$ and a constant C such that, for all $\tau \in (0, \tau_0)$, there exists a symmetric operator $S(\tau)$ such that*

$$\exp(i\tau V)\exp(-iA_0) = \exp(-iS(\tau)).$$

Moreover, we have

$$S(\tau) = -\frac{1}{\tau}\beta(\tau\Delta) + V(\tau) + \tau W(\tau),$$

where $V(\tau), W(\tau)$ satisfy

$$\|V(\tau)\|_\alpha \leq C\|V\|_\alpha, \qquad \|W(\tau)\|_\alpha \leq C\|V\|_\alpha^2.$$

The heavy lifting, so to speak, is largely done in the set up of this theorem. The loose synopsis is that we use the bounds given above, and Cauchy estimates to bound terms and define a formal series, which solves a related ODE, that converges in some radius of convergence. In particular, this proof explicitly gives a value, $\tau_0 = \frac{\pi}{48MC_\alpha\|V\|_\alpha}$, which can be viewed as the maximal timestep.

This constructs our modified operator, from which we have a corresponding modified energy. We give two results relating to properties of this modified energy. The first result shows that under certain conditions the modified energy is close to the actual energy. The second result shows an exact conservation of modified energy for the Lie-Trotter splitting.

Lemma 3.4.4 (First Energy Result) *Let $\nu \in [0,1]$, and suppose that A_0 is associated with the filter function $\beta(x) = 2\arctan\left(\frac{x}{2}\right)$. Assume that $u \in H^{1+\nu}(\mathbb{T}^d)$, then we have for $\tau \in (0, \tau_0)$,*

$$|\langle u, S(\tau)u\rangle - \langle u, (-\Delta + V)u\rangle| \leq C\tau^\nu \|u\|^2_{H^{1+\nu}},$$

where C depends on ν and V.

The proof of this result relies on the form of the filter function, and uses bounds on arctan. Two other notable aspects of this proof are that in the proof for the modified operator one obtains explicit expressions for $V(\tau), W(\tau)$, and these expressions are used in this proof. Lastly, there are a few properties of Sobolev norms required at the end of the proof, such as $\|v\|_{H^{2\nu}} \leq \|v\|_{H^{1+\nu}}$ for $\nu \in [0,1]$, and a relation between the norm in terms of Fourier coefficients and the Sobolev norm $\|\cdot\|_{H^{1+\nu}}$.

Corollary 3.4.4.1 (Second Energy Result) *Assume that $u_0 \in L^2(\mathbb{T}^d;\mathbb{C})$, and $\tau \in (0,\tau_0)$. We define the following*

$$u_n := (\exp(-itV)\exp(-iA_0))^n u_0.$$

Then for all $n \geq 0$, we have the conservation of modified energy,

$$\langle u_n, S(\tau)u_n\rangle = \langle u_0, S(\tau)u_0\rangle.$$

The proof of this result is very straightforward. We write

$$\langle u_1, S(\tau)u_1\rangle = \langle \exp(-i\tau S(\tau))u_0, S(\tau)\exp(-i\tau S(\tau))u_0\rangle,$$

and note (as observed when we defined the functional calculus), that $S(\tau)$, and $exp(\pm i\tau S(\tau))$ commute. Thus we find

$$\langle u_1, S(\tau)u_1\rangle = \langle u_0, \exp(i\tau S(\tau))S(\tau)\exp(-i\tau S(\tau))u_0\rangle = \langle u_0, S(\tau)u_0\rangle,$$

and then one repeats this process inductively for the result.

We can compare these results with the finite dimensional results of chapter 2. The first result can be viewed in the same way that the modified Hamiltonian for a symplectic method, of order p, can be written as (2.9). Similarly the second result is analogous to the Theorem 2.3.5. Moreover, the quantity h_0 plays a similar role to τ_0 and we can interpret them both as maximal timesteps[4].

3.5 Faou's Technique Applied to the Strang Splitting

Faou remarks that these results can extend to both the Lie Trotter scheme where we first consider the potential, i.e. $\exp(-i\tau A_0)\exp(-i\tau V)$, and the Strang splitting $\exp(-i\tau\frac{V}{2})\exp(-i\tau A_0)\exp(-i\tau\frac{V}{2})$. However, this analysis is not presented in [2]. We investigate this remark in greater detail. In the case of the Lie-Trotter splitting with the operators permuted, this is very straightforward. As the Strang splitting is a more complex splitting there is some concern that this may be reflected in the analysis.

Indeed, we immediately run into an issue. We consider the same set up with finding $t \mapsto Z(t)$ such that

$$\exp\left(\frac{-itV}{2}\right)\exp(-iA_0)\exp\left(\frac{-itV}{2}\right) = \exp(-iZ(t)),$$

with $Z(0) = A_0$. Now we differentiate this, to obtain

$$-\frac{iV}{2}\exp(-iZ(t)) - \exp(-iZ(t))\frac{iV}{2} = -i\left[\left(\frac{\exp(ad_{-iZ(t)})-1}{ad_{-iZ(t)}}\right)\frac{dZ}{dt}\right]\exp(-iZ(t)),$$

which we can simplify to

$$\frac{V}{2} + \exp(-iZ(t))\frac{V}{2}\exp(iZ(t)) = \left[\left(\frac{\exp(ad_{-iZ(t)})-1}{ad_{-iZ(t)}}\right)\frac{dZ}{dt}\right].$$

Now applying the dexp^{-1} formula one obtains an ODE for $Z(t)$,

$$Z'(t) = \sum_{k=0}^{\infty}\frac{B_k}{k!}(-1)^k ad_{iZ(t)}\left(\frac{V}{2} + \exp\left(-\frac{itV}{2}\right)\exp(-iA_0)\frac{V}{2}\exp(iA_0)\exp\left(\frac{itV}{2}\right)\right),$$

which is considerably more complicated than the ODE arising from the Lie Trotter splitting. We assume that the argument in $ad_{iZ(t)}(\cdot)$, which we will denote $\tilde{V}(t)$, can be bounded as follows

$$\|\tilde{V}(t)\|_\alpha \leq \frac{1}{2}\left(\|V\|_\alpha + \left\|\exp\left(-\frac{itV}{2}\right)\exp(-iA_0)V\exp(iA_0)\exp\left(\frac{itV}{2}\right)\right\|_\alpha\right) \leq K\|V\|_\alpha,$$

[4]In fact, in Theorem 2.3.4 we see that the maximal timestep is actually $\frac{h_0}{4}$, but this is still a valid comparison.

where K is a constant which will be discussed later. In particular, this gives us the important bound

$$\|ad_{A_0}(\tilde{V}(t))\|_\alpha \leq \pi \|\tilde{V}(t)\|_\alpha \leq \pi K \|V\|_\alpha,$$

which will be required for some of the early bounds in the proof.

Proceeding as Faou does, if we define the formal series,

$$Z(t) = \sum_{l=0}^{\infty} t^l Z_l,$$

then one can rewrite the ODE, and compare the coefficients of t to obtain, for $l \geq 1$

$$(l+1)Z_{l+1} = \sum_{k=0}^{\infty} \frac{B_k}{k!}(-i)^k \left(\sum_{l_1+\ldots+l_k=l} ad_{Z_{l_1}} \ldots ad_{Z_{l_k}}(\tilde{V}(t)) \right).$$

In particular, at $l = 1$ we obtain

$$Z_1 = \sum_{k=0}^{\infty} \frac{B_k}{k!}(-i)^k ad_{A_0}^k(\tilde{V}(t)).$$

These expressions are more or less identical to those presented by Faou, where the only difference is the term $\tilde{V}(t)$. From here, it is relatively straightforward to see that the analysis will carry over. I encourage the reader to look through the proof of interest (Theorem V.7 of [2]), and it should be clear that the analysis transfers if one replaces the term $\|V\|_\alpha$ with $K\|V\|_\alpha$ when analysing the Strang splitting.

One may be interested in the value of this constant K, as it is the main difference in these results. Using the properties of $\|\cdot\|_\alpha$ it is easy to see that

$$\|\tilde{V}(t)\|_\alpha \leq \frac{1}{2}\left(1 + C_\alpha^4 \left\|exp\left(-\frac{itV}{2}\right)\right\|_\alpha \left\|exp\left(\frac{itV}{2}\right)\right\|_\alpha \|exp\left(-iA_0\right)\|_\alpha \|exp\left(iA_0\right)\|_\alpha\right) \|V\|_\alpha.$$

Appealing to the definition of $\|\cdot\|_\alpha$ and by observing A_0 is a diagonal operator, we see that

$$\|exp(\pm iA_0)\|_\alpha = \sup_{a\in\mathbb{Z}^d} |e^{\pm i\beta(\tau|a|^2)}| = 1.$$

The potential term is less convenient to work with as it isn't a diagonal operator. However, this does still give a possible value of K as

$$K = \sup_{t\in\mathbb{R}} \frac{1}{2}\left(1 + C_\alpha^4 \left\|exp\left(-\frac{itV}{2}\right)\right\|_\alpha \left\|exp\left(\frac{itV}{2}\right)\right\|_\alpha\right).$$

It appears that this constant K is dependent on the potential V, and that different potentials will give rise to different behaviour. For a given potential, V, if it can be shown that $K < 1$ then this gives the Strang splitting a higher value of τ_0 than the Lie-Trotter splitting, which would imply it is a valid method for more timesteps.

Conversely, if $K > 1$ this reduces the value of τ_0, which would mean that the higher order of the Strang splitting comes at the cost of a tighter restriction on the timestep.

A quick numerical experiment with the energy error, with the same set up as used in section 1.3, seems to suggest that the Strang splitting is well behaved for a wider range of timesteps. This is indicative of $K < 1$ for $V(x) = x^4 - 10x^2$, but the integration here is on a relatively small timescale, so this isn't particularly conclusive.

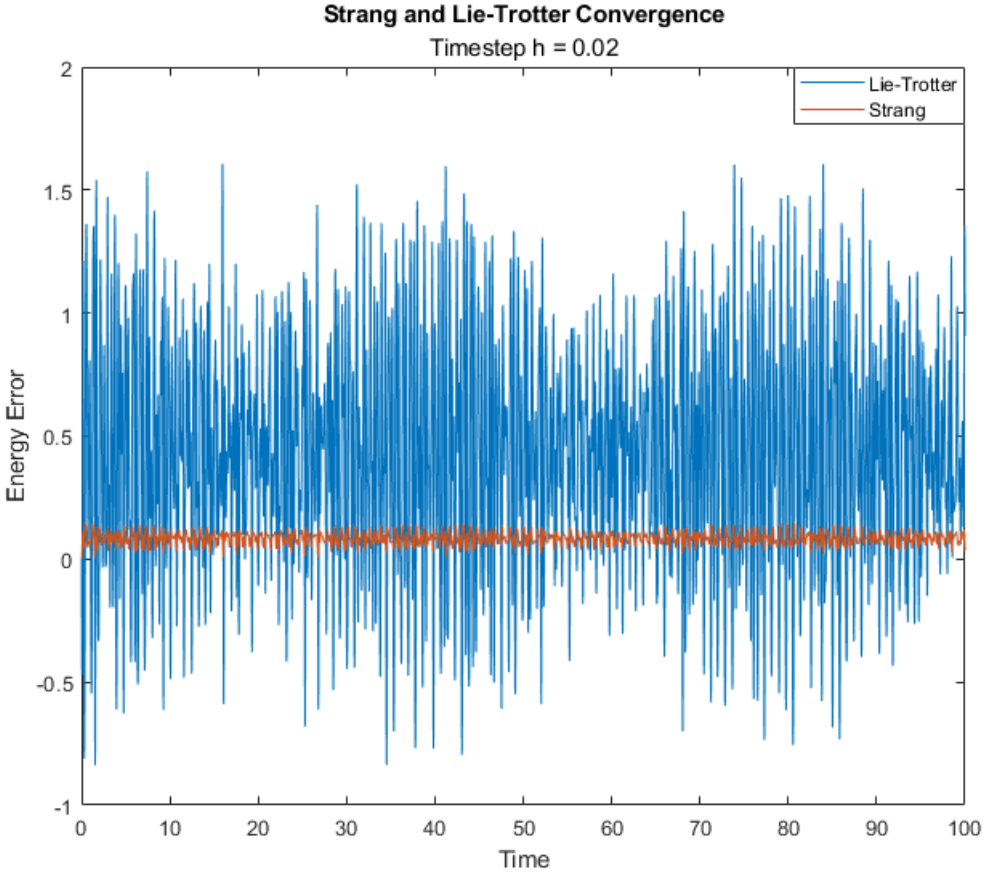

Figure 3.1: Experiment, with $V(x) = x^4 - 10x^2$, showing Strang behaving well and Lie-Trotter being comparatively large in error.

Chapter 4

Time Dependence and the Magnus Expansion

This final chapter considers the problem (1.1) with a time dependent potential $V(x,t)$, for which we cannot apply the exponential splittings from the previous chapters. Section 4.1 introduces a technique for this problem via the Magnus expansion, and we refer the reader to [4] and [15] for further details. Section 4.2 discusses the long term behaviour of a Magnus based solution, and illustrates a new phenomenon in the energy error bounds from [5]. Finally, section 4.3 utilises the content of section 3.3, and the techniques of [4] and [5] to achieve error bounds on the exponential splittings seen in the previous chapters. Moreover, we obtain a new result for the near norm conservation for a family of exponential splitting methods for (1.1).

4.1 The Magnus Expansion

As noted earlier, exponential splitting methods, such as Lie-Trotter (1.3) and Strang (1.4), aren't immediately applicable when one has a time dependent Hamiltonian. This is because the solution of (1.1) is no longer $e^{itH}u_0$. In this scenario, a common approach is to use Magnus expansion based methods. The idea of the Magnus expansion is to find a solution of a differential equation of the form

$$\frac{du}{dt} = A(t)u, \qquad u(0) = u_0,$$

as an exponential $u(h) = \exp(\Theta(h))u_0$, for some sufficiently small timestep, h, and where $\Theta(t)$ is some time dependent operator. If $\Theta(t)$ exists [17], it can be expressed as an infinite series,

$$\Theta(t) = \sum_{k=1}^{\infty} \Theta_k(t),$$

where the first two terms are

$$\Theta_1(t) = \int_0^t A(\xi_1)\, d\xi_1, \quad \Theta_2(t) = \frac{1}{2}\int_0^t \left[\int_0^{\xi_1} A(\xi_2)\, d\xi_2, A(\xi_1)\right] d\xi_1,$$

with square brackets denoting the commutator of operators. This power series is known as the Magnus expansion for $\Theta(t)$.

As the Magnus expansion is an infinite series, there are immediately questions about its convergence. In fact, it is known [20] that the Magnus expansion will converge absolutely for $0 \leq t < T$, where

$$T = \max \left\{ t \geq 0 : \int_0^t \|A(s)\|_2 \, ds < \pi \right\}.$$

Moreover, there is an example of divergence with $\int_0^t \|A(s)\|_2 \, ds = \pi$ (see [21]). This gives us some idea on how small we require our timestep h to be. We refer the reader to [17] for further details on the Magnus expansion.

A Magnus based method then involves some truncation of this series, often followed by some discretization. An explicit example of such a Magnus based method is the exponential midpoint method

$$u_{n+1} = \exp\left(hA\left(nh + \frac{h}{2} \right) \right) u_n. \tag{4.1}$$

The Magnus expansion is known to be have a time symmetric property (see [17]). That is, for h sufficiently small we have $\Theta(-h)\Theta(h) = \mathrm{id}$. Hence, if the operator $A(t)$ is analytic, this allows one to express the Magnus expansion of $\Theta(h)$ exclusively in odd powers of h. A remarkable property of Magnus based methods, is that if the operator $A(t)$ is analytic, then they inherit this time symmetric property, as proven in [16]. In particular, this gives rise to methods gaining higher order in their error, as noted in [4].

Some of the initial error analysis of Magnus based methods was presented in [5], where the operators considered were matrices, and an extension of this to the Schrödinger equation was considered in [4]. We will consider the error analysis of these methods in the following sections. In particular, a useful feature of these methods is that they have a straightforward construction of a modified equation which the truncated operator solves exactly.

4.2 Long Term Behaviour of Magnus Based Methods

We now investigate the long term behaviour of Magnus based methods, in particular the exponential midpoint method (4.1). Contrary to the case where we have considered a time independent potential, the energy of this system isn't a conserved quantity. As such, we cannot expect an integrator to have near energy conservation for this system. The question then becomes how does the error in energy behave as

time increases. In the following plot we consider the same domain and initial data as in section 1.4, but now we use a laser potential $V(x,t) = x^4 - 10x^2 + 10\sin(8\pi t)x$.

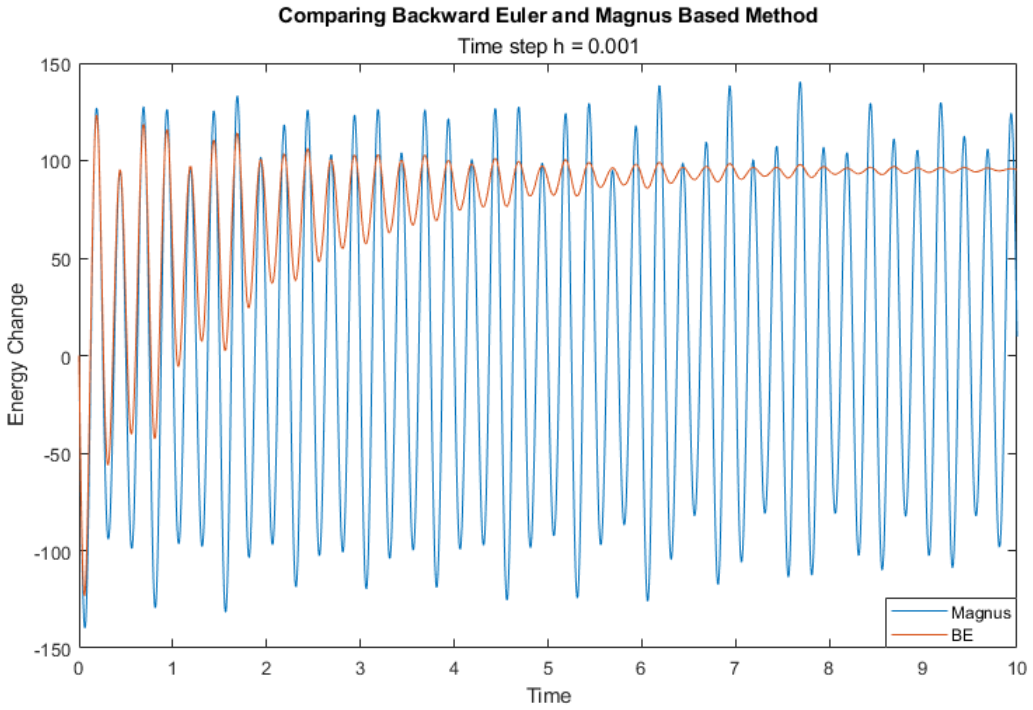

Figure 4.1: Energy Change for Backward Euler and Midpoint Method

This potential corresponds to, as the name suggests, the emission of photons by a laser. In this system energy is being pumped in to cause an excitation of electrons, which then causes photons to be emitted, and the electrons return to a lower energy level. From a physical view point, this suggests the oscillatory behaviour displayed by the Magnus method is accurate, and the backward Euler method is again inaccurate. This shows some of the promising features of these Magnus based methods, however their analysis presents an issue with their energy error.

In [5], Hochbruck and Lubich present analysis for the case of matrix functions, that is a Hamiltonian of the form $H(t) = U + V(t)$ for U symmetric positive definite and $V(t)$ Hermitian. This matrix case corresponds to a discretization of the problem (1.1). In this paper it is shown that the exponential midpoint method, used above, is expected to accumulate energy error as time progresses. More specifically, we assume that the Hamiltonian operator satisfies

$$\|[H(t), H(\sigma)]v\| \leq K_1 h \|Dv\|, \text{ for } |t-\sigma| \leq h \tag{4.2}$$

and

$$\left\| \frac{d^m V}{dt^m} \right\| \leq M_m, \tag{4.3}$$

where K_1 and M_m are some constants and $D = U^{\frac{1}{2}}$.

Under these assumptions it can be shown that

$$\|u_n - u(t_n)\| \leq Ch^2 t_n \max_{0 \leq t \leq t_n} \|u(t)\|.$$

This can in be generalised for higher order Magnus integrators, under further assumptions. This can be applied to obtain an error bound on the energy,

$$\langle u_n, Hu_n \rangle - \langle u(t_n), Hu(t_n) \rangle = \mathcal{O}(t_n h^2).$$

This suggests that the energy error grows in time, which is not ideal.

Some numerical experiments suggest this bound may be overly pessimistic. In particular, by considering potentials where the time dependence becomes less significant as time progresses we can observe, numerically, that these bounds are not too tight. Specifically we considered potentials of the form $V(x,t) = x^4 - 10x^2 + e(t)x$, for some function $e(t)$ which tends to zero at t goes to infinity.

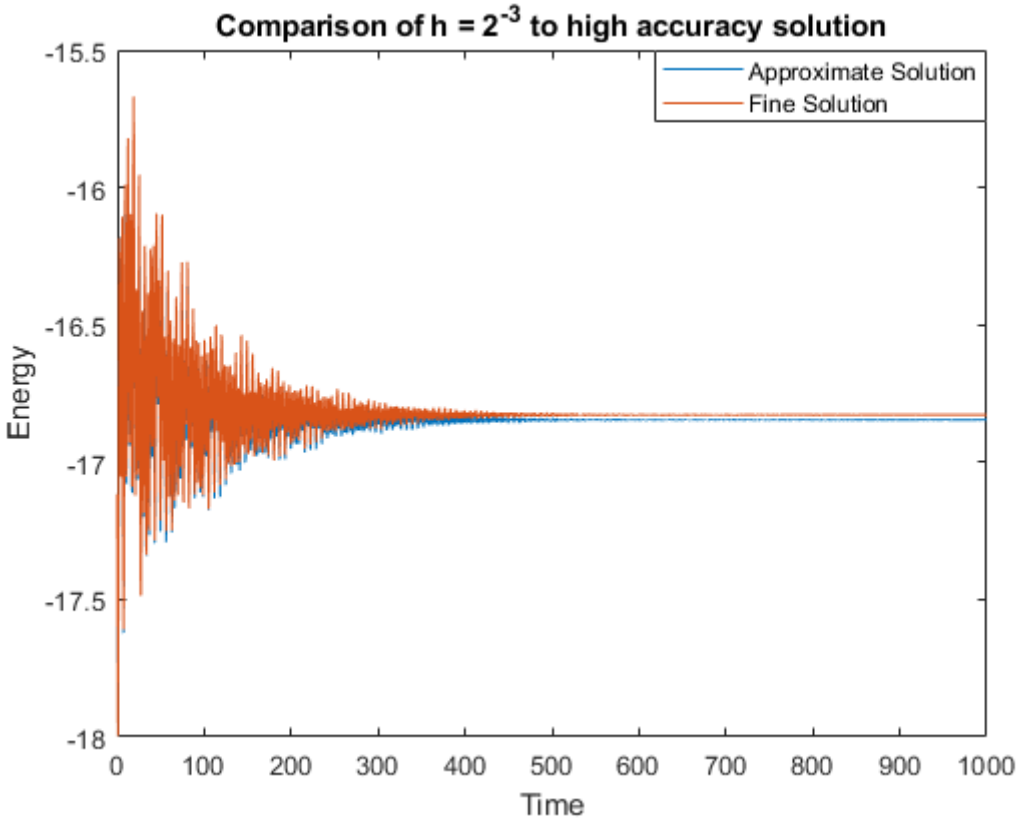

Figure 4.2: Long time integration of with $e(t) = e^{-0.1t}\sin(2\pi t)$.

In the above figure, we compare two solutions, one with $h = 2^{-3}$, and the other with $h = 2^{-7}$. It is clear that for large time, that the difference in energy does not

increase with time, and in fact at the end time we have the difference is approximately 0.02.

This process was repeated with a family of similar potentials figures (4.3) and (4.4). We consider a rough solution ($h = 2^{-3}$), and a fine solution ($h = 2^{-6}$), with end time $T = 500$. Denoting the energy of the rough solution at time t by $E_1(t)$, and the energy for the fine solution by $E_2(t)$. We observe that there is roughly a linear relation between the difference in energy at a final time (i.e. $E_1(T) - E_2(T)$) and the total change in energy of the system (i.e. $E_2(T) - E_2(0)$).

This behaviour seems to show that if the net gain/loss of energy at the end time is sufficiently small, then the energy error will remain small too. In particular, this suggests that if we consider a potential where the time dependent part stays bounded, then there shouldn't be an accumulation of error as time progresses, contrary to the bounds from [5].

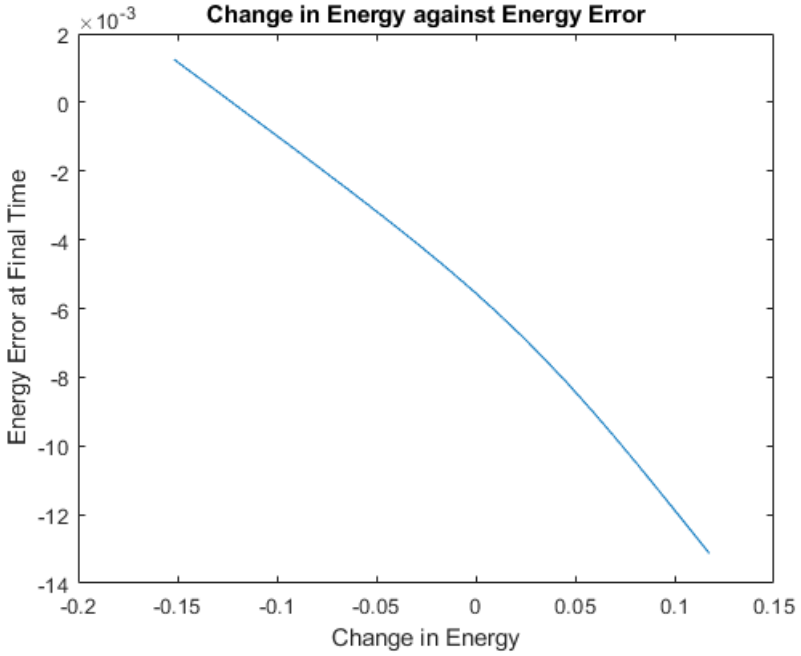

Figure 4.3: Family of exponentials, going from $e(t) = e^{-0.5t}\sin(2\pi t)$ to $e(t) = e^{-0.1t}\sin(2\pi t)$.

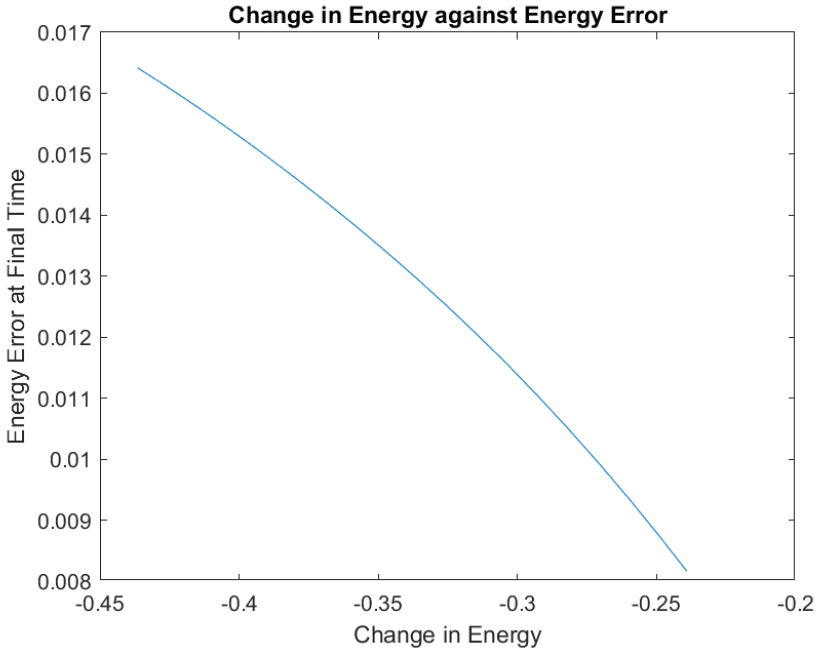

Figure 4.4: Family of rational functions, going from $e(t) = \frac{\sin(2\pi t)}{1+10t^2}$ to $e(t) = \frac{\sin(2\pi t)}{1+50t^2}$.

4.3 A Magnus Based Approach to Exponential Splittings

An entirely different approach to the error analysis of exponential splittings is to use bounds for Magnus based methods for a well chosen operator $A(t)$. To do this for the Lie-Trotter Splitting, we could define the operator

$$A(t) = \begin{cases} -2iV, & t \in \left[0, \frac{h}{2}\right) \\ 2i\Delta, & t \in \left[\frac{h}{2}, h\right] \\ 0, & \text{otherwise} \end{cases},$$

and consider the differential equation

$$u'(t) = A(t)u(t), \qquad u(0) = u_0.$$

Then by completing the integration over $[0, h]$ will give our solution $u_1 = u(h) = e^{i\Delta}e^{-iV}u_0$. That is, the Lie-Trotter splitting is the flow of this differential equation at time h. Similarly, for the Strang splitting one can consider the operator,

$$A^{[S]}(t) = \begin{cases} -2iV, & t \in \left[0, \frac{h}{4}\right) \\ 2i\Delta, & t \in \left[\frac{h}{4}, \frac{3h}{4}\right) \\ -2iV, & t \in \left[\frac{3h}{4}, h\right] \\ 0, & \text{otherwise} \end{cases},$$

in the same way as we did above. We note that the operator $A(t)$ is not analytic, and so we cannot expect to gain a time symmetry property from this. This is not surprising, as it is clear that the Lie Trotter splitting isn't time symmetric. One can easily observe this, as generally $e^{ihA}e^{ihB}e^{-ihA}e^{ihB} \neq \mathrm{id}$, for operators A, B.

One may similarly express this solution in terms of the Magnus expansion. This gives us a solution of the form, $u_1 = e^{\Theta(h)}u_0$, and direct comparison of these lets us see

$$e^{\Theta(h)} = e^{ih\Delta}e^{-ihV}.$$

That is to say, the operator $\Theta(h)$ is playing the role of the modified Hamiltonian operator. Much like the BCH formula, the Magnus expansion is considered as a formal series, as the operator $\Theta(h)$ isn't necessarily well defined.

We initially present the analysis in the simpler matrix case, and then move on to the case with unbounded operators. In the latter case, it is worth mentioning that we instead consider the splitting of the form $e^{-ihV}e^{ih\Delta}$ and

$$A(t) = \begin{cases} 2i\Delta, & t \in \left[0, \frac{h}{2}\right) \\ -2iV, & t \in \left[\frac{h}{2}, h\right] \\ 0, & \text{otherwise} \end{cases},$$

for technical reasons which we discuss later.

To begin the analysis of such a method, we follow the process from [5]. We assume the Hamiltonian, $H = -U + V$, is in terms of matrices, so the operator $A(t)$ is now

$$A(t) = \begin{cases} -2iV, & t \in \left[0, \frac{h}{2}\right) \\ 2iU, & t \in \left[\frac{h}{2}, h\right] \\ 0, & \text{otherwise} \end{cases},$$

for U symmetric positive definite, and V Hermitian. We will be looking at the first order Magnus method, $\Theta_1(t) = \int_0^t A(s)\,ds$. The piecewise form of A is useful here, as we find that this becomes

$$\Theta_1(h) = \int_0^h A(s)\,ds = \int_0^{\frac{h}{2}} -2iV\,ds + \int_{\frac{h}{2}}^h 2iU\,ds = -ihV + ihU = -ihH.$$

We do not assume the bounds (4.2) and (4.3) used in [5].

We define $u(t) = e^{itU}e^{-itV}u_0$, and $\tilde{u}(t) = e^{\Theta_1(t)}u_0$. Then it can be shown that $\tilde{u}$ exactly solves the modified equation

$$\frac{d\tilde{u}}{dt} = \tilde{A}(t)\tilde{u}, \qquad \tilde{u}(0) = u_0,$$

where $\tilde{A}(t) = \varphi(ad_{\Theta_1})(\Theta_1')$, for $\varphi(z) = \frac{e^z-1}{z}$. Following the method in [5], one can express this function as $\varphi(z) = 1 + \frac{1}{2}z\, r_2(z)$, where $r_2(z)$ is some residual function. This then allows us to express $\tilde{A}$ as

$$\tilde{A}(t) = A(t) + \frac{1}{2} r_2(ad_{\Theta_1})(\Theta_1'(t)),$$

where we consider $0 \leq t \leq h$. We now use the form of Θ_1 to see that $\Theta_1'(t) = A(t)$, and hence the difference between these operators is

$$\tilde{A}(t) - A(t) = \frac{1}{2} r_2(ad_{\Theta_1})(A(t)).$$

Our goal is to bound the error, $\tilde{\epsilon}(t) = \tilde{u}(t) - u(t)$, as

$$\|\tilde{\epsilon}(t)\| \leq \int_0^t \|(\tilde{A}(t) - A(t))u(t)\|\, ds. \tag{4.4}$$

This inequality is given as Lemma 4.1 of [5], where they require $A(t)$ is skew Hermitian, which our A is since it is a multiple of i of a Hermitian matrix[1] on each piece. This proof requires some amount of differentiability, and so in the proof we consider the derivative on each piece.

From this inequality it is clear that we want a bound for $\|r_2(ad_{\Theta_1})(A(t))u(t)\|$. To obtain a bound of this form, we use the same bound as in [5], which is

$$\|r_2(ad_{\Theta_1})(A(t))u(t)\| \leq \|\hat{r}_2\|_{L^1} \sup_{\xi \in \mathbb{R}} \|ad_{\Theta_1}(A(t)) \exp(\xi\Theta_1)u(t)\|\,,$$

where the hat denotes the Fourier transform. Now, we note that

$$\|ad_{\Theta_1}(A(t))\| = \left\| \int_0^t A(s)\, ds A(t) - A(t) \int_0^t A(s)\, ds \right\| \leq 2h\|A(t)\|^2$$

and one can write

$$A(t) = 2i\left(U\hat{I}\left(t - \frac{h}{2}\right) - V\hat{I}(t)\right),$$

where $\hat{I}(t) = \mathbf{1}_{0 \leq t < h/2}\mathrm{id}$. From this it is clear to see that $\|A(t)\| \leq 2(\|U\| + \|V\|)$, and so we may write $\|ad_{\Theta_1}(A(t))\| \leq Ch$, where C is a constant independent of h. Now, as noted in [5], as $A(t)$ is Hermitian on each piece we know that $\Theta_1(t)$ is skew-Hermitian, and hence $\exp(\xi\Theta_1(t))$ is unitary. Thus we may combine these results to see

$$\|(\tilde{A}(t) - A(t))u(t)\| = \|r_2(ad_{\Theta_1})(A(t))u(t)\| \leq Ch\|u(t)\|.$$

Finally, we see that as $0 \leq t \leq h$ one may apply (4.4) with the above bound to see

$$\|\tilde{u}(h) - u(h)\| = \|\tilde{\epsilon}(h)\| \leq Ch^2. \tag{4.5}$$

[1] It can be shown that U is Hermitian as it is positive definite.

From this we look at the error at further time steps. We let $\tilde{u}_n = (\exp(\Theta_1(h)))^n u_0$, and $u_n = (\exp(ihU)\exp(-ihV))^n u_0$. It should be clarified that $\tilde{u}_n$ is the exact solution, and u_n is the Lie-Trotter solution, at time $t = nh$.

We want to use (4.5) to attain a bound for $\|u_n - \tilde{u}_n\|$. Using the definitions of u_n and $\tilde{u}_n$, and the triangle inequality, one obtains

$$\|u_n - \tilde{u}_n\| \leq \|\exp(ihU)\exp(-ihV)u_{n-1} - \exp(\Theta_1(h))u_{n-1}\| + \|\exp(\Theta_1(h))u_{n-1} - \exp(\Theta_1(h))\tilde{u}_{n-1}\|.$$

By considering the scheme with initial data $u_0 = u_n$, and by using (4.5), it is clear that $\|\exp(ihU)\exp(-ihV)u_n - \exp(\Theta_1(h))u_n\| \leq Ch^2$. For the second term, we observe that $\exp(\Theta_1(h))$ is unitary, and hence

$$\|u_n - \tilde{u}_n\| \leq Ch^2 + \|u_{n-1} - \tilde{u}_{n-1}\|.$$

Iterating this result over each time step, and noting that $\tilde{u}_0 = u_0$, we arrive at the bound

$$\|u_n - \tilde{u}_n\| \leq Ct_n h.$$

This agrees with the known results for the Lie-Trotter splitting, and indeed the first order behaviour observed in section 1.4.

Now we turn our focus towards the energy conservation of this method. We know from section 1.3 that we have the following conservation of energy, for all $n \in \mathbb{N}$,

$$\langle \tilde{u}_n, H\tilde{u}_n \rangle = \langle u_0, Hu_0 \rangle.$$

We would like to show that u_n has an approximate conservation of this form. To do this, we take sufficiently small h to ensure the convergence of the Magnus expansion (see [20]) so that $\Theta(h)$ is well defined. Then, as we saw for the truncated Magnus expansion, Θ_1, $\Theta(h)$ is a skew-Hermitian matrix and hence $e^{\Theta(h)}$ is unitary. We define the modified Hamiltonian as $\tilde{H} = -i\Theta(h)$, and we observe that

$$\langle u_n, \tilde{H}u_n \rangle = \langle e^{n\Theta(h)}u_0, \tilde{H}e^{n\Theta(h)}u_0 \rangle = \langle e^{n\Theta(h)}u_0, e^{n\Theta(h)}\tilde{H}u_0 \rangle = \langle u_0, \tilde{H}u_0 \rangle.$$

From the form of the Magnus expansion, we may write $\Theta(h) = \Theta_1(h) + E(h)$, where $\|E(h)\| = \mathcal{O}(h)$. Thus we may write

$$\langle u_0, \tilde{H}u_0 \rangle = \langle u_0, (H - iE(h))u_0 \rangle = \langle u_0, Hu_0 \rangle - i\langle u_0, E(h)u_0 \rangle,$$

and hence by use of the Cauchy-Schwarz inequality

$$|\langle u_0, \tilde{H}u_0 \rangle - \langle u_0, Hu_0 \rangle| = |\langle u_0, E(h)u_0 \rangle| \leq \|u_0\|\|E(h)u_0\| = \mathcal{O}(h).$$

Thus, for all times we see that the Lie-Trotter splitting has an approximate conservation of energy, with $\mathcal{O}(h)$ error. This is again consistent with our earlier observations, and known results.

It is worth noting that while we required a restriction on the timestep for the energy result, we did not for the norm preservation. Moreover, this proof did not explicitly use the form of the Lie-Trotter splitting. Indeed, we only required an exponential splitting of the form $e^{a_1 ihU}e^{-b_1 ihV}...e^{a_m ihU}e^{-b_m ihV}$ such that

$$\sum_{k=1}^{m} a_k = \sum_{k=1}^{m} b_k = 1,$$

as in this case $\Theta_1(h) = -ihH$. Thus one sees for the matrix case, that any consistent exponential splitting for this problem has near norm and energy preservation properties, as seen in figure (4.5).

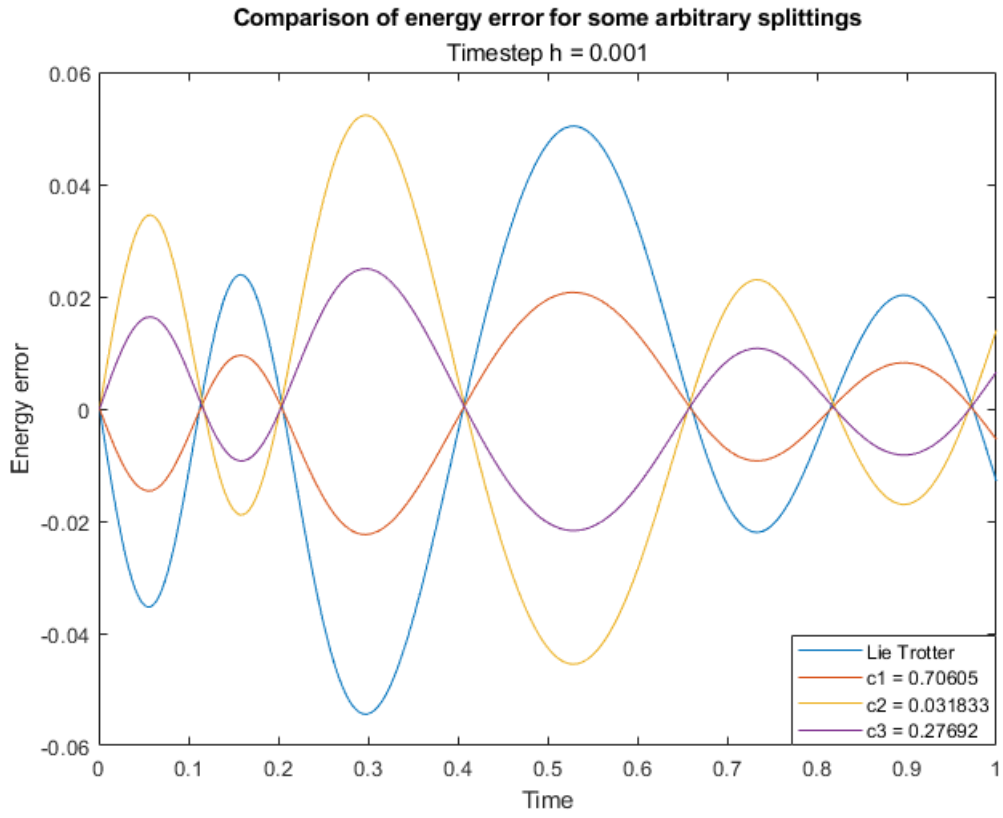

Figure 4.5: Comparison of energy error for splittings of the form $e^{-ic_k hV}e^{ihU}e^{-i(1-c_k)hV}$, using the same data as in section 1.4.

However, it seems that this technique might prove challenging to extend to higher order methods like the Strang splitting (1.4). This is as showing the higher accuracy of higher order methods one would require commutator bounds, like (4.2). Such bounds may be difficult to obtain, as our operator $A(t)$ is not analytic.

A common theme throughout this dissertation is that one encounters many problems when considering unbounded operators. For example, when we considered $A(t)$

as a matrix valued function, as the operators are matrices we have a very obvious domain. However, when we consider the operator,

$$A(t) = \begin{cases} 2i\Delta, & t \in \left[0, \frac{h}{2}\right) \\ -2iV, & t \in \left[\frac{h}{2}, h\right] \\ 0, & \text{otherwise} \end{cases},$$

this domain of definition is less obvious. To counter this, we take inspiration from the Kato-Rellich theorem (Theorem 3.3.6). We assume throughout that the operators, Δ and V, satisfy the conditions of the Kato-Rellich Theorem, such that $D_\Delta \subseteq D_V$. We refer the reader to section 3.3, Theorem 3.3.5 for the specific domains of these operators. We choose this domain so that the operator $-i\Theta_1(t)$ is self adjoint for $t \in [0, h]$, and so we may use results from section 3.3. We note that it is important that we define $A(t)$ by taking Δ first, as the condition $D_\Delta \subseteq D_V$ is less restrictive than the converse inclusion, and we require self adjointness on each piece. Using this domain of definition, we may begin analysis as we did above.

We use similar notation as before, defining $u(t) = e^{-itV}e^{it\Delta}u_0$, and $\tilde{u}(t) = e^{\Theta_1(t)}u_0$, where u_0 is our initial condition. As seen earlier, this will give us $u(h)$ being the Lie-Trotter solution for time $t = h$, and $\tilde{u}(h)$ the exact solution for time $t = h$. Then, as before, it can be shown that $\tilde{u}(t)$ is the exact solution to the modified equation,

$$\frac{d\tilde{u}}{dt} = \tilde{A}(t)\tilde{u}, \qquad \tilde{u}(0) = u_0,$$

for some modified operator $\tilde{A}(t)$. Moreover, the operators A and $\tilde{A}$ have the same relation as before, $\tilde{A}(t) - A(t) = \frac{1}{2}r_2(ad_{\Theta_1})(A(t))$, although here we require the functional calculus to define the operator $r_2(ad_{\Theta_1})(A(t))$. As we know $\Theta_1(t) = \int_0^t A(s)\,ds$, we observe that $-i\Theta_1(t)$ is a self adjoint operator on D_Δ and so we may use the results from the spectral theorem to define the operator $r_2(ad_{\Theta_1})(A(t))$.

We would like to proceed as we did in the matrix case, and to do so we follow similar arguments to those presented in Chapter 9 of [4]. Indeed, Singh shows that all the results we required from [5] are applicable with the Hamiltonian $H = -\Delta + V$. In particular the error bound, (4.4), is proven to still hold, where the norm is now the L^2 norm $\|\cdot\|_{L^2}$, under the condition that the operator $A(t)$ is skew-Hermitian. Our piecewise defined $A(t)$ satisfies this condition from the same logic as in the matrix case, where one considers the operator on each piece. Hence, we now look at bounding the term $\|r_2(ad_{\Theta_1})(A(t))u_0\|_{L^2}$, following the process in [4].

We know from [5] that we may express the residual function r_2 in terms of its Fourier transform, as

$$r_2(ix) = \int_{\mathbb{R}} e^{i\xi x}\hat{r_2}(\xi)\,d\xi.$$

Hence, this can be used to give us the operator expression

$$r_2(ad_{\Theta_1})(A(t)) = \int_{\mathbb{R}} \hat{r_2}(\xi)e^{\xi\, ad_{\Theta_1}}(A(t))\,d\xi,$$

where symbolically we have taken $x = -i\,ad_{\Theta_1}$, and formally used the functional calculus as this is a self adjoint operator under our assumptions. Now we would like to use the following Lie group result,

$$e^{s\,ad_X}(Y) = e^{sX}Ye^{-sX}.$$

This is proven for matrices in [14], and we briefly prove this for unbounded operators.

We assume that the operators X, Y have the same domain, and that $-iX$ is self adjoint. Fixing some u in the domain we define $v(s) = e^{sX}Ye^{-sX}u$. Now using Lemma 3.3.9, and writing $e^{\pm sX} = e^{\pm is(-iX)}$, we see that

$$\frac{dv}{ds} = \left(Xe^{sX}Ye^{-sX} - e^{sX}Ye^{-sX}X\right)u = \left[X, e^{sX}Ye^{-sX}\right]u.$$

We note that $v(0) = Yu$, and thus by uniqueness of solutions to this problem we see that we must have $v(s) = e^{s\,ad_X}(Y)u$. ■

This is sufficient for our needs with $X = \Theta_1(t)$ and $Y = A(t)$ for a fixed t. Moreover, we note that the operators $e^{\pm\xi\Theta_1}$ are bounded from results for the functional calculus of $-i\Theta_1$. All of this allows us to see that

$$r_2(ad_{\Theta_1})(A(t))u = \int_{\mathbb{R}} \hat{r_2}(\xi)e^{\xi\Theta_1}A(t)e^{-\xi\Theta_1}u\,d\xi,$$

provided that u is in the domain of $A(t)$, i.e. D_Δ.

This puts us in a position so that we continue as in [4] obtaining

$$\|r_2(ad_{\Theta_1})(A(t))u(t)\|_{L^2} \leq \|\hat{r_2}\|_{L^1} \sup_{\xi\in\mathbb{R}} \|ad_{\Theta_1}(A(t))e^{-\xi\Theta_1}u(t)\|_{L^2}. \tag{4.6}$$

Now we require some further assumptions to proceed. We assume that the exact solution is such that $e^{-\xi\Theta_1}u(t) \in H^2(\mathbb{R}^d;\mathbb{C})$ for $\xi \in \mathbb{R}$, and that the potential function is such that $V \in H^2(\mathbb{R}^d;\mathbb{C})$. For ease of notation we define $v_\xi(t) := e^{-\xi\Theta_1}u(t)$. We now separate into three cases depending on t to obtain a bound on the above.

- **Case 1**: $0 \leq t < \frac{h}{2}$

  In this case we find that $ad_{\Theta_1}(A(t)) = (-4it\Delta^2 + 4it\Delta^2) = 0$.

- **Case 2**: $\frac{h}{2} \leq t \leq h$

  This is the only non-trivial case, and it is straightforward to see that

  $$ad_{\Theta_1}(A(t))v_\xi = (2h\Delta V - 2hV\Delta)\,v_\xi.$$

  This motivates the above assumptions, as they are sufficient so that we obtain the bound

  $$\|ad_{\Theta_1}(A(t))v_\xi(t)\| \leq Ch,$$

  for some constant C dependent on various Sobolev norms of V and v_ξ.

- **Case 3**: $t \notin [0, h]$.

  This case requires no computation as we know that $A(t) = 0$ and thus we must have $ad_{\Theta_1}(A(t))v_\xi = 0$.

Hence we obtain a bound $\|ad_{\Theta_1}(A(t))v_\xi(t)\| \leq Ch$ in all three cases. Thus, we may use the bounds (4.6) and (4.4) to obtain

$$\|\tilde{u}(h) - u(h)\|_{L^2} = \|\tilde{\epsilon}(h)\|_{L^2} \leq \tilde{C}h^2.$$

From this result, we can bound the error at an arbitrary timestep as before. We define $\tilde{u}_n = (\exp(\Theta_1(h)))^n u_0$, and $u_n = (\exp(-ihV)\exp(ih\Delta))^n u_0$. Proceeding as we did for matrices, one obtains the intermediate bound

$$\|\tilde{u}_n - u_n\|_{L^2} \leq \tilde{C}h^2 + \|\tilde{u}_{n-1} - u_{n-1}\|_{L^2},$$

and iterating this over each timestep and noting $\tilde{u}_0 = u_0$, one obtains

$$\|\tilde{u}_n - u_n\|_{L^2} \leq \tilde{C}t_n h,$$

as we would expect from the Lie-Trotter splitting.

Moreover, as we observed in the matrix case, this proof does not explicitly use the form of the Lie-Trotter splitting. As before, we require are that the coefficients give rise to a consistent splitting, i.e. one of the form $e^{-a_1 ihV}e^{b_1 ih\Delta}...e^{-a_m ihV}e^{b_m ih\Delta}$ such that

$$\sum_{k=1}^{m} a_k = \sum_{k=1}^{m} b_k = 1.$$

Notice however, that now we require that our splitting begins with a term involving Δ, i.e. $b_m \neq 0$, due to domain considerations. Furthermore, when one introduces a more complicated splitting of this form one doesn't require any further assumptions on $u(t)$ and V, as only the coefficients change, not the operators involved in expressing $ad_{\Theta_1}(A(t))v_\xi$. One could try to further this technique to obtain an energy result, as we did for a matrix Hamiltonian. This is not done here, due to time constraints.

We end this chapter, and report, by outlining some possible future work. First, and foremost, the omitted energy conservation in the above method should be addressed. This may prove difficult, as since the operators involved are unbounded, we cannot use the same approach we did for matrices, as the Magnus expansion will not converge for all timesteps, and so one cannot define $\Theta(h)$. Similarly, it should be investigated whether one can consider a higher order truncation of the Magnus expansion, and further assumptions, to obtain higher order bounds on splitting methods using this technique. This is particularly interesting as it is known that the Strang splitting, (1.4), is a second order method (see [6]), but this technique does not indicate this behaviour. Moreover, another extension of this technique is to see whether the condition that the splitting begins with a term in Δ is required. It should be the case that one can be careful enough with the domain consideration to obtain the same result for the Lie-Trotter splitting as $e^{ih\Delta}e^{-ihV}$.

Bibliography

[1] Hairer, E., Lubich, C. and Wanner, G., 2004. *Geometric Numerical Integration, Structure Preserving Algorithms, for Ordinary Differential Equations*. 2nd ed. Heidelberg : Springer.

[2] Faou, E., 2012 *Geometric Numerical Integration and Schrödinger Equations*. Zurich : European Mathematical Society.

[3] Lubich, C., 2008 *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*. Zurich : European Mathematical Society.

[4] Singh, P., 2018. *High accuracy computational methods for the semiclassical Schrödinger equation*. Thesis (PhD). King's College, University of Cambridge.

[5] Hockbruck, M. and Lubich, C., 2004. *On Magnus integrators for time-dependent Schrödinger equations*. SIAM Journal on Numerical Analysis , 2004, 41(3) (2004), pp. 945-963.

[6] Jahnke, T. and Lubich, C., 2000. *Error bounds for exponential operator splittings*. BIT, 2000, 40(3), pp. 735-744.

[7] Hall, B.C., 2013. *Quantum Theory for Mathematicians*. New York : Springer.

[8] Rudin, W., 1987. *Real and Complex Analysis*. 3rd ed. Singapore : McGraw - Hill Book Company.

[9] Lax, P.D., 2002. *Functional Analysis*. New York : John Wiley & Sons, Inc.

[10] Blanes, S., Casas, F. and Murua, A., 2006. *Symplectic splitting operator methods for the time-dependent Schrödinger equation*. The Journal of Chemical Physics, 124 (234105).

[11] Hairer, E., Lubich, C. and Wanner, G., 2003. *Geometric numerical integration illustrated by the Störmer–Verlet method*. Acta Numerica, pp. 399-450.

[12] Hairer, E., 2006. Long-time energy conservation of numerical integrators. In: L.M. Pardo, ed. *Foundations of Computational Mathematics, Santander 2005*. Cambridge: Cambridge University Press, pp. 162-180.

[13] Blanes, S. and Casas, F., 2003. *On the convergence and optimization of the Baker-Campbell-Hausdorff formula.* Linear Algebra and its Applications, 378(2004), pp.135-158.

[14] Hall, B.C., 2015. *Lie Groups, Lie Algebras, and Representations.* 2nd ed. New York: Springer.

[15] Iserles, A., Munthe-Kaas, H.Z., Norsett,. S.P. and Zanna, A., 2000. *Lie group methods.* Acta Numerica, pp. 215 - 365.

[16] Iserles, A., Norsett,. S.P. and Rasmussen, A.F., 2001. *Time symmetry and high-order Magnus methods.* Applied Numerical Mathematics, 39, pp. 379–401.

[17] Blanes, S., Casas, F., Oteo, J.A. and Ros. J., 2009. *The Magnus expansion and some of its applications.* Physics Reports, 470, pp. 151-238.

[18] Ito, T., and Tanikawa, K., 2002 *Long-term integrations and stability of planetary orbits in our Solar system.* Monthly Notices of the Royal Astronomical Society, 336 (2), pp. 483-500.

[19] Moler, C. and Van Loan, C., 2003. *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later.* SIAM Review, 45 (1), pp. 3-000.

[20] Casas, F., 2007. *Sufficient conditions for the convergence of the Magnus expansion.* Journal of Physics A: Mathematical and Theoretical, 40 (50), 15001.

[21] Moan, P.C., Niesen, J., 2008. *Convergence of the Magnus expansion.* Foundations of Computational Mathematics, 8(3), pp. 291-301.